mathematics

Quantitative Methods for Economics and Finance

Edited by
J.E. Trinidad-Segovia and Miguel Ángel Sánchez-Granero

Printed Edition of the Special Issue Published in *Mathematics*

www.mdpi.com/journal/mathematics

MDPI

Quantitative Methods for Economics and Finance

Quantitative Methods for Economics and Finance

Editors

J.E. Trinidad-Segovia
Miguel Ángel Sánchez-Granero

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin

Editors

J.E. Trinidad-Segovia
University of Almería
Spain

Miguel Ángel Sánchez-Granero
University of Almería
Spain

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Mathematics* (ISSN 2227-7390) (available at: https://www.mdpi.com/journal/mathematics/special_issues/Quantitative_Methods_Economics_Finance_2020).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-0365-0196-3 (Hbk)
ISBN 978-3-0365-0197-0 (PDF)

Cover image courtesy of Miguel A. Sánchez Granero.

Contents

About the Editors

J.E. Trinidad-Segovia (Ph .D.) is an Associate Professor of Finance at the Department of Economics and Business, University of Almería, Spain. His research interests include financial modeling, portfolio selection, CAPM, and applications of statistical mechanics to financial markets. He has published over 30 papers in peer-reviewed journals.

Miguel Ángel Sánchez-Granero (Ph. D.) is an Associate Professor at the Department of Mathematics, University of Almería, Spain. He is the author of several publications in international journals on asymmetric topology, self-similarity, fractal structures, and fractal dimension with applications in financial series.

Preface to "Quantitative Methods for Economics and Finance"

Since the mid-twentieth century, it has been clear that the more classical mathematical models were not enough to explain the complexity of financial and economic series. Since then, the effort to develop new tools and mathematical models for their application to economics and finance has been remarkable. However, it is still necessary to continue developing new tools, as well as continue studying the latest tools developed for the study of the financial and economic series. These tools can come from techniques and models taken from physics or new branches of mathematics such as fractals and dynamical systems, or new statistical techniques such as big data.

This book is a collection of the 19 papers that appeared in the Special Issue "Quantitative Methods for Economics and Finance" for the journal Mathematics. The purpose of this Special Issue is to gather a collection of articles reflecting the latest developments in different fields of economics and finance where mathematics plays an important role.

Khan et al. reexamine the relationship between policy uncertainty and stock prices in the United States, using a dynamically simulated autoregressive distributed lag setting.

Bengoa et al. use structural gravity model theory to study if the co-existence of bilateral investment treaties and two major regional agreements exert any effects on foreign direct investment in eleven Latin American countries.

Martín Cervantes et al. apply SRA copulas to analyze the relationship between price and trading volume of four stock market indexes.

Kim et al. use copula models to study the relationship between Bitcoin, gold, and the S&P500 index.

Kao et al. propose a forecasting framework based on the ensemble empirical mode decomposition, and hybrid models including autoregressive integrated moving average, support vector regression, and the genetic algorithm, to predict the primary energy consumption of an economy.

Schneider et al. propose a dispersion trading strategy based on a statistical stock selection process, which determines appropriate subset weights by exploiting a principal component analysis to specify the individual index explanatory power of each stock.

Díaz et al. study efficiency scores of pharmaceutical firms based on non-parametric data envelopment analysis techniques.

Isohätälä et al. study the impact of financing constraints on the relationship between net worth and investment.

Yan et al. propose a framework of a financial early warning system for listed companies in China, combining the unconstrained distributed lag model and widely used financial distress prediction models such as the logistic model and the support vector machine.

Lamothe-Fernández et al. compare deep learning methodologies for forecasting Bitcoin price and introduce a new prediction model.

Nikolova et al. provide a methodology, based on the Hurst exponent, to calculate the probability of volatility clusters and apply it to different assets including stocks, indexes, forex, and cryptocurrencies.

García Mirantes et al. analyze how many factors (including short and long-term components) should be considered when modeling the risk-management of energy derivatives.

Salmerón Gómez et al. analyze the detection of near-multicollinearity in a multiple linear regression from auxiliary centered and non-centered regressions.

Naimy et al. analyze the volatility dynamics in the financial markets of the United States, Russia, and China during their intervention in the Syrian war.

Cruz Rambaud et al. introduce a novel concept of an abstract derivative with applications in intertemporal choice when trying to characterize moderately and strongly decreasing impatience.

Salmerón Gómez et al. extends the concept of variance inflation factor to be applied in a raise regression in a model that presents collinearity.

Cruz Rambaud et al. study the relationship between the Time Trade-Off, Expected Utility, and Discounted Utility models.

Ramos-Requena et al. introduce different models to calculate the amount of money to be allocated in each stock in a pairs trading strategy.

Druică et al. study the relationship between academic dishonesty and dishonest and fraudulent behavior, such as tax evasion, social insurance fraud, and piracy.

J.E. Trinidad-Segovia, Miguel Ángel Sánchez-Granero
Editors

mathematics

MDPI

Article

US Policy Uncertainty and Stock Market Nexus Revisited through Dynamic ARDL Simulation and Threshold Modelling

Muhammad Asif Khan [1,*], Masood Ahmed [1,2,*], József Popp [3,4] and Judit Oláh [4,5]

1 Faculty of Management Sciences, University of Kotli, Azad Jammu and Kashmir, Kotli 11100, Pakistan
2 Lee Kuan Yew School of Public Policy, National University of Singapore, Kent Ridge, 469C Bukit Timah Road, Singapore 259772, Singapore
3 Faculty of Economics and Social Sciences, Szent István University, 2100 Gödölő, Hungary; Popp.Jozsef@szie.hu
4 TRADE Research Entity, Faculty of Economic and Management Sciences, North-West University, Vanderbijlpark 1900, South Africa; olah.judit@econ.unideb.hu
5 Faculty of Economics and Business, University of Debrecen, 4032 Debrecen, Hungary
* Correspondence: khanasif82@uokajk.edu.pk (M.A.K.); masoodahmed@u.nus.edu (M.A.)

Received: 30 October 2020; Accepted: 16 November 2020; Published: 20 November 2020

Abstract: Since the introduction of the measure of economic policy uncertainty, businesses, policymakers, and academic scholars closely monitor its momentum due to expected economic implications. The US is the world's top-ranked equity market by size, and prior literature on policy uncertainty and stock prices for the US is conflicting. In this study, we reexamine the policy uncertainty and stock price nexus from the US perspective, using a novel dynamically simulated autoregressive distributed lag setting introduced in 2018, which appears superior to traditional models. The empirical findings document a negative response of stock prices to 10% positive/negative shock in policy uncertainty in the short-run, while in the long-run, an increase in policy uncertainty by 10% reduces the stock prices, which increases in response to a decrease with the same magnitude. Moreover, we empirically identified two significant thresholds: (1) policy score of 4.89 (original score 132.39), which negatively explain stock prices with high magnitude, and (2) policy score 4.48 (original score 87.98), which explains stock prices negatively with a relatively low magnitude, and interestingly, policy changes below the second threshold become irrelevant to explain stock prices in the United States. It is worth noting that all indices are not equally exposed to unfavorable policy changes. The overall findings are robust to the alternative measures of policy uncertainty and stock prices and offer useful policy input. The limitations of the study and future line of research are also highlighted. All in all, the policy uncertainty is an indicator that shall remain ever-important due to its nature and implication on the various sectors of the economy (the equity market in particular).

Keywords: policy uncertainty; stock prices; dynamically simulated autoregressive distributed lag (DYS-ARDL); threshold regression; United States

1. Introduction

The field of mathematical finance is one of the most rapidly emerging domains in the subject of finance. Dynamically Simulated Autoregressive Distributed Lag (DYS-ARDL) [1] is an influential tool that may help an investor analyze and benefit by understanding the positive and negative shocks in policy indicators. This strategy enables investors to observe the reaction of equity prices to positive and negative shocks of various magnitude (1%, 5%, 10%, and others). More importantly, it may assist the diversification of potential portfolios across various equities based on predicted reaction. Coupled with DYS-ARDL, the few other effective strategies include statistical arbitrage strategies (SAS) and

pairs trading strategy (PTS) that are empirically executed in mathematical finance literature; see for example [1–6]. Stübinger and Endres [2] developed and applied PTS to minute-by-minute data of oil companies constituting the S&P 500 market index for the US and revealed that the statistical arbitrage strategy enables intraday and overnight trading. Similarly, Stübinger, Mangold, and Krauss [6] developed SAS (based on vine copulas), which is a highly flexible instrument with multivariate dependence modeling under the linear and nonlinear setting. The authors find it promising in the context of the S&P 500 index of the United States (US) equities. Using SAS, Avellaneda and Lee [3] related the performance of mean-reversion SAS with the stock market cycle and found it effective in studying stock performance during the liquidity crisis. Empirical evidence from the US equity market on PTS links trading cost documents that PTS is profitable among well-matched portfolios [5]. Liu, Chang, and Geman [4] argue that PTS can facilitate stakeholders to capture inefficiencies in the local equity market using daily data. Interestingly, we find that SAS and PTS strategies are successfully employed in the context of the US, while to the best of our knowledge, we do not find the use of DYS-ARDL, which is surprising. Each of the described strategies has unique features in a given scenario in which they are used, yet it worthwhile to add little value to mathematical finance literature by empirically examining the DYS-ARDL specification in the US context.

Given the economic implications of financial markets and eventual behavior [7–10], this piece of research empirically examines the short- and long-run impacts of policy uncertainty (hereafter PU) on stock prices of the US using a novel DYS-ARDL setting proposed by Jordan and Philips [1] and of threshold relation using the Tong [11] model. The study is motivated by conflicting literature on policy risk stock price and shortcomings associated with the traditional cointegration model (e.g., Autoregressive Distributed Lag (ARDL)).

Since the introduction of the measure of PU by Baker et al. [12], the effects of the PU on macro variables have gained substantial attention. PU is closely monitored and analyzed by businesses, policymakers, and academic scholars, as the global economy is now more closely interconnected than ever [13]. Intuitively, an increase in PU is expected to negatively influence the stock market, while on the contrary, stock market indicators may react positively to a decline in PU [14]. This intuition is consistent with the findings of Baker, Bloom, and Davis [12], who have shown its adverse effects on economic activities, which is confirmed by the recent literature [14–27].

According to Baker, Bloom, and Davis [12], economic PU refers to "*a non-zero probability of changes in the existing economic policies that determine the rules of the game for economic agents*". The impact of changes in PU may potentially rout the following channels:

- First, it can change or delay important decisions made by companies and other economic actors, such as investment [28], employment, consumption, and savings decisions [14,29].
- Second, it increases financing and production costs by affecting supply and demand channels, exacerbating investment decline, and economic contraction [14,17,30,31].
- Third, it can increase the risks in financial markets, especially by reducing the value of government protection provided to the market [17].
- Lastly, PU also affects inflation, interest rates, and expected risk premiums [32,33].

Importantly, in the context of the US, the phenomena are also captured by a few studies [14,21–23,27] with conflicting findings. For example, some of them [14,15,21,23,34] found a negative relationship, while others reported no effect [22,27]. The conflicting referred literature on the US [14,15] relies on the classical approach [35] to capture the cointegration relationship. From the symmetry assumption perspective drawn on this approach [35], it follows that an increase in PU will negatively affect the other macroeconomic variables and that a decrease in PU will increase this variable. However, this may not be the case, as investors' responses may differ from increasing PU versus decreasing PU. It is possible that, due to an increase in uncertainty, investors move their equity assets to safer assets and that a decrease in uncertainty may cause them to shift their portfolio towards the stock market (assume

the change in PU is less than increase) if they expect that a decrease in uncertainty is short-lived and then that asymmetry originates.

The shortcomings associated with Pesaran, Shin, and Smith [35] are, to some extent, addressed by nonlinear extension by Shin et al. [36], which generates two separate series (positive and negative) from the core explanatory variable. Thus, the asymmetric impact may be estimated; however, this approach overlooks the simulation features while estimating the short- and long-run asymmetries. The package is given by Jordan and Philips [1], known as the DYS-ARDL approach, which takes into account the simulation mechanism and liberty to use positive and negative shocks in an explanatory variable and captures the impact in a variable of interest. According to recent literature [37,38], this novel approach is capable of predicting the actual positive and negative changes in the explanatory variable and its subsequent impact on the dependent variable. Moreover, it can stimulate, estimate, and automatically predict, and graph said changes. The authors also believe that classical ARDL can only estimate the long-term and short-term relationships of the variables. Contemplating the limitations associated with traditional estimators, this study uses Jordan and Philips [1] inspirational DYS-ARDL estimator to examine the relationship between PU and US stock prices.

In addition, this study extends the analysis beyond the DYS-ARDL estimator [1] by using Tong [11] threshold regression Although DYS-ARDL [1] is a powerful tool to capture the dynamic cointegration between an independent variable and dependent variable, and its unique feature automatically generates the simulation-based graph of changes to SP as a result of a certain positive/negative shock in PU, it is beyond its capacity to figure out a certain level (point) where the relationship (magnitude of coefficient) changes. For example, literature shows that the general stock market is linearity correlated with the changes in PU [14–27]. An increase in PU brings a negative influence on SP, in which a decrease translates into a positive change.

Threshold models have recently paid attention to modeling nonlinear behavior in applied economics. Part of the interest in these models is in observable models, followed by many economic variables, such as asymmetrical adjustments to the equilibrium [39]. By reviewing a variety of literature, Hansen [40] recorded the impact of the Tong [11] threshold model on the field of econometrics and economics and praised Howell Tong's visionary innovation that greatly influenced the development of the field of econometrics and economics.

Concisely, this small piece of research extends the financial economics and mathematical finance literature on PU and SP in the context of the United States, which is the world's top-ranked equity market [41] in three distinct ways. First, the novelty stems from the use of DYS-ARDL [1], which produces efficient estimation using simulations mechanism (which traditional ARDL departs), and auto-predicts the relationship graphically alongside empirical mechanics. To the best of our knowledge, this is the first study to verify traditional estimation with this novel and robust method. The empirical findings of DYS-ARDL document a negative response of stock prices in the short-run for a 10% positive and negative change (shock) in PU, while a linear relationship is observed in case of the long-run in response to said change.

Second, coupled with novel DYS-ARDL, this study adds value to relevant literature by providing evidence from threshold regression [11], which provides two significant thresholds in the nexus of PU-SP that may offer useful insight into policy matters based upon identified threshold(s). It is worth noting that SP negatively reacts to PU until a certain level (threshold-1), where the magnitude of such reaction changes (declines) to another point (threshold-2) with relatively low magnitude (still negative). Interestingly, below threshold-2, the PU became irrelevant to the US-SP nexus.

Third, this is a compressive effort to provide a broader picture of the US stock market reaction to policy changes. In this regard, prior literature is confined to the New York Stock Exchange Composite Index and S&P 500, while this study empirically tested seven major stock indices: S&P 500, Dow Jones Industrial Average, Dow Jones Composite Average, NASDAQ composite, NASDAQ100, and Dow Jones Transpiration Average. Expending analysis of these indices potentially provides useful insights to investors and policymakers because all are not equally exposed to adverse changes in PU. Some of them

are nonresponsive to such changes, which may help a group of investors diversifying their investments to avoid unfavorable returns and to construct the desired portfolio with low risk. On the other hand, risk-seeking investors may capitalize on risk premiums, where understanding the identified thresholds may help to diversify their investments reasonably.

The rest of the work is organized as follows. Section 2 outlines the related literature; Section 3 illustrates the material and methods; results and discussion are covered by Section 4; the study is concluded in Section 5.

2. Literature Review

Bahmani-Oskooee and Saha [15] assessed the impact of PU on stock prices in 13 countries, including the United States, and find that, in almost all 13 countries, increased uncertainty has negative short-term effects on stock prices but not in the long term. Sum [18] utilized the ordinary least squares method to analyze the impact of PU on stock markets (from January 1993 to April 2010) of Ukraine, Switzerland, Turkey, Norway, Russia, Croatia, and the European Union. The study finds that PU negatively impacts EU stock market returns, except for in Slovenia, Slovakia, Latvia, Malta, Lithuania, Estonia, and Bulgaria. The analysis does not identify any negative impact of stock market returns of non-EU countries included in the study. Sum [24] used a vector autoregressive model with Granger-causality testing and impulse response function and founds that PU negatively impacts stock market returns for most months from 1985 to 2011.

Another study [34] analyzed monthly data of PU and stock market indices of eleven economies, including China, Russia, the UK, Spain, France, India, Germany, the US, Canada, Japan, and Italy. The study found that PU negatively impacts stock prices mostly except periods of low-to-high frequency cycles. The study used data from 1998 to 2014. Using data from 1900 to 2014, Arouri, Estay, Rault, and Roubaud [14] measured PU's impact on the US stock market and found a weak but persistent negative impact of PU on stock market returns. Inflation, default spread, and variation in industrial production were the control variables used. The study also found that PU has a greater negative impact on stock market returns during high volatility.

Pastor and Veronesi [17] estimated how the government's economic policy announcement impacts stock market prices and reported that stock prices go up when the government makes policy announcements and that more unexpected announcement brings in greater volatility. Li, Balcilar, Gupta, and Chang [19] found a weak relationship between PU and stock market returns in China and India. For China, the study used monthly data from 1995 to 2013, and for India, it used monthly data from 2003 to 2013. The study employed two methods (i) bootstrap Granger full-sample causality testing and (ii) subsample rolling window estimation. The first method did not find any relationship between stock market returns and PU, while the second method showed a weak bidirectional relationship for many sub-periods. Employing the time-varying parameter factor-augmented vector autoregressive (VAR) model on data from January 1996 to December 2015, Gao, Zhu, O'Sullivan, and Sherman [20] estimated the impact of PU on the UK stock market returns. The study considered both domestic and international economic PU factors. The paper maintains that PU explains the cross-section of UK stock market returns.

Wu, Liu, and Hsueh [22] analyzed the relationship between PU and performance of the stock markets of Canada, Spain, the UK, France, Italy, China, India, the US, and Germany. Analyzing monthly data from January 2013 to December 2014, the study found that not all stock markets under investigation react similarly to PU. According to the study, the UK stock market falls most with negative PU, but the markets of Canada, the US, France, China, and Germany remain unaffected. Asgharian, Christiansen, and Hou [21] measured the relationship between PU and the US (S&P 500) and the UK (FTSE 100) stock markets. The study used daily data for stock market indices and monthly data for PU. The paper found that stock market volatility in the US depends on PU in the US and that stock market volatility in the UK depends on PU in both the US and UK.

Christou, Cunado, Gupta, and Hassapis [23] estimated the impact of PU on the stock markets of the US, China, Korea, Canada, Australia, and Japan. Using monthly data from 1998 to 2014 and employing a panel VAR model with impulse response function, the study found that own country PU impacts stock markets negatively in all aforementioned countries. The study also found that PU in the US also negatively impacts all other countries' stock markets in the analysis, except Australia. Debata and Mahakud [25] found a significant relationship between PU and stock market liquidity in India. The study used monthly data from January 2013 to Granger 2016 and employed VAR Granger causality testing, variance decomposition analysis, and impulse response function. The impulse response function showed that PU and stock market liquidity are negatively related.

Liu and Zhang [42] investigated PU's impact on stock market volatility of the S&P 500 index from January 1996 to June 2013. The study found that PU and stock market volatility are interconnected and that PU has significant predictive power on stock market volatility. Pirgaip [27] focused on the relationship between stock market volatility for fourteen OECD countries, subject to monthly data from March 2003 to April 2016 for Japan, France, Germany, Chile, Canada, Italy, Australia, the US, UK, Sweden, Spain, Netherlands, Australia, and South Korea. Employing the bootstrap panel Granger causality method, the study found that PU impacts stock prices in all countries except the US, Germany, and Japan.

Škrinjarić and Orlović [26] estimated the spillover effects of PU shocks on stock market returns and risk for nine Eastern and Central European countries, including Bulgaria, Estonia, Lithuania, Croatia, Slovenia, Hungary, Czech Republic, Poland, and Slovakia. The paper employed a rolling estimation of the VAR model and the spillover indices. The study's findings suggest that Poland, the Czech Republic, Slovenia, and Lithuania are more sensitive to PU shocks compared to other markets in the study. In contrast, the Bulgarian stock market is least impacted by PU shocks. Other countries' stock markets have an individual reaction to PU shocks.

Ehrmann and Fratzscher [43] examined how the US monetary policy shocks are transmitted stock market returns over February 1994 to December 2004, with a weak association in India, China, and Malaysia's stock markets while strong on Korea, Hong Kong, Turkey, Indonesia, Canada, Finland, Sweden, and Australia. Brogaard and Detzel [44] examined the relationship between PU and asset prices using a monthly Center for Research in Security Prices (CRSP) value-weighted index as the US stock market's performance measure and PU. The findings suggest that a one standard deviation increase in PU decreases stock returns by 1.31% and increases 3-month log excess returns by 1.53%. The study also found that dividend growth is not affected by PU. Antonakakis et al. [45] estimated co-movements between PU and the US stock market returns and stock market volatility using S&P 500 stock returns data and S&P 500 volatility index data. The study found a negative dynamic correlation between PU and stock returns except during the financial crisis of 2008, for which the correlation became positive.

Stock market volatility also negatively impacts the stock market returns, according to the study. Dakhlaoui and Aloui [46] scrutinized the relationship between the US PU and Brazil, Russia, India, and China stock markets, estimating daily data from July 1997 to July 2011. The study found a negative relationship between the US PU and the returns, but the volatility spillovers were found to oscillate between negative and positive making, it highly risky for investors to invest in US and BRIC stock markets simultaneously. Yang and Jiang [47] used data from the Shanghai stock index from January 1995 to December 2014 to investigate the relationship between PU and china stock market returns and suggest that stock market returns and PU are negatively correlated and that the negative impact of PU lasts for about eight months after the policy announcement.

Das and Kumar [16] estimated the impacts of domestic PU and the US PU on the economies of 17 countries. The analysis included monthly data from January 1998 to February 2017 and found that emerging markets are less prone and vulnerable to domestic and US PU than developed economies while Chile and Korea are relatively more sensitive to both Domestic PU and US PU, whereas China is least affected. Estimation reveals that except Canada and Australia, stock prices and all other developed

economies in the analysis are quite sensitive to US PU. Australia and Canada stock prices are more reliant on domestic PU. Stock prices of all the emerging economies are more reliant on domestic PU except for the marginal exception of Russia and Brazil.

We conclude that the reviewed literature on policy-stock prices is conflicting. See, for example, Bahmani-Oskooee and Saha [15]; Asgharian, Christiansen, and Hou [21]; Christou, Cunado, Gupta, and Hassapis [23]; Ko and Lee [34]; and Arouri, Estay, Rault, and Roubaud [14], who found that the US stock market is negatively correlated to changes in PU, and Wu, Liu, and Hsueh [22], and Pirgaip [27], who documented no effect of US. Sum [13] revealed a cointegration relationship that exists between the economic uncertainty of the US and Europe, showing a spillover effect across financial markets across the national borders. The literature referred to the US with few exceptions including Arouri, Estay, Rault, and Roubaud [14], and Bahmani-Oskooee and Saha [15], who assumed a linear relationship between PU and stock prices and relied on Pesaran, Shin, and Smith [35] for the traditional cointegration approach to finding the long-run dynamics of the PU and stock prices. Amongst these, Arouri, Estay, Rault, and Roubaud [14] found a long-run weak negative impact in general and persistent negative impact during high volatility regimes. However, Bahmani-Oskooee and Saha [15] found short-run negative impacts and no effect in the long-run.

The strand of literature relied on traditional cointegration [35] for modeling policy-stock price connection follows the symmetry assumption perspective holding that an increase in PU will negatively affect the other macroeconomic variable and a decrease in PU will increase this variable. However, this may not be the case, as investors' responses may differ from increasing PU versus decreasing PU. It is possible that, due to an increase in uncertainty, investors move their equity assets to safer assets and that a decrease in uncertainty may cause them to shift their portfolio towards the stock market (assume the change in PU is less than increase) if they expect that a decrease in uncertainty is short-lived and then that asymmetry originates. Figure 1 plots the theoretical framework based on reviewed papers [14,15].

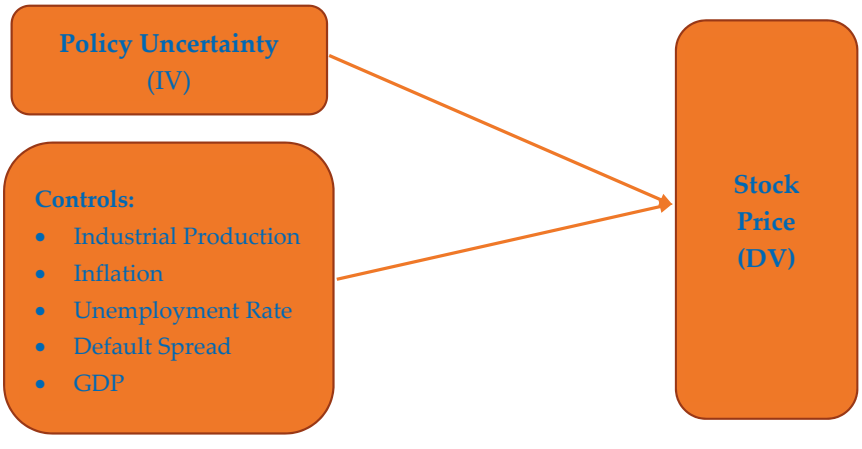


Figure 1. Theoretical framework, IV: independent variable (measured by news-based policy uncertainty—PU_NB) and DV: dependent variable. Source: Drawn from the literature (Arouri et al., [14], and Bahmani-Oskooee and Saha, [15]).

We conclude that empirical literature on PU and stock prices is conflicting, with no consensus on its empirical impact, as the literature shows mixed results (positive, negative, and no effect). This may be attributable to the differences in methodological strategies used, time coverage, and other controls used in the estimation process. Among empirical methods used, ARDL is commonly used to arrive at short- and long-run cointegration relationships. Moreover, it is surprising that threshold identification

in PU and stock price connection is an unaddressed phenomenon. Thus, it is imperative to go ahead and comprehensively examine the short- and the long-run association between PU and stock prices using an updated dataset coupled with DYS-ARDL and the threshold strategy in the context of the United States, the world top-ranked financial market (in terms of market size) [41].

3. Materials and Methods

3.1. Description of Variables and Data Source

The independent variable is economic PU, which is an index built by Baker, Bloom, and Davis [12] from three types of underlying components. The first component quantifies the news coverage of policy-related economic uncertainties. A second component reflects the number of federal tax legislation provisions that will expire in the coming years. The third component uses the disagreement among economic meteorologists as an indicator of uncertainty. The first component is an index of the search results of 10 major newspapers, which includes *US Today, Miami Herald, Chicago Tribune, Washington Post, Los Angeles Times, Boston Globe, San Francisco Chronicle, Dallas Morning News, New York Times*, and *Wall Street Diary*. From these documents, Baker, Bloom, and Davis [12] create a normalized index of the volume of news articles discussing PU. The second component of this index is based on reports from the congressional budget office, which compiles lists of temporary provisions of federal tax legislation. We compile an annual number of weighted US dollar tax laws that expire over the next 10 years and provide a measure of uncertainty on which path federal tax laws will follow in the future.

The third component of the PU index is based on the Federal Reserve Bank of Philadelphia's survey of professional meteorologists. Here, it uses the dispersion between individual meteorologists' forecasts of future consumer price index levels, federal spending, and state and local spending to create indices of uncertainty for policy-related macroeconomic variables. This study uses news-based PU (PU_NB) for the main analysis, while robustness is performed using three component-based PU index (PU_3C). Figure 2 provides a glimpse of historical US monthly PU indices (both PU_NB and PU_3C). A rise in PU indices is shown by the second half of the sample period. We may attribute it to a series of incidents, such as the 9/11 attack on the World Trade Centre, followed by US coalition attack on Afghanistan in 2001, mounted tension between the US and North Korea in 2003, and the US attack on Iraq in 2003. Later, the major event was the 2007–2009 recession in the form of a US economic crash caused by the subprime crisis in 2008, and aftershocks in subsequent years have significantly raised the PU. The efforts to reduce carbon emission have caused an economic downturn in 2010–2011, US economic slowdown was heavily weighted by the global economic slowdown in 2015–2016, and finally, the sizable swaths of US economic shutdown were a result of COVID-19.

This study utilizes monthly data ranging from January 1985 to August 2020. For this period, data on PU is downloaded from the economic PU website (http://www.policyuncertainty.com/), which is open-source and commonly used in related literature, while data on seven US stock market indices are accessed from Yahoo Finance [48]. This includes monthly adjusted closing stock prices of the New York Stock exchange composite index (NYSEC) as a dependent variable for baseline analysis; however, the same measure of S&P500, Dow Jones Industrial Average (DJI), Dow Jones Composite Average (DJA), NASDAQ composite, NASDAQ100, and Dow Jones Transpiration Average (DJT) are utilized as robustness. Following Arouri, Estay, Rault, and Roubaud [14], and Bahmani-Oskooee and Saha [15], we consider Industrial production (IP), default spread (DS), inflation (INF), and unemployment rate (UE) as potential controls, and monthly data are sourced from Federal Reserve Economic Data (FRED https://fred.stlouisfed.org/) [49]. FRED is an open-source database that maintains various frequency datasets on more than 0.07 million in the United States and international time-series data from above 100 date sources.

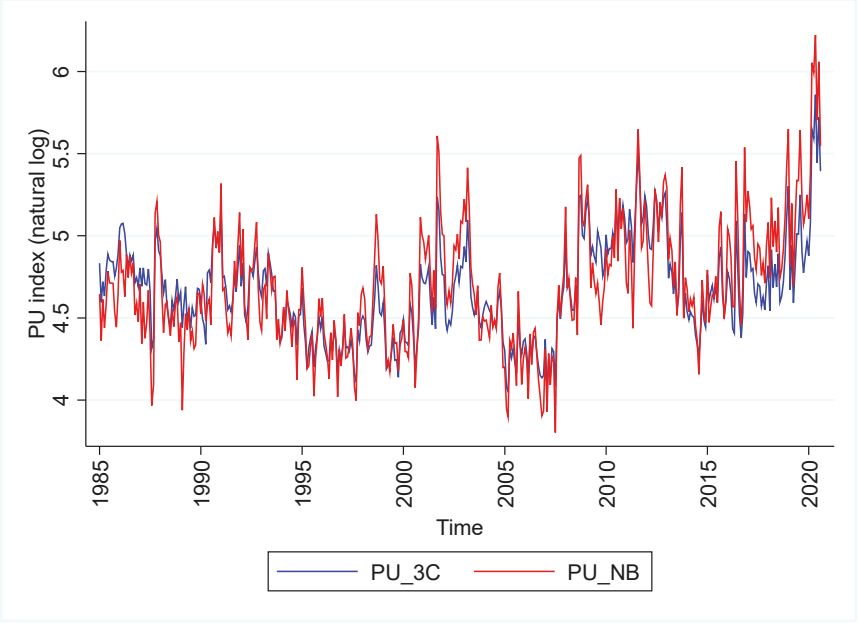


Figure 2. US monthly policy uncertainty index. Source: Baker, Bloom, and Davis [12].

The choice of the US's stock market as a potential unit for analysis is led by its top rank in terms of size among world exchanges [41]. Figure 3 shows a list of the world's largest stock exchanges as per the market capitalization of listed companies. The New York stock exchange and NASDAQ are ranked first and second with 25.53 and 11.23 trillion US dollars, respectively, among world exchanges [41]. Table 1 shows the descriptive properties of the underlying variables.

Figure 3. World largest stock exchanges by market. Source: Szmigiera [41].

Table 1. Descriptive Statistics.

Abbreviation	Description	Measurement	Mean	Std.Dev.	Min	Max	Data Source
SP	Stock price	Natural log of adjusted closing prices of NYSE composite index	8.538	0.702	7.000	9.541	Yahoo Finance https://finance.yahoo.com/
PU_NB	News-based policy uncertainty	Natural log of news-based policy uncertainty index	4.692	0.378	3.802	6.223	http://www.policyuncertainty.com/
INF	Inflation	Natural log	5.181	0.260	4.659	5.560	
IP	Industrial production	Natural log	4.439	0.248	1.716	4.705	FRED https://fred.stlouisfed.org/
UE	Unemployment rate	%	5.947	1.647	3.500	14.700	
DS	Default spread	Moody's Seasoned Baa Corporate Bond Minus Federal Funds Rate	3.847	1.596	0.500	8.820	

Jordan and Philips's [1] DYS-ARDL can be expressed in the following standard pathway:

$$\Delta(y)_t = \alpha_0 + \theta_0(y)_{t-1} + \theta_1(x_1)_{t-1} + \ldots + \theta_k(x_k)_{t-1} + \sum_{i=1}^{p} \alpha_{0i}\Delta(y)_{t-1} + \sum_{j=0}^{q1} \beta_{1j}\Delta(x_1)_{t-j} + \ldots + \sum_{j=0}^{qk} \beta_{kj}\Delta(x_k)_{t-j} + \varepsilon_t \quad (1)$$

where the change in the dependent variable (y) is a function of the intercept (α_0), and all the independent variables at time $t-1$ are the levels of the maximum of p and qk lags in respective first difference (Δ) along with error term (ε) at time t. The study uses Pesaran, Shin, and Smith's [35] ARDL bounds testing approach for a level relationship using Kwiatkowski–Phillips–Schmidt–Shin (2018) critical values as a benchmark. The null hypothesis for no level relationship obtained by joint F-statistics from the estimation is rejected against the critical bounds, in particular, when estimated F-statistics is greater than the upper bound I(1). Drawn on the empirical specification illustrated in Equation (1), the error correction transformation of the ARDL bounds estimators are estimated under the following:

The change in the dependent variable (y) is a function of intercept (α_0), and all the independent variables at time $t-1$ are levels to the maximum of p and qk lags in respective first difference (Δ) along with error term (ε) at time t. The study uses Pesaran, Shin, and Smith's [35] ARDL bounds testing approach for a level relationship using Kwiatkowski–Phillips–Schmidt–Shin (2018) critical values as a benchmark. The null hypothesis for no level relationship obtained by joint F-statistics from the estimation is rejected against the critical bounds, in particular, when estimated F-statistics is greater than the upper bound I(1). Drawn on the empirical specification illustrated in Equation (1), the error correction transformation of the ARDL bounds estimators are estimated as under the following:

$$\begin{aligned}\Delta ln(SP)_t = \alpha_0 \quad & +\theta_0\, ln\,(SP)_{t-1} + \beta_1\Delta\, ln\,(PU\_NB)_t + \theta_1\, ln\,(PU\_NB)_{t-1} + \beta_2\Delta\, ln\,(INF)_t \\ & +\theta_2\, ln\,(INF)_{t-1} + \beta_3\Delta\, ln\,(IP)_t + \theta_3\, ln\,(IP)_{t-1} + \beta_4\Delta\,(UE)_t + \theta_4(UE)_{t-1} \\ & +\beta_5\Delta\,(DS)_t + \theta_5(DS)_{t-1} + \varepsilon_t\end{aligned} \quad (2)$$

In Equation (2), θ_0 denotes error correction term (ECT), $\beta_1 - \beta_5$ capture the short-tun coefficient, and $\theta_1 - \theta_5$ indicate a long-run coefficient for each of the regressors respectively.

3.2. Threshold Regression

The DYS-ARDL [1] is a powerful tool for capturing the dynamic cointegration between the independent variable and the dependent variable. Its unique function is to simulate the automatic generation of graphs based on changes in the dependent variable as a result of a certain positive/negative shock in the explanatory variable. However, any particular degree (point) of change in the relationship cannot be imagined. For example, the literature shows that linearity on the general stock market is related to changes in PU. The increase in PU harms the stock market, while the decrease in PU leads to positive changes.

The threshold model extends linear regression to allow the coefficients to vary among regions/regimes. These regions are identified by threshold variables that are greater or less than

the threshold value. The model can have multiple thresholds; you can specify a known number of thresholds, or you can allow it to use Bayesian Information Criteria (BIC), Akaike Information Criteria (AIC), or Hannan–Quinn Information Criteria (HQIC to determine the number for you). It includes region-varying coefficients for specified covariates for each identified threshold, and it is efficient in automatically estimating the possible thresholds using *the n thresholds(#)* function. Moreover, it creates a variable with a sum of squared residuals for each tentative threshold. Thus, single threshold regression [11] is modeled by Equation (3).

The threshold model provides a systematic method of tracking the turning point in the relationship that can help decision-makers make better decisions [50–53]. Therefore, the threshold regression model for a single threshold [11] is modeled by Equation (3) for regions defined by threshold γ.

$$\begin{aligned} y_t &= X_t\beta + Z_t\delta_1 + \varepsilon_t \; if \; w_t \; \leq \; \gamma \\ y_t &= X_t\beta + Z_t\delta_2 + \varepsilon_t \; if \; \gamma < \; w_t \end{aligned} \tag{3}$$

where y_t is a dependent variable (SP in our case), X_t represents the vector consisting of region-invariant parameters (INF, IP, UE, and DS), Z_t is a vector of exogenous variables with region-specific coefficient vectors (δ_1 *and* δ_2), and w_t is the threshold variable, PU_NB (that may be one of the variables in X_t or Z_t.

4. Results

4.1. Unit-Root Analysis

The preliminary step to test the level relationship between the dependent variable (SP) and respective regressors is to satisfy the stationarity condition of the individual series; in particular, the dependent variables must be integrated at the first difference, I(1). Furthermore, all the independent variables must not be stationary at the second difference, I(2). To determine the integration order aligned with recent literature [37], this study uses the Augmented Dickey–Fuller (ADF) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test [54]. The null hypothesis for ADF assumes unit-root, while Baum [55] STATA module for KPSS, on the other hand, is tested under the null hypothesis of stationarity. The results are reported in Table 2, which reveals that the null hypothesis for most of the underlying variables cannot be rejected at level, which is rejected at first difference. Both tests (ADF and KPSS) witness that the dependent variable is stationary at the first difference and that none of the regressors is integrated at second order. The situation calls for the potential use of ARDL bounds testing for cointegration. After determining the stationarity conditions, the selection of optimal lag is another essential and challenging task, for which there is no hard and fast rule. However, following Bahmani-Oskooee and Saha [15], and Bahmani-Oskooee and Saha [56], we allow 8–12 lags for monthly data in this study. The lag-order is determined using AIC, SC, and HQ.

Table 2. Unit-root analysis.

Variable	Augmented Dickey–Fuller Test		Kwiatkowski–Phillips–Schmidt–Shin Test	
	Level	1st Difference	Level	1st Difference
SP	−2.450	−10.413 ***	0.425	0.037 ***
PU_NB	−2.676	−4.950 ***	0.336	0.011 ***
INF	−1.270	−10.619 ***	0.879	0.030 ***
IP	−3.513 **	−4.637 ***	0.141 **	0.097 ***
UE	−2.798	−12.726 ***	0.499	0.041 ***
DS	−3.477 ***	−8.201 ***	0.134 **	0.063 ***

***, and ** indicate that the null hypothesis of unit-root rejected at 1%, 5%, and 10% levels of significance, respectively.

4.2. Baseline Analysis—ARDL Bounds Test

The study estimates the baseline relationship between PU_NB and SP using ordinary least squared regression (OLS) in a series of estimation. Referring to Table A1 (Appendix A), initially in the model (1), PU_NB shows a positive impact on SP; however, when controls are introduced, the relationship

becomes negative and consistent across all estimators (2–5). This shows the relevance of controls, and the inclusion of each control has increased the R-squared.

Pesaran, Shin, and Smith;s [35] bounds test reveals the existence of a cointegration relationship between PU_NB and SP along with controls. Table 3 shows the ARDL bounds test and diagnostic testing for baseline estimation. Computed absolute F-statistics and t-statistics in Table 3 (upper part) are greater than upper bounds at all significance levels (10%, 5%, and 1%) of Kripfganz and Schneider (2018) critical values.

Table 3. ARDL bounds test and diagnostic testing.

Test Statistics	**Value**					
F-statistics	6.574					
t-statistics	−4.654					
Confidence interval	10%		5%		1%	
Bounds	I(0)	I(1)	I(0)	I(1)	I(0)	I(1)
F-stat *	2.130	3.240	2.459	3.640	3.155	4.470
t-stat *	−2.557	−4.065	−2.857	−4.399	−3.439	−5.023
Diagnostic testing						
Diagnostic test	Jarque-Bera		ARCH	Breusch-Pagan/Cook-Weisberg	Breusch-Godfrey LM	Ramsey RESET
p-value	0.188		0.101	0.102	0.418	0.452
Inference	Estimated residuals are normal		No heteroskedasticity problem		No serial correlation problem	Model correctly specified
Variance inflation factor						
Variable	INF	IP	UE	DS	PU_NB	Mean VIF
VIF	2.94	2.8	2.18	1.85	1.43	2.15

* indicates Kripfganz and Schneider (2018) critical values.

After establishing the ARDL bounds test, we proceed to check for diagnostic testing. Table 3 (middle part) shows the *p*-values for each of the estimated tests, which affirms that baseline, and ARDL estimation satisfies the diagnostic properties, such as normality of estimated residuals, heteroskedasticity, serial correlation, and correct specification of the model. The bottom part of Table 3 incorporates the multicollinearity results of the variance inflation factor (VIF). The benchmark to draw inference is the individual VIF value for each regressor not being greater than 5. In our case, none of the regressors violate these criteria, which signifies that the multicollinearity problem does not exist in our estimation and that the obtained results are correctly estimated.

4.3. Dynamic ARDL Simulations and Robustness

Table 4 documents the results of DYS-ARDL [1], using PU_NB as a baseline measure (Table 4, model-1) and PU_3C as its near alternative for robustness (Table 4, model-2); 5000 simulations are estimated for both models. Khan, Teng, and Khan [37], and Khan et al. [57] affirm that the novel DYS-ARDL model can stimulate, estimate, and graph to automatically predict the graphs of negative and positive changes occurring in variables and their short- and long-term relationships, which are beyond the capacity of classical ARDL. The significant and negative coefficient of error correction term (ECT) in each of the estimated model stratifies the existence of a cointegration relationship between variables under consideration. The system corrects the previous period disequilibrium at a monthly rate of 6.7%. The negative and significant PU_NB coefficient illustrates that equity markets in the US do not like a mounting risk in the form of PU.

Table 4. Dynamically simulated autoregressive distributed lag (DYS-ARDL) results.

	(1)	(2)
Variables	**SP**	**SP**
	PU_NB	**Robustness: PU_3C**
ECT(−1)	−0.067 ***	−0.072 ***
	(0.021)	(0.021)
ΔPU	−0.044 ***	−0.068 ***
	(0.008)	(0.012)
PU	−0.014 **	−0.024 **
	(0.007)	(0.010)
ΔINF	−0.096	−0.224
	(0.651)	(0.649)
INF	0.088 **	0.092 **
	(0.035)	(0.043)
ΔIP	−0.008	−0.010
	(0.015)	(0.015)
IP	0.128 **	0.129 **
	(0.056)	(0.055)
ΔUE	0.008 **	0.008 **
	(0.004)	(0.004)
UE	0.005 **	0.006 ***
	(0.002)	(0.002)
ΔDS	−0.030 ***	−0.029 ***
	(0.007)	(0.007)
DS	−0.007 ***	−0.007 ***
	(0.002)	(0.002)
Constant	−0.368 ***	−0.360 ***
	(0.134)	(0.134)
Obs.	427	427
F-statistics	10.95 ***	10.21 ***
Simulations	5000	5000

Standard errors are in parenthesis. *** $p < 0.01$, and ** $p < 0.05$. All the abbreviations are defined in Table 1. SP refers to the NYSE composite index.

This study performed several robustness checks using an alternative measure of PU_NB and SP. A three component-based measure of PU is used as an alternative measure (it is explained in detail in the variable description section of the Material and Methods section). Table 4 (model-2) incorporates the results of DYS-ARDL, where the dependent variable is SP from the NYSEC index and the independent variable is PU_3C. The results are consistent with the main analysis. The disequilibrium is corrected at a monthly speed of adjustment of 7%, while both in the short- and long-run, increasing PU negatively drives the SP. Similar to the main analysis, the behavior of defaults spread is aligned to the implication of PU both in the short- and long-run. We find inflation, industrial production, and unemployment as stock-friendly indicators in the long-run with positive stimulus.

The impulse response function of the PU_NB impact on SP in the United States for the sampling period is shown in Figure 4. The results of the impulse response graph show that a 10% increase in PU_NB negatively affects the SP in the US both in the short- and long term while a 10% drop in the PU_NB shows negative effects in the US in the short term and positive effects in the long term. A short-term decline in SP because of PU_NB may be attributable to standard investment behavior, which is depicted by declining risk premium, which constitutes a substantial part of security prices. However, in the long run, the disequilibrium is corrected as shown by ECT in Table 4, and investment patterns are adjusted according to the risk.

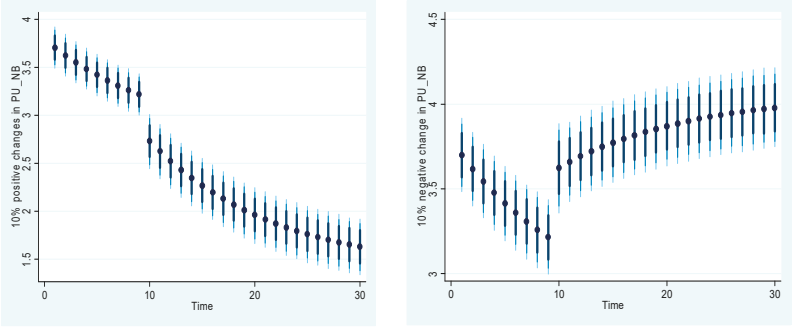


Figure 4. Graphical illustration of the response of stock prices (NYSE composite index) to a 10% increase (in the left portion of the Figure) and 10% decrease (in the right portion of the Figure) in PU_NB.

4.4. Robustness: Alternative Measures of Stock Prices

Apart from the main analysis, the study explores the matter in depth by using six alternative stock indices (S&P 500, Dow Jones Industrial Average, Dow Jones Composite Average, NASDAQ Composite, NASDAQ 100, and Dow Jones Transpiration Average) from the United States for robustness. Table A2 (Appendix B) uncovers interesting scenarios across the alternative measure so stock prices. Out of six alternatives, DJI, DJA, and DJT show a cointegration relationship to changes in PU, while other measures show negative but insignificant coefficients. One important phenomenon is the persistent negative reaction of all these measures to policy change in the short-run, while in the long-run, the response is negligible, and in particular, S&P 500, NASDAQ composite, and NASDAQ 100 are even not exposed to such a risk. These findings are encouraging for potential investors to diversify the investment across potential portfolios considering these reactions.

4.5. Threshold Regression

Table 5 summarizes the threshold regression results for both measures of PU (PU_NB, and PU_3C). Columns (1–2) carry the results of PU_NB regressed as an independent variable on NYSEC, and S&P 500, while columns (3–4) incorporate the robustness with an alternative measure, PU_3C. First, we estimate a single threshold model for PU_NB as a threshold variable and found 4.89 to be a significant threshold, which enables us to go ahead and estimate, double, and tribble thresholds. Second, a threshold of 4.48 was also significant, while the third one is found insignificant. Single threshold shows that a policy score of 4.89 (PU $\geq$ Th-1) negatively explains the stock prices in the US with coefficients -0.291 and -0.270 for models (1–2), respectively. For a double threshold level of 4.48 (PU $\geq$ Th-2& $\leq$ TH-1), the PU still negatively translates the stock prices, but the magnitude has relatively declined with coefficients of -0.074, and -0.005 for models (1–2), respectively. Interestingly, below 4.48 (PU $\leq$ TH-2), the relationship between PU and stock price becomes irrelevant, which has no statistical and economic implications. Likewise, the results are consistent (in terms of the direction of relationship and significance) and robust across alternative measure, PU_3C.

Identified thresholds through threshold regression are also shown in Figure 5, where the black horizontal line (a) indicates the single threshold of $\geq$4.89 (132.39 PU score), for which PU negatively explains stock prices with relatively high magnitude, whereas the second horizontal black line (b) denotes the double threshold value of 4.48 (87.98 PU score), the point where the coefficient of PU translates to a negative reaction in stock prices (less than the single threshold magnitude which remains unchanged between 4.89–4.48—the area covered by two lines). On the Y-axis, both thresholds corresponding to a) and b) are indicated by solid black circles. Interestingly, the area below 4.48 is the region in which PU becomes irrelevant to stock prices in our sample period.

Table 5. Threshold regression results.

	PU_NB		Robustness: PU_3C	
	(1)	(2)	(3)	(4)
Variables	SP	SP_S&P 500	SP	SP_S&P 500
TH-1		4.89 **		4.89 **
TH-2		4.48 **		4.48 **
INF	0.873 ***	1.017 ***	0.877 ***	0.987 ***
	(0.123)	(0.184)	(0.125)	(0.187)
IP	2.291 ***	2.080 ***	2.278 ***	2.098 ***
	(0.150)	(0.223)	(0.152)	(0.228)
GDP	−4.167 ***	−7.293 ***	−3.921 ***	−6.560 ***
	(0.983)	(1.467)	(0.914)	(1.366)
DS	−0.016 ***	0.008	−0.015 **	−0.043 ***
	(0.006)	(0.009)	(0.006)	(0.010)
UE	−0.023 *	−0.104 ***	−0.020 *	−0.048 ***
	(0.012)	(0.017)	(0.011)	(0.009)
PU ≥ Th-1	−0.291 ***	−0.270 ***	−0.174 **	−0.092 ***
	(0.077)	(0.093)	(0.093)	(0.016)
PU ≥ Th-2& ≤ TH-1	−0.074 **	−0.005 ***	−0.071 ***	−0.038 ***
	(0.034)	(0.002)	(0.011)	(0.014)
PU < TH-2	0.031	0.104	0.026	0.067
	(0.046)	(0.069)	(0.077)	(0.055)
Constant	13.086 ***	25.987 ***	12.111 ***	22.783 ***
	(4.459)	(6.657)	(4.219)	(6.304)
Obs.	428	428	428	428

Standard errors are in parenthesis. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$. The economic meanings of thresholds 4.89 and 4.48 are equivalent to the original PU_NB scores of 132.39, and 87.98, respectively. SP refers to the NYSE composite index.

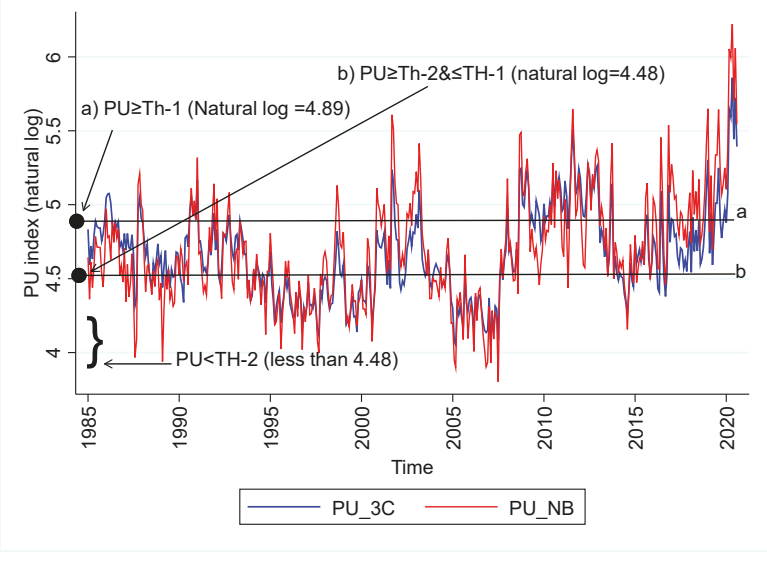


Figure 5. Illustration of identified thresholds (PU). Source: Baker, Bloom, and Davis [12].

5. Discussion

The empirical findings document a declining trend in stock prices in the short-run for both an increase and a decrease in PU. We have found that increased PU hurts stock prices while decreasing uncertainty increases them in the long-run. Following relevant literature, the study uses the New York Stock exchange composite index for baseline analysis and provides a comprehensive insight by extending the analysis to alternative stock indices (S&P 500, Dow Jones Industrial Average, Dow Jones Composite Average, NASDAQ Composite, NASDAQ 100, and Dow Jones Transpiration Average) for the United States. Moreover, besides the news-based measure of PU, we use three component-based uncertainties to affirm the baseline results. Interestingly, the findings produced by the alternative measures of stock prices (Dow Jones Industrial Average, Dow Jones Composite Average, and Dow Jones Transpiration Average) and PU are found consistent and robust.

For convenient discussion, the overall findings are categorized into three groups, namely, (1) DYS-ARDL output, (2) threshold points, and (3) channels following which PUs influence the stock prices.

(1) It is observed that a 10% shock in PU_NB (both positive/negative) negatively drives the stock prices in the short- and long-run (as depicted by Figure 4). This may be attributable to standard investment behavior differentials, depicted by declining risk premiums, which constitutes a substantial part of security prices. More specifically, a decline in PU also reduces the risk premium, which was part of security prices before the decline in PU. In this scenario, risk-seeking investors may shift investments to relatively high-risk securities while risk-averse investors may continue trading in existing securities. This behavior causes disequilibrium to the traditional demand and supply metaphor; however, in the long run, this disequilibrium is automatically rectified with monthly rates of around 6.7, and 7%, respectively (see Table 4), and investment patterns are corrected accordingly.

(2) The threshold(s) levels identified through threshold regression are interesting for policy matters. The PU score above the threshold point of 4.89 (natural log—equals 132.39 of original score) compels the pessimistic investors to be involved in selling their securities, which results in high supply and low demand, which causes a decline in stock prices and vice-versa for a decline in such risk. It is important to understand that a high level of PU appears to include most of the investors in shifting investments to relatively safe heavens, which, in contrast, behave differently for the second threshold. This difference denotes a relatively low magnitude in the explanation power of PU, which is still negative. In this stream between two threshold points (4.89–4.48 (132.39–87.98, original score)—the area covered between points a and b in Figure 2), the policy risk is not extremely high, which eventually influence the stock prices with relatively low magnitudes (Table 5, model (1), where coefficient changes from −0.291 to −0.07). This channel holds for models (2–4). While the stock price reaction to changes in PU below the second threshold appears to be irrelevant to decision making, it still carries a positive coefficient (which is statistically insignificant).

(3) It may take any one or a combination of more than one to influence the stock prices. The impact of changes in PU may theoretically take any combination or one of the following avenues to trigger stock prices. It can cause a delay in important decisions (e.g., employment, investment, consumption, and savings) by stakeholders (policymakers, regulators, and businesses, and economic agents) [29]. It increases financing and production costs by affecting the supply and demand channels, exacerbating the decline in investments and economic contraction [17,30,31]. Finally, the financial risk may be amplified due to such changes, as it is argued that the jump risk premium associated with policy decisions should be positive on average [17], which also influences inflation, interest rates, and expected risk premiums [32,33]. Therefore, the firms facing increased uncertainty in economic policy will reduce their investments in the short- and long-term [28]. This argument is supported by [58], who recorded a negative reaction in

accounting-based performance measures of firm performance in response to an increase in PU in the context of listed non-financial corporations in the United States.

The readers may carefully interpret the results of the threshold by understanding the original PU scores of 132.39 and 87.98 for single and double thresholds, respectively.

Summing up this section, we find that the literature supports our findings, for example, Asgharian, Christiansen, and Hou [21]; Bahmani-Oskooee and Saha [15]; Christou, Cunado, Gupta, and Hassapis [23]; Ko and Lee (2015); and Arouri, Estay, Rault, and Roubaud [14]. However, the findings of Wu, Liu, and Hsueh [22]; Pirgaip [27]; and Bahmani-Oskooee and Saha [15] are not in the same line, documenting no effect of the US stock market to changes in PU. Therefore, the PU-stock market dilemma shall remain debatable in the future.

6. Conclusions

This study revisits the PU-stock prices nexus and extends the equity market modeling beyond traditional cointegration specification by providing short- and long-run implications of PU on stock prices of the United Stated using the novel DYS-ARDL setting proposed by Jordan and Philips [1]. The next vital contribution of this study is the identification of two significant thresholds of PU (see Figure 2 and Table 5), which is a useful addition to the equity market literature. Finally, we provide a comprehensive picture of the PU and stock price relationship in the US perspective by expanding the analysis across seven stock market indices. Because related literature mostly opts for S&P 500 and NYSEC indices as representative of the US equity market, it is worth mentioning that all indices of the United States stock market are not equally exposed to rising PU, which may help effective diversification of portfolio and associated riskiness.

One important conclusion is the significant negative impact of PU on stock prices across all measures in the short-run, which needs to be considered by stakeholders while making an investment decision or policy formulation. Nevertheless, under dynamic ARDL simulations perspectives, the study extends the equity market literature by producing evidence from the world's largest equity by size and tosses the debate to be examined across other major financial markets of the world.

This research empirically examines the PU and stock price connection in the context of the United States, which is ranked first in terms of market size [41], using Jordan and Philips's [1] novel estimator and threshold setting. Particularly, there may be some other economic factors that influence stock prices; this study followed recent literature [14,15] to include industrial production, default spread, inflation, and the unemployment rate as potential controls. To be specific in capturing the PU impact on US stock indices, this study departs from examining the response of stock prices to positive/negative shocks in each of the control variables. Besides controls, the study excludes other medium and small equity indices in the present analysis, which may be of sound interest for policy matters for small investors. In future research, such considerations may produce interesting findings. Another grey area may be the extension of the present methodology to a cross-country level, based on the development and/or income level. Such an extension, with a comparative image, may be beneficial to those who want to diversify the security market investment across national borders.

Author Contributions: Conceptualization, M.A.K., and M.A.; methodology, software and data curation, M.A.K.; writing—original draft preparation, M.A.K., and M.A.; writing—review and editing, J.P. and J.O. All authors have read and agreed to the published version of the manuscript.

Funding: Project no. 132805 has been implemented with support provided from the National Research, Development, and Innovation Fund of Hungary, financed under the K_19 funding scheme.

Acknowledgments: We are grateful to Kamran Khan, a doctoral candidate at Northeast Normal University, School of Economics and Management for guidance in estimating the DYS-ARDL model and its technicalities. We are also appreciative of Juan, guest editor, for consistent guidance throughout the review process and finally of the constructive and encouraging suggestions by anonymous reviewers at the core of this manuscript, which we endorse vigorously.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. OLS baseline results.

	(1)	(2)	(3)	(4)	(5)
	SP	**SP**	**SP**	**SP**	**SP**
PU_NB	0.443 ***	−0.220 ***	−0.186 ***	−0.124 ***	−0.105 ***
	(0.072)	(0.021)	(0.029)	(0.022)	(0.022)
INF		0.717 ***	0.481 ***	0.531 ***	0.556 ***
		(0.023)	(0.228)	(0.186)	(0.150)
IP			0.301	0.269	0.176
			(0.280)	(0.226)	(0.189)
DS				−0.053 ***	−0.029 ***
				(0.005)	(0.007)
UE					−0.040 ***
					(0.009)
Constant	0.460 ***	−0.505 ***	−0.776 ***	−0.986 ***	−0.646 ***
	(0.343)	(0.130)	(0.205)	(0.165)	(0.208)
Obs.	428	428	428	428	428
R-squared	0.057	0.944	0.948	0.961	0.965

Standard errors are in parenthesis. *** $p < 0.01$. All the abbreviations are defined in Table 1.

Appendix B

Table A2. Robustness: alternative measures of SP.

	(1)	(2)	(3)	(4)	(5)	(6)
	SP_DJI	**SP_DJA**	**SP_SP**	**SP_NASDAQC**	**SP_NASDAQ100**	**SP_DJT**
ECT(−1)	−0.027 *	−0.048 ***	−0.016	−0.012	−0.011	−0.075 ***
	(0.016)	(0.018)	(0.014)	(0.012)	(0.011)	(0.023)
ΔPU_NB	−0.045 ***	−0.041 ***	−0.042 ***	−0.066 ***	−0.068 ***	−0.054 ***
	(0.008)	(0.008)	(0.008)	(0.012)	(0.014)	(0.012)
PU_NB	−0.006 ***	−0.007 **	0.004	0.006	0.006	−0.008 ***
	(0.002)	(0.003)	(0.007)	(0.010)	(0.012)	(0.003)
Constant	−0.162	−0.338 **	−0.127	−0.167	−0.195	−0.510 ***
	(0.127)	(0.137)	(0.125)	(0.154)	(0.212)	(0.187)
Obs.	427	427	427	427	418	343
F-statistics	5.92 ***	6.11 ***	5.95 ***	5.48 ***	4.51***	4.52 ***
Simulations	5000	5000	5000	5000	5000	5000

Standard errors are in parenthesis. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$. DJI: Dow Jones Industrial Average, DJA: Dow Jones Composite Average, S&P: S&P 500, NASDAQ Composite, NASDAQ100, and DJT: Dow Jones Transpiration Average. Controls are not shown in this table as they are the same as estimated in the baseline model.

References

1. Jordan, S.; Philips, A.Q. Cointegration testing, and dynamic simulations of autoregressive distributed lag models. *Stata J.* **2018**, *18*, 902–923. [CrossRef]
2. Stübinger, J.; Endres, S. Pairs trading with a mean-reverting jump—Diffusion model on high-frequency data. *Quant. Financ.* **2018**, *18*, 1735–1751. [CrossRef]
3. Avellaneda, M.; Lee, J.-H. Statistical arbitrage in the US equities market. *Quant. Financ.* **2010**, *10*, 761–782. [CrossRef]
4. Liu, B.; Chang, L.-B.; Geman, H. Intraday pairs trading strategies on high frequency data: The case of oil companies. *Quant. Financ.* **2017**, *17*, 87–100. [CrossRef]
5. Do, B.; Faff, R. Are Pairs Trading Profits Robust to Trading Costs? *J. Financ. Res.* **2012**, *35*, 261–287. [CrossRef]
6. Stübinger, J.; Mangold, B.; Krauss, C. Statistical arbitrage with vine copulas. *Quant. Financ.* **2018**, *18*, 1831–1849. [CrossRef]

7. Khan, M.A.; Domician, M.; Abdulahi, M.E.; Sadaf, R.; Khan, M.A.; Popp, J.; Oláh, J. Do Institutional Quality, Innovation and Technologies Promote Financial Market Development? *Eur. J. Int. Manag.* **2020**, *14*. [CrossRef]
8. Khan, M.A.; Ilyas, R.M.A.; Hashmi, S.H. Cointegration between Institutional Quality and Stock Market Development. *NUML Int. J. Bus. Manag.* **2018**, *13*, 90–103.
9. Khan, M.A.; Khan, M.A.; Abdulahi, M.E.; Liaqat, I.; Shah, S.S.H. Institutional quality and financial development: The United States perspective. *J. Multinatl. Financ. Manag.* **2019**, *49*, 67–80. [CrossRef]
10. Shah, S.S.H.; Khan, M.A.; Meyer, N.; Meyer, D.F.; Oláh, J. Does Herding Bias Drive the Firm Value? Evidence from the Chinese Equity Market. *Sustainability* **2019**, *11*, 5583. [CrossRef]
11. Tong, H. Threshold models in time series analysis—30 years on. *Stat. Interface* **2011**, *4*, 107–118. [CrossRef]
12. Baker, S.R.; Bloom, N.; Davis, S.J. Measuring economic policy uncertainty. *Q. J. Econ.* **2016**, *131*, 1593–1636. [CrossRef]
13. Sum, V. Economic policy uncertainty in the United States and Europe: A cointegration test. *Int. J. Econ. Financ.* **2013**, *5*, 98–101. [CrossRef]
14. Arouri, M.; Estay, C.; Rault, C.; Roubaud, D. Economic policy uncertainty and stock markets: Long-run evidence from the US. *Financ. Res. Lett.* **2016**, *18*, 136–141. [CrossRef]
15. Bahmani-Oskooee, M.; Saha, S. On the effects of policy uncertainty on stock prices. *J. Econ. Financ.* **2019**, *43*, 764–778. [CrossRef]
16. Das, D.; Kumar, S.B. International economic policy uncertainty and stock prices revisited: Multiple and Partial wavelet approach. *Econ. Lett.* **2018**, *164*, 100–108. [CrossRef]
17. Pastor, L.; Veronesi, P. Uncertainty about government policy and stock prices. *J. Financ.* **2012**, *67*, 1219–1264. [CrossRef]
18. Sum, V. Economic policy uncertainty and stock market performance: Evidence from the European Union, Croatia, Norway, Russia, Switzerland, Turkey and Ukraine. *J. Money Invest. Bank.* **2012**, *25*, 99–104. [CrossRef]
19. Li, X.-l.; Balcilar, M.; Gupta, R.; Chang, T. The causal relationship between economic policy uncertainty and stock returns in China and India: Evidence from a bootstrap rolling window approach. *Emerg. Mark. Financ. Trade* **2016**, *52*, 674–689. [CrossRef]
20. Gao, J.; Zhu, S.; O'Sullivan, N.; Sherman, M. The role of economic uncertainty in UK stock returns. *J. Risk Financ. Manag.* **2019**, *12*, 5. [CrossRef]
21. Asgharian, H.; Christiansen, C.; Hou, A.J. Economic Policy Uncertainty and Long-Run Stock Market Volatility and Correlation. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3146924 (accessed on 29 October 2020).
22. Wu, T.-P.; Liu, S.-B.; Hsueh, S.-J. The causal relationship between economic policy uncertainty and stock market: A panel data analysis. *Int. Econ. J.* **2016**, *30*, 109–122. [CrossRef]
23. Christou, C.; Cunado, J.; Gupta, R.; Hassapis, C. Economic policy uncertainty and stock market returns in PacificRim countries: Evidence based on a Bayesian panel VAR model. *J. Multinatl. Financ. Manag.* **2017**, *40*, 92–102. [CrossRef]
24. Sum, V. Economic Policy Uncertainty and Stock Market Returns. *SSRN Electron. J.* **2012**. [CrossRef]
25. Debata, B.; Mahakud, J. Economic policy uncertainty and stock market liquidity: Does financial crisis make any difference? *J. Financ. Econ. Policy* **2018**, *10*, 112–135. [CrossRef]
26. Škrinjarić, T.; Orlović, Z. Economic policy uncertainty and stock market spillovers: Case of selected CEE markets. *Mathematics* **2020**, *8*, 1077. [CrossRef]
27. Pirgaip, B. The causal relationship between stock markets and policy uncertainty in OECD countries. In Proceedings of the RSEP International Conferences on Social Issues and Economic Studies, Barcelona, Spain, 7–10 November 2017.
28. Chen, P.-F.; Lee, C.-C.; Zeng, J.-H. Economic policy uncertainty and firm investment: Evidence from the U.S. market. *Appl. Econ.* **2019**, *51*, 3423–3435. [CrossRef]
29. Gulen, H.; Ion, M. Policy uncertainty and corporate investment. *Rev. Financ. Stud.* **2016**, *29*, 523–564. [CrossRef]
30. Julio, B.; Yook, Y. Corporate financial policy under political uncertainty: International evidence from national elections. *J. Financ.* **2012**, *67*, 45–84. [CrossRef]
31. Leduc, S.; Liu, Z. Uncertainty shocks are aggregate demand shocks. *J. Monet. Econ.* **2016**, *82*, 20–35. [CrossRef]
32. Pástor, L'.; Veronesi, P. Political uncertainty and risk premia. *J. Financ. Econ.* **2013**, *110*, 520–545. [CrossRef]

33. Bernal, O.; Gnabo, J.-Y.; Guilmin, G. Economic policy uncertainty and risk spillovers in the Eurozone. *J. Int. Money Financ.* **2016**, *65*, 24–45. [CrossRef]
34. Ko, J.-H.; Lee, C.-M. International economic policy uncertainty and stock prices: Wavelet approach. *Econ. Lett.* **2015**, *134*, 118–122. [CrossRef]
35. Pesaran, M.H.; Shin, Y.; Smith, R.J. Bounds testing approaches to the analysis of level relationships. *J. Appl. Econom.* **2001**, *16*, 289–326. [CrossRef]
36. Shin, Y.; Yu, B.; Greenwood-Nimmo, M. Modelling asymmetric cointegration and dynamic multipliers in a nonlinear ARDL framework. In *Festschrift in Honor of Peter Schmidt*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 281–314.
37. Khan, M.I.; Teng, J.Z.; Khan, M.K. The impact of macroeconomic and financial development on carbon dioxide emissions in Pakistan: Evidence with a novel dynamic simulated ARDL approach. *Environ. Sci. Pollut. Res.* **2020**, *27*, 39560–39571. [CrossRef] [PubMed]
38. Khan, M.K.; Teng, J.-Z.; Khan, M.I.; Khan, M.O. Impact of globalization, economic factors and energy consumption on CO_2 emissions in Pakistan. *Sci. Total Environ.* **2019**, *688*, 424–436. [CrossRef] [PubMed]
39. Zapata, H.O.; Gauthier, W.M. Threshold models in theory and practice. In Proceedings of the 2003 Annual Meeting of the Southern Agricultural Economics Association, Mobile, AL, USA, 1–5 February 2003.
40. Hansen, B.E. Threshold autoregression in economics. *Stat. Interface* **2011**, *4*, 123–127. [CrossRef]
41. Szmigiera, M. Largest Stock Exchange Operators, Listed by Market Cap of Listed Companies 2020. 2020. Available online: https://www.statista.com/statistics/270126/largest-stock-exchange-operators-by-market-capitalization-of-listed-companies/ (accessed on 20 May 2020).
42. Liu, L.; Zhang, T. Economic policy uncertainty and stock market volatility. *Financ. Res. Lett.* **2015**, *15*, 99–105. [CrossRef]
43. Ehrmann, M.; Fratzscher, M. Global financial transmission of monetary policy shocks. *Oxf. Bull. Econ. Stat.* **2009**, *71*, 739–759. [CrossRef]
44. Brogaard, J.; Detzel, A. The asset-pricing implications of government economic policy uncertainty. *Manag. Sci.* **2015**, *61*, 3–18. [CrossRef]
45. Antonakakis, N.; Chatziantoniou, I.; Filis, G. Dynamic co-movements of stock market returns, implied volatility and policy uncertainty. *Econ. Lett.* **2013**, *120*, 87–92. [CrossRef]
46. Dakhlaoui, I.; Aloui, C. The interactive relationship between the US economic policy uncertainty and BRIC stock markets. *Int. Econ.* **2016**, *146*, 141–157. [CrossRef]
47. Yang, M.; Jiang, Z.-Q. The dynamic correlation between policy uncertainty and stock market returns in China. *Phys. A Stat. Mech. Appl.* **2016**, *461*, 92–100. [CrossRef]
48. Yahoo-Finance. 2020. Available online: https://finance.yahoo.com/ (accessed on 1 September 2020).
49. FRED, Federal Reserve Economic Data. 2020. Available online: https://fred.stlouisfed.org/ (accessed on 28 September 2020).
50. Abdulahi, M.E.; Shu, Y.; Khan, M.A. Resource rents, economic growth, and the role of institutional quality: A panel threshold analysis. *Resour. Policy* **2019**, *61*, 293–303. [CrossRef]
51. Khan, M.A.; Gu, L.; Khan, M.A.; Oláh, J. Natural Resources and Financial Development: The Role of Institutional Quality. *J. Multinatl. Financ. Manag.* **2020**, *56*, 100641. [CrossRef]
52. Khan, M.A.; Islam, M.A.; Akbar, U. Do economic freedom matters for finance in developing economies: A panel threshold analysis. *Appl. Econ. Lett.* **2020**, 1–4. [CrossRef]
53. Liu, H.; Islam, M.A.; Khan, M.A.; Hossain, M.I.; Pervaiz, K. Does financial deepening attract foreign direct investment? Fresh evidence from panel threshold analysis. *Res. Int. Bus. Financ.* **2020**, *53*, 101198. [CrossRef]
54. Kwiatkowski, D.; Phillips, P.C.B.; Schmidt, P.; Shin, Y. Testing the null hypothesis of stationarity against the alternative of a unit root. *J. Econom.* **1992**, *54*, 159–178. [CrossRef]
55. Baum, C. *KPSS: Stata Module to Compute Kwiatkowski-Phillips-Schmidt-Shin Test for Stationarity*; Boston College Department of Economics: Boston, MA, USA, 2000.
56. Bahmani-Oskooee, M.; Saha, S. On the effects of policy uncertainty on stock prices: An asymmetric analysis. *Quant. Financ. Econ.* **2019**, *3*, 412–424. [CrossRef]

57. Khan, M.K.; Teng, J.-Z.; Khan, M.I. Effect of energy consumption and economic growth on carbon dioxide emissions in Pakistan with dynamic ARDL simulations approach. *Environ. Sci. Pollut. Res.* **2019**, *26*, 23480–23490. [CrossRef]
58. Iqbal, U.; Gan, C.; Nadeem, M. Economic policy uncertainty and firm performance. *Appl. Econ. Lett.* **2020**, *27*, 765–770. [CrossRef]

mathematics

MDPI

Article

Do Trade and Investment Agreements Promote Foreign Direct Investment within Latin America? Evidence from a Structural Gravity Model

Marta Bengoa [1,2,*], Blanca Sanchez-Robles [3] and Yochanan Shachmurove [4,5]

1 Colin Powell School, City University of New York (CUNY-CCNY), New York, NY 10031, USA
2 SARChI College of Business and Economics, University of Johannesburg South Africa, Senior Fellow at CIRANO, Montreal, QC H2Y1C6, Canada
3 Department of Economic Analysis at UNED University, 28040 Madrid, Spain; bsanchez-robles@cee.uned.es
4 The City College and Graduate Center of the City University of New York (CUNY-CCNY), New York, NY 10031, USA; yshachmurove@ccny.cuny.edu
5 Faculty of Management at The University of Warsaw, 00-927 Warsaw, Poland
* Correspondence: mbengoa@ccny.cuny.edu

Received: 28 September 2020; Accepted: 21 October 2020; Published: 30 October 2020

Abstract: Latin America has experienced a surge in foreign direct investment (FDI) in the last two decades, in parallel with the ratification of major regional trade agreements (RTAs) and bilateral investment treaties (BITs). This paper uses the latest developments in the structural gravity model theory to study if the co-existence of BITs and two major regional agreements, Mercosur and the Latin American Integration Association (ALADI), exerts enhancing or overlapping effects on FDI for eleven countries in Latin America over the period 1995–2018. The study is novel as it accounts for variations in the degree of investment protection across BITs within Latin America by computing a quality index of BITs. It also explores the nature of interactions (enhancing/overlapping effects) between RTAs and BITs. The findings reveal that belonging to a well-established regional trade agreement, such as Mercosur, is significantly more effective than BITs in fostering intra-regional FDI. Phasing-in effects are large and significant and there is evidence of enhancing effects. Results within the bloc are heterogeneous: BITs exert a positive, but small effect, for middle income countries. However, BITs are not effective in attracting FDI in the case of middle to low income countries, unless these countries ratify BITs with a high degree of investment protection.

Keywords: foreign direct investment; bilateral investment treaties; regional trade agreements; structural gravity model

1. Introduction

Foreign Direct Investment (FDI) into and across Latin America has experienced a dynamic performance in the last few decades. In parallel, many countries in the area have taken part in regional trade agreements (RTAs) and bilateral investment treaties (BITs). RTAs foster trade by facilitating access to foreign markets. Sometimes they contain provisions about FDI which enhance these flows. BITs specify conditions under which foreign investment operates in the host country. This paper addresses the following questions: What is the relative importance of RTAs versus BITs as a way of attracting FDI? Do effects differ depending on the nature of the BITs and the presence of other agreements? Do they complement or substitute for economic and political institutions? There is substantial controversy in the empirical literature about these matters.

Some contributions have argued that host countries should exhibit a certain minimum level of income or other forms of social capacity in order to profit from FDI [1,2]. Nevertheless, it is frequent for

developing countries to lack, at least partially, the necessary environment (human capital, rule of law, institutions, etc.) for efficient activity of multinational enterprises (MNEs). This raises the question of whether the impact of RTAs and BITs on FDI might be contingent on the institutional framework of the host country.

Specifically, this paper addresses three questions:

(i) What is the impact of regional agreements and BITs on intraregional FDI in Latin America? In particular, the paper focuses on the FDI creation and diversion effects of the two main RTAs in the region, Mercosur (Southern Common Market) and ALADI (The Latin American Integration Association). The study explores the relative effectiveness—interaction and complementarity—of trade and investment treaties regarding their impact on FDI.
(ii) Do the qualitative aspects of BITs, as measured by a new index, matter for their efficacy?
(iii) Do the specific institutional characteristics of recipient countries determine or condition the effectiveness of major RTAs and BITs?

The framework of a structural gravity model for FDI is used for this purpose (see [3], for a thorough exposition). The structural gravity model in economics, inspired by its physic counterpart, has recently acquired popularity as an appropriate tool for analyzing international trade and investment. According to this model, bilateral flows/stocks among countries are directly related to their sizes (usually captured by their Gross Domestic Product or GDPs) and inversely related to the distance between them. This framework combines an intuitive appeal, which can be rigorously founded on theoretical propositions, with strong predicting capabilities.

The study constructs a panel data detailing intra-regional bilateral FDI stocks among eleven Latin American countries over the period 1995–2018. It takes into account BITs' quality in terms of the degree of investment protection they warrant, cross-country differences in endowments, and level of developed institutions. The paper employs the Poisson pseudo-maximum likelihood estimator (PPML), as recommended by [4]. It addresses the potential endogeneity derived from the establishment of RTAs and BITs as well as the robustness of the results to alternative estimations. The main findings suggest that Mercosur exerts a larger impact than either ALADI or the presence of BITs on intra-bloc foreign direct investment. This is also true when controlling for investment protection strength. However, BITs are effective in fostering intra-regional FDI in Latin America when there is a sizeable degree of institutional development in the host country or a high degree of investment protection entailed by the treaty. Furthermore, the interaction of Mercosur, ALADI and BITs positively impacts FDI. The effects of BITs are larger for middle income than in low-middle income economies. The results are consistent with those obtained for other areas of the world [5–7].

This paper focuses on Latin America since the subcontinent is composed of developing and middle-income countries, whose population have the potential to substantially improve their living conditions. Therefore, it is of the utmost importance from a development viewpoint. Furthermore, the area is in the midst of undergoing an active and complex integration process, in which countries have engaged historically in different, often conflicting, approaches to trade liberalization and FDI. In addition, the empirical literature exploring FDI in the area is scarce as well as reaching ambiguous results. More generally, South-South empirical studies of the impact of RTAs and BITs on FDI for developing countries are still sparse. Finally, the area comprises countries which exhibit a substantial degree of heterogeneity, while sharing common aspects. It is, therefore, an appropriate sample for an empirical investigation.

This paper lies at the intersection of several strands of the literature. On one hand, it is closely related to contributions exploring FDI from a theoretical and empirical viewpoint [8–11]. On the other hand, it builds upon analyses conducted within the structural gravity model see [3,12–15]. Additionally, it is similar in spirit to papers which examine the impact of BITs on FDI [7,16–18]. Finally, Dixon and Haslam's work [17] analyze the impact of BITs on FDI in several samples, one of them of intraregional Latin American flows. Their results for this subsample are somewhat puzzling, since they suggest that

the interaction of weak BITs and RTAs has a negative impact on FDI for this area. This study extends and complements these previous studies.

This paper contributes to the literature in four dimensions. First, and in terms of methodology, we work with a fully specified structural gravity model expressly accounting for multilateral resistance terms among countries. Second, and as a consequence of methodological differences, the results are more in accordance with the underlying theoretical framework than other contributions, and very robust to alternative specifications. Third, the study accounts for variations in the quality of BITs through computing a quality index. Refs [7,18] analyze the effects of BITs by means of an index capturing dispute settlement mechanisms. This study extends and complements these analyses by designing a more thorough index of BITs, encompassing a wider set of BIT clauses potentially important for investors. The samples are also different, since they work with a large number of developed and developing countries, while the focus of this study is Latin America. Additionally, this paper explores the possibility of enhancing or overlapping effects when there are major RTAs in place coexisting with BITs. Finally, the analysis here covers the largest temporal horizon possible by working with bilateral data over 1995–2018.

In particular, the paper suggests that the interaction of Mercosur and BITs has a positive and significant impact on FDI whereas the combination of ALADI and BITs also displays a positive effect, although smaller in size and less significant. The study also explores the heterogeneity within the country sample according to their income levels.

The structure of the paper is as follows: Section 2 provides insights about the main RTAs and BITs in Latin America and discusses the links between RTAs, BITs, and foreign direct investment, summarizing the relevant literature. Section 3 describes the theoretical model that underlies our empirical work. Section 4 presents the data and develops the empirical methodology. Section 5, presents the results. Section 6 concludes.

2. An Overview of Integration Agreements in Latin America and the Links between RTAs, BITs and FDI

Mercosur was signed in 1991 by Argentina, Brazil, Paraguay and Uruguay. Mercosur was designed to be a customs union, with a free intra-zone trade and a common trade policy. Bolivia entered in 2006 but has not been recognized as a full member by the other members. Venezuela joined in 2015 and has since been suspended from the Treaty. Chile, Colombia, Ecuador, Peru, Guyana and Suriname are associated members. Mexico signed a deep FTA in 2006 and it has the status of observer. Members started to lower their tariffs in 1991 and the tariff schedule varies in the range 0%–20%. Approximately 90% of trade was liberalized by 1997. Mercosur included protocols for BIT protection and investor-state dispute settlement (ISDS) mechanisms; however, they have never been enforced. Mercosur engaged in negotiations with other Latin American countries to establish free trade agreements (FTAs). The FTAs provide substantial levels of integration between Mercosur and third countries, named associated members. The associated members are Chile, Colombia, Ecuador, Peru, Guyana, and Suriname. FTAs aimed to reduce tariffs to the same levels as Mercosur and diminish non-tariff barriers, facilitating trade and investment. The online Appendix A exhibits the sequence of the signatories of the FTAs.

The Latin America Free Trade Association was created in 1962 as the first component of an intended large integration project. It was superseded by ALADI in 1980. ALADI provides a general framework which intends to foster integration in the region and guarantee its economic and social development. It is not a deep integration mechanism per se. From that starting point, individual countries have strengthened their integration process by gradually engaging in bilateral or multilateral treaties with other ALADI members. De facto, ALADI now encompasses various free trade agreements within its framework. Not all countries within ALADI have yet established substantial/deep trade agreements among themselves. The countries under the ALADI umbrella are: Argentina, Bolivia, Brazil, Chile, Colombia, Cuba, Ecuador, Mexico, Paraguay, Panama, Peru, Uruguay, and Venezuela.

Hence the gradual integration that ALADI envisaged is still in process. Table A1 of the online Appendix A summarizes the pairs of countries, under the general ALADI scheme, involved over time in FTA agreements.

Latin American countries have signed an increasing number of BITs between 1995 and 2007, although the attitude of countries towards these agreements is not uniform. Chile has been very active and takes part in seven BITs, which include clauses with a high degree of protection to foreign investors. Colombia and Mexico, instead, have only signed one and two BITs respectively, with a lower degree of investment protection. During our period of analysis Brazil has not ratified any BIT, but since 2015 the country has negotiated, but not concluded, BITs with Mexico, Chile, Colombia and Peru (for a more detailed account see the United Nations Commission for Trade and Development (UNCTAD), 2017). Table A2 of the online Appendix A offers the list of BITs and their dates of entry into force.

2.1. The Relationship between Regional Trade Agreements and Foreign Direct Investment

The literature has traditionally focused on the distinction between horizontal and vertical types of foreign investment. Horizontal (or proximity concentration) FDI undertakes all production in the country whose market it intends to serve, thus substituting trade between the parent firm and affiliates Refs [8,19]. High trade barriers incentivize the setup of proximity-concentration FDI and therefore RTAs do not necessarily impact horizontal FDI positively [10].

Vertical FDI, instead, carries out each step of the production process in a different location, in order to minimize costs through scale economies and/or benefit from the low cost of inputs [20]. This implies an active exchange of intermediate and final goods between parents and affiliates. Vertical MNEs benefit from the trade liberalization which an agreement entails; hence RTAs exert a positive impact on vertical FDI.

On the other hand, US multinationals abroad seem to be organized in a horizontal pattern, although frequently MNE parent firms own vertically linked affiliates [11]. The combination between horizontal and vertical features follows what has been named a hybrid-knowledge capital model [9]. A particular case is the export-platform strategy [21]: companies set up a plant in a country belonging to a trade bloc in order to improve their access to other markets in the bloc. In these instances, net effects of trade integration are more nuanced and non-predictable a priori, since they depend on the relative importance and complex relationships between horizontal and vertical integration patterns within the firms.

Additionally, RTAs may also alter the macroeconomic environment where firms operate by strengthening fiscal discipline, macroeconomic stability and the rule of law in the host country [22–24]. Therefore, they might provide a more favourable setting to attract FDI.

Since the net impact of RTAs on aggregate FDI is a priori ambiguous, due to the intertwining of different forces described previously, the connection between RTAs and FDI is an empirical issue. Related evidence is not unanimous, though. A number of contributions report a positive impact of RTAs on FDI. With respect to the case of Mexico and NAFTA, ref [24] find a non-significant effect. For the European Union [25] documents a similar result (although the effect gradually decays). Ref [13] asserts that trade liberalization favoured the relocation of FDI from West to East Europe in the 90s, partly based upon export platform motives. Finally, these studies document a positive impact of trade agreements on FDI for a sample of OECD and non-OECD countries (see [10,26]).

In the case of the Canada–US Free Trade Agreement (CUSFTA) on FDI flows [24], no significant effect was found. For Latin America, Ref [22] identify pro-market and stable macroeconomic policies as the main factors attracting FDI, but an important role for Mercosur was not found. Finally, there are contributions suggesting that the impact of RTAs on FDI is contingent on other factors, such as skill differences in the home and host country [27], the institutional, financial, and macroeconomic framework of the host country [28], and the features of the agreement itself [16].

The lack of consensus in the literature might be due to the use of different identification strategies and econometric methods, such as the set of control variables included in the equations, the specification of fixed effects, and the estimation procedures. This paper contributes to shedding light on these issues.

2.2. Bilateral Investment Treaties and Foreign Direct Investment

The literature has identified three main theoretical reasons why BITs may impact FDI: they *signal* that signatory governments are willing to create an adequate institutional and economic environment for FDI [29,30]; they provide an *insurance* for foreign investors by establishing compensation schemes and conflict resolution procedures; they *deter* non-compliance because of the potential reputation costs for countries breaching the treaties [7]. When BITs are considered as signals, they may attract FDI from both partner and non-partner countries. If BITs are considered as insurers or deterrents, though, the attraction of FDI from partners will be higher than from non-partners, albeit both will be positive.

The first wave of empirical contributions about the impact of BITs on FDI addressed the link between the presence of a BIT and FDI, while more recent research focuses on the association between the nature of the BIT (as summarized by a set of characteristics) and the attraction of flows. In general, results from both cases are mixed. These studies [6,29,31–33] report a positive impact of investment treaties on FDI. In addition, Ref [12] found that the (positive) long-run effect of BITs on FDI is larger than the short-run impact because of phasing in effects. Following refs [30,33], BITs act as substitutes for weak legal and regulatory institutions in the host country.

Other studies, refs [34–37], instead, do not find a link between BITs and FDI. Tobin and Rose-Ackerman [38] suggest that BITs do not impact FDI when considered in isolation, but that they exert a positive effect when interacting with institutional or fundamental variables, thus concluding that BITs complement institutions in the host country.

The strand of the literature which deals with the nature of BITs is more recent in time and sparser. These contributions consider not only the number of BITs in place but also their key qualitative aspects. They usually work with bilateral data (as opposed to aggregate), which allows an investigation the link between a particular BIT signed by a pair of countries and the inflows between that same pair of countries.

A key feature of BITs is the treatment of dispute settlement procedures but results here are not uniform either. The impact of BITs on North-South and South-South FDI flows over the period 1990–2008 is the focus of the study by Dixon and Haslam [17]. They construct an index to capture more thoroughly the degree of FDI protection entailed by each BIT, including, but not restrained to, dispute settlement procedures (see their online Appendix A for classification). Their empirical results suggest that only treaties providing a strong degree of protection to investment impact FDI.

Frenkel and Walter's work [18] focuses on dispute settlement procedures. In the spirit of [17], they construct an indicator for each BIT by adding the assigned scores to various features. They show that the strength of the dispute settlement mechanisms is positively correlated with FDI. A recent study [7] concentrates on the effects of dispute settlement mechanisms on both partner and non-partner countries. They suggest that BITs have a positive impact on FDI from partner countries if a dispute with an investor has not affected the host country.

This paper builds partially on these contributions. Alternatively to the Dixon and Haslam study [17], it uses a gravity model with bilateral data which enables us to identify the impact of particular BITs on their own signatories. The analysis is not restricted uniquely to the quality of BITs dispute settlement mechanisms, as in previous studies [7,18].

It follows from the above that results regarding the connection between the existence and/or characteristics of BITs and FDI are mixed so far. As in the case of RTAs, some of the discrepancies may relate to econometric aspects such as the definition of the variable capturing the BITs, the strategy of controlling for endogeneity and phasing-in effects, the estimation techniques, and the use of aggregate versus bilateral data. Moreover, there is no consensus about the key characteristics of a BIT and how to measure them. Finally, it is not clear either if BITs have a differential impact in countries with weak

versus strong institutions. On the one hand, they may substitute for institutions by giving credibility to governments, (as claimed by [30–33]). On the other, they act as complements since strong institutions lend support to treaties [38]. Ultimately, this is a multifaceted issue related to geography, the quality of institutions, the nature of the agreements, the rule of law, and other idiosyncratic aspects of the country itself.

3. Theoretical Framework: The Gravity Model for FDI

The gravity model has been extensively used to study international trade flows. Further theoretical and empirical advances allow its use as a framework to study FDI [39]. The gravity model is compatible with a theoretical model of heterogeneous multinational firms. Moreover, it conveniently allows us to capitalize on the rich information embedded in databases organized around dyads of countries (home/parent country origin of FDI flows and host/receptor of FDI flows); this feature is especially relevant for this paper because we study how RTAs and bilateral BITs signed across Latin American countries impact intra-bloc bilateral FDI.

In addition, the disaggregation by country pairs over time increases the number of observations available to explore the panel dimension. Finally, recent contributions have been actived in designing adequate techniques to circumvent econometric issues associated with the gravity equation, such as the inclusion of fixed effects [40], or the presence of many zeros in the data [4].

While the gravity equation for trade now has a solid theoretical background, the development of theoretical gravity FDI models is more recent. The Head and Ries' model [41] serves as baseline for this paper, but we consider technology as non-rival as in [9]. This analysis follows, as well, Anderson and Yotov's model [42] in developing an intuitive FDI gravity equation, similar to the structural gravity system for trade, which is possible to estimate directly.

The FDI structural system is also similar in spirit to the trade structural gravity model [43,44]. It departs from a definition of bilateral FDI:

$$FDI_{ij,t}^{stock} \equiv \omega_{ij,t}^{\varepsilon} M_{i,t} \tag{1}$$

where $FDI_{ij,t}$ represents FDI stocks between countries i and j at a time t, $M_{i,t}$ is the non-rival aggregate technology capital stock in a particular time, ω_{ijt} represents openness (or barriers to FDI) for foreign technology coming from country i to country j, and ε is the elasticity of FDI with respect to openness. To transform FDI stocks into values we multiply Equation (1) by its marginal product:

$$FDI_{ij,t}^{stock,value} \equiv \omega_{ij,t}^{\varepsilon} M_{i,t} \frac{\partial Y_{j,t}}{\partial M_{i,t}} \tag{2}$$

in which $M_i = \phi_i \frac{E_i}{P_i}$, where E_i represents total expenditures of country i which equal the country's output plus net rents from foreign investment, P_i stands for consumer prices (which can be considered as a multilateral resistance since higher prices for goods and inputs can affect FDI), and Y is nominal output. The production function is Cobb Douglas. Therefore $\frac{\partial Y_j}{\partial M_i} = \phi_j \frac{Y_j}{M_i}$; Y_j equals $Y_t = \sum_j Y_{j,t}$. Solving the representative agent's problem delivers a structural system for the steady-state. The gravity system of equations is given by:

$$FDI_{ij} = \phi_i \phi_j \omega_{ij}^{\varepsilon} \frac{E_i}{P_i} \frac{Y_j}{M_i} \tag{3}$$

$$P_i = \left[\sum_{j=1}^{N} \left(\frac{\tau_{ji}}{\Pi_j} \right)^{1-\sigma} \frac{Y_j}{Y} \right]^{\frac{1}{1-\sigma}}, \tag{4}$$

$$\prod_j = \left[\sum_{i=1}^{N} \left(\frac{\tau_{ji}}{P_i} \right)^{1-\sigma} \frac{E_i}{Y} \right]^{\frac{1}{1-\sigma}}. \tag{5}$$

where P_i denotes the aggregate price index or the multilateral resistance, as defined in [43], and τ_{ji} represents standard iceberg trade costs. Equation (3) establishes that FDI between two countries depends positively on the home country size E_i and the size of the host economy Y_j. According to Equation (3), FDI depends negatively on FDI barriers $\phi_i\phi_j$. Higher multilateral resistances (MR) in the country of origin or higher opportunity cost of investing in technology should lead to lower FDI. ω_{ij} takes the form:

$$\omega_{ij,t} = d_{ij}^{\beta_1} \cdot \exp\left(\beta_2 X_{ij,t}\right) \tag{6}$$

where d_{ij} stands for the bilateral distance between the countries and $X_{ij,t}$ is a set of variables that capture both deterrents and incentives for FDI. This includes the variables Mercosur, ALADI, and BITs, as well as the BIT investment protection index (details are included below). The analysis adds control for other common institutional characteristics, differences in factor endowments, and labor costs. It also includes time-invariant covariates such as distance and adjacency. It is possible to transform Equation (3) into a baseline specification that includes time-varying bilateral determinants, factors that are specific to the country of origin or destination, and time-invariant variables affecting FDI:

$$X_{ij,t} = \sum_{h=1}^{H} \alpha_h Z_{i,t}^h + \sum_{r=1}^{R} \alpha_r Z_{j,t}^r + \sum_{m=1}^{M} \alpha_m Z_{ij}^m + \sum_{k=1}^{K} \alpha_k Z_{ij,t}^k + \varepsilon_{ij,t} \tag{7}$$

where $X_{ij,t}$ represents FDI stocks from country i to country j in year t, $Z_{i,t}$ is a set of H variables which are specific for the country I, $Z_{j,t}$ is a set of R variables specific for the country j, Z_{ij} stands for the M time invariant variables, $Z_{ij,t}$ is a set of K time-varying variables for both countries, and $\varepsilon_{ij,t}$ is the error component. The structure of fixed effects (FE) determines which variables could be included in the regression to avoid collinearity. There are N cross sections units observed for T periods (1995–2018). An estimation caveat in Equation (7) is that the MRs are not directly observable. The following section discusses this further.

4. Data and Empirical Strategy

4.1. Descriptive Analysis

This analysis focuses on a panel dataset of 11 Latin American countries over the period 1995–2018. The countries in our sample are: Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Ecuador, Mexico, Paraguay, Peru, and Uruguay. Data on FDI stocks come from the United Nations Commission for Trade and Development (UNCTAD). Our panel is organized in dyads of countries. If country i does not invest in country j in time t, the correspondent observation is zero. This structure entails that a number of observations are zero. The World Trade Organization (WTO) provides extensive data on RTAs, and BIT data comes from UNCTAD. The rest of the gravity variables come from CEPII (Centre d'Etudes Prospectives et d´Informations Internationals), Penn World Tables 8.0, World Bank, and UNCTAD (see Table A3 in the online Appendix A for definition of variables and sources). The countries in the sample account for more than 90% of GDP and FDI within the region.

Figure 1 compares the countries in our sample in terms of their FDI within Latin America. The main investor in the area is Chile, followed by Mexico, Brazil and Argentina, while the main recipients of FDI are Brazil, Argentina, and Chile. Economies with high GDP have greater capacity of attracting FDI, in accord with the gravity model.

Figure 2 displays a slight negative correlation between the total numbers of BITs in force in a country and the total FDI (R-square = 0.0987). This suggests that the link between FDI and number of BITs is more complex than expected. Therefore, the use of number of BITs per country to capture the relationship between BITs and FDI presents shortcomings and, as established in the literature review, might reflect only a signaling effect. In the empirical approach the focus turns to an analysis of the impact of a BIT signed by a pair of countries (instead of the total number of BITs a country has signed over the period) and the evolution of FDI among that pair over time. In this way it is possible to capture more thoroughly the effectiveness of establishing a particular BIT between two countries.

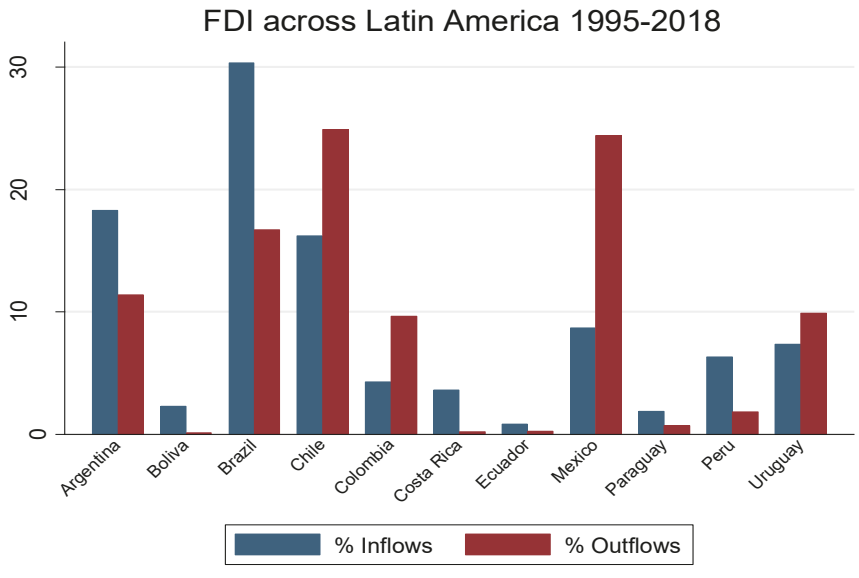


Figure 1. Intra-regional foreign direct investment (FDI) inflows/outflows in Latin America. Source: own elaboration. Vertical axis represents inflows/outflows by country over total FDI inflows/outflows in the area.

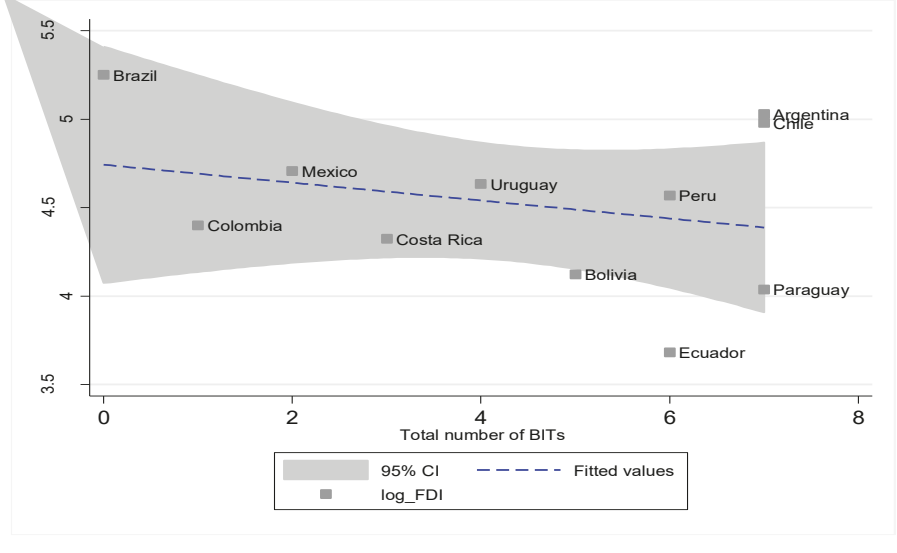


Figure 2. Intra-regional FDI vs number of bilateral investment treaties (BITs) in force (1995–2018). Source: Authors' calculations based on the United Nations Commission for Trade and Development (UNCTAD). Horizontal axis: total number of BITs in force in country j over the period 1995–2018. Vertical axis: sum of the stocks (constant dollars of 2012) received by each country j from the rest of countries in the sample.

The descriptive analysis offers some useful insights: it conveys the idea of differential patterns of intraregional FDI, a non-perfect matching between origin and destination countries across Latin America, and the need to control for size and other country characteristics.

4.2. Empirical Strategy: Estimation of the Gravity Equation for FDI

The general specification in Equation (7) raises the question about the appropriate estimation technique. One possible approach is to estimate the log-linearized gravity model by ordinary least squares (OLS). However, estimating using OLS presents several issues. First, the OLS estimation introduces bias associated with the presence of zero-FDI bilateral observations, due to the nonexistence of the natural log of zero. Therefore, estimating the model without taking into account zero observations generates biased estimated coefficients. Our FDI stock presents zeros since there are years and pair of countries that do not have FDI. Second, the log-linearization of the gravity equation changes the properties of the error term, leading to inefficient estimations in the presence of heteroscedasticity. FDI data are subject to heteroscedasticity, thus, estimating by OLS will give a log-linear residual that depends on the vector of covariates, generating inconsistent estimators.

Given the limitations that FDI presents regarding the existence of observations with zero values and the presence of heteroscedasticity, the seminal paper of Santos-Silva and Tenreyro [4] proposes to use the Poisson pseudo-maximum likelihood estimator (PPML). The PPML estimator circumvents the shortcomings of a linear model and estimates the gravity equation in its multiplicative form. Their study conveys that even when controlling for fixed effects, the presence of heteroscedasticity can generate strikingly different estimates when the gravity equation is log-linearized, rather than estimated in levels. The PPML is a special case of the Generalized Nonlinear Linear Model (GNLM) in which the variance is proportional to the mean. The authors show that this method is robust to different patterns of heteroscedasticity and resolves the inefficiency problem since it changes the distribution of the error term.

Despite the proven robustness of the PPML estimator, there are still some limitations, as heteroscedasticity might persist. The PPML estimator relies on the assumption that the variance is proportional to the mean ($\exp(x\beta)$), which may pose questions about its optimality. Additionally, PPML may present limited-dependent variable bias when a significant part of the observations is censored (which is not the case in this study).

Figure 3 sums up the main aspects of our empirical investigation.

Alternatively, and for robustness, in this study the structural gravity model is estimated using two alternative methods: Hausman-Taylor and the inverse hyperbolic sine transformation. The Hausman-Taylor method allows for the estimation of parameters of variables such as GDP, which vary only in a single dimension, and for the selection of variables considered endogenous in the model without uniquely controlling for heterogeneity by the structure of fixed effects. The inverse hyperbolic sine transformation is also adequate for estimating bilateral FDI across countries because its distribution is defined at zero.

To estimate the structural gravity model, the empirical analysis follows the methodology exposed in [3,40,45]. Therefore, this analysis includes fixed effects (FE), as opposed to random effects, since FE provide a better fit with samples encompassed by countries selected on a priori grounds [46]. It is also an appropriate technique to handle the unobservable heterogeneity potentially remaining among country pairs [47]. The variable BITs refers to those agreements that are in force, in line with the works of [13,23]. For the gravity variables, the study follows [39], who conclude that the variables more robustly correlated with FDI are the home and host country GDPs, RTAs, BITs, distance, and differences in endowments.

One of the problems when estimating Equation (7) is how to control for multilateral resistances. One natural way is to include fixed effects. Including time fixed effects and home and host fixed effects [45] is sufficiently adequate. Time fixed effects capture the business cycle, whereas country fixed effects control for all time invariant country characteristics. However, the omission of specific

effects capturing the bilateral interaction between countries could bias the estimation [40]. It is proposed to complement the main effects (time, home country and host country) with interaction effects, defined for country pairs and characterized for being time invariant, together with other country specific characteristics such as distance, contiguity, or common language.

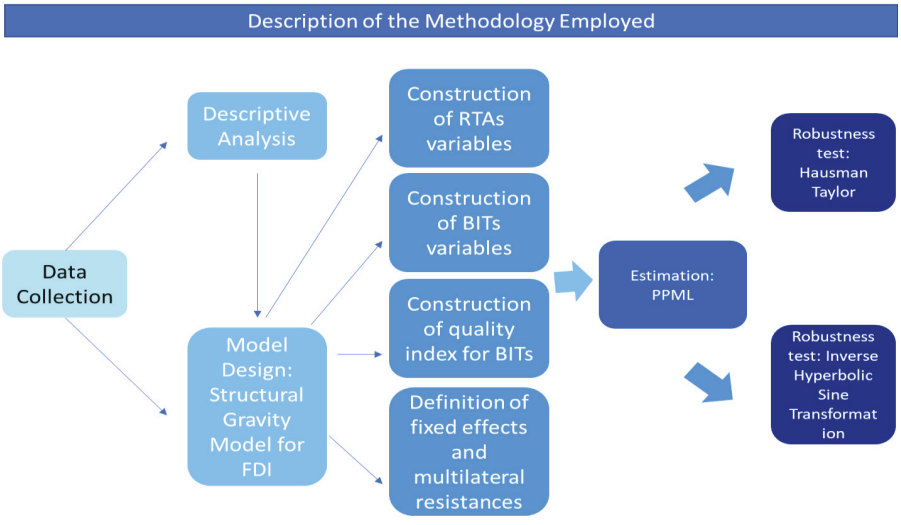


Figure 3. Empirical Model Description. Source: own elaboration. Notes: this figure displays some relevant features of our empirical investigation. Its starting point has been the collection of data from the sources detailed in Table A3. At first, descriptive analysis of the data has been carried out and discussed to inform the empirical model. The structural gravity model has been chosen as the most appropriate theoretical framework for the data. The next step has been the design and construction of different variables intended to capture the effects of regional trade agreements (RTAs) and BITs and the fixed effects. To design, code and compute these variables it was necessary to map all the agreements within Mercosur, The Latin American Integration Association (ALADI) and all the BITs, together with the design and computation of a quality index for BITs. The Poisson Pseudo-Likelihood Estimator (PPML) has been identified as the most convenient method of estimation for the model and it has been complemented with two additional robustness tests.

With respect to multilateral resistances, refs [3,48,49] suggest the inclusion of exporter time and importer time FE, taking into account that these will already capture the national output or expenditure of both countries.

In this setting, estimations may present endogeneity due to unobserved heterogeneity and reverse causality. The problem is tractable by alternative methods; this is how this study addresses this issue:

1. The analysis uses time invariant country pair fixed effects to correct for endogeneity due to unobserved heterogeneity in country pairs. In cross sections of country-pairs observed repeatedly over time, previous empirical studies suggest the inclusion of country-pair fixed effects in order to absorb the potential correlation between one or several regressors and the error term [13,47]. This way of controlling for country-pair and multilateral resistances with country-time fixed effects leads to estimates that can be interpreted as a difference-in-difference direct effect of the Mercosur, ALADI and BIT agreements on bilateral FDI.
2. If a group of countries interchange vast amounts of FDI, it is possible that they may be prompted to set up a RTA or a BIT. In this case, refs [3,47] recommend testing for endogeneity associated with reverse causation by including future leads of the RTA and BIT variables in the estimation. The empirical analysis also incorporates this approach in the estimations.

3. Additionally, the Hausman-Taylor method is used to control for endogeneity. Thus, it explicitly considers the Mercosur, ALADI and BITs variables as endogenous. Previous studies [46,50–52] apply this methodology to gravity models.

Therefore, Equation (7) can be estimated by PPML as:

$$X_{ij,t} = exp\left[\sum_{h=1}^{H} \alpha_h Z_{i,t}^h + \sum_{r=1}^{R} \alpha_r Z_{j,t}^r + \sum_{m=1}^{M} \alpha_m Z_{ij}^m + \sum_{k=1}^{K} \alpha_k Z_{ij,t}^k + \nu_i + \nu_j + \nu_t\right] + \varepsilon_{ij,t} \quad (8)$$

$$X_{ij,t} = exp\left[\sum_{h=1}^{H} \alpha_h Z_{i,t}^h + \sum_{r=1}^{R} \alpha_r Z_{j,t}^r + \sum_{m=1}^{M} \alpha_m Z_{ij}^m + \sum_{k=1}^{K} \alpha_k Z_{ij,t}^k + \nu_{i,t} + \nu_{j,t} + \nu_{ij}\right] + \varepsilon_{ij,t} \quad (9)$$

The difference between Equations (8) and (9) lies in the structure of the fixed effects. The analysis accounts for MR in Equation (9) with origin country-time (annual) and destination country-time (annual) FE. When we use country-year FE, the variables which proxy for the size of the countries drop out from the equation to avoid collinearity. The same effect applies for time invariant variables that are collinear with the country-pair FE structure.

$X_{ij,t}$ is the bilateral FDI stock in year t from country i to country j. The set of H observable variables $Z_{i,t}^h$ includes gross domestic product (GDP) in origin, and political risk in origin (we use this as control variable). The set of R observable variables $Z_{j,t}^r$ includes GDP in the recipient country, the sum of the GDP of the countries linked to the recipient by a trade agreement, and the political risk index of the recipient. The set of M observable variables Z_{ij}^m includes distance between the two countries, and adjacency (if both countries share a common border). Finally, the set of K observable variables $Z_{ij,t}^k$ includes our variables of main interest, capturing RTAs, Mercosur and ALADI, bilateral investment agreements, diversion effects, factor endowment differentials and labor cost differentials for each dyad of countries. Factor endowment differentials are computed as $\left|ln\left(\frac{K_{i,t}}{L_{i,t}}\right) - ln\left(\frac{K_{j,t}}{L_{j,t}}\right)\right|$, K being capital stock and L labor. The labor cost differential between pairs of countries is computed as: $\left|ln\left(\frac{Y_{i,t}}{L_{i,t}}\right) - ln\left(\frac{Y_{j,t}}{L_{j,t}}\right)\right|$, Y being gross domestic product.

The economic intuitions for the observable variables are as follows:

1. The GDP of the home and host country are expected to have a positive impact on FDI. In the case of the destination country, a larger level of income is tantamount to a more dynamic market.
2. The variable, $\left(\sum_j GDP_{j,t}^{RIA}\right)$, which is the sum of the GDP of the countries linked to the recipient by a trade agreement, is intended to capture the extended market effect [10]. Note that the gravity equation is an expenditure function and we must use variables in nominal terms, to avoid what is called the bronze medal mistake (see [3]).
3. The analysis uses two dyad variables, one to capture the FDI creation and the other the FDI diversion effect of Mercosur, in the spirit of [10,53]. The variables are Mercosur and One Mercosur, respectively. The Mercosur variable reflects the original treaty and the subsequent creation of free trade agreements with the associate members. Mercosur is a dummy which takes the value one when both countries, i and j, are part of Mercosur (either as original signatories, as an associated member or as signing a FTA with full scope, as described in the Appendix A) in a particular year t, and 0 otherwise. Since associated members incorporated in Mercosur in different years, it follows that the variable Mercosur exhibits time and country-pair variation.

The variable One Mercosur takes the value one when the recipient country j—the host country—belongs to Mercosur in the way defined above, while the country of origin of the flows, i, does not; it takes the value zero otherwise. Note that in this case the dummy equals one when the recipient country in the dyad belongs in the RTA as an original signatory, as an associated member (Chile, Colombia, Ecuador and Peru) or as a signatory of a non-partial scope FTA agreement within the Mercosur framework (Mexico from 2006 onwards) at a particular time. This captures the FDI diversion

effect (see [54]). FDI diversion occurs when investment flows from a Mercosur non-member country to a Mercosur member decline after that host country joined the RTA.

The variable ALADI follows the same construction as Mercosur. The variable ALADI reflects the FTAs (free trade zones) established over time across the countries within the ALADI framework. It takes value one for a pair of countries i, j —that were original signatories of ALADI—when the two countries entered into a free trade zone in time t (t being the year when the agreement entered into force) and zero otherwise. Therefore, the variable ALADI exhibits time and country-pair variation. The variable One ALADI is similar to One Mercosur.

4. The empirical analysis captures the effect *of* BITs in two ways. The variable BIT takes the value 1 if a ratified agreement is in place between the pair of countries at time t, and 0 otherwise. This variable captures the signaling effect. Second, the variable BIT-index is a continuous variable in the interval (0,1) which captures the degree of investment protection conferred by the said BIT. It is constructed in the spirit of [17,18].

To create the BIT investor protection index, it is necessary first to map all the BITs signed among the countries in our sample. BITs are classified according to 14 different clauses (see online Appendix A). We agree with the assessment of Berger et al. [55] in the sense that investors worry not only about dispute settlement arrangements but also about other aspects, such as policies on transfers of funds, treatment before and after the establishment, and performance requirements.

Ultimately, a sound dispute settlement mechanism is non-effective if the foreign firm in the host country is not profitable enough in the first place, because the treatment it receives prevents the consolidation of earnings and/or the contention of operational costs. In the sample, all BITs include ISDS provisions as well as State-State Dispute Settlement (SSDS) and similar clauses; focusing only on variations within these two clauses, therefore, would have given us less variability across BITs (for a complete definition of all clauses mapped see Appendix A).

Next, each indicator has a score associated with it. The score reflects how each particular aspect is covered in each BIT: 1 meaning maximum protection for investors and 0 non-protection. All individual scores add up (assuming equal weights, following [17,18,56]), to be normalized. Figure 4 displays the relationship between the BITs' pro-investment index and FDI.

The bulk of the treaties have a score of around 0.64–0.72, although there are a few values in the neighbourhood of 0.8. The median BIT protection index score is 0.68 and we use this value as a cut-off point (the average is 0.69). The assumption of equal weights could be problematic, but any other method of weighing individual indicators could also be contested.

5. To capture the enhancing or overlapping effects between trade agreements and BITs, the empirical analysis includes interactions in the model, i.e., Mercosur and ALADI are multiplied by the BIT investor protection index.
6. The relative availability and costs of inputs, which may have an important role in location decisions are FactorEndow and laborCost [9]. Following [57], FactorEndow measures the difference in the capital/labor ratio between the two countries. LaborCost captures the gap between countries in the price of labor. Because of the absence of good data on this issue, the analysis uses official data from the International Labor Organization (ILO). ILO assumes that real wages are equal to productivity, defined as the ratio GDP/number of workers.
7. The Political Risk variable accounts for the degree of consolidation of the rule of law, social, economic and political institutions and political stability. It has been constructed from the World Bank Aggregate Governance Indicator 1995–2018, complemented with the International Country Risk Indexes from Princeton University.

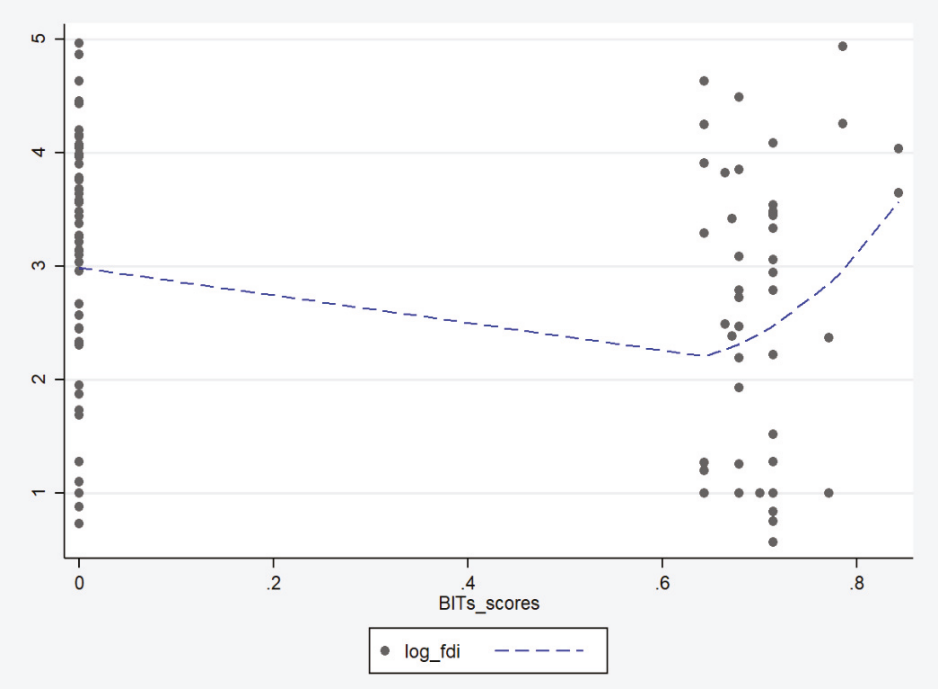


Figure 4. FDI intra-regional inflows per country (as percentage of total intra-regional inflows in Latin America) vs BITs pro-investor index (1995–2018). Source: Author's calculations using data from UNCTAD and WTO.

5. Empirical Results

Table 1 displays the results of the PPML estimation of the baseline gravity model described in Equation (8). In these regressions, the main focus lies on the impact of Mercosur, ALADI, and BIT related agreements. Regressions include fixed effects and time fixed effects to control for national heterogeneity and business cycles, respectively. The fixed effects absorb all observable and unobservable characteristics that are country-specific.

The variables which capture the purchasing power or market power of the host and home country are positive and significant, as expected. Their signs and magnitudes are similar to those reported by [39,58] in their meta-analyses of robust FDI determinants. The empirical studies analogous to ours use a large variety of samples, methodologies, and specifications. This implies some degree of heterogeneity in the size of the point estimates of the main variables, albeit results are typically the same in spirit. Therefore, both the market size of the investor and, more importantly, according to the point estimate, the market of the host country exerts a large influence on the decision to invest in the countries in our sample. This result makes economic sense: a big market represents a dynamic expected demand by potential consumers which, in turn, entails a higher level of revenues for firms offering that particular good or service. In addition, in this kind of setting, fixed costs related to the establishment of a firm in a foreign market can be supported by a larger number of products, hence reducing total costs per unit.

The market potential is captured by the sum of the GDPs of the countries belonging to the same RTA as the host country. This variable exhibits a positive and significant point estimate, in the neighborhood of 0.9, which is similar or even slightly higher than the coefficient of the host country

GDP. The analysis of Yeyati et al. [10] also finds a positive and significant impact of the extended market on FDI. The results regarding this variable suggests that foreign firms choose their location in Latin America not only to satisfy the local market of the destination country but also to access their neighbors more easily via an export platform strategy, as in [13,21]. Furthermore, these findings provide preliminary evidence about the effectiveness of the trade agreements in attracting FDI from third countries.

Distance is negative and significant while adjacency is positive and significant, as expected, with magnitudes similar to those reported by [23,58,59]. The latter find a lower point estimate for distance, between −0.38 and −0.46, but the impact of contiguity is quite akin to ours. A recent paper [60] reports distance estimates between −0.68 and −1.5. Although their level of analysis is slightly different, since they work with disaggregated data by industries and with a sample of 243 countries, it is reassuring that our distance coefficients are closer to theirs than those reported by [23]. Our results of magnitude and signs of traditional variables included in gravity models are thus comparable to those displayed in the literature.

The variable Mercosur is positive, significant and quite stable, with point estimates that vary around 0.136–0.147. We can recover the impact of this dummy from the expression $\left[e^{\hat{\beta}} - 1\right]$. A point estimate of 0.147 (column 4), means that, since Mercosur entered into force, FDI flows to its members have increased on average by 15.83% per year. Thus, the effect of Mercosur ranges between 14.56% and 15.83% per year. This is similar to the result reported by [23] for Economic Integration Agreements and Custom Unions (0.11), and also to [7] when they perform their estimation using PPML. It is also in line with the results reported by [3] but is smaller than the 0.41–0.5 point estimate documented by [13] for European agreements. The difference in the estimates found by [13] and by this paper can be attributed to two reasons: firstly, European countries have been traditionally very connected; secondly, European agreements bring about very strong ties among their members, implying ultimately free movement of goods, services and resources, to an extent not reached yet by Latin American treaties.

The variable One Mercosur displays a negative and significant coefficient. The joint consideration of the Mercosur and One-Mercosur variables indicates that, when the host country enters Mercosur, FDI within the members of the blocs increase, but FDI between a member and a third country declines. This sort of FDI diversion is not negligible since it amounts to 8.43–6.60%. Empirical evidence for this phenomenon can be found in [10]. The dummy capturing the ALADI agreements is positive, although less significant than Mercosur. The point estimate of ALADI is also considerably smaller, around 0.049–0.062. Thus, the effect of ALADI on FDI is between 5.02 to 6.39%. The diversion effect of ALADI is only significant at the 10% level (column 4) and amounts approximately to 3%. In the rest of the estimations the coefficient of ONE_ALADI is non-significant.

The different degrees of integration provided by the agreements can explain the dissimilar creation and diversion effects of Mercosur and ALADI. Mercosur is a consolidated, common market RTA that has promoted a certain degree of stability within the region. The FTAs within ALADI, (in contrast to the Mercosur type agreement) do not necessarily exert a significant impact on the macroeconomic environment of the host country, nor contribute to the harmonization of legislation and standards between them. Their impact on FDI, therefore, while relevant, is more subdued than that of Mercosur. As mentioned above, ref [13] estimate this effect to be around 0.4–0.5 for European agreements, thus lending countenance to our claim that the impact of a trade agreement is positively correlated with the degree of integration it provides.

Table 1. Estimates of RTAs and BITs on FDI stocks. Gravity-PPML with country and time fixed effects.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$Y_{i,t}$ (log)	0.726		0.764	0.701		0.693	0.752		0.628
	(0.000) ***		(0.003) ***	(0.006) ***		(0.007) ***	(0.003) ***		(0.005) ***
$Y_{j,t}$ (log)	0.815		0.804	0.831		0.809	0.802		0.813
	(0.002) ***		(0.003) ***	(0.003) ***		(0.004) ***	(0.002) ***		(0.003) ***
$\left(SumGDP_{j,t}^{RIA}\right)(log)$		0.935	0.947		0.895	0.906		0.926	0.985
		(0.490) **	(0.486) **		(0.488) **	(0.475) **		(0.468) **	(0.506) **
Distance_{ij} (log)	−0.702	−0.713	−0.724	−0.743	−0.782	−0.736	−0.801	−0.751	−0.762
	(0.104) ***	(0.096) ***	(0.109) ***	(0.113) ***	(0.095) ***	(0.079) **	(0.108) ***	(0.123) ***	(0.114) ***
Adjacency_{ij}	0.216	0.261	0.272	0.260	0.206	0.214	0.236	0.221	0.274
	(0.065) ***	(0.053) ***	(0.048) ***	(0.047) ***	(0.052) ***	(0.050) ***	(0.044) ***	(0.039) ***	(0.049) ***
$\text{Mercosur}_{ij,t}$	0.143	0.146	0.138	0.147	0.136	0.142	0.136	0.138	0.140
	(0.033 ***	(0.031) ***	(0.029) ***	(0.026) ***	(0.027) ***	(0.022) ***	(0.019) ***	(0.017) ***	(0.021) ***
$\text{ALADI}_{ij,t}$	0.061	0.049	0.055	0.062	0.056	0.053	0.054	0.060	0.049
	(0.024) **	(0.026) *	(0.029) *	(0.030) *	(0.031) *	(0.026) **	(0.024) **	(0.026) **	(0.027) **
$\text{BITs}_{ij,t}$	0.086		0.082	0.074	0.069	0.065	0.063		0.052
	(0.046) *		(0.044) *	(0.039) **	(0.036) *	(0.035) *	(0.033) *		(0.028) *
$\text{BIT index}_{ij,t}$		0.045	0.043			0.042		0.039	
		(0.024) *	(0.023) *			(0.022) *		(0.021) *	
$\text{ONE\_Mercosur}_{ij,t}$				−0.081	−0.076	−0.074	−0.070	−0.065	−0.064
				(0.022) ***	(0.032) ***	(0.026) ***	(0.024) ***	(0.028) ***	(0.031) ***
$\text{ONE\_ALADI}_{ij,t}$				−0.026	−0.023	−0.028	−0.025	−0.018	−0.021
				(0.025)	(0.023)	(0.032)	(0.037)	(0.042)	(0.041)
$\text{Factor Endow}_{ij,t}$							0.104	0.113	0.107
							(0.031) ***	(0.039) ***	(0.042) ***
$\text{Political Risk}_{j,t}$							−0.043	−0.052	−0.053
							(0.010) ***	(0.009) ***	(0.014) ***
$\text{Labor Cost dif}_{ij,t}$							0.037	0.042	0.045
							(0.020) *	(0.0321) *	(0.022) *
No. Observations	2230	2230	2230	2230	2230	2230	2230	2230	2230
Adj. R^2	0.537	0.521	0.546	0.557	0.560	0.529	0.587	0.579	0.601
Individual Country Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time Fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: This table reports panel gravity estimates for FDI stocks from country i to country j for 1995–2018. It reports Poisson pseudo-maximum likelihood estimation with individual country effects and time FE. Robust standard errors are in parenthesis and clustered by country pair. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

The variable capturing BITs is positive and significant at least at the 10% level in all specifications. The point estimate ranges between 0.052 and 0.086, indicating a yearly effect between 5.3% and 8.5%, consistent with results of previous studies employing a similar methodology [6]. Qualitative results are similar in spirit to those of [7]. The main message here is that BITs also enhance FDI inflows to host countries; The differences in the point estimates of the BIT indicator in their contribution and ours can be attributed to alternative econometric specifications, since we use PPML as our baseline technique and they do not. Interestingly, the point estimate is smaller than the coefficients obtained for Mercosur, but larger than that of ALADI. The variable that captures the impact of the quality of BITs (the BIT index) on FDI is significant and ranges between 0.039 and 0.045. The economic interpretation of this result is straightforward. Those treaties granting a higher level of protection to foreign investments, in the form of less stringent establishment prerequisites, assurance of fair and equitable treatment, allowance for transfer of funds, and design of sound dispute settlements mechanisms, among others, facilitate the smooth operation of foreign firms and hence help attract new investment.

The effect of the variable capturing differences in factor endowments is positive and highly significant. Intuitively, if labor is relatively more abundant in the host country than in the home country, it will also be cheaper, thus creating incentives for foreign firms intending to rationalize their employee costs. Labor cost differentials are only marginally significant. The economic interpretation of this variable is the same as for differences in factor endowments but its effect is captured with less precision. This variable is constructed as the difference in GDP over total employment in the home and the host countries. In other words, it captures differences in productivities, which should be closely linked to dissimilarities in labor costs, under the hypothesis of efficient labor markets. This assumption, however, may not hold in some developing countries, thus negatively affecting the accuracy of the indicator.

The political risk variable exerts a significant and negative impact on FDI stocks. Since it proxies for the institutional environment, our results suggest that more stable countries attract higher amounts of FDI, while political unrest acts as a deterrent. Countries with larger degrees of macroeconomic and political stability and well-established political institutions provide more predictable environments and reduce the uncertainty associated with new investments, thus fostering the attraction of FDI. The size of effect is akin to that reported by [28,58], in this last case for Middle East and North African countries, respectively.

These results support the hypothesis of an intra-regional FDI in Latin America driven both by considerations about market size and relative endowments. These findings suggest, therefore, that the pattern of FDI within Latin America can be represented by a hybrid, knowledge capital model in the particular case of the export-platform strategy [13,21]. Intuitively, according to this analytical framework, firms make decisions taking into account the opportunities provided by a potentially dynamic demand for the goods they offer (as proxied by market size), and by differentials in input costs (captured by relative endowments), which translate into more efficient production processes.

Table 2 summarizes the results obtained from the estimation of Equation (9). Now the structure of the fixed effects has changed and includes country-time fixed effects to account for multilateral resistances. This specification [47,48,61] addresses the potential endogeneity in the model by using country-pair fixed effects (or first-differencing) in panel data structures. This structure of fixed effects (country-year and country-pair) absorbs the host-home GDPs, the market purchasing power, and the distance and adjacency variables.

The estimated coefficient for the BIT index is positive and significant at the 90% level. The point estimate is around 0.042. Both the coefficients of the dummy for the BIT and the degree of protection index are quite similar to the results obtained by [18] when they work with bilateral FDI flows. We have estimated the model specified in Table 2 including diversion effect variables and the results are similar to those in Table 1. The magnitude of the coefficients slightly changes but the interpretation remains. Since we are working now with a continuous function over an interval, instead of usinga dummy, the index allows the testing of several interesting hypotheses. It is possible, as ref [17] show, that the relationship between BIT–FDI varies for different values of the index (notice that the

score attributed to each BIT does not change once the BIT enters into force unless withdrawal occurs; in this framework, BIT index is equivalent to BIT index t-5). In order to test this hypothesis, we have considered two scenarios and introduced them separately in the equations: a value of the index smaller than 0.68 (meaning that the corresponding BIT is less pro-investor) and an index larger than 0.68 (which represents a treaty with a more pro-investor orientation).

Results suggest that the impact of the index does change with its value: those BITs whose indexes show a less favorable stance for the investor are non-significant whereas those which are more pro-investor are positive and significant at 90% (see column 3 and 5). Notice, however, that the impact of BITs (as captured either by the variable BITs or BIT index) is still lower than that of Mercosur but larger than ALADI's, as in the previous specifications.

In order to further explore the heterogeneity suggested by the index, the model includes interactions of the BIT index with the proxies for the RTAs. When interacted with Mercosur, the pro-investor index is positive and significant at 99%, with a point estimate of 0.068. The result carries over to the subset of BITs with higher index values (with an impact of 7%). Instead, the interaction of ALADI with the index, although showing the expected signs, is not significant. A higher value of the index combined with participation in Mercosur is, hence, more effective in attracting FDI than just a higher index score (columns 4 and 5). The interaction term, both positive and significant, indicates that the effect of Mercosur is enhanced for higher values of the BIT investment protection variable.

Table 2. Impact of RTA and BIT investor protection index on FDI. Gravity-PPML estimation with country-year and country-pair FE.

	Stock FDI$_{ij,t}$				
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
Mercosur	0.129 (0.018) ***	0.122 (0.023) ***	0.115 (0.019) ***	0.103 (0.013) ***	0.096 (0.010) ***
ALADI	0.030 (0.016) *	0.032 (0.017) *	0.027 (0.016) *	0.025 (0.014) *	0.031 (0.013) *
BITs	0.042 (0.021) **			0.039 (0.019) **	0.034 (0.019) **
BIT index		0.034 (0.018) *		0.039 (0.019) *	
BIT Index < 0.68 (less pro−investor)			0.026 (0.029)		0.018 (0.036)
BIT Index ≥ 0.68 (pro−investor)			0.037 (0.020) *		0.042 (0.019) *
Mercosur * BIT Index				0.062 (0.013) ***	
Mercosur * BIT Index ≥ 0.68					0.068 (0.020) ***
Aladi*BIT Index				0.023 (0.017)	
ISDS concluded				−0.008 (0.023)	−0.007 (0.029)
Controls	Yes	Yes	Yes	Yes	Yes
No. Observations	2230	2230	2230	2230	2230
Individual country−year fixed effect	Yes	Yes	Yes	Yes	Yes
Country−pair fixed effect	Yes	Yes	Yes	Yes	Yes

Note: This table reports panel gravity estimates for FDI stocks from country i to country j for the period 1995–2018. It reports Poisson pseudo-maximum likelihood estimation with country time FE (η_{it}, η_{jt}) and pair country FE (η_{ij}). We control for the effect of country outliers as part of FE. We control for labor cost differentials, endowment differences, and political risks. Country pair fixed effects are used to address the issue of endogeneity in RTAs and BITs [47]. Robust standard errors are in parenthesis, clustered by country pair. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Panel A of Table 3 presents two additional specifications of the basic estimation in order to test for the impact of the policy variables. The first column includes a country-pair fixed effects structure

and estimates coefficients for the leads of the policy variables to test for endogeneity due to reverse causation [3,47]. In other words, this specification aims to assess whether abundant FDI among some countries prompts the constitution of a trade agreement among them, so investment at time t translates into RTAs or BITs at time t + h. If RTAs/BITs are exogenous to FDI, then the estimated coefficient for the lead variables should not be significantly different from zero. Indeed, the results show a non-significant effect of the lead variables.

The lags of the Mercosur, ALADI and BITs capture the impact of phasing-in effects. Now the question is whether trade and investment agreements take some time to fully display their effects. It is reasonable to think that firms need a period of adaptation to changing circumstances in order to consider alternatives and make location decisions. The lagged coefficients, also positive and significant, display a significant impact on FDI, meaning that their effect comes about gradually. We have worked with five lags for several reasons. First, effects of trade and investment agreements seem to peak in four–five years. Second, we want to capture the heterogeneous effects of implementation processes across countries. These results are consistent with [12]. In general, the impact of (lagged) Mercosur lies at around 13.42%. The effect of (lagged) BITs is, again, smaller than that of Mercosur but larger than that of ALADI. These results are in accord with the reported pattern in [3]: the lagged impact of Mercosur, ALADI and BITs is larger than the contemporaneous impact. In other words, medium- and long-term effects of these treaties exceed short-term effects.

Panel B of Table 3's estimations presents the results using the Hausman-Taylor method to further address the possible endogeneity of the policy variables. This method uses the deviations from the mean of the exogenous time-varying variables for each country pair as instruments [62]. This procedure removes the part of the error term correlated with the endogenous time-varying variables. Since the deviations from the group means are by definition uncorrelated with the error term, they do not alter the estimation of the exogenous time varying variables when estimating the coefficient of the rest of the variables.

The main policy variables are treated as endogenous. The estimated coefficients are slightly higher in magnitude than those obtained by PPML with country-time and country-pair fixed effects. These results are consistent with the larger coefficients obtained by [3,47] when estimating the effects of RTAs on exports for a different set of countries. The Hausman test gives us a chi-square tests with a probability below 0.05. Therefore, we reject the null hypothesis of random effects. The instruments pass a conventional test for overidentifying restrictions. As a robustness check the online Appendix A presents estimations using an inverse hyperbolic sine transformation of the FDI dependent variable.

It is plausible that BITs exert heterogeneous effects not only because of their own characteristics but also due to the nature of the signatories in terms of economic and institutional development. To test this hypothesis, a further analysis first classifies the countries in our sample in two categories, according to their income level, and works with subsamples.

Table 4 reports very suggestive results. The impact of BITs per se is large in the subsample encompassed by medium income countries, and non-significant for middle-low income countries. The BIT index is positive and significant in the first subsample, but only marginally significant in the second.

Similarly, a BIT index above 0.68 is significant for middle income countries; it is only marginally significant, however, for middle-low income countries. When the BIT index is below 0.68, it is only significant at the 90% confidence level for the first subsample and non-significant for the second. The interaction between Mercosur and BITs displays results consistent with previous estimations (Table 2). Notice that the interaction term is also positive and significant for middle-low income countries. This model specification seems to confirm the hypothesis of complementarity between BITs and institutions. BITs are primarily effective in reducing the perceived risk by investors, and thus help attract FDI, when the host country has sounder and more stable institutions. From an economic point of view, a more developed institutional framework in a particular country provides credibility to the

agreements signed by that country, enhancing their efficacy. In [10], the authors find that the effect of RTA is higher in those countries which are more attractive for FDI.

Table 3. Impact of RTAs and BITs on FDI. PPML and Hausman-Taylor Estimations.

	Panel A: PPML		**Panel B: Hausman−Taylor**		
	(1)	**(2)**	**(1)**	**(2)**	**(3)**
$Y_{i,t}$ (log)			0.564		0.537
			(0.062) ***		(0.071 ***
$Y_{j,t}$ (log)			0.792		0.805
			(0.069) ***		(0.042 ***
$\left(SumGDP_{j,t}^{RIA}\right)(log)$				0.901	0.875
				(0.489) **	(0.425) **
Distance (log)			−0.503	−0.498	−0.485
			(0.257) **	(0.250) **	(0.248) **
Adjacency			0.273	0.253	0.261
			(0.081) ***	(0.073) ***	(0.062) ***
Mercosur	0.123	0.118	0.153	0.145	0.169
	(0.032) ***	(0.026) ***	(0.070) ***	(0.062) ***	(0.073) ***
ALADI	0.031	0.026	0.056	0.048	0.052
	(0.016) *	(0.014) *	(0.028) *	(0.025) **	(0.026) **
BITs	0.047	0.039	0.063		0.072
	(0.024) **	(0.020) **	(0.032) **		(0.035) **
BIT index				0.041	0.046
				(0.022) *	(0.021) *
Mercosur_LEAD5	0.072				
	(0.048)				
ALADI_ LEAD5	0.015				
	(0.032)				
BITs_ LEAD5	0.029				
	(0.054)				
Mercosur_LAG5		0.126			
		(0.011) ***			
ALADI_ LAG5		0.038			
		(0.020) *			
BITs_ LAG5		0.046			
		(0.021) **			
Controls	Yes	Yes	Yes	Yes	Yes
No. Observations	1896	1896	1670	1670	1670
Adj. R^2			0.426	0.443	0.521
Sigma_u			0.729	0.931	0.834
Sigma_e			1.317	1.320	1.320
Rho (fraction of variance due to u_i)			0.234	0.332	0.285
F−stat overidentification restriction			1.345	1.421	1.415
			(0.426)	(0.432)	(0.447)

Note: Panel A represents Poisson pseudo-maximum likelihood estimation with country time FE $(\eta_{it,}\ \eta_{jt})$ and pair country FE (η_{ij}). Country pair fixed effects are used to address endogeneity in RTAs and BITs (see [47]). As control variables we include labor cost differentials, endowment differences, and political risk. Panel B reports Hausman-Taylor estimates for bilateral FDI in logs. GDPs and sum GDP are considered as time-varying and exogenous. Distance is considered as endogenous, time-invariant (proxy for trade costs). Adjacency is considered exogenous, time-invariant. Mercosur, ALADI, and BITs are considered as endogenous, time-varying. The F-statistic should not be different from zero, and refers to the test for over-identifying restrictions in the corresponding log-linear instrumental variable (IV) model (p-value of over-identifying restrictions in parenthesis). Robust standard errors in parenthesis, clustered by country pair *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 4. Impact of RTA and BIT pro-investor index for country subsets. Gravity-PPML, with country-time and country-pair fixed effects.

	Middle Income Countries				**Middle-Low Income Countries**			
	(1)	**(2)**	**(3)**	**(4)**	**(1)**	**(2)**	**(3)**	**(4)**
Mercosur	0.106	0.097	0.115	0.119	0.092	0.101	0.095	0.087
	(0.016)***	(0.023)***	(0.020)***	(0.016)***	(0.024)***	(0.013)***	(0.021)***	(0.025)***
Aladi	0.036	0.042	0.031	0.028	0.030	0.019	0.022	0.015
	(0.020)*	(0.021)*	(0.025)	(0.015)*	(0.016)*	(0.011)*	(0.012)*	(0.018)
BITs	0.039	0.036		0.041	0.032	0.027		0.037
	(0.022)*	(0.020)*		(0.022)*	(0.045)	(0.038)		(0.042)
BIT Index		0.043				0.029		
		(0.020)**				(0.015)*		
BIT Index < 0.68	0.023		0.028	0.019	0.021		0.025	0.018
	(0.011)*		(0.012)*	(0.011)*	(0.018)		(0.029)	(0.027)
BIT Index ≥ 0.68	0.042		0.045	0.047	0.036		0.039	0.042
	(0.021)**		(0.020)**	(0.021)**	(0.019)*		(0.021)*	(0.019)**
Mercosur * BIT Index		0.055				0.057		
		(0.027)**				(0.025)**		
Mercosur * BIT Index ≥ 0.68	0.057		0.059	0.055	0.042		0.040	0.036
	(0.019)***		(0.012)***	(0.010)***	(0.011)***		(0.013)***	(0.015)***
Aladi*BIT Index		0.017	0.018	0.014		0.021	0.019	0.023
		(0.030)	(0.037)	(0.026)		(0.010)*	(0.009)*	(0.011)*
ISDS concluded		−0.018	−0.012	-0.015		−0.006	−0.009	−0.011
		(0.047)	(0.052)	(0.061)		(0.039)	(0.042)	(0.058)
Mercosur_LAG5			0.084	0.093			0.090	0.092
			(0.012)***	(0.024)***			(0.022)***	(0.031)***
Aladi_LAG5			0.023	0.016			0.014	0.012
			(0.008)*	(0.009)*			(0.007)*	(0.006)*
BITs_LAG5				0.039				0.030
				(0.020)**				(0.014)**
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adj-R^2	0.439	0.520	0.561	0.536	0.410	0.421	0.438	0.446
Wald Test	32.15	31.29	31.42	27.31	28.22	31..05	27.31	32.91
Observations	1160	1160	1042	1042	960	960	852	852

Source: Own elaboration. Notes: Middle income countries division according to GDP per capita in constant dollars using the World Bank's approach. First group of countries includes: Argentina, Brazil, Chile, Costa Rica, Mexico, and Uruguay. Middle-low income countries include: Bolivia, Colombia, Ecuador, Peru, and Paraguay. Dependent variable accounts for bilateral FDI stocks received in country j (subset) from all 11 economies. All models include origin and destination time fixed effects to control for MR and country-pair FE to control for endogeneity. We control for labor cost differentials, endowment differences, and political risk. Country-pair clustered robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Wald test p-value for the difference of coefficients in both samples is 0.00.

6. Concluding Remarks

When it comes to institutional strategies promoting trade agreements and bilateral investment treaties to attract FDI to areas in development, there are still empirical questions that remain unclear. This study is novel in addressing the following questions: (i) Does the effect on FDI depend on key characteristics (clauses) of the BITs? (ii) If there is co-existence of BITs in force with major trade agreements, which ones are more effective in attracting foreign investment? Are there overlapping or enhancing effects? (iii) Finally, it is not clear if BITs have a differential impact in countries with weak versus strong institutions. This paper uses a flexible econometric specification derived from a structural gravity model of FDI to assess the impact of RTAs and BITs on intra-regional FDI stocks in Latin America over the years 1995–2018.

The analysis shows that the participation in deep integration agreements, such as Mercosur, has the strongest impact on intra-regional FDI inflows, of between 14.56% and 15.83%. Taking part in less ambitious trade agreements as ALADI does help attract FDI inflows but the effect is more modest and amounts to 3–5% on average. Finally, BITs enhance the attraction of foreign investment, with an average effect of 3.7% and 4.6%, which lies in between those estimated for Mercosur and ALADI.

To disentangle the mere presence of a BIT between two countries from that associated with the quality of the BIT, we have constructed an index in order to capture the level of protection provided to foreign investors by a BIT. The estimations suggest that higher levels of protection are associated with greater capacities of attraction of FDI in Latin America, when the index is above a certain threshold. Additionally, the analysis shows that when RTAs are combined with investment agreements offering a high degree of protection, they help attract foreign investment within Latin America.

Not only the features of the BIT influence the capacity to attract foreign investment. The institutional characteristics of the host country are crucial as well. According to estimations, BITs are associated with larger FDI inflows in a subsample made up of the countries with the highest income. For middle-low income countries (Bolivia, Colombia, Ecuador, Paraguay, and Peru), however, the mere presence of BITs per se is not associated with an increase in FDI. Furthermore, the impact of higher quality BITs is smaller and statistically less significant for middle-low income countries than for more developed nations. Thus, BITs appear as complements of a sound institutional environment in the host country.

These findings have implications for policy making. First, belonging to solid, consolidated trade agreements which imply high levels of integration is beneficial not only from the point of view of external trade but also because it helps attract FDI inflows. Countries seeking higher levels of foreign investment might want to consider membership in these types of accords, together with a reinforcement of the treaties they have already signed.

Second, empirical analysis has shown that BIT exerts a signal effect but its impact is conditional upon the degree of integration and consolidation of the RTA, the level of development—or institutional advancement—of the host country, and the extent to which a particular BIT protects foreign investment. Policy makers interested in attracting FDI should engage in investment treaties that warrant a reasonable degree of protection to foreign investors. The signing of these agreements may entail a costs, in the form of restrictions to the autonomy of the host country, for example. In line with other contributions our results imply that policymakers should weigh carefully the trade-offs associated with entry into a BIT. Otherwise, the political costs associated with the signing of these agreements may not be worthwhile. Finally, Latin American economies striving to attract foreign investment should keep in mind that the combined effect of belonging to consolidated RTAs together with BITs enhances FDI.

The main limitation of this paper is the use of aggregate data instead of disaggregated FDI by sector. It would be interesting to ascertain whether the impact of Mercosur, ALADI and BITs varies across different industries. Further research is necessary to gauge in more detail the potentially different effect of treaties according to their sectoral distribution, but unfortunately sectoral/industry data are not yet available on a bilateral basis at a cross-country level for Latin America.

Author Contributions: M.B.: conceptualization, data curation, formal analysis, methodology, writing, review and editing. B.S.-R.: conceptualization, investigation, methodology, writing, review and editing. Y.S.: conceptualization, investigation, writing, review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We thank the CIRANO data lab, Thierry Warin, Joseph Pelzman, Zadia Feliciano, Mona Pinchis-Paulsen, Federico Ortino, Thilo Huning, and Kevin Foster for helpful comments. David Dam provided excellent research assistance. We are grateful to participants in the VIII La Laguna Workshop on International Economics 2019, the Global Research on Emerging Economies Conference 2018, the Lipsey Panel at the Western Economic Association 2017, the International Panel Data Conference 2016, the Workshop of Economic Integration, the European Trade Study Group 2016, the International Trade and Finance Association meeting 2016 and seminar participants at Department of Economics at University of Cape Town, SARChI and Dept. of Economics at University of Johannesburg, CIRANO, and the Institute of Economics at University of

Republic (Uruguay) for the comments received during presentations of previous versions of this work. All errors are our own.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ALADI	The Latin American Integration Association/Asociación Latinoamericana de Integración
ASEAN	Association of Southeast Asian Nations
BITs	Bilateral investment treaties
CEPII	French Centre d'Etudes Prospectives et d'Informations Internationales
CUSFTA	The Canada-United States Free Trade Agreement
FDI	Foreign direct investment
FE	Fixed effects
FET	Fair and equitable treatment
GDP	Gross Domestic Product
GNLM	Generalized Nonlinear Linear Model
IHS	Inverse hyperbolic sine transformation
ILO	International Labor Organization
IMF	International Monetary Fund
ISDS	Investor-state dispute settlement
Mercosur	Mercado comun del sur
MFN	Most-favored-nation clause
MNEs	Multinational enterprises
MRs	Multilateral resistances
NAFTA	North American Free Trade Agreement
NT	National treatment
OECD	The Organisation for Economic Co-operation and Development
OLS	Ordinary least square
PPML	Poisson pseudo-maximum likelihood
RTAs	Regional trade agreements
SSDS	State-State dispute settlement
UNCTAD	United Nations Commission for Trade and Development
WTO	World Trade Organization

Appendix A

Appendix A.1. Clauses Mapped in the BIT Investment Protection Index

Individual provisions are coded for all the BITs between the 11 Latin American countries in our sample, i.e., 14 provisions that UNCTAD and the American Bar Association consider of substance to protect any investment to which a developing host country is a signatory. Low scores indicate less investment protection. Values of 0.5 are assigned to medium level investor protection [17,18].

1. National treatment (NT) pre-establishment: Ensures that requirements for foreign firms upon entry in the host country, such as establishment and participation in existing enterprises, are no greater than those for domestic firms. If the clause is present, the index in this section sums 1 in this category, 0 otherwise.
2. National treatment post establishment: The same as 1 but associated with "the treatment of the investment after its entry". If present, the index in this category takes value 1, 0 otherwise.
3. Most-favored-nation (MFN) treatment pre-establishment: MFN treatment of the foreign firm regarding entry, establishment and participation in existing enterprises. If present, the index in this category takes value 1, 0 otherwise.
4. Most-favored-nation (MFN) treatment post-establishment: The same as 3, after the entry of the foreign firm. If the clause is present it takes value 1.

5. Fair and equitable treatment (FET): Can be qualified either by reference to International Law (General International Law, Principles of International Law or Customary International Law) or "by listing the elements of the FET obligation". In the last case, the FET obligation may "include an indicative or exhaustive list of more specific elements" to avoid.
6. Full protection and security: This commitment, in turn, may be:

    - Standard "if the treaty contains an unqualified obligation to provide full protection and security" (with formulations such as most constant protection, legal protection and security, and so forth).
    - Referenced to domestic law of the host country

7. No general security exception: Ensures that the host country does not prevent investment in a particular sector for security reasons.
8. Indirect expropriation: Treaties may refer to this issue under two approaches: the scope of measures covered, and/or refining expropriation clauses.

On the scope of measures covered: The options under this classification are the following:
Indirect expropriation not mentioned "if the treaty's expropriation clause does not contain an explicit reference to indirect expropriation".

- Indirect expropriation mentioned, whatever the formulae it employs ("measures having effect equivalent to nationalization or expropriation", measures tantamount to expropriation, de facto expropriation).
- No expropriation clause "if the treaty does not include a provision that protects foreign investors against non-compensated dispossession of their investments".

9. Transfer of funds: A "provision regarding the free transfer of funds relating to investments (covering outward and/or inward transfers". If present, the index in this category takes a value of 1.
10. Performance requirements: If the treaty includes a provision that restricts the use of performance requirements, the measure in these clause takes value 1.
11. Umbrella clause: requiring the signatories "to respect or observe any obligation assumed by it with regard to a specific investment", hence protecting it de facto under its umbrella. The measure in these clause takes value 1.
12. State-State Dispute Settlement (SSDS): If the treaty provides for a dispute settlement procedure (e.g., arbitration) between States, the measure in this clause takes value 1.
13. Investor-State Dispute Settlement (ISDS): If the treaty establishes a mechanism for the settlement of disputes between covered investors and the host State (arbitration and/or domestic courts of the host State) the measure in these clause takes value 1.
14. Alternatives to arbitration: The more options at investors' disposal the better, although we assign a value of 0.5 to a treaty that establishes that the investor needs to go first through a local court before international arbitration.

    - "Voluntary Alternative Dispute Resolution (conciliation/mediation)" "If the treaty mentions the possibility of such procedures (e.g., non-binding, third-party procedures) but does not prescribe them as a necessary step".
    - "Compulsory Alternative Dispute Resolution (conciliation/mediation) If the treaty prescribes the use of compulsory conciliation or mediation.
    - "None" if the treaty does not refer to alternative means of settling investor–State disputes (conciliation/mediation or similar non-binding procedures).

Appendix A.2. Construction of the Mercosur and ALADI Variables

The sequence of the signature of the FTAs within Mercosur, considered deep integration agreements, and the incorporation of associated members is as follows:

1. All Mercosur original countries (Argentina, Brazil, Paraguay and Uruguay) and Chile entered into a FTA (Free Trade Agreement named AAP.CE number 35) in 1996.
2. All Mercosur original countries and Bolivia entered into a FTA (Free Trade Agreement AAP.CE number 36) in 1997.
3. All Mercosur original countries and Colombia and Ecuador entered into a FTA (Free Trade Agreement named AAP.CE number 59) in 2005.
4. All Mercosur original countries and Mexico entered into a FTA (Free Trade Agreement named AAP.CE number 54 and 55) in 2006.
5. All Mercosur original countries and Peru entered into a FTA (Free Trade Agreement, named AAP.CE number 58) in 2006.
6. Paraguay was suspended as a member during 2012.

Every FTA within ALADI are mapped. Table A2 shows the FTAs within the ALADI framework together with the year in which the FTA entered into force.

Table A1. Free Trade Zones across the signatories of original ALADI treaty.

	Argentina	Bolivia	Brazil	Chile	Colombia	Ecuador	Mexico	Paraguay	Peru	Uruguay
Argentina			1995							
Bolivia				2006						
Brazil	1995									
Chile		2006			2007		1999		2007	
Colombia				2007			1995			
Costa Rica										
Ecuador				2010						
Mexico				1999	1995					2004
Paraguay										
Peru				2007						
Uruguay							2004			

Source: own elaboration based on ALADI records. See http://www.aladi.org/nsfaladi/textacdos.nsf/vACEWEB?OpenView&Start=1&Count=800&Expand=7#7.

1. Argentina and Brazil signed the RTA called AAP.CE number 14 in 1999.
2. Bolivia and Chile signed the RTA called AAP.CE number 22 in 2006.
3. Chile and Mexico signed the RTA called AAP.CE number 41 in 1999.
4. Chile and Bolivia signed the RTA called AAP.CE number 22 in 2006.
5. Chile and Colombia signed the RTA called AAP.CE number 22 in 2007 (they had a previous partial agreement since 1995).
6. Chile and Mexico signed the RTA called AAP.CE number 41 in 1999.
7. Chile and Peru signed the RTA called AAP.CE number 38 in 2007 (they had a previous partial scope agreement since 1998).
8. Colombia and Mexico signed the RTA called AAP.CE number 33 in 1995.
9. Ecuador has not engaged in any free trade zone with other Latin American countries. It has a partial scope agreement with Chile since 2010.
10. Mexico and Uruguay signed the RTA called AAP.CE number 60 in 2004.

Table A2. BITs in force across Latin American countries.

	Argentina	Bolivia	Brazil	Chile	Colombia	Costa Rica	Ecuador	Mexico	Paraguay	Peru	Uruguay
Argentina		2005		1995		2001	1996	1998	1995	1996	
Bolivia	2005			1999			1997		2003	1995	
Brazil											
Chile	1995	1999				2000	1996		1998	2001	2010
Colombia										2004	
Costa Rica	2001			2000					2001		
Ecuador	1996	1997		1996					1995–2008	2000	1985–2008
Mexico	1998										2002
Paraguay	1995	2003		1998		2001	1995–2008			1995	1994
Peru	1996	1995		2001	2004		2000		1995		
Uruguay				2010			1985-2008	2002	1994		

Source: own elaboration based on UNCTAD (2017). The years reflects when the agreement entered in force and if terminated before 2018.

Appendix A.3. Gravity Model for FDI

The gravity model for FDI departs from a definition of bilateral FDI:

$$FDI_{ij,t}^{stock} \equiv \omega_{ij,t}^{\varepsilon} M_{i,t} \tag{A1}$$

where $FDI_{ij,t}$ represents FDI stocks between countries i and j at a time t. $M_{i,t}$ is the non-rival aggregate technology capital stock in a particular time. ω_{ijt} represents openness (or barriers to FDI) for foreign technology coming from country i to country j. ε is the elasticity of FDI with respect to openness. To transform FDI stocks into values we multiply Equation (A1) by its marginal product:

$$FDI_{ij,t}^{stock,value} \equiv \omega_{ij,t}^{\varepsilon} M_{i,t} \frac{\partial Y_{j,t}}{\partial M_{i,t}} \tag{A2}$$

in which $M_i = \phi_i \frac{E_i}{P_i}$, where E_i represents total expenditures of country i which equals the country output plus net rents from foreign investment. P_i stands for consumer prices (which can be considered as a multilateral resistance since higher prices for goods and inputs can affect FDI). Y is nominal output. Solving the representative agent's problem delivers a structural system for the steady-state. Then, the gravity structural system is given by:

$$FDI_{ij} = \phi_i \phi_j \omega_{ij}^{\varepsilon} \frac{E_i}{P_i} \frac{Y_j}{M_i} \tag{A3}$$

$$P_i = \left[\sum_{j=1}^{N} \left(\frac{\tau_{ji}}{\Pi_j} \right)^{1-\sigma} \frac{Y_j}{Y} \right]^{\frac{1}{1-\sigma}}, \tag{A4}$$

$$\prod_j = \left[\sum_{i=1}^{N} \left(\frac{\tau_{ji}}{P_i} \right)^{1-\sigma} \frac{E_i}{Y} \right]^{\frac{1}{1-\sigma}}. \tag{A5}$$

where P_i denotes the aggregate price index or the multilateral resistance, as defined by Anderson and van Wincoop [43]. τ_{ji} represents standard iceberg trade costs. Equation (A3) can be transformed to estimate the parameters of interest empirically (see Section 4.2 of the manuscript).

Appendix A.4. Data and Robustness Tests

Table A3. Variables and sources.

Variable	Description	Source
FDI_{ij}	Bilateral Foreign Direct Investment stocks	UNCTAD (proprietary data from 1990–2008 and 2012–2018). Foreign Direct Investment Statistics database.
$Y_{i/j,t}$	GDP home/host country (dollars 2010)	Balance of Payments, IMF
D_{ij}	Bilateral distance between two countries based on distances between their biggest cities	CEPII dataset available http://www.cepii.fr/CEPII/en/bdd_modele/presentation.asp?id=6
$\left(\Delta GDP_{ij,t}^{RIA}\right)$	Sum of GDP to which the host country has tariff-free access.	Own elaboration from IMF http://www.imf.org/external/ns/cs.aspx?id=28
Adjacency	Dummy, takes value 1 when countries share border, 0 otherwise.	Own elaboration
Mercosur	Dummy, 1 when both countries belong to Mercosur, including associated members and deep FTA within Mercosur framework, 0 otherwise.	Own elaboration
ONE Mercosur	Dummy: 1 if recipient country belongs to Mercosur, and sender country does not, 0 otherwise.	Own elaboration
ALADI	Dummy: 1 when both countries belong to a FTA within the *ALADI framework*, 0 otherwise.	Own elaboration
BITs	Dummy: 1 when there is a BIT in force among the two countries, 0 otherwise	World Trade Organization database (WTO)
BIT index	Continuous variable in the interval (0,1). See text for details and Appendix A below	Own elaboration with data from UNCTAD (2017)
FactEndow	Difference between home and host country ratio of gross fixed capital formation over labor force	World Development Indicators; labor force from ILO, UN
LaborCost Dif	Difference in the Relative Cost of Labor among the home and host country	International Labor Organization http://www.ilo.org/global/statistics-and-databases/lang--en/index.htm
PolitRisk	Role of Institutions, law enforcement and government stability. Complemented with International Country Risk	World Bank and Princeton University

Source: own elaboration.

Estimations with the Inverse Hyperbolic Sine (IHS) Transformation Method

This method is also adequate for zeros and negative values. The inverse hyperbolic sine transformation for FDI is defined as: $log\left(FDI_{ij,t} + \left(FDI_{ij,t}^2 + 1\right)^{\frac{1}{2}}\right)$. Except for very small values of FDI, the IHS is approximately equal to $log\left(2FDI_{ij,t}\right)$ or $log(2) + log\left(FDI_{ij,t}\right)$ and it can be interpreted in exactly the same way as a standard logarithmic dependent variable (see Aisbett et al., 2018 for discussion).

Table A4. Impact of RTAs and BITs investor protection index on FDI. Inverse hyperbolic sine transformation, two-way fixed effects.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Mercosur	0.132	0.127	0.122	0.112	0.117	0.119	0.121
	(0.028) ***	(0.031) ***	(0.028) ***	(0.036) ***	(0.029) ***	(0.027) ***	(0.031) ***
ALADI	0.025	0.021	0.017	0.023	0.018	0.015	0.019
	(0.011) *	(0.009) *	(0.008) *	(0.012) *	(0.008) *	(0.007) *	(0.012)
BITs	0.047	0.045	0.038	0.032	0.036	0.035	0.039
	(0.015) **	(0.022) **	(0.019) **	(0.014) **	(0.018) **	(0.018) *	(0.020) *
ONE_Mercosur	−0.064	−0.061	−0.066	−0.058	−0.052	−0.054	−0.056
	(0.020) ***	(0.015) ***	(0.018) ***	(0.013) ***	(0.023) ***	(0.020) ***	(0.018) ***
ONE_ALADI	−0.012	−0.014	−0.016	−0.015	−0.011	−0.010	−0.009
	(0.005) *	(0.007) *	(0.008) *	(0.006) *	(0.015)	(0.015)	(0.019)
BIT index		0.044		0.039			0.043
		(0.020) *		(0.019) *			(0.021) *
BIT Index < 0.68 (less pro-investor)			0.016		0.018	0.014	
			(0.022)		(0.026)	(0.029)	
BIT Index ≥ 0.68 (pro-investor)			0.032		0.030	0.028	
			(0.017) *		(0.015) *	(0.013) *	
Mercosur * BIT Index				0.042			0.036
				(0.012) ***			(0.009) ***
Mercosur * BIT Index ≥ 0.68					0.048	0.052	
					(0.010) ***	(0.006) ***	
Aladi*BIT Index				0.016		0.014	0.011
				(0.023)		(0.028)	(0.031)
ISDS concluded				−0.015	−0.012	−0.006	−0.009
				(0.048)	(0.052)	(0.043)	(0.039)
Mercosur_LAG5						0.096	0.102
						(0.033) ***	(0.038) ***
ALADI_LAG5						0.008	0.010
						(0.003) *	(0.005) *
BITs_LAG5							0.036
							(0.018) **
Controls		Yes	Yes	Yes	Yes	Yes	Yes
No. Observations		1962	1962	1962	1962	1632	1632
Individual country−year fixed effect		Yes	Yes	Yes	Yes	Yes	Yes
Country−pair fixed effect		Yes	Yes	Yes	Yes	Yes	Yes

Note: This table reports estimates for FDI flows from country i to country j at the aggregate level for the period 1995–2018. Estimation with host year FE and pair-country FE (η_{ij}). We control for labor cost differentials, endowment difference, political risk, and GDP in the home economy. Robust standard errors in parenthesis, clustered by country pair. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A5. Correlation Matrix.

	fdi stock	**gdp_d**	**gdp_o**	**Distance**	**Adjacency**	**Mercosur**	**ONE_Mercosur**
fdistock	1						
gdp_d	0.1534 *	1					
gdp_o	0.1921 *	0.0109	1				
distance	−0.0300	0.1615 *	0.1615 *	1			
adjacency	0.0687 *	0.1115 *	0.0541 *	−0.5483 *	1		
Mercosur	0.2109 *	0.1511 *	0.1511 *	−0.0853 *	0.1681 *	1	
ONE_Mercosur	−0.1201 *	0.2538 *	0.0334	−0.0191	0.1101 *	0.4601 *	1
Aladi	0.1389 *	0.1356 *	0.1356 *	0.1188 *	0.0679 *	0.1637	0.0902
ONE_Aladi	−0.1370 *	0.5079 *	0.0292	0.1104 *	0.0775 *	0.2341	0.357
l5_Mercosur	0.2194 *	0.1190 *	0.1190 *	−0.2246 *	0.2270 *	0.3857 *	0.2604
l5_Aladi	0.1133 *	0.1598 *	0.1598 *	0.0942 *	0.0441	0.1682	0.0843
Mercosur_LEAD5	0.1743	0.1386 *	0.1386 *	−0.0069	0.1322 *	0.4438	0.3889
Aladi_LEAD5	0.1006	0.1261	0.1261	0.1569	0.069	0.0838	0.0496
bit_enforced	0.0251 *	0.2224 *	0.2138 *	0.0235	0.0481 *	0.1113	0.0974
bit_index	0.0151 *	0.2066 *	0.2061 *	0.0527 *	0.1042 *	0.0244	−0.0164
BIT_LAG5	0.0307 *	0.2202 *	0.2113 *	0.0164	0.0165	0.1657	0.1119
BIT_LEAD5	0.0119	0.2506	0.2422	0.0558	0.0674	0.0211	0.0677
sum_gdp_partner	0.2215 *	0.0832 *	0.0927 *	0.1982	0.0408	0.1879 *	0.0264
factendow	0.0828 *	0.2192 *	0.1093 *	0.0000	−0.0203	0.0000	0.2208
laborcostdif	0.0495 *	0.1286 *	0.1014 *	0.0000	−0.0122	0.0000	−0.2745 *
politrisk	−0.1233 *	−0.0046 *	−0.0273	0.0890 *	−0.1127	0.0451	−0.0245
	Aladi	ONE_Aladi	Mercosur_LAG5	Aladi_LAG5	Mercosur_LEAD5	Aladi_LEAD5	bit_enforced

Table A5. *Cont.*

	fdi stock	gdp_d	gdp_o	Distance	Adjacency	Mercosur	ONE_Mercosur
Aladi	1						
ONE_Aladi	0.2386 *	1					
Mercosur_LAG5	0.0838 *	0.1541 *	1				
Aladi_LAG5	0.4325 *	0.1784 *	0.1251 *	1			
Mercosur_LEAD5	0.1706 *	0.2052 *	0.3531 *	0.1647 *	1		
Aladi_LEAD5	0.4713 *	0.2274 *	−0.0018	0.6186 *	0.1637 *	1	
bit_enforced	0.004	−0.0921	0.0211	−0.0816 *	0.1657 *	0.0185	1
bit_index	0.0265	−0.0457	−0.1609 *	−0.0879 *	0.0393	0.0561 *	0.4251 *
BIT_LAG5	0.0185	−0.0935	0.1145 *	−0.0530 *	0.1620 *	−0.0581	0.4461 *
BIT_LEAD5	−0.0816 *	−0.1800	0.0242	−0.1479 *	0.0663 *	0.0067	0.3105 *
sum_gdp_partner	0.1561 *	−0.0068	0.1460 *	0.1762 *	0.1507 *	0.1419 *	−0.1323 *
factendow	0.000	−0.0562	0.0000	0.0000	0.0000	0.000	0.0210
laborcostdif	0.000	−0.0095	0.0000	0.0000	0.0000	0.000	0.0311
politrisk	0.1008 *	−0.0283	−0.0202 *	0.0478	−0.0204	0.1297	−0.0032
	bit_index	BIT_LAG5	BIT_LEAD5	sum_gdp_partner	factendow	laborcostdif	politrisk
bit_index	1						
BIT_LAG5	0.3493 *	1					
BIT_LEAD5	0.3778 *	0.3466 *	1				
sum_gdp_partner	0.2232	0.2205	0.2735	1			
factendow	0.0121	0.0193	0.0166	0.1268 *	1		
laborcostdif	−0.0116	0.0332	0.0274	0.1217 *	0.2015 *	1	
politrisk	−0.0612 *	−0.0302	0.0380	0.0270	0.2530 *	0.1381 *	1

Source: own elaboration. Variables gdp_o and gdp_d stand for origin and destination. The rest of variables are self-explanatory. * $p < 0.05$.

References

1. Borensztein, E.; De Gregorio, J.; Lee, J. How does FDI affect economic growth? *J. Int. Econ.* **1998**, *45*, 115–135. [CrossRef]
2. Alvarado, R.; Iñiguez, M.; Ponce, P. FDI and economic growth in Latin America. *Econ. Anal. Policy* **2017**, *56*, 176–187. [CrossRef]
3. Yotov, Y.; Piermartini, R.; Monteiro, J.A.; Larch, M. *An Advanced Guide to Trade Policy Analysis: The Structural Gravity Model*; World Trade Organization: Geneva, Switzerland, 2016.
4. Santos-Silva, J.; Tenreyro, S. The Log of gravity. *Rev. Econ. Stat.* **2006**, *88*, 641–658. [CrossRef]
5. Jang, Y. The Impact of bilateral FTA on bilateral FDI among developed countries. *World Econ.* **2011**, *34*, 1628–1651. [CrossRef]
6. Egger, P.; Merlo, V. BITs bite: An anatomy of the impact of BITs on multinational firms. *Scand. J. Econ.* **2012**, *114*, 1240–1266. [CrossRef]
7. Aisbett, E.; Busse, M.; Nunnenkamp, P. BITs as deterrents of host-country discretion: The impact of investor-state disputes on FDI in developing countries. *Rev. World Econ.* **2018**, *154*, 119–155. [CrossRef]
8. Markusen, J.R.; Venables, A. The theory of endowment, intra-industry and multi-national trade. *J. Int. Econ.* **2000**, *52*, 209–234. [CrossRef]
9. Markusen, J.R.; Maskus, K. Discriminating among alternative theories of the multinational enterprise. *Rev. Int. Econ.* **2002**, *104*, 694–707. [CrossRef]
10. Levy-Yeyati, E.; Stein, E.; Daude, C. *Regional Integration and the Location of FDI. Working Paper 492*; Inter-American Development Bank, Research Department: Washington, DC, USA, 2003.
11. Ramondo, N.; Rappoport, V.; Ruhl, K. Intra-firm trade and vertical fragmentation in U.S. multinational corporations. *J. Int. Econ.* **2016**, *98*, 51–59. [CrossRef]
12. Egger, P.; Merlo, V. The impact of BITs on FDI dynamics. *World Econ.* **2007**, *30*, 1536–1549. [CrossRef]
13. Baltagi, B.H.; Egger, P.; Pfaffermayr, M. Estimating Regional Trade Agreement effects on FDI in an interdependent world. *J. Econom.* **2008**, *145*, 194–208. [CrossRef]
14. Chenaf-Nicet, D.; Rougier, E. The effect of macroeconomic instability on FDI flows: A gravity estimation of the impact of regional integration in the case of Euro-Mediterranean agreements. *Int. Econ.* **2016**, *145*, 66–91. [CrossRef]
15. Felbermayr, G.; Toubal, F. Cultural Proximity and Trade. *Eur. Econ. Rev.* **2010**, *54*, 279–293. [CrossRef]
16. Büthe, T.; Milner, H. FDI and institutional diversity in trade agreements: Credibility, commitment and economic flows in the developing world, 1971–2007. *World Politics* **2014**, *66*, 88–122. [CrossRef]
17. Dixon, J.; Haslam, P.A. Does the quality of investment protection affect FDI flows to Developing Countries? Evidence from Latin America. *World Econ.* **2016**, *39*, 1080–1108. [CrossRef]
18. Frenkel, M.; Walter, B. Do bilateral investment treaties attract FDI? The role of international dispute settlement provisions. *World Econ.* **2019**, *42*, 1316–1342. [CrossRef]
19. Markusen, J.R. Multinationals, multiplant economies and the gains from trade. *J. Int. Econ.* **1984**, *16*, 205–226. [CrossRef]
20. Helpman, E. A simple theory of international trade with multinational corporations. *J. Political Econ.* **1984**, *92*, 451–471. [CrossRef]
21. Neary, J.P. FDI and the single market. *Manch. Sch.* **2002**, *70*, 291–314. [CrossRef]
22. Castilho, M.; Zignago, S. Trade, FDI and regional integration in Mercosur. *Revue Econ.* **2000**, *51*, 761–774.
23. Anderson, J.E.; Larch, M.; Yotov, Y. Trade liberalization, growth and FDI: A structural estimation framework. *ETSG Work. Pap.* 2017. Available online: http://www.etsg.org/ETSG2016/Papers/052.pdf (accessed on 1 November 2018).
24. Blomström, M.; Kokko, A. *Regional Integration and FDI*; NBER Working Paper 1997; NBER: Cambridge, MA, USA, 1997.
25. Egger, P.; Pfaffermayr, M. FDI and European integration in the 1990s. *World Econ.* **2004**, *27*, 99–110. [CrossRef]
26. Büge, M. *Do PTA Increase Their Members' FDI?* German Development Institute Discussion Paper; DIE-GDI: Bonn, Germany, 2014.
27. Carr, D.L.; Markusen, J.R.; Maskus, K.E. Estimating the knowledge-capital model of the multinational enterprise. *Am. Econ. Rev.* **2001**, *91*, 693–708. [CrossRef]

28. Cherif, M.; Dreger, C. Institutional determinants of financial development in MENA countries. *Rev. Dev. Econ.* **2016**, *20*, 670–680. [CrossRef]
29. Kerner, A. Why should I believe you? Costs and consequences of BITs. *Int. Stud. Q.* **2009**, *53*, 73–102. [CrossRef]
30. Busse, M.; Königer, J.; Nunnenkamp, P. FDI promotion through bilateral investment treaties: More than a bit? *Rev. World Econ.* **2010**, *146*, 147–177. [CrossRef]
31. Neumayer, E.; Spess, L. Do BITs increase FDI to developing countries? *World Dev.* **2005**, *33*, 1567–1585. [CrossRef]
32. Salacuse, J.W.; Sullivan, N. Do BITs really work? An evaluation of BITs and their grand bargain. *Harv. Int. Law J.* **2005**, *46*, 67–130.
33. Sirr, G.; Garvey, J.; Gallagher, L. BITs and FDI: Evidence of asymmetric effects on vertical and horizontal investments. *Dev. Policy Rev.* **2017**, *35*, 93–113. [CrossRef]
34. Hallward-Driemeier, M. *Do Bilateral Investment Treaties Attract FDI? Only a Bit ... and They Could Bite*; Policy Research Working Paper WPS3121 2003; World Bank: Washington, DC, USA, 2003.
35. Rose-Ackerman, S.; Tobin, J. FDI and the business environment in developing countries: The impact of BITs. *Yale Law Econ. Res. Paper* **2005**, *293*. Available online: http://ssrn.com/abstract=557121 (accessed on 1 November 2018).
36. Gallagher, K.P.; Birch, M.B. Do investment agreements attract investment? Evidence from Latin America. *J. World Investig. Trade* **2006**, *7*, 961–974. [CrossRef]
37. Yackee, J.W. BITs, credible commitment, and the rule of (International) Law: Do BITs promote foreign direct investment? *Law Soc. Rev.* **2008**, *42*, 805–832. [CrossRef]
38. Tobin, J.L.; Rose-Ackerman, S. When BITs have some bite: The political-economic environment for bilateral investment treaties. *Rev. Int. Organ.* **2011**, *6*, 1–32. [CrossRef]
39. Blonigen, B.A.; Piger, J. Determinants of FDI. *Can. J. Econ.* **2014**, *47*, 775–812. [CrossRef]
40. Egger, P.; Pfaffermayr, M. The proper panel econometric specification of the gravity equation: A three-way model with bilateral interaction effects. *Empir. Econ.* **2003**, *28*, 571–580. [CrossRef]
41. Head, K.; Ries, J. FDI as an outcome of the market for corporate control: Theory and evidence. *J. Int. Econ.* **2008**, *74*, 2–20. [CrossRef]
42. Anderson, J.E.; Yotov, Y. Terms of trade and global efficiency effects of free trade agreements, 1990–2002. *J. Int. Econ.* **2016**, *99*, 279–298. [CrossRef]
43. Anderson, J.E.; van Wincoop, E. Gravity with gravitas: A solution to the border puzzle. *Am. Econ. Review* **2003**, *93*, 170–192. [CrossRef]
44. Head, K.; Mayer, T. Gravity equations: Workhorse, toolkit, and cookbook. In *Handbook of International Economics*; Gopinath, G., Helpman, E., Rogoff, K., Eds.; Elsevier: Amsterdam, The Netherlands, 2014; Volume 4, pp. 131–195.
45. Mátyás, L. Proper econometric specification of the gravity model. *World Econ.* **1997**, *20*, 363–368. [CrossRef]
46. Egger, P. A note on the proper econometric specification of the gravity equation. *Econ. Lett.* **2000**, *66*, 25–31. [CrossRef]
47. Baier, S.L.; Bergstrand, J.H. Do free trade agreements actually increase members' international trade? *J. Int. Econ.* **2007**, *71*, 72–95. [CrossRef]
48. Olivero, M.P.; Yotov, Y. Dynamic gravity: Endogenous country size and asset accumulation. *Can. J. Econ.* **2012**, *45*, 64–92. [CrossRef]
49. Feenstra, R. *Advanced International Trade: Theory and Evidence*; Princeton University Press: Princeton, NJ, USA, 2016.
50. Egger, P.; Pfaffermayr, M. Distance, trade and FDI: A Hausman–Taylor SUR approach. *J. Appl. Econom.* **2004**, *19*, 227–246. [CrossRef]
51. Babetskaia-Kukharchuk, O.; Maurel, M. Russia's accession on to the WTO: The potential for trade increase. *J. Comp. Econ.* **2004**, *32*, 680–699. [CrossRef]
52. McPherson, M.; Trumbull, W. Rescuing observed fixed effects: Using the Hausman-Taylor method for out-of-sample trade projections. *Int. Trade J.* **2008**, *22*, 315–340. [CrossRef]
53. Frankel, M.; Funke, K.; Stadtmann, G. A panel Analysis of bilateral FDI flows to emerging economies. *Econ. Syst.* **2004**, *28*, 281–300. [CrossRef]
54. Yang, S.; Martinez-Zarzoso, I. A panel data analysis of trade creation and trade diversion effects: The case of ASEAN-China Free Trade Area. *China Econ. Rev.* **2014**, *29*, 138–151. [CrossRef]

55. Berger, A.; Busse, M.; Nunnenkamp, P.; Roy, M. Do trade and investment agreements lead to more FDI? Accounting for key provisions inside the black box. *Int. Econ. Econ. Policy* **2013**, *10*, 247–275. [CrossRef]
56. Kohl, T.; Brakman, S.; Garretsen, H. Do trade agreements stimulate international trade differently? Evidence from 296 trade agreements. *World Econ.* **2016**, *39*, 97–131. [CrossRef]
57. Janicki, H.P.; Warin, T.; Wunnava, P. Endogenous OCA theory: Using the gravity model to test Mundell's intuition. *CES Work. Pap.* **2005**, 125.
58. Eicher, T.S.; Papageorgiou, C.; Raftery, A.E. Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *J. Appl. Econom.* **2011**, *26*, 30–55. [CrossRef]
59. Bergstrand, J.; Egger, P. A knowledge-and-physical-capital model of international trade flows, FDI and multinational enterprises. *J. Int. Econ.* **2007**, *73*, 278–308. [CrossRef]
60. Borchert, I.; Larch, M.; Shikher, S.; Yotov, Y. *Disaggregated Gravity: Benchmark Estimates and Stylized Facts from a New Database*; School of Economics Working Paper Series 2020-8; LeBow College of Business, Drexel University: Philadelphia, PA, USA, 2020.
61. Egger, P.; Nigai, S. Structural gravity with dummies only: Constrained ANOVA-type estimation of gravity models. *J. Int. Econ.* **2015**, *97*, 86–99. [CrossRef]
62. Hausman, J.A.; Taylor, W.E. Panel data and unobservable individual effects. *Econometrica* **1981**, *49*, 1377–1398. [CrossRef]

mathematics

MDPI

Article

An Application of the SRA Copulas Approach to Price-Volume Research

Pedro Antonio Martín Cervantes †, Salvador Cruz Rambaud *,† and María del Carmen Valls Martínez †

Department of Economics and Business, University of Almería, La Cañada de San Urbano, 04120 Almería, Spain; pmc552@ual.es (P.A.M.C.); mcvalls@ual.es (M.d.C.V.M.)

* Correspondence: scruz@ual.es; Tel.: +34-950-01-51-84

† These authors contributed equally to this work.

Received: 18 September 2020; Accepted: 19 October 2020; Published: 26 October 2020

Abstract: The objective of this study was to apply the Sadegh, Ragno, and AghaKouchak (SRA) approach to the field of quantitative finance by analyzing, for the first time, the relationship between price and trading volume of the securities using four stock market indices: DJIA, FOOTSIE100, NIKKEI225, and IBEX35. This procedure is a completely new methodology in finance that consists of the application of a Bayesian framework and the development of a hybrid evolution algorithm of the Markov Chain Monte Carlo (MCMC) method to analyze a large number (26) of parametric copulas. With respect to the DJIA, the Joe's copula is the one that most efficiently models its succinct dependence structures. One of the copulas included in the SRA approach, the Tawn's copula, is jointly adjusted to the FOOTSIE100, NIKKEI225, and IBEX 35 indices to analyze the asymmetric relationship between price and trading volume. This adjustment can be considered almost perfect for the NIKKEI225, and a relatively different characterization for the IBEX35 seems to indicate the existence of endogenous patterns in the price and volume.

Keywords: copulas; Markov Chain Monte Carlo simulation; local optima vs. local minima; financial markets; SRA approach

MSC: 62H05; 62F15; 60J22; 62P05

1. Introduction

Current trends in quantitative finance reveal that econophysics has become an economic analysis discipline characterized not by its multidisciplinary but by its transdisciplinary nature [1], contributing to the formation of a common framework in the research of financial phenomena [2]. Traditionally, the link has been strong between the stochastic analysis of hydrological phenomena and the study of time series, especially in the field of quantitative finance. The best-known example is likely represented by the Hurst exponent, a procedure inspired by the floods of the Nile River [3], which is of unquestionable efficiency when estimating the long-term memory of time series. Hydrological phenomena are completely different to financial ones but, in general, they present certain common patterns of analysis. Thus, several works have transferred the applicability of the theory of copulas from the field of hydrology to finance [4–7]. Recently, Sadegh et al. [8] developed a specific methodology based on the joint use of 26 multivariate copulas applied in hydrology (hereinafter, SRA), which, in our opinion, offers huge potential for the analysis of the price–volume relationship. Therefore, our aim was to introduce this methodological approach within quantitative finance, summarizing its fundamental aspects as a step prior to its practical implementation.

The analysis of the joint dependence between economic and financial variables has found important support in the Sklar's theorem, through which it has been possible to specify, define, and contrast the latent or redundant dependence structures present in the bivariate and multivariate time series. Notably, Sklar's theorem, the starting point from which this theory departs, has been subject to continuous extensions that have improved the analysis of the structures of dependence between random variables or, in other words, of their succinct relationships when these are schematized in their minimal mathematical expression.

The emergent interest in copulas, detailed by [9], which increased in the field of finance after the paper by [10], does not correspond in reality to the use of several of the numerous types of pre-existing copulas, but to the systematic implementation of certain copulas types, either in economics and quantitative finance or in any other field. According to the compendium of copulas by [11], nonparametric and semiparametric models represent a minority that is largely surpassed by parametric models, amongst which almost 100 different types could be distinguished. Some of them have not been yet fully spread by the literature or, at least, they are not sufficiently well known, since most empirical studies opt for the application of a narrow number of copulas that could be classified as classic copulas.

Conversely, the analysis of the price–volume relationship (hereinafter, PVR) continues being a specific area of the financial literature that has not yet received a conclusive solution. In our opinion, the relationship between prices and trading volume can be derived by dissecting the dependence structure of both variables through the Sklar's theorem, that is, through the implementation of copulas. To accomplish this task, we followed the suggestion of [11] when implementing as many parametric copulas as possible to jointly analyze the same relationship, prices vs. trading or transaction volume, from different points of view (or dependence structures). Therefore, through this empirical work, we aimed to provide a new approach to the application of copulas in the context of PVR, implementing a large number of copulas that, to the best of our knowledge, have not been previously applied in the area of quantitative finance with the aim that these types of transdisciplinary approaches will transcend from the study of PVR to other areas of financial research in the future. This study was mainly based on [8], whose 26 parametric copulas, estimated according to a Bayesian uncertainty framework, were replicated in the price–volume variables of the DJIA, FTSE100, NIKKEI225, and IBEX35 indices.

The SRA was implemented in accordance with two different guidelines focused on two respective scenarios: first, this procedure was applied per se to price–volume data of the DJIA index over the period 1928–2009. Second, one of the 26 copulas included in this methodology, the Tawn's copula [12], was used to jointly compare the dependence structures derived from the PVR in the FTSE100, NIKKEI225, and IBEX35 indices using the period 2000–2018 as the time horizon (also in per se values). This copula was expressly used as it can be considered one of the new-generation copulas whose knowledge is not yet broadly applied in the literature and whose contribution to the analysis of the PVR may be crucial given its exhaustiveness in the estimation of parameters.

The rest of this article is organized as follows: first, Section 2 describes the current state of this research by outlining a literature review concerning the theory of copulas and the analysis of the PVR, detailing the works that expressly employed copulas in the determination of the relationship between prices and trading volume. In our opinion, with few exceptions such as [13,14], most of the works usually offer an excessively summarized and, in some cases, incomplete literature review of the PVR. For this reason, an extensive review of the literature was conducted by listing the four explanatory hypotheses that were mostly addressed in its study. Similarly, this section summarizes the plausible shortcomings derived from the utilization of copulas, pointing out a series of sociological weaknesses. In Section 3, the different databases used as well as a brief review of the theoretical bases presented in the SRA are described: its Bayesian perspective, later developed in Appendix A, and the Markov Chain Monte Carlo simulation used by this methodology. In Section 4, the results obtained are contextualized, finishing this investigation with Section 5, which is dedicated to the discussion of the results. The paper finishes with Section 6, which reflect our conclusions, supplemented with

a proposal for future lines of investigation, congruent with the methodological scheme implemented in this manuscript, emphasizing the practical usefulness of the PVR analysis, both for investors and practitioners, from the perspective of the scheme proposed by Karpoff [13]. To ensure the maximum possible exhaustiveness, Appendix B provides an introductory summary of the main basis of the theory of copulas.

2. State of the Art

2.1. Related to the Theory of Copulas

From Sklar [15] until now, the theory of copulas has not stopped being an area under continuous development, to the point that copulas, as a concept, as well as their proven ability to determine parametric and nonparametric dependence measures, have been discovered and rediscovered during the last 50 years [16]. In this sense, Genest et al. [9] applied bibliometric methods to fix the end of the 1990s as the starting point of a growing interest, practically exponential, which, according to [17–19], was due to the seminal repercussion of several works of singular importance for its popularization. This would mean the rediscovery of Sklar's works. In the opinion of [20], this would include its involvement in quantitative finance areas and the opening of new lines of research in this field, which would serve as a trigger for its gradual generalization toward numerous multidisciplinary areas such as the insurance sector, actuarial science, meteorology, hydrology, and many other disciplines [5].

Daníelsson [21] highlighted three stylized findings commonly detected when implementing copulas: the volatility clustering, the phenomenon of fat tails [22], and the analysis of a nonlinear dependence between a given dataset of variables [23–26]. More generically, the application of copulas in economic-financial fields can be structured around a series of predominant research lines such as the valuation of collateralized debt obligations (CDOs) [10], the analysis of financial time series [27–29] (reinforced by the time-varying copulas approach [30,31]), the interpretation of the implicit asymmetries in the exchange rates [32], the successive contributions to the context of the portfolio management either from the construction of a simplified portfolio based on the theory of copulas [33] or from the application of the value-at-risk (VaR) methodology [30,34], or to the study of contingent claims, especially the valuation of financial options in turbulent environments, characterized by risk [35–37]. In addition to these research lines, the theory of copulas has been employed to address all kinds of specific aspects like the methodology proposed by [38] to obtain new copulas based on a given one or the creation of a new class of semiparametric copula-based multivariate dynamic models (SCOMDY), introduced by [39]. Analogously, García et al. [40] focused on building copulas in the contexts of marked uncertainty; the elaborated goodness-of-fit testing procedure for copulas suggested by [41] are also remarkable, as well as the development promoted by the vine-copulas to model dependence structures [42–44] in which the copulas are directly linked with the decision processes.

Limitations of the Copulæ Approach

Strictly, a complete literature review of the theory of copulas would not be objective enough if some of its perceptible limitations are not highlighted, often given by an erroneous conception and misuse of its theoretical basis and, to a lesser extent, by sociological factors. Embrechts et al. [45] listed three conceptual fallacies linked to the relative understanding and abuse when implementing copulas. However, although these have gradually been solved, the main limitation of copulas is the breach of the continuity condition [46], which a priori establishes a univocal relationship between any continuous multivariate distribution and a single resulting copula C [45]. So, in any case, Equation (A7) (see Appendix A) must be satisfied if all distribution functions $F_1(x_1), F_2(x_2), \ldots, F_n(x_n)$ are continuous. Schweizer and Sklar [47] showed that if there is at least one discrete F_i, the joint distribution function can continue being expressed as a function, as shown in Equation (A7); however, this would not be defining a copula per se, but a possible (or feasible) copula C. Several works have furthered the mitigation of this inconvenience; for example, Genest and Nešlehová [48] related copulas with discrete

distribution functions, demonstrating how such links can invalidate some basic precepts of the theory of copulas (evidently, in the continuous case) or Mayor et al. [49], who performed a discrete extension of the Sklar's theorem in function of some operators similar to copulas, defined as a finite chain that they denominates "discrete copulas".

Similarly, others [50,51] emphasized that the justification of modeling the relationship of dependence between variables via copulas does not always have to be obvious or completely necessary as, in many cases, it may be more convenient to directly adjust the variables to a given multivariate distribution function (i.e., Gaussian or lognormal) to delimit the predictable stylized findings relative to their dependence structures. Another impediment, according to [52], is that copulas do not entirely correspond with the pre-existing stochastic framework because they are static models and, therefore, they are not completely adequate for modeling dependence structures over time.

The misuse of the Gaussian copula as a general indicator of credit risk should also be considered during the most recent period of economic boom, called "irrational exuberance" by Shiller [53], in whose case the procedure introduced by [10] practically became a standardized measure of the risk level of certain assets with high levels of volatility, being one of the indirect triggers in the expansion of the subprime mortgage crisis. Donnelly and Embrechts [54] metaphorically stated that "the devil is in the tails" when describing the main limitation of the models based on Gaussian copulas to fit extreme data values or outliers if compared with others like the Gumbel copula [55]. According to [45,56,57], there were many voices that, long enough in advance, warned about these models' inconsistencies that ignored the fact that the application of Gaussian copulas could be more or less viable in relatively stable financial environments but would be completely inefficient in detecting joint extreme events. This conclusion was personally confirmed by P. Embrechts to one of the coauthors of this work (November 2017):

> "[. . .] I insisted from the beginning, back in 1998, that credit risk models based on Gaussian copula are not capable of capturing joint credit defaults in a sufficiently realistic way. The mathematical result underlying this statement of mine dates back to the late fifties [. . .]"

Mikosch [52], Daníelsson [58], and Zimmer [59] also criticized the widespread application of this procedure and even Salmon [60] deduced that the interests, aims, and objectives of the banking industry overlapped with those of mathematics, pointing out a sociological limitation born from considering the mathematical methodology implicit in the theory of copulas as a *factotum* in the determination of the risk of financial assets. In this sense, Rogers [61] stated:

> "The problem is not that mathematics was used by the banking industry, the problem was that it was abused by the banking industry. Quants were instructed to build models which fitted the market prices. Now if the market prices were way out of line, the calibrated models would just faithfully reproduce those wacky values, and the bad prices get reinforced by an overlay of scientific respectability".

Daníelsson [21] considered that the a priori use of copulas can arbitrarily determine any structure of dependence so that an "optimal" adjustment of a copula does not mean an obligatory a sine qua non condition that leads to an optimal fit from the original distribution of the data. As no economic theory is explicitly linked to copulas, it is difficult to specify in advance what type of copulas are the most appropriate for each specific analysis given the total freedom in the choice of the underlying structures of dependence, which, in no case, are subrogated in a preliminary way to any economic theory.

2.2. Related to the Analysis of the Price-Volume Relationship

Osborne [62] was the first to address the concurrent relationship between prices and trading volume from a strictly quantitative perspective, estimating that the logarithm of the price of financial assets follows a diffusion process with a trend whose variance depends on the trading volume. Samuelson [63] was inspired by this research to infer that the prices of financial assets describe a

specific random trajectory based on the Geometric Brownian motion. Thus, the primary roots of modern quantitative finance are based in the preliminary studies of the analysis of the PVR. Others [62,64,65] applied spectral analysis to determine that, in principle, there is no a significant relationship between prices and volumes (or it is too meager to take it into consideration).

These initial works provided the background to justify and empirically test the reconsidered theory of demand [66], a new conceptualization of the theory of supply and demand, openly contrary to classical postulates, which would anticipate the empirical basis of the Granger causality test [67]. Based on Godfrey et al. [65], Ying [68] presented a complete disagreement with the theory of conventional demand, performing a series of statistical tests whose results defined five empirical patterns that characterize the joint evolution of the price and volume variables. Clark [69] used a mixture of probabilistic distributions to describe what would be considered the first explanatory hypothesis of the PVR, the MDH (Mixture of Distribution Hypothesis), proposing that the number of operations that occur per unit of time is a random variable and the variation in prices per unit of time is the sum of the increments of the intraday price equilibrium. Thus, the mixed variable is hypothesized according to the information rate periodically reached by the markets, inferring that, in principle, price and volume must be positively correlated, varying in a contemporary basis, just before the arrival of new information. Others [70–74] used the basis of this approach, which were further expanded [75,76] by inputting the information rate into the GARCH (Generalized Autoregressive Conditional Heteroskedasticity models) primary specification of Bollerslev [77], hypothesizing that the daily trading volume behaves like a representative proxy variable when explaining the evolution of prices growth depending on the GARCH effects, or on the persistence of transitory volatility shocks. Practically as a counterpart to the MDH, the SAIH (Sequential Information Arrival Hypothesis) [78,79] arose as a probabilistic model based on a binomial distribution, according to which the information arrives the markets generating a noncontinuous or fragmented flow. Per Darrat et al. [80], this hypothesis should be only contrastable in those periods in which the information is public and whose empirical evidence is ascertained by all market participants. Copeland [78] argued that, as more than an effective explanatory hypothesis of the PVR, it should be reconsidered as "a new technique for the analysis of demand".

The DBH (Dispersion of Beliefs Hypothesis) and the NTH (Noise Trader Hypothesis) would complete, together with the MDH and the SIAH, the four major explanatory hypotheses of the PVR, being the common denominator of all information that reaches the markets, although analyzed from opposite points of view and finally convergent [81]. The NTH [82] states that prices and volumes are the result of positive and negative feedback strategies that degenerate into noise in the sense stated by Black [83], on which passive, rational, and speculative investors react positively to a feedback strategy. In other words, according to this hypothesis, all information of interest that arrives to the markets, or relevant in any investment process, would be equivalent to the paradigmatic [83] noise. In contrast, the DBH [84,85] defines an antagonistic theoretical scenario in which investors who interact exclusively for speculative reasons and their degree of risk aversion is neutral, collectively receive public information, which, in principle, is common and perceived in the form of market signals. Consequently, the consecutive changes in prices exhibit a negative serial correlation and trading volume is positively correlated [84].

Use of Copulas in the Price-Volume Research

Amongst the works that explicitly opted for the implementation of copulas in the study of the PVR are those by Gurgul, who focused on the Polish and central European stock markets (Austria and Germany). Gurgul and Syrek [86] implemented the family of Archimedean copulas to demonstrate that the volatility of (daily) returns of the companies listed on the DAX was positively related to the trading volume. Gurgul et al. [87] introduced a measure of dependences based on copulas to quantify the relationship between performance and volume, volatility and volume, and yield and performance of the benchmark Polish stock market (WIG) compared to three indices corresponding

to other international financial markets (ATX, DAX, and DJIA). They concluded that each one of the proposed relationships is significant except for the volume traded in the Polish market vs. the volatility of DJIA returns. Gurgul et al. [88] used a Granger's nonlinear causality model based on the Bernstein's copula by applying the nonparametric test of conditional independence between two vector processes [89] in five selected ordinary shares of the ATX index, confirming the existence of several well-defined causal guidelines between the performance of shares, the volatility, and the trading volume (both expected and unexpected). This same copula, in conjunction with Hellinger's distance, was implemented [90] to study the high-frequency data of 10 central European companies (Austria and Poland), detecting a high degree of unidirectional causality, both linear and nonlinear, of the returns to the expected volume, which was not appreciable in the opposite direction. They also observed the existence of a linear causality from the volatility realized to the expected trading volume that, once again, was negligible in the opposite direction.

Gurgul and Syrek [91] studied the dependence structures of ordinary stock returns, volatility, and transaction volumes of several companies listed in the CAC40 and FTSE100 indices to verify the long-term memory of the MDH through the fractional cointegration of these series according to the procedure previously described [92]. In most cases, there is no structure of common dependence whereby the analyzed series would not be caused by a process of reaching a common information with long-term memory. Gurgul et al. [93] investigated the high-frequency data of 13 German companies included in the DAX index for a period of 33 days by selecting the copulas t and Gumbel to analyze their different underlying dependence structures according to the inference function for margins (IFM) method [94]. These scholars inferred that the contemporary relationship between the price duration and its associated trading volume depends on the distribution tails as unusual high volume accumulations tend to coincide with long durations and, conversely, dependence is minimal when any of the variables are delayed.

The Asian Financial Crisis of 1997 provided a empirical scenario from which Ning and Wirjanto [95] analyzed the structure of dependence between prices and volumes in a context of extreme volatility by examining the evolution of the most representative stock indices of the six countries in southern Asia, which were more seriously affected by the crisis. Gallant et al. [96] implemented several mixtures of copulas (Clayton, survival Clayton, and Frank) expressly focused on both tails. They obtained two conclusions: (1) In general terms, volume positively depends on the return exclusively in the upper tail of the distribution but not in the lower, which can be interpreted as volume is a key piece able to explain the periodical booms of the market, not its eventual collapses. (2) A marked asymmetric dependence exists between return and volume in the extremes of the distribution, evidenced by extremely high returns tending to be attached to extremely large volumes, but extremely low returns tending not to be associated with disproportionate trading volumes, whether high or low.

Naeem et al. [97] focused on the study of the PVR from the analysis of the asymmetric relationship between returns and trading volumes based on four stock indices also in Asia, developing an alternative measure of dependence by combining several copulas (Clayton, Survival Clayton, and Gumbel) with the univariate GARCH and FIGARCH (Fractionally Integrated GARCH models) in which the marginal distributions of the respective series of returns and volumes are adjusted, proving that the FIGARCH specification substantially improves the estimation of the parameters of each of the proposed copulas. As in [95], we remark that extraordinarily high trading volumes are often related to significant returns, which is due to sudden and sharp declines in the value of financial assets and, more specifically, within financial crisis environments.

3. Materials and Methods

Our objective was to present a multi-perspective design of Larkin's research [98] that enables the analysis of the PVR from different standpoints, depending on the use of different datasets, time horizons, and analytical tools (copulas). The SRA was applied to two different scenarios to provide

a generic and a specific image of this methodology. Instead of using a representative hydrological or meteorological index as an empirical basis (i.e., the standardized precipitation index (SPI) [99]), per se values of four stock market indices commonly employed by the literature in the study of the PVR were selected: DJIA, FTSE100, NIKKEI225, and IBEX35.

In the first case, or generic scenario, all available copulas (26) were applied to a single index (DJIA). Later, in the specific scenario, a single copula was adjusted to three indices (FTSE100, NIKKEI225, and IBEX35). The copula chosen in the second case was the Tawn copula, a family of new-generation copulas derived from the Khoudraji's device copula [100]. In this way, we contribute to the analysis of the PVR with the inclusion of new copulas never or rarely implemented in this research, such as some of those included in the SRA approach. In relation to the construction of the generic scenario, we decided to use a wide database consisting of 20,219 stock trading sessions of the DJIA index, covering the period from 10 January 1928 to 4 August 2009, which were consecutively subdivided into quarterly periods until obtaining 490 observations representing the adjusted closing values of the DJIA at the end of each corresponding session and the final volume of the shares traded at each date.

This temporal accrual as well as the use of data per se allowed us to adapt the original datasets to the methodology proposed by [8]. The analysis of the specific scenario corresponding to the FTSE100, NIKKEI225, and IBEX35 indices involved monthly data of per se price and volume collected during the period from 31 October 2000 to 30 November 2018, which included 218 monthly observations for each stock index. The most representative descriptive statistics of the generic scenario, shown in Table 1, reveal a fundamental aspect: the huge level of variability of variables "price" and "trading volume" when both are measured in per se terms (especially in the latter case).

Table 1. Descriptive statistics and dependence evaluation of DJIA price and volume per se (1928–2009).

(A) Descriptive Statistics							
Variable	**T. Count**	**Mean**	**SEM**	**T. Mean**	**St. Dev.**	**Variance**	**CV**
Price (DJIA)	241	2366	227	1973	3528	12,445,239	149.10
Volume (DJIA)	241	310,592,490	52,3001,976	1.61×10^8	8.1×10^8	6.57×10^{22}	260.92
Variable	**Sum**	**SS**	**Min**	**Q1**	**Median**	**Q3**	**Max**
Price (DJIA)	570,211	4,335,989,110	60	218	827	2448	13,502
Volume (DJIA)	7.49×10^{10}	1.80865×10^{25}	210,000	1,970,000	10,710,000	1.68×10^8	5,531,290,000
Variable	**Range**	**IQR**	**Mode**	**Skewness**	**Kurtosis**	**MSSD**	
Price (DJIA) Volume (DJIA) *	13,441 5.53×10^9	2.3030 166,350,000	240,96 770,000; 880,000; 990,000; 1,440,000	1.77 3.99	1.69 18.10	55,250 1.44×10^{21}	
(B). Dependence Evaluation							
Correlation Price (DJIA)-Volume (DJIA)							
Correlation type		**Correlation Coefficient**		***p*-value**		**Significant at 5%?**	
Kendall rank		0.8279		0		Yes	
Spearman's rank-order		0.9559		0		Yes	
Pearson product-moment		0.7365		0		Yes	

Subtable (A): (*) The analyzed data contain at least five mode values. Only the smallest four have been selected. T. Count: Total Count; SEM: Standard Error of the Mean; T. Mean: Trimmed Mean; CV: Coefficient of Variation; SS: Sum of Squares; IQR: Interquartile Range; MSSD: Mean of the Squared Successive Differences. Subtable (B): Source: Own elaboration.

In the same way, the values per se of the variables "price" and "trading volume" denote a relatively high degree of correlation in terms of the Pearson, Kendall, and Spearman correlation coefficients (0.7365, 0.8279, and 0.9559, respectively), which a priori could be considered significant measures

of dependence. However, as underlined by Frey et al. [101], a high degree of correlation does not necessarily imply real dependence between the involved variables.

Figure 1 shows the huge level of dispersion and variability of both variables. The first two subfigures, elaborated according to Patton [29], exhibit a normalized time series plot of price (DJIA)–volume (DJIA) as well as a scatter plot of log-increments, both series normalized in base 100, according to the equality $100 \times \exp\left\{\sum_{i=1}^{n} \frac{\ln X_i}{\ln X_{i-1}}\right\}$. The third subfigure represents the Pearson regression coefficient of per se prices and volumes of the DJIA over the analyzed time horizon, showing a quasicyclical relationship between prices and transactional volume within this index, which a priori do not appear to be connected with the evolution of the economic cycle. Several phases or trends can be distinguished: relative decline (1934–1957, 1979–1984, and 2000 onwards), stabilization (1967–1977), and increase (1929–1933, 1958–1966, and 1985–1999) in the relationship between the variables in terms of Pearson's linear correlation coefficient (ρ).

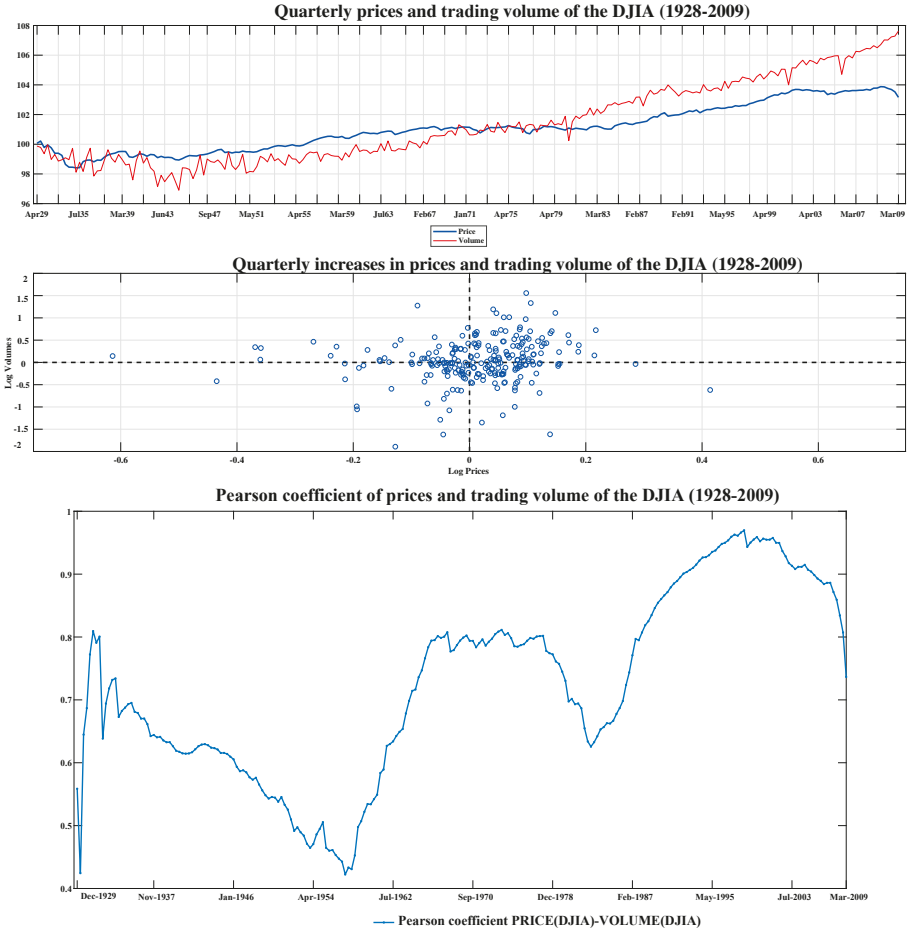


Figure 1. Three different representations of DJIA price-volume evolution and variability during the period 1928–2009. Source: Own elaboration.

Considering per se magnitudes, Figure 2 presents a three-dimensional scatter plot of the DJIA index that links variables X (volume) and Y (price) to the Pearson linear correlation coefficient ($Z = \rho$).

Simply, it can be observed that this chart mostly associates the highest correlation levels of P and V to high per se values of P. Low trading volume per se usually fluctuates within a range from 5.00×10^9 to 15.00×10^9, although sometimes a relatively high degree of correlation between price and low trading volume can be detected (close to 5.00×10^9).

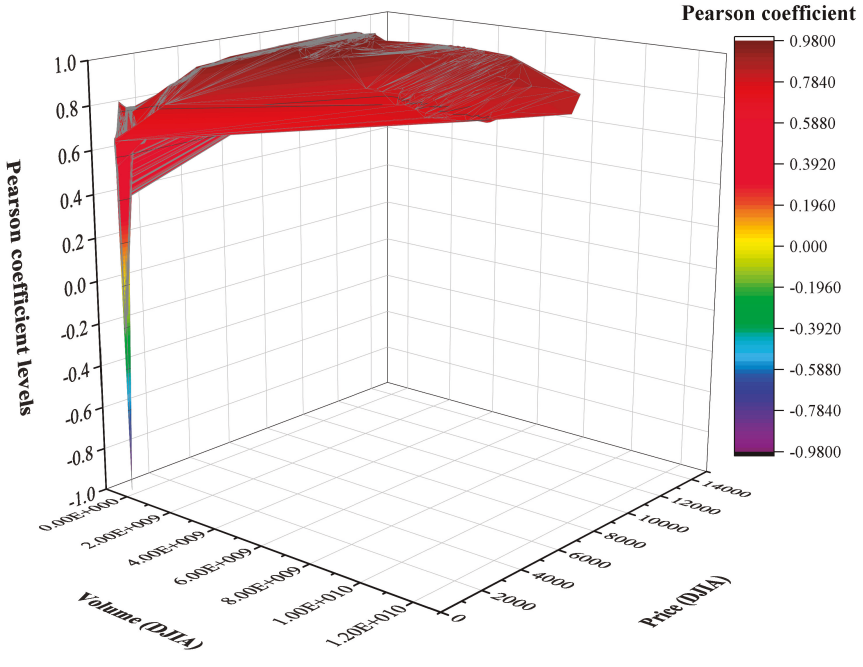


Figure 2. Scatter plot of ρ vs. Price (DJIA)-Volume (DJIA). Source: Own elaboration.

The aim of this paper is to highlight the key aspects of the SRA as an optimal methodological approach for the analysis of the PVR from an empirical perspective that is completely different from the rest of the predominant lines of research. In summary, this methodology can be characterized by: (1) the use of a high number of bivariate copulas (26, see Table 2), especially recommended to simultaneously represent different dependence structures and to conduct prospective inferences based on the chosen variables (not necessarily related to hydrology), such as the variables price and trading volume of a given financial asset or stock index. Notably, to the best of our knowledge, the large number of copulas jointly implemented in the SRA was employed for the first time in the investigation of the PVR. (2) This methodology is based on a unitary reference framework (Bayesian analysis, see Appendix A) in which the hybrid evolution algorithm of the Monte Carlo Markov Chain simulation (MCMCS) was introduced, focusing on the numerical estimation of the subsequent distribution of copula parameters within a context of uncertainty that is relatively similar to the uncertainty observable in financial markets, especially when the different volatility ranges can be conveniently delimited.

Table 2. Families of copulas used in this analysis, specifying their corresponding most common mathematical specifications.

Denomination	Mathematical Representation [A]	Parametric Range	Reference
Gaussian	$\int_{-\infty}^{\phi^{-1}(u)} \int_{-\infty}^{\phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\theta^2}} \exp\left(\frac{2\theta xy - x^2 - y^2}{2(1-\theta^2)}\right) dxdy\, B$	$\theta \in [-1,1]$	[102]
t	$\int_{-\infty}^{t_{\theta_2}^{-1}(u)} \int_{-\infty}^{t_{\theta_2}^{-1}(v)} \frac{\Gamma(\frac{\theta_2+2}{2})}{\Gamma(\frac{\theta_2}{2})\pi\theta_2\sqrt{1-\theta_1^2}} \left(1 + \frac{x^2 - 2\theta_1 xy + \gamma^2}{\theta_2}\right)^{(\frac{\theta_2+2}{2})} dxdy\, C$	$\theta_1 \in [-1,1]$ and $\theta_2 \in (0,\infty)$	[102]
Clayton	$\max(u^{-\theta} + y^{-\theta} - 1, 0)^{-\frac{1}{\theta}}$	$\theta \in [-1,\infty)\backslash 0$	[103]
Frank	$-\frac{1}{\theta}\ln\left[1 + \frac{(\exp(-\theta u) - 1)(\exp(-\theta v) - 1)}{\exp(-\theta) - 1}\right]$	$\theta \in \mathbb{R}\backslash 0$	[102]
Gumbel	$\exp\{-[(-\ln(u))^\theta + (-\ln(v))^\theta]^{\frac{1}{\theta}}\}$	$\theta \in [1,\infty)$	[102]
Independence	uv		[104]
Ali-Mikhail-Haq (AMH)	$\frac{uv}{1-\theta(1-u)(1-v)}$	$\theta \in [-1,1)$	[105]
Joe	$1 - [(1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta]^{\frac{1}{\theta}}$	$\theta \in [1,\infty)$	[102]
Farlie-Gumbel-Morgenstern (FGM)	$uv + [\theta(1-u)(1-v)]$	$\theta \in [-1,1]$	[18]
Gumbel-Barnett	$u + v - 1 + (1-u)(1-v)\exp[-\theta\ln(1-u)\ln(1-v)]$	$\theta \in [0,1]$	[55,106]
Plackett	$\frac{1 + (\theta-1)(u+v) - \sqrt{[1+(\theta-1)(u+v)]^2 - 4\theta(\theta-1)uv}}{2(\theta-1)}$	$\theta \in (0,\infty)$	[107]
Cuadras-Auge	$[\min(u,v)]^\theta (uv)^{(1-\theta)}$	$\theta \in [0,1]$	[108]
Raftery	$\begin{cases} u - \frac{1-\theta}{1+\theta} u^{\frac{1}{1-\theta}} (v^{\frac{-\theta}{1-\theta}} - v^{\frac{1}{1-\theta}}) & \text{if } u \le v \\ v - \frac{1-\theta}{1+\theta} v^{\frac{1}{1-\theta}} (u^{\frac{-\theta}{1-\theta}} - u^{\frac{1}{1-\theta}}) & \text{if } v \le u \end{cases}$	$\theta \in [0,1)$	[18]
Shih-Louis	$\begin{cases} (1-\theta)uv + \theta\min(u,v), \text{ if } \theta \in (\theta,\infty) \\ (1+\theta)uv + \theta(u+v^{-1})\Psi(u+v^{-1}), \text{ if } \theta \in (-\infty,\theta] \\ \Psi(a) = 1, \text{ if } a \ge 0 \\ \Psi(a) = 0, \text{ if } a < 0 \end{cases}$		[109]

Table 2. *Cont.*

Denomination	Mathematical Representation[A]	Parametric Range	Reference
Linear-Spearman	$\begin{cases} ([u+\theta(1-u)]v, & \text{if } v \le u \text{ and } \theta \in [0,1] \\ [v+\theta(1-v)]u, & \text{if } u < v \text{ and } \theta \in [0,1] \\ (1+\theta)uv, \text{ if } (u+v) < 1 \text{ and } \theta \in [-1,0] \\ (uv+\theta(1-u)(1-v), \text{ if } (u+v) \ge 1 \text{ and } \theta \in [-1,0] \end{cases}$	$\theta \in [-1,1]$	[110]
Cubic	$uv[1+\theta(u-1)(v-1)(2u-1)(2v-1)]$	$\theta \in [-1,2]$	[111]
Burr	$u+v-1+[(1-u)^{-\frac{1}{\theta}}+(1-v)^{-\frac{1}{\theta}}-1]^{-\theta}$	$\theta \in (0,\infty)$	[50]
Nelse n	$-\frac{1}{\theta}\log\left\{1+\frac{[\exp(-\theta u)-1][\exp(-\theta v)-1]}{\exp(-\theta)-1}\right\}$	$\theta \in (0,\infty)$	[18]
Galambos	$uv\exp\{(-\ln(u))^{-\theta}+(-\ln(v))^{-\theta}\}^{-\frac{1}{\theta}}$	$\theta \in (0,\infty)$	[12]
Marshall-Olkin	$\min[u^{(1-\theta_1)}v, uv^{(1-\theta_2)}]$	$\theta_1,\theta_2 \in (0,\infty)$	[12]
Fischer-Hinzmann	$\{\theta_1[\min(u,v)]^{\theta_2}+(1-\theta_1)[uv]^{\theta_2}\}^{\frac{1}{\theta_2}}$	$\theta_1 \in [0,1], \theta_2 \in \mathbb{R}$	[112]
Roch-Alegre	$\exp\{1-[(((1-\ln(u))^{\theta_1}-1)^{\theta_2}+((1-\ln(v))^{\theta_1}-1)^{\theta_2})^{\frac{1}{\theta_2}}+1]^{\frac{1}{\theta_1}}\}$	$\theta_1 \in (0,\infty), \theta_2 \in [1,\infty)$	[113]
Fischer-Kock	$uv[1+\theta_2(1-u^{\frac{1}{\theta_1}})(1-v^{\frac{1}{\theta_2}})]^{\theta_1}$	$\theta_1 \in [1,\infty), \theta_2 \in [-1,1]$	
BB1	$\{1+[(u^{-\theta_1}-1)^{\theta_2}+(v^{-\theta_1}-1)^{\theta_2}]^{\frac{1}{\theta_2}}\}^{-\frac{1}{\theta_1}}$	$\theta_1 \in (0,\infty), \theta_2 \in (1,\infty)$	[4]
BB5	$\exp\{-[(-\ln(u))^{\theta_1}+(-\ln(v))^{\theta_1}-((-\ln(u))^{-\theta_1\theta_2}+(-\ln(v))^{-\theta_1\theta_2})^{-\frac{1}{\theta_2}}]^{\frac{1}{\theta_1}}\}$	$\theta_1 \in [1,\infty), \theta_2 \in (0,\infty)$	[4]
Tawn	$\exp\{\ln(u^{(1-\theta_1)})+\ln(v^{(1-\theta_2)})-[(-\theta_1\ln(u))^{\theta_3}+(-\theta_2\ln(v))^{\theta_3}]^{\frac{1}{\theta_3}}\}$	$\theta_1,\theta_2 \in [0,1], \theta_3 \in [1,\infty)$	[12]

A. The different formulations of this table do not necessarily have to be unique. B. ϕ represents the distribution *Gaussian* or standard normal. C. t_{θ_2} denotes the *t student* distribution with θ_2 degrees of freedom. Source: Specifically readapted to this study from [8].

As stated by Johannes and Polson [114], the key aspect of the MCMCS is its ability to easily characterize the complete conditional distributions, $p(\theta|X,Y)$ and $p(X|\theta,Y)$, instead of analyzing the higher-dimensional joint distribution $p(\theta,X|Y)$. The SRA belongs to the class of econometric methods usually applied to the sampling of high-dimensional complex distributions, which implement a hybrid-evolution MCMCS algorithm to infer posterior parameter regions within a Bayesian context. This algorithm is considered a hybrid since it includes a combination of Gibbs steps and Metropolis–Hastings steps [114].

The hybrid-evolution MCMCS algorithm starts with an intelligent starting point selection, structured according to the use of adaptive metropolis (AM), differential evolution (DE), and snooker update. Table 3 summarizes, in descending order, the working schema implemented in the algorithm developed by Sadegh et al. [8]. For the sake of brevity, intermediate iterative conditions (i.e., end do, end if, etc.) have been omitted from the table.

Table 3. Description of the basis scheme of the hybrid MCMCS algorithm implemented in the SRA approach.

Intelligent prior sampling.
Draw $LN(\geq N)$ samples from prior $(p(\theta))$ using Latin Hypercube Sampling (*LHS*).
Randomly assign the LHS samples to N complexes.
Selecting the best sample in each complex as the starting point of a Markov chain (CH).
Snooker update (with a 10% of probability).
Drawing 3 samples, r_{1-3}, from parameter space $\{1:D\}\backslash\{i\}$.
Finding the update direction $Z = CH_j - CH_{r_1}$.
Projecting CH_{r2} and CH_{r_3} onto Z, to get Z_{p_1} and Z_{p_2}.
Creating a proposal $CH^* = CH_j + \gamma_1(Z_{p_2} - Z_{p_1})$.
Computing the Metropolis ratio (1) $\quad MR = \dfrac{\mathcal{L}(CH^*)\|CH^* - CH_{r_1}\|^{D-1}}{\mathcal{L}(CH_i)\|CH_i - CH_{r_1}\|^{D-1}}$.
Adaptive Metropolis and differential evolution updating (with a 90% of probability).
Randomly select d dimensions from D-dimensional parameter space to update (within Gibbs sampling).
Creating a proposal sample (1) $\quad CH^*(d) = CH_i(d) + (1-\beta)N(0_d, \gamma_2^2\Sigma_d) + \beta N(0_d, \gamma_3^2 I_d)$.
Creating a proposal sample (2) $\quad CH^*(d) = CH_i(d) + \gamma_4(CH_{r_2}(d) - CH_{r_1}(d)) + e$.
Computing the Metropolis ratio (2) $\quad MR = \dfrac{\mathcal{L}CH^*}{\mathcal{L}CH_i}$.
Accepting proposal CH , with probability $\max(MR, 1)$, and update current chain, CH_i.
Checking for Gelman-Rubin $\hat{R}$ convergence diagnostic.

LN = number of samples drawn from the prior distribution $[p(\theta)]$, using Latin Hypercube Sampling (LHS) and N = number of Markov chains (CH). D = the dimension of the entire parameter space, d = the dimension of the subspace of the parameters randomly selected for update (Metropolis within Gibbs sampling), T = the total number of iterations, and N_{AM} = the number of chains selected for the Adaptive Metropolis algorithm. $\gamma_1 - \gamma_4$ = "jump factors", where γ_1 is randomly selected, $\gamma_2 = 2.38/\sqrt{d}, \gamma_3 = 0.1/\sqrt{d}$ and $\gamma_4 = 2.38/\sqrt{2d}$. Σ_d = adaptive covariance matrix, based on the last 50% samples of the Markov chains. Source: Specifically readapted to this study from [8].

4. Results

Despite the SRA employing a good number of new generation copulas, with some of them complex in mathematical terms (i.e., Plackett or Shih-Louis), Table 4 shows that two copulas with a not very analytically complex, Li et al. [102] and Frees and Valdez [50] best fit the price–volume time series

of the DJIA during the considered period (1928–2009), emphasizing that, in all cases, the specified selection criteria coincide except for three copulas: Galambos, BB1, and BB5.

Complementarily, Table 5 provides estimations of the parameters of each copula (Par) by fixing a range of 95% of uncertainty in their estimation (Unc-Range) through the application of local optimization and MCMCS. The copulas with best performance (Rank) are defined in terms of the root mean square error (RMSE) and the Nash–Sutcliff Efficiency (NSE) criteria. At this point, the existing literature usually employs local optimization algorithms when estimating the parameters of copulas with the consequent risk of being trapped in local optima, thus often obtaining unbiased and nonsignificant results [8]. Conversely, the hybrid-evolution MCMCS algorithm used in the SRA overcomes this initial limitation by determining an efficient estimator of the global optimum as well as an accurate approximation of uncertainties in the content of a Bayesian conceptual framework in the form of isolines, which is another of the improvements provided by this methodology to PVR analysis.

Table 4. Selection of copulas fitted to the DJIA index (1928–2009) based on three different criteria. Performance-criterion ranking amongst the implemented copulas.

Rank	Max-Likelihood	AIC	BIC	Criteria Coincidence
1	*Joe*	*Joe*	*Joe*	YES
2	*Burr*	*Burr*	*Burr*	YES
3	*Fischer-Hinzmann*	*Fischer-Hinzmann*	*Fischer-Hinzmann*	YES
4	*Roch-Alegre*	*Roch-Alegre*	*Roch-Alegre*	YES
5	*Tawn*	*Tawn*	*Tawn*	YES
6	*Gumbel*	*Gumbel*	*Gumbel*	YES
7	*BB5*	*Galambos*	*Galambos*	NO
8	*BB1*	*BB5*	*BB5*	NO
9	*Galambos*	*BB1*	*BB1*	NO
10	*Marshal-Olkin*	*Marshal-Olkin*	*Marshal-Olkin*	YES
11	*Cuadras-Auge*	*Cuadras-Auge*	*Cuadras-Auge*	YES
12	*Nelsen*	*Nelsen*	*Nelsen*	YES
13	*Frank*	*Frank*	*Frank*	YES
14	*Linear-Spearman*	*Linear-Spearman*	*Linear-Spearman*	YES
15	*Shih-Louis*	*Shih-Louis*	*Shih-Louis*	YES
16	*t*	*t*	*t*	YES
17	*Gaussian*	*Gaussian*	*Gaussian*	YES
18	*Raftery*	*Raftery*	*Raftery*	YES
19	*Clayton*	*Clayton*	*Clayton*	YES
20	*Plackett*	*Plackett*	*Plackett*	YES
21	*AMH*	*AMH*	*AMH*	YES
22	*FGM*	*FGM*	*FGM*	YES
23	*Fischer-Kock*	*Fischer-Kock*	*Fischer-Kock*	YES
24	*Cubic*	*Cubic*	*Cubic*	YES
25	*Independence*	*Independence*	*Independence*	YES
26	*Gumbel-Barnet*	*Gumbel-Barnet*	*Gumbel-Barnet*	YES

Souce: Own elaboration.

The analysis of the SRA applied to the NIKKEI225, FTSE100, and IBEX35 indices using the Tawn's copula is summarized in Table 6, similarly to Table 5. The price–volume dependence structure of the per se NIKKEI225 index is optimal in accordance with the NSE criterion, as it is very close to unity (0.9914), indicating an almost perfect model fitting. The per se IBEX35 adjustment is relatively optimal (0.9737), being lower for the FTSE100 (0.8235). The range of uncertainty of the parameters defining the Tawn's copula (θ_1, θ_2, and θ_3, Table 2) is considerably lower in the Nippon index than in the other two stock market indices.

Figure 3 shows that each stock exchange index corresponds to a certain typology of its probability isolines. Rows 1 to 3 refer to the analyzed indices, whereas columns correspond to the following specifications: (A) fitted empirical copulas probabilities, (B) fitted empirical copulas, and (C) return

period copulas, calculated according to [115] by considering the joint return $\left(\frac{1}{1-C(u,v)}\right)$ as a measure of the dependence structure between the observed price peaks and trading volumes.

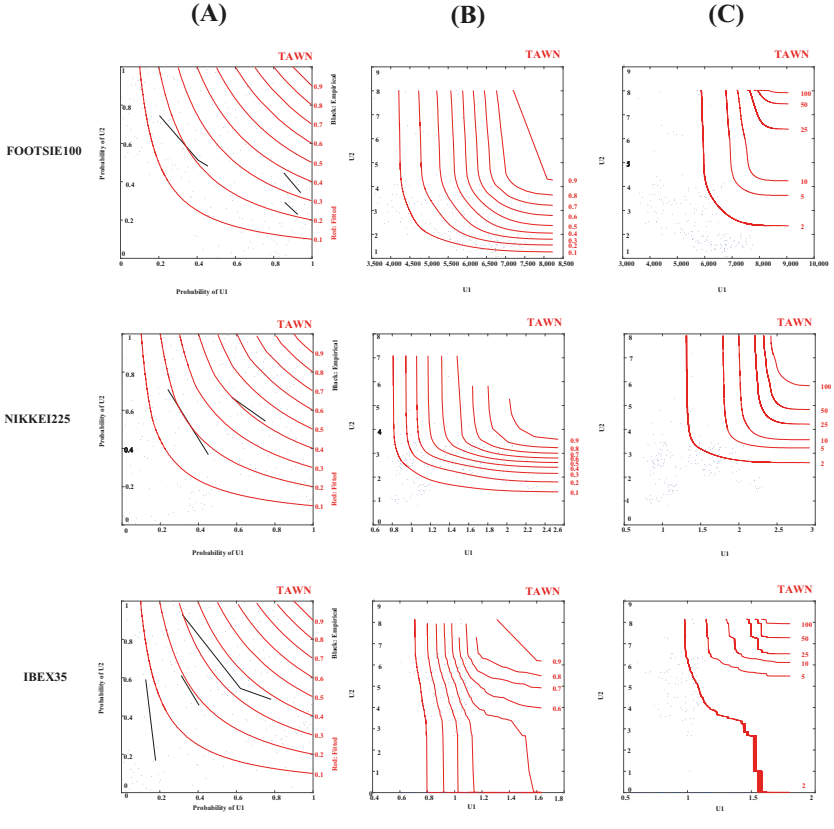


Figure 3. Probability isolines of Tawn's copula for FOOTSIE100, NIKKEI225, and IBEX35 indices. Source: Own elaboration.

The isolines derived from the application of Tawn's copula are ostensibly biased toward the upper left corner, which seems to indicate a low probability of occurrence of the price (U_1) synchronously linked to a high probability of occurrence of the trading volume (U_2) (both measured in magnitudes per se). Likewise, given the joint representation of the probability isolines and the empirical estimates of the joint probability distributions, the trends of the FOOTSIE100 and NIKKEI225 indices are fairly similar, although in the former index, high prices use to be related to trading volumes lower than those shown in the Japanese stock market. The *P*—*V* relationship in the IBEX35, although following a similar pattern, differs to some extent from the analysis of the other two indices, as low prices seem to be more related to high trading volumes quotas. This type of asymmetric and skewed dependence structure can be considered a common pattern of the three indices analyzed, equally extrapolated to the analysis of the fitted empirical copulas and return period copulas.

Table 5. Copula parameters estimation: DJIA (1928–2009).

Copula Name	Rank	RMSE [A]	NSE [B]	Par#1-Local	Par#2-Local	Par#3-Local	Par#1-MCMC	95%-Par#1-Unc-Range	Par#2-MCMC	95%-Par#2-Unc-Range	Par#3-MCMC	95%-Par#3-Unc-Range
AMH	21	1.4462	0.9051	1,0000	NaN	NaN	1.0000	[0.9793;0.9998]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
BB1	8	0.1864	0.9984	0.0001	6.8711	NaN	0.0012	[0.0007;0.0890]	6.8738	[6.3355;7.4168]	NaN	[NaN;NaN]
BB5	7	0.1864	0.9984	1.0075	6.1177	NaN	6.8666	[1.0085;7.2476]	0.0434	[0.0260;6.3875]	NaN	[NaN;NaN]
Burr	2	0.1564	0.9989	0.0884	NaN	NaN	0.0884	[0.0832;0.0948]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
Clayton	19	0.2617	0.9969	2.6215	NaN	NaN	27.7501	[22.3225;34.3061]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
Cuadras-Auge	11	0.1919	0.9983	0.9401	NaN	NaN	0.9401	[0.9327;0.9474]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
Cubic	24	2.1634	0.7877	2.0000	NaN	NaN	1.9998	[1.3902;1.9954]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
FGM	22	1.6844	0.8713	1.0000	NaN	NaN	1.0000	[0.9550;0.9999]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
Fischer-Hinzmann	3	0.1653	0.9988	0.9745	−1.7302	NaN	0.9738	[0.9622;0.9814]	−1.7073	[−2.0944;−1.2435]	NaN	[NaN;NaN]
Fischer-Kock	23	1.6846	0.8713	1.0000	1.0000	NaN	1.0005	[1.0015;1.1530]	0.9996	[0.9498;0.9996]	NaN	[NaN;NaN]
Frank	13	0.2041	0.9981	20.1073	NaN	NaN	25.7416	[23.5320;28.7390]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
Galambos	9	0.1865	0.9984	6.1687	NaN	NaN	6.1688	[5.6654;6.7940]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
Gaussian	17	0.2131	0.9979	0.9140	NaN	NaN	0.9785	[0.9735;0.9827]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
Gumbel	6	0.1864	0.9984	4.3567	NaN	NaN	6.8717	[6.3810;7.5823]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
Gumbel-Barnet	26	2.2788	0.7645	0.0000	NaN	NaN	0.0000	[0.0003;0.0321]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
Independence	25	2.2788	0.7645	NaN	NaN	NaN	NaN	[NaN;NaN]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
Joe	1	0.1563	0.9989	12.0596	NaN	NaN	12.0583	[11.3142;12.9402]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
Linear-Spearman	14	0.2079	0.9980	0.9238	NaN	NaN	0.9238	[0.9113;0.9340]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
Marshal-Olkin	10	0.1893	0.9984	8.6736	0.1012	NaN	0.9484	[0.9382;0.9714]	0.9298	[0.9019;0.9410]	NaN	[NaN;NaN]
Nelsen	12	0.2041	0.9981	24.0724	NaN	NaN	25.7465	[23.6645;28.8592]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
Plackett	20	0.4651	0.9902	35.0000	NaN	NaN	34.9994	[34.0284;34.9944]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
Raftery	18	0.2602	0.9969	0.9522	NaN	NaN	0.9522	[0.6394;0.9791]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
Roch-Alegre	4	0.1707	0.9987	0.0001	9.0334	NaN	0.0007	[0.0029;0.1435]	9.0521	[8.3601;9.7642]	NaN	[NaN;NaN]
Shih-Louis	15	0.2079	0.9980	0.9238	NaN	NaN	0.9237	[0.9128;0.9348]	NaN	[NaN;NaN]	NaN	[NaN;NaN]
t	16	0.2085	0.9980	0.9363	3.9764	NaN	0.9809	[0.9745;0.9849]	0.5611	[0.2827;30.9390]	NaN	[NaN;NaN]
Tawn	5	0.1793	0.9985	0.9787	0.9466	11.5247	0.9786	[0.9580;0.9957]	0.9457	[0.9233;0.9662]	11.6332	[8.4142;23.0476]

A. Root Mean Square Error. B. Nash-Sutcliff Efficiency. Source: Own elaboration.

Table 6. Tawn copula parameters estimation: NIKKEI225, IBEX35, and FTSE100 (2000–2018).

Index	RMSE	NSE	Par#1-Local	Par#2-Local	Par#3-Local	Par#1-MCMC	95%-Par#1-Unc-Range	Par#2-MCMC	95%-Par#2-Unc-Range	Par#3-MCMC	95%-Par#3-Unc-Range
NIKKEI 225	0.2621	0.9914	0.0744	0.2153	34.9816	0.0738	[0.0551–0.1248]	0.219	[0.0977–0.4725]	25.9197	[2.2342–34.4984]
IBEX 35	0.3901	0.9737	0.005	0.8691	11.1339	0.0049	[0.0000–0.9876]	0.9969	[0.0000-0.9969]	22.6851	[1.0000–34.4965]
FTSE 100	0.9005	0.8235	0.3857	0	13.1732	0.0097	[0.0000–0.9896]	0.5024	[0.0000–0.9881]	1	[1.0000–35.0000]

Source: Own elaboration.

Figure 4 shows the degree of uncertainty associated with the three parameters defining the Tawn's copula. Figure 4 exhibits the specification of the copula parameters generated by the MCMCS through a Bayesian framework. Blue bins represent the MCMC-obtained parameters, blue crosses (bottom of each plot) denote the maximum likelihood estimation parameters, and red asterisks (top of each plot) indicate the copula parameter value obtained by local optimization.

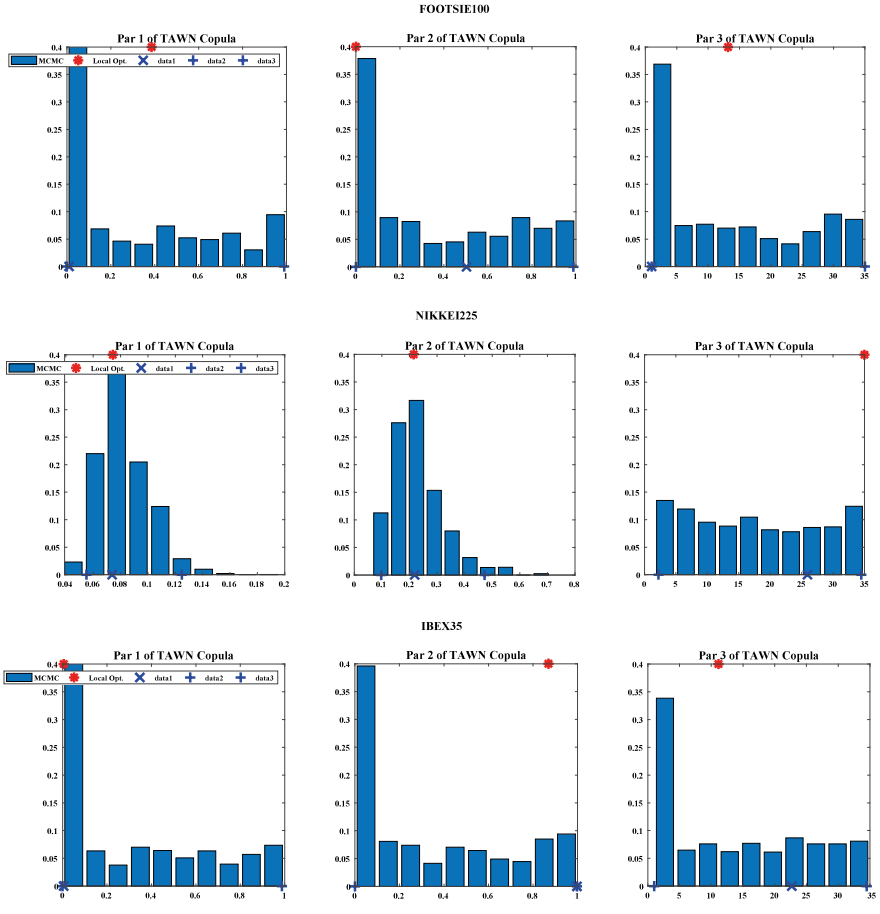


Figure 4. Posterior distribution of fitted Tawn's copula on FOOTSIE100, NIKKEI225, and IBEX35 indices obtained by the MCMCS. Source: Own elaboration.

In a context characterized by minimal uncertainty when specifying the parameters of the copula in each market, the parameters obtained by the local optimization algorithm should coincide with the mode of the distribution calculated through the MCMCS. However, this was only observed in the NIKKEI225 (parameters 1 and 2) and IBEX35 (parameter 1) and was not contrastable for any of the three parameters obtained from the Tawn's copula to the FOOTSIE100 index. These results are consistent with the previously calculated uncertainty ranges and with the delimitation of the degree of goodness of the adjustment performed by the NSE criteria (Table 6), according to which the NIKKEI225 index represented a quasiperfect fitting to this copula, followed by the IBEX35, and, to a lesser extent, the FOOTSIE100. This can be justified by the different range of variation of the parameters obtained

for each index, where the FOOTSIE100 index is associated with a higher level of uncertainty compared with NIKKEI225 and IBEX35.

5. Discussion

The application of the SRA provides an alternative and innovative approach to PVR based on the simultaneous application of 26 copulas, which facilitated the analysis of their dependence structures and implicit morphology according with their probability isolines. Many of these copulas are dissimilar in form, although quite similar in performance. This also allowed us to model the relationship between prices and trading volumes from different points of view, quantifying the uncertainty underlying to the specification of the parameters defining each copula. The PVR, usually characterized by a markedly asymmetric relationship [13], is reinforced by the application of the SRA, since several of the copulas used in this methodology (e.g., Galambos, Bernstein, Tawn, etheory of copulas.) are especially effective in the study of phenomena with underlaying asymmetric skewed dependence structures.

From an empirical point of view, the joint implementation of the 26 copulas in the DJIA (generic scenario) confirmed that Joe's copula is able to more efficiently model the dependence structures of this index. Framing our findings with the existing literature, the use of Tawn's copula in the FOOTSIE100, NIKKEI225, and IBEX35 indices (specific scenario) confirms Ying [68]'s findings in their seminal analysis of the S&P 500 index. Similarly, our results confirm the analysis of the NIKKEI225 completed by Bremer and Kato [116], according to which an asymmetric relationship could be observed (negative correlation between past prices and current trading volume). This relationship was explained in the FOOTSIE100 by Huang and Masulis [117] based on the existence of a minority of informed-trading investors who simply sought immediate liquidity. The asymmetric PVR detected in the IBEX35 aligns with that already reported in the literature (see, for example, [118]). Its differentiated nature with respect to the other two indices is probably due to, according to [119] in the Spanish financial markets (Mercado Continuo), a strong linear causal relationship from returns to trading volume. Specifically, periods with high returns are usually followed by periods with particularly high trading volume. Such guidelines are comparable to those detected in other works [95,97], which explicitly used copulas in the study of PVR in the Asian financial markets, repeatedly verifying the existence of an inverse relationship between prices and volumes traded, in both cases foreseeably increased by the effects of the 1997 Asian financial crisis.

One of the improvements associated with the application of the SAR approach to the PVR is facilitating the analysis of both variables from a large number of copulas by defining the relationship based on ranges of uncertainty applying the RMSE and NSE criteria (see Tables 5 and 6), independent of the degree or sign of the linear correlation exhibited by the Pearson correlation coefficient (ρ), which, to the best of our knowledge, is an entirely new application in the field of quantitative finance of future utility for researchers and investors.

6. Conclusions

The main contribution of this research is the analysis of the existing relationship between prices and trading volume from multiple copulas, which allowed us to comparatively abstract the underlying dependency structures of both variables to establish possible analogies or differences. One of the most important limitations related to the empirical application of copulas is solved, which is employing a limited number of standard copulas when, in reality, there are multiple copulas not yet well extended in the literature [11]. Through the empirical methodology introduced in this article, the versatility of copulas increases when they are simultaneously combined with the polyvalence of the Bayesian analysis and with the hybrid-evolution MCMCS algorithm proposed by Sadegh et al. [8]. We are the first to implement the SRA, not just in PVR analysis, but in the ambit of quantitative finance. More specifically, for practical purposes, PVR analysis is decisive for both academics and practitioners, since, following the scheme constructed by Karpoff [13], it has the following implications: (1) it generates additional information regarding the structure of financial markets; (2) from an

empirical point of view, it is fundamental in the generation of case studies that jointly use prices and trading volume, facilitating the implementation of analyses and inferences; (3) it is a crucial element in the study of the empirical distribution of speculative prices; and (4) its research would be particularly indicated in the futures markets where, a priori, the variability of prices used to affect the trading volume.

Additionally, we tried to answer and reconcile three questions linked to the theory of copulas: which copula is the "right one" [120], which copula should be used [20], and why copulas have been successful in many practical applications [44]? Versatility is the key term that best defines a copula; therefore, the most appropriate copula for analyzing a particular issue is the one that best summarizes its implicit dependence structures. Hence, copulas have been so successful in different fields of study.

A first conclusion to be drawn from this work is that the potential of the theory of copulas could be significantly reduced if certain copulas-type are systematically used in the analysis of bivariate time series. Precisely, this was the factor that caused a certain reluctance toward the use of copulas when the Gaussian copula [10] was employed massively in almost any scientific field, without considering either the intrinsic nature of the phenomena analyzed or that the use of a large number of copulas can substantially improve the knowledge of the different relations of dependence observable in a given dataset [50]. Thus, the SRA is not simply limited to the task of choosing and fitting the copula [121], but following the transdisciplinary perspective of econophysics, it supposes a new framework in the analysis of the PVR, extrapolated from the field of hydrology, which is directly applicable to many other areas such as quantitative finance.

Since Karpoff [13], the PVR has been practically subsumed to the generalization of the significance of Pearson's linear correlation coefficient of the price and trading volume variables. However, an alternative is provided in this methodology since the classical optimization methods applied to copulas often get trapped in local minima. The SRA is able to conveniently overcome this limitation by accurately describing the dependence structure of variables P and V and, importantly, by allowing the analysis of uncertainties given a determined time horizon (or length of record, see [8]). Another contribution of this work that may be important for future lines of research is the incorporation in the methodology of copula probability isolines in the analysis of PVR, which approximates this research to the multifractal models of Mandelbrot [22].

In our opinion, other future lines of research related to this work include, for example, the analysis of the role played by floating capital (outstanding shares vs. restricted shares) in the context of PVR, an aspect which has been often overlooked in the literature, or the rigorous enunciation and detailed compilation of those empirical stylized facts defining the price–volume time series, as well as the definitive consolidation of the works that have analyzed PVR from the perspective of the market microstructure of Garman [122]. This research could be gradually applied to the area of behavioral finance following the path of works such as Gomes [123], in which the analysis of PVR is directly connected to the prospective theory of Kahneman and Tversky [124].

Author Contributions: Conceptualization, methodology, software, writing—original draft preparation, P.A.M.C.; resources, writing—review and editing, funding acquisition, S.C.R.; investigation, data curation, supervision, M.d.C.V.M. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was partially supported by the project "La sostenibilidad del sistema nacional de salud: reformas, estrategias y propuestas" (Ministry of Economy and Competitiveness, DER2016-76053-R).

Acknowledgments: The authors would like to thank P. Embrechts (ETH, Zürich) for making us aware of some limitations derived from the misuse of copulas in certain specific assets, as well as E. Ragno (University of California at Irvine) and M. Sadegh (Boise State University, Idaho), two coauthors of the SRA, who advised and encouraged us in the use of this methodology given its first implementation in the area of quantitative finance.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PVR	price–volume relationship.
SRA	SRA approach.
MCMS	Markov Chain Monte Carlo simulation.
MDH	Mixture of Distribution Hypothesis.
SIAH	Sequential Information Arrival Hypothesis.
DBH	Dispersion of Beliefs Hypothesis.
NTH	Noise Trader Hypothesis.

Appendix A. The Bayesian Perspective of the SRA Approach

The Bayesian methodology constitutes one of the most recurrent approaches in economic and financial research, considered as an alternative third way to traditional perspectives. In previous studies [125,126], we can find an extensive introduction to Bayesian methods applied to finance, which created an important field of implementation in those contexts characterized by a high uncertainty, such as stress testing study cases [121,127] and the risk management optimization or the estimation of GARCH models in environments of extreme volatility [128,129]. Shemyakin [130] is essential for studying the Bayesian estimate of copulas based on the simplicity of Bayes' theorem (A1), that is, univocally assigning the corresponding uncertainties representatives of each parameter to the model and estimating its posterior distribution, the starting point of the SRA:

$$p(\theta|\tilde{Y}) = \frac{p(\theta)p(\tilde{Y}|\theta)}{p(\tilde{Y})}, \tag{A1}$$

where $p(\theta)$, $p(\theta|\tilde{Y})$, $p(\tilde{Y}|\theta) \cong \mathcal{L}(\theta|\tilde{Y})$, and $p(\tilde{Y}) = \int_\theta p(\tilde{Y}|\theta)\mathrm{d}\theta$ denote prior and posterior distribution (of parameters), likelihood function and coined (or real) evidence, respectively. Under the hypothesis that error residuals are Gaussian-distributed with mean zero, uncorrelated, and homoscedastic, the likelihood function can be reformulated as:

$$\mathcal{L}(\theta|\tilde{Y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \exp\left\{-\frac{1}{2}\tilde{\sigma}^{-2}[\tilde{y}_i - y_i(\theta)]^2\right\}, \tag{A2}$$

and logarithmically transformed into the formula [8]:

$$\ell(\theta|\tilde{Y}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\tilde{\sigma}^2 - \frac{1}{2}\tilde{\sigma}^{-2}\sum_{i=1}^{n}[\tilde{y}_i - y_i(\theta)]^2, \tag{A3}$$

where $\tilde{\sigma}$ is an estimate of the standard deviation of the measurement error given by:

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n}[\tilde{y}_i - y_i(\theta)]^2}{n}, \tag{A4}$$

which allows us to simplify (A2) into:

$$\ell(\theta|\tilde{Y}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2} - \frac{n}{2}\ln\frac{\sum_{i=1}^{n}[\tilde{y}_i - y_i(\theta)]^2}{n} \tag{A5}$$

Finally, eliminating the constant terms of (A5), we would obtain a simplified equivalent log-likelihood function as:

$$\ell(\theta|\tilde{Y}) \approx -\frac{n}{2}\ln\frac{\sum_{i=1}^{n}[\tilde{y}_i - y_i(\theta)]^2}{n}. \tag{A6}$$

Once the data have been modeled according to any of the available copulas (Table 2), the SRA evaluates the goodness of fit using three different criteria: max-likelihood [131], Akaike information criterion (AIC) [132,133], and Bayesian information criterion (BIC) [134], taking the primary error residuals function as a reference under the assumption that, given a set of parameters, its maximum likelihood level completely minimizes the residuals between the model simulations and their linked observations. Notably, these assumptions are explicitly referred to the distribution of residual error that is applied to construct the likelihood function that summarizes the distance between the given observations and the prospective model simulations.

Appendix B. Brief Insight into the Theory of Copulas

The background of the theory of copulas can be traced to the works of Fréchet, Hoeffding, Menger, Féron, Gumbel, and Dell'Aglio, most of them analyzing the relationships between bivariate and trivariate distributions with their corresponding univariate marginal distributions. According to Sempi [135], the basis of the theory of copulas was established by Fréchet [136] and can be synthesized schematically according to the dimensions of Fréchet [136] and Hoeffding [137].

An n-dimensional *copula* C is a multivariate distribution function on the n-dimensional hypercube $[0,1]^n$ with uniformly distributed marginals.

The Sklar's theorem [15] is the starting point for the construction, development, and modeling of a new class of functions (or dependence functions, according to Galambos [138]), which have been generically denominated copulas since Sklar [15], who nominalized them using the Latin term *copulæ* ("couples") [4].

In short, this theorem [139] states that given a n-dimensional random vector $X = (X_1, X_2, \ldots, X_n)$ with joint distribution function F and marginal distribution functions $F_1, F_2, \ldots, F_n$, there exists an n-dimensional copula C, such that for every $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, Equation (A7) is satisfied. For absolutely continuous distributions, the copula C is unique.

Conversely, if C is the n-dimensional copula corresponding to a multivariate distribution function F with marginal distribution functions $F_1, F_2, \ldots, F_n$, then C can be expressed as:

$$C(u_1, \ldots, u_n) = F(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n)) \tag{A7}$$

and its copula density or probability function is given by:

$$c(u_1, \ldots, u_n) = \frac{f(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n))}{f_1(F_1^{-1}(u_1)) \cdots f_n(F_n^{-1}(u_n))}. \tag{A8}$$

If the joint distribution function is n times differentiable, the partial derivatives of order n can be calculated in (A7), by obtaining:

$$f(x) \equiv \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F(x) = \prod_{i=1}^{n} f_i(x_i) \times \frac{\partial^n}{\partial u_1 \partial u_2 \cdots \partial u_n} C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n))$$

$$\equiv \prod_{i=1}^{n} f_i(x_i) \times c(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)), \tag{A9}$$

from where:

$$\log f(x) = \sum_{i=1}^{n} \log f_i(x_i) + \log c(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)). \tag{A10}$$

That is, the joint density function is equal to the product of the marginal densities and the density of the copula (represented by c [27]), from which it follows that the joint logarithmic probability is equal to the sum of the univariate logarithmic likelihoods and the copula logarithmic likelihood, which is a feature of extreme utility for the parametric estimation of multivariate model. Therefore, according to the Sklar's theorem (A7) and considering the equivalence relation (A9), given any couple of variables X and Y with respective marginal distributions $u = F(x_t)$ and $v = G(y_t)$ and joint distribution function $J(x_t, y_t)$, there is a copula C for all (x_t, y_t) in $\mathbb{R}^2$, which relates them according to the equation:

$$J(x_t, y_t) = C(F(x_t), G(y_t)). \tag{A11}$$

Again, calculating the partial derivatives in both terms of Equation (A11), we obtain:

$$\frac{\partial^2 J(x_t, y_t)}{\partial x_t \partial y_t} = \frac{\partial^2 C(F(x_t), G(y_t))}{\partial F \partial G} f(x_t) g(y_t)), \tag{A12}$$

which allows us to model the marginal distributions and the dependence structure between the variables separately from a certain copula [95].

Thus, the Sklar's theorem implies that the dependence relation between different variables can be completely subsumed to the construction of a copula, a process that can be summarized in two consecutive steps [51,139]: (1) identification of associated marginal distributions, and (2) election of a certain copula that appropriately represents the interrelations between the variables, so that the dependence between n random variables $X_1, X_2, \ldots, X_n$ is theoretically explained in its entirety from its joint distribution function $F(x_1, \ldots, x_n) = \mathbb{P}[X_1 \leq x_1, \ldots, X_n \leq x_n]$ [56].

References

1. Jovanovic, F.; Schinckus, C. The History of Econophysics' Emergence: A New Approach in Modern Financial Theory. *Hist. Political Econ.* **2013**, *45*, 443–474. [CrossRef]
2. Schinckus, C.; Jovanovic, F. Towards a transdisciplinary econophysics. *J. Econ. Methodol.* **2013**, *20*, 164–183. [CrossRef]
3. Hurst, H.E. Long Term Storage Capacity of Reservoirs. *Trans. Am. Soc. Civ. Eng.* **1951**, *116*, 770–799.
4. Genest, C.; Favre, A.C. Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask. *J. Hydrol. Eng.* **2007**, *12*, 347–368. [CrossRef]
5. Reiß, R.D.; Thomas, M. *Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields*; Birkhäuser: Basel, Switzerland, 2007.
6. Schweizer, B. Introduction to Copulas. *J. Hydrol. Eng.* **2007**, *12*, 346. [CrossRef]
7. Genest, C. Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask. In Proceedings of the NIPS 2011 Workshop on Copulas in Machine Learning, Sierra Nevada, Spain, 16 December 2011; Technical Report; American Society of Civil Engineers: Reston, VA, USA, 2011.
8. Sadegh, M.; Ragno, E.; AghaKouchak, A. Multivariate Copula Analysis Toolbox (MvCAT): Describing dependence and underlying uncertainty using a Bayesian framework. *Water Resour. Res.* **2017**, *53*. [CrossRef]
9. Genest, C.; Gendron, M.; Bourdeau-Brien, M. The Advent of Copulas in Finance. *Eur. J. Financ.* **2009**, *15*, 609–618. [CrossRef]
10. Li, D.X. On Default Correlation: A Copula Function Approach. *J. Fixed Income* **2000**, *9*, 4. [CrossRef]
11. Nadarajah, S.; Afuecheta, E.; Chan, S. A Compendium of Copulas. In *Probability and Statistics Group School of Mathematics, Research Reports*; Technical Report 10; The University of Manchester: Manchester, UK, 2016.
12. Huynh, V.N.; Inuiguchi, M.; Denoeux, T. (Eds.) *Integrated Uncertainty in Knowledge Modelling and Decision Making: Proceedings of the 4th International Symposium, IUKM 2015, Nha Trang, Vietnam, 15–17 October 2015*; Lecture Notes in Artificial Intelligence; Springer International Publishing: Cham, Switzerland, 2015; Volume 9376.

13. Karpoff, J.M. The Relation between Price Changes and Trading Volume: A Survey. *J. Financ. Quant. Anal.* **1987**, *22*, 109–126. [CrossRef]
14. Martín Cervantes, P.A. Hacia un Modelo Estocástico Eficiente Para la Valoración de Activos Financieros Basado en el Volumen de Negociación: Fundamentos Teóricos e Implementación Práctica. Ph.D. Thesis, Universidad de Almería, Almería, Spain, 2017.
15. Sklar, A. Fonctions de répartition à n dimensions et leurs marges. *Publ. L'Institut Stat. L'Université Paris* **1959**, *8*, 229–231.
16. Quesada, J.J.; Rodríguez, J.A.; Úbeda, M. What are copulas? In *Monografías del Seminario Matemático García de Galdeano*; Universidad de Zaragoza: Zaragoza, Spain, 2003; pp. 499–506.
17. Joe, H. *Multivariate Models and Multivariate Dependence Concepts*; Chapman & Hall/CRC Monographs on Statistics & Applied Probability; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1997; Volume 73.
18. Nelsen, R.B. *An Introduction to Copulas*, 1st ed.; Lecture Notes in Statistics; Springer: New York, NY, USA, 1999; Volume 139.
19. Durante, F.; Sempi, C. *Principles of Copula Theory*; CRC Press: Boca Raton, FL, USA, 2015.
20. Embrechts, P. Copulas: A Personal View. *J. Risk Insur.* **2009**, *76*, 639–650. [CrossRef]
21. Daníelsson, J. *Financial Risk Forecasting: The Theory and Practice of Forecasting Market Risk with Implementation in R and Matlab*; The Wiley Finance Series; John Wiley & Sons: Chichester, UK, 2011.
22. Mandelbrot, B.B. Heavy Tails in Finance for Independent or Multifractal Price Increments. In *Handbook of Heavy Tailed Distributions in Finance*; Rachev, S.T., Ed.; North-Holland: Amsterdam, The Netherlands, 2003; Volume 1, Chapter 1, pp. 1–34.
23. Rachev, S.T.; Menn, C.; Fabozzi, F.J. Copulas. In *Fat-Tailed & Skewed Asset Return Distributions: Implications for Risk Management, Portfolio Selection, and Option Pricing*; Chapter 6; John Wiley & Sons: Hoboken, NJ, USA, 2005; pp. 71–80.
24. Junker, M.; Szimayer, A.; Wagner, N. Nonlinear term structure dependence: Copula functions, empirics, and risk implications. *J. Bank. Financ.* **2006**, *30*, 1171–1199. [CrossRef]
25. Ning, C.; Xu, D.; Wirjanto, T.S. *Modeling Asymmetric Volatility Clusters Using Copulas and High Frequency Data*; Technical Report 6, Working Papers; Ryerson University, Department of Economics: Toronto, ON, Canada, 2009.
26. Ibragimov, R.; Mo, J.; Prokhorov, A. *Fat Tails and Copulas: Limits of Diversification Revisited*; Technical Report 2015-06, Working Papers; University of Sydney Business School, Discipline of Business Analytics: Sydney, Australia, 2015.
27. Patton, A.J. Copula-Based Models for Financial Time Series. In *Handbook of Financial Time Series*; Andersen, T.G., Davis, R.A., Kreiß, J.P., Mikosch, T.V., Eds.; Mathematics and Statistics; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009; pp. 767–786.
28. Patton, A.J. A review of copula models for economic time series. *J. Multivar. Anal.* **2012**, *110*, 4–18. [CrossRef]
29. Patton, A.J. Copula Methods for Forecasting Multivariate Time Series. In *Handbook of Economic Forecasting*; Elliott, G., Timmermann, A., Eds.; Elsevier/North Holland: Amsterdam, The Netherlands, 2013; Volume 2, Chapter 16, pp. 899–960.
30. Giacomini, E.; Härdle, W.; Spokoiny, V. Inhomogeneous Dependence Modeling with Time-Varying Copulae. *J. Bus. Econ. Stat.* **2009**, *27*, 224–234. [CrossRef]
31. Manner, H.; Reznikova, O. A Survey on Time-Varying Copulas: Specification, Simulations, and Application. *Econom. Rev.* **2012**, *31*, 654–687. [CrossRef]
32. Patton, A.J. Modelling asymmetric exchange rate dependence. *Int. Econ. Rev.* **2006**, *47*, 527–556. [CrossRef]
33. Patton, A.J. On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. *J. Financ. Econom.* **2004**, *2*, 130–168. [CrossRef]
34. Embrechts, P.; Höing, A.; Juri, A. Using copulae to bound the Value-at-Risk for functions of dependent risks. *Financ. Stochastics* **2003**, *7*, 145–167. [CrossRef]
35. Cherubini, U.; Luciano, E. Bivariate option pricing with copulas. *Appl. Math. Financ.* **2002**, *9*, 69–85. [CrossRef]
36. Van den Goorbergh, R.W.; Genest, C.; Werker, B.J. Bivariate option pricing using dynamic copula models. *Insur. Math. Econ.* **2005**, *37*, 101–114. [CrossRef]

37. Chiou, S.C.; Tsay, R.S. A copula-based approach to option pricing and risk assessment. *J. Data Sci.* **2008**, *6*, 273–301.
38. Morillas, P.M. A method to obtain new copulas from a given one. *Metrika* **2005**, *61*, 169–184. [CrossRef]
39. Chen, X.; Fan, Y. Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. *J. Econom.* **2006**, *135*, 125–154. [CrossRef]
40. García, C.; Herrerías, J.M.; Trinidad, J.E. Making Copulas Under Uncertainty. In *Distribution Models Theory*; Herrerías, R., Callejón, J., Herrerías, J.M., Eds.; World Scientific Publishing: Singapore, 2006; Chapter 2, pp. 27–53.
41. Genest, C.; Remillard, B.; Beaudoin, D. Goodness-of-fit tests for copulas: A review and a power study. *Insur. Math. Econ.* **2009**, *44*, 199–213. [CrossRef]
42. Joe, H.; Kurowicka, D. *Dependence Modeling: Vine Copula Handbook*; World Scientific Publishing: Singapore, 2011.
43. Kiatmanaroch, T.; Sriboonchitta, S. Relationship between Exchange Rates, Palm Oil Prices, and Crude Oil Prices: A Vine Copula Based GARCH Approach. In *Modeling Dependence in Econometrics: Selected Papers of the 7th International Conference of the Thailand Econometric Society, Faculty of Economics, Chiang Mai University, Thailand, 8–10 January 2014*; Huynh, V.N., Kreinovich, V., Sriboonchitta, S., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Switzerland, 2014; Volume 251, pp. 399–413.
44. Kreinovich, V.; Nguyen, H.T.; Sriboonchitta, S.; Kosheleva, O. Why Copulas Have Been Successful in Many Practical Applications: A Theoretical Explanation Based on Computational Efficiency. In *Integrated Uncertainty in Knowledge Modelling and Decision Making: Proceedings of the 4th International Symposium, IUKM 2015, Nha Trang, Vietnam, 15–17 October 2015*; Huynh, V.N., Inuiguchi, M., Denoeux, T., Eds.; Lecture Notes in Artificial Intelligence; Springer International Publishing: Cham, Switzerland, 2015; Volume 9376, pp. 112–126.
45. Embrechts, P.; Mcneil, A.J.; Straumann, D. Correlation and dependence in risk management. In Proceedings of the ASTIN Colloquium, Tokyo, Japan, 22–25 August 1999; Cambridge University Press: Cambridge, UK, 1999; pp. 176–223.
46. McNeil, A.J.; Frey, R.; Embrechts, P. *Quantitative Risk Management: Concepts, Techniques, and Tools: Concepts, Techniques, and Tools*; Princeton Series in Finance; Princeton University Press: Princeton, NJ, USA, 2005.
47. Schweizer, B.; Sklar, A. *Probabilistic Metric Spaces*; North Holland Series in Probability and Applied Mathematics; North Holland: Amsterdam, The Netherlands, 1983.
48. Genest, C.; Nešlehová, J. A Primer on Copulas for Count Data. *Astin Bull.* **2007**, *37*, 475–515. [CrossRef]
49. Mayor, G.; Suñer, J.; Torrens, J. Sklar's Theorem in Finite Settings. *IEEE Trans. Fuzzy Syst.* **2007**, *15*, 410–416. [CrossRef]
50. Frees, E.; Valdez, E. Understanding Relationships Using Copulas. *N. Am. Actuar. J.* **1998**, 2, 1–25. [CrossRef]
51. Bouyé, E.; Durrleman, V.; Nikeghbali, A.; Riboulet, G.; Roncalli, T. *Copulas for Finance—A Reading Guide and Some Applications*; Technical Report; Crédit Lyonnais, Groupe de Recherche Opérationnelle: Paris, France, 2000.
52. Mikosch, T. Copulas: Tales and facts. *Extremes* **2006**, *9*, 3–20. [CrossRef]
53. Shiller, R.J. *Irrational Exuberance*; New York Times Bestseller, Broadway Books: New York, NY, USA, 2001.
54. Donnelly, C.; Embrechts, P. The Devil is in the Tails: Actuarial Mathematics and the Subprime Mortgage Crisis. *Astin Bull. J. Int. Actuar. Assoc.* **2010**, *40*, 1–33. [CrossRef]
55. Gumbel, E.J. Bivariate Exponential Distributions. *J. Am. Stat. Assoc.* **1960**, *55*, 698–707. [CrossRef]
56. Embrechts, P.; Mcneil, A.J.; Straumann, D. Correlation and dependence in risk management: Properties and pitfalls. In *RISK Management: Value at Risk and Beyond*; Cambridge University Press: Cambridge, UK, 2002; pp. 176–223.
57. Schachermayer, W. Mathematics and Finance. In Proceedings of the 7th European Congress of Mathematics (7ECM), Berlin, Germany, 18–22 July 2016. Technical report, The German Mathematical Society (DMV), the International Association of Applied Mathematics and Mechanics (GAMM), the Research Center Matheon, the Einstein Center ECMath and the Berlin Mathematical School (BMS).
58. Daníelsson, J. The emperor has no clothes: Limits to risk modelling. *J. Bank. Financ.* **2002**, *26*, 1273–1296. [CrossRef]
59. Zimmer, D.M. The Role of Copulas in the Housing Crisis. *Rev. Econ. Stat.* **2012**, *94*, 607–620. [CrossRef]

60. Salmon, F. The formula that killed Wall Street. *Significance* **2012**, *9*, 16–20. [CrossRef]
61. Rogers, L.C.G. Document in response to questions posed by Lord Drayson, UK Science and Innovation Minister. *Financ. Math. Credit. Crisis* **2009**.
62. Osborne, M.F.M. Brownian Motion in the Stock Market. *Oper. Res.* **1959**, *7*, 145–173. [CrossRef]
63. Samuelson, P.A. Rational Theory of Warrant Pricing. *Ind. Manag. Rev.* **1965**, *6*, 13–39.
64. Granger, C.W.J.; Morgenstern, O. Spectral analysis of New York stock market prices. *Kyklos* **1963**, *16*, 1–27. [CrossRef]
65. Godfrey, M.D.; Granger, C.W.J.; Morgenstern, O. The Random Walk Hypothesis of Stock Market Behavior. *Kyklos* **1964**, *17*, 1–30. [CrossRef]
66. Morgenstern, O. Demand Theory Reconsidered. *Q. J. Econ.* **1948**, *62*, 165–201. [CrossRef]
67. Granger, C.W.J. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica* **1969**, *37*, 424–438. [CrossRef]
68. Ying, C.C. Stock Market Prices and Volumes of Sales. *Econometrica* **1966**, *34*, 676–685. [CrossRef]
69. Clark, P.K. A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices. *Econometrica* **1973**, *41*, 135–155. [CrossRef]
70. Epps, T.W. Security Price Changes and Transaction Volumes: Theory and Evidence. *Am. Econ. Rev.* **1975**, *65*, 586–597.
71. Epps, T.W.; Epps, M.L. The Stochastic Dependence of Security Price Changes and Transaction Volumes: Implications for the Mixture-of-Distributions Hypothesis. *Econometrica* **1976**, *44*, 305–321. [CrossRef]
72. Tauchen, G.E.; Pitts, M. The Price Variability-Volume Relationship on Speculative Markets. *Econometrica* **1983**, *51*, 485–505. [CrossRef]
73. Andersen, T.G. Return volatility and trading volume: An information flow interpretation of stochastic volatility. *J. Financ.* **1996**, *51*, 169–204. [CrossRef]
74. García, J. Volumen y volatilidad en mercados financieros: El caso del mercado de futuros español. *Rev. Espa NOla Financ. Contab.* **1998**, *XXVII*, 367–393.
75. Lamoureux, C.G.; Lastrapes, W.D. Heteroskedasticity in Stock Return Data: Volume versus GARCH Effects. *J. Financ.* **1990**, *45*, 221–229. [CrossRef]
76. Lamoureux, C.G.; Lastrapes, W.D. Endogenous Trading Volume and Momentum in Stock-Return Volatility. *J. Bus. Econ. Stat.* **1994**, *12*, 253–260.
77. Bollerslev, T. Generalized Autoregressive Conditional Heteroskedasticity. *J. Econom.* **1986**, *31*, 307–327. [CrossRef]
78. Copeland, T.E. A Model of Asset Trading under the Assumption of Sequential Information Arrival. *J. Financ.* **1976**, *31*, 1149–1168. [CrossRef]
79. Copeland, T.E. A Probability Model of Asset Trading. *J. Financ. Quant. Anal.* **1977**, *12*, 563–578. [CrossRef]
80. Darrat, A.F.; Zhong, M.; Cheng, L.T. Intraday volume and volatility relations with and without public news. *J. Bank. Financ.* **2007**, *31*, 2711–2729. [CrossRef]
81. Chen, Z.; Daigler, R.T. An Examination of the Complementary Volume-volatility Information Theories. *J. Futur. Mark.* **2008**, *28*, 963–992. [CrossRef]
82. DeLong, J.B.; Shleifer, A.; Summers, L.H.; Waldmann, R.J. Positive Feedback Investment Strategies and Destabilizing Rational Speculation. *J. Financ.* **1990**, *45*, 379–395. [CrossRef]
83. Black, F. Noise. Papers and Proceedings of the Forty-Fourth Annual Meeting of the America Finance Association. *J. Financ.* **1986**, *41*, 529–543.
84. Harris, M.; Raviv, A. Differences of Opinion Make a Horse Race. *Rev. Financ. Stud.* **1993**, *6*, 473–506. [CrossRef]
85. Shalen, C.T. Volume, Volatility, and the Dispersion of Beliefs. *Rev. Financ. Stud.* **1993**, *6*, 405–434. [CrossRef]
86. Gurgul, H.; Syrek, R. Archimedean copulas for price-volume dependencies of DAX companies. *Syst. Sci.* **2006**, *32*, 63–90.
87. Gurgul, H.; Mestel, R.; Syrek, R. Polish stock market and some foreign markets—Dependence analysis by copulas. *Oper. Res. Decis.* **2008**, *2*, 17–35.
88. Gurgul, H.; Mester, R.; Syrek, R. The testing of causal Stock returns-trading Volume Dependencies with the Aid of copulas. *Manag. Econ.* **2013**, *13*, 21–44. [CrossRef]
89. Bouezmarni, T.; Rombouts, J.V.; Taamouti, A. Nonparametric Copula-Based Test for Conditional Independence with Applications to Granger Causality. *J. Bus. Econ. Stat.* **2012**, *30*, 275–287. [CrossRef]

90. Gurgul, P.; Syrek, R. Testing of Dependencies between Stock Returns and Trading Volume by High Frequency Data. *Manag. Glob. Transit.* **2013**, *11*, 353–373.
91. Gurgul, H.; Syrek, R. The structure of contemporaneous price-volume relationships in financial markets. *Manag. Econ.* **2013**, *14*, 39–60. [CrossRef]
92. Rossi, E.; de Magistris, P.S. Long memory and tail dependence in trading volume and volatility. *J. Empir. Financ.* **2013**, *22*, 94–112. [CrossRef]
93. Gurgul, H.; Syrek, R.; Mitterer, C. Price duration versus trading volume in high-frequency data for selected DAX companies. *Manag. Econ.* **2016**, *17*, 241–260. [CrossRef]
94. Joe, H.; Xu, J.J. *The Estimation Method of Inference Functions for Margins for Multivariate Models*; Technical Report #166; University of British Columbia, Department of Statistics: Vancouver, BC, Canada, 1996.
95. Ning, C.; Wirjanto, T.S. Extreme Return-Volume Dependence in East-Asian Stock Markets: A Copula Approach. *Financ. Res. Lett.* **2009**, *6*, 202–209. [CrossRef]
96. Gallant, A.R.; Rossi, P.E.; Tauchen, G.E. Stock Prices and Volume. *Rev. Financ. Stud.* **1992**, *5*, 199–242. [CrossRef]
97. Naeem, M.; Hao, J.; Brunero, L. Negative return-volume relationship in Asian stock markets: FIGARCH-Copula Approach. *Eurasian J. Econ. Financ.* **2014**, *2*, 1–20. [CrossRef]
98. Larkin, M.; Shaw, R.; Flowers, P. Multiperspectival designs and processes in interpretative phenomenological analysis research. *Qual. Res. Psychol.* **2019**, *16*, 182–198. [CrossRef]
99. McKee, T.B.; Doesken, N.J.; Kliest, J. The relationship of drought frequency and duration to time scales. In Proceedings of the Eighth Conference on Applied Climatology, Anaheim, CA, USA, 17–22 January 1993; pp. 179–184.
100. Khoudraji, A. Contributions à L'etude des Copules et à la Modélisation de Valeurs Extrêmes Bivariées. Ph.D. Thesis, Université de Laval, Laval, QC, Canada, 1995.
101. Frey, R.; McNeil, A.J.; Nyfeler, M. *Modelling Dependent Defaults: Asset Correlations Are Not Enough*! Technical Report, Working Papers; ETH Zürich, Department of Mathematics: Zürich, Switzerland, 2001.
102. Li, C.; Singh, V.P.; Mishra, A.K. A bivariate mixed distribution with a heavy-tailed component and its application to single-site daily rainfall simulation. *Water Resour. Res.* **2013**, *49*, 767–789. [CrossRef]
103. Clayton, D.G. A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biometrika* **1978**, *65*, 141–151. [CrossRef]
104. Nelsen, R.B. Properties and applications of copulas: A brief survey. In Proceedings of the First Brazilian Conference on Statistical Modeling in Insurance and Finance, Maresias, Brazil, 25–30 March 2003; Dhaene, J., Kolev, N., Morettin, P.A., Eds.; University of Sao Paulo: Sao Paulo, Brazil, 2003; pp. 10–28.
105. Ali, M.M.; Mikhail, N.N.; Haq, M.S. A class of bivariate distributions including the bivariate logistic. *J. Multivar. Anal.* **1978**, *8*, 405–412. [CrossRef]
106. Barnett, V. Some bivariate uniform distributions. *Commun. Stat. Theory Methods* **1980**, *9*, 453–461. [CrossRef]
107. Plackett, R.L. A Class of Bivariate Distributions. *J. Am. Stat. Assoc.* **1965**, *60*, 516–522. [CrossRef]
108. Cuadras, C.M.; Augé, J. A Continuous General Multivariate Distribution and its Properties. *Commun. Stat. Theory Methods* **1981**, *10*, 339–353. [CrossRef]
109. Shih, J.H.; Louis, T.A. Inferences on the Association Parameter in Copula Models for Bivariate Survival Data. *Biometrics* **1995**, *51*, 1384–1399. [CrossRef]
110. Joe, H. *Dependence Modeling with Copulas*; CRC Press: Boca Raton, FL, USA, 2014.
111. Durrleman, V.; Nikeghbali, A.; Roncalli, T. *A Note about the Conjecture about Spearman's rho and Kendall's Tau*; Technical Report; Crédit Lyonnais, Groupe de Recherche Opérationnelle: Paris, France, 2000.
112. Fischer, M.J.; Hinzmann, G. *A New Class of Copulas With Tail Dependence and a Generalized Tail Dependence Estimator*; Discussion Papers; Friedrich-Alexander University Erlangen-Nuremberg: Nuremberg, Germany, 2006.
113. Roch, O.; Alegre, A. Testing the bivariate distribution of daily equity returns using copulas. An application to the Spanish stock market. *Comput. Stat. Data Anal.* **2006**, *51*, 1312–1329. [CrossRef]
114. Johannes, M.; Polson, N. MCMC Methods for Continuous-Time Financial Econometrics. In *Handbook of Financial Econometrics: Applications*; Aït-Sahalia, Y., Hansen, L.P., Eds.; Handbooks in Finance; Elsevier: Amsterdam, The Netherlands, 2010; Volume 2, Chapter 13, pp. 1–72.
115. AghaKouchak, A. Water Resources Research. *J. Hydrometeorol.* **2014**, *15*, 1944–1973.

116. Bremer, M.; Kato, H.K. Trading Volume for Winners and Losers on the Tokyo Stock Exchange. *J. Financ. Quant. Anal.* **1996**, *31*, 127–142. [CrossRef]
117. Huang, R.D.; Masulis, R. Trading activity and stock price volatility: Evidence from the London Stock Exchange. *J. Empir. Financ.* **2003**, *10*, 249–269. [CrossRef]
118. Quiroga García, R.; Sánchez Álvarez, I. Intraday volatility and information arrival in the IBEX-35 futures markets. *Span. J. Financ. Account. Rev. EspañOla Financ. Contab.* **2006**, *35*, 523–540.
119. Zárraga Alonso, A. Análisis de causalidad entre rendimiento y volumen. *Investig. EconÓmicas* **1998**, *XXII*, 45–67.
120. Durrleman, V.; Nikeghbali, A.; Roncalli, T. *Which Copula Is the Right One*? Technical Report; Crédit Lyonnais, Groupe de Recherche Opérationnelle: Paris, France, 2000.
121. Rebonato, R.; Denev, A. Choosing and fitting the copula. In *Portfolio Management under Stress: A Bayesian-Net Approach to Coherent Asset Allocation*; Cambridge University Press: Cambridge, UK, 2014; Chapter 19, pp. 278–290.
122. Garman, M.B. Market microstructure. *J. Financ. Econ.* **1976**, *3*, 257–275. [CrossRef]
123. Gomes, F. Portfolio Choice and Trading Volume with Loss-Averse Investors. *J. Bus.* **2005**, *78*, 675–706. [CrossRef]
124. Kahneman, D.; Tversky, A. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* **1979**, *47*, 263–291. [CrossRef]
125. Rachev, S.T.; Hsu, J.S.J.; Bagasheva, B.S.; Fabozzi, F.J. *Bayesian Methods in Finance*; The Frank J. Fabozzi Series; John Wiley & Sons: Hoboken, FL, USA, 2008; Volume 163.
126. Jacquier, E.; Polson, N. Bayesian Methods in Finance. In *The Oxford Handbook of Bayesian Econometrics*; Geweke, J., Koop, G., van Dijk, H., Eds.; Oxford Handbooks in Economics, Oxford University Press: Oxford, UK, 2011.
127. Rebonato, R. *Coherent Stress Testing: A Bayesian Approach to the Analysis of Financial Stress*; The Wiley Finance Series; John Wiley & Sons: Chichester, UK, 2010.
128. Ardia, D. *Financial Risk Management with Bayesian Estimation of GARCH Models: Theory and Applications*; Lecture Notes in Economics and Mathematical Systems; Springer: Berlin/Heidelberg, Germany, 2008; Volume 612.
129. Sekerke, M. *Bayesian Risk Management: A Guide to Model Risk and Sequential Learning in Financial Markets*; Wiley Finance, John Wiley & Sons: Hoboken, NJ, USA, 2015.
130. Shemyakin, A.; Kniazev, A. *Introduction to Bayesian Estimation and Copula Models of Dependence*; John Wiley & Sons: Hoboken, NJ, USA, 2017.
131. Barnard, G.A.; Jenkins, G.M.; Winsten, C.B. Likelihood Inference and Time Series. *J. R. Stat. Soc. Ser.* **1962**, *125*, 321–372. [CrossRef]
132. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
133. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*; Parzen, E., Tanabe, K., Kitagawa, G., Eds.; Springer Series in Statistics; Springer Science+Business Media: New York, NY, USA, 1998; Chapter 4, pp. 199–213.
134. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 416–464. [CrossRef]
135. Sempi, C. An introduction to Copulas. Technical report. In Proceedings of the 33rd Finnish Summer School on Probability Theory and Statistics, Tampere, Finland, 6–10 June 2011.
136. Fréchet, M.R. *Sur les Tableaux de Corrélation Dont les Marges Sont Données*; Annales de l'Université de Lyon, Sciences: Lyon, France, 1951; pp. 13–84.
137. Hoeffding, W. Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin (Reprinted in english as "Scale-invariant correlation theory" in The Collected Works of Wassily Hoeffding, pp. 57–108, 1994). In *Maszstabinvariante Korrelationstheorie*; Fisher, N.I., Sen, P.K., Eds.; Springer Series in Statistics. Perspectives in Statistics; Springer: Berlin/Heidelberg, Germany, 1940; Volume 5, pp. 179–233.

138. Galambos, J. *The Asymptotic Theory of Extreme Order Statistics*; John Wiley & Sons: New York, NY, USA, 1978.
139. Molanes, E.M.; Romera, R. *Copulas in Finance and Insurance*; Technical Report ws086321, Universidad Carlos III. UC3M Working Papers, Statistics and Econometrics 08-21; Universidad Carlos III de Madrid, Departamento de Estadística: Madrid, Spain, 2008.

mathematics

MDPI

Article

On the Relationship of Cryptocurrency Price with US Stock and Gold Price Using Copula Models

Jong-Min Kim [1], Seong-Tae Kim [2] and Sangjin Kim [3,*]

[1] Statistics Discipline, University of Minnesota at Morris, Morris, MN 56267, USA; jongmink@morris.umn.edu
[2] Department of Mathematics, North Carolina A&T State University, Greensboro, NC 27411, USA; skim@ncat.edu
[3] Department of Management and Information Systems, Dong-A University, Busan 49236, Korea
* Correspondence: skim10@dau.ac.kr

Received: 15 September 2020; Accepted: 20 October 2020; Published: 23 October 2020

Abstract: This paper examines the relationship of the leading financial assets, Bitcoin, Gold, and S&P 500 with GARCH-Dynamic Conditional Correlation (DCC), Nonlinear Asymmetric GARCH DCC (NA-DCC), Gaussian copula-based GARCH-DCC (GC-DCC), and Gaussian copula-based Nonlinear Asymmetric-DCC (GCNA-DCC). Under the high volatility financial situation such as the COVID-19 pandemic occurrence, there exist a computation difficulty to use the traditional DCC method to the selected cryptocurrencies. To solve this limitation, GC-DCC and GCNA-DCC are applied to investigate the time-varying relationship among Bitcoin, Gold, and S&P 500. In terms of log-likelihood, we show that GC-DCC and GCNA-DCC are better models than DCC and NA-DCC to show relationship of Bitcoin with Gold and S&P 500. We also consider the relationships among time-varying conditional correlation with Bitcoin volatility, and S&P 500 volatility by a Gaussian Copula Marginal Regression (GCMR) model. The empirical findings show that S&P 500 and Gold price are statistically significant to Bitcoin in terms of log-return and volatility.

Keywords: cryptocurrency; gold; S&P 500; GARCH; DCC; copula

1. Introduction

Knowing the relationships of the cryptocurrency market with either the US stock market or commodity market will be very useful to manage investors' portfolios and how many portions of their investment money will be allocated to cryptocurrency for their secure and profitable investment plan. Cryptocurrency is a digital or virtual currency that is exchanged between peers without the need for a third party [1]. The key features of the cryptocurrency include that there is no central system to manage the transactions of cryptocurrencies, and they are classified as a commodity by the U.S. Commodity Futures Trading Commission (CFTC). The first cryptocurrency, Bitcoin, operates with block-chain technology, in which a secure system of accounting is used that transfers ownership. The cryptocurrency market is an attractive emerging market for investment, but this market revealed downfalls such as cryptocurrency hacking news. For example, in May 2019, hackers stole $40 million worth of Bitcoin from Binance, one of the largest cryptocurrency exchanges in the world. Therefore, investors themselves have to take a high risk from cryptocurrency investment. However, the recent cryptocurrency market is a bull market where the Bitcoin price is equal to the USD 10,806.90 as of 30 September 2020, but the Bitcoin price has severely fluctuated since the maximum Bitcoin price at the USD 19,783.06 on 17 December 2017. Despite a series of negative events in this market, investing cryptocurrency is gaining popularity among investors to make their own money. Consequently, economic entities are interested in the dynamic relationships among the cryptocurrency market, commodity market, and stock market.

There have been many studies on the analysis of the exchange rates of cryptocurrency [2]. Recently, Hyun et al. [3] examined dependence relationships among the five well-known cryptocurrencies (Bitcoin, Ethereum, Litecoin, Ripple, and Stella) using a copula directional dependence. Kim et al. [4] studied the volatility of nine well-known cryptocurrencies—Bitcoin, XRP, Ethereum, Bitcoin Cash, Stella, Litecoin, TRON, Cardano, and IOTA using several GARCH models and Bayesian Stochastic Volatility (SV) models. Klein et al. [5] employed the BEKK [6] GARCH model to estimate time-varying conditional correlations between gold and Bitcoin. In terms of portfolio management, Aslanidisa, Barivierab and Martínez-Ibañeza [7] considered Dynamic Conditional Correlation (DCC) with daily price data (21 May 2014, to 27 September 2018), pairs of four cryptocurrencies (Bitcoin, Dash, Monero, and Ripple), and three traditional financial assets (Standard & Poors 500 Composite (SP500), S&P US Treasury bond 7-10Y index (BOND), and Gold Bullion LBM) [8–11]. Guesmi et al. [12] examined the dynamics of Bitcoin and other financial assets using the VARMA (1, 1)-DCC-GJR-GARCH model and found that Bitcoin provides diversification and hedging opportunities for investment. Hyun et al. [3] already applied the copula approach to cryptocurrency because no assumption is needed such as normality, linearity, and independence of the errors from the proposed model.

In this study, we aim to apply the copula-based GARCH-DCC models [3,13,14] to see the recent time varying correlations between the cryptocurrency market and US stock price or between the cryptocurrency market and commodity market price after the slump of the cryptocurrency market price since 2018. The copula-based GARCH-DCC models are compared to the GARCH-DCC models in the empirical data analysis [8,15–17] which shows that copula-based GARCH-DCC models has better model than GARCH-DCC models. A copula is a multivariate distribution function described on the unit $[0,1]^n$ with uniformly distributed marginal [18]. Our result also leaded to the same conclusion as the previous researches. Furthermore, because of the failure of the ordinary least regression to capture the heteroscedasticity with high volatility financial data, we use the Gaussian Copula Marginal Regression (GCMR) models [19] which can consider the heteroscedasticity and non-normality of the financial data to test our alternative hypothesis that Bitcoin is statistically significant by log-returns of S&P 500 and Gold price in terms of log-return. We also test the current volatility of log-returns of Bitcoin can be statistically significant with the current and lagged volatilities of the other assets (S&P 500 and Gold price). We also test that the time varying correlations of log-returns of Bitcoin and S&P 500 can be statistically significant with the current volatilities of the Bitcoin and S&P 500.

The paper is organized as follows. Section 2 reviews econometric methodologies that will be used in this paper. Section 3 describes data and discusses empirical data analysis. Section 4 provides the conclusion and our related future study.

2. Econometrical Methods

This section introduces the volatility model, dynamic correlation coefficient, copula, and their combinations. The description of econometric models is not comprehensive but selective to understand the dynamic relationships among the three markets.

2.1. GARCH Models

Let S_t be a price time series at time t. For a log return series $r_t = log\left(\frac{S_t}{S_{t-1}}\right)$, we let $a_t = r_t - E_{t-1}[r_t]$ be the innovation at time t. Then a_t follows a GARCH (p, q) model if $a_t = h_t\epsilon_t$

$$h_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i a_{t-i}^2 + \sum_{j=1}^{p} \beta_j h_{t-j}^2 \tag{1}$$

where $\{\epsilon_t\}$ is a sequence of independent and identically distributed random variables with mean 0 and variance 1, $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, and $\sum_{i=1}^{\max(p,q)} \left(\alpha_i + \beta_j\right) \leq 1$. All members of the family of GARCH models can be obtained from a transformation of the conditional standard deviation, h_t, determined by the transformation of the innovations, a_t, and lagged transformed conditional standard

deviations. An extensive discussion on the nested GARCH models is given in Hentschel [20]. Since the conditional variance in the GARCH model did not properly respond to positive and negative shocks, Engel and Ng [21] also proposed one of the popular nonlinear asymmetric GARCH (NAGARCH) models as follows:

$$h_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i (a_{t-i} - \gamma_i h_{t-i})^2 + \sum_{j=1}^{q} \beta_j h_{t-j}^2, \tag{2}$$

where $\alpha_0 > 0,\ \alpha_i \geq 0,\ \beta_j \geq 0$ for $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, q$. In the model, the distance $\gamma_i h_{t-i}$ moves the news impact curve to the right, and the parameter γ_i of stock returns is estimated to be positive. It indicates that negative returns increase future volatility with larger amounts than positive returns of the same magnitude.

The T-GARCH model, which can capture the asymmetric effect in the volatility is given by

$$h_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i (|a_{t-i}| - \eta_i a_{t-i})^2 + \sum_{j=1}^{p} \beta_j h_{t-j}^2, \tag{3}$$

where the asymmetric parameter η satisfies the condition $-1 < \eta < 1$. For the model selection of the GARCH (1, 1) models considered, we use the Akaike Information Criterion (AIC). Besides, this study also considers the Student-t errors to take into account the possible fatness of the distribution tails of a_t.

2.2. DCC and Copula DCC Models

To investigate the time-varying correlations among multivariate returns, we adopt the DCC model, which incorporates the flexibility of univariate GARCH models and the harmonicity of correlation estimation functions. In the DCC model in [6,22], the correlation matrix is time-varying, and the covariance matrix can be decomposed into:

$$H_t = D_t R D_t = \rho_{ij} \sqrt{h_{ij,t} h_{ij,t}}, \ \text{ where } D_t = diag\left(\sqrt{h_{11,t}}, \ldots, \sqrt{h_{nn,t}} \right) \tag{4}$$

containing the time-varying standard deviations is obtained from GARCH models, and R is the constant conditional correlation (CCC) proposed by Bollerslev [23], which is defined as $R = T^{-1} \sum_{i=t}^{T} v_t v_t'$, where $v_t = \frac{r_t - \mu}{\sigma_t}$, and μ is a vector of expected returns. The DDC in [24] is a time-varying extension of the CCC, which has the following structure:

$$R_t = diag(Q_t)^{-\frac{1}{2}} Q_t diag(Q_t)^{-\frac{1}{2}}, \tag{5}$$

where $Q_t = R + \alpha\left(v_{t-1} v_{t-1}' - R\right) + \beta(Q_{t-1} - R)$.

Note that to ensure stationarity, nonnegative α and β satisfy the constraint $\alpha + \beta < 1$, and Q_t is positive definite which makes R_t positive definite. Off-diagonal elements in the covariance matrix Q_t are the correlation coefficients between pairwise indexes among Bitcoin, Gold, and S&P 500 at time t. In this paper, we use the "*dcc.estimation*" function in the "*ccgarch*" on R package [24,25] to estimate each conditional correlation.

We consider another statistical approach to address the correlation among multivariate time series. Sklar [26] suggested copular functions to build joint multivariate distributions. The copula models we consider here are Gaussian copulas which are used to estimate the time-varying correlation matrix of the DCC model. A copula is an efficient way to characterize and model correlated multivariate random variables. Therefore, we consider the time-varying conditional correlation in the copula

framework. Let a random vector $\left(X_1, \ldots, X_p\right)$ have marginal distribution functions $F_i(x_i) = P(X_i \leq x_i)$ for $i = 1, \ldots, p$. The dependence function, C, for all $u_1, \ldots, u_n \in [0,1]^n$ can be defined as:

$$\begin{array}{c} C(u_1, \ldots, u_n) = P(F_1(X_1) \leq u_1, \ldots, F_n(X_n) \leq u_n) \\ C(u_1, \ldots, u_n) = F\left(F_1^{-1}(u_1), \ldots, F_1^{-1}(u_n)\right). \end{array} \tag{6}$$

In this study, we estimate DCC($\hat{\rho}_t$) by a Gaussian copula function whose conditional density is defined as:

$$c_t(u_{1t}, \ldots, u_{nt}|R_t) = \frac{f_t(F_1^{-1}(u_{1t}), \ldots, F_1^{-1}(u_{nt})\big|R_t)}{\prod_{i=1}^{n} f_i(F_1^{-1}(u_{it}))}, \tag{7}$$

where R_t is the correlation matrix implied by the covariance matrix, $u_{it} = F_{it}(r_{it}|\mu_{it}, h_{it}, \nu_t, \tau_i)$ is the probability integral transformed values estimated by the GARCH process, and $F_1^{-1}(u_{it}|\tau)$ represents the quantile transformation. We estimate each conditional correlation via the "*cgarchspec*" function in the R package "*rmgarch*" implementing the Gaussian copula [27,28]. In particular, our model applies the Gaussian copula to estimate the conditional covariance matrix. We propose four different DCC-related models: the GARCH-DCC (DCC) model, Nonlinear Asymmetric-GARCH-DCC (NA-DCC) model, Gaussian copula-based GARCH-DCC (GC-DCC) model, and Gaussian copula-based nonlinear asymmetric GARCH-DCC (GCNA-DCC) model to see the dynamic conditional correlations between Bitcoin and S&P 500 and between Bitcoin and Gold.

2.3. Gaussian Copula Marginal Regression (GCMR) Model

Gaussian Copula Marginal Regression (GCMR) is another methodology used in this study to capture the relationship, where dependence is expressed in the correlation matrix of a multivariate Gaussian distribution [19,29]. Let $F(\cdot|x_i)$ be a marginal cumulative distribution depending on a vector of covariates x_i. If a set of n dependent variables in Y_i is considered, then the joint cumulative distribution function is in the Gaussian copula regression defined by

$$\Pr(Y_1 \leq y_1, \ldots, Y_1 \leq y_1) = \Phi_n\{\varepsilon_1, \ldots, \varepsilon_n; P\}, \tag{8}$$

where $\varepsilon_i = \Phi^{-1}\left\{F(y_i|x_i)\right\}$. $\Phi(\cdot)$ and $\Phi_n(\cdot\,; P)$ indicate the univariate and multivariate standard normal cumulative distribution functions, respectively. P denotes the correlation matrix of the Gaussian copula. Masarotto and Varin [19] propose an equivalent formulation of the Gaussian copula model linking each variable Y_i to a vector of covariates x_i as follows:

$$Y_i = h(x_i, \varepsilon_i), \tag{9}$$

where ε_i indicates a stochastic error. In particular, the Gaussian copula regression model assumes that $h(x_i, \varepsilon_i) = F^{-1}\{\Phi(\varepsilon_i)|x_i\}$ and ε has a multivariate standard normal distribution with correlation matrix P. The advantages of using GCMR are to keep the marginal univariate distributions for each variable and to have multivariate normal errors for the joint distribution.

3. Empirical Analysis and Results

In this section, we apply the proposed methods to the three selected price time series. Given the sensitivity of the periods in predicting the volatility of financial time-series return data such as cryptocurrencies, we examine two different periods, more recent and short- and long-term periods. The sample consists of the daily log-returns of the nine cryptocurrencies over the period from 2 January 2018 to 21 September 2020. The log-returns of Bitcoin (BTC) and S&P 500 are denoted by LBTC and LSP, respectively. We obtained our Bitcoin data from a financial website [30], Gold data from Prof. Werner Antweiler's website [31] at the University of British Columbia Sauder School of Business, and S&P 500 data from the Yahoo finance website [32].

Figure 1 compares the pattern of prices of Bitcoin, Gold, and S&P 500 at the original scale since January 2018. The graphs appear to have a significant pairwise positive relationship after the COVID-19 pandemic occurrence. Therefore, with the log-returns of prices of Bitcoin, Gold, and S&P 500 (LBTC, LGD, LSP), we test if there is the significant pairwise correlation among LBTC, LGD, and LSP in this period using three correlations measures, the Pearson correlation method with the linear relationship assumption and Spearman and Kendall rank correlations as non-parametric methods. The data provided no statistically significant pairwise relationships among the three variables of prices as seen in Table 1. We also summarized descriptive statistics of the log return data of the cryptocurrencies such as mean, skewness, and kurtosis as well as the five-number summary statistics in Table 2. In Table 2, it is recognized that the standard deviation of LBTC is larger than those of LGD and LSP, which means that LBTC has a higher risk than LGD and LSP in terms of investment. Besides, the value of kurtosis in LBTC is greater than 3, meaning heavy tails while LGD and LSP have values less than 3, meaning light tails compared to a normal distribution. The LBTC and LSP are left-skewed while LGD is right-skewed. It means that the prices of Bitcoin and the S&P 500 will more likely be decreased soon, but the price of Gold will more likely be increased shortly.

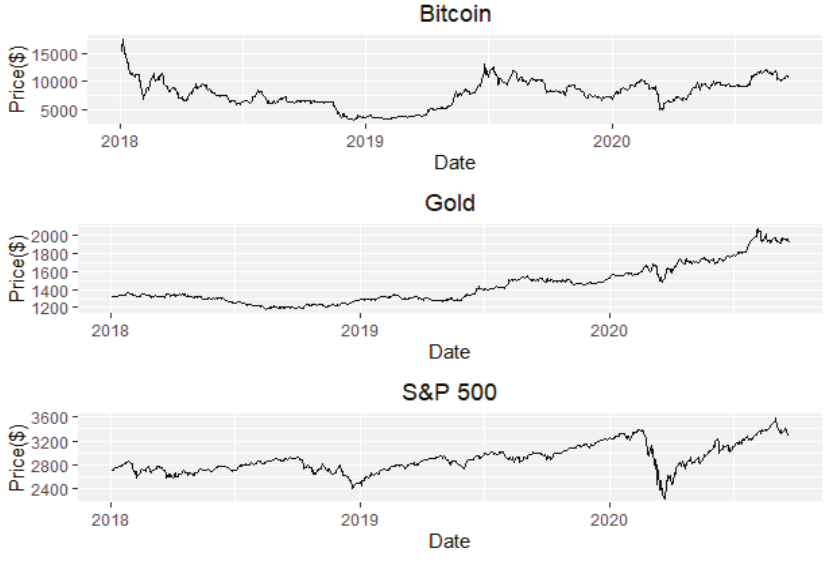


Figure 1. Prices of Bitcoin, Gold, and S&P 500.

Table 1. Correlation coefficients of log-return of Bitcoin (LBTC), log-return of Gold (LGD), and log-returns of S&P 500 (LSP) with Pearson, Spearman, and Kendall.

		LBTC	LGD	LSP
Pearson	LBTC	1	0.202	0.240
	LGD	0.202	1	0.255
	LSP	0.240	0.255	1
Spearman	LBTC	1	0.134	0.095
	LGD	0.134	1	0.084
	LSP	0.095	0.084	1
Kendall	LBTC	1	0.090	0.064
	LGD	0.090	1	0.060
	LSP	0.064	0.063	1

Table 2. Summary statistics of the log-return of Bitcoin (LBTC), log-return of Gold (LGD), and log-returns of S&P 500 (LSP).

	LBTC	**LGD**	**LSP**
Min	−0.465	−0.053	−0.128
Q1	−0.016	−0.004	−0.004
Q2	0.001	0.0004	0.001
Mean	−0.0005	0.0006	0.0003
Q3	0.019	0.005	0.007
Max	0.203	0.051	0.090
SD	0.049	0.009	0.015
Skewness	−1.579	−0.268	−1.067
Kurtosis	17.662	9.311	18.960

Since a causality between two variables may exist although there is no correlation as in Table 1, we tested if there is linear Granger causality with each lag of 1, 2, and 3 using the "*grangertest*" function in the "*lmtest*" R package [33]. That is, we consider the causality from LBTC to LSP and vice versa and from LBTC to LGD and vice versa. Table 3 shows the results of linear Granger causality tests at lag 1, 2, and 3, respectively. As seen in Table 3, there is no statistically significant causality among LBTC, LSP and LGD at the lag 1 but there is statistically significant causality among (LBTC, LSP) and (LBTC, LGD) at the lag 2 and the lag 3.

Table 3. The result of linear Granger causality with lag of 1, 2, and 3. There is no Granger causality between the log-returns of Bitcoin (LBTC) and S&P 500 (LSP) and Bitcoin (LBTC) and Gold (LGD), respectively.

	Lag 1		**Lag 2**		**Lag 3**	
Causality	F-stat	*p*-val	F-stat	*p*-val	F-stat	*p*-val
Bitcoin → S&P 500	2.656	0.104	8.153	0.000	5.520	0.001
S&P 500 → Bitcoin	0.530	0.467	0.120	0.887	0.257	0.857
Bitcoin → Gold	0.034	0.854	4.217	0.015	3.788	0.010
Gold → Bitcoin	0.868	0.352	1.882	0.153	2.413	0.066

Figure 2 shows the volatilities of log-returns of Bitcoin, Gold, and S&P 500 with the models of GARCH and NAGARCH. The GARCH volatilities are larger than those of the NAGARCH, while the pattern of volatility is similar between the two models. In each of the two plots, the level of volatilities (or risk) among the log-returns of Bitcoin, Gold, and S&P 500 is in the order of Bitcoin, S&P 500, and Gold.

To investigate the volatilities of the LBTC, LGD, LSP, we consider three different GARCH models which include two asymmetric GARCH models, T-GARCH (1, 1), and Nonlinear Asymmetric-GARCH (1, 1), and one standard-GARCH (1, 1). Table 4 reports the result of log-likelihood to choose an optimal model among the three models. The standard-GARCH (1, 1) model achieved the minimum AIC scores meaning a better fit across LBTC, LSP, and LGD.

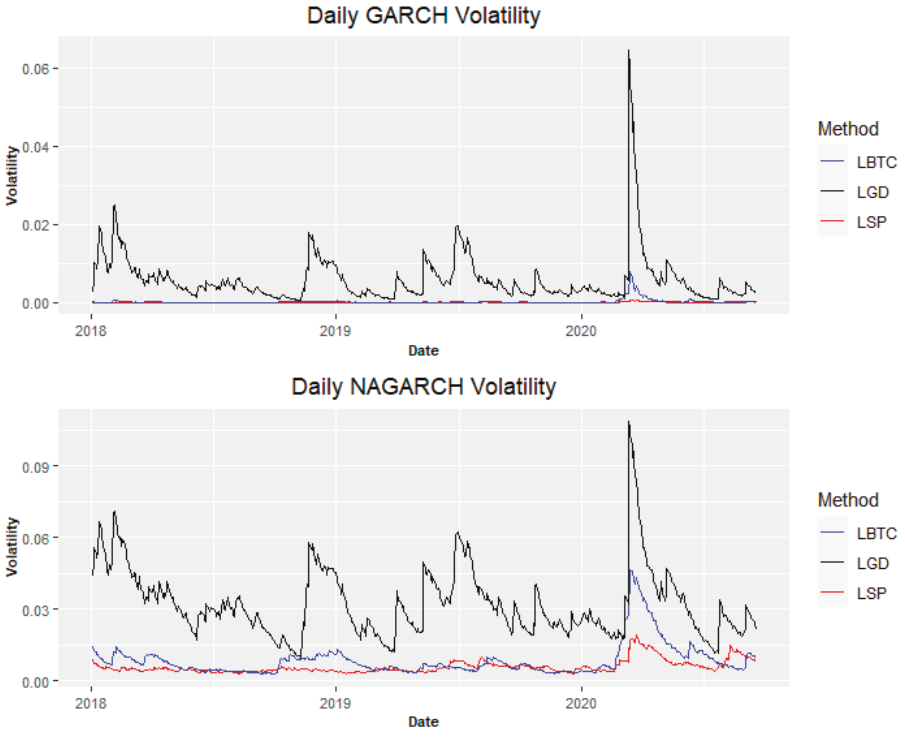


Figure 2. Daily volatility plots for LBTC, LSP, LGD with GARCH (1, 1), and Nonlinear Asymmetric-GARCH (1, 1).

Table 4. The result of Akaike Information Criterion (AIC) scores to select the best model among three different GARCH models. The Standard-GARCH (1, 1) model has maximum values of log-likelihood (LH) from the log-return of Bitcoin (LBTC), log-return of Gold (LGD), and log-returns of S&P 500 (LSP). A higher LH indicates a better fit.

	LBTC	**LSP**	**LGD**
T-GARCH (1, 1)	1568.405	1702.325	797.6006
NA-GARCH (1, 1)	2014.815	2139.722	1163.852
Standard-GARCH (1, 1)	2106.239	2246.397	1206.569

We apply a standard-GARCH (1, 1) model to LBTC, LSP, and LGD to check if there exists volatility clustering. Table 5 shows the results of the model fits based on the standard-GARCH (1, 1) model. The coefficient β_1 is the effect of the conditional variance at time t-1 on the conditional variance at time t, so a high value close to one indicates a longer persistency of the volatility shock. Hence, the estimates of $\beta_1^{'}s$ in the table explain the amount of volatility clustering. Likewise, there exist consistent volatility clusterings throughout all models since all p values of β_1s are closed to 0 at $\alpha = 0.05$.

Table 5. The results of the standard-GARCH (1, 1) model with the log-return of Bitcoin (LBTC), log-return of Gold (LGD), and log-returns of S&P 500 (LSP) where α_0, α_1, and β_1 are from Equation (1).

Standard-GARCH Model Fit with LBTC				
	Estimate	**S.E**	***t*-Value**	***p*-Value**
α_0	0.000	0.000	2.700	0.007
α_1	0.244	0.051	4.794	0.000
β_1	0.755	0.040	18.707	0.000
t-distribution parameter	5.547	1.167	4.752	0.000
Standard-GARCH Model Fit with LSP				
α_0	0.000	0.000	2.026	**0.043**
α_1	0.097	0.029	3.319	**0.000**
β_1	0.867	0.039	22.494	**0.000**
t-distribution parameter	6.260	1.522	4.111	**0.000**
Standard-GARCH Model Fit with LGD				
α_0	0.000	0.000	0.964	0.335
α_1	0.277	0.148	1.560	0.084
β_1	0.878	0.033	32.097	0.000
t-distribution parameter	2.315	0.337	10.516	0.000

Note: β_1 is statistically significant in the table. It means there exists consistent volatility clustering.

Furthermore, we checked the normality of the data and determined if a good fit had been achieved based on the Ljung-Box test which is a classical hypothesis test whose null hypothesis is that the autocorrelations between the population series values are zero. Table 6 shows the results of the Jarque–Bera and Sapiro–Wilk tests for normality and the Ljung–Box and LM-ARCH conditional heteroscedasticity tests for residuals. According to the statistical tests in the table, the residuals appear to be non-normal since the *p*-values of the two normality tests are less than $\alpha = 0.05$, and they show no serial correlations in the series since the *p*-values of the Ljung–Box tests are greater than $\alpha = 0.05$.

Table 6. The residual test results of the standard-GARCH (1, 1) model. It shows that the residuals are not normal and there is volatility clustering.

Standardized Residuals (R) Tests	**Statistic**	***p*-Value**
Jarque-Bera Test on R	6760.294	0.000
Shapiro-Wilk Test on R	0.824	0.000
Ljung-Box Test on R Q(10)	10.668	0.384
Ljung-Box Test on R Q(15)	12.946	0.606
Ljung-Box Test on R Q(20)	16.334	0.696
Ljung-Box Test on R Squared Q(10)	10.223	0.421
Ljung-Box Test on R Squared Q(15)	11.416	0.723
Ljung-Box Test on R Squared Q(20)	12.765	0.887
LM-ARCH Test on R	10.217	0.597

We also consider nonlinear asymmetric GARCH to model LBTC, LSP, and LGD. Table 7 reports that there exists consistent volatility clustering since the p values of $\beta_1 s$ are all significant at $\alpha = 0.05$, which is consistent with the results in Table 5, and there is no volatility asymmetry in leverage effect in this period because all p values of $\gamma_1 s$ are not significant over each of the LBTC, LGD, LSP.

We built four different dynamic conditional correlation (DCC) models for LBTC and LSP and three different DCC models for LBTC and LGD. Figure 3 represents the DCC of four different models with DCC, NA-DCC, GC-DCC, and GCNA-DCC for log-returns of Bitcoin and S&P 500. The patterns of the four models are almost similar to each other. However, the top two graphs for DCCs without Gaussian copulas are slightly different from the bottom two graphs for DCCs with Gaussian copulas for which NA-DCC using Gaussian copulas has relatively smaller values than those of using NA-DCC alone. In

Figure 3, the highest positive DCC between LBTC and LSP was observed during the cryptocurrency crash in early 2018. In particular, we need to pay attention to that there exists a postive time-varying correlation between LBTC and LSP from March 2020 to September 2020 which is the COVID-19 pandemic period.

Table 7. Model fit of NA-GARCH (1, 1) where α_0, α_1, and β_1 are from Equation (2). Each of all β_1s has significance indicating there exists consistent volatility clustering and all γ_1s have no significance meaning there is no leverage effect (not asymmetric).

	Estimate	**S.E**	***t*-Value**	***p*-Value**
NA-GARCH Model Fit with LBTC				
α_0	0.000	0.000	0.298	0.765
α_1	0.050	0.006	8.385	0.000
β_1	0.900	0.010	89.068	0.000
γ_1	0.050	0.068	0.737	0.461
t-distribution parameter	4.000	0.215	18.605	**0.000**
NA-GARCH Model Fit with LSP				
α_0	0.000	0.000	0.073	0.942
α_1	0.050	0.006	8.266	0.000
β_1	0.900	0.012	73.973	0.000
γ_1	0.051	0.059	0.861	0.389
t-distribution parameter	4.000	0.208	19.215	0.000
NA-GARCH Model Fit with LGD				
α_0	0.000	0.000	0.923	0.356
α_1	0.050	0.009	5.753	0.000
β_1	0.900	0.019	48.154	0.000
γ_1	−0.003	0.089	−0.037	0.971
t-distribution parameter	4.000	0.268	14.921	0.000

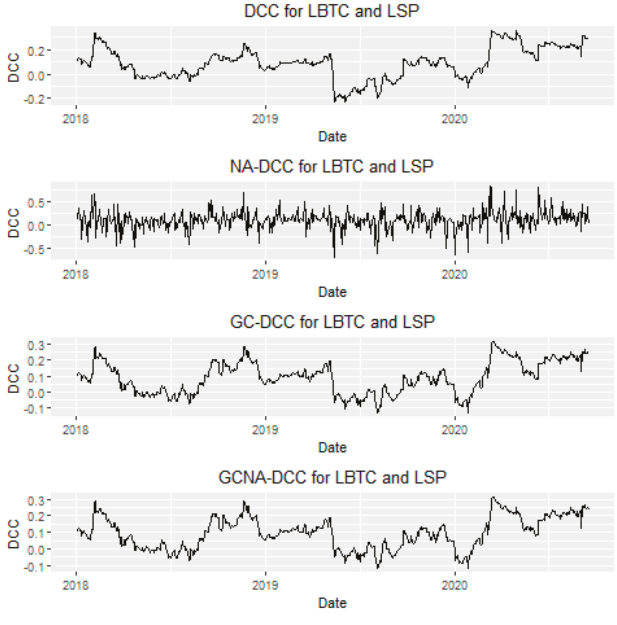


Figure 3. Dynamic conditional correlation between LBTC and LSP with GARCH-DCC (DCC),

Nonlinear Asymmetric GARCH-DCC (NA-DCC), Gaussian copula-based GARCH-DCC (GC-DCC), and Gaussian copula-based Nonlinear Asymmetric GARCH-DCC (GCNA-DCC).

Figure 4 shows the plots describing the three models of DCC, GC-DCC, and GCNA-DCC for log-returns of Bitcoin and Gold. From the patterns of GC-DCC and GCNA-DCC in Figure 4, we also found that there exists a postive time-varying correlation between LBTC and LGD from March 2020 to September 2020 which is the COVID-19 pandemic period.

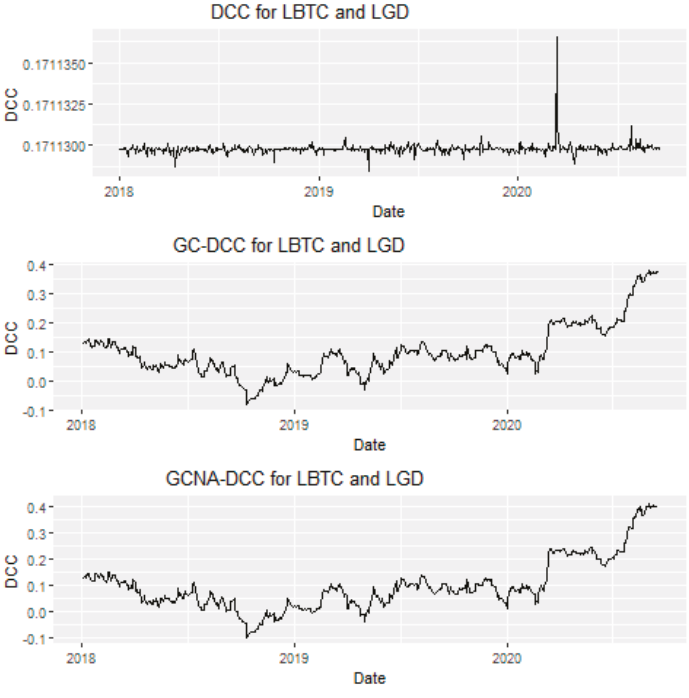


Figure 4. Dynamic conditional correlation between LBTC and LGD with GARCH-DCC (DCC), Gaussian copula-based GARCH-DCC (GC-DCC), and Gaussian copula-based Nonlinear Asymmetric GARCH-DCC (GCNA-DCC).

Log-likelihood is a measure of model fit. The higher the value, the better the fit. This is usually obtained from statistical output. For the pair of LBTC and LSP, the log-likelihood values of the DCC and NA-DCC models are smaller than the values of GC-DCC and GCNA-DCC in Table 8. Therefore, we can say that GC-DCC and GCNA-DCC are better models than DCC and NA-DCC to show relationship of Bitcoin with Gold and S&P 500 in terms of log-likelihood. In addition, there is a computation difficulty to compute NA-DCC with LBTC and LGD. Therefore, we can conclude that our proposed method is a better statistical method to look at the relationship among financial assets compared with DCC and NA-DCC. In addition, the estimates of alpha and beta for GC-DCC and GCNA-DCC are statistically significant at the 5% significance level but the estimates of alpha and beta for DCC and NA-DCC are not statistically significant at the 5% significance level. We can see that there is a computation difficulty to apply DCC and NA-DCC to high volatility financial data. The standard errors of the estimates from the DCC and NA-DCC models are much smaller than the standard errors from GC-DCC and GCNA-DCC. Especially, NA-DCC for Bitcoin and Gold cannot be computed from the *"fGarch"* R package [34] even though the log-likelihood value of DCC is larger than GC-DCC and GCNA-DCC. Based on these results, GC-DCC and GCNA-DCC are better models than DCC and NA-DCC. It is a

strong motivation to apply the Gaussian copula DCC models for cryptocurrency to US stock and Gold market prices. We also investigate the relationship of the volatilities of cryptocurrency and US stock market with the GC-DCC or GCNA-DCC.

Table 8. The results of DCC with LBTC and LSP and with LBTC and LGD. Alpha and beta are the parameters for DCC, NA-DCC, GC-DCC and GCNA-DCC.

DCC		**DCC Alpha**	**DCC Beta**
Bitcoin and S&P 500	Estimate	0.025	0.953
	S.E	0.038	0.067
	Log-likelihood	3179.215	
Bitcoin and Gold	Estimate	0.000	0.227
	S.E	0.009	59,386
	Log-likelihood	3326.434	
NA-DCC		**NA-DCC Alpha**	**NA-DCC Beta**
Bitcoin and S&P 500	Estimate	0.334	0.188
	S.E	0.109	0.591
	Log-likelihood	2754.319	
Bitcoin and Gold	Estimate	NA	NA
	S.E	NA	NA
	Log-likelihood	NA	
GC-DCC		**GC-DCC Alpha**	**GC-DCC Beta**
Bitcoin and S&P 500	Estimate	0.069	0.910
	S.E	0.013	0.019
	Log-likelihood/AIC	3286.216/−10.0129	
Bitcoin and Gold	Estimate	0.069	0.910
	S.E	0.013	0.019
	Log-likelihood/AIC	3437.459/−10.475	
GCNA-DCC		**GCNA-DCC Alpha**	**GCNA-DCC Beta**
Bitcoin and S&P 500	Estimate	0.068	0.899
	S.E	0.015	0.063
	Log-likelihood/AIC	3293.257/−10.028	
Bitcoin and Gold	Estimate	0.068	0.899
	S.E	0.015	0.063
	Log-likelihood/AIC	2143.902/−10.028	

NA means no computational result because of an optimization error from the "*fGarch*" R package.

We have two hypotheses from this research. The first hypothesis is that we want to test the alternative hypothesis that Bitcoin is statistically significant by log-returns of S&P 500 and Gold price in terms of log-return. The second hypothesis is that we also test another alternative hypothesis that the current volatility of log-returns of Bitcoin can be statistically significant with the current and lagged volatilities of the other assets (S&P 500 and Gold price).

To perform the first alternative hypothesis that Bitcoin is statistically significant by log-returns of S&P 500 and Gold price in terms of log-return, we consider building an optimal Autoregressive Moving Average (ARMA) model based on AIC criteria among four different combinations of p and q: (0, 0), (0, 1), (1, 0), and (1, 1). Table 9 shows the result of the selection of p and q for the ARMA model. The ARMA (0, 0) turned out to be the best model with a minimum AIC value and Table 9 shows the result of the GCMR model fit of LBTC with LSP and LGD with error dependence structure of ARMA (0, 0). The reason we employ GCMR for the modeling is that GCMR has a Sigma dispersion parameter which accounts for heteroscedasticity of error. The GCMR model is more flexible to model the data which do not follow normality or heteroscedasticity of errors. Table 9 shows that there exists a statistical significance between LSP and LGD to LBTC in terms of price. And the Sigma dispersion parameter is statistically significant at the 5% significance level.

Table 9. Selection of p and q for ARMA based on AIC of 4 cases of (0, 0), (0, 1), (1, 0), and (1, 1). ARMA (0, 0) is selected based on AIC criteria. CMR model fit of LBTC with LSP and LGD with error dependence structure ARMA (0, 0).

Model	**LBTC = Intercept + $\alpha_1 \times$ LSP + $\beta_1 \times$ LGD**			
ARMA (p,q)	ARMA (0,0)	ARMA (0,1)	ARMA (1,0)	ARMA (1, 1)
AIC	−2140.6	−2139.7	−2139.8	−2138.5
ARMA (0,0)	LBTC = Intercept + $\alpha_1 \times$ LSP + $\beta_1 \times$ LGD			
	Estimate	S.E	Z-value	*p*-value
Intercept	−0.001	0.002	−0.653	0.514
LSP	0.645	0.124	5.182	0.000
LGD	0.803	0.208	3.860	0.000
Sigma	0.047	0.001	36.163	0.000

With volatilities by both standard-GARCH (1, 1) and nonlinear asymmetric GARCH (1, 1), we compare the values of both AIC and Log Likelihood for LBTC Volatility(t) = Intercept + $\alpha_1 \times$ LSP Volatility(t) + $\alpha_2 \times$ LGD Volatility(t)+ $\alpha_3 \times$ LSP Volatility(t-1) + $\alpha_4 \times$ LGD Volatility(t-1) where t-1 is one day before and t = 2, . . . , 401 in Tables 10 and 11.

Table 10. With standard-GARCH (1, 1) volatilities, GCMR model fit of LBTC volatility with LSP volatility(t), LGD volatility (t), volatility (t-1) and LGD volatility (t-1) with error dependence structure ARMA (0, 0).

Model	**LBTC Volatility (t) = Intercept + $\alpha_1 \times$ LSP Volatility (t) + $\alpha_2 \times$ LGD Volatility (t) + $\alpha_3 \times$ LSP Volatility (t-1) + $\alpha_4 \times$ LGD Volatility (t-1)**			
	Estimate	S.E.	Z-value	*p*-value
Intercept	−6.698	0.262	−25.570	0.000
LSP Volatility (t)	1341.294	0.015	90993.019	0.000
LGD Volatility (t)	17.942	3.303	5.431	0.000
LSP Volatility (t-1)	6350.534	0.008	771671.098	0.013
LGD Volatility (t-1)	−11.556	3.662	−3.156	0.001
Shape	0.800	0.209	3.818	0.000
Log Likelihood		−5799.7		
AIC		−11583		

Table 11. With nonlinear asymmetric GARCH (1, 1) volatilities, GCMR model fit of LBTC volatility with LSP volatility(t), LGD volatility (t), volatility (t-1) and LGD volatility (t-1) with error dependence structure ARMA (0, 0).

Model	**LBTC Volatility (t) = Intercept+$\alpha_1 \times$ LSP Volatility (t) + $\alpha_2 \times$ LGD Volatility (t) + $\alpha_1 \times$ LSP Volatility (t-1) + $\alpha_2 \times$ LGD Volatility (t-1)**			
	Estimate	S.E.	Z-value	*p*-value
Intercept	−6.085	0.109	−55.757	0.000
LSP Volatility (t)	24.313	0.032	760.170	0.000
LGD Volatility (t)	5.092	0.583	8.739	0.000
LSP Volatility (t-1)	89.167	0.021	4291.506	0.000
LGD Volatility (t-1)	−2.085	0.636	−3.279	0.001
Sigma	1.388	0.091	15.295	0.000
Log Likelihood		−3837.7		
AIC		−7661.5		

The GCMR model fit of LBTC volatility (t) with LSP volatility (t), LGD volatility (t), volatility (t-1), and LGD volatility (t-1) with standard-GARCH (1, 1) volatilities and error dependence structure ARMA (0, 0) is better than the GCMR model fit of LBTC volatility (t) with LSP volatility (t), LGD volatility (t), volatility (t-1), and LGD volatility (t-1) with nonlinear asymmetric GARCH (1, 1) volatilities and error dependence structure ARMA (0, 0).

We chose the statistical output from Table 10 so that LSP volatility (t), LGD volatility (t) and LSP volatility (t-1) are statistically significant, and they have a positive statistical effect to LBTC volatility (t), but LGD volatility (t-1), one day before volatilities, has a statistically significant negative effect to LBTC volatility (t) at the 5% significance level. The Sigma dispersion parameter is also statistically significant at the 5% significance level in both Tables 10 and 11.

The following statistical output is another interesting result in our paper. We want to see the relationship of the Gaussian copula time-varying correlation (GC-DCC or GCNA-DCC) with the volatilities of LBTC and LSP. With volatilities by both standard-GARCH (1, 1) and nonlinear asymmetric GARCH (1, 1) with an error dependence structure of ARMA (1, 0), we also compared the log-likelihood of GC-DCC = Intercept + $\alpha_1 \times$ LBTC Volatility + $\beta_1 \times$ LSP Volatility with GCNA-DCC = Intercept + $\alpha_1 \times$ LBTC Volatility + $\beta_1 \times$ LSP Volatility in Tables 12 and 13.

Table 12. Gaussian Copula Marginal Regression (GCMR) with standard-GARCH (1, 1) volatilities. Selection of p and q for ARMA based on AIC of 4 cases of (0, 0), (0, 1), (1, 0), and (1, 1). ARMA (0, 0) is selected based on AIC criteria. GCMR Model fit of GC-DCC with LBTC Volatility, and LSP Volatility of Error dependence structure ARMA (1, 1).

Model	**GC-DCC = Intercept + $\alpha_1 \times$ LBTC Volatility + $\beta_1 \times$ LSP Volatility**			
ARMA(p,q)	ARMA(0, 0)	ARMA(0, 1)	ARMA(1, 0)	ARMA(1, 1)
AIC	−1312.2	−1960.7	NA	−3014.6
ARMA (1, 1)	GC-DCC = Intercept + $\alpha_1 \times$ LBTC Volatility + $\beta_1 \times$ LSP Volatility			
	Estimate	S.E	z-value	*P*-value
Intercept	0.068	0.026	2.580	0.010
LBTC Volatility	1.912	4.698	0.407	0.684
LSP Volatility	371.367	0.269	1379.714	0.000
Sigma	0.089	0.013	6.625	0.000
Log-likelihood		−1513.3		

Table 13. Gaussian Copula Marginal Regression with nonlinear asymmetric GARCH (1, 1) volatilities. Selection of p and q for ARMA based on AIC of 4 cases of (0, 0), (0, 1), (1, 0), and (1, 1). ARMA (1, 0) is selected based on AIC criteria. GCMR Model fit of GCNA-DCC with LBTC Volatility, and LSP Volatility of Error dependence structure ARMA (1, 1).

Model	**GCNA-DCC = Intercept + $\alpha_1 \times$ LBTC Volatility + $\beta_1 \times$ LSP Volatility**			
ARMA(p,q)	ARMA(0, 0)	ARMA(0, 1)	ARMA(1, 0)	ARMA(1, 1)
AIC	−678.24	−1994.3	NA	−3070.4
ARMA(1, 1)	GCNA-DCC = Intercept + $\alpha_1 \times$ LBTC Volatility + $\beta_1 \times$ LSP Volatility			
	Estimate	S.E	z-value	*P*-value
Intercept	0.087	0.027	3.223	0.001
LBTC Volatility	16.698	4.572	3.652	0.000
LSP Volatility	131.869	0.264	499.807	0.000
Sigma	0.088	0.014	6.377	0.000
Log-likelihood		−1541.2		

From the relationship among time-varying conditional correlation with LBTC volatility, and LSP volatility by the Gaussian Copula Marginal Regression (GCMR) Model in Tables 12 and 13, we find that there exists a statistically significant and positive effect to time-varying conditional correlation by the volatility of LBTC and the volatility of LSP.

4. Conclusions

We applied the copula-based GARCH-DCC models to the financial assets, Bitcoin, Gold, and S&P 500. We showed that the proposed method for the relationships among time-varying conditional correlation with Bitcoin volatility, and S&P 500 can overcome the difficulty which cannot be computed by the GARCH-DCC models. Our empirical study showed the time-varying relationship between

the cryptocurrency market and the US stock market or the gold market price. Recent data showed that there was a positive time-varying relationship between these two markets since the COVID-19 occurrence. Our Gaussian copula marginal regression modeling the volatility of the most popular cryptocurrency, Bitcoin, with Gold price and US stock market price has more performance compared to competitors such as DCC and NA-DCC to show that a volatility relationship exists among the three market prices with the current day and one-day lagged prices. Our findings provide important implications for both investors and policymakers. In our future study, we will apply state-space modeling for the most popular cryptocurrency with the Gold price and US stock market to see a time-varying relationship in terms of a time-varying intercept and slope. The limitation of this research is that our proposed copula DCC methodology to the high volatility finance assets is not multivariate data analysis but pairwise data analysis. In order to overcome this limitation, our future study will be based on multivariable time series data by using vine copula based multivariate time varying correlation analysis so that we will be able to look at the multivariate time varying correlation behavior among several financial assets simultaneously.

Author Contributions: Formal analysis and investigation, J.-M.K. and S.K.; writing—original draft preparation, S.K., J.-M.K. and S.-T.K.; supervision and reviewing, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Dong-A University, South Korea.

Acknowledgments: We would like to thank the editor and reviewers for their insightful comments and helpful suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nakamoto, S. Bitcoin: A Peer-to-Peer Electronic Cash System. Available online: https://bitcoin.org/bitcoin.pdf (accessed on 3 May 2018).
2. Katsiampa, P. An empirical investigation of volatility dynamics in the cryptocurrency market. *Res. Int. Bus. Finance* **2019**, *50*, 322–335. [CrossRef]
3. Hyun, S.; Lee, J.M.; Kim, J.; Jun, C. What coins lead in the cryptocurrency market? Using Copula and Neural Network Models. *J. Risk Financ. Manag.* **2019**, *12*, 132. [CrossRef]
4. Kim, J.-M.; Jun, C.; Lee, J. Forecasting the volatility of the cryptocurrency market using GARCH and Stochastic Volatility. *Econ. Model.* **2019**. under review.
5. Klein, T.; Thu, H.P.; Walther, T. Bitcoin is not the New Gold—A comparison of volatility, correlation, and portfolio performance. *Int. Rev. Financ. Anal.* **2018**, *59*, 105–116. [CrossRef]
6. Baba, Y.; Engle, R.F.; Kraft, D.F.; Kroner, K.F. Multivariate simultaneous generalized ARCH. *Econ. Theory* **1995**, *11*, 122–150.
7. Aslanidisa, N.; Barivierab, A.F.; Martínez-Ibañeza, O. An analysis of cryptocurrencies conditional cross correlations. *Financ. Res. Lett.* **2019**, *31*, 130–137. [CrossRef]
8. Ghosh, I.; Sanyal, M.K.; Jana, R.K. Co-movement and Dynamic Correlation of Financial and Energy Markets: An Integrated Framework of Nonlinear Dynamics, Wavelet Analysis and DCC-GARCH. *Comput. Econ.* **2020**. [CrossRef]
9. Maraqa, B.; Bein, M. Dynamic Interrelationship and Volatility Spillover among Sustainability Stock Markets, Major European Conventional Indices, and International Crude Oil. *Sustainability* **2020**, *12*, 3908. [CrossRef]
10. Chen, Y.; Qu, F. Leverage effect and dynamics correlation between international crude oil and China's precious metals. *Phys. A Stat. Mech. Appl.* **2019**, 534. [CrossRef]
11. Lee, N.; Kim, J.-M. Dynamic functional connectivity analysis of functional MRI based on copula time-varying correlation. *J. Neurosci. Methods* **2019**, *323*, 32–47. [CrossRef]
12. Guesmi, K.; Saadi, S.; Abid, I.; Ftiti, Z. Portfolio diversification with virtual currency: Evidence from bitcoin. *Int. Rev. Financ. Anal.* **2019**, *63*, 431–437. [CrossRef]
13. Denkowska, A.; Wanat, S. A Tail Dependence-Based MST and Their Topological Indicators in Modeling Systemic Risk in the European Insurance Sector. *Risks* **2020**, *8*, 39. [CrossRef]

14. Chen, H.; Liu, Z.; Zhang, Y.; Wu, Y. The Linkages of Carbon Spot-Futures: Evidence from EU-ETS in the Third Phase. *Sustainability* **2020**, *12*, 2517. [CrossRef]
15. Lee, N.; Kim, J.-M. Dynamic functional connectivity analysis based on time-varying partial correlation with a copula-DCC-GARCH model. *Neurosci. Res.* **2020**. [CrossRef]
16. John, M.; Wu, Y.; Narayan, M.; John, A.; Ikuta, T.; Ferbinteanu, J. Estimation of Dynamic Bivariate Correlation Using a Weighted Graph Algorithm. *Entropy* **2020**, *22*, 617. [CrossRef]
17. Amrouk, E.M.; Grosche, S.C.; Heckelei, T. Interdependence between cash crop and staple food international prices across periods of varying financial market stress. *Appl. Econ.* **2020**, *52*. [CrossRef]
18. Kim, J.M.; Jun, S. Graphical causal inference and copula regression model for apple keywords by text mining. *Adv. Eng. Inform.* **2015**, *29*, 918–929. [CrossRef]
19. Masarotto, G.; Varin, C. Gaussian Copula Marginal Regression. *Electron. J. Stat.* **2012**, *6*, 1517–1549. [CrossRef]
20. Hentschel, L. All in the Family Nesting Symmetric and Asymmetric GARCH Models. *J. Financ. Econ.* **1995**, *39*, 71–104. [CrossRef]
21. Engle, R.F.; Ng, V.K. Measuring and Testing the Impact of News on Volatility. *J. Financ.* **1993**, *48*, 1749–1778. [CrossRef]
22. Tse, Y.K.; Tsui, A.K.C. A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. *J. Bus. Econ. Stat.* **2002**, *20*, 351–362. [CrossRef]
23. Bollerslev, T. Modeling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. *Rev. Econ. Stat.* **1990**, *72*, 498–505. [CrossRef]
24. Engle, R. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *J. Bus. Econ. Stat.* **2002**, *20*, 339–350. [CrossRef]
25. Engle, R.F.; Sheppard, K. *Theoretical and Empirical Properties of Dynamic Conditional Correlation Multivariate GARCH*; Working Paper 8554; National Bureau of Economic Research: Cambridge, MA, USA, 2011. [CrossRef]
26. Sklar, M. *Fonctions de Répartition À N Dimensions et Leurs Marges*; Université Paris: Paris, France, 1959.
27. Joe, H. *Multivariate Models and Dependence Concepts*; Chapman & Hall: London, UK, 1997.
28. Genest, C.; Ghoudi, K.; Rivest, L. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrica* **1995**, *82*, 543–552. [CrossRef]
29. Song, P. Multivariate Dispersion Models Generated from Gaussian Copula. *Scand. J. Stat.* **2000**, *27*, 305–320. [CrossRef]
30. Crypto—Defi Wallet—CoinMarketCap. Available online: https://coinmarketcap.com/coins/ (accessed on 2 April 2018).
31. University of British Columbia, Sauder School of Business. Pacific Exchange Rate Service. Available online: http://fx.sauder.ubc.ca/data.html (accessed on 5 January 2020).
32. Yahoo Finance. Available online: https://finance.yahoo.com/ (accessed on 5 January 2020).
33. Dumitrescu, E.I.; Hurlin, C. Testing for Granger non-causality in heterogeneous panels. *Econ. Model.* **2012**, *29*, 1450–1460. [CrossRef]
34. Bollerslev, T. Generalized Autoregressive Conditional Heteroscedasticity. *J. Econ.* **1986**, *31*, 307–327. [CrossRef]

mathematics

MDPI

Article

Predicting Primary Energy Consumption Using Hybrid ARIMA and GA-SVR Based on EEMD Decomposition

Yu-Sheng Kao [1], Kazumitsu Nawata [1] and Chi-Yo Huang [2,*]

[1] Department of Technology Management for Innovation, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan; sunkao1035@gmail.com (Y.-S.K.); nawata@tmi.t.u-tokyo.ac.jp (K.N.)
[2] Department of Industrial Education, National Taiwan Normal University, Taipei 106, Taiwan
* Correspondence: cyhuang66@ntnu.edu.tw; Tel.: +886-277-493-357

Received: 6 September 2020; Accepted: 27 September 2020; Published: 7 October 2020

Abstract: Forecasting energy consumption is not easy because of the nonlinear nature of the time series for energy consumptions, which cannot be accurately predicted by traditional forecasting methods. Therefore, a novel hybrid forecasting framework based on the ensemble empirical mode decomposition (EEMD) approach and a combination of individual forecasting models is proposed. The hybrid models include the autoregressive integrated moving average (ARIMA), the support vector regression (SVR), and the genetic algorithm (GA). The integrated framework, the so-called EEMD-ARIMA-GA-SVR, will be used to predict the primary energy consumption of an economy. An empirical study case based on the Taiwanese consumption of energy will be used to verify the feasibility of the proposed forecast framework. According to the empirical study results, the proposed hybrid framework is feasible. Compared with prediction results derived from other forecasting mechanisms, the proposed framework demonstrates better precisions, but such a hybrid system can also be seen as a basis for energy management and policy definition.

Keywords: ensemble empirical mode decomposition (EEMD); autoregressive integrated moving average (ARIMA); support vector regression (SVR); genetic algorithm (GA); energy consumption; forecasting

1. Introduction

Research on energy supply and demand has become critical since the 1973 oil crisis. In the past decades, the average annual worldwide energy consumption grew due to the rapid economic growth of major economies. Based on the forecast by BP [1], the worldwide energy consumption will increase by 34% between 2014 and 2035. Over 90% of the world's energy consumption comes from coal, oil, natural gas, and nuclear sources [1]. Furthermore, energy consumption always plays a dominant role in countries' long-term sustainability. For most industries, e.g., heavy industries, much more energy will be required for sustainable growth. Therefore, the understanding and prediction of energy consumption in general, and of a specific economy in particular, are critical from an economic perspective.

Recently, scholars have started to forecast the demand and supply of energy by integrating various models into a hybrid one [2–4]. In general, such hybrid forecasting methods can be divided into two: causal models and time series models. On the one hand, causal models are mainly constructed based on one or more independent variables. Then the dependent variables can be predicted. Strict assumptions and theoretical bases are required for constructing such causal models. On the other hand, time series models are based on historical data. Whether linear or nonlinear, such models are used for estimating future values. These approaches are always regarded as the most feasible ways to predict energy consumption. The fundamental purpose of time series is to derive trends or patterns that can be modeled by econometric methods such as the autoregressive integrated moving average (ARIMA).

However, the models for nonlinear time series predictions are always difficult to realize due to the uncertainty and volatility of the time series. Since the linear models cannot be used to predict complex time series, nonlinear approaches are more suitable for such purpose. Thus, this study aims to predict energy consumption by using integrated methods that incorporate linear and nonlinear methods. Due to issues that arise in time series forecasting, accurate predictions are essential. Conventional linear approaches are effective in the event of forecasting issues. However, more studies are finding that, compared to nonlinear methods, such as support vector machines (SVMs) and tree-based algorithms, linear methods do not perform well in the event of various time series problems, especially complex time sequences. This is because linear methods cannot be used to detect the complex implicit patterns in time series. This study adopts a hybrid model by incorporating linear and nonlinear methods to predict energy consumption and overcome this problem.

However, in accordance with previous studies, no single prediction model is applicable to all scenarios. Therefore, many researchers have introduced hybrid models for predicting time series; such models incorporate both linear and nonlinear models or combine two linear models [5]. Earlier works have also revealed that such hybridization of prediction frameworks not only shows the complementary nature of the frameworks with respect to predictions but also enhances the accuracy of predictions. Thus, models in hybrid forms have become a common practice in forecasting. However, noise and unknown factors exist in time series. These factors influence the volatility of time series and cannot be easily solved by hybridizing linear and nonlinear patterns only. Such hybridization probably produces an overfitting problem; thus, the optimal parameters required to model a prediction framework cannot be derived. Fortunately, such difficulties can be partially solved by leveraging the ensemble empirical mode decomposition (EEMD) proposed by Wu and Huang [6] because the EEMD can solve the noise problems and enhance the prediction performance. Noises, such as trend, seasonality, and unknown factors, which often exist in time series, influence forecasting performance. To make the prediction more precise, noise problems should be carefully dealt with. There are several ways to approach noise problems and enhance forecasting performance. One way is to tune the hyperparameter for algorithms (such as the support vector regression, SVR). As a result, the prediction's performance will improve. Another way is to first deal with the time sequence using a decomposition method, such as the EEMD method. If a decomposition approach is used, the time series can be split into several stable sequences for prediction. The more stable the sequences, the better the model's prediction performance. The feasibility of EEMD has been verified by various works (e.g., [7,8]) in solving nonstationary time series and complex signals.

Given the abovementioned advantages regarding the advantages of the hybrid prediction system as well as the EEMD for time series, this research proposes a novel hybrid framework integrating the EEMD, ARIMA, genetic algorithm (GA), and the SVR to predict primary energy consumption. The concept of the proposed model comes from several sources. First, in conventional time series methods, ARIMA is popular and powerful. It has been extensively used to deal with various forecasting issues. Therefore, the ARIMA method is suitable for this study. Second, based on past research, univariate or single methods used to deal with forecasting problems cannot yield a high forecasting performance when compared to hybrid methods. Therefore, this study simultaneously uses another nonlinear method, SVR, to enhance the prediction performance. In past decades, SVR has been useful in a wide variety of prediction domains. Therefore, SVR is suitable for this study. Although SVR has obtained several important prediction records, it has a significant problem: hyperparameter tuning. Hyperparameter tuning will affect forecasting performance. It is thus necessary to find a way to select the ideal hyperparameter for SVR. Due to computing speed, the authors cannot spend much time searching for the ideal hyperparameters. Consequently, the greedy algorithm (grid search) will be abandoned. Instead, the authors intend to adopt a heuristic algorithm, such as the GA approach, to find the best hyperparameters in SVR. Finally, decomposition methods' effectiveness in enhancing models' forecasting performance has been verified by past literature. The EEMD, Wavelet, and other

methods have been hugely successful in the signal processing field. Due to the innovativeness and power of EEMD, recent studies have widely adopted the method in signal processing research.

Based on above-mentioned reasons, the authors attempt to combine the EEMD, ARIMA, GA, and SVR into a prediction model for energy consumption. First, the time series of energy consumption is divided into several intrinsic mode functions (IMFs) and a residual term. Here, IMFs stand for different frequency bands of time series, which range from high to low, while each IMF represents a series of oscillatory functions [9,10]. That way, each time series component can be identified and modeled accordingly. The characteristics of the time series can also be captured in detail. The ARIMA is then introduced to predict the future values of all extracted IMFs and the residue independently. Since the accuracy of the nonlinear time series derived using the ARIMA may be unacceptable, the SVR is utilized based on the nonlinear pattern to further improve prediction performance. In addition, the accuracy of the SVR-based prediction models completely depends on the control parameters, and the parameters should be optimized. Therefore, the GA is leveraged to derive the optimal parameters. In general, the prediction model fuses EEMD, ARIMA, SVR, and GA into a hybrid prediction framework. The predicted IMFs and the residue will be split into a final ensemble result. The proposed hybrid framework will be verified by a prediction of Taiwanese primary energy consumption for the next four years. Meanwhile, the accuracy of the prediction results will be compared with the ones derived by other forecast methods, which include ARIMA, ARIMA-SVR, ARIMA-GA-SVR, EEMD-ARIMA, EEMD-ARIMA-SVR, and EEMD-ARIMA-GA-SVR. Based on the empirical study results, the effectiveness of the prediction framework can be verified.

The remaining part of this work is structured as follows. Section 2 reviews the related literature regarding the consumption and forecasting methods of energy. Section 3 introduces the methods used in this paper, which include ARIMA, SVR, and GA. Section 4 describes the background of the empirical study case, the dataset, and the empirical study process. Section 5 concludes the whole work, with major findings and opportunities for future studies.

2. Literature Review

For decision-makers to effectively understand the trend of energy fluctuation, which may generate far-reaching implications, the precise forecasting of primary energy consumption is indispensable. On the one hand, with an accurate forecasting of energy consumption, the government can establish energy plans for fluctuations in oil supply. On the other hand, energy predictions can be useful for the investment of firms. Albeit important, predicting energy consumption is not always simple. Therefore, a robust forecasting model will be necessary.

In the literature on energy prediction, several researchers have completed accurate predictions [2–4,11,12]. Some researchers adopted economic indicators by mixing various energy indicators for predicting energy consumption. Other researchers used only time series data for forecasting. While these forecast methods are different, the prediction results of the two categories of models can serve as solid foundations for further investigations on energy consumption.

Furthermore, to enhance the accuracy of predictions, some authors employed hybrid models. Of such models, linear and nonlinear ones were integrated. The fusion of linear and nonlinear models can overcome the shortage of adopting only one kind of method and provide more accurate results [3,5,13]. In addition to the hybrid methods involving both linear and nonlinear models, some studies have attempted to transform the data by integrating data preprocessing and post-processing procedures [14,15]. By doing so, the forecasting capabilities of the hybrid models with data preprocessing and postprocessing procedures can show superior performance in energy predictions.

Further, several studies have proposed machine learning methods for predicting energy consumption. Al-Garni, Zubair, and Nizami [16] used weather factors as explanatory variables of a regression model for predicting the consumption of electric energy in eastern Saudi Arabia. Azadehet al. [12] modeled Turkey's electricity consumption on the sector basis by utilizing the

artificial neural network (ANN). Wong, Wan, and Lam [17] developed an ANN-based model for analyzing the energy consumption of office buildings. Fumo and Biswas [11] employed simple and multiple regression models as well as the quadratic regression model to predict residential energy consumption. Ahmadet al. [13] attempted to review the applications of ANN and the SVM for energy consumption prediction and found that both seem to show better performance in energy forecasting [13]. Ardakani and Ardehali [3] applied regressive methods consisting of linear, quadratic, and ANN models by incorporating an optimization algorithm into the model to achieve better performance in predicting long-term energy consumption.

Although many scholars have empirically verified the effectiveness of machine learning methods for dealing with time series problems, no single prediction model seems applicable to all scenarios. That is, even if the machine learning model outperforms other traditional linear methods, using a single machine learning model to address all time series issues would be problematic and unrealistic.

Many researchers have thus employed hybrid time series forecasting models, which incorporate linear with nonlinear models or combine two kinds of linear models [5]. Previous studies have also revealed that such hybrid frameworks not only complement each other in prediction but also enhance prediction accuracy. Thus, models in hybrid forms have become a common practice in forecasting.

For example, Yuan and Liu [2] proposed a composite model that combined ARIMA and the grey forecasting model, GM (1,1), to predict the consumption of primary energy in China. Based on their findings, the results obtained when using the hybrid model were far superior to those obtained when only using the ARIMA or GM (1,1) models. Zhang [18] developed a hybrid prediction model consisting of both ARIMA and ANN. Zhu et al. [4] developed a hybrid prediction model of energy demands in China, which employed the moving average approach by integrating the modified particle swarm optimization (PSO) method for enhancing prediction performance. Wang, Wang, Zhao, and Dong [19] combined the PSO with the seasonal ARIMA method to forecast the electricity demand in mainland China and obtain a more accurate prediction. Azadeh, Ghaderi, Tarverdian, and Saberi [20] also adopted the GA and ANN models to predict energy consumption based on the price, value-add, number of customers, and energy consumption. Lee and Tong [21] proposed a model that combined ARIMA and genetic programming (GP) to improve prediction efficiency by adopting the ANN and ARIMA models.

Further, Yolcuet et al. [5] developed a linear and nonlinear ANN model with the modified PSO approach for time series forecasting. They achieved prediction results superior to those of conventional forecasting models. According to the analytical results, the hybrid model is more effective because it adopts a single prediction method and can thus improve the prediction accuracy.

Though hybrid models based on ARIMA and ANN have achieved great success in various fields, they have several limitations. First, this hybrid approach requires sufficient data to build a robust model. Second, the parameter control, uncertainties in weight derivations, and the possibility of overfitting must often be discussed when using ANN models. Because of these limitations, more researchers started to adopt SVR in forecasting since it can mitigate the disadvantages of ANN models. SVR is suitable for forecasts based on small datasets.

Pai and Lin's [22] work is a representative example of adopting SVR methods in hybrid models for forecasting. They integrated ARIMA and SVR models for stock price predictions. Patel, Chaudhary, and Garg [23] also adopted the ARIMA-SVR for predictions and derived optimal results based on historical data. Alwee, Hj Shamsuddin, and Sallehuddin [24] optimized an ARIMA-SVR-based model, using the PSO for crime rate predictions. Fan, Pan, Li, and Li [25] employed independent component analysis (ICA) to examine crude oil prices and then used an ARIMA-SVR-based model to predict them.

Based on the results of the literature review, hybrid models, including both the SVR and the ANN, have achieved higher prediction accuracies than traditional prediction techniques. However, the invisible and unknown factors which can influence the volatility of time series cannot be addressed easily by hybridizing linear and nonlinear patterns. The problem of overfitting can emerge; thus, the optimal parameters cannot be derived.

Fortunately, such difficulties can be partially resolved by leveraging the EEMD proposed by Wu and Huang [6]. The method has been feasible and effective in solving problems consisting of nonstationary time series and complex signals [7,8]. Wang et al. [7] integrated the EEMD method with the least square SVR (LSSVR) and successfully predicted the time series of nuclear energy consumption. Prediction performance has increased significantly and outperformed some well-recognized approaches based on level forecasting and directional prediction. Zhang, Zhang, and Zhang [26] predicted the prices of crude oil by hybridizing PSO-LSSVM and EEMD decomposition. The work demonstrated that the EEMD technique can decompose the nonstationary and time-varying components of times series of crude oil prices. The hybrid model can be beneficial to model the different components of crude oil prices and enhance prediction performance.

The previous studies in the literature review section aimed to develop a model that could effectively and accurately predict energy consumption and demand. In their methodologies, these works being reviewed attempted to use linear or nonlinear methods to predict energy consumption. Furthermore, they tried to use the parameter search algorithm in their model to enhance its prediction accuracy. Based on the review results, complex time series can be split by the EEMD into several relatively simple subsystems. The hidden information behind such complex time series can be explored more easily. Thus, in the following section, a hybrid analytical framework consisting of ARIMA, SVR, and GA will be proposed. The framework will be adopted to predict primary energy consumption.

3. Research Methods

This section first introduces the data processing method. Next, the individual models including ARIMA and SVR will be introduced. Afterward, the optimization approach based on GA will be introduced. Finally, the analytical process of the proposed hybrid model will be described.

3.1. EEMD

Empirical mode decomposition (EMD), an adaptive approach based on the Hilbert–Huang transformation (HHT), is often used to deal with time series data including ones with nonlinear and nonstationary forms [8]. Since such time series are complicated, various fluctuation modes may coexist. The EMD technique can be used to decompose the original time series into several simple IMFs, which correspond to different frequency bands of the time series and range from high to low; each IMF stands for a series of oscillatory functions [9,10]. Moreover, the IMFs must satisfy two conditions [6]: (1) in the whole data series, the number of extrema and zero crossings must either be equal or differ at most by 1; and (2) at any point, the mean value of the envelope (envelope, in mathematics, is a curve that is tangential to each one of a family of curves in a plane) defined by the local minima is 0.

Based on the above definitions, IMFs can be extracted from the time series $y(t)$ according to the following shifting procedures [27]: (1) identify the local maxima and the minima; (2) connect all local extrema points to generate an upper envelope $e_{max}(t)$ and connect all minima points to generate a lower envelope $e_{min}(t)$ with the spline interpolation, respectively; (3) compute the mean of the envelope, $a(t)$, from the upper and lower envelopes, where $a(t) = (e_{max}(t) + e_{min}(t))/2$; (4) extract the mean from the time series and define the difference between $y(t)$ and $a(t)$ as $c(t)$, where $c(t) = y(t) - a(t)$; (5) check the properties of $c(t)$: (i) if $c(t)$ satisfies the two conditions illustrated above, an IMF will be extracted and replace $y(t)$ with the residual, $r(t) = y(t) - c(t)$; (ii) if $c(t)$ is not an IMF, then $y(t)$ will be replaced by $c(t)$; and (6) the residue $r_1(t) = y(t) - c_1(t)$ is regarded as the new data subjected to the same shifting process, which was described above for the next IMF from $r_1(t)$. When the residue $r(t)$ becomes a monotonic function or at most has one local extrema point from which no more IMF can be extracted [27], the shifting processes can be terminated.

Through the abovementioned shifting process, the original data series $y(t)$ can be expressed as a sum of IMFs and a residue, $y(t) = \sum_{i=1}^{m} c_i(t) + r_m(t)$, where m is the number of IMFs, $r_m(t)$ is the final

residue, and $c_i(t)$ is the ith IMF. All the IMFs are nearly orthogonal to each other, and all have nearly zero means.

Although the EMD has been widely adopted in handling data series, the mode-mixing problem still exists. The problem can be defined as either a single IMF consisting of components of widely disparate scales or a component of a similar scale residing in different IMFs. To overcome this problem, Wu and Huang [6] proposed the ensemble EMD (EEMD), which adds white noise to the original data, and thus the data series of different scales can be automatically assigned to proper scales of reference built by the white noise [7]. The core concept of the EEMD method is to add the white noise into the data processing. White noise can be viewed as a sequence with zero mean value; this sequence does not fall under any distribution. Based on different algorithms, the white noise can be assigned to specific distributions for calculation. In EEMD, the purpose of this method is to make the original sequence the stable sequence. Hence, this method employs simulation, using the original sequence to generate various sequences of normal distributions—this is the white noise concept. With this method, the original sequence can be split into several different sequences. Meanwhile, the sum of the decomposed sequences equals the original sequence. These decomposed sequences are called IMFs. This way, the mode-mixing problem can be easily solved. The EEMD procedure is developed as follows [6]: (1) add a white noise series to the original data; (2) decompose the data with added white noise into IMFs; (3) repeat steps 1 and 2 iteratively, but with different white noise each time; and (4) obtain ensemble means of corresponding IMFs as the final results.

In addition, Wu and Huang [6] established a statistical rule to control the effect of added white noise: $e_n = \varepsilon / \sqrt{n}$, where n is the number of ensemble members, ε represents the amplitude of the added noise, and e_n is the final standard deviation of error, which is defined as the difference between the input signal and the corresponding IMFs. Based on previous studies, the number of ensemble members is often set to 100, and the standard deviation of white noise is set to 0.2.

3.2. The ARIMA Model

The ARIMA model for forecasting time series was proposed by Box, Jenkins, and Reinsel [28]. The model consists of the autoregressive (AR) and the moving average (MA) models. The AR and MA models were merged into the ARMA model, which has already become matured in predictions. The future value of a variable is a linear function of past observations and random errors. Thus, the ARMA can be defined as

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (1)$$

where y_t is the forecasting value; ϕ_i is the coefficient of the ith observation; y_{t-i} is the ith observation; θ_i is the parameter associated with the ith white noise; ε_t is the white noise, whose mean is zero; and ε_{t-i} is the noise terms.

The ARMA model can satisfactorily fit the original data when the time series data is stationary. However, if the time series are nonstationary, the series will be transformed into a stationary time series using the dth difference process, where d is usually set as 0, 1, or 2. ARIMA is used to model the differenced series. The process is called ARIMA (p, d, q), which can be expressed as

$$w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \cdots + \phi_p w_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (2)$$

where w_t is denoted as $\nabla^d y_t$. When d equals zero, the model is the same as Equation (1) of ARMA.

The ARIMA model was developed by using the Box–Jenkins method. The procedure of the ARIMA involves three steps: (1) Model identification: Since the stationary series is indispensable for the ARIMA model, the data needs to be transformed from a nonstationary one to a stable one. For the stability of the series, the difference method is essential for removing the trends of the series. This way, the d parameter is determined. Based on the autocorrelation function (ACF) and the partial autocorrelation function (PACF), a feasible model can be established. Through the parameter estimation

and diagnostic checking process, the proper model will be established from all the feasible models. (2) Parameter estimation: Once the feasible models have been identified, the parameters of the ARIMA model can be estimated. The suitable ARIMA model based on Akaike's information criterion (AIC) and Schwarz's Bayesian information criterion (BIC) can be further determined. (3) Diagnostic checking: After parameter estimation, the selected model should be tested for statistical significance. Meanwhile, hypothesis testing is conducted to examine whether the residual sequence of the model is a white noise. Based on the above procedures, the forecasting model will be determined. The derived model will be appropriate as the training model for predictions.

In this research, the ARIMA models will be built by feeding each decomposed data. The separated ARIMA models will then be integrated with an SVR model for further analysis. Fitting performance is expected to be enhanced further.

3.3. SVR

SVR was proposed by Vapnik [29], who thought that theoretically, a linear function f exists to define the nonlinear relationship between the input and output data in the high-dimensional-feature space. Such a method can be used to solve the function with respect to fitting problems. Based on the concepts of SVR, the basis function can be described as

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \cdot \varphi(\boldsymbol{x}) + b, \tag{3}$$

where $f(\boldsymbol{x})$ denotes the forecasting values, $\boldsymbol{x}$ is the input vector, $\boldsymbol{w}$ is the weight vector, b is the bias, and $\varphi(\boldsymbol{x})$ stands for a mapping function to transform the nonlinear inputs into a linear pattern in a high-dimensional-feature space.

Conventional regression methods take advantage of the square error minimization method for modeling the forecasting patterns. Such a process can be regarded as an empirical risk in accordance with the loss function [29]. Therefore, the ε − insensitive loss function (T_ε) is adopted in the SVR and can be defined as

$$T_\varepsilon(f(\boldsymbol{x}), y) = \begin{Bmatrix} \left|f(\boldsymbol{x}) - y\right| - \varepsilon \\ 0 \end{Bmatrix} \quad \begin{matrix} \text{if} \left|f(\boldsymbol{x}) - y\right| \geq \varepsilon \\ \text{otherwise} \end{matrix}, \tag{4}$$

where y is the target output and ε is expressed as the region of ε − insensitive. When the predicted value falls into the band area, the loss is equal to the difference between the predicted value and the margin [30]. $T_\varepsilon(f(\boldsymbol{x}), y)$ is leveraged to derive an optimum hyperplane on the high-feature space to maximize the distance which can divide the input data into two subsets. The weight vector (w) and constant (b) in Equations (4) can be estimated by minimizing the following regularized risk function:

$$P(c) = c\frac{1}{n}\sum_{i=1}^{n} T_\varepsilon(f(x_i), y_i) + \frac{1}{2}\|\mathrm{w}\|^2, \tag{5}$$

where $T_\varepsilon(f(\boldsymbol{x}_i), y_i)$ is the ε − insensitive loss function in Equation (5). Here, $1/2\|\mathrm{w}\|^2$ plays the regularizer role, which tackles the problem of trade-off between the complexity and approximation accuracy of the regression model.

Equation (5) above aims to ensure that the forecasting model has an improved generalized performance. In the regularization process, c is used to specify the trade-off between the empirical risk and the regularization terms. Both c and ε can be defined by hyper-parameter search algorithms and users. These parameters significantly determine the prediction performance of the SVR.

In addition, based on the concept of the tube regression, if the predicted value is within the ε − tube, the error will be zero. However, if the predicted value is located outside the ε − tube, the error will be produced. Such an error, the so-called ε − insensitive error, is calculated in terms of the distance between the predicted value and the boundary of the tube. Since some predicted values exist outside the tube, the slack variables $(\xi + \xi_i^*)$ are introduced and defined as tuning parameters. These variables

stand for the distance from actual values to the corresponding boundary values of the tube. Given the synchronous structural risk, Equation (5) is transformed into the following constrained form by using the slack variables:

Minimize $\frac{1}{2}\|w\|^2 + c\sum_{i=1}^{n}(\xi + \xi_i^*)$

Subject to

$$\left\{ \begin{array}{ll} y_i - (\mathrm{w}^T \cdot \varphi(\mathrm{x}_i)) - b \le \varepsilon + \xi_i & \\ (\mathrm{w}^T \cdot \varphi(\mathrm{x}_i)) + b - y_i \le \varepsilon + \xi_i^* & \\ \xi_i, \xi_i^* \ge 0, \quad \text{for} \quad i = 1, \ldots, n & \end{array} \right\} \tag{6}$$

The constant c determines the trade-off between the flatness of f and the amount up to which deviations larger than ε are tolerated. To solve the above problem, the Lagrange multiplier and the Karush–Kuhn–Tucker (KKT) conditions will be leveraged. After the derivation, the general form of the SVR function can be expressed as

$$f(\mathrm{x}, \mathrm{w}) = f(\mathrm{x}, \alpha_i, \alpha_i^*) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \times K(\mathrm{x}, \mathrm{x}_i) + b \tag{7}$$

where α_i and α_i^* are the Lagrange multipliers, and $K(\mathrm{x}, \mathrm{x}_i)$ is the inner product of two vectors in the feature space, $\varphi(\mathrm{x}_i)$ and $\varphi(\mathrm{x}_j)$. Here, $K(\mathrm{x}, \mathrm{x}_i)$ is called the kernel function. In general, the most popular kernel function is the Gaussian radial basis function (RBF), which can be defined as $K(\mathrm{x}_i, \mathrm{x}_j) = \exp(-\|\mathrm{x}_i - \mathrm{x}_j\|^2/2\sigma^2)$ [29]. In this work, the RBF is employed for predictions. In addition, since σ is a free parameter, the RBF kernel can be described as $K(\mathrm{x}_i, \mathrm{x}_j) = \exp(-\gamma\|\mathrm{x}_i - \mathrm{x}_j\|^2)$, where γ is a parameter of the RBF kernel.

3.4. Optimization by GA

While defining a prediction model, the enhancement of prediction accuracy and the avoidance of overfitting are the most important tasks. By doing so, the training model can achieve far better performance in predictions when testing data are inputted. In the SVR-based models, the c, ε, and γ parameters play a dominant role in determining modeling performance. That is, if these parameters can be defined correctly and appropriately, the forecasting model will be efficient. To select the best parameters, the method for searching these parameters will be indispensable. In this work, GA will be adopted in selecting the optimal values for these three SVR parameters.

GA, a concept first proposed by Holland [31], is a stochastic search method based on the ideas of natural genetics and the principle of evolution [32,33]. GA works with a population of individual strings (chromosomes), each of which stands for a possible solution to a given problem. Each chromosome is assigned a fitness value based on the result of the fitness function [34]. GA allows more opportunities to fit chromosomes to reproduce the shared features originating from their parent generation. It has been regarded as a useful tool in many applications and has been extensively applied to derive global solutions to optimization problems. The algorithm is also applicable to large-scale and complicated nonlinear optimal problems [35]. The GA procedure is summarized below based on the work by [36]:

Step 1: Randomly generate an initial population of n agents; each agent is an n – bit genotype (chromosome).

Step 2: Evaluate the fitness of each agent.

Step 3: Repeat the following procedures until n offspring has been created.

(a) Select a pair of parents for mating: A proportion of the existing population is selected to create a new generation. Thus, the most appropriate members of the population survive in this process while the least appropriate ones are eliminated.

(b) Apply variation operators (crossover and mutation): Such operators are inspired by the crossover of the deoxyribonucleic acid (DNA) strands that occur in the reproduction of biological organisms. The subsequent generation is created by the crossover of the current population.

Step 4: Replace the current population with the new one.

Step 5: The whole process is finished when the stopping condition is satisfied. Then the best solution is returned to the current population. Otherwise, the process will go back to Step 2 until the terminating condition can be satisfied.

3.5. The Proposed Hybrid Model

The abovementioned prediction approaches including ARIMA and SVR deliver good performance in general since these methods deal with regression problems effectively. For linear models, ARIMA performs quite well in forecasting time series. However, its prediction performance is limited since it cannot appropriately predict highly nonlinear and nonstationary time series. Therefore, the data stabilization technique based on EEMD decomposition is introduced to handle nonlinear data; the nonstationary process is introduced as well. Complexities such as randomness and intermittence often exist in the time series. Even if the series have been proceeded by the EEMD, unknown factors that influence the series remain. To enhance the forecasting accuracy of the EEMD-ARIMA, the SVR method is integrated. SVR based on the nonlinear method is good at coping with small data and unstable data series. Thus, the EEMD-ARIMA, integrated with the SVR, will be useful in predicting nonlinear time series generally and energy consumption specifically. Meanwhile, GA is adopted to derive the best parameters for SVR, which can improve the performance of the hybrid model.

In general, the proposed EEMD-ARIMA-GA-SVR prediction framework (Figure 1) is composed of the following steps:

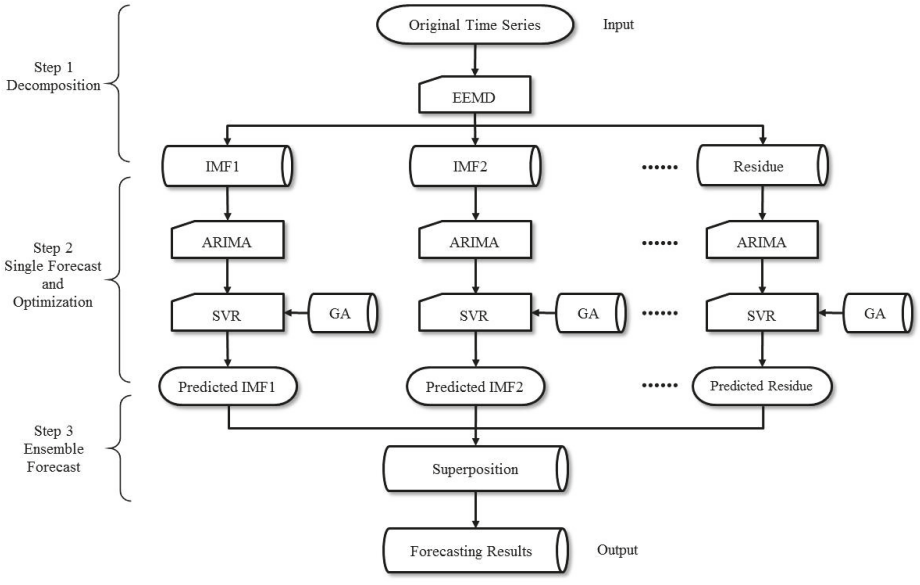


Figure 1. General procedure of the proposed ensemble empirical mode decomposition- autoregressive integrated moving average-genetic algorithm-support vector regression (EEMD)-(ARIMA)-(GA)-(SVR) modeling framework for primary energy consumption forecasting.

Step 1: The original time series of primary energy consumptions, y_t, $t = 1, 2, \ldots, n$, is decomposed into m IMF components $c_i(t)$, $i = 1, 2, \ldots, m$, and a residual component $r_m(t)$ using the EEMD method.

Step 2: ARIMA and SVR are introduced as stand-alone prediction methods to extract the IMFs and residual of the time series, respectively. Meanwhile, GA is introduced to optimize the parameters

associated with SVR. Accordingly, the corresponding prediction results for all components can be obtained.

Step 3: The independent prediction results of all the IMFs and the residual are aggregated as an output, which can be regarded as the final prediction results of the original time series $y(t)$.

Thus, the fitted values can be accordingly derived from this proposed hybrid prediction framework. Further, to demonstrate the effectiveness of the proposed hybrid EEMD-ARIMA-GA-SVR framework, the time series of Taiwanese primary energy consumption will be adopted to verify the feasibility of the proposed framework. Meanwhile, the prediction results based on ARIMA, ARIMA-SVR, ARIMA-GA-SVR, EEMD-ARIMA, and EEMD-ARIMA-SVR will be introduced for comparison.

3.6. Performance Measures for Predictions

Different measures for prediction errors will be adopted to evaluate the accuracy of the prediction models. In this research, the mean absolute error (MAE), mean absolute percentage error (MAPE), mean square error (MSE), and root mean square error (RMSE) will be adopted. The four metrics are used here to evaluate the forecasting performance. The MAPE and RMSE metrics can be useful to explain the performance of predictions. To yield more accurate evaluation results, we further provide results being derived by MAE and MSE as references. These performance measures are defined below:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_t - \hat{y}_t\right| \tag{8}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100 \tag{9}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_t - \hat{y}_t)^2 \tag{10}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_t - \hat{y}_t)^2} \tag{11}$$

where n stands for the size of the test data and y_t as well as $\hat{y}_t$ denote the actual value and the predicted value. Based on these measures, the lower values of all performance measures represent superior forecasting. The MAE reveals how similar the predicted values are to the observed values while the MSE and RMSE measure the overall deviations between the fitted and predicted values. Generally, the MAPE value should be less than 10%. To prove the effectiveness of the proposed hybrid EEMD-ARIMA-GA-SVR model in forecasting, alternative methodologies consisting of ARIMA, ARIMA-SVR, ARIMA-GA-SVR, EEMD-ARIMA, and EEMD-ARIMA-SVR will be used as benchmark models to compare with proposed model.

To test the differences of means between fitted and actual values, the paired-sample Wilcoxon signed-rank test is introduced in this research, where μ_1 stands for the mean of the actual data while μ_2 represents the mean of the fitted data. The null hypothesis is defined as H_0: $\mu_1 - \mu_2 = 0$ while the alternative hypothesis is defined as H_1: $\mu_1 - \mu_2 \neq 0$. The null hypothesis cannot be rejected while the result of the subtraction between the mean value of the actual data and the mean value of the fitted data is 0. Such a case means that there is no mean difference between the fitted data and the actual data. In other words, the training model is suitable as a prediction model since it is consistent with the real situation.

4. Forecasting Taiwan's Primary Energy Consumption

In this section, an empirical case based on Taiwanese primary energy consumption is presented to verify the feasibility and effectiveness of the proposed hybrid framework. Comparisons with

other benchmark models will be provided to demonstrate the forecasting capabilities of the proposed framework. The background of the Taiwanese primary energy consumption will be presented first. Then the raw data and analytic process will be introduced in Section 4.2. Afterward, the time series will be decomposed by the EEMD for the predictions in Section 4.3. Modeling via ARIMA will be introduced in Section 4.4. The predictions of energy consumptions using the EEMD-ARIMA-SVR model optimized by the GA method will be discussed in Section 4.5. Finally, the evaluations of hybrid models and forecasting results will be described in Section 4.6.

4.1. Background

Because of its shortage of natural resources, Taiwan relies heavily on energy imports. In managing energy imports, energy consumption predictions are indispensable. Such energy predictions can help the government sector define relevant energy policies for sustainable development.

4.2. Raw Data and Modeling Method

In this study, the time series of the Taiwanese primary energy consumption (Figure 2) was adopted to verify the effectiveness of the prediction models. Annual primary energy consumption from 1965 to 2014 was provided by BP [1]. The time series was separated into two subsets where 90% (46 samples) of the dataset were chosen as the training set while the remaining 10% (4 samples) were selected as the test set for verifying the prediction efficiency.

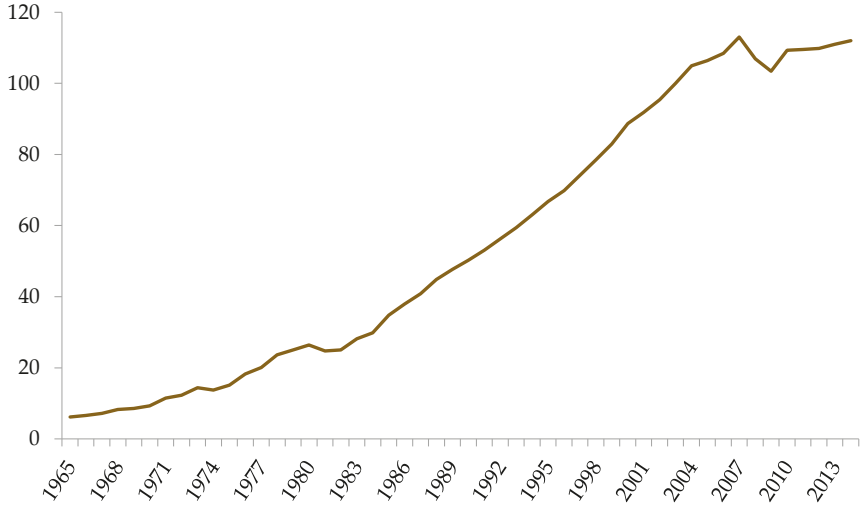


Figure 2. Taiwan's primary energy consumption (million barrels of oil equivalent). Source: BP Statistical Review of World Energy 2015.

For accurate predictions, the original dataset will be transformed by the EEMD method. After the decomposition of the time series, the results of decompositions can be further predicted by the ARIMA method. Then, the prediction results derived from the ARIMA method will be aggregated as the final prediction results for energy consumption.

Meanwhile, such procedures can be extended by adopting SVR. Moreover, to help build the training model and avoid overfitting risks, the *k*-fold cross-validation was introduced into model construction within the prediction process. The *k*-fold cross-validation is used in SVR prediction error examinations based on selected hyperparameters from GA. In this study, the authors adopt five-fold cross-validation. A five-fold validation entails that all samples are divided into five portions;

four of the five portions are used for training, and one of the five portions is for testing. The process is repeated five times. At first, the GA algorithm generates a set of hyperparameters. Through the five-fold cross-validation, we can obtain a performance. Next, the GA algorithm will continue to generate different hyperparameters based on cross-validation, thus obtaining a performance. Finally, we can select the best hyperparameters in terms of the best performances. The training data was randomly divided into k subsamples. Among the k samples, the $k-1$ subsamples were selected as the training data, and the remaining subsample was considered as validation data for testing the model. This research adopted 5-fold cross-validation into the model construction process.

4.3. Data Preprocessing Using EEMD Decomposition

Before conducting the prediction using the proposed hybrid framework, the time series of energy consumptions will be processed using the EEMD decomposition method, which separates the original time series into several IMFs and a residue. The results are depicted in Figure 3. The four independent IMF components correspond to different frequency bands of the time series, which range from high to low.

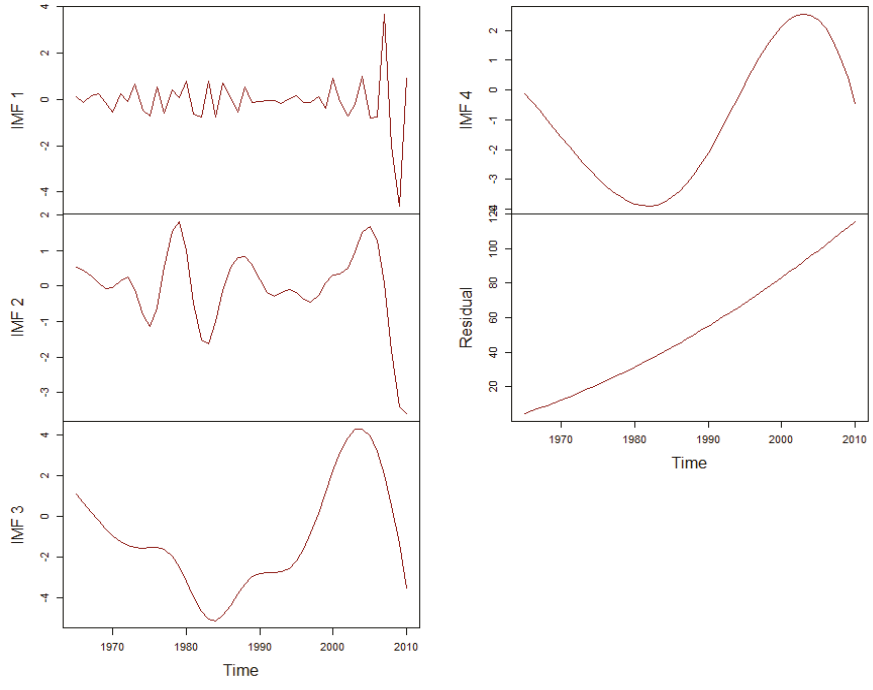


Figure 3. Decomposition results for primary energy consumption via EEMD.

In Figure 3, the IMFs stand for changing frequencies, amplitudes, and wavelengths. IMF_1 represents the highest frequency, maximum amplitude, and the shortest wavelength. The frequencies and amplitudes associated with the rest of the IMFs are lower while the wavelengths are longer. The residue represents a mode slowly varying around the long-term average.

4.4. Forecasting with the ARIMA Model

To establish the prediction model of primary energy consumption in terms of the historical time series dataset, the ARIMA method is introduced. Since the ARIMA model must be built into the series

stationarity, the d times difference needs to be obtained to have an ARIMA (p, d, q) model with d as the order of differencing used. To test the stationarity of d times differencing, the augmented Dickey–Fuller (ADF) test is utilized.

The construction of the ARIMA model depends on model identification. Here, differencing will be important for solving non-stationarity. Moreover, the order of AR (p) and MA (q) needs to be identified. Through ACF and PACF, the order of AR (p) and MA (q) can further be determined [28]. However, the ACF and PACF method may not be useful when performing hybrid ARMA processes. Commonly, AIC or BIC measures can be used to easily inspect the appropriateness of the ARMA model. In this research, the best forecasting model is determined based on AIC measures and statistically significant results. Once the forecasting model has been selected, the nonlinear method based on the optimization approach will be integrated into this model.

First, the stationary test for the decomposed data series and the first difference of the original time series derived from the ADF test is implemented and shown in Table 1. According to the results of the ADF test, all the data series belonged to the stationary. That is, all the transformed data can be used to model constructions.

Table 1. Stationary analysis of primary energy consumption in Taiwan.

Difference and Decomposition	Specification	*t*-Value	Critical Value		
			1%	5%	10%
First difference of original values	Trend	−4.267	−4.150	−3.500	−3.180
	Drift	−2.869	−3.580	−2.930	−2.600
	None	−1.792	−2.620	−1.950	−1.610
IMF1	Trend	−10.045	−4.150	−3.500	−3.180
	Drift	−10.000	−3.580	−2.930	−2.600
	None	−10.022	−2.620	−1.950	−1.610
IMF2	Trend	−9.793	−4.150	−3.500	−3.180
	Drift	−9.954	−3.580	−2.930	−2.600
	None	−10.066	−2.620	−1.950	−1.610
IMF3	Trend	−4.431	−4.150	−3.500	−3.180
	Drift	−4.797	−3.580	−2.930	−2.600
	None	−4.801	−2.620	−1.950	−1.610
IMF4	Trend	−46.468	−4.150	−3.500	−3.180
	Drift	−20.728	−3.580	−2.930	−2.600
	None	−7.099	−2.620	−1.950	−1.610

To further determine the parameters of the ARIMA order, ACF and PACF will be utilized. Likewise, such a procedure can also be followed for the decomposed data set using EEMD. To simplify the analytical procedure of EEMD-ARIMA, the corresponding ACF and PACF diagrams are not presented here.

Through the AIC and the statistical significance test, the suitable ARIMA models can be derived for the first difference time series and decomposed time series via the EEMD. Then the optimal form is specified as ARIMA(1,1,1). Table 1 shows the test results in the first difference of the original sequence. The original sequence is not stationary. Based on the first difference in the original difference, the sequence shows a stable status. Therefore, the difference d will be set as 1. IMF1 ~ IMF4 and the residual are derived from the original sequence using the EEMD method. The sum of the IMFs and the residual is equal to the original sequence. The rationality has already been explained in the fourth paragraph of Section 3.1, where the original sequence was split into several different sequences. The sum of decomposed sequences equals to original sequence. Finally, after determining the proper parameters of the ARIMA models, whether the residual of the selected model possesses the autocorrelation problem should be confirmed. Therefore, the ACF and PACF tests were conducted to verify the selected model. Figure 4 demonstrates the estimated residuals using the ACF and PACF tests. According to the test results, no autocorrelation and partial autocorrelation exist within the residuals.

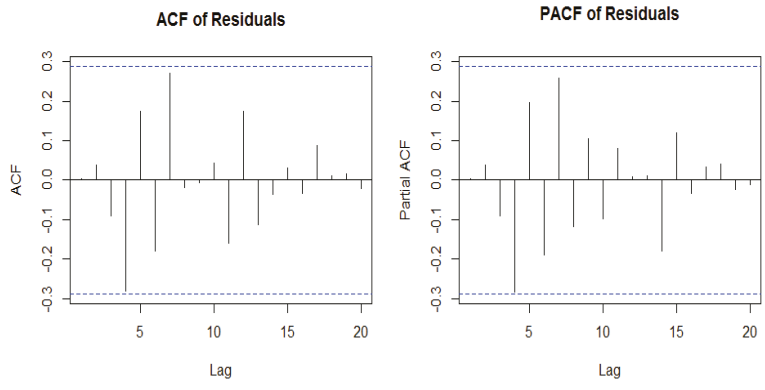


Figure 4. Residual error by autocorrelation function (ACF) and partial autocorrelation function (PACF) for ARIMA(1,1,1).

4.5. Forecasting with the EEMD-ARIMA-SVR Model Optimized by the GA Method

After building the EEMD-ARIMA model, the SVR method will be introduced to reduce the errors produced by ARIMA and enhance forecasting accuracy. According to earlier works, the prediction performance of the SVR method is outstanding; further, it can be fused with other nonlinear or linear methods successfully. Thus, after building the EEMD-ARIMA model, the SVR method will be introduced to reduce the errors produced by ARIMA and enhance forecasting accuracy.

Regarding the hybrid models in this research, the ARIMA initially served as a preprocessor to filter the linear pattern of the decomposed data series. Then, the error terms of the ARIMA model were fed into the SVR model to improve prediction accuracy. Generally, three parameters, c, ε, and γ, can influence the accuracy of the SVR model. Currently, no clear definition and standard procedure are available for determining the above three parameters [21]. However, the improper selection of the three parameters will cause either overfitting or underfitting. To prevent these, the GA method with cross-validation will be introduced to derive the best parameters for constructing the forecasting model. Meanwhile, some studies have pointed out that the utilization of the RBF can yield better prediction performance [21,22]. Thus, the RBF kernel with the 5-fold cross-validation based on the RMSE measure is adopted to help derive the best parameters of the SVR using GA. The above procedures will be applied to the ARIMA-SVR and EEMD-ARIMA-SVR models.

In GA, the number of iterations is set as 100, the population size is defined as 50, and the maximum number of iterations is defined as 50. The search boundaries for c, ε, and γ are within the intervals $[\ 10^{-4},\ \ 10^{2}\ ]$, $[\ 10^{-4},\ \ 2\ ]$, and $[\ 2^{-4},\ \ 2^{2}\ ]$, respectively. The optimal values for c, ε, and γ can thus be 0.441, 3.813, and 0.219, respectively. Further, the c, ε, and γ parameters of the decomposed time series belonging to the four independent IMFs and the residual (Figure 3) are summarized in Table 2. Once the parameters have been derived using GA, the optimal hybrid prediction model can be established.

Table 2. The parameter selection of forecasting model optimized by GA.

Models		c	γ	ε
ARIMA-GA-SVR		0.441	3.813	0.219
EEMD-ARIMA-GA-SVR	IMF1	0.995	1.072	0.195
	IMF2	4.775	0.852	0.100
	IMF3	0.751	1.134	0.461
	IMF4	0.304	3.308	0.470
	Residuals	0.022	1.501	0.352

4.6. Evaluations of Hybrid Models and Forecasting Results

After the construction of the hybrid models, the effectiveness of predictions will be compared further with those of different models including ARIMA, ARIMA-SVR, ARIMA-GA-SVR, EEMD-ARIMA, and EEMD-ARIMA-SVR. The superiority of the proposed EEMD-ARIMA-GA-SVR model of forecasting capability will be verified accordingly. The parameters derived using the GA with the fivefold cross-validation will be utilized to construct the SVR model. The parameters can yield better forecasting performances in related ARIMA-SVR models.

Based on the proposed hybrid framework and the five models for comparisons, the primary energy consumption in Taiwan for 2010–2014 is predicted and summarized in Table 3. According to the prediction results, EEMD-ARIMA-SVR and EEMD-ARIMA-GA-SVR outperformed the other four models. Meanwhile, the prediction results derived using ARIMA and ARIMA-SVR were unsatisfactory from the aspect of inconsistencies between predicted versus actual values. The four prediction performance measures, MAE, MAPE, MSE, and RMSE, derived from the six training models versus the actual values are illustrated in Figure 5 and summarized in Table 4. Based on the results of comparisons, the proposed model outperformed ARIMA. MAE and MSE decreased by 70.43% and 93.28% in the training stage, respectively; further, the two measures decreased by 71.89% and 88.51% in the testing stage, respectively. From the aspects of MAPE and RMSE, both measures improved by 64.68% and 74.07% in the training stage, respectively; they also improved by 71.85% and 66.10% in the testing stage, respectively.

Table 3. Data Test by four sample ranging from 2011 to 2014.

Year	2011	2012	2013	2014
Actual	109.542	109.797	110.959	112.019
ARIMA	111.896	114.527	117.200	119.914
ARIMA-SVR	110.579	113.137	115.735	118.372
ARIMA-GA-SVR	110.071	112.591	115.160	117.777
EEMD-ARIMA	113.174	110.733	111.934	115.542
EEMD-ARIMA-SVR	112.222	110.112	110.809	115.061
EEMD-ARIMA-GA-SVR	112.057	110.075	110.678	114.912

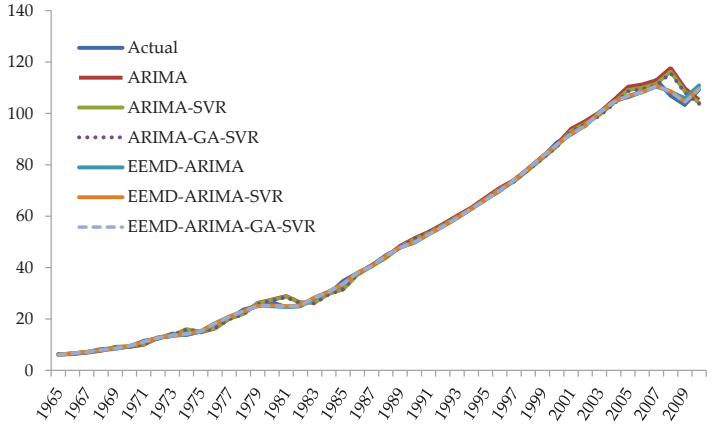


Figure 5. Fittings of the six models.

The proposed model outperformed ARIMA-SVR. MAE and MSE decreased by 67.85% and 91.59% in the training stage, respectively; further, the two measures decreased by 61.52% and 80.31% in the

testing stage, respectively. Meanwhile, MAPE and RMSE improved by 62.26% and 71.01% in the training stage and by 61.44% and 55.63% in the testing stage, respectively.

Table 4. Comparison of forecasting indices.

Models	Training				Testing			
	MAE	**MAPE** (*)	**MSE**	**RMSE**	**MAE**	**MAPE** (*)	**MSE**	**RMSE**
ARIMA	1.397	3.704	5.669	2.381	5.305	4.782	32.301	5.683
ARIMA-SVR	1.285	3.467	4.535	2.130	3.877	3.491	18.851	4.342
ARIMA-GA-SVR	1.257	3.408	4.246	2.061	3.320	2.988	14.723	3.837
EEMD-ARIMA	0.499	1.429	0.611	0.781	2.266	2.048	6.856	2.618
EEMD-ARIMA-SVR	0.425	1.351	0.388	0.623	1.547	1.396	4.139	2.034
EEMD-ARIMA-GA-SVR	0.413	1.308	0.381	0.617	1.492	1.346	3.711	1.926

Remark: *: numbers in percentage.

Compared with ARIMA-GA-SVR, the proposed model performed better as well. MAE and MSE decreased by 67.14% and 91.02%, in the training stage and by 55.08% and 74.79% in the testing stage, respectively. Further, MAPE and RMSE improved by 61.61% and 70.04% in the training stage and by 54.96% and 49.79% in the testing stage, respectively. Based on the above comparison results, the hybrid model integrated with the EEMD method showed better forecasting performance than other models without the data decomposition process.

The proposed model also outperformed EEMD-ARIMA. MAE and MSE decreased by 17.17% and 37.57% in the training stage and by 34.19% and 45.87% in the testing stage, respectively. Meanwhile, MAPE and RMSE were enhanced by 8.41% and 20.98% in the training stage and by 34.27% and 26.43% in the testing stage, respectively.

Compared with EEMD-ARIMA-SVR, the proposed framework performed better. From the aspect of MAE and MSE, the proposed framework outperformed EEMD-ARIMA-SVR by reducing both measures by 2.76% and 1.80% in the training stage and by 3.57% and 10.33% in the testing stage, respectively. At the same time, MAPE and RMSE were enhanced by 3.19% and 0.91% in the training stage and by 3.58% and 5.30% in the testing stage, respectively.

More specifically, from the aspect of training models, the hybrid models including EEMD-ARIMA-SVR and EEMD-ARIMA-GA-SVR performed better, with relatively smaller residual errors in comparison with results derived from any stand-alone or hybrid models. The predictions based on the models without the EEMD decomposition show the limited forecasting capability of the training models. Similarly, from the perspective of the testing model, the proposed model achieved better performance in predicting primary energy consumption. In this study, the comparison results show that the hybrid models with EEMD decomposition can significantly reduce overall forecasting errors. That is, the EEMD method is useful in manipulating the nonstationary time series; thus, better prediction results can be derived by integrating other forecasting methods. Further, GA can reduce forecasting errors by ARIMA-SVR. Based on the analytical results, the proposed EEMD-ARIMA-GA-SVR is a powerful tool and model for energy consumption prediction.

Finally, to identify the significant differences between the prediction results of any two models adopted in this work, the Wilcoxon signed-rank test is performed. This test was adopted extensively in examining the prediction results of two different models and justifying whether these results are significantly different based on small samples [37,38]. The null and alternative hypotheses are described as follows:

$$H_0\text{: } \mu_1 - \mu_2 = 0 \text{ (null hypothesis) and}$$

$$H_1\text{: } \mu_1 - \mu_2 \neq 0 \text{ (alternative hypothesis),}$$

where μ_1 represents the mean of the actual data and μ_2 stands for the mean of the predicted data. The Wilcoxon test can easily determine whether the mean differences are significant or not. All the

p-values derived from the testing of the six pairs of methods were higher than 0.05. That is, no mean differences are observed between the test and predicted values derived from each method.

5. Conclusions

This study presented a hybrid prediction model consisting of ARIMA, SVR, EEMD, and GA. The empirical results have verified the feasibility of the proposed method. Such a hybrid model that combines linear and nonlinear patterns based on the ARIMA and the SVR models as well as adopts data preprocessing and parameter optimization for time series predictions can produce more precise prediction results. From the aspect of limitations and future research possibilities, different datasets on energy consumption can be used at the same time to evaluate whether the forecasting performance of the proposed model will be the best among all prediction models. The data post-processing procedure can be integrated; the differences of the prediction results derived from the proposed model and the framework consisting of the post-processing procedure can be compared. In the future, a dynamic procedure in terms of multiple-step-ahead forecasting can be adopted to replace the one-step-ahead forecasting techniques being used in this work. Much more meaningful and valuable information can be derived for decision-makers in energy predictions.

Author Contributions: Y.-S.K. designed, performed research, analyzed the data, and wrote the paper. K.N. advised on the research methods. C.-Y.H. advised on the research methods, re-wrote, and proof-read the whole article. All authors have read and agreed to the published version of the manuscript.

Funding: This article was subsidized by the Taiwan Normal University (NTNU), Taiwan and the Ministry of Science and Technology, Taiwan under Grant numbers T10807000105 and MOST 106-2221-E-003-019-MY3.

Conflicts of Interest: The authors declare no conflict of interests.

References

1. BP Statistical Review of World Energy. June 2015. Available online: http://www.bp.com/content/dam/bp/pdf/Energy-economics/statistical-review-2015/bpstatistical-review-of-world-energy-2015-full-report.pdf (accessed on 1 July 2017).
2. Yuan, C.; Liu, S.; Fang, Z. Comparison of china's primary energy consumption forecasting by using arima (the autoregressive integrated moving average) model and GM (1, 1) model. *Energy* **2016**, *100*, 384–390. [CrossRef]
3. Ardakani, F.; Ardehali, M. Long-term electrical energy consumption forecasting for developing and developed economies based on different optimized models and historical data types. *Energy* **2014**, *65*, 452–461. [CrossRef]
4. Zhu, S.; Wang, J.; Zhao, W.; Wang, J. A seasonal hybrid procedure for electricity demand forecasting in china. *Appl. Energy* **2011**, *88*, 3807–3815. [CrossRef]
5. Yolcu, U.; Egrioglu, E.; Aladag, C.H. A new linear & nonlinear artificial neural network model for time series forecasting. *Decis. Support Syst.* **2013**, *54*, 1340–1347.
6. Wu, Z.; Huang, N.E. Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Adv. Adapt. Data Anal.* **2009**, *1*, 1–41. [CrossRef]
7. Wang, T.; Zhang, M.; Yu, Q.; Zhang, H. Comparing the applications of emd and eemd on time–frequency analysis of seismic signal. *J. Appl. Geophys.* **2012**, *83*, 29–34. [CrossRef]
8. Tang, L.; Yu, L.; Wang, S.; Li, J.; Wang, S. A novel hybrid ensemble learning paradigm for nuclear energy consumption forecasting. *Appl. Energy* **2012**, *93*, 432–443. [CrossRef]
9. Fei, S.-W. A hybrid model of emd and multiple-kernel rvr algorithm for wind speed prediction. *Int. J. Electr. Power Energy Syst.* **2016**, *78*, 910–915. [CrossRef]
10. Bagherzadeh, S.A.; Sabzehparvar, M. A local and online sifting process for the empirical mode decomposition and its application in aircraft damage detection. *Mech. Syst. Signal Process.* **2015**, *54*, 68–83. [CrossRef]
11. Fumo, N.; Biswas, M.R. Regression analysis for prediction of residential energy consumption. *Renew. Sustain. Energy Rev.* **2015**, *47*, 332–343. [CrossRef]
12. Azadeh, A.; Ghaderi, S.; Sohrabkhani, S. Annual electricity consumption forecasting by neural network in high energy consuming industrial sectors. *Energy Convers. Manag.* **2008**, *49*, 2272–2278. [CrossRef]

13. Ahmad, A.; Hassan, M.; Abdullah, M.; Rahman, H.; Hussin, F.; Abdullah, H.; Saidur, R. A review on applications of ann and svm for building electrical energy consumption forecasting. *Renew. Sustain. Energy Rev.* **2014**, *33*, 102–109. [CrossRef]
14. Xiong, T.; Bao, Y.; Hu, Z. Does restraining end effect matter in emd-based modeling framework for time series prediction? Some experimental evidences. *Neurocomputing* **2014**, *123*, 174–184. [CrossRef]
15. Lin, C.-S.; Chiu, S.-H.; Lin, T.-Y. Empirical mode decomposition–based least squares support vector regression for foreign exchange rate forecasting. *Econ. Model.* **2012**, *29*, 2583–2590. [CrossRef]
16. Al-Garni, A.Z.; Zubair, S.M.; Nizami, J.S. A regression model for electric-energy-consumption forecasting in eastern saudi arabia. *Energy* **1994**, *19*, 1043–1049. [CrossRef]
17. Wong, S.L.; Wan, K.K.; Lam, T.N. Artificial neural networks for energy analysis of office buildings with daylighting. Appl. *Energy* **2010**, *87*, 551–557.
18. Zhang, G.P. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing* **2003**, *50*, 159–175. [CrossRef]
19. Wang, Y.; Wang, J.; Zhao, G.; Dong, Y. Application of residual modification approach in seasonal arima for electricity demand forecasting: A case study of china. *Energy Policy* **2012**, *48*, 284–294. [CrossRef]
20. Azadeh, A.; Ghaderi, S.; Tarverdian, S.; Saberi, M. Integration of artificial neural networks and genetic algorithm to predict electrical energy consumption. *Appl. Math. Comput.* **2007**, *186*, 1731–1741. [CrossRef]
21. Lee, Y.-S.; Tong, L.-I. Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming. *Knowl. Based Syst.* **2011**, *24*, 66–72. [CrossRef]
22. Pai, P.-F.; Lin, C.-S. A hybrid arima and support vector machines model in stock price forecasting. *Omega* **2005**, *33*, 497–505. [CrossRef]
23. Patel, M.; Chaudhary, S.; Garg, S. Machine learning based statistical prediction model for improving performance of live virtual machine migration. *J. Eng.* **2016**, *2016*, 3061674. [CrossRef]
24. Alwee, R.; Hj Shamsuddin, S.M.; Sallehuddin, R. Hybrid support vector regression and autoregressive integrated moving average models improved by particle swarm optimization for property crime rates forecasting with economic indicators. *Sci. World J.* **2013**, *2013*, 951475. [CrossRef]
25. Fan, L.; Pan, S.; Li, Z.; Li, H. An ica-based support vector regression scheme for forecasting crude oil prices. *Technol. Forecast. Soc. Chang.* **2016**, *112*, 245–253. [CrossRef]
26. Zhang, J.-L.; Zhang, Y.-J.; Zhang, L. A novel hybrid method for crude oil price forecasting. *Energy Econ.* **2015**, *49*, 649–659. [CrossRef]
27. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.-C.; Tung, C.C.; Liu, H.H. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **1998**, *454*, 903–995. [CrossRef]
28. Box, G.E.; Jenkins, G.M.; Reinsel, G.C. *Time Series Analysis: Forecasting and Control*; Holdenday: San Francisco, CA, USA, 1976.
29. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 2000.
30. Kao, L.-J.; Chiu, C.-C.; Lu, C.-J.; Yang, J.-L. Integration of nonlinear independent component analysis and support vector regression for stock price forecasting. *Neurocomputing* **2013**, *99*, 534–542. [CrossRef]
31. Holland, J.H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*; University of Michigan Press: Ann Arbor, MI, USA, 1975.
32. Davis, L. *Handbook of Genetic Algorithms*; Van Nostrand Reinhold: New York, NY, USA, 1991.
33. Goldberg, D.E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley Reading: Menlo Park, CA, USA, 1989.
34. Mitchell, M. *An Introduction to Genetic Algorithms*; MIT Press: Cambridge, MA, USA, 1998.
35. Jin, Y.; Branke, J. Evolutionary Optimization in Uncertain Environments-A Survey. *IEEE Trans. Evol. Comput.* **2005**, *9*, 303–317. [CrossRef]
36. Huang, C.-F. A hybrid stock selection model using genetic algorithms and support vector regression. *Appl. Soft Comput.* **2012**, *12*, 807–818. [CrossRef]

37. Wang, Y.; Gu, J.; Zhou, Z.; Wang, Z. Diarrhoea outpatient visits prediction based on time series decomposition and multi-local predictor fusion. *Knowl. Based Syst.* **2015**, *88*, 12–23. [CrossRef]
38. Yan, W. Toward automatic time-series forecasting using neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2012**, *23*, 1028–1039. [PubMed]

mathematics

MDPI

Article

Dispersion Trading Based on the Explanatory Power of S&P 500 Stock Returns

Lucas Schneider * and Johannes Stübinger

Department of Statistics and Econometrics, University of Erlangen-Nürnberg, Lange Gasse 20, 90403 Nürnberg, Germany; johannes.stuebinger@fau.de

* Correspondence: lucas.schneider@fau.de

Received: 16 August 2020; Accepted: 17 September 2020; Published: 20 September 2020

Abstract: This paper develops a dispersion trading strategy based on a statistical index subsetting procedure and applies it to the S&P 500 constituents from January 2000 to December 2017. In particular, our selection process determines appropriate subset weights by exploiting a principal component analysis to specify the individual index explanatory power of each stock. In the following out-of-sample trading period, we trade the most suitable stocks using a hedged and unhedged approach. Within the large-scale back-testing study, the trading frameworks achieve statistically and economically significant returns of 14.52 and 26.51 percent p.a. after transaction costs, as well as a Sharpe ratio of 0.40 and 0.34, respectively. Furthermore, the trading performance is robust across varying market conditions. By benchmarking our strategies against a naive subsetting scheme and a buy-and-hold approach, we find that our statistical trading systems possess superior risk-return characteristics. Finally, a deep dive analysis shows synchronous developments between the chosen number of principal components and the S&P 500 index.

Keywords: dispersion trading; option arbitrage; volatility trading; correlation risk premium; econometrics; computational finance

1. Introduction

Relative value trading strategies, often referred to as statistical arbitrage, were developed by Morgan Stanley's quantitative group in the mid-1980s and describe a market neutral trading approach [1]. Those strategies attempt to generate profits from the mean reversion of two closely related securities that diverge temporarily. Pairs trading, which in its plain form tries to exploit mispricing between two co-moving assets, is probably the most popular delta-one trading approach amongst relative value strategies. Several studies show that those procedures generate significant and robust returns (see [2–6]).

In the non-linear space, relative value strategies are also prominent. Dispersion approaches are one of the most common trading algorithms and attempt to profit from implied volatility spreads of related assets and changes in correlations. Since index options usually incorporate higher implied volatility and correlation than an index replicating basket of single stock options, returns are generated by selling index options and buying the basket. Ultimately, the trader goes long volatility and short correlation [7]. As shown by [8,9], volatility based strategies generate meaningful and reliable returns. Dispersion trades are normally conducted by sophisticated investors such as hedge funds [10]. In 2011, the Financial Times speculated that Och-Ziff Capital Management (renamed to Sculptor Capital Management, Inc. New York, NY, USA in 2019 with headquarter in New York City, United States), an alternative asset manager with $34.5bn total assets under management [11], set up a dispersion trade worth around $8.8bn on the S&P 100 index [12]. In academia, References [13–15] examined the profitability of dispersion trades and delivered evidence of substantial returns across markets.

However, Reference [16] reported that returns declined after the year 2000 due to structural changes in options markets. References [14–16] enhanced their returns by trading dispersion based on an index subset. All of those studies try to replicate the index with as few securities as possible, but neglect the individual explanatory power of stocks in their weighting schemes. This provides a clear opportunity for further improvements of the subsetting procedure.

This manuscript contributes to the existing literature in several aspects. First, we introduce a novel statistical approach to select an appropriate index subset by determining weights based on the individual index explanatory power of each stock. Second, we provide a large-scale empirical study on a highly liquid market that covers the period from January 2000 to December 2017 and therefore includes major financial events such as 9/11 and the global financial crisis. Third, we benchmark our statistical trading approaches against baseline dispersion trading algorithms and a buy-and-hold strategy. Fourth, we conduct a deep dive analysis, including robustness, risk factor, and sub-period analysis, that reports economically and statistically significant annual returns of 14.51 percent after transaction costs. Our robustness analysis suggests that our approach produces reliable results independent of transaction costs, reinvestment rate, and portfolio size. Fifth, we evaluate in depth our innovative selection process and report the number of required principal components and selected sector exposures over the study period. We find a synchronous relationship between the number of principal components that are necessary to describe 90 percent of the stock variance and the S&P 500 index performance, i.e., if the market performs well, more components are required. Finally, we formulate policy recommendations for regulators and investors that could utilize our approach for risk management purposes and cost-efficient dispersion trade executions.

The remainder of this paper is structured as follows. Section 2 provides the underlying theoretical framework. In Section 3, we describe the empirical back-testing framework followed by a comprehensive result analysis in Section 4. Finally, Section 5 concludes our work, provides practical policy recommendations, and gives an outlook of future research areas.

2. Theoretical Framework

This section provides an overview of the theoretical framework of our trading strategy. Section 2.1 describes the underlying methodology and the drivers of dispersion trades. Different dispersion trading structures and enhancement methods are elaborated in Section 2.2.

2.1. Dispersion Foundation and Trading Rational

The continuous process of stock price S is defined as [17]:

$$\frac{dS_t}{S_t} = \mu(t)dt + \sigma(t)dZ_t, \tag{1}$$

where the dt term represents the drift and the second term denotes the diffusion component. In one of its simplest forms, continuous processes follow a constant drift μ and incorporate a constant volatility σ. In various financial models, for example in the well-known Black–Scholes model (see [18]), it is assumed that the underlying asset follows a geometric Brownian motion (GBM):

$$dS_t = \mu S_t dt + \sigma S_t dW_t. \tag{2}$$

Following [19,20], we define return dispersion for an equity index at time t as:

$$RD_t = \sqrt{\sum_{i=1}^{N} w_i \left(R_{i,t} - R_{I,t}\right)^2} = \sqrt{\sum_{i=1}^{N} w_i R_{i,t}^2 - R_{I,t}^2}, \tag{3}$$

where N represents the number of index members, w_i the index weight, and $R_{i,t}$ the return of stock i at time t. Moreover, $R_{I,t}$ denotes the index return with $R_{I,t} = \sum_{i=1}^{N} w_i R_{i,t}$. Return dispersion statistically

describes the spread of returns in an index. References [21–23] showed in their seminal works that the realized variance (RV) is an accurate estimator of the actual variance (σ^2) as RV converges against the quadratic variation. Therefore, RV and σ^2 are used interchangeably in the following. Applying the definition of realized variance of the last J returns $RV_{i,t} = \sum_{k=1}^{J} R_{i,t-k}^2$, Equation (3) can be rewritten as:

$$RD_t^2 = \sum_{i=1}^{N} w_i RV_{i,t} - RV_{I,t}. \tag{4}$$

Expanding Equation (4) by the index variance of perfectly correlated index constituents yields:

$$RD_t^2 = \left(\sum_{i=1}^{N} w_i RV_{i,t} - \left(\sum_{i=1}^{N} w_i \sqrt{RV_{i,t}} \right)^2 \right) + \left(\left(\sum_{i=1}^{N} w_i \sqrt{RV_{i,t}} \right)^2 - RV_{I,t} \right). \tag{5}$$

The first term represents the variance dispersion of the single index constituents. This component is independent of the individual correlations. However, the second expression depends on realized correlations and describes the spread between the index variance under perfectly positively correlated index members and the realized correlations. Therefore, dispersion trades are exposed to volatility and correlation. This is reasonable as the variance of the index is eventually a function of the realized correlations between index members and their variances. Reference [24] already quantified this relationship in 1952:

$$\sigma_I^2 = \sum_{i=1}^{N} w_i^2 \sigma_i^2 + \sum_{i=1}^{N} \sum_{j \neq i}^{N} w_i w_j \sigma_i \sigma_j \rho_{i,j}, \tag{6}$$

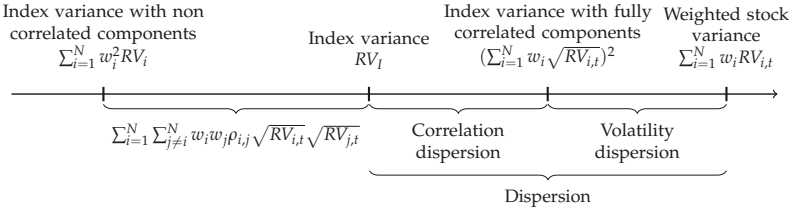
where σ_i (σ_j) represents the volatility of stock i (j) and $\rho_{i,j}$ denotes the correlation coefficient between shares i and j.

Figure 1 illustrates the dispersion drivers graphically. Dispersion consists of volatility and correlation dispersion, i.e., the deviation of the realized correlation from perfectly positively correlated index members. The missing diversification benefits of the index compared to uncorrelated constituents decrease the profit of a dispersion trade, highlighting the short correlation characteristic of a long dispersion trade. Hence, a long dispersion trade profits from an increase in individual stock volatility and a decrease in index volatility, which itself implies a decline in correlation. Shorting dispersion leads to a gain when index volatility rises while the constituent's volatility remains unchanged or falls. In practice, multiple ways exist to structure dispersion trades that possess distinct merits. In our empirical study in Section 3, we develop a cost-efficient way to trade dispersion. Within the scope of this work, we focus on two trading approaches, namely at-the-money (ATM) straddles with and without delta hedging. Both strategies rely on options to trade volatility. When expressing a view on dispersion using non-linear products, one has to consider the implied volatility (IV), which essentially reflects the costs of an option. Profits from owning plain vanilla options are generally achieved when the realized volatility exceeds the implied one due to the payoff convexity. The contrary applies to short positions since these exhibit a concave payoff structure. Thus, a long dispersion trade would only be established if the expected realized volatility exceeds the implied one.

Two essential metrics exist for measuring the implied costs of dispersion trades. First, the costs can be expressed as implied volatility. To assess the attractiveness of a trade, the IV of the index has to be compared with that of the index replicating basket. This method has to assume an average correlation coefficient in order to calculate the IV of the basket. To estimate the average correlation, historical realizations are often used. Second, implied costs can also be directly expressed as average implied correlation, which is computed based on the IVs of the index and its constituents. This measurement simply backs out the average implied correlation so that the IV of the basket equals the index IV. Through modification of the Markowitz portfolio variance equation (see Equation (6)) and assuming

$\rho = \rho_{i,j}$ for $i \neq j$ and $i, j = 1, ..., N$, the implied volatility can be expressed as average implied correlation (see [16]):

$$\overline{\rho} = \frac{\sigma_I^2 - \sum_{i=1}^{N} w_i^2 \sigma_i^2}{\sum_{i=1}^{N} \sum_{j \neq i}^{N} w_i w_j \sigma_i \sigma_j}. \tag{7}$$



For a better understanding, we give the following example: Assume an equal weighted portfolio of two stocks with a pairwise correlation of 0.5 and RV of 0.04 and 0.08, respectively. The arisen dispersion of 0.0159 can be split as followed:

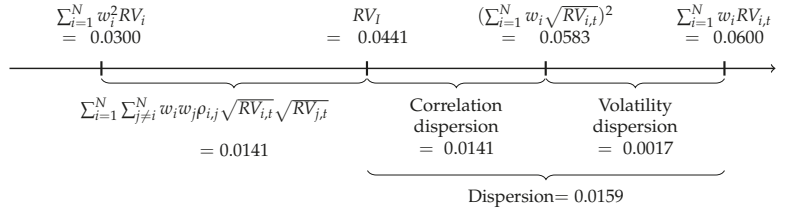


Figure 1. Visualization of dispersion trade performance drivers. Source: authors' calculations.

Figure 2 reports the effect of correlation on index volatility in a simple two stock portfolio case. It is easily perceivable that the index variance decreases with a lower correlation of the two stocks. If the two assets are perfectly negatively correlated, the index variance is virtually zero. This represents the best case for a long dispersion trade as the single index constituents variance is unaffected at 0.0064%, while the index is exposed to 0% variance. This state generates in the simulation a return dispersion of 0.0800%. Nevertheless, this extreme case is not realistic as indices normally incorporate more than two securities. Adding an additional stock leads inevitably to a positive linear relationship with one or the other constituent. Moreover, stocks are typically positively correlated as they exhibit similar risk factors.

The representation of implied costs as average implied correlation is a useful concept as it enables us to assess deviations of index and basket IV by solely computing one figure that is independent of the index size. Hence, this metric is always one-dimensional. Typically, the implied average correlation is compared with historical or forecasted correlations to identify profitable trading opportunities [13,16]. We included this trading filter in our robustness check in Section 4.4 to examine if it is economically beneficial to trade conditional on correlation. Overall, investors would rather sell dispersion in an environment of exceptionally low implied correlation than to build a long position since correlation is expected to normalize in the long run.

However, most of the time investors engage in long dispersion trades. This has mainly two reasons. First, a historical overpricing of implied volatility of index options compared to that of its constituents is persisting. Thus, selling index options is more attractive than buying them. Second, index option sellers can generate profits from the embedded correlation risk premium. Buying index options is ultimately an insurance against decreasing diversification as correlations tend to rise in market

downturns. An increase in correlation negatively affects any portfolio diversification. Therefore, many investors are willing to pay a premium to hedge against rising correlations, making long dispersion trades more attractive [25,26]. Nonetheless, Reference [16] showed that the correlation risk premium seems to play only a minor role in explaining the performance of dispersion trades and concluded that returns depend mainly on mispricing and market inefficiency.

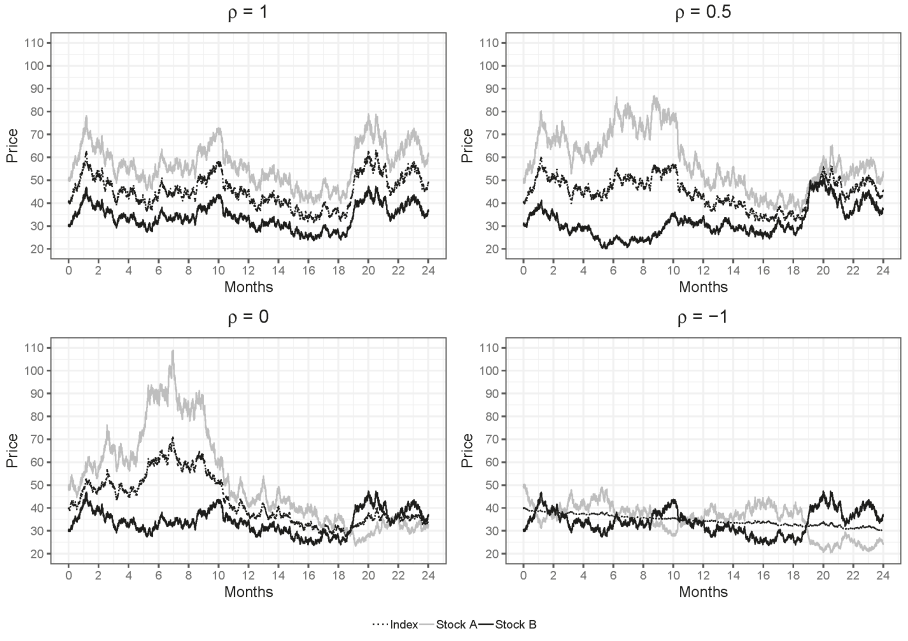


Figure 2. Illustration of the effect of different correlation levels on the variance of an equally weighted index of two stocks. Source: authors' calculations.

Over the years, several explanations for the overpricing of index options have emerged. The most prominent argument is related to the supply and demand of index and stock options. Changes in implied volatility are rooted in net buying and selling pressure. Amongst institutional investors, there is usually high demand for index options, especially puts, to hedge their portfolios. This creates net buying pressure, resulting in a higher implied volatility for index options [27–29]. Hedge funds and other sophisticated investors engage in call and put underwriting on single stock options to earn the negatively embedded volatility premium [30]. Due to consistent overpricing, selling both insurance (puts) and lottery tickets (calls) generates positive net returns in the long run, despite a substantial crash risk [31]. Hence, sophisticated market participants would sell volatility especially when the implied volatility is high. This was for example the case after the bankruptcy of Long-Term Capital Management in 1998 when several investors sold the high implied volatility [32]. The net buying pressure for index options and net selling pressure for stock options create the typical mispricing between index IV and the IV of a replicating portfolio. Dispersion trades therefore help balance the supply and demand of index and single name options as the strategy involves buying the oversupplied single stock options and selling the strongly demanded index options. In a long dispersion trade, investors ultimately act as the liquidity provider to balance single stock and index volatility.

2.2. Dispersion Trading Strategies

In the market, a variety of structures to capture dispersion are well established. In this subsection, we give an overview of the two most common variations, namely at-the-money straddles (Section 2.2.1) and at-the-money straddles with delta hedging (Section 2.2.2).

2.2.1. At-The-Money Straddle

One of the traditional and most transparent, as well as liquid ways to trade volatility is by buying and selling ATM straddles. A long straddle involves a long position in a call and put with the same strike price (K) and maturity (T). Selling both options results in a short straddle. The payoff of a long straddle position at T with respect to the share price (S_T) is given by:

$$P_T = \underbrace{[S_T - K]^+}_{\text{Call}} + \underbrace{[K - S_T]^+}_{\text{Put}}, \tag{8}$$

where $[x]^+$ represents $\max(x, 0)$. Translating an ATM straddle into the context of a long dispersion trade, the payoff equals:

$$P_{T,Dispersion} = \sum_i^N w_i \Big([S_{i,T} - K_i]^+ + [K_i - S_{i,T}]^+ \Big) - \Big([S_{I,T} - K_I]^+ + [K_I - S_{I,T}]^+ \Big). \tag{9}$$

The first part denotes the single stock option portfolio while the second term corresponds to the index leg. When setting up an at-the-money straddle, the initial delta of the position is approximately zero as the put and call delta offset each other [7]. Despite the low delta at inception, this structure is exposed to directional movements and is therefore not a pure volatility play. This is rooted in changes of the position's overall delta as the underlying moves or the time to maturity decreases.

2.2.2. At-the-Money Straddle with Delta Hedging

Delta hedging eliminates the directional exposure of ATM straddles. Through directional hedging, profits from an increase in volatility are locked in. Therefore, this structure generates income even in the case that the underlying ends up at expiry exactly at-the-money. The volatility profits are generated through gamma scalping. A long gamma position implies buying shares of the asset on the way down and selling them on the way up [7]. This represents every long investor's goal: buying low and selling high. A net profit is generated when the realized volatility is higher than the implied volatility at inception as more gamma gains are earned as from the market priced in. Hence, ATM straddles with frequent delta hedging are better suited to trade volatility than a plain option.

Assuming the underlying time series follows a GBM (Equation (2)) and the Black–Scholes assumptions hold, it can be shown that the gamma scalping gains are exactly offset by the theta bleeding when the risk-free rate equals zero [18,33]. Taking the Black–Scholes PDE:

$$\frac{\partial \Pi}{\partial t} + rS\frac{\partial \Pi}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 \Pi}{\partial S^2} = r\Pi \tag{10}$$

and substituting the partial derivatives with the Greeks yields:

$$\Theta + rS\Delta + \frac{1}{2}\sigma^2 S^2 \Gamma = r\Pi. \tag{11}$$

Invoking Itô's lemma, this can be expressed for an infinitesimal small time change dt as:

$$\Theta dt = -\frac{1}{2}\sigma^2 S^2 \Gamma dt. \tag{12}$$

In the above-mentioned equations, Π describes the delta neutral portfolio, S represents the asset price, r illustrates the risk-free interest rate, and Θ and Γ denote the option's price sensitivities with respect to the passage of time and delta.

In the real world when using a discrete delta hedging method, theta and gamma do not necessarily offset each other. This results partially from the risk and randomness that an occasionally hedged straddle exhibits. In fact, the profit-and-loss (P&L) of a long straddle can be approximated as [17,33]:

$$P\&L_{t_0,T} = \sum_{i=0}^{T-t_0} \underbrace{\frac{1}{2}\Gamma_{t_i} S_{t_i}^2}_{\Gamma_{\$,ti}} \underbrace{\left[\left(\frac{\delta S}{S}\right)^2 - \sigma_{implied}^2 \delta t\right]}_{\text{Volatility spread}}, \tag{13}$$

where $\sigma_{implied}$ denotes the implied volatility and δ describes the change in a variable. The first term $\Gamma_{\$,ti}$ is known as dollar gamma, and the last expression illustrates the difference between realized and implied variance, implying that a hedged ATM straddle is indeed a volatility play. However, this approach is still not a pure volatility trade due to the interaction of the dollar gamma with the volatility spread. This relationship creates a path dependency of the P&L. Noticeable, the highest P&L will be achieved when the underlying follows no clear trend and rather oscillates in relatively big movements around the strike price. This is driven by a high gamma exposure along the followed path since the straddle is most of the time relatively close to ATM. A starting trend in either directions would cut hedging profits, despite the positive variance difference. Concluding, delta hedged ATM straddles provide a way to express a view on volatility with a relatively low directional exposure.

3. Back-Testing Framework

This section describes the design of the back-testing study. First, an overview of the software and data used is provided (Section 3.1). Second, Section 3.2 introduces a method to subset index components (formation period). Third, we present our trading strategy in Section 3.3 (trading period). Following [2,34], we divide the dataset into formation-trading constellations, each shifted by approximately one month. Finally, Section 3.4 describes our return calculation method.

3.1. Data and Software

The empirical back-testing study is based on the daily option and price data of the S&P 500 and its constituents from January 2000 to December 2017. This index represents a highly liquid and broad equity market of leading U.S. companies. Hence, this dataset is suitable for examining any potential mispricing since investor scrutiny and analyst coverage are high. The stock price data, including outstanding shares, dividends, and stock split information, were retrieved from the Center for Research in Security Prices (CRSP). Information about the index composition and Standard Industrial Classification (SIC) codes were obtained from Compustat. For the options data of the S&P 500 and its members, we rely on the OptionMetrics Ivy database, which offers a comprehensive information set. From that database, all available one month options were extracted. We mention that OptionMetrics provides for single stocks only data on American options. Thus, the returns in Section 4 are probably understated. This conclusion arises from two facts. First, the constructed strategies could also be established with European options that usually trade at a discount to American ones due to the lack of the early exercise premium. Second, a dispersion strategy is typically long individual stock options, therefore resulting in higher initial costs. Nonetheless, this study provides a conservative evaluation of the attractiveness of dispersion trades. The above-mentioned data provider was accessed via Wharton Research Data Services (WRDS). The one month U.S. Dollar LIBOR, which is in the following used as a risk-free rate proxy, was directly downloaded from the Federal Reserve Economic Research database and transformed into a continuously compounded interest rate. Moreover, we use the Kenneth R. French data library to obtain all relevant risk factors.

To keep track of the changes in the index composition, a constituent matrix with appropriate stock weights is created. As the weights of the S&P 500 are not publicly available, the individual weights are reconstructed according to the market capitalization of every stock. This approach leads not to a 100% accurate representation of the index as the free-float adjustment factors, used by the index provider, are not considered. However, it provides us with a reliable proxy. The options data was cleaned by eliminating all data points with missing quotes. Furthermore, the moneyness was calculated for every option to determine ATM options.

All concepts were implemented in the general-purpose programming language Python. For some calculations and graphs, the statistical programming language R was used as a supplementary [35]. Computationally intensive calculations were outsourced to cloud computing platforms.

3.2. Formation Period

Transaction costs play a major role in trading financial securities (see [36,37]). In particular, when traded products are complex or exotic, transaction costs might be substantial. Reference [38] showed that the portfolio construction is of great importance in order to execute strategies cost efficiently in the presence of transaction costs. Therefore, portfolio building represents a material optimization potential for trading strategies. One simple way to reduce transaction fees is to trade less. In dispersion trades, the single option basket is the main trading costs driver. Therefore, it is desirable to reduce the number of traded assets, especially when the portfolio is delta hedged. However, trading less stocks as the index incorporates could result in an insufficient replication, which might not represent the desired dispersion exposure accurately.

To determine an appropriate subset of index constituents that acts as a hedge for the index position, a principal component analysis (PCA) is used. This statistical method converts a set of correlated variables into a set of linearly uncorrelated variables via an orthogonal transformation [39,40]. The main goal of this method is to find the dominant pattern of price movements, as well as stocks that incorporate this behavior most prominently. Selecting those assets leads to a portfolio that explains the majority of the index movement.

The fundamentals of the applied procedure are based on [15,16,41], but we improve the selection process by identifying stocks with the highest explanatory power. Therefore, we are in a position to get a basket of stocks that explains the index in a more accurate way. To be more specific, our method is comprised of six steps that recur every trading day:

1. Calculate the covariance matrix of all index members based on the trailing twelve month daily log returns.
2. Decompose the covariance matrix into the eigenvector and order the principal components according to their explanatory power.
3. Determine the first I principal components that cumulatively explain 90% of the variance.
4. Compute the explained variation of every index constituent through performing Steps 1 and 2 while omitting the specific index member and comparing the new explained variance of the I components to that of the full index member set.
5. Select the top N stocks with the highest explained variation.
6. Calculate the individual weights as the ratio of one index member's explained variation to the total explained variation of the selected N stocks.

To illustrate our approach in more detail, we report an example of our portfolio construction methodology for the trading day 29/11/2017 below.

1. Calculating the trailing 12 month return covariance matrix, after excluding stocks with missing data points, yields the following 410×410 matrix.

	AAPL	**MSFT**	**AMZN**	**FB**	**...**	**SIG**	**DISCA**	**VIA**	**LEN**
AAPL	0.00012	0.00004	0.00007	0.00006	...	-0.00003	0.00000	-0.00002	0.00002
MSFT	0.00004	0.00008	0.00007	0.00005	...	-0.00002	0.00000	-0.00001	0.00002
AMZN	0.00007	0.00007	0.00017	0.00009	...	-0.00004	0.00002	-0.00001	0.00003
FB	0.00006	0.00005	0.00009	0.00011	...	-0.00004	0.00003	0.00002	0.00003
⋮	⋮	⋮	⋮	⋮	...	⋮	⋮	⋮	⋮
SIG	-0.00003	-0.00002	-0.00004	-0.00004	...	0.00119	0.00010	0.00003	0.00004
DISCA	0.00000	0.00000	0.00002	0.00003	...	0.00010	0.00030	0.00017	0.00005
VIA	-0.00002	-0.00001	-0.00001	0.00002	...	0.00003	0.00017	0.00048	0.00004
LEN	0.00002	0.00002	0.00003	0.00003	...	0.00004	0.00005	0.00004	0.00020

2. Decomposing the covariance matrix from 1 into the eigenvectors results in 410 principal components. To keep this section concise, we report the selected components in Table 1 below.
3. Examining the cumulative variance of the principal components in the last column of Table 1 shows that we have to set $I = 36$ to explain 90% of the variance.
4. After repeating Steps 1 and 2 while omitting one index member at a time, we receive the new cumulative variance of the 36 components for every stock that enables us to calculate the individual explanatory power by comparing the new cumulative variance to that of the full index member set (Columns 3–5 in Table 2).
5. We select the top five stocks with the highest explained variation, which in our example are Xcel Energy Inc. (XEL), Lincoln National Corporation (LNC), CMS Energy Corporation (CMS), Bank of America Corporation (BAC), and SunTrust Banks, Inc. (STI).
6. Based on the explained variation of the top five stocks, we calculate the appropriate portfolio weights as the ratio of individually explained variation and total explained variation of the selected stocks. Eventually, we arrive in Table 2 at the reported portfolio weights.

Table 1. Total explained variance of the selected principal components.

Principal Component	**% of Variance**	**Cumulative Variance %**
1	32.01	32.01
2	17.02	49.03
3	6.32	55.35
4	5.65	61.00
5	4.74	65.73
⋮	⋮	⋮
35	0.34	89.82
36	0.33	90.16
37	0.31	90.47
38	0.30	90.77
39	0.29	91.06
40	0.28	91.34
⋮	⋮	⋮
410	0.00	100.00

Source: authors' calculations.

Table 2. Ranking of the index constituents by explanatory power and portfolio weights.

		Cumulative Variance %			
Rank	**Stock Ticker**	**All Constituents**	**Omitting Stocks**	**Explanatory Power %**	**Portfolio Weight %**
1	XEL	90.1576	90.1358	0.0219	20.2518
2	LNC	90.1576	90.1359	0.0217	20.1303
3	CMS	90.1576	90.1360	0.0216	20.0108
4	BAC	90.1576	90.1361	0.0215	19.9274
5	STI	90.1576	90.1364	0.0212	19.6797
⋮	⋮	⋮	⋮	⋮	⋮
406	ULTA	90.1576	90.2146	−0.0570	0.0000
407	ADM	90.1576	90.2171	−0.0595	0.0000
408	NRG	90.1576	90.2177	−0.0601	0.0000
409	HRB	90.1576	90.2207	−0.0631	0.0000
410	EFX	90.1576	90.2333	−0.0757	0.0000

Source: authors' calculations.

3.3. Trading Period

As [42] showed that even professional traders that are normally considered as rational and sophisticated suffer from behavioral biases, our strategy is based on predefined and clear rules to alleviate any unconscious tendencies. In line with [15,16], we implement our trading strategies based on one month ATM options. The following rules specify our trading framework:

1. Whenever one month ATM options are available, a trading position is established.
2. A trading position consists always of a single stock option basket and an index leg.
3. Every position is held until expiry.

Whenever a new position is established, we invest 20% of our capital. The remaining capital stock acts as a liquidity buffer to cover obligations that may emerge from selling options. All uncommitted capital is invested at LIBOR. In total, we construct the four trading systems PCA straddle delta hedged (PSD), PCA straddle delta unhedged (PSU), largest constituents straddle delta hedged (LSD), and largest constituents straddle delta unhedged (LSU). In order to benchmark our index subsetting scheme (Section 3.2), we also apply our strategies to a naive subset of the index, consisting of the five largest constituents. As a point of reference, a simple buy-and-hold strategy on the index (MKT) is reported. Details can be found in the following lines.

- PCA straddle delta hedged (PSD): The replicating portfolio of this strategy consists of the top five stocks that are selected with the statistical approach outlined in Section 3.2. A long straddle position of the basket portfolio is established while index straddles are sold. The overall position is delta hedged on a daily basis.
- PCA straddle delta unhedged (PSU): The selection process and the trade construction are similar to PSD. However, the directional exposure remains unhedged during the lifetime of the trade. This illustrates a simpler version of PSD.
- Largest constituents straddle delta hedged (LSD): The replicating portfolio of this strategy follows a naive subsetting approach and contains the top five largest constituents of the index at trade inception. Basket straddles are bought while index straddles are shorted. The directional exposure is delta hedged on a daily basis.
- Largest constituents straddle delta unhedged (LSU): The trade construction and selection process is identical to LSD. However, the overall delta position remains unhedged, hence representing a less sophisticated approach than LSD.
- Naive buy-and-hold strategy (MKT): This approach represents a simple buy-and-hold strategy. The index is bought in January 2000 and held during the complete back-testing period. MKT is the simplest strategy in our study.

Transaction costs are inevitable when participating in financial markets. Thus, transaction fees have to be considered in evaluating trading strategies to provide a more realistic picture of profitability. Besides market impact, bid-ask spreads, and commissions, slippages are the main cost driver. Over the last few years, trading costs decreased due to electronic trading, decimalization, increased competition, and regulatory changes (see [43,44]). The retail sector also benefited from this development as brokers such as Charles Schwab first decreased and then completely eliminated fees for stocks, ETFs, and options that are listed on U.S. or Canadian exchanges [45]. However, transaction costs are hard to estimate as they depend on multiple factors such as asset class and client profile. To account for asset specific trading fees in our back-testing study, we apply 10 bps for every half turn per option, to which 2.5 bps are added if delta hedging is performed. In light of our trading strategy in a highly liquid equity market, the cost assumptions appear to be realistic.

3.4. Return Calculation

In contrast to [2,37,46], who constructed a fully invested portfolio, we base our returns on a partially invested portfolio. Our calculation is similar to the concept of return on risk-adjusted capital (RORAC). There are two reasons for choosing this method. First, short selling options incorporates extreme payoffs that can easily outstrip option premiums. As a result, investors need substantial liquidity to cover any future cash outflows to honor their obligations. Second, writing options requires a margin to enter and maintain the position. For example, the initial margin for a one month short index ATM call position at the Chicago Board Options Exchange (CBOE) amounts to 100% of the options proceeds plus 15% of the aggregated underlying index value minus the out-of-the-money amount [47].

The return of a dispersion trade is calculated as:

$$R_{t,T} = \begin{cases} \dfrac{P_{T,Stocks}}{V_{t,Stocks}} - \dfrac{P_{T,Index}}{V_{t,Index}}, & \text{if long dispersion,} \\ -\dfrac{P_{T,Stocks}}{V_{t,Stocks}} + \dfrac{P_{T,Index}}{V_{t,Index}}, & \text{if short dispersion,} \end{cases} \tag{14}$$

where V_t represents the initial costs of the individual legs:

$$V_{t,Stocks} = \sum_{i}^{N} w_i \Big(C_{i,t}(K_i, T) + P_{i,t}(K_i, T) \Big), \tag{15}$$

$$V_{t,Index} = C_{I,t}(K_I, T) + P_{I,t}(K_I, T). \tag{16}$$

C_t and P_t denote the prices for calls and puts for a specific time to maturity (T) and strike (K). When delta hedging is conducted, the generated P&L is added to the nominator of Equation (14) for both legs. As [18] assumed continuous hedging is impracticable in reality due to transaction costs and a lack of order execution speed, we undertake daily delta hedging at market close. To calculate the delta exposure (Δ), the Black–Scholes framework is used. $IV_j \, \forall j \in (t, T)$ is assumed to be the annualized one month trailing standard deviation of log returns. Any proceeds (losses) from delta hedging are invested (financed) at dollar LIBOR (r). This rate also serves as borrowing rate when cash is needed to buy shares (S). Hence, the delta P&L at T for a portfolio of N options is determined by:

$$P\&L_{\Delta,Portfolio} = \sum_{t=1}^{T} \sum_{i=1}^{N} w_i \left[(\Delta_{i,t} - \Delta_{i,t-1}) S_{i,t} * e^{r(T-t)} - \Delta_{i,t-1} (S_{i,t} + D_{i,t} - S_{i,t-1}) \right], \tag{17}$$

where $D_{i,t}$ denotes the present value at time t of the dividend payment of stock i.

4. Results

We follow [48] and conduct a fully-fledged performance evaluation on the strategies PSD, PSU, LSD, and LSU from January 2000 to December 2017—compared to the general market MKT. The key results for the options portfolio of the top five stocks are depicted in two panels—before and after transaction costs. First, we evaluate the performance of all trading strategies (Section 4.1), conduct a sub-period analysis (Section 4.2), and analyze the sensitivity to varying market frictions (Section 4.3). Second, Section 4.4 checks the robustness, and Section 4.5 examines the exposure to common systematic risk factors. Finally, we investigate the number of principal components and the corresponding sector exposure of our PCA based selection process (Section 5).

4.1. Strategy Performance

Table 3 shows the risk-return metrics per trade and the corresponding tradings statistics for the top five stocks per strategy from January 2000 to December 2017. We observe statistically significant returns for PSD and PSU, with Newey–West (NW) t-statistics above 2.20 before transaction costs and above 1.90 after transaction costs. A similar pattern also emerges from the economic perspective: the mean return per trade is well above zero percent for PSD (0.84 percent) and PSU (2.14 percent) after transaction costs. In contrast, LSD produces a relatively small average return of 0.30 percent per trade. It is very interesting that LSU achieves 1.28 percent, but this is not statistically significant. The naive buy-and-hold strategy MKT achieves identical results before and after transaction costs because the one-off fees are negligible. The range, i.e., the difference of maximum and minimum, is substantially lower for the delta hedged strategies PSD and LSD, which reflects the lower directional risk of those approaches. Furthermore, the standard deviation of PSU (15.23 percent) and LSU (13.53 percent) is approximately two times higher than that of PSD (6.78 percent) and LSD (7.25 percent). We follow [49] and report the historical value at risk (VaR) figures. Overall, the tail risk of the delta hedged strategies is greatly reduced, e.g., the historical VaR 5% after transaction costs for PSD is -11.13 percent compared to -20.90 percent for PSU. The decline from a historical peak, called maximum drawdown, is at a relatively low level for PSD (70.53 percent) compared to PSU (94.70 percent), LSD (82.01 percent), and LSU (85.64 percent). The hit rate, i.e., the number of trades with a positive return, varies between 52.15 percent (LSD) and 56.71 percent (PSD). Across all systems, the number of actually executed trades is 395 since none of the strategies suffers a total loss. Consequently, the average number of trades per year is approximately 22; this number is well in line with [50].

In Table 4, we report annualized risk-return measures for the strategies PSD, PSU, LSD, and LSU. The mean return after transaction costs ranges from 0.77 percent for LSD to 26.51 percent for PSU. As anticipated from Table 3, the standard deviation of both delta hedged strategies amounts to approximately 20%—half of the unhedged counterparts. Notably, the Sharpe ratio, i.e., the excess return per unit of standard deviation, of PSD clearly outperforms the benchmarks with a value of 0.40 after transaction costs. Concluding, PSD generates promising risk-return characteristics, even after transaction costs.

Table 3. Return characteristics, risk metrics, and trading statistics per trade for PCA straddle delta hedged (PSD), PCA straddle delta unhedged (PSU), largest constituents straddle delta hedged (LSD), and largest constituents straddle delta unhedged (LSU) from January 2000 until December 2017. NW denotes Newey–West standard errors and CVaR the conditional value at risk.

	Before Transaction Costs					**After Transaction Costs**				
	PSD	**PSU**	**LSD**	**LSU**	**MKT**	**PSD**	**PSU**	**LSD**	**LSU**	**MKT**
Mean return	0.0100	0.0228	0.0044	0.0140	0.0022	0.0084	0.0214	0.0030	0.0128	0.0022
t-statistic (NW)	2.2413	2.2086	0.8608	1.5083	1.2897	1.9069	2.0933	0.5809	1.3852	1.2897
Standard error (NW)	0.0044	0.0103	0.0051	0.0093	0.0017	0.0044	0.0102	0.0051	0.0093	0.0017
Minimum	−0.2560	−0.4539	−0.2399	−0.4798	−0.2506	−0.2556	−0.4529	−0.2396	−0.4781	−0.2506
Quartile 1	−0.0238	−0.0640	−0.0288	−0.0704	−0.0097	−0.0251	−0.0648	−0.0300	−0.0711	−0.0097
Median	0.0097	0.0163	0.0037	0.0201	0.0037	0.0082	0.0150	0.0022	0.0188	0.0037
Quartile 3	0.0489	0.1006	0.0374	0.0884	0.0170	0.0470	0.0988	0.0356	0.0866	0.0170
Maximum	0.3241	1.1694	0.3414	0.4270	0.1315	0.3202	1.1626	0.3373	0.4258	0.1315
Standard deviation	0.0678	0.1523	0.0730	0.1361	0.0363	0.0673	0.1515	0.0725	0.1353	0.0363
Skewness	−0.0818	1.2340	0.2528	−0.1259	−1.5218	−0.0806	1.2366	0.2546	−0.1249	−1.5218
Kurtosis	2.7042	8.9553	2.3025	0.9795	9.2676	2.7051	8.9824	2.3021	0.9806	9.2676
Historical VaR 1%	−0.1955	−0.3580	−0.1820	−0.3584	−0.1431	−0.1955	−0.3570	−0.1820	−0.3574	−0.1431
Historical CVaR 1%	−0.2166	−0.4080	−0.2133	−0.4068	−0.1788	−0.2165	−0.4069	−0.2132	−0.4056	−0.1788
Historical VaR 5%	−0.1106	−0.2091	−0.1178	−0.2052	−0.0486	−0.1113	−0.2090	−0.1184	−0.2051	−0.0486
Historical CVaR 5%	−0.1573	−0.2889	−0.1632	−0.2977	−0.1000	−0.1576	−0.2884	−0.1635	−0.2971	−0.1000
Maximum drawdown	0.6731	0.9358	0.8074	0.8534	0.4990	0.7053	0.9470	0.8201	0.8564	0.4990
Trades with return ≥ 0	0.5747	0.5519	0.5342	0.5671	0.5975	0.5671	0.5468	0.5215	0.5646	0.5975
Number of trades	395	395	395	395	1	395	395	395	395	1
Avg. trades per year	21.9444	21.9444	21.9444	21.9444	0.0556	21.9444	21.9444	21.9444	21.9444	0.0556

Source: authors' calculations.

Table 4. Annualized risk-return measures for PSD, PSU, LSD, and LSU from January 2000 until December 2017.

	Before Transaction Costs					**After Transaction Costs**				
	PSD	**PSU**	**LSD**	**LSU**	**MKT**	**PSD**	**PSU**	**LSD**	**LSU**	**MKT**
Mean return	0.1841	0.2988	0.0401	0.1049	0.0347	0.1452	0.2651	0.0077	0.0779	0.0347
Mean excess return	0.1646	0.2794	0.0206	0.0854	0.0152	0.1257	0.2456	−0.0117	0.0585	0.0152
Standard deviation	0.3192	0.7169	0.3438	0.6403	0.1710	0.3169	0.7128	0.3414	0.6367	0.1710
Downside deviation	0.2074	0.4110	0.2271	0.4264	0.1305	0.2090	0.4114	0.2287	0.4266	0.1305
Sharpe ratio	0.5157	0.3897	0.0599	0.1334	0.0890	0.3967	0.3446	−0.0344	0.0918	0.0890
Sortino ratio	2.3481	2.0620	0.4254	0.6190	0.6389	1.7935	1.7911	0.0799	0.4518	0.6389

Source: authors' calculations.

4.2. Sub-Period Analysis

Investors are concerned about the stability of the results, potential drawdowns, and the behavior in high market turmoils. Inspired by the time-varying returns of [51–53], we analyze the performance of our implemented trading strategies over the complete sample period. Figure 3 illustrates the different dispersion structures across the three different sub-periods January 2000–December 2005, January 2006–December 2011, and January 2012–December 2017. The graphs show the development of an investment of 1 USD at the beginning of each sub-period.

The first sub-period ranges from 2000 to 2005 and includes the dot-com crash, the 9/11 attacks, and the time of moderation. For PSD and LSD, 1 USD invested in January 2000 grows to more than 2.00 USD at the end of this time range; both show a steady growth without any substantial drawdowns. PSU, LSU, and MKT exhibit a similar behavior, which is clearly worse than the hedged strategies. The second sub-period ranges from 2006 to 2011 and describes the global financial crisis and its aftermath. The strategy PSD seems to be robust against any external effects since it copes with the global financial crisis in a convincing way. As expected, the unhedged trading approaches PSU and LSU show strong swings during high market turmoils. The third sub-period ranges from 2012 to 2017 and specifies the period of regeneration and comebacks. The development of all strategies does not decline even after transaction costs; profits are not being arbitraged away. Especially PSU outperforms in this time range with a cumulative return up to 300 as a consequence of too high index

option premiums and high profits on single stock options, e.g., Goldman Sachs Group, Inc. and United Rentals, Inc.

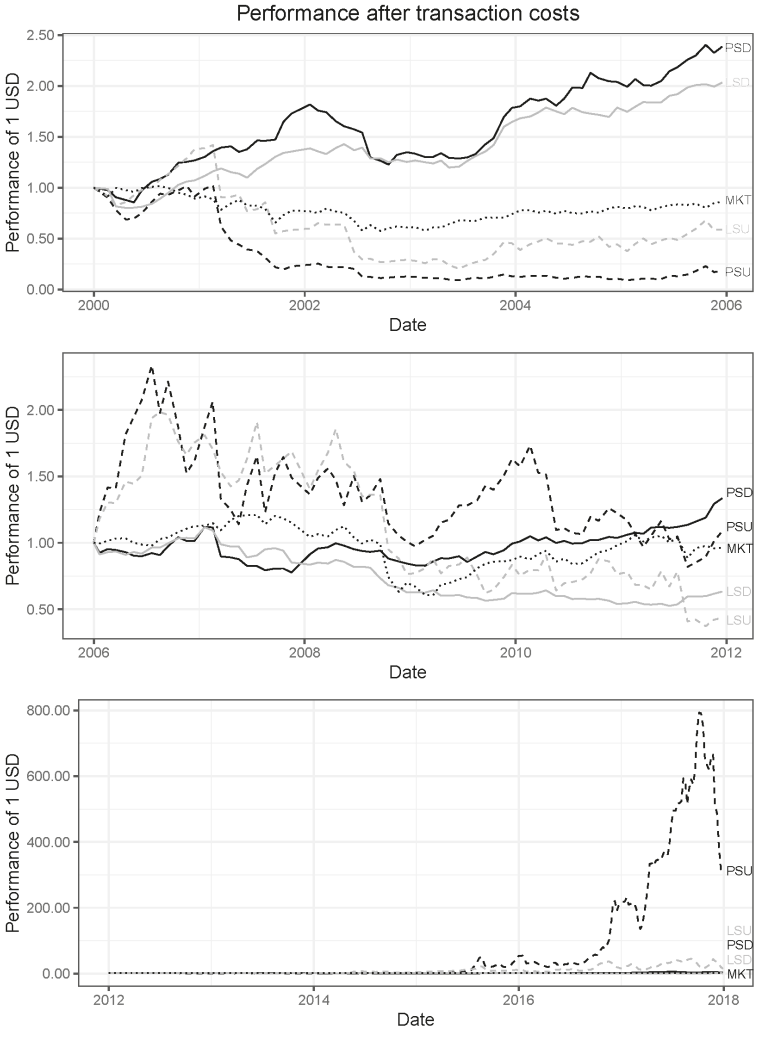


Figure 3. Development of an investment of 1 USD in PSD, PSU, LSD, and LSU after transaction costs compared to the S&P 500 (MKT). The time period is divided intro three sub-periods from January 2000–December 2005, January 2006–December 2011, and January 2012–December 2017. Source: authors' calculations.

4.3. Market Frictions

This subsection evaluates the robustness of our trading strategies in light of market frictions. Following [54], we analyze the annualized mean return and Sharpe ratio for varying transaction cost levels (see Table 5). Motivated by the literature, our back-testing study supposes transaction costs of 10 bps per straddle and 2.5 bps per stock for delta hedging. The results for 0 bps and 10 bps are identical

to Table 4. Due to the same trading frequencies, all strategies are similarly affected by transaction costs; the naive buy-and-hold strategy MKT is constant (see Section 4.1). Taking annualized returns into account, we observe that the breakeven point for PSD and LSU is between 40 bps and 80 bps. As expected, PSU has a higher level where costs and returns are equal; the breakeven point is around 90 bps. The additional consideration of the risk side leads to the Sharpe ratio, with the results being similar to before. As expected, the breakeven points are reached for lower transactions costs, as the Sharpe ratio is calculated on the basis of excess returns. The thresholds vary between approximately 5 bps for LSD and 90 bps for PSU. Concluding, the delta hedged strategies PSD and PSU provide promising results in the context of risk-return measures, even for investors that are exposed to different market conditions and thus higher transaction costs.

Table 5. Annualized mean return and Sharpe ratio for PSD, PSU, LSD, and LSU from January 2000 until December 2017 for different transaction costs per trade per straddle in bps.

		Annualized Return					**Sharpe Ratio**				
		PSD	**PSU**	**LSD**	**LSU**	**MKT**	**PSD**	**PSU**	**LSD**	**LSU**	**MKT**
Transaction costs	0	0.1841	0.2988	0.0401	0.1049	0.0347	0.5157	0.3897	0.0599	0.1334	0.0890
	1	0.1802	0.2954	0.0368	0.1022	0.0347	0.5037	0.3851	0.0504	0.1292	0.0890
	5	0.1645	0.2818	0.0238	0.0913	0.0347	0.4560	0.3670	0.0126	0.1126	0.0890
	10	0.1452	0.2651	0.0077	0.0779	0.0347	0.3967	0.3446	−0.0344	0.0918	0.0890
	15	0.1262	0.2485	−0.0080	0.0647	0.0347	0.3380	0.3222	−0.0809	0.0712	0.0890
	20	0.1075	0.2321	−0.0236	0.0516	0.0347	0.2798	0.3001	−0.1271	0.0508	0.0890
	40	0.0356	0.1687	−0.0836	0.0007	0.0347	0.0521	0.2130	−0.3082	−0.0299	0.0890
	50	0.0014	0.1381	−0.1122	−0.0238	0.0347	−0.0589	0.1703	−0.3967	−0.0696	0.0890
	60	−0.0318	0.1083	−0.1400	−0.0478	0.0347	−0.1681	0.1283	−0.4839	−0.1088	0.0890
	90	−0.1252	0.0232	−0.2185	−0.1166	0.0347	−0.4852	0.0054	−0.7380	−0.2238	0.0890
	100	−0.1544	−0.0038	−0.2431	−0.1384	0.0347	−0.5876	−0.0345	−0.8203	−0.2613	0.0890

Source: authors' calculations.

4.4. Robustness Check

As stated previously, the 20 percent reinvestment rate, the top stocks option portfolio, which is traded unconditionally on correlation, and the number of five target stocks was motivated based on the existing literature (Section 3.3). Since data snooping is an important issue in many research studies, we examine the sensitivity of our PSD results with respect to variations of these hyperparameters.

In contrast to the reinvestment rate and the number of target stocks, trading based on unconditional correlation is not a hyperparameter in our framework. However, it is beneficial to examine if changes in the entry signal lead to substantial changes in our strategy performance. In Section 2.1, the concept of measuring implied costs as average implied correlation is introduced. Based on this approach, the relative mispricing between the index and the replicating basket can be quantified. Applying Equation (7) to forecasted or historical volatility yields an average expected correlation. Comparing this metric to the implied average correlation provides insights regarding the current pricing of index options. As discussed in Section 2.1, index options trade usually richer in terms of implied volatility than the replication portfolio. However, there are times at which the opposite might be the case. Following, an investors would set up a short dispersion trade: selling the basket and buying the index. A simple trading signal can be derived from average correlation levels:

$$\begin{aligned} &\text{Long dispersion if} \quad \overline{\rho}_{implied} > \overline{\rho}_{historical,} \\ &\text{Short dispersion if} \quad \overline{\rho}_{implied} < \overline{\rho}_{historical.} \end{aligned}$$

For our robustness check, we rely on historical volatility as a baseline approach. However, more advanced statistical methods could be applied to forecast volatility and determine current pricing levels (see [55–57]). Table 6 reports the annualized return of PSD for a variety of replicating portfolio sizes, reinvestment rates, and correlation based entry signals. First of all, a higher reinvestment rate leads to higher annualized mean returns; concurrently, higher risk aggravates the Sharpe ratio and

the maximum drawdown. We observe several total losses for reinvestment rates of 80 percent and 100 percent, which is a result of the corresponding all-or-nothing strategy. Higher annualized returns and Sharpe ratios can generally be found at lower numbers of top stocks indicating that our selection algorithm introduced in Section 3.2 is meaningful. Regarding the correlation filter, we recognize that the strategy based on conditional correlation leads to worse results than the unconditional counterpart, e.g., the annualized mean return for the top five stocks and an investment rate of 20 percent is 9.53 percent (conditional correlation) vs. 18.41 percent (unconditional correlation). Summarizing, our initial hyperparameter setting does not hit the optimum; our selection procedure identifies the right top stocks; and considering unconditional correlation has a positive impact on the trading results.

Table 6. Annualized mean return, Sharpe ratio, and maximum drawdown before transaction costs for a varying number of top stocks, the amount of reinvestment, and the correlation filter from January 2000 until December 2017.

	Reinvestment	**5%**	**10%**	**20%**	**40%**	**80%**	**100%**
Mean return							
Top 5	Uncond.Correlation	0.07120	0.11478	0.18411	0.23003	NA	NA
	Cond.Correlation	0.05096	0.07288	0.09532	0.03780	NA	NA
Top 10	Uncond. Correlation	0.06976	0.11226	0.18090	0.23413	−0.11755	NA
	Cond. Correlation	0.04677	0.06481	0.08092	0.02113	NA	NA
Top 20	Uncond. Correlation	0.07394	0.12131	0.20171	0.28533	0.00145	−0.59641
	Cond. Correlation	0.05318	0.07825	0.11008	0.08608	NA	NA
Sharpe ratio							
Top 5	Uncond. Correlation	0.6467	0.5969	0.5157	0.3298	NA	NA
	Cond. Correlation	0.3902	0.3321	0.2361	0.0285	NA	NA
Top 10	Uncond. Correlation	0.6492	0.6004	0.5226	0.3475	−0.1109	NA
	Cond. Correlation	0.3496	0.2912	0.1976	0.0027	NA	NA
Top 20	Uncond. Correlation	0.7226	0.6774	0.6064	0.4424	−0.0150	−0.4100
	Cond. Correlation	0.4432	0.3879	0.2994	0.1101	NA	NA
Maximum drawdown							
Top 5	Uncond. Correlation	0.2146	0.4017	0.6731	0.9258	1.0000	1.0000
	Cond. Correlation	0.2520	0.4943	0.8124	0.9908	1.0000	1.0000
Top 10	Uncond. Correlation	0.2363	0.4332	0.7053	0.9447	1.0000	1.0000
	Cond. Correlation	0.2898	0.5403	0.8399	0.9920	1.0000	1.0000
Top 20	Uncond. Correlation	0.2282	0.4198	0.6885	0.9294	0.9999	1.0000
	Cond. Correlation	0.2744	0.5187	0.8232	0.9897	1.0000	1.0000

Source: authors' calculations.

4.5. Risk Factor Analysis

Table 7 evaluates the exposure of PSD and PSU after transaction costs to systematic sources of risk (see [48]). Therefore, we apply the Fama–French three factor model (FF3) and the Fama–French 5-factor model (FF5) of [58,59], and the Fama–French 3+2 factor model (FF3+2) presented by [2]. Hereby, FF3 measures the exposure to the general market, small minus big capitalization stocks (SMB), and high minus low book-to-market stocks (HML). Next, FF3+2 enlarges the first model by a momentum factor and a short-term reversal factor. Finally, FF5 extends FF3 by a factor capturing robust minus weak (RMW5) profitability and a factor capturing conservative minus aggressive (CMA5) investment behavior.

Across all three models, we observe statistically significant monthly alphas ranging between 0.74% and 0.79% for PSD and between 2.01% and 2.19% for PSU. The highest explanatory content is given for FF3+2 with adjusted R^2 of 0.0142 and 0.0118; probably, the momentum and reversal factor possess a high explanatory power. Not significant loadings for SMB5, HML5, RMW5, and CMA5 confirm our long–short portfolio we are constructing. Exposure to HML (PSD) and the reversal factor (PSU) underlies our selection and trading process (see Section 3).

Table 7. Risk factor exposure of PSD and PSU. The standard error is reported in brackets. FF3, Fama–French three factor model; SMB, small minus big capitalization stocks; HML, high minus low book-to-market stocks; RMW, robust minus weak; CMA, conservative minus aggressive.

	PSD			**PSU**		
	FF3	**FF3+2**	**FF5**	**FF3**	**FF3+2**	**FF5**
Intercept	0.0075 **	0.0079 **	0.0074 **	0.0201 **	0.0222 ***	0.0219 ***
	(0.0035)	(0.0035)	(0.0035)	(0.0078)	(0.0080)	(0.0080)
Market	0.1200	0.1316	0.1517 *	0.2128	0.4157 **	0.0560
	(0.0834)	(0.0926)	(0.0909)	(0.1894)	(0.2084)	(0.2056)
SMB	0.1626	0.1736		0.1325	0.1920	
	(0.1359)	(0.1374)		(0.3085)	(0.3092)	
HML	0.2424 **	0.2360 **		0.0117	0.0767	
	(0.1074)	(0.1146)		(0.2440)	(0.2579)	
Momentum		−0.0225			0.0711	
		(0.0893)			(0.2008)	
Reversal		–0.0656			−0.5416 **	
		(0.0968)			(0.2178)	
SMB5			0.1168			0.0847
			(0.1551)			(0.3506)
HML5			0.1383			0.3636
			(0.1396)			(0.3157)
RMW5			−0.0216			−0.2720
			(0.1815)			(0.4102)
CMA5			0.2194			−0.8783
			(0.2362)			(0.5340)
R^2	0.0256	0.0267	0.0264	0.0050	0.0243	0.0146
Adj.R^2	0.0181	0.0142	0.0139	−0.0026	0.0118	0.0019
No. obs.	395	395	395	395	395	395
RMSE	0.0672	0.0674	0.0674	0.1527	0.1516	0.1523

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Source: authors' calculations.

4.6. Analysis of PCA Components and Market Exposure

Following [60], we report the Kaiser–Meyer–Olkin criterion (KMO) and Bartlett's sphericity test in Table 8 to examine the suitability of our data for a PCA analysis. Bartlett's test of sphericity, which tests the hypothesis that the correlation matrix is an identity matrix and therefore the variables are unrelated to each other and not suitable to detect any structure, is for all of our trading dates at a significance level of 1% [61]. KMO is a sampling adequacy measure that describes the proportion of variance in the variables that might be caused by shared underlying factors [62,63]. A low KMO score is driven by high partial correlation in the underlying data. The threshold for an acceptable sampling adequacy for a factor analysis is normally considered to be 0.5 [64]. However, in our context of finding the top five stocks that describe the index as closely as possible, a low KMO value is actually favorable as it indicates that strong relationships amongst the S&P 500 constituents exist that we aim to exploit with our methodology. As expected, our KMO values are on almost all trading days below the threshold with an average and median of 0.2833 and 0.2619, respectively. There are two determinants that explain the low KMO values in our study and support our approach. (i) Stocks are often dependent on the same risk factors, one of the most prominent factors being the market (for more, see [65,66]), and therefore exhibit rather high partial correlations. (ii) On every trading day, we analyze around 450 S&P 500 index members, which results in 101,250 different partial correlations. Due to the substantial number of combinations, you will find structure and high partial correlations in the data.

Table 8. Kaiser–Meyer–Olkin (KMO) criterion and Bartlett's sphericity test results. The KMO values and *p*-values of Bartlett's test are segmented into six and three buckets, respectively.

	Kaiser-Meyer-Olkin Criterion						**Bartlett's Sphericity Test**		
	KMO Value						***p*-Value**		
Segment	0–0.1	0.1–0.2	0.2–0.3	0.3–0.5	≥ 0.5	Total	$< 1\%$	$\geq 1\%$	Total
Count	7	84	148	138	18	395	395	0	395
Average	0.0895	0.1600	0.2445	0.3725	0.5701	0.2833	0.0000	NA	0.0000
Median	0.0878	0.1668	0.2431	0.3530	0.5502	0.2619	0.0000	NA	0.0000

Source: authors' calculations.

Figure 4 reports the number of required principal components to explain 90 percent of the S&P 500 constituents' return variation and the standardized S&P 500 index. Overall, we observe synchronous developments of both time series, i.e., if one variable increases, the other increases, and vice versa. Specifically, the number of PCA components ranges between 15 and 45 from 2000 until 2007. The value decreases to approximately five with the beginning of the financial crisis in 2008. This fact is not surprising since downsides of the S&P 500 lead to a higher pair-wise correlation of the stock constituents. Consequently, a lower number of shares is in a position to describe 90 percent of stock price variations. After 2011, the general market possesses a positive trend without any strong swings. Thus, the number of PCA components increases to approximately 20.

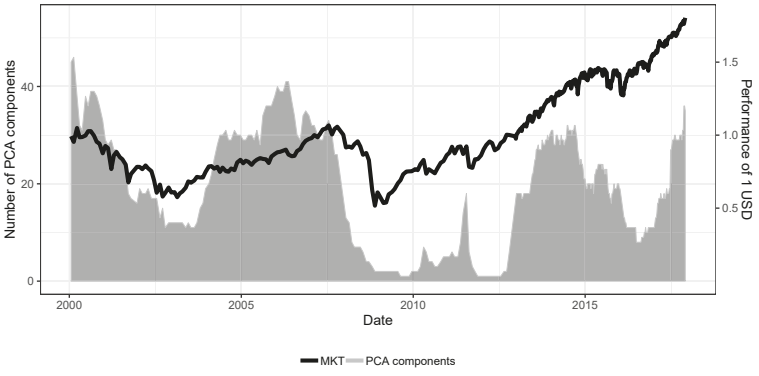


Figure 4. Number of required principal components to explain 90% of S&P 500 constituents return variation and S&P 500 index performance (MKT) from January 2000–December 2017. Source: authors' calculations.

Last, but not least, Figure 5 shows the yearly sector exposure based on our PCA selection process from January 2000 to December 2017. In line with the Standard Industrial Classification, all companies are categorized into the following nine economic sectors: "Mining", "Manufacturing", "Wholesale Trade", "Finance, Insurance, Real Estate", "Construction", "Transportation and Public Utilities", "Real Trade", "Services". Each point on the horizontal axis refers to the sectors that are used in the year after to the respective point. First of all, we observe that the sum of positive sector exposures increases in times of high market turmoil. Especially, the cumulative value exceeds three in 2013 because of a high exposure in "Finance, Insurance, Real Estate", especially for Bank of America Corporation and Goldman Sachs Group, Inc. Furthermore, the exposure of each sector varies over time, i.e., industry branches are preferred and avoided in times of bull and bear markets. As such, the percentage of financial stocks increases in 2009 and 2013, the peak of the global financial and European debt crisis.

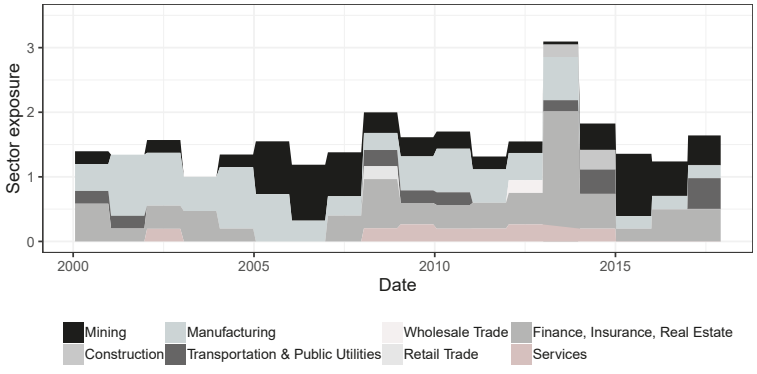


Figure 5. Yearly sector exposure based on the PCA selection process from January 2000–December 2017. Sector short positions are for illustrative purposes not included. Source: authors' calculations.

5. Conclusions and Policy Recommendations

In this manuscript, we developed a dispersion trading strategy based on a statistical stock selection process and applied our approach to the S&P 500 index and its constituents from January 2000 to December 2017. We contributed to the existing literature in four ways. First, we developed an index subsetting procedure that considers the individual index explanatory power of stocks in the weighting scheme. Therefore, we are in a position to build a replicating option basket with as little as five securities. Second, the large-scale empirical study provides a reliable back-testing for our dispersion trades. Hence, the profitability and robustness of those relative value trades can be examined across a variety of market conditions. Third, we analyzed the added value of our strategies by benchmarking them against a naive index subsetting approach and a simple buy-and-hold strategy. The trading frameworks that employ the PCA selection process outperformed its peers with an annualized mean return of 14.52 and 26.51 percent for PSD and PSU, respectively. The fourth contribution focuses on the conducted deep dive analysis of our selection process, i.e., sector exposure and number of required principal components over time, and the robustness checks. We showed that our trading systems possess superior risk-return characteristics compared to the benchmarking dispersion strategies.

Our study reveals two main policy recommendations. First of all, our framework shows that advanced statistical methods can be utilized to determine a portfolio replicating basket and could therefore be used for sophisticated risk management. Regulatory market risk assessments of financial institutions often rely on crude approaches that stress bank's capital requirements disproportionately to the underlying risk. However, regulators should explore the possibility of employing more advanced statistical models in their risk assessments, such as considering PCA built replicating baskets to hedge index exposures, to adequately reflect risk. Finally, investors should be aware that a principal component analysis can be used to cost-efficiently set up dispersion trades, and they should use their comprehensive datasets to improve the stock selection process further.

For future research endeavors, we identify the following three areas. First, the back-testing framework should be applied to other indices and geographical areas, i.e., price weighted equity markets and emerging economies, to shed light on any idiosyncrasies related to geographical or index construction differences. Second, efforts could be undertaken to improve the correlation filter from Section 4.4, so that profitable short dispersion opportunities can be spotted more accurately. Third, financial product innovations in a dispersion context should be subject to future studies, particularly the third generation of volatility derivatives such as variance, correlation, and gamma swaps.

Author Contributions: L.S. conceived of the research method. The experiments are designed and performed by L.S. The analyses were conducted and reviewed by L.S. and J.S. The paper was initially drafted and revised by L.S. and J.S. It was refined and finalized by L.S. and J.S. All authors read and agreed to the published version of the manuscript.

Funding: We are grateful to the "Open Access Publikationsfonds", which covered 75 percent of the publication fees.

Acknowledgments: We are further grateful to four anonymous referees for many helpful discussions and suggestions on this topic.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pole, A. *Statistical Arbitrage: Algorithmic Trading Insights and Techniques*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
2. Gatev, E.; Goetzmann, W.N.; Rouwenhorst, K.G. Pairs trading: Performance of a relative-value arbitrage rule. *Rev. Financ. Stud.* **2006**, *19*, 797–827. [CrossRef]
3. Do, B.; Faff, R. Does simple pairs trading still work? *Financ. Anal. J.* **2010**, *66*, 83–95. [CrossRef]
4. Ramos-Requena, J.; Trinidad-Segovia, J.; Sánchez-Granero, M. Introducing Hurst exponent in pair trading. *Phys. Stat. Mech. Its Appl.* **2017**, *488*, 39–45. [CrossRef]
5. Ramos-Requena, J.P.; Trinidad-Segovia, J.E.; Sánchez-Granero, M.Á. Some notes on the formation of a pair in pairs trading. *Mathematics* **2020**, *8*, 348. [CrossRef]
6. Sánchez-Granero, M.; Balladares, K.; Ramos-Requena, J.; Trinidad-Segovia, J. Testing the efficient market hypothesis in Latin American stock markets. *Phys. Stat. Mech. Its Appl.* **2020**, *540*, 123082. [CrossRef]
7. Bennett, C. *Trading Volatility: Trading Volatility, Correlation, Term Structure and Skew*; CreateSpace: Charleston, SC, USA, 2014.
8. Nasekin, S.; Härdle, W.K. Model-driven statistical arbitrage on LETF option markets. *Quant. Financ.* **2019**, *19*, 1817–1837. [CrossRef]
9. Glasserman, P.; He, P. Buy rough, sell smooth. *Quant. Financ.* **2020**, *20*, 363–378. [CrossRef]
10. Gangahar, A. Smart Money on Dispersion. Financial Times. 2006. Available online: https://www.ft.com/content/a786ce1e-3140-11db-b953-0000779e2340 (accessed on 16 August 2020)
11. Sculptor Capital Management Inc. Sculptor Capital Management Inc. website. 2020. Available online: https://www.sculptor.com/ (accessed on 16 August 2020)
12. Alloway, T. A $12bn Dispersion Trade. Financial Times. 2011. Available online: https://ftalphaville.ft.com/2011/06/17/597511/a-12bn-dispersion-trade/ (accessed on 16 August 2020)
13. Marshall, C.M. Dispersion trading: Empirical evidence from U.S. options markets. *Glob. Financ. J.* **2009**, *20*, 289–301. [CrossRef]
14. Maze, S. Dispersion trading in south africa: An analysis of profitability and a strategy comparison. *SSRN Electron. J.* **2012**. . [CrossRef]
15. Ferrari, P.; Poy, G.; Abate, G. Dispersion trading: An empirical analysis on the S&P 100 options. *Invest. Manag. Financ. Innov.* **2019**, *16*, 178–188.
16. Deng, Q. Volatility dispersion trading. *SSRN Electron. J.* **2008**. . [CrossRef]
17. Wilmott, P. *Paul Wilmott Introduces Quantitative Finance*, 2nd ed.; John Wiley: Chichester, UK, 2008.
18. Black, F.; Scholes, M. The pricing of options and corporate liabilities. *J. Political Econ.* **1973**, *81*, 637–654. [CrossRef]
19. Jiang, X. Return dispersion and expected returns. *Financ. Mark. Portf. Manag.* **2010**, *24*, 107–135. [CrossRef]
20. Chichernea, S.C.; Holder, A.D.; Petkevich, A. Does return dispersion explain the accrual and investment anomalies? *J. Account. Econ.* **2015**, *60*, 133–148. [CrossRef]
21. Andersen, T.G.; Bollerslev, T. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *Int. Econ. Rev.* **1998**, *39*, 885. [CrossRef]
22. Andersen, T.G.; Bollerslev, T.; Diebold, F.X.; Labys, P. The distribution of realized exchange rate volatility. *J. Am. Stat. Assoc.* **2001**, *96*, 42–55. [CrossRef]
23. Barndorff-Nielsen, O.E.; Shephard, N. Estimating quadratic variation using realized variance. *J. Appl. Econom.* **2002**, *17*, 457–477. [CrossRef]
24. Markowitz, H. Portfolio selection. *J. Financ.* **1952**, *7*, 77.

25. Driessen, J.; Maenhout, P.J.; Vilkov, G. Option-implied correlations and the price of correlation risk. *SSRN Electron. J.* **2013**. [CrossRef]
26. Faria, G.; Kosowski, R.; Wang, T. The correlation risk premium: International evidence. *SSRN Electron. J.* **2018**. [CrossRef]
27. Bollen, N.P.B.; Whaley, R.E. Does net buying pressure affect the shape of implied volatility functions? *J. Financ.* **2004**, *59*, 711–753. [CrossRef]
28. Shiu, Y.M.; Pan, G.G.; Lin, S.H.; Wu, T.C. Impact of net buying pressure on changes in implied volatility: before and after the onset of the subprime crisis. *J. Deriv.* **2010**, *17*, 54–66. [CrossRef]
29. Ruan, X.; Zhang, J.E. The economics of the financial market for volatility trading. *J. Financ. Mark.* **2020**, 100556. [CrossRef]
30. Bakshi, G.; Kapadia, N. Delta-hedged gains and the negative market volatility risk premium. *Rev. Financ. Stud.* **2003**, *16*, 527–566. [CrossRef]
31. Ilmanen, A. Do financial markets reward buying or selling insurance and lottery tickets? *Financ. Anal. J.* **2012**, *68*, 26–36. [CrossRef]
32. Gatheral, J. *The Volatility Surface: A Practitioner's Guide*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2006.
33. Crépey, S. Delta-hedging vega risk? *Quant. Financ.* **2004**, *4*, 559–579. [CrossRef]
34. Jegadeesh, N.; Titman, S. Returns to buying winners and selling losers: Implications for stock market efficiency. *J. Financ.* **1993**, *48*, 65–91. [CrossRef]
35. R Core Team. *Stats: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2019.
36. Do, B.; Faff, R. Are pairs trading profits robust to trading costs? *J. Financ. Res.* **2012**, *35*, 261–287. [CrossRef]
37. Stübinger, J.; Schneider, L. Statistical arbitrage with mean-reverting overnight price gaps on high-frequency data of the S&P 500. *J. Risk Financ. Manag.* **2019**, *12*, 51.
38. Korajczyk, R.A.; Sadka, R. Are momentum profits robust to trading costs? *J. Financ.* **2004**, *59*, 1039–1082. [CrossRef]
39. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
40. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
41. Su, X. Hedging Basket Options by Using A Subset of Underlying Assets (Working Paper). 2006. Available online: https://www.econstor.eu/bitstream/10419/22959/1/bgse14_2006.pdf (accessed on 16 August 2020).
42. von Beschwitz, B.; Massa, M. Biased short: Short sellers' disposition effect and limits to arbitrage. *J. Financ. Mark.* **2020**, *49*, 100512. [CrossRef]
43. Voya Investment Management. The Impact of Equity Market Fragmentation and Dark Pools on Trading and Alpha Generation. 2016. Available online: https://investments.voya.com (accessed on 16 August 2020).
44. Frazzini, A.; Israel, R.; Moskowitz, T.J. Trading costs. *SSRN Electron. J.* **2018**. [CrossRef]
45. Henderson, R. Schwab Opens New front in Trading War by Slashing Rates to Zero. Financial Times. 2019. Available online: https://www.ft.com/content/cf644610-e45a-11e9-9743-db5a370481bc (accessed on 16 August 2020).
46. Avellaneda, M.; Lee, J.H. Statistical arbitrage in the US equities market. *Quant. Financ.* **2010**, *10*, 761–782. [CrossRef]
47. Cboe Global Markets, I. Chicago Board Options Exchange Margin Manual. 2000. Available online: https://www.cboe.com/learncenter/pdf/margin2-00.pdf (accessed on 16 August 2020).
48. Stübinger, J.; Endres, S. Pairs trading with a mean-reverting jump-diffusion model on high-frequency data. *Quant. Financ.* **2018**, *18*, 1735–1751. [CrossRef]
49. Mina, J.; Xiao, J.Y. Return to RiskMetrics: The evolution of a standard. *Riskmetrics Group* **2001**, 1, 1–11.
50. Stübinger, J.; Mangold, B.; Krauss, C. Statistical arbitrage with vine copulas. *Quant. Financ.* **2018**, *18*, 1831–1849. [CrossRef]
51. Liu, B.; Chang, L.B.; Geman, H. Intraday pairs trading strategies on high frequency data: The case of oil companies. *Quant. Financ.* **2017**, *17*, 87–100. [CrossRef]
52. Knoll, J.; Stübinger, J.; Grottke, M. Exploiting social media with higher-order factorization machines: Statistical arbitrage on high-frequency data of the S&P 500. *Quant. Financ.* **2019**, *19*, 571–585.

53. Stübinger, J. Statistical arbitrage with optimal causal paths on high-frequency data of the S&P 500. *Quant. Financ.* **2019**, *19*, 921–935.
54. Endres, S.; Stübinger, J. Regime-switching modeling of high-frequency stock returns with Lévy jumps. *Quant. Financ.* **2019**, *19*, 1727–1740. [CrossRef]
55. Degiannakis, S.; Filis, G.; Hassani, H. Forecasting global stock market implied volatility indices. *J. Empir. Financ.* **2018**, *46*, 111–129. [CrossRef]
56. Fang, T.; Lee, T.H.; Su, Z. Predicting the long-term stock market volatility: A GARCH-MIDAS model with variable selection. *J. Empir. Financ.* **2020**, *58*, 36–49. [CrossRef]
57. Naimy, V.; Montero, J.M.; El Khoury, R.; Maalouf, N. Market volatility of the three most powerful military countries during their intervention in the Syrian War. *Mathematics* **2020**, *8*, 834. [CrossRef]
58. Fama, E.F.; French, K.R. Multifactor explanations of asset pricing anomalies. *J. Financ.* **1996**, *51*, 55–84. [CrossRef]
59. Fama, E.F.; French, K.R. A five-factor asset pricing model. *J. Financ. Econ.* **2015**, *116*, 1–22. [CrossRef]
60. Yoshino, N.; Taghizadeh-Hesary, F. Analysis of credit ratings for small and medium-sized enterprises: Evidence from Asia. *Asian Dev. Rev.* **2015**, *32*, 18–37. [CrossRef]
61. Bartlett, M.S. The effect of standardization on a χ^2 approximation in factor analysis. *Biometrika* **1951**, *38*, 337–344. [CrossRef]
62. Kaiser, H.F. A second generation little jiffy. *Psychometrika* **1970**, *35*, 401–415. [CrossRef]
63. Kaiser, H. An index of factorial simplicity. *Psychometrika* **1974**, *39*, 31–36. [CrossRef]
64. Kaiser, H.F.; Rice, J. Little jiffy, Mark Iv. *Educ. Psychol. Meas.* **1974**, *34*, 111–117. [CrossRef]
65. Sharpe, W.F. Capital asset prices: A theory of market equilibrium under conditions of risk. *J. Financ.* **1964**, *19*, 425–442.
66. Lintner, J. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Rev. Econ. Stat.* **1965**, *47*, 13–37. [CrossRef]

mathematics

MDPI

Article

Non-Parametric Analysis of Efficiency: An Application to the Pharmaceutical Industry

Ricardo F. Díaz and Blanca Sanchez-Robles *

Department of Economic Analysis, Facultad CC Económicas y Empresariales, UNED, Senda del Rey 11, 28040 Madrid, Spain; rr_dos@hotmail.com

* Correspondence: bsanchez-robles@cee.uned.es

Received: 29 July 2020; Accepted: 31 August 2020; Published: 7 September 2020

Abstract: Increases in the cost of research, specialization and reductions in public expenditure in health are changing the economic environment for the pharmaceutical industry. Gains in productivity and efficiency are increasingly important in order for firms to succeed in this environment. We analyze empirically the performance of efficiency in the pharmaceutical industry over the period 2010–2018. We work with microdata from a large sample of European firms of different characteristics regarding size, main activity, country of origin and other idiosyncratic features. We compute efficiency scores for the firms in the sample on a yearly basis by means of non-parametric data envelopment analysis (DEA) techniques. Basic results show a moderate average level of efficiency for the firms which encompass the sample. Efficiency is higher for companies which engage in manufacturing and distribution than for firms focusing on research and development (R&D) activities. Large firms display higher levels of efficiency than medium-size and small firms. Our estimates point to a decreasing pattern of average efficiency over the years 2010–2018. Furthermore, we explore the potential correlation of efficiency with particular aspects of the firms' performance. Profit margins and financial solvency are positively correlated with efficiency, whereas employee costs display a negative correlation. Institutional aspects of the countries of origin also influence efficiency levels.

Keywords: pharmaceutical industry; scale economies; profitability; biotechnological firms; non-parametric efficiency; productivity; DEA

JEL Classification: I15; O32; L6

1. Introduction

Pharmaceutical companies contribute crucially to the health and welfare of individuals. This issue is particularly relevant nowadays: as the Covid-19 pandemic has shown, no country is immune to the emergence of new diseases. Furthermore, the population in many countries is experiencing deep demographic transformations which increase life expectancy and raise new challenges for policymakers. Not surprisingly, the performance of the industry directly affects some of the Sustainable Development Goals of the 2030 Agenda for Sustainable Development.

The economic importance of the industry is also paramount. The pharmaceutical sector employs highly skilled labor and exhibits one of the largest figures of research and development (R&D) intensity (defined as expenditure in R&D as a share of sales). As recent contributions in the field of macroeconomics have shown, human capital and R&D are key drivers of economic growth, productivity and prosperity [1–3].

The pharmaceutical industry is facing new challenges because of several factors. New diseases as the Covid-19 demand quick, pathbreaking solutions. R&D costs grow because conditions become chronic and more complicated. Paradoxically, the progress in molecular biology which increases the range of potential innovations also raises the complexity of decisions related to the R&D strategy.

New investments seek increasingly *high risk/high premium* drugs [4]. Official agencies accumulate requirements for drug approvals. Firms must cope with the expiration of patents and with reductions in public expenditure in healthcare due to stability measures and fiscal adjustments.

Meanwhile the business model in the industry has experienced deep transformations over the last decades. Some firms have specialized in particular steps of the value chain, as R&D in the biotechnological sphere or clinical research, this last in the case of contract research organizations (CROs). Reductions in R&D productivity have brought about mergers and acquisitions, partly to profit from the expertise in research and the pipeline of other companies. Reference [5] argue that Japanese firms engaged in mergers and acquisitions over 1980–1997 to handle the declining productivity of R&D. Other firms outsource activities or engage in technological alliances [6,7]. In this context, firms must strive to increase their levels of productivity and efficiency, which may become a strategic asset [8].

In parallel, empirical research on productivity and efficiency (defined as output per unit of inputs) has grown over the last decades. Mathematical techniques such as data envelopment analysis (DEA) have facilitated the empirical assessment of efficiency at the country, entity or firm level. The literature has explored the levels and trends of efficiency in many activities and areas such as banking [9], farming [10], food [11], universities [12], airlines [13], shipping [14], oil [15], electricity distribution [16,17] and energy consumption [18,19], to quote just a few examples.

Recent meta-analyses and compilations of DEA exercises can be found in [20] for the public sector, [21] for energy and the environment, [22] for seaports, [23] for microfinance institutions and [24] for rail transport. [25] provide a thorough list of the main journal articles on DEA methodology and applications published between 1978 and 2016.

Researchers have also dealt with more theoretical aspects of the DEA model. Examples are [26], which describes a dynamic version of DEA that allows intertemporal links between inputs and outputs to be considered, and [27] which provides an alternative to the inverse DEA model. Furthermore, [28] explore the features of the model when the data are imprecise and [29] devise a DEA algorithm suitable to deal with Big Data.

The analysis of efficiency in the pharmaceutical industry has also been addressed in the recent past [8,30] although the number of contributions in this regard is comparatively sparse. Most of the studies in this area perform their analyses at the country level and/or focus on a (usually small) sample of companies. Examples are [31] for China; [32,33] for Japan; [6] for US; [34] for Jordan and [35] for India.

We intend to complement this literature with a two-stages analysis of efficiency within a relatively large sample of European firms. In the first stage we compute efficiency levels for the firms in our sample. In the second stage we explore by statistical modelling the connection between the efficiency scores obtained in the first stage and a set of variables potentially correlated with efficiency.

We are especially interested in the assessment of efficiency by type of activity and firm size. More specifically, we want to explore whether large firms exhibit higher levels of efficiency, which would be consistent with the potential presence of scale economies in the industry. Furthermore, it is feasible that firms which primarily operate in the R&D niche enjoy a different level of efficiency, on average, than companies with activities along the entire value chain. Finally, we want to explore the data to find common patterns and detect possible features of the economic and institutional framework and firm management strategy which can be correlated with efficiency.

In parallel, our empirical exercise may prove useful to illustrate how to apply modern mathematical, non-parametric techniques in order to get insights about the performance of firms in a particular industry, and how these tools are related to more traditional, parametric approaches.

Our paper is closely related to three DEA explorations of the pharmaceutical industry: [6,8,35].

Reference [8] analyze efficiency in a sample of 37 large firms from different countries over 2008–2013. They report an average level of efficiency in their sample of 0.9345 and find that firms with higher level of efficiency carry out more financial transactions with other companies.

We complement this exploration in several dimensions. First, our sample is different, broader and more heterogeneous, since it encompasses a large group of European firms, of different sizes and profiles. Second, we report an average efficiency score of 0.34. We think that this figure is a more accurate reflection of the mean efficiency for the whole industry, at least for the European case.

Third, we carry out a two-stage exploration of efficiency whereby in the second stage we look at variables potentially correlated with the efficiency levels obtained in the first stage. Reference [8] omit the second stage because it is somehow controversial. It is true that the literature has not reached a consensus yet on the right specification for the second stage; nonetheless, we think that this analysis can still provide some valid insights about efficiency.

Fourth, we work with a more recent time horizon, 2010–2018, and examine the dynamic performance of efficiency over time; they look at data from 2008–2013 but perform their analysis on average terms, so they do not uncover the pattern of efficiency over time.

Another related investigation is [6]. They employ proprietary data from a sample encompassed by 700 US pharmaceutical firms over the period 2001–2016. They assess the connection between open innovation methods and efficiency.

Reference [35] utilize data from a financial database to examine the performance of a group of Indian firms over the years 1991–2005. They perform a two-stage analysis. In the second stage they examine the determinants of efficiency in their sample by regression tools.

In contrast to [6,35], we work with a sample made up of European firms and explore the potential impact of alternative aspects of firm management and country characteristics. While [35] employ only a Tobit specification in the second stage of their analysis, we utilize also a pure random-effects and a Simar–Wilson procedure, and perform a comparison of the three methods.

We contribute to the literature in several ways. To the best of our knowledge, we are the first to perform a DEA analysis for a relatively large sample of European pharmaceutical firms, of different sizes and main activities, fully exploiting the time dimension of the data.

The inclusion of biotechnological companies in our sample and the exploration of their specific performance are also novel features of our investigation.

We introduce in the second stage of our empirical work a set of variables potentially correlated with efficiency, capturing different aspects of firm management and the macroeconomic environment where companies operate. Employing these variables is original as well in these kinds of analysis.

Finally, we compare the results for the second stage of three different estimation procedures (Tobit, pure random-effects, Simar–Wilson [36]). While the estimates yielded by the Tobit and the pure random-effects specifications are rather close, the Simar–Wilson tool provides larger point estimates. Nonetheless, the quantification of the marginal effects of the main covariates are more similar, and therefore the Simar–Wilson method may also be useful in applied research.

Our investigation suggests that the level of efficiency in the European pharmaceutical industry is moderate and has displayed a decreasing trend over the period 2010–2018. We find a connection between size and efficiency for the firms in our sample, where larger and very small firms tend to perform better as far as efficiency is concerned. Instead, efficiency is smaller for medium and small firms.

In terms of activity, companies operating over the complete value chain register higher levels of efficiency than firms that specialize in the R&D area. Moreover, the geographical market where firms operate seems to matter for their efficiency. Higher margins, sound financial management and lower levels of employee cost are also positively correlated with efficiency according to our results.

The structure of this paper is the following: Section 2 describes the theoretical background of our investigation. Section 3 describes the data and empirical strategy pursued. Sections 4 and 5 discuss the main results of our analysis and Section 6 concludes.

2. Theoretical Background

Conventional microeconomic theory assumes that firms optimize by producing the maximum possible quantity of output for a given input endowment or, equivalently, by producing a given amount of output with the minimum feasible inputs; this is tantamount to presupposing that they are efficient.

Empirical evidence and casual observation suggest that this is not necessarily the case. Inefficiencies exist and may arise due to managerial practices [37] or cultural beliefs [38]. Moreover, some features of the macroeconomic environment where companies operate, as information asymmetries or market rigidities, may also be detrimental for firms' productivity, as some important breakthroughs in macroeconomics in the last decades have pointed out.

Modern applied research pursues productivity analyses through two main avenues: stochastic frontier analysis (SFA) and DEA. While the intuition of both approaches is similar, the procedures are different.

In both cases the starting point is the idea of an efficient combination of inputs and outputs which encompasses a production function or *frontier*. The units of analysis are the so-called *decision-making units* or DMUs, i.e., the firms, organizations, institutions etc. whose efficiency is explored. The main difference between SFA and DEA lies in their methodology. SFA estimates the (continuous) production function by statistical techniques; DEA fits a piecewise hull enveloping the data which is assumed to approximate the true frontier, without making any statistical assumption about the data-generating process.

SFA originated with the pathbreaking contributions of [39,40]. In this setting, deviations from the estimated production function can be decomposed in statistical noise and inefficiency. Therefore, the error term in these models is usually composite [41].

An SFA model may be described by Equation (1)

$$\begin{aligned} y_i &= m(x_i;\ \beta) + \epsilon_i \\ \epsilon_i &= v_i - u_i \\ v_i &\sim N\left(0,\ \sigma_v^2\right) \\ u_i &\sim \mathcal{F} \end{aligned} \tag{1}$$

where y_i is (log) output for the ith decision-making unit or DMU, x_i is a vector of inputs for the ith DMU, ε_i the vector of parameters to be estimated, u_i captures the (one sided) inefficiency of the ith DMU and v_i represents stochastic shocks. $m(.)$ is the production function, usually assumed to be Cobb Douglas or Translog. The estimation is ordinarily implemented by maximum likelihood or other appropriate methodologies.

The stochastic shock is usually considered normal with zero mean and known variance, whereas different distributions have been advocated and estimated in the literature for the term capturing inefficiency (for a thorough review, see [41]).

The assumption about the error term may be too restrictive. Sometimes it may be preferable to work with a more flexible specification which involves fewer hypotheses. This is why non-parametric techniques, and in particular DEA, have been developed and used increasingly in recent years.

In the applied work, nonetheless, parametric and non-parametric tools sometimes intertwine: the non-parametric approach may be complemented by some statistical analyses, usually by regression procedures, which explore the output of DEA and employ inference to generalize its results to a non-deterministic setting.

Data Envelopment Analysis

The seminal paper for DEA is [42]. This technique computes efficiency by linear programming. The technique operates in two steps: first, it constructs the frontier from the data; second, it computes the distance of each unit to the frontier. It is assumed that the DMUs with the greatest efficiency determine the frontier and have efficiency of 1.

Not all efficient DMUs, however, need to be real: they can be fictitious, i.e., linear combinations of other units. This assumes, in turn, that inputs can be used continuously, i.e., they are divisible. Moreover, it presupposes that the efficiency frontier is a convex set, and hence the linear combination of two points belonging to the feasible set are also feasible. The efficient DMUs which generate a fictitious unit are called referees.

The ideas of frontier and distance encompass an intuitively appealing way to address the study of efficiency. Consider a simple example, firms from an industry which produce a single output y by means of an input x (Figure 1) (this example can be immediately generalized to the case of a vector of outputs and a vector of inputs). There are several firms or DMUs dubbed A, B, C, D, and E. The coordinates for each point in the x, y, space symbolize the input employed and the output produced by each firm. The frontier (solid line) represents *optimal* combinations of inputs and outputs. It is immediate to notice that B provides more output than A, $y_B > y_A$, while using the same amount of input since $x_A = x_B$. Alternatively, D and E produce the same output, $y_D = y_E$, but firm D consumes a smaller amount of input than E, $x_D < x_E$.

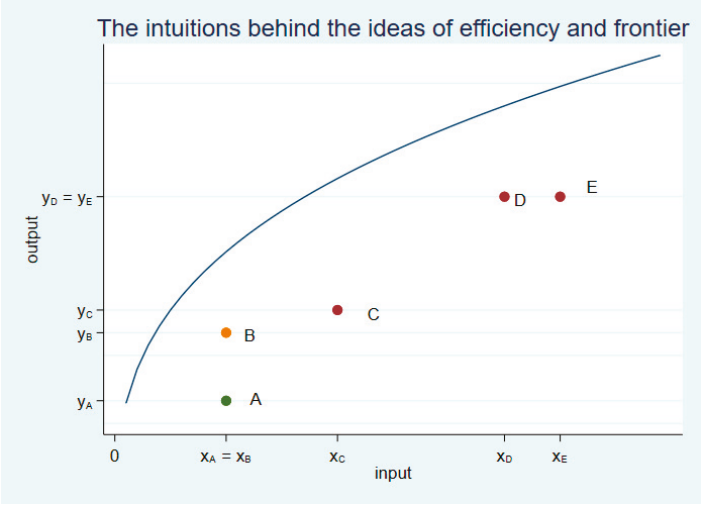


Figure 1. The intuitions behind the ideas of efficiency and frontier. Note: The figure portrays the ideas of efficiency and frontier. x is input and y is output. The concave solid line represents the technology or frontier of possibilities of production, the maximum attainable amount of output for each value of the input endowment. The dots A, B, C, D and E represent decision-making units or DMUs, i.e., firms, organizations, institutions, etc., whose efficiency is considered. Intuitively, B is more efficient than A because it produces more output than A ($y_B > Y_A$) with the same amount of input ($x_B = x_A$). Similarly, D is more efficient than E since D uses a smaller amount of input ($x_D < X_E$) to produce the same amount of output ($Y_D = Y_E$). The closer a DMU is to the frontier, the larger its level of efficiency. Source: own elaboration.

We say than B is more efficient than A and that D is more efficient than E. The closer a firm to the frontier, the larger its efficiency. Conversely, the deviations from the frontier can be understood as inefficiencies.

It is clear from Figure 1 that optimality can be defined in two alternative ways, maximum output per unit of input or minimal consumption of resources to attain a certain level of output. The first approach is named *output oriented* while the second is called *input oriented*.

Suppose there are N DMUs with a technology characterized by constant returns to scale. For the ith firm we can define the following ratio of outputs to inputs:

$$ratio\ i = \frac{\alpha' y_i}{\beta' x_i}$$

$$i = 1, \ldots, N$$

where y_i is a vector of M outputs and x_i a vector of K inputs.

The maximization of efficiency implies the following problem:

$$\max_{\alpha,\beta} \frac{\alpha' y_i}{\beta' x_i}$$

subject to the following constraints:

$$\frac{\alpha' y_s}{\beta' x_s} \leq 1, \quad s = 1, \ldots, N \tag{2}$$

$$\alpha_m \geq 0, \quad m = 1, \ldots, M \tag{3}$$

$$\beta_k \geq 0, \quad k = 1, \ldots, K \tag{4}$$

The restriction given by Equation (2) implies that the efficiencies of all firms have to be less or equal that 1. Restrictions given by (3) and (4) rule out negative weights of outputs and inputs.

Intuitively, the problem seeks the optimal weights such that the efficiency of the firm i is maximized, while operating within the feasible set implied by the constraints.

Imposing the restriction $\beta' x_i = 1$, this fractional programming problem can be linearized ([43]) and transformed into the following:

$$\max_{\alpha,\beta} \alpha' y_i$$

subject to:

$$\beta' x_i = 1$$

$$\alpha' y_s - \beta' x_s \leq 0, \qquad s = 1, \ldots, N$$

$$\alpha \geq 0$$

$$\beta \geq 0$$

which can be written in the envelopment form as:

$$\min_{\theta,\lambda} \theta_i$$

subject to:

$$\sum_{s=1}^{N} \lambda_s y_s - y_i \geq 0$$

$$\theta_i x_i - \sum_{s=1}^{N} \lambda_s x_s \geq 0$$

$$\lambda_s \geq 0$$

where θ_i is the input oriented *efficiency score* for the ith firm.

λ stands for the set of multipliers in the linear combinations of the DMUs' inputs and outputs, i.e., the weight of each DMU within the peer group of DMUs.

This set up can also be applied to a technology exhibiting variable returns to scale by adding the convexity condition:

$$\sum_{s=1}^{N} \lambda_s = 1$$

This is an optimization problem, with linear objective function and constraints, solvable by linear programming.

The value of θ_i, the input-oriented technical efficiency score for the ith firm, indicates to what extent the inputs can be reduced in percent while keeping the output constant. For example, if DMU i has an efficiency score of 90%, it can reduce all inputs by 10% while offering the same amount of output.

Notice the difference between this set up and the statistical approach of SFA as presented in Equation (1) above.

The empirical exercise described in this paper employs the non-parametric, DEA formulation of the optimization problem as the baseline for analysis.

3. Material and Method: Data and Empirical Strategy

Data have been gathered primarily from Amadeus [44] a rich database comprising disaggregated economic and financial information from a large number of European companies. [8,35] employ also financial information from similar databases for their analyses.

Within the pharmaceutical industry, we have selected two main categories of firms in Amadeus according to their main activity:

(i) Manufacture of basic pharmaceutical products and pharmaceutical preparations;
(ii) Research and experimental development on biotechnology.

They will be labelled henceforth *manufacturers* and *R&D firms*, respectively. The two subgroups correspond to NACE (Nomenclature statistique des Activités Économiques dans la Communauté Européenne) codes 2110, 2120 (for manufacturers) and 7211 (for R&D firms). This is equivalent to NAICS (North American Industry Classification System) codes 541714 and 541715.

We work with yearly observations over the time horizon 2010–2018.

Following part of the literature on DEA, our research design has two stages (see Appendix A for an explanatory diagram of the design of our empirical exercise). The stages are detailed in Sections 4 and 5, respectively. In the first stage we compute the efficiency scores of the firms in our sample by DEA. In the second stage we design and estimate several statistical models to explore potential variables correlated with the efficiency scores; these models provide information regarding the sign of the correlation between the efficiency score and each variable, its statistical significance and its size.

Ordinarily, non-parametric techniques cannot be applied to data structured in panels because of tractability considerations, as is common, instead, with other methodologies which allow for an explicit time dimension and have been successfully employed with panels. We circumvent this problem computing measures of efficiency year by year. This feature may be regarded as a drawback on a priori grounds; nonetheless, the estimation of efficiency measures performed on a yearly basis has been useful to uncover interesting patterns in their evolution over time.

We have started to work with a sample encompassed by more than 4000 observations from 482 firms over the nine years in the period 2010–2018, evenly split among manufacturers and R&D firms.

For the computation of efficiency for a particular year, however, we have dismissed those observations corresponding to firms which do not report data of turnover, employees and/or assets for that same year. After discarding the firms with missing values, we end up with samples comprising around 200 companies for each year, of different sizes, geographical origins and performances over time. The samples, therefore, are quite representative of the industry.

In the case of multinationals, firms correspond to headquarters. In our selection of companies we have discarded local affiliates because internal accounting procedures of multinationals may reduce their degree of comparability.

Nominal variables have been deflated using the Harmonized European Index from Eurostat [45].

Our measure of output is turnover in real terms (in constant euros of 2015). The inputs labor and capital are proxied by the number of employees and total assets in real terms, respectively. Total assets in real terms are also measured in constant euros of 2015. The choice of these variables has been made in accordance with other contributions performing similar analyses, as [6,8,32].

Economic and financial conditions have been captured by cash flow over turnover, profit margin and average cost of employees, among others (see Appendix B).

We have constructed dummies for size, country of origin, main activity and years. The specific details will be provided in Sections 4 and 5 below.

Figure 2 conveys some information for selected variables, disaggregated in manufacturers and R&D firms. Real turnover is expressed in constant euros of 2015. It is apparent from the Figure that the firms encompassing the first category are considerably larger than those in the second, as shown by the average real turnover and average number of employees.

Figure 2. Average real turnover (in constant euros of 2015) and average number of employees over time by main activity. Notes: The figure displays the time pattern for average real turnover and average number of employees over 2010–2018, disaggregated by main activity of firms. Averages have been computed from the data year by year. Two main categories are considered: firms whose main activity is the manufacture of basic pharmaceutical products (manufacturers), and companies focused on research and experimental development on biotechnology (research and development (R&D) firms). Average turnover exhibits a decreasing trend over the period, with a big drop in 2012 for manufacturers, and an increasing trend for R&D firms since 2013. Average number of employees decreases over the period for the first category of firms and increases since 2016 for the second. Source: own elaboration with data from the Amadeus data basis).

It is also clear that both variables have experienced a decreasing pattern over time for manufacturers, with a very pronounced drop in 2012 in the case of real turnover. This is consistent with the increasingly difficult environment in which they operate. For R&D firms, the pattern is less straightforward.

Average real turnover has also plummeted in 2012 but has increased thereafter. Average number of employees falls until 2016 and rises in the last years of the period.

These trends may be associated to the progressive outsourcing of some stages of the value chain, which were traditionally performed by manufacturers and now are increasingly implemented by CROs and other biotechnological firms.

Two more considerations about our empirical strategy are in order. First, and as stated above, the DEA analysis can be implemented in an output oriented or input oriented setting. We have followed this second approach since it seems intuitively more appealing and conforming with firms' experience: their plans to increase efficiency are usually linked to reduction in costs, rather than to expansions in output.

Secondly, the relevant role played by R&D in this industry suggests that scale economies might be prevalent, but this is a controversial issue which the literature has not been able to settle yet. Reference [46] found evidence in favor of this hypothesis; Reference [47], however, did not, although they did suggest that economies of scope and accumulated knowhow were important for the firms in the sector. Reference [48] encountered knowledge spillovers among firms in Phase I of clinical research and diseconomies of scope in later phases. Reference [32] find that 60% of the firms in their sample of Japanese chemical and pharmaceutical companies operate with either increasing or decreasing returns to scale.

There is no consensus yet, therefore, on the degree of homogeneity of the production function in the industry. Anyhow, since the existence of increasing returns to scale cannot be ruled out, we have chosen to employ a variable returns to scale model as our theoretical framework, rather than a constant returns to scale. Reference [8] follow a similar approach.

4. Stage 1: Computation of Efficiency Scores

Pharmaceutical and biotechnological firms share some activities and hence compete with each other in certain stages of the value chain. We are interested in assessing whether the companies specialized in R&D activities are more or less efficient, being thus better or more poorly positioned to succeed and survive, than companies which are mainly producers and sellers. Hence, we analyze the firms in the industry jointly, i.e., with respect to an efficient frontier common for all of them (nonetheless, we have performed the analysis separately in each of the subgroups and basic results carry over).

Tables 1–5 and Figures 3 and 4 summarize some summary statistics about the efficiency of the firms that encompass our sample, as obtained employing DEA in our sample on a yearly basis.

Table 1. Efficiency in the pharma and biotechnological European industry by activity, 2010–2018.

	Efficiency Mean	Standard Deviation	Coefficient of Variation
Whole sample	0.341	0.265	0.777
Manufacturers	0.381	0.266	0.698
R&D firms	0.281	0.251	0.893

Note: the table summarizes selected statistics for efficiency levels, computed as described in the main text. We classify firms in two groups, manufacturers and R&D firms, according to their main activity. Source: Own elaboration.

The mean efficiency for the entire sample and over the period 2010–2018 is 0.341. Thus, firms in our sample could increase their efficiency on average in 0.659 points or 65.9%. It seems a reasonable figure. Reference [6] report values of efficiency between 0.42 and 0.58. Their sample is made up by US firms; it seems sensible to think that US firms are, by and large, more efficient than their European counterparts because the general level of efficiency of the US economy is larger and its regulatory burden is smaller. Furthermore, US pharmaceutical firms are larger, on average, than European firms and, as we shall argue below, our results suggest that larger firms are more efficient. The standard deviation is 0.265, which suggests a noticeable degree of dispersion in the sample.

The results are not very different from those obtained by [33]; they find that the average efficiency for a sample of Japanese firms is 0.68 for 1983–1987 and 0.47 for 1988–1993.

If we classify the firms according to their main activity, we find that the mean efficiency for the manufacturers is 0.381 whereas for the R&D firms the figure is smaller, 0.281. This is a somewhat surprising result: the common practice in the industry whereby manufacturers outsource some activities to R&D and biotechnological specialized companies like CROs would suggest on a priori grounds that the former be more efficient that the latter. Otherwise, the outsourcing could be questioned on economic grounds. This is not what we find, however.

One possible explanation for our results is that many manufacturers have been in the market longer, and their historical performance have endowed them with expertise, knowhow and managerial practices which have increased their productivity. This is related to the phenomenon called learning curve in engineering or learning by doing in economics. A classical example is provided by [49], who noticed that the number of hours necessary to produce an airframe was a decreasing function of the number of airframes already produced. Instead, many R&D firms are still relatively young; it is feasible, therefore, that there is still room for them to optimize their processes and value chains and improve their productivity and efficiency.

In addition, the R&D activity in order to develop new drugs is very risky. Success rates are low. Only a modest percentage of molecules are able to complete clinical phases successfully and enter the final market. Reference [50] report that only 10.4% of the drugs entering the clinical stage gain approval by the US Food and Drug Administration (FDA). Biotechnological firms displaying small sizes and relatively reduced pipelines may thus be very affected by failures in the R&D stage. These episodes, in turn, will entail lower levels of productivity.

Notice also that the standard deviation for R&D firms is comparatively high, 0.251. In fact the coefficient of variation, as measured by the ratio standard deviation to mean, is higher for this category. This implies that heterogeneity is more pronounced for this kind of firm.

In order to assess the connection between relative efficiency and size, we have created six categories of firms. Five of these categories (from very big to very small) are linked to the intervals delimited by the 95, 75, 50 and 25 percentiles of real turnover over the period. In particular, the classification is as follows:

- Huge: if the average real turnover over the period exceeds 2000 million euros.
- Very big: if the average real turnover is less or equal than 2000 million euros and higher than 426.92 million euros.
- Quite big: if the average real turnover is less or equal than 426.92 million euros and higher than 38.86 million euros.
- Medium: if the average real turnover is less or equal than 38.86 million euros and higher than 8.10 million euros.
- Small: if the average real turnover is less or equal than 8.10 million euros and higher than 2.10 million euros.
- Very small: if the average real turnover is less or equal than 2.10 million euros.

Table 2 displays summary statistics for relative efficiency classified according to these categories. The largest companies in the sample, those with turnover larger than 2000 million euros, have the highest level of efficiency in the sample, 0.98. In other words, most of them encompass the efficient frontier or are very close to it. There is very little dispersion within this category and the coefficient of variation is almost negligible.

For very big companies, with turnover roughly between 500 and 2000 million euros, efficiency is also remarkably high, 0.765 in average terms. The potential gains in efficiency for this category are only around 25% on average. Firms in the next turnover interval have a smaller record, 0.425. Medium-size firms register lower levels of efficiency on average, 0.312; this is slightly below the figure for the whole sample and period, 0.341.

Small firms, with turnover between 2.10 and 8.10 million euros, register the smallest value of average efficiency, only 0.267. Interestingly, their record is worse than that of the very small firms, with turnover below 2.10 million euros: this last category attains an indicator of 0.318, slightly above medium size firms. This result is consistent with [35], which find that small pharmaceutical firms display smaller levels of efficiency for the case of India.

Higher degrees of flexibility and capacity to adapt to the environment, more agile management and lower levels of conflicts among partners which characterize very small firms may be behind this result. The comparatives advantages provided by specialization may also play a role.

The performance within those categories, as reported by the coefficient of variation, is not uniform. Dispersion is maximum for the very small firms (0.9), whereas more limited for very big firms (0.267). Dispersion in the other categories is similar and quite high: between 0.6 and 0.71.

The implications of these results are interesting. There is not a monotonic, clear cut relationship between size, as captured by turnover, and relative efficiency. Our findings suggest that larger firms are more efficient but only beyond a certain threshold of income, located around 500 million euros. Companies above this figure are considerably more efficient, suggesting the possibility of scale economies for high levels of turnover. Firms with turnover between 38 and 500 thousand million euros also perform better than the whole sample, although their particular advantage amounts just to less than 10 points.

Intermediate and small firms do not profit from scale economies neither from the flexibility and specialization associated to very small firms, and therefore register the poorest results as far as efficiency is concerned.

Table 2. Relative efficiency in the pharma and biotechnological European industry by size, 2010–2018.

	Mean	Standard Deviation	Coefficient of Variation
Huge	0.98	0.039	0.039
Very big	0.765	0.205	0.267
Quite big	0.425	0.266	0.625
Medium	0.312	0.218	0.698
Small	0.267	0.19	0.71
Very Small	0.318	0.288	0.9

Note: the table summarizes selected statistics for efficiency levels, disaggregated by size of the firms (proxied by real turnover). The thresholds are described in the main text. Source: Own elaboration.

Table 3 and Figure 3 provide the dynamic context to these results by detailing the performance over the years 2010–2018. Average efficiency plummets from the beginning of the period until 2015, to recover thereafter. In year 2017, efficiency falls again, to increase in 2018, but it does not recover to the levels attained before 2010. Between 2010 and 2018 efficiency diminishes by almost 10 points. The decrease is especially acute for manufacturers, whereas R&D firms only lose 4 points on average.

These results are consistent with [6], who also document a decrease in efficiency for most of the firms in their sample for 2010–2015.

Table 3. Efficiency in the pharma and biotechnological European industry by activity, yearly results, 2010–2018.

	2010	2011	2012	2013	2014	2015	2016	2017	2018
Whole sample	0.428	0.392	0.348	0.308	0.304	0.292	0.383	0.311	0.334
Manufacturers	0.481	0.449	0.391	0.351	0.334	0.335	0.409	0.34	0.367
R&D firms	0.338	0.277	0.263	0.243	0.267	0.231	0.345	0.272	0.294

Note: the table details average levels of efficiency by year and main activity of firms. Efficiency is computed as described in the main text. Source: Own elaboration.

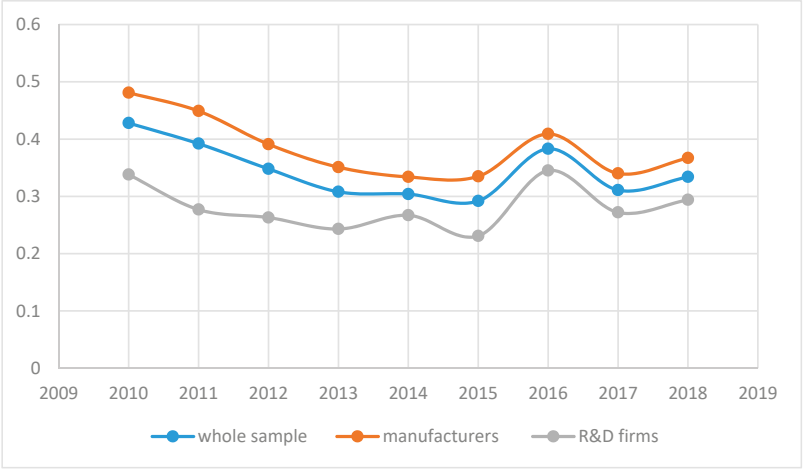


Figure 3. Efficiency in the pharma and biotechnological European industry by main activity, 2010–2018. Note: the figure summarizes the yearly trend of average efficiency, for the whole sample and disaggregated by categories corresponding to the main activity of firms. Efficiency decreases over the period, with a partial recovery in 2015–2016. Source: own elaboration.

Figure 4 portrays the behavior of firms over time classified according to their size. The largest companies exhibit a fairly consistent performance over time. Instead, for quite big companies the fall of efficiency between the beginning and end of the period is almost 20 points.

At the beginning of the period, in 2010, the efficiency of quite large firms was well above that of the entire sample, while this is not the case anymore in 2018. This category has been affected the most by the drop of efficiency over time.

Medium-sized and small firms exhibit a reduction of 10 points over time, whereas very small firms register a rather stable performance.

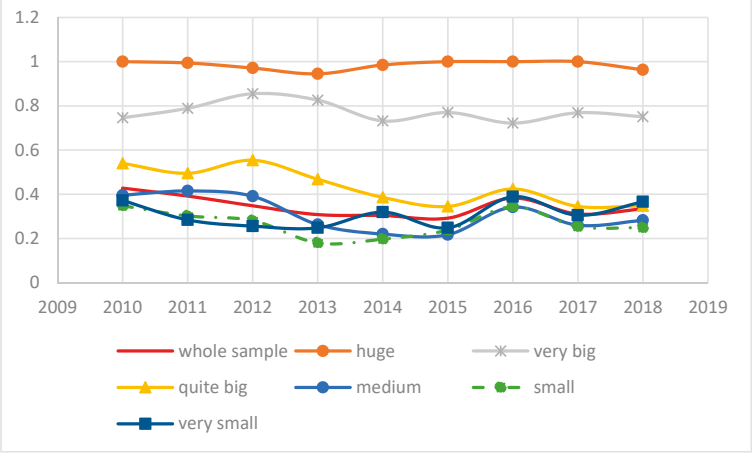


Figure 4. Average efficiency, pharma and biotechnological industry by size, 2010–2018. Note: the figure summarizes the yearly trend of average efficiency of the firms in our sample, disaggregated by size of firms. Size is proxied by real turnover. Efficiency decreases over the period for all categories except for the huge and very big firms. The thresholds are detailed in the main text. Source: own elaboration.

5. Stage 2: Variables Correlated with Efficiency

5.1. Overview

In the second stage of this research we have performed a regression analysis in order to explore several aspects of the firms' economic setting and management which may be correlated with efficiency. Efficiency is proxied by the efficiency scores obtained in the first stage, as detailed in Section 4.

The basic framework is a statistical model described in very general terms by Equation (5):

$$\theta = f(x; \upsilon) \tag{5}$$

where θ is a vector containing the efficiency scores, x is a matrix of covariates and υ is the error term.

There are several statistical issues to be considered here.

First, the literature has not reached a consensus about the data generation process underlying Equation (5). Researchers have widely used the Tobit model and ordinary least squares (OLS) (see, for example, [35,51]).

Since the efficiency scores are censored at a maximum of 1 by construction, the Tobit specification seems especially appropriate for this analysis. In addition, References [52,53] argue that OLS provide consistent estimates which are quite similar to those obtained with Tobit and are, therefore, a convenient procedure. Reference [54] show, by means of Monte Carlo simulations, that OLS and Tobit outperform other procedures when employed in the second stage of DEA analyses.

Reference [36], however, have argued that the true data generation process for the efficiency scores is not a censored but a truncated distribution; they discard the analysis of the efficiency scores performed according to Tobit or ordinary least squares because this assessment would not rely on the *true* distribution of the data. With censored data, the *true* value of the variable is not known because of the measurement scale; in this particular case, since efficiency has an upper bound of 1. With truncated data, instead, the true value of the variable is unknown because of the sample limitations. The difference in practice between a censored and a truncation distribution may be unclear.

Furthermore, they claim that the efficiency scores are affected by serial correlation. Since the Tobit procedure does not correct for this problem, the estimates obtained from the Tobit model are, in their view, biased. This issue is also controversial, since [54] have argued that OLS and Tobit procedures are valid even if the X variables are correlated.

Reference [36] propose an alternative estimation technique which employs a truncated model, computes new standard errors by bootstrapping the data and corrects the biases in the estimates. There are downsides for this procedure. Reference [53] argues that the Simar–Wilson estimates lack robustness. Furthermore, the Simar–Wilson technique is convoluted and intensive in computing time. Furthermore, as we shall show below, the point estimates computed by the Simar–Wilson method are bigger than those obtained by Tobit or ordinary least squares, although the difference may not be very relevant in applied research.

The debate is still open. According to [53], the controversy about the correct statistical model underlying the data is ultimately methodological and exceeds the scope of our research. By and large, we agree with [53] and think that Tobit and ordinary least squares have helped obtain valid insights about the efficiency in numerous industries or activities, and thus can be employed in applied research.

Meanwhile, since the controversy has not been settled yet, we have decided to adopt a conservative strategy, employ the three methods and compare their results.

Second, the data we are going to use to estimate Equation (5) encompass a panel and hence comprises observations from firms at different points in time.

As is well known, panel data can be assessed by fixed effects or random effects models. [55] shows that Tobit models with fixed effects produce coefficients which are overestimated and asymptotic variances which are biased downwards. Moreover, our specification includes as regressors time-invariant characteristics of firms (such as country of origin, for example); these characteristics

would be perfectly collinear with the terms capturing the idiosyncratic features of firms in a fixed effects model. In this case we cannot employ a Hausman test to compare the fixed effects and random effects models because our model cannot be specified within a fixed effects setting.

These considerations advise the utilization of random-effects models. This is the approach followed, for example, by [35].

Finally, at this point we are searching for correlations among efficiency and different aspects of firm idiosyncrasies and management. Looking for causality relationships exceeds the scope of this paper and is left for future research.

We shall start by discussing the main qualitative implications of this exercise, for reasons which will be apparent below.

5.2. Qualitative Implications

5.2.1. Tobit Estimation

Typically, a Tobit model distinguishes between the latent or unobservable dependent variable and the observable dependent variable, where the observed variable is a censored version of the unobserved.

Equation (6) represents a random-effects Tobit specification for the second stage of our analysis:

$$\begin{aligned} \theta^*_{it} &= x_{it}\beta + u_i + \varepsilon_{it} \\ \theta_{it} &= 1 \quad if\ \theta^*_{it} \geq 1 \\ \theta_{it} &= \theta^*_{it} \quad if\ 0 \leq \theta^*_{it} \leq 1 \\ \theta_{it} &= 0 \quad if\ \theta^*_{it} \leq 0 \\ i &= 1,\ 2,\ \ldots,\ n. \\ t &= 2010,\ 2011,\ \ldots,\ 2018 \end{aligned} \tag{6}$$

wre θ_{it}* is the latent or unobservable efficiency, θ_{it} is the observable efficiency, x_{it} is a matrix of covariates, β is a vector of coefficients, u_i is the time invariant component of the error term, ε_{it} is the time-varying component of the error term, i indexes firms and t time.

In the estimation of Equation (6) we have included several indicators as covariates in order to capture different dimensions of firms, such as main activity, size, margins, financial management and personnel costs. We have also included time dummies to capture the impact of the business cycle and country dummies to allow for idiosyncratic aspects related to the markets where firms operate. The data are structured in a panel over the period 2010–2018 in order to exploit both the cross section and time variations.

Table 4 shows a first set of results obtained from the estimation by maximum likelihood of the model described by Equation (6). In order to avoid multicollinearity among the regressors, we have not included all covariates simultaneously; instead, we have added them sequentially, conforming different specifications of the baseline Equation (6). In other words, Equation (6) describes Models 1–4, the differences among them being the variables considered in x_{it} in each case.

To correct for heteroskedasticity, estimations have been performed with the observed information matrix (OIM) corrected standard errors. In this particular case, the variance-covariance matrix of the estimators is the matrix of second derivatives of the likelihood function. This correction for heteroskedasticity is robust to the violation of normality if distribution is symmetrical.

The last lines of Table 4 include the results from a Lagrange multiplier Breusch–Pagan likelihood ratio test of whether the variance of the time invariant component of the error term is equal to zero. This test is can be regarded as an indirect text of the appropriateness of the random effect model. The null hypothesis of equality to 0 of the variance of the u_i component of the error term is rejected at the 99% significance level for the four models, hence supporting the utilization of the random-effects model.

Dummies for countries capture different aspects: on the one hand, cultural and institutional aspects and managerial practices ([38]). On the other, regulatory and microeconomic and macroeconomics conditions of the particular markets where the firms operate. Regulatory aspects and institutional and macroeconomic conditions in the host country have been shown to impact the performance of multinational firms ([56,57]).

Dummies for the United Kingdom (UK), Italy and Sweden are positive and highly significant in all specifications, implying that the institutional framework in these countries, the size of their markets and/or their macroeconomic and institutional conditions affect the efficiency of firms positively. The dummy for Germany is also positive and significant in two specifications (models 2 and 4), although in one of them at a smaller significance level (90% in model 4).

Instead, the dummies for Spain and France display positive and negative signs and are not significant.

UK pharmaceutical firms feature a swift decision-making process which facilitates a successful and fast adjustment to changing market conditions ([58]). Moreover, the level of distortions in the UK economy is low and factor markets are relatively flexible. In addition, the dynamic biotechnological landscape of the country has allowed the surge of alliances and collaborations. These facts may explain the positive sign of the UK dummy.

German firms typically work in less-flexible environments than their British counterparts; their access to bank funding, though, is comparatively easy. Since sound finance is one important determinant of firms' success, as will be detailed below, the availability of funding seems quite relevant for the performance of companies in the sector and help explain the positive sign of the dummy.

The Italian industry is populated by highly skilled, agile firms, with a large component of exports and close ties to US companies. These companies encompass an important hub for foreign investment in the industry, which in turn enhances the productivity of local firms through technology diffusion and *learning by watching*.

Swedish pharmaceutical and biotechnological firms benefit from a market with limited regulation where bureaucracy is kept at a minimum, government support and a highly skilled workforce. These aspects would explain the successful performance of the Swedish pharmaceutical industry.

The positive signs of the country dummies, therefore, are in accordance with particular features of their institutional frameworks and/or industries.

These features, however, are not present in the French and Spanish cases. The French pharmaceutical market has historically been very protected by an outdated industrial policy. Spanish companies have been damaged by a rigid labor market and a low level of interaction between universities, research centers and firms.

We have also captured the main activity of the firms by means of dummies variables. The dummy *manufacturers* is equal to 1 for those firms whose main activity corresponds to NACE codes 2110 and 2120, and 0 otherwise. Conversely, the dummy *biotech* is 1 for firms included under the 7211 NACE code and 0 otherwise.

The dummy manufacturers are positively and significantly correlated with efficiency (columns 1 and 3), while biotech displays a negative and significant correlation in one model (column 2) and is not significant in the other (column 4). Overall, these findings are in accordance with those reported in Section 4 above, which suggest consistently higher levels of efficiency for firms engaged in the production and commercialization of pharmaceutical articles.

Dummies for size have been assigned according to the thresholds detailed in Section 4 above. Again, the results for the estimations agree with the trends reported in the previous Section. Firms characterized by large sizes, as conveyed by their levels of turnover, are more efficient than their counterparts, since the dummies huge and very big are positively and significantly correlated with efficiency (Models 1 and 4). The dummy that is quite big is positive but not significant.

The positive correlation between size and efficiency, however, holds only for the first two categories we defined, i.e., for sales larger than 426.92 million euros or the 95 percentile in the distribution. For companies with real turnover between 38.86 and 426.92 million euros results are inconclusive.

Those companies whose level of sales is less or equal than 38.86 million and more than 2.10 million euros register smaller efficiency figures *ceteris paribus*, since the dummies medium and small are negative and significant (column 2). Finally, we do not find a significant correlation between the dummy capturing the *very small* level of sales and efficiency (column 2). This is not surprising since firms with sales lower than the 25% percentile register poor levels of efficiency in some years but are capable of surpassing the figure attained by medium and small others.

The results for the dummy variables reflecting size and activity are thus consistent with those reported in the previous section. They are also in accord with [35], who disclose a negative correlation between size and efficiency for a sample of Indian pharmaceutical firms.

Let us turn to the discussion of the variables capturing other aspects of firms in the industry.

As portrayed by column 1 of Table 4, the profit margin is positively and significantly correlated, at the 99% significance level, with efficiency. This means that more efficient firms operate with higher margins. This result makes sense because the industry we are scrutinizing provides goods and services characterized by high added value which can be reflected in large margins. In fact, Reference [59] argues that deviations from trend in profit margins are highly correlated with expenditure in R&D for pharmaceutical companies, thus confirming the links between efficiency, margins and R&D.

Interestingly, this finding suggests that successful firm strategies in this sector are featured by both high margins and high intensity of resource utilizations, at the same time. It is common to see that companies tend to choose to focus either on the achievement of high profits per unit or in the optimization of the installed capacity. This dichotomy, however, is not present in the companies in the pharmaceutical industry, according to our results.

Table 4. Variables correlated with efficiency, Tobit estimations. Dependent variable is efficiency.

	Model 1	**Model 2**	**Model 3**	**Model 4**
profit_margin	0.1539 ***			
	(7.00)			
Germany	0.0799	0.1178 **	0.0818	0.1045 *
	(1.38)	(2.02)	(1.49)	(1.89)
Spain	−0.0232	−0.0614	0.0305	0.0435
	(0.44)	(1.10)	(0.54)	(0.87)
France	0.0100	0.0544	−0.0117	0.0167
	(0.17)	(0.94)	(0.21)	(0.31)
Sweden	0.2977 ***	0.2615 ***	0.2389 ***	0.3008 ***
	(3.93)	(3.44)	(3.45)	(4.13)
Italy	0.1421 ***	0.1549 ***	0.1587 ***	0.1374 ***
	(2.62)	(2.85)	(2.85)	(2.79)
UK	0.1389 ***	0.1637 ***	0.1128 ***	0.1356 ***
	(3.62)	(4.16)	(2.88)	(3.64)
Manufacturers	0.0660 **		0.1599 ***	
	(2.00)		(4.96)	
Huge	0.2054 **			0.2393 ***
	(2.49)			(3.02)
Verybig	0.1276 ***			0.1163 ***
	(3.61)			(3.24)
Quitebig	0.0215			0.0179
	(1.39)			(1.09)
year2014	−0.0795 ***	−0.0744 ***	−0.0633 ***	−0.0738 ***
	(6.90)	(6.37)	(5.88)	(6.16)
year2015	−0.1733 ***	−0.0758 ***	−0.0556 ***	−0.1537 ***
	(4.97)	(4.69)	(5.31)	(4.35)
year2016	0.0373 ***	0.0465 ***	0.0439 ***	0.0441 ***
	(3.32)	(4.12)	(4.34)	(3.79)

Table 4. *Cont.*

	Model 1	Model 2	Model 3	Model 4
cash_flow		0.1650 *** (6.86)		
Biotech		−0.1384 *** (4.21)		−0.0494 (1.59)
Medium		−0.0300 * (1.81)		
Small		−0.0614 *** (3.22)		
Verysmall		−0.0069 (0.29)		
collection_period			−0.0226 *** (4.44)	
employee_cost				−0.3739 *** (10.22)
_cons	0.2808 *** (7.95)	0.3788 *** (12.50)	0.2245 *** (6.37)	0.4390 *** (15.49)
Likelihood Ratio test of $\sigma^2_u = 0$: $X^2(1)$	928.17 ***	980.9 ***	1505.81 ***	771.79 ***
Likelihood Ratio test of $\sigma^2_u = 0$: p value	0	0	0	0
Number observations	1547	1344	1850	1353

Notes: The table summarizes the results from the Tobit estimation of Equation (6). Dependent variable is efficiency computed in Stage 1. Cons stands for the intercept. For the rest of variables, see main text. Data are organized in a panel varying across firms and time over 2010–2018. In order to circumvent heteroskedasticity, estimations have been performed with corrected standard errors; the variance-covariance matrix of the estimators is the matrix of second derivatives of the likelihood function. LR test of $\sigma^2_u = 0$ distributed as $X^2(1)$. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

The literature has documented that cash flow influences R&D expenditure in the case of the industry we are considering ([60]). Reference [61] provide some additional evidence since they find that, for the Spanish firms, the proportion of expenditure in R&D financed with internal resources is 75% for pharmaceuticals and 40% for the rest of the industries. Again, we are confronted with another differential feature of this industry. Whereas it is commonly accepted that firms should heavily rely on external funding and increase their profitability through financial leverage, the empirical evidence for this industry suggests that successful companies enjoy comparatively low ratios of indebtedness. This prudent financial structure is consistent with the high risk and long maturing period associated with the R&D activity.

To test this idea in our sample, we have included in the analysis some variables which capture particular elements of financial management. Column 2 shows that cashflow (as a percentage of sales) is indeed positively and significantly correlated with efficiency. The level of significance is very high, 99%.

Column 3, in turn, displays the estimation results when the variable collection period is included as a regressor in the baseline specification. The point estimate is negative and significant at the 99% level. Higher collection periods increase the amount of working capital necessary to run the daily activity of the firm, while shorter spans imply a sounder financial management. Our findings, therefore, are consistent with the literature, and stress the importance of exhibiting solid, well-financed balance sheets in order to register high levels of productivity. In more detail, Reference [35] argue that the low efficiency scores achieved by some firms in their sample is associated to their inability to access financial resources.

Column 4 includes a variable capturing the cost of labor, average cost per employee, as a percentage of sales. It is highly significant and negatively correlated with efficiency.

In terms of the validations of Models 1–4, and as stated above, the literature has shown that the Tobit model provides consistent estimates ([52–54,62]).

Moreover, it has been argued that the severity of the problem implied by the presence of heteroskedasticity in Tobit models is a function of the degree of censoring. In our case, censoring is limited, and affects only to 6–7% of the data.

Since the estimations have been performed with OIM corrected standard errors, they are robust to the presence of heteroskedasticity. These standard errors are also robust to the violation of normality if the distribution is symmetric.

Finally, and as detailed below, results from Tobit are quite similar to those obtained by random-effects models. All these considerations lend countenance to the models described in this subsection.

5.2.2. Classical Estimation

In order to assess the robustness of these findings we have performed two complementary analyses. First, we have considered a pure random-effects model, as described by Equation (7).

$$\theta_{it} = x_{it}\beta + u_i + \varepsilon_{it} \tag{7}$$

where θ_{it} is efficiency, x_{it} is a matrix of covariates, β is a vector of coefficients, u_i is the time invariant component of the error term, ε_{it} is the time-varying component of the error term, i indexes firms and t time.

The estimation has been carried out with robust standard errors, in the spirit of [63–65], clustered at the firm level. This procedure is widely recommended in the literature in these types of estimations ([66]).

Table 5 summarizes the specification and results for Models 5–8, estimated according to (7). We see that the main conclusions obtained from the Tobit specification regarding the correlation of efficiency with selected variables carry over to the classical, pure random effects specification. The only remarkable differences are related to the dummy for Spain, which is now negative and significant at the 95% level (Model 6), and the dummy quite big, now significant at the 90% level.

Furthermore, the point estimates of the coefficients are very similar in the censored and the non-censored model. These results are reassuring and consistent with [52,53], who document this kind of similarity when Tobit and ordinary least squares are employed in the second stage analysis.

The last two lines of Table 5 display the results from the Lagrange multiplier Breusch–Pagan test for the presence of random effects. The null hypothesis of no random effects is rejected at conventional levels.

In terms of the validation of Models 5–8, we can invoke the result according to which OLS produces unbiased and consistent estimates because of the central limit theorem for large enough samples. In addition, the literature has also shown the consistency of OLS second-stage estimators for the particular case of DEA analyses. Moreover, cluster robust standard errors yield estimates that are robust to the presence of heteroskedasticity and correlation in the error term.

Table 5. Variables correlated with efficiency, random effects estimations. Dependent variable is efficiency.

	Model 5	**Model 6**	**Model 7**	**Model 8**
profit_margin	0.1531 ***			
	(5.33)			
Germany	0.0702	0.1064 *	0.0740	0.0908 *
	(1.23)	(1.85)	(1.41)	(1.70)
Spain	−0.0210	−0.0598 **	0.0304	0.0427
	(0.55)	(2.04)	(0.66)	(0.87)
France	0.0093	0.0504	−0.0150	0.0128
	(0.19)	(1.08)	(0.32)	(0.28)
Sweden	0.2779 ***	0.2420 ***	0.2280 ***	0.2922 ***
	(3.07)	(3.23)	(3.32)	(3.78)

Table 5. *Cont.*

	Model 5	**Model 6**	**Model 7**	**Model 8**
Italy	0.1375 **	0.1488 ***	0.1529 ***	0.1336 ***
	(2.52)	(2.74)	(2.86)	(2.78)
UK	0.1340 ***	0.1538 ***	0.1041 **	0.1295 ***
	(3.44)	(3.91)	(2.51)	(3.70)
Manufacturers	0.0619 *		0.1534 ***	
	(1.93)		(4.81)	
Huge	0.1189 ***			0.1649 ***
	(2.63)			(3.72)
Verybig	0.1277 ***			0.1247 ***
	(4.28)			(4.06)
Quitebig	0.0233 *			0.0218
	(1.71)			(1.56)
year2014	−0.0778 ***	−0.0736 ***	−0.0625 ***	−0.0728 ***
	(8.64)	(7.67)	(7.14)	(7.90)
year2015	−0.1712 ***	−0.0792 ***	−0.0558 ***	−0.1596 ***
	(6.11)	(6.84)	(7.03)	(5.51)
year2016	0.0366 ***	0.0450 ***	0.0430 ***	0.0430 ***
	(4.09)	(4.52)	(4.78)	(4.46)
cash_flow		0.1661 ***		
		(5.83)		
Biotech		−0.1298 ***		−0.0421
		(4.42)		(1.44)
Medium		−0.0349 **		
		(2.48)		
Small		−0.0648 ***		
		(3.62)		
Verysmall		−0.0092		
		(0.33)		
collection_period			−0.0226 ***	
			(3.97)	
employee_cost				−0.3701 ***
				(7.83)
_cons	0.2786 ***	0.3767 ***	0.2257 ***	0.4312 ***
	(7.92)	(15.55)	(6.51)	(19.16)
LR test of $\sigma^2_u = 0$: $X^2(1)$	1306.01 ***	1656.88 ***	2561.80 ***	1156.37 ***
LR test of $\sigma^2_u = 0$: p value	0	0	0	0
Number of observations	1547	1344	1850	1353

Notes: The table summarizes the results from a pure Random-effects estimation of Equation (5). Dependent variable is efficiency computed in Stage 1. Cons stands for the intercept. For the rest of variables, see main text. Data are organized in a panel varying across firms and time over 2010–2018. Robust standard errors clustered at the firm level. LR test of $\sigma^2_u = 0$ distributed as $X^2(1)$. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

5.2.3. Simar–Wilson Estimation

We have employed the [36] methodology as a further robustness test. Accordingly, we have replicated the estimations described above, this time employing their technique. These are Models 9–12, whose detailed specifications and results are displayed in Table 6.

Once again, we see that the basic findings obtained by the Tobit and classical random effects estimations regarding the sign and significance of covariates carry over when the [36] procedure, based upon a truncated distribution for the data and bootstrapping, is employed.

As reported above, this tool aims to remove the alleged bias in the estimation due to correlation among residuals. It computes new standard errors and corrected parameters. In contrast to the Tobit and classical frameworks, the literature has not provided enough evidence yet to illustrate the properties of this estimator.

Table 6. Variables correlated with efficiency, Simar–Wilson estimations. Dependent variable is efficiency.

	Model 9	**Model 10**	**Model 11**	**Model 12**
profit_margin	0.3089 ***			
	(9.31)			
Germany	0.1287 ***	0.1562 ***	0.1405 ***	0.1204 ***
	(4.43)	(4.93)	(3.64)	(4.09)
Spain	−0.0356	−0.0971 ***	−0.0098	0.0053
	(1.36)	(3.25)	(0.24)	(0.20)
France	0.0342	0.0684 *	−0.1606 ***	−0.0364
	(0.92)	(1.72)	(3.06)	(0.97)
Sweden	0.2958 ***	0.2602 ***	0.3352 ***	0.3539 ***
	(8.04)	(6.04)	(7.06)	(8.41)
Italy	0.1548 ***	0.1539 ***	0.2307 ***	0.1506 ***
	(6.13)	(5.67)	(6.41)	(5.99)
UK	0.1439 ***	0.1596 ***	0.1370 ***	0.1396 ***
	(6.89)	(7.33)	(4.97)	(6.81)
Manufacturers	0.0859 ***		0.3157 ***	
	(4.54)		(10.51)	
Huge	0.7812 **			0.8741 **
	(2.06)			(2.45)
Verybig	0.4284 ***			0.3849 ***
	(9.48)			(8.47)
Quitebig	0.0552 ***			0.0678 ***
	(2.98)			(3.63)
year2014	−0.0994 ***	−0.1072 ***	−0.0982 ***	−0.0977 ***
	(4.20)	(4.16)	(2.80)	(3.77)
year2015	−0.4586 ***	−0.1540 ***	−0.1161 ***	−0.4179 ***
	(9.43)	(5.15)	(3.40)	(8.48)
year2016	0.0652 ***	0.0651 ***	0.0902 ***	0.0785 ***
	(3.04)	(2.76)	(3.07)	(3.66)
cash_flow		0.3351 ***		
		(8.31)		
Biotech		−0.1790 ***		−0.0229
		(8.07)		(1.15)
Medium		−0.1294 ***		
		(6.36)		
Small		−0.1091 ***		
		(4.63)		
Verysmall		0.0192		
		(0.60)		
collection_period			−0.0568 ***	
			(4.11)	
employee_cost				−0.6340 ***
				(10.95)
_cons	0.1328 ***	0.3237 ***	−0.0524	0.3822 ***
	(6.01)	(15.50)	(1.24)	(20.23)
Number of observations	1446	1257	1741	1264

Notes: The table summarizes the results from the Simar–Wilson estimation of Equation (5). Dependent variable in the estimations is efficiency. Data are set in a panel varying across firms and time. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

5.3. Quantitative Implications

From the comparisons of Tables 4–6 we observe that Tobit and pure random-effects models yield point estimates which are rather similar. Instead, estimates obtained by the Simar–Wilson methodology are larger.

In contrast to what happens in the classical regression model, the marginal effect or impact of the individual regressor x_j on the dependent variable, defined as:

$$\frac{\partial \theta}{\partial x_j}$$

is not directly measured by the point estimates of regressions estimated by Tobit or Simar–Wilson methodologies, since they are non-linear models.

In order to extract more quantitative implications of the different estimations described in Section 5.2 above, we have computed the marginal effects of selected variables on efficiency implied by these two methods.

Basic results are displayed in Table 7. In order to facilitate comparisons, we have added the point estimates obtained by the pure random-effects estimation.

The variable exerting the highest impact on efficiency is employee cost. According to our results, an increase of one unit in the employee cost reduces efficiency in an amount which is comprised in the interval (0.368, 0.42).

If the profit margin rises in one unit, the correspondent increase in efficiency is around 0.15–0.2. The improvement of the financial position (as captured by cash flow/income) in one unit brings about a positive change in efficiency of 0.162–0.218. Finally, the increase of the collection period in one unit reduces efficiency around 0.02.

In our view, these findings have some interesting economic implications and may be useful for managers, owners and other stakeholders of firms in the industry. The efforts to contain personnel costs and increase margins translate directly into higher levels of productivity. Firms in the industry should also strive to achieve an adequate combination of external and internal finance, aligned with the risky and slow-paced nature of R&D activities.

There are implications for policymakers and policy analysts as well. Efficiency in the pharmaceutical sector, according to the empirical evidence presented here, hinges on the sound functioning of labor markets and financial markets. Measures to improve their behavior may have a noticeable impact on the performance of the firms in the industry.

It is apparent from Table 7 that the marginal effects obtained by the Tobit and the classical specifications are remarkably close, whereas those yielded by the Simar–Wilson procedure are slightly larger. It is important to notice that the difference among the Tobit/pure random effects results, on the one hand, and the Simar–Wilson, on the other, is smaller regarding the marginal effects (Table 7) that if we compare the point estimates (Tables 4–6).

This fact has several interesting implications:

- As far the particular goal of this subsection is concerned, the Simar–Wilson tool implies marginal effects slightly larger (about 15–35%) but of the same order of magnitude than those obtained from Tobit/pure random-effects model.
- In general terms, more research at the theoretical level and probably Monte Carlo simulations are necessary to know in more detail the properties of the Simar–Wilson estimator. This exceeds the scope of this paper.
- The Simar–Wilson procedure may be useful for applied research, especially in conjunction with other methodologies, although it has a higher cost in computing time if compared with Tobit or classical models.

Table 7. Comparison of marginal effects, Tobit, Simar–Wilson and random effects estimations.

Variable	Tobit	Simar–Wilson	Random Effects
Profit margin	0.1511	0.2053	0.1531
Cash flow/income	0.1628	0.2189	0.1661
Collection period	−0.0223	−0.026	−0.0226
Employee cost	−0.3683	−0.4215	−0.3701

Notes: The table details the marginal effects on efficiency levels of each one of the variables displayed in the first column. These marginal effects have been recovered from the Tobit (Models 1–4) and the Simar–Wilson (Models 9–12) estimations. The last column displays the marginal effects obtained in the pure random-effects models (Models 5–8) to facilitate the comparison; since this framework is linear, the marginal effects coincide with the point estimates of the variables as reported in Table 5.

6. Concluding Remarks

The pharmaceutical industry has experienced deep changes in the last few decades. The cost of R&D has soared while market conditions have become tougher. Companies have confronted these challenges by different strategies such as mergers, acquisitions, outsourcing and alliances. It remains an open question whether these transformations have brought about an increase in the efficiency of the firms that make up the industry.

We examine this issue employing disaggregated microdata from a large sample of European medium and large firms belonging to the pharmaceutical and biotechnological industry. In the first stage of our research, we perform a non-parametric DEA analysis of efficiency over the period 2010–2018. In the second stage we analyze which potential features of the environmental framework and management are correlated with efficiency by regression techniques.

The consideration of a large sample of European firms, disaggregating by main activity and isolating the performance of biotechnological firms is a novel feature of this paper. The comparison of the results provided by the Tobit, classical and Simar–Wilson frameworks for the second stage is also a contribution of the investigation presented here.

The main insights from our analysis are the following:

- The average level of efficiency in the industry is moderate, 0.341. This figure is not far from results obtained by other studies for alternative samples. Efficiency exhibits a decreasing trend over the years 2010–2018.
- Efficiency levels display a large level of heterogeneity when particular dimensions of companies are considered. Efficiency is higher for those companies whose main activity is manufacturing of pharmaceutical products than for firms focused on R&D activities. This result may be traced to the relative youth of R&D firms, which cannot fully exploit the learning curve yet. The specialization of this kind of firms in a few projects, characterized by low rates of success, may also be a relevant factor in this respect.
- We find a complex relationship between size and efficiency. By and large, bigger firms are more efficient, but only beyond the threshold of 426.92 million euros of turnover per year. Medium-size and small firms register the poorest levels of efficiency, whereas very small firms perform slightly better. This suggests that firms may benefit from either scale economies or high levels of specialization, while the middle ground does not yield good results.
- Our findings suggest that sound financial structures, lower employee costs and higher margins are correlated with higher levels of efficiency. Moreover, the idiosyncratic aspects of the country of origin of the firms may foster or jeopardize productivity.

Our results convey some messages for policymakers. The survival and buoyancy of companies in the pharmaceutical industry seems closely linked to the sound functioning of the labor and capital markets. The experience of selected countries, in particular the UK, suggests as well that the existence of agile, dynamic biotechnological firms is beneficial for the whole sector.

Finally, the higher levels of efficiency obtained for larger firms suggest that mergers and acquisitions may enhance the performance of pharmaceutical companies due to the influence of scale economies. These financial transactions should not be discouraged or jeopardized by policymakers on the basis of an allegedly anti-competitive strategy. It is important to keep in mind that the pharmaceutical and biotechnological industry relies heavily on R&D, and that R&D is only feasible for firms if their size is big enough.

We have also found that very small firms display a sounder behavior than medium size companies. The link between size and performance for the sector is thus nuanced. This suggests that industrial policies intending to enhance the sector should be horizontal rather than vertical: instead of featuring active interventions in favor of a particular firm size, it is better to adopt a less activist stance since it is hard to determine on an a priori basis which is the efficient scale of operations.

Our investigation has several limits. The time horizon is relatively short; it would be convenient to increase it whenever new data are available. We have computed efficiency scores in Stage 1 only by a non-parametric technique, DEA; another computation by means of parametric SFA would be useful to check whether efficiency scores are very sensible to the tool employed.

In stage 2 we have investigated the correlations among efficiency scores and other variables, but we have not explored the direction of causality among them. This last issue could be addressed by introducing lags and leads of the variables and/or employing other econometric techniques, such as general methods of moment or instrumental variables.

One of the techniques we have employed in Stage 2 is the Simar–Wilson estimation. It seems to be useful in applied work, especially in combination with other techniques. More evidence about its performance would be convenient, nonetheless.

Finally, and although country dummies have provided useful information about the potential impact of institutional and economic aspects on efficiency, they are ultimately dummies or *the measure of our ignorance*; it would be interesting to go one step further and characterize the specific features of the various countries which enhance or jeopardize efficiency. This could be done by introducing macroeconomic and institutional variables in the Stage 2 models.

These limitations suggest promising directions for new research.

Author Contributions: Conceptualization, R.F.D. and B.S.-R.; Methodology, B.S.-R.; Formal Analysis, R.F.D. and B.S.-R.; Data Curation, R.F.D. and B.S.-R.; Writing—Original Draft Preparation, R.F.D. and B.S.-R.; Writing-Review and Editing, R.F.D. and B.S.-R.; Supervision: B.S.-R. Both authors have read and agreed to the published version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We are very grateful to José María Labeaga, Teresa Herrador and three anonymous referees for helpful suggestions and comments.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

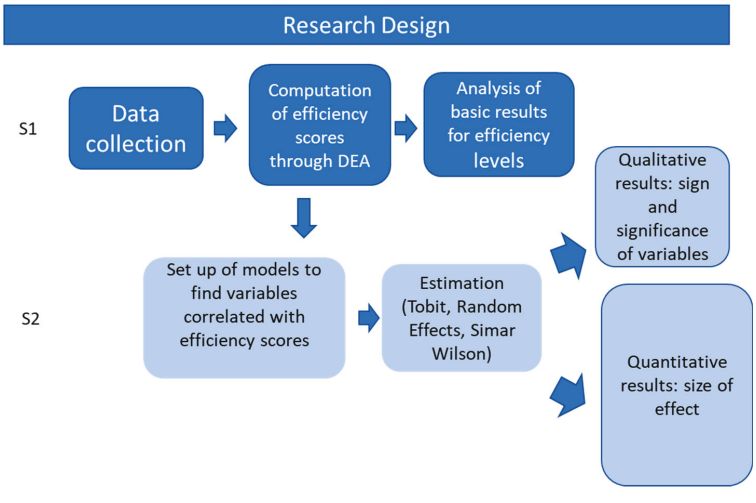


Figure A1. Explanatory diagram of our research design (S1 and S2 are Stage 1 and Stage 2).

Appendix B

Table A1. Variables definition and sources.

Variable	Description	Source
OPRE—Operating Revenue (Turnover)	Total Operating Revenues (Net Sales + Other Operating Revenues + Stock Variations)	Amadeus
TOAS—Total Assets	Total Assets (Fixed Assets + Current Assets)	Amadeus
PRMA—Profit Margin (%)	(Profit Before Tax/Operating Revenue) * 100	Amadeus
EMPL—Number of Employees	Total Number of Employees included in the Company's payroll	Amadeus
CFOP—Cash Flow/Operating Revenue (%)	(Cash Flow/Operating Revenue) * 100	Amadeus
SCT—Cost of Employees/Operating Revenue (%)	(Cost of Employees/Operating Revenue) * 100	Amadeus
COLL—Collection Period (days)	(Debtors/Operating Revenue) * 360	Amadeus
Yearly deflator	Computed from the Harmonized European Index	Eurostat

References

1. Lucas, R.E. On the mechanics of economic growth. *J. Monet. Econ.* **1988**, *22*, 3–42. [CrossRef]
2. Romer, P.M. Increasing Returns and Long-Run Growth. *J. Political Econ.* **1986**, *94*, 1002–1037. [CrossRef]
3. Romer, P.M. Endogenous Technological Change. *J. Political Econ.* **1990**, *98*, S71–S102. [CrossRef]
4. Pammolli, F.; Magazzini, L.; Riccaboni, M. The productivity crisis in pharmaceutical R&D. *Nat. Rev. Drug Discov.* **2011**, *10*, 428–438. [CrossRef] [PubMed]
5. Shimura, H.; Masuda, S.; Kimura, H. A lesson from Japan: Research and development efficiency is a key element of pharmaceutical industry consolidation process. *Drug Discov. Ther.* **2014**, *8*, 57–63. [CrossRef] [PubMed]
6. Shin, K.; Lee, D.; Shin, K.; Kim, E. Measuring the Efficiency of U.S. Pharmaceutical Companies Based on Open Innovation Types. *J. Open Innov. Technol. Mark. Complex.* **2018**, *4*, 34. [CrossRef]
7. Rafols, I.; Hoekman, J.; Siepel, J.; Nightingale, P.; Hopkins, M.M.; O'Hare, A.; Perianes-Rodriguez, A. Big Pharma, Little Science? A Bibliometric Perspective on Big Pharma's R&D Decline. *SSRN Electron. J.* **2012**, *81*, 22–38. [CrossRef]
8. Gascón, F.; Lozano, J.; Ponte, B.; De La Fuente, D. Measuring the efficiency of large pharmaceutical companies: An industry analysis. *Eur. J. Health Econ.* **2016**, *18*, 587–608. [CrossRef]
9. Jiang, H.; He, Y. Applying Data Envelopment Analysis in Measuring the Efficiency of Chinese Listed Banks in the Context of Macroprudential Framework. *Mathematics* **2018**, *6*, 184. [CrossRef]
10. Kumbhakar, S.C.; Lien, G.; Hardaker, J.B. Technical efficiency in competing panel data models: A study of Norwegian grain farming. *J. Prod. Anal.* **2012**, *41*, 321–337. [CrossRef]
11. Wang, C.-N.; Nguyen, M.N.; Le, A.L.; Tibo, H. A DEA Resampling Past-Present-Future Comparative Analysis of the Food and Beverage Industry: The Case Study on Thailand vs. Vietnam. *Mathematics* **2020**, *8*, 1140. [CrossRef]
12. Chen, C.F.; Soo, K.T. Some university students are more equal than others: Efficiency evidence from England. *Econ. Bull.* **2010**, *30*, 2697–2708.
13. Lozano, S.; Gutiérrez, E. A slacks-based network DEA efficiency analysis of European airlines. *Transp. Plan. Technol.* **2014**, *37*, 623–637. [CrossRef]

14. Lin, B.-H.; Lee, H.-S.; Chung, C.-C. The Construction and Implication of Group Scale Efficiency Evaluation Model for Bulk Shipping Corporations. *Mathematics* **2020**, *8*, 702. [CrossRef]
15. Zhou, Z.; Jin, Q.; Peng, J.; Xiao, H.; Wu, S. Further Study of the DEA-Based Framework for Performance Evaluation of Competing Crude Oil Prices' Volatility Forecasting Models. *Mathematics* **2019**, *7*, 827. [CrossRef]
16. Kuosmanen, T.; Saastamoinen, A.; Sipiläinen, T. What is the best practice for benchmark regulation of electricity distribution? Comparison of DEA, SFA and StoNED methods. *Energy Policy* **2013**, *61*, 740–750. [CrossRef]
17. Cherchye, L.; De Rock, B.; Walheer, B. Multi-output efficiency with good and bad outputs. *Eur. J. Oper. Res.* **2015**, *240*, 872–881. [CrossRef]
18. Orea, L.; Llorca, M.; Filippini, M. A new approach to measuring the rebound effect associated to energy efficiency improvements: An application to the US residential energy demand. *Energy Econ.* **2015**, *49*, 599–609. [CrossRef]
19. Alarenan, S.; Gasim, A.A.; Hunt, L.C.; Muhsen, A.R. Measuring underlying energy efficiency in the GCC countries using a newly constructed dataset. *Energy Transit.* **2019**, *3*, 31–44. [CrossRef]
20. Ahn, H.; Afsharian, M.; Emrouznejad, A.; Banker, R.D. Recent developments on the use of DEA in the public sector. *Socio-Econ. Plan. Sci.* **2018**, *61*, 1–3. [CrossRef]
21. Sueyoshi, T.; Yuan, Y.; Goto, M. A literature study for DEA applied to energy and environment. *Energy Econ.* **2017**, *62*, 104–124. [CrossRef]
22. Odeck, J.; Bråthen, S. A meta-analysis of DEA and SFA studies of the technical efficiency of seaports: A comparison of fixed and random-effects regression models. *Transp. Res. Part A Policy Pract.* **2012**, *46*, 1574–1585. [CrossRef]
23. Fall, F.; Akim, A.-M.; Wassongma, H. DEA and SFA research on the efficiency of microfinance institutions: A meta-analysis. *World Dev.* **2018**, *107*, 176–188. [CrossRef]
24. Marchetti, D.; Wanke, P.F. Efficiency in rail transport: Evaluation of the main drivers through meta-analysis with resampling. *Transp. Res. Part A Policy Pract.* **2019**, *120*, 83–100. [CrossRef]
25. Emrouznejad, A.; Yang, G.-L. A survey and analysis of the first 40 years of scholarly literature in DEA: 1978–2016. *Socio-Econ. Plan. Sci.* **2018**, *61*, 4–8. [CrossRef]
26. Emrouznejad, A.; Thanassoulis, E. A mathematical model for dynamic efficiency using data envelopment analysis. *Appl. Math. Comput.* **2005**, *160*, 363–378. [CrossRef]
27. Hu, X.-Y.; Li, J.; Li, X.; Cui, J. A Revised Inverse Data Envelopment Analysis Model Based on Radial Models. *Mathematics* **2020**, *8*, 803. [CrossRef]
28. Wei, G.-W.; Wang, J. A comparative study of robust efficiency analysis and Data Envelopment Analysis with imprecise data. *Expert Syst. Appl.* **2017**, *81*, 28–38. [CrossRef]
29. Khezrimotlagh, D.; Zhu, J.; Cook, W.D.; Toloo, M. Data envelopment analysis and big data. *Eur. J. Oper. Res.* **2019**, *274*, 1047–1054. [CrossRef]
30. You, T.; Chen, X.; Holder, M.E. Efficiency and its determinants in pharmaceutical industries: Ownership, R&D and scale economy. *Appl. Econ.* **2010**, *42*, 2217–2241. [CrossRef]
31. Mao, Y.; Li, J.; Liu, Y. Evaluating business performance of China's pharmaceutical companies based on data envelopment analysis. *Stud. Ethno-Med.* **2014**, *8*, 51–60. [CrossRef]
32. Sueyoshi, T.; Goto, M. DEA radial measurement for environmental assessment: A comparative study between Japanese chemical and pharmaceutical firms. *Appl. Energy* **2014**, *115*, 502–513. [CrossRef]
33. Hashimoto, A.; Haneda, S. Measuring the change in R&D efficiency of the Japanese pharmaceutical industry. *Res. Policy* **2008**, *37*, 1829–1836. [CrossRef]
34. Al-Refaie, A.; Wu, C.-W.; Sawalheh, M. DEA window analysis for assessing efficiency of blistering process in a pharmaceutical industry. *Neural Comput. Appl.* **2018**, *31*, 3703–3717. [CrossRef]
35. Mazumdar, M.; Rajeev, M.; Ray, S.C. *Output and Input Efficiency of Manufacturing Firms in India: A Case of the Indian Pharmaceutical Sector*; Institute for Social and Economic Change: Bangalore, India, 2009.
36. Simar, L.; Wilson, P.W. Estimation and inference in two-stage, semi-parametric models of production processes. *J. Econ.* **2007**, *136*, 31–64. [CrossRef]
37. Bloom, N.; Lemos, R.; Sadun, R.; Scur, D.; Van Reenen, J. International Data on Measuring Management Practices. *Am. Econ. Rev.* **2016**, *106*, 152–156. [CrossRef]

38. Bénabou, R.; Tirole, J. Mindful Economics: The Production, Consumption, and Value of Beliefs. *J. Econ. Perspect.* **2016**, *30*, 141–164. [CrossRef]
39. Aigner, D.; Lovell, C.; Schmidt, P. Formulation and estimation of stochastic frontier production function models. *J. Econ.* **1977**, *6*, 21–37. [CrossRef]
40. Meeusen, W.; Broeck, J.V.D. Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error. *Int. Econ. Rev.* **1977**, *18*, 435. [CrossRef]
41. Kumbhakar, S.C.; Parmeter, C.F.; Zelenyuk, V. Stochastic frontier analysis: Foundations and advances. In *Handbook of Production Economics*; Springer: New York, NY, USA, 2017.
42. Charnes, A.; Cooper, W.; Rhodes, E. Measuring the efficiency of decision making units. *Eur. J. Oper. Res.* **1978**, *2*, 429–444. [CrossRef]
43. Banker, R.D.; Charnes, A.; Cooper, W.W. Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. *Manag. Sci.* **1984**, *30*, 1078–1092. [CrossRef]
44. Van Dijk, B. *Amadeus Database*; Bureau van Dijk Electronic Publishing: Brussels, Belgium, 2020.
45. Eurostat. Available online: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=prc_hicp_aind&lang=en (accessed on 5 May 2020).
46. Henderson, R.; Cockburn, I. Scale, Scope and Spillovers: The Determinants of Research Productivity in the Pharmaceutical Industry. *RAND J. Econ.* **1993**, *27*, 32–59. [CrossRef]
47. Cockburn, I.; Henderson, R.M. Scale and scope in drug development: Unpacking the advantages of size in pharmaceutical research. *J. Health Econ.* **2001**, *20*, 1033–1057. [CrossRef]
48. Danzon, P.M.; Nicholson, S.; Pereira, N.S. Productivity in pharmaceutical-biotechnology R&D: The role of experience and alliances. *J. Health Econ.* **2005**, *24*, 317–339. [CrossRef] [PubMed]
49. Arrow, K.J. The Economic Implications of Learning by Doing. *Rev. Econ. Stud.* **1962**, *29*, 155. [CrossRef]
50. Hay, M.; Thomas, D.W.; Craighead, J.L.; Economides, C.; Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* **2014**, *32*, 40–51. [CrossRef]
51. Bravo-Ureta, B.E.; Solís, D.; López, V.H.M.; Maripani, J.F.; Thiam, A.; Rivas, T. Technical efficiency in farming: A meta-regression analysis. *J. Prod. Anal.* **2006**, *27*, 57–72. [CrossRef]
52. Hoff, A. Second stage DEA: Comparison of approaches for modelling the DEA score. *Eur. J. Oper. Res.* **2007**, *181*, 425–435. [CrossRef]
53. McDonald, J. Using least squares and tobit in second stage DEA efficiency analyses. *Eur. J. Oper. Res.* **2009**, *197*, 792–798. [CrossRef]
54. Banker, R.D.; Natarajan, R. Evaluating Contextual Variables Affecting Productivity Using Data Envelopment Analysis. *Oper. Res.* **2008**, *56*, 48–58. [CrossRef]
55. Greene, W. The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *Econ. J.* **2004**, *7*, 98–119. [CrossRef]
56. Bengoa, M.; Sanchez-Robles, B. Policy shocks as a source of endogenous growth. *J. Policy Model.* **2005**, *27*, 249–261. [CrossRef]
57. Bengoa-Calvo, M.; Sanchez-Robles, B.; Shachmurove, Y. Back to BITs and Bites: Do Trade and Investment Agreements Promote Foreign Direct Investment within Latin America? *SSRN Electron. J.* **2017**, 3083980. [CrossRef]
58. Casper, S.; Matraves, C. Institutional frameworks and innovation in the German and UK pharmaceutical industry. *Res. Policy* **2003**, *32*, 1865–1879. [CrossRef]
59. Scherer, F.; Kleinke, J. The Link Between Gross Profitability and Pharmaceutical R&D Spending. *Health Aff.* **2001**, *20*, 216–220. [CrossRef]
60. Lakdawalla, D.N. Economics of the Pharmaceutical Industry. *J. Econ. Lit.* **2018**, *56*, 397–449. [CrossRef]
61. Mondrego, A.; Barge-Gil, A. La I+D en el sector farmacéutico español en el periodo 2003–2015. *Pap. Econ. Esp.* **2019**, *160*, 76–93.
62. Greene, W.H. *Econometric Analysis Fifth Edition*; Prentice Hall: New York, NY, USA, 2003.
63. Eicker, F. Limit theorems for regressions with unequal and dependent errors. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Davis, CA, USA, 21 June–18 July 1965.
64. Huber, P.J. The behavior of maximum likelihood estimates under nonstandard conditions. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Davis, CA, USA, 21 June–18 July 1965; pp. 221–233.

65. White, H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* **1980**, *48*, 817. [CrossRef]
66. Stock, J.; Watson, M. Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression. *NBER Tech. Work. Pap.* **2006**, *323*. [CrossRef]

mathematics

MDPI

Article

The Net Worth Trap: Investment and Output Dynamics in the Presence of Financing Constraints

Jukka Isohätälä [1], Alistair Milne [2,*] and Donald Robertson [3]

[1] Institute of Operations Research and Analytics, National University of Singapore, 3 Research Link, Innovation 4.0 04-01, Singapore 117602, Singapore; jukka.isohatala@gmail.com
[2] School of Business and Economics, Loughborough University, Epinal Way, Loughborough LE11 3TU, UK
[3] Faculty of Economics, University of Cambridge, Cambridge CB3 9DD, UK; dr10011@cam.ac.uk
* Correspondence: a.k.l.milne@lboro.ac.uk

Received: 13 July 2020; Accepted: 7 August 2020; Published: 10 August 2020

Abstract: This paper investigates investment and output dynamics in a simple continuous time setting, showing that financing constraints substantially alter the relationship between net worth and the decisions of an optimizing firm. In the absence of financing constraints, net worth is irrelevant (the 1958 Modigliani–Miller irrelevance proposition applies). When incorporating financing constraints, a decline in net worth leads to the firm reducing investment and also output (when this reduces risk exposure). This negative relationship between net worth and investment has already been examined in the literature. The contribution here is providing new intuitive insights: (i) showing how large and long lasting the resulting non-linearity of firm behaviour can be, even with linear production and preferences; and (ii) highlighting the economic mechanisms involved—the emergence of shadow prices creating both corporate prudential saving and induced risk aversion. The emergence of such pronounced non-linearity, even with linear production and preference functions, suggests that financing constraints can have a major impact on investment and output; and this should be allowed for in empirical modelling of economic and financial crises (for example, the great depression of the 1930s, the global financial crisis of 2007–2008 and the crash following the Covid-19 pandemic of 2020).

Keywords: cash flow management; corporate prudential risk; the financial accelerator; financial distress; induced risk aversion; liquidity constraints; liquidity risk; macroeconomic propagation; multiperiod financial management; non-linear macroeconomic modelling; Tobin's q; precautionary savings

JEL Classification: E44

1. Introduction

This paper examines the impacts of financing constraints on firm operations and finances using the tools of continuous time dynamic stochastic optimisation. The introduction of a threshold whereat the firm faces costly refinancing or liquidation, changes the behaviour of a firm, even when the firm's shareholders are risk neutral. As the threshold is approached, there emerges an increasing premium on the value of cash held inside the firm (relative to the outside cost of capital) and an increasing aversion to risk.

This modelling builds on the dynamic analyses of [1,2]. Following [2] firms have constant returns to scale, i.e., linear production technology, and seek to maximise the value of cash dividends paid to share- holders whilst facing convex costs of adjusting their capital stock. With these assumptions the solution to the firm's optimisation problem can be expressed in terms of a single state variable, the ratio of balance sheet net worth to productive capital. Cash held internally reduces shareholder returns but

also lowers the expected future costs of refinancing or liquidation. It is the interplay between these two forces that drives behaviour.

The financing constraint is assumed to be an exogenous (to the firm) lower boundary for this state variable at which the firm must undertake costly refinancing or bankruptcy occurs and the firm is liquidated. An upper bound appears as part of the optimal solution and marks the threshold where the firm pays out cash flow dividends to its impatient shareholders. As shocks drive the firm closer to the liquidation threshold, its presence increasingly affects the optimal decisions of the firm.

An analogy for this mechanism is provided by Whittle [3] pages 287–288:

> This might be termed the "fly-paper" effect... A deterministic fly, whose path is fully under its own control, can approach arbitrarily closely to the fly-paper with impunity, knowing he can avoid entrapment ... A stochastic fly cannot guarantee his escape; the nearer he is to the paper, the more certain it is that he will be carried onto it. This also explains why the fly tries so much harder in the stochastic case than in the deterministic case to escape the neighbourhood of the fly-paper... One may say that the penalty of ending on the fly-paper "propagates" into the free-flight region in the stochastic case, causing the fly to take avoiding action while still at a distance from the paper.

Whittle is pointing out that in dynamic settings with (i) uncertainty in the equations of motion for state variables (the position of the fly); and (ii) constraints on state (the fly paper), then (in the language of economic theory) non-zero shadow prices appear even for values of the state where the constraints are not currently binding.

In the setting of this paper there are two routes by which this mechanism affects firm decisions:

(1) A shadow price of internal funds creates a wedge between the internal and external cost of capital. This reduces the marginal valuation of investment in terms of internal funds (Tobin's marginal-q). A firm invests less and less as its net worth declines. The consequence is corporate prudential saving, analogous to the household prudential saving extensively discussed in the literature on the consumption function (surveyed in [4]).
(2) A shadow price of risk creates an "induced risk aversion" leading to firms reducing their risk exposure. A variation of the model allows firms to respond by renting out more and more of their capital as net worth declines below a threshold level. It is this mechanism which, if sufficiently powerful, creates the "net worth trap."

Since the firm cannot raise new equity capital, declines in net worth are financed by increases in firm borrowing. When this state variable is comparatively high, near the upper boundary where dividends are paid, then these shadow prices are close to those that would apply for a financially unconstrained firm that can raise new equity. As the state variable falls closer to its minimum level at which no further borrowing is possible, these shadow prices rise. Even with the assumed linear production technology and the risk-neutrality of firm shareholders, the resulting dynamics of corporate output and investment are highly non-linear, depending on both the direction and size of shocks: negative shocks to productivity or net worth have a larger impact than positive shocks; small negative shocks self-correct relatively quickly; larger negative shocks (or a succession of smaller negative shocks) can result in extended periods of high shadow prices and contractions of output and investment.

The remainder of the paper is set out as follows. Section 2 locates the paper in the economics, finance and mathematical insurance literature. Section 3 presents a simplified version of the model in which capital cannot be rented out. For high values of the fixed cost of recapitalisation, firms do not recapitalise instead liquidating on the lower net worth boundary; but for lower values firms choose to exercise their option to recapitalise on the lower boundary and so avoid liquidation. In either case, investment is reduced below unconstrained levels by the state dependent shadow price of internal funds.

Section 4 then introduces the possibility that firms, by mothballing or renting capital to outsiders, are able to reduce their risk exposure, but at the expense of a decline in their expected output.

The extent to which this is done depends on the magnitude of a shadow price of risk capturing an effective induced risk aversion for other wise risk-neutral shareholders. This is where the possibility of a net worth trap emerges. Section 5 provides a concluding discussion considering the macroeconomic implications of these findings. While the model solution is numerical, not closed form, we have developed convenient and rapid solution routines in Mathematica. (We have created a standalone module which can be used by any interested reader to explore the impact of parameter choice on model outcomes. This, together with the Mathematica notebook used for creating the Figures reported in the paper, can be downloaded via http://leveragecycles.lboro.ac.uk/networthtrap.html and run using the free Mathematica Player software https://www.wolfram.com/player/). Four appendices contain supporting technical details. (Appendix A solves the situation in which there is no non-negativity constraint on dividend payments; or equivalently when uncertainty vanishes. Appendix B provides proofs of the propositions in the main text. Appendix D derives the asymptotic approximations used to incorporate the singularities that arise in the model with rent. Appendix C details the numerical solution, noting how this must be handled differently in the two possible cases, wherein a "no Ponzi" condition applies to the unconstrained model of Appendix A; and when this condition does not).

2. Related Literature

There is a substantial body of literature examining firm operations, financing and risk management over multiple periods. Central to this work is the inventory theoretic modelling (initiated by [5,6])) of both financial (cash, liquidity and capitalisation) and operational (inventory, employment, fixed capital investment) decisions subject to fixed (and sometimes also proportional or convex) costs of replenishment or investment.

Most dynamic models of corporate behaviour focus either on financial or operational decisions without considering their interaction. Well known contributions include work on the dynamics of fixed capital investment in the presence of adjustment costs (including [7,8]); and on applying standard tools of inventory modelling to study corporate cash holdings and money demand [9–11]. Dynamic modelling methods are also employed in the contingent claims literature, to examine both the pricing of corporate liabilities [12] and the possibility of strategic debt repudiation [13,14] and the interaction of the choice of asset risk and capital structure, taking account of the implications for the cost of debt [15]. However, this line of research does not address the dynamic interaction of financing and investment.

The interaction of financial and operational decisions is often considered in a static framework. This allows an explicit statement of the informational asymmetries and strategic interactions that lead to departures from the [16] irrelevance proposition (for a unified presentation of much of this literature, see [17]). This is widely used in the corporate finance literature. Take, for example, the pecking order theory of capital structure in which costs of equity issuance result in discrepancies between the costs of inside funds (retained earnings), debt and outside equity [18,19]. In [20] such a static framework was applied to develop a joint framework of the determination of investment and risk management decisions.

There is a smaller body of literature on the dynamic interactions of financial and operational decisions. The negative relationship between net worth and the shadow price of internal funds that appears in the present paper is not a new finding; it appears in a number of other contributions to the literature. In [21] the costly state verification problem of [22] is extended into a recursive model of dynamic stochastic control wherein one period debt contract can be refinanced through a new debt contract. His analysis does not establish an explicit solution for the optimal contract, but it does show how if debt contracts are used to dynamically finance a productive investment opportunity, then the value function has a "characteristic" convex shape, with a negative second derivative with respect to net worth, reflecting a departure from [16] capital structure irrelevance and a resulting shadow price of internal funds. In consequence, as net worth declines so does investment and output.

Progress has been made more recently on analysing optimal financial contracts in a dynamic principal agent context (see [23,24] and references therein), yielding similar divergence between the

cost of funds. In [24] it is shown that it can be optimal for a firm to use, simultaneously, both long term debt and short term lines of credit, in order to create incentives for managerial effort, but this work has not been extended to modelling the interactions of financial and operational decisions.

Most other work on the dynamic interactions of financial and operational decisions has proceeded, as in this paper, by imposing costly financial frictions (rather than establishing an optimal contract). Many of these papers employ continuous time modelling techniques. An early example is [25] exploring the bond financing of a project subject to fixed costs both of opening and shutting the project (hence creating real option values) and of altering capital structure through bond issuing. Four papers written independently [1,26–28] explore cash flow management and dividend policy in a context where cash holdings evolve stochastically (as a continuous time diffusion) resulting in a need for liquidity management. This leads to the simple boundary control for dividends that is inherited by the model of the present paper: paying no dividends when net cash holdings are below a target level and making unlimited dividend payments on this boundary.

This cash flow management framework was subsequently employed in a variety of different contexts. These include the risk exposure decisions of both insurance companies and non-financial corporates (see [29–31]). In a sequence of papers [32–35] bank capital regulation and bank behaviour are analysed. Other related work examined how intervention rules affect market pricing in exchange rates and in money markets; for example, [36–38] provides a survey article linking this work to portfolio allocation and cash management problems faced by companies and insurance firms.

Other recent and closely related studies exploring the interactions of financing, risk management and operational decisions include [39–44]. While employing differing assumptions, these papers have a great deal in common. The resulting dynamic optimisation again yielded a value function with the "characteristic" convex shape reported by [21] and appearing also in this paper, and hence resulted in internal "shadow prices" which reduce risk exposure, output and investment as net worth or cash holdings decline.

Similar findings emerged in discrete time models, such as those of [45,46], who considered risk management and firm decision making in the presence of taxation and imposed costs of financial transactions. They incorporated a wide range of determining factors and firm decision variables, again finding that a reduction in net worth leads to reduced of risk exposure and increased incentives to hedge risks.

While the literature offers a consistent account of the dynamic interactions of corporate financing and operational decisions, the macroeconomic implications are less fully explored. Capital market frictions, in particular the high costs of external equity finance and the role of collateral values have been proposed as an explanation of macroeconomic dynamics (see, e.g., [47,48]). The most widely used implementation of these ideas is the "financial accelerator" introduced into macroeconomics by [49,50]. This is based on a static model of underlying capital market frictions, in which the macroeconomic impact of financing constraints comes through assuming a costly state verification (as in [22]) and modelling the resulting difficulties entrepreneurs face in obtaining external finance for new investment projects. The resulting propagation mechanism operates through an "external financing premium", i.e., a additional cost that must be paid by investors in fixed capital projects in order to overcome the frictional costs of external monitoring whenever they raise external funds, rather than internal shadow prices such as appear in this paper.

An alternative perspective on the propagation of macroeconomic shocks is found in the literature on endogenous risk in traded asset markets (see [51–53] in which asset price volatility limits access to external finance. In their overview of the impacts of financing constraints on macrofinancial dynamics using continuous-time modelling [54] highlighted some further recent analysis of this kind [2,55,56] that focused on the impacts of financial constraints on asset prices.

In [55,56] "specialists" were able to better manage financial assets but their ability to invest in these assets was constrained by their net worth. The outcome was two regimes: one of higher net worth wherein the constraint does not bind and the pricing and volatility of assets are determined

by fundamental future cash flows; the other of lower net worth wherein asset prices fall and asset price volatility rises relative to fundamental levels. This approach is developed further in [2], treating physical capital as a tradable financial asset and showing how endogenous volatility can then limit investment and create the possibility of a "net worth" trap with extended declines in output and investment following a major negative shock.

The analysis of this paper shows that such a net worth trap can also emerges through the operation of internal shadow prices, rather than, as in [2], through external market pricing. Thus there is a potential net worth trap for all firms, large and small and regardless of the extent to which they participate in external financial and asset markets. This paper also extends [2] by allowing not just for shocks to capital productivity (these can be interpreted as supply shocks such as those that have resulted from the Covid-19 pandemic) but also for shocks to net worth (these can be interpreted as demand and financial market shocks, such as those that occurred during the global financial crisis). As in [2], in order to limit the number of state variables and obtain tractable results, this paper considers only serially uncorrelated shocks represented in continuous time by a Wiener process.

This paper contributes to this literature in two further ways. First, building on [1] it distinguishes the impact on firm decisions of the shadow price of internal funds (the first derivative of the value function in net worth) from the shadow price of risk (the scaled second derivative of the value function). Second, it develops efficient numerical solution methods which support convenient exploration of the effect of changing parameters on firm decisions.

3. A Basic Model

This section solves a basic model in which firms decide only on investment and dividend payments. We delay to Section 4 consideration of a broader model in which firms can reduce their risk exposure by selling or renting capital to outsiders, and the circumstances in which a net worth trap might then appear.

Section 3.1 sets out the model assumptions. Section 3.2 discusses its numerical solution. Section 3.3 presents some illustrative simulations of the numerical solution.

3.1. Model Assumptions

Firms produce output sold at price unity through a constant returns to scale production function ($y = ak$) and seek to maximise a flow of cash dividends ($\lambda \geq 0$) to risk neutral owners/shareholders. There are two state variables, capital k, which is augmented by investment at rate i and reduced by depreciation at rate δ, and cash, which increases with sale of output and is reduced by capital investment (subject to quadratic adjustment costs) and dividend payouts. Cash c held internally (which can be negative if borrowing) attracts an interest rate r. Additionally, the firm can recapitalise (increase its cash) at any time τ by an amount $\epsilon(\tau)$ where τ can be chosen by the firm. Cash holdings are disturbed by an amount $\sigma k dz$ with dz a Wiener process.

The state variables evolve according to:

$$dc = \left[-\lambda + ak + rc - ik - \frac{1}{2}\theta (i-\delta)^2 k\right] dt + \epsilon(\tau) + \sigma k dz \tag{1a}$$

$$dk = (i-\delta)k dt \tag{1b}$$

where the coefficient θ captures costs of adjustment of the capital stock (increasing with the net rate of investment $i - \delta$).

The firm seeks to maximise the objective:

$$\Omega = \max_{\{i_t\},\{\lambda_t\},\{\epsilon_\tau\}} E \int_{t=0}^{\infty} e^{-\rho t} \lambda dt - \sum_{\tau=\tau_1}^{\tau_\infty} e^{-\rho \tau} \left(\epsilon_\tau + \chi k\right) \tag{2}$$

where $\chi > 0$ represents the cost to shareholders of recapitalisation, arising from any associated due-diligence or dilution of interests. These are assumed proportional to k.

The only other agents are outside investors ("households" in the terminology of [2] who lend to firms, but do not take credit risks; instead, they require that lending is secured against the firm's assets, limiting the amount of credit available to the firm, and they become the residual owner of the firm's assets if and when the debt is not serviced. Like firm owners, these investors are risk-neutral and seek to maximise the present discounted value of current and future consumption. Unlike firms, there is no non-negativity constraint on their cash flow. Since they are the marginal suppliers of finance, and there is no risk of credit losses, they lend to or borrow from firms at a rate of interest r reflecting their rate of time discount. Investors are more patient than firms, i.e., $r < \rho$ (without this assumption firms will build up unlimited cash holdings instead of paying dividends). Fixed capital held directly by outside investors generates an output of $\bar{a}k$.

Further assumptions are required in order to obtain a meaningful solution: (i) capital is less productive in the hands of outside investors than when held by firms (otherwise, firms will avoid using capital for production), $\bar{a} < a$; (ii) upper bounds on both a and $\bar{a}$ to ensure that the technology does not generate sufficient output to allow self-sustaining growth faster than the rates of shareholder or household discount; (iii) a further technical condition (a tighter upper bound on a) ensuring that there is a solution in which dividends are paid to firm shareholders.

3.2. Solution

3.2.1. Characterisation of the Solution

The form of the solution is summarised in the following propositions:

Proposition 1. *The firm can borrow (hold $c < 0$) up to a maximum amount determined by the valuation of the firm by outside investors:*

$$c > \bar{\eta} k = -\left[1 + \theta\left(r - \sqrt{r^2 - 2\theta^{-1}[\bar{a} - \delta - r]}\right)\right] k \tag{3}$$

Proof. Appendix A. □

This maximum amount of borrowing ($-\bar{\eta}k$) is exogenous to the the firm but endogenous to the model. It increases with the net productivity of firms' assets in the hands of outside investors ($\bar{a} - \delta$) and decreases with the costs of investing in new capital (θ) and with the discount rate of outside investors (r) which is also the available rate of interest on cash holding/cash borrowing by the firm. It would be possible to generalise the model by enforcing either a lower exogenous limit on borrowing or a higher exogenous limit on borrowing with an interest rate on borrowing of $\bar{r}(c) > r$ for $c < 0$ that compensates lenders for the liquidated value of the firm—that being less than the amount borrowed at liquidation. We do not explore these extensions.

If c reaches this bound then the firm has a choice: either liquidate (in which case its assets are acquired by the lenders and there is no further payment to shareholders); or recapitalise (at a cost to shareholders of χk).

Proposition 2. *Sufficient conditions for an optimal policy for choice of* $\{i_t\}$, $\{\lambda_t\}$, $\{\epsilon_\tau\}$ *as functions of the single state variable* $\eta = ck^{-1}$ *to exist and satisfy* $i_t - \delta < \rho$, $\forall t$ *are*

$$a - \delta < \rho + \frac{1}{2}\theta\rho^2 - (\rho - r)\left[1 + \theta\left(r - \sqrt{r^2 - 2\theta^{-1}[\bar{a} - \delta - r]}\right)\right], \tag{4a}$$

$$\bar{a} - \delta < r + \frac{1}{2}\theta r^2. \tag{4b}$$

Further, if Equation (4a) is satisfied, the growth rate of the capital stock $g(\eta)$ *and the optimal investment rate* $i(\eta)$ *always satisfies the constraints*

$$\bar{g} = \left(r - \sqrt{r^2 - 2\theta^{-1}[\bar{a} - \delta - r]}\right) \leq g(\eta) = i(\eta) - \delta < \rho.$$

Proof. Appendix B. □

If a solution exists then it is characterised by the following further proposition:

Proposition 3. *An optimal policy choice for* $\{i_t\}, \{\lambda_t\}, \{\epsilon_\tau\}$ *as functions of the single state variable* $\eta = ck^{-1}$, *if it exists and takes the following form: (i) making no dividend payments a long as* $\bar{\eta} < \eta < \eta^*$ *for some value* η^* *of* η, *while making dividend payments at an unlimited rate if* $\eta > \eta^*$; *(ii) investing at a rate*

$$i = \delta + \theta^{-1}(q - 1)$$

where $W(\eta)k$ *is the value of* Ω *under optimal policy; and* $q(\eta)$ *representing the valuation of fixed assets by the firm (the cash price it would be willing to pay for a small increase in k) is given by:*

$$q = \frac{W}{W'} - \eta, \qquad q' = -\frac{WW''}{W'W'}, \tag{5}$$

with $q' > 0$ *whenever* $\eta < \eta^*$; *and* $W(\eta)$ *is the unique solution to the second-order differential equation over* $\eta \in [\bar{\eta}, \eta^*]$:

$$\rho\frac{W}{W'} = a - \delta + r\eta - \frac{1}{2}\sigma^2\left(-\frac{W''}{W'}\right) + \frac{1}{2}\theta^{-1}\left(\frac{W}{W'} - \eta - 1\right)^2 \tag{6}$$

obtained subject to three boundary conditions: (i) an optimality condition for payment of dividends at η^* $W''(\eta^*) = 0$ *(ii) a scaling condition* $W'(\eta^*) = 1$; *and (iii) the matching condition:*

$$W(\bar{\eta}) = \max\left[W(\eta^*) - (\eta^* - \bar{\eta} + \chi), 0\right]. \tag{7}$$

Finally, the firm recapitalises only on the lower boundary and only if $W(\bar{\eta}) > 0$ *in which case it recapitalises by increasing* η *immediately to* η^*.

Proof. Appendix B. □

This solution combines barrier control at an upper level of the state variable with either impulse control or absorption at a lower level.

- Barrier control is applied at an upper level of cash holding/borrowing $\eta^* k$, retaining all earnings when below this level and paying out all earnings that would take it beyond this level (a form of barrier control). It never holds more cash (or conducts less borrowing) than this targeted amount, and below this level no dividends are paid (as discussed in Section 2, similar barrier control appears in a number of earlier papers studying corporate decision making subject to external financing constraints).
- Impulse control through recapitalisation at a lower boundary $\bar{\eta} k$, but only if the cost to shareholders of recapitalisation is less than their valuation of the recapitalised firm. Net worth is then restored to the upper impulse control level $\eta^* k$. The value of the firm at the lower boundary $W(\bar{\eta})$ is the value at the upper boundary $W(\eta^*)$ less the total costs of recapitalisation $(\eta^* - \bar{\eta} + \chi)$.
- Absorption if instead the cost of recapitalisation at the lower boundary representing the maximum level of borrowing exceeds the valuation of the recapitalised firm. It then liquidates and the value obtained by shareholders is zero.

In the absence of financing constraints (as discussed in Appendix A), impulse control is exerted for all values of $\eta \neq 0$ to immediately enforce $\eta = 0$, leverage is no longer relevant to the decisions of the firm (the Modigliani–Miller [16] proposition applies) and the value function is linear in η and given by $\Omega = k\,(W_0 + \eta)$, $W' = 1$ and $q = W/W' - \eta = W_0$.

The outcome is very different in the presence of financing constraints. The value function is then distorted downward. As η declines towards the maximum level of borrowing, an increasing marginal valuation of cash results (the slope of W) because $W'' < 0$. This increasing marginal valuation of cash as the firm comes closer to liquidation is reflected in a curvature of the value function $\Omega = kW\,(\eta)$ characteristic of dynamic models of financing constraints (see the upper panel of Figure 1 and discussion in Section 2). See [1] for further discussion).

This higher marginal valuation of cash results in a reduction of q, the marginal or the internal cost of cash (Ω_c) relative to the marginal benefits of capital (Ω_k). The further η falls below the target η^*, the more investment is reduced in order to realise cash and stave off costly liquidation or recapitalisation.

The implications of the model for dynamic behaviour can be analysed using the steady state "ergodic distribution." The ergodic distribution, if it exists, represents both the cross-sectional distribution of many firms subject to independent shocks to cash flow and the unconditional time distribution of a single firm across states. It indicates the relative amount of time in which a firm stays in any particular state. When this is high, it visits this state often; when it is low then it visits this state rarely.

If a firm is liquidated at the lower boundary, i.e., if there is no recapitalisation, and it is not replaced by new firms, then no ergodic density exists. In order to compute an ergodic distribution and for comparability with the case of recapitalisation, an additional assumption is required: that liquidated firms are replaced at the upper dividend paying boundary. The following proposition then applies:

Proposition 4. *The pdf of the ergodic distribution is described the following first-order ODE:*

$$\frac{1}{2}\sigma^2 f' - \left[a + r\eta - \delta - \theta^{-1}(1+\eta)(q-1) - \frac{1}{2}\theta^{-1}(q-1)^2\right] f = -d. \tag{8}$$

and can be computed subject to the boundary conditions

$$f(\bar{\eta}) = 0 \tag{9}$$

and $F(\eta^*) = 1$ *where* $F(\eta) = \int_{u=\bar{\eta}}^{\eta} f(u)\, du$.

Proof. Appendix B. □

Here d is a constant representing the net flow of companies through the non-dividend paying region, until they exit at the lower boundary $\bar{\eta}$ through liquidation or recapitalisation and are replaced at the upper boundary η^*.

The interpretation of this ergodic distribution is slightly different in the two cases of recapitalisation and of liquidation. In the case of recapitalisation this represents the steady state cross-sectional distribution of firm net worth for firms hit by independent shocks *and* the proportion of time spent by a firm at each level of net worth. In the case of liquidation, it represents only the cross-sectional distribution of firm net worth and only when liquidated firms are indeed replaced at the upper boundary. (Other replacement assumptions are possible, for example, replacement at different levels of net worth in proportion to the steady state distribution of firms in which case the right-hand side of Equation (8) is replaced by $-df$. The ergodic density still represents only a cross-sectioaln distribution).

3.2.2. Numerical Solution

Appendix C presents the methods of numerical solution. The outline of these is as follows, utilising the function $q(\eta)$. Equation (6) can be written as:

$$q' = \frac{2}{\sigma^2}\left[a - \delta - (\rho - r)\eta - \rho q + \frac{1}{2}\theta^{-1}(q-1)^2\right](q + \eta). \tag{10}$$

requiring only two boundary conditions for solution: the optimality condition locating the upper boundary $q'(\eta^*) = 0$ together with the condition on the lower boundary Equation (7).

In the case of liquidation no iteration is necessary. This is because $W(\bar{\eta}) = 0$ implying from Equation (5) that $q(\bar{\eta}) = -\bar{\eta}$, i.e., the maximum amount of lending is the valuation of capital by outsiders and this determines the value of q on the lower boundary. Equation (10) is simply computed directly beginning from the lower boundary with $q = -\eta$ and continuing for higher values of η until $q' = 0$ and the upper boundary, if it exists, is located.

Iteration is required when there is recapitalisation rather than liquidation. This is because in this case $q(\bar{\eta})$ is not known, but must be determined from the matching condition $W(\bar{\eta}) = W(\eta^*) - (\eta^* - \bar{\eta} + \chi)$. Given any initial starting value for $q(\bar{\eta})$ it is possible to jointly compute both $q(\eta)$ and the accompanying value function $W(\eta)$. Iteration on the starting value $q(\bar{\eta})$ then yields the solution with recapitalisation (if one exists) with $W(\bar{\eta}) > 0$. While numerical solution is straightforward, it may fail to locate an upper boundary η^* for some combinations of parameters. This happens, for example, when the productivity of capital a is so high, and the adjustment costs of capital increase θ are so low, that output can be reinvested to increase the stock of capital faster than the discount rate of firms. (See Appendix A) for a discussion of the parameter restrictions required to prevent this in the deterministic case $\sigma = 0$). In this case the value function is unbounded and there is no meaningful solution. Extreme parameter values, for example, very low values of σ, can also result in numerical instability and failure to find a solution.

3.3. Simulation Results

Numerical solution is rapid, allowing extensive simulations of the model equations. Focusing on the shape of the ergodic distribution $f(\eta)$, one question is whether it has two peaks and can therefore help explain a transition from a high output boom to a low output slump, or instead has a single peak. In this first version of the model in this section there is always a single peak located at the maximum

value η^*, i.e., the model without rental or sale of capital does not create long lasting periods with output and investment below normal levels.

Typical value functions W together with the corresponding ergodic densities f are presented in Figure 1. For that, the chosen parameters were:

$$\begin{aligned} \rho = 0.06, \quad r = 0.05, \quad \sigma = 0.2, \\ \theta = 15.0, \quad \chi = 0.75, \\ a = 0.1, \quad \bar{a} = 0.04, \quad \delta = 0.02. \end{aligned} \tag{11}$$

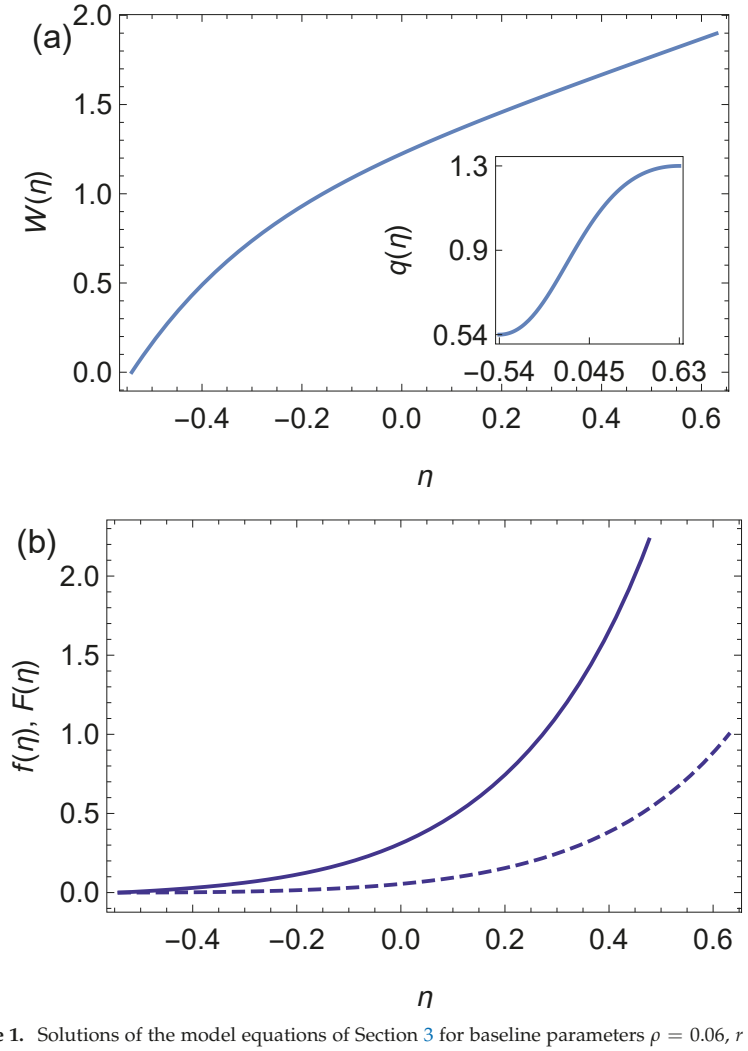


Figure 1. Solutions of the model equations of Section 3 for baseline parameters $\rho = 0.06$, $r = 0.05$, $\sigma = 0.2$, $a = 0.1$, $\bar{a} = 0.04$, $\delta = 0.02$, $\theta = 15$ and $\chi = 0.75$. Subfigure (**a**): value function W, inset shows the function q over the same η range; (**b**) the ergodic density f (solid curve) and the cumulative density function F (dashed).

Note the shape of these plots, with a monotonically increasing value function W, and a single-peaked ergodic density with a maximum at η^*. Across a wide range of parameter space search, only single-peaked distributions of this kind emerge. Although double peaks were not found, in some simulations the main peak normally at η^* can migrate into the central part of the η range. This occurs when choosing parameters for which cash-flows are non-positive $(\mathrm{d}\eta \leq 0)$. We do not report these simulations here.

4. An Extended Model

This section extends the model of Section 4 by assuming that capital can be rented by firms to outside investors. The structure of this section parallels that of Section 3, with subsections on assumptions (Section 4.1), the solution (Section 4.2) and simulation results (Section 4.3).

4.1. Additional Assumptions

In this extended setting, firms continue to manage the same two "state" variables, net cash c and capital k, but these now evolve according to:

$$\mathrm{d}c = \left\{ -\lambda + [\psi a + (1-\psi)\bar{a}]k + rc - ik - \frac{1}{2}\theta(i-\delta)^2 k \right\} \mathrm{d}t \tag{12a}$$

$$+ \epsilon(\tau) + \psi\sigma_1 k\, \mathrm{d}z_1,$$

$$\mathrm{d}k = (i-\delta)k\, \mathrm{d}t + \psi\sigma_2 k\, \mathrm{d}z_2. \tag{12b}$$

There are now two independent diffusion terms ($\psi\sigma_1 k\, \mathrm{d}z_1$ and $\sigma_2\psi k\, \mathrm{d}z_2$) and an additional third control variable, the proportion of capital ψ firms themselves manage (with remaining capital $1-\psi$ rented to households). All the other assumptions of Section 3 continue to apply.

Due to competition amongst households to acquire this capital, the amount households are willing to pay and hence the income from renting out a unit of capital is $\bar{a}$ the productivity of capital when managed by households. A special case is when $\bar{a} = 0$. In this case the renting of capital can be understood as "mothballing," taking a proportion $1-\psi$ capital out of production. In either case, whether renting or mothballing, the firm benefits from a reduction in the diffusion terms from σk to $\psi\sigma k$, protecting it from the risk of fluctuations in net worth.

The introduction of a second diffusion term is a modest extension of the model. This introduces a dependency of diffusion on the level of net worth with $\sigma^2(\eta) = \sigma_1^2 + \sigma_2^2\eta^2$ instead of a constant σ^2. The introduction of renting is a more fundamental change, leading to the possibility of a double-peaked ergodic density and the possibility of persistence of a sequence of negative shocks that push net worth down to very low levels (the "net worth trap").

4.2. Solution

4.2.1. Characterisation of Solution

Propositions 1 and 2 apply to the generalised model with renting. Proposition 3 applies in the following amended form:

Proposition 5. *An optimal policy choice for* $\{i_t\}, \{\psi_t\}, \{\lambda_t\}, \{\epsilon_\tau\}$*as functions of the single state variable* $\eta = ck^{-1}$*, if it exists, takes the following form. The rules for* $i(\eta)$,$\lambda(\eta)$ *are exactly as stated in Proposition 3;*

optimal policy for $\psi(\eta)$ renting of fixed capital is that for lower values of η, in the range $\bar{\eta} \leq \eta < \tilde{\eta}$ where $\bar{\eta} \leq \tilde{\eta} < \eta^*$, firms retain a proportion $\psi < 1$ of fixed capital given by:

$$\psi = \frac{a - \bar{a}}{\sigma^2(\eta)} \left[-\frac{W''}{W'} \right]^{-1} \tag{13}$$

and rent the remaining proportion $1 - \psi > 0$ to firms; and for $\eta \geq c$, firms retain all fixed capital, i.e., $\psi = 1$ and none is rented out. Here $\sigma^2(\eta) = \sigma_1^2 + \sigma_2^2 \eta^2$ $W(\eta)$; the unique solution to the second-order differential equation over $\eta \in [\bar{\eta}, \eta^*]$, now obeys:

$$\begin{aligned} \rho \frac{W}{W'} = \bar{a} + (a - \bar{a})\psi - \delta + r\eta - \frac{\sigma^2(\eta)}{2} \psi^2 \left[-\frac{W''}{W'} \right] \\ + \frac{1}{2\theta} \left[\frac{W}{W'} - 1 - \eta \right]^2 \end{aligned} \tag{14}$$

Solution for W is found subject to same boundary conditions as in Proposition 3

Proof. Appendix B. □

Here $-W''/W'$ expresses the induced risk aversion created by the presence of financing constraints (Section 4 of [1] has further discussion of this induced risk aversion and a comparison with the risk loving behaviour that emerges in many standard discrete time models as a result of moral hazard). $-W''/W'$ appears also in Proposition 3, the solution for the model with no option to rent out capital. There though, while it appears in Equation (6) the second-order differential equation for the value function, it has no direct impact on firm decisions. Now in Proposition 5, induced risk aversion $-W''/W'$ has a direct impact on firm decisions once net worth η falls below $\tilde{\eta}$. Renting out productive capital to households then reduces both the drift and the diffusion of η.

The introduction of the option to rent out capital introduces a second component to the behaviour of the firm. Now, and with the reduction of investment in the basic model because of a higher internal cost of capital, they can also reduce their employment of capital as a response to higher induced risk aversion. As a consequence of reduction in the employment of capital, in effect a "shrinking" of the size of operations, the firm can get "stuck" near the bankruptcy threshold, leading to a second peak in the ergodic distribution.

The resulting ergodic density can be computed using this Proposition (an indirect statement is used because of the dependency of ψ and σ on η. While ϕ can be substituted out from Equation (15) the resulting ODE for f is rather cumbersome):

Proposition 6. *The pdf of the ergodic distribution is described by the following first-order ode:*

$$\phi' - \left[\frac{1}{2} \psi^2 \sigma^2(\eta) \right]^{-1} \left[a + (a - \bar{a})\psi + r\eta - \delta - \theta^{-1}(1 + \eta)(q - 1) - \frac{1}{2}\theta^{-1}(q - 1)^2 \right] \phi \;=\; -d \tag{15}$$

where $\phi = \psi^2 \sigma^2 f/2$ and satisfies the boundary conditions

$$\begin{cases} f(\bar{\eta}) = 0, & \text{if } W(\bar{\eta}) > 0 \\ d = 0 & \text{if } W(\bar{\eta}) = 0 \end{cases} \tag{16}$$

and $F(\eta^*) = 1$*where* $F(\eta) = \int_{u=\bar{\eta}}^{\eta} f(u)\, du$.

Proof. Appendix B. □

4.2.2. Numerical Calculation

Numerical solution methods are again detailed in Appendix C. This proceeds in the same way as for the first model without renting of Section 3, by re-expressing Equation (14) as a differential equation in q. Over the lower region $\eta < \bar{\eta}$(Equation (14)) becomes:

$$q' = -\frac{1}{2}\frac{(a-\bar{a})^2}{\sigma_1^2 + \eta^2\sigma_2^2}\frac{q+\eta}{\bar{a} - \delta + r\eta - \rho(q+\eta) + \frac{1}{2}\theta^{-1}(q-1)^2} \tag{17}$$

while in the upper region Equation (10) continues to apply (except that now $\sigma^2 = \sigma_1^2 + \sigma_2^2\eta^2$ is a function of η).

If there is no recapitalisation then the model can again be solved without iteration, commencing the calculation at $\eta = \bar{\eta}$ and continuing until the intermediate values $\eta = \tilde{\eta}$ and $\eta = \eta^*$ are located. However, in this case $q(\bar{\eta}) = -\bar{\eta}$ and hence $\psi(\bar{\eta}) = 0$, with the consequence that there are singularities in f, q and W at $\bar{\eta}$. We incorporate these singularities using asymptotic approximations summarised in the following further proposition.

Proposition 7. *W, q and ϕ close to $\bar{\eta}$ are described by:*

$$W = C_W(\eta - \bar{\eta})^{\beta}(1 + \mathcal{O}(\eta - \bar{\eta})), \tag{18}$$

where C_W is a constant and $\beta = 1/(1 + q'(\bar{\eta})) \in (0, 1]$;

$$q = \bar{q} + q'(\bar{\eta})(\eta - \bar{\eta}). \tag{19}$$

and;

$$\phi = C_\phi(\eta - \bar{\eta})^{\alpha}, \tag{20}$$

where α is given by Equation (A34) of Appendix D and C_ϕ is another constant.

This further implies that $-W''/W'$ (our measure of induced risk aversion) is divergent at $\bar{\eta} = -\bar{q}$,

$$-\frac{W''}{W'} \simeq \frac{1-\beta}{\eta - \bar{\eta}}.$$

(consistent with $\psi(\bar{\eta}) = 0$), and the ergodic density is approximated by

$$f \propto (\eta - \bar{\eta})^{\alpha - 2} \tag{21}$$

and thus diverges if $\alpha < 2$ and becomes degenerate, with the entire probability mass at $\bar{\eta}$ if $\alpha \leq 1$.

Proof. Appendix D. □

.

In the case of recapitalisation $q(\bar{\eta}) > -\bar{\eta}$ and there is are no singularities in the solution; so, while iteration ia again required to determine $q(\bar{\eta})$, this can be conducted in exactly the same way as described in Section 3.2.2 for the model without renting.

4.3. Simulation Results

As expected from the power-law shape of f, Equation (21), the option to rent can have a strong impact on the shape of the ergodic density. As an example of this, in Figure 2 plots the value function W together with q, and the probability and cumulative densities, now using baseline parameters $\rho = 0.06$, $r = 0.05$, $\sigma_1 = 0.2$, $\sigma_2 = 0.0$, $a = 0.1$, $\bar{a} = 0.04$, $\delta = 0.02$, $\theta = 15$ and $\chi = 0.75$ (identical parameters to those used in Figure 1). Whereas the value function W and q show little change when renting is introduced, the density function f changes dramatically. This time a second peak is clearly present near the left-hand side range of η values. Note that with these particular parameter values the firm chooses not to recapitalise, with χ being slightly above the critical value of around 0.74 at which recapitalisation is not worthwhile, and $W(\bar{\eta}) \approx 0.05$. The interested reader can observe, using our standalone application, how increasing χ to above this critical level results in the emergence of singularities and the divergence of $f(\eta)$ to $+\infty$ at $\eta = \bar{\eta}$.

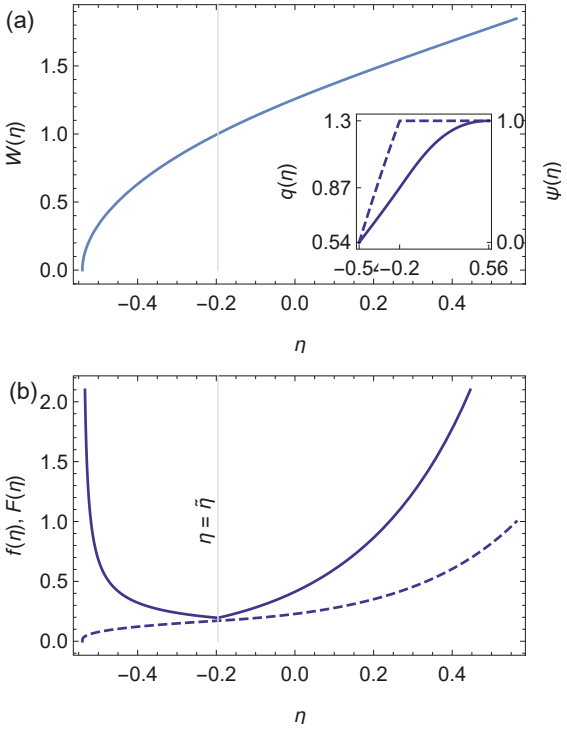


Figure 2. Solutions of the model of Section 4 with option to rent, using baseline parameters $\rho = 0.06$, $r = 0.05$, $\sigma_1 = \sigma = 0.2$, $\sigma_2 = 0.0$, $a = 0.1$, $\bar{a} = 0.04$, $\delta = 0.02$, $\theta = 15$ and $\chi = 0.75$. Contrast this to Figure 1 where identical parameters were used, but without renting. Subfigure (**a**): value function W, inset shows the functions q and ψ over the same η range; (**b**) the ergodic density f (solid curve) and the cumulative density function F (dashed). Notice the prominent peak in f towards the left-hand side boundary.

This ergodic instability (a second peak towards in the ergodic density associated with low values of the state variable η representing the ratio of cash-to-capital) is parameter dependent. This parameter

dependence emerges in two different ways: (i) through the power-law exponent α, and (ii) dependence on the cost of recapitalisation χ. The ability to recapitalise or not has a major impact on the ergodic distribution. For any given parameters, there is a threshold χ, $\bar{\chi}$, above which recapitalisation is no longer worthwhile. If χ is equal to or greater than this value, then $\psi(\bar{\eta}) = 0$, and the density diverges and the ergodic density follows the power-law $f \propto (\eta - \bar{\eta})^{\alpha-2}$ near $\bar{\eta}$, which in turn can lead to infinite densities. Hence, the strength of the instability (*i.e.* the amount of probability mass near $\bar{\eta}$) is strongly controlled by the parameter χ.

This is illustrated in Figure 3 showing how the ergodic density changes as χ is varied. For low values of χ, there is no left-hand side peak in the model with rent (Figure 3a) and f largely resembles that of the model without the option to rent (Figure 3b). As χ approaches $\bar{\chi}$ (indicated by the dotted lines on the floor of the two panels of this figure, where$\bar{\chi} \simeq 0.74$ with rent, $\bar{\chi} \simeq 0.73$ without), in the model with renting a probability mass starts to appear near $\bar{\eta}$. Crossing $\bar{\chi}$, recapitalisation becomes no longer an option, and the density at $\bar{\eta}$ diverges. Above $\bar{\chi}$ there is no longer χ dependence. Note that the distribution f changes quite sharply; when approaching $\bar{\chi}$ is crossed, with a second peak of the distribution emerging close to $\eta = \bar{\eta}$, a robust result across a variety of simulations.

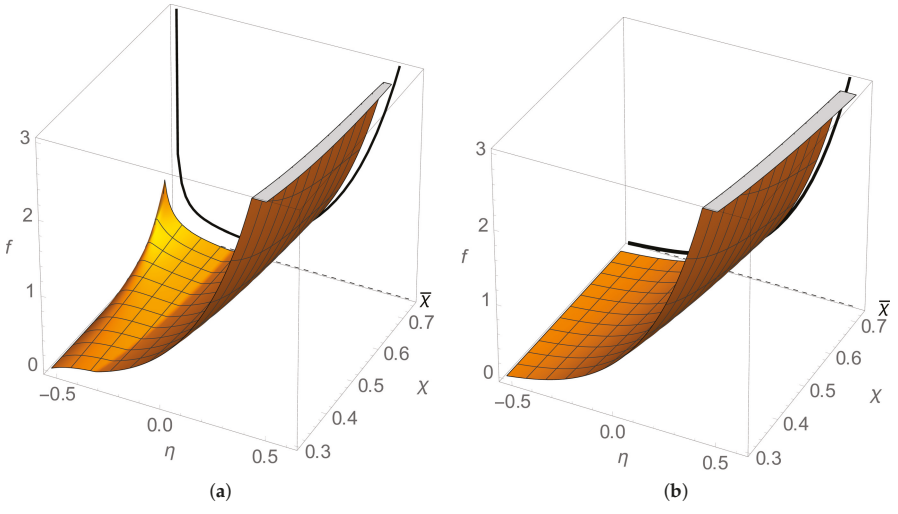


Figure 3. Comparison of ergodic densities f given the option to rent (**a**) and no option to rent (**b**) as the financing constraint χ is varied. Other parameters were set to baseline values. The lower boundary is recapitalising unto $\chi = \bar{\chi}$ ($\bar{\chi} \simeq 0.74$ in (**a**) and 0.73 in (**b**)), indicated by the thick solid line on the graph and dashed line on the axis. In (**a**) a left-boundary peak emerges for χ just less than $\bar{\chi}$. Density is infinite at $\bar{\eta}$ for $\chi > \bar{\chi}$. Note the complete absence of the left-hand side peak in (**b**).

To further explore this parameter dependence consider how the median of f depends on various parameters. Since the values of $\bar{\eta}$ and η^*, the range on which the distribution is defined, also vary with the parameters, it is convenient to scale the median on to the interval $[0, 1]$: Let m be the median; then the scaled median is defined as

$$\tilde{m} = \frac{m - \bar{\eta}}{\eta^* - \bar{\eta}}, \qquad F(m) = \frac{1}{2}. \tag{22}$$

A value of $\tilde{m} \sim 0$ implies that most of the probability mass is concentrated near $\bar{\eta}$, while $\tilde{m} \sim 1$ suggests that firms are more probably found near η^*. While this is a somewhat crude measure (e.g., the median cannot distinguish between distributions that are $\cup$ or $\cap$-shaped), nonetheless, $\tilde{m} \lesssim 1/2$ is a strong indicator of large mass of probability near the lower boundary, hence the long lasting response to a large initial shock found by [2].

In Figure 4 presents a contour plot $\tilde{m}$ as a function of the financing constraint χ and the volatility σ (note that Figure 3 represents a small slice of data presented in this figure). Three roughly distinct regimes can be seen:

(i) The low volatility range $\sigma \lesssim 0.2$, in which the firm always prefers to recapitalise and where $\tilde{m} \gtrsim 0.8$ and so most of the probability is found near the dividend paying boundary.
(ii) A region where $\sigma \gtrsim 0.3$ and at the same time $\chi \gtrsim 0.5$, i.e., red region to the top right, where $\tilde{m} \sim 0$, and much of the probability mass is located near the left-hand boundary.
(iii) An intermediate transition range wherein small changes in either σ or χ result in a very substantial change in $\tilde{m}$. This transition is especially abrupt for high values of σ.

Exploring the behaviour of $\tilde{m}$ as a function of other model parameters yields remarkably similar contour plots. For example, as the relative impatience of shareholders $\rho - r$ is increased from relatively low to high values, there are also two distinct regions similar to those of Figure 4, with a relatively sharp transition in the balance of the probability distribution from near the upper boundary η^* to the lower boundary $\bar{\eta}$.

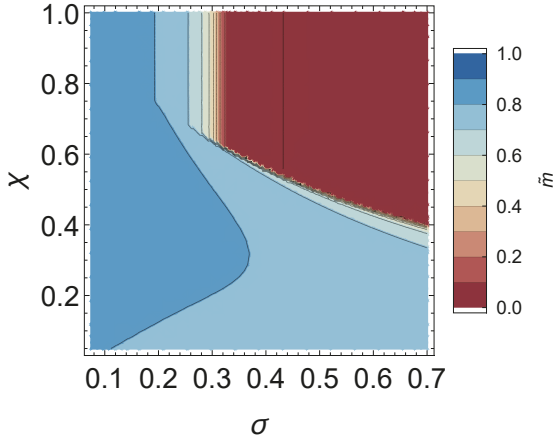


Figure 4. The scaled median $\tilde{m}$ as a function of χ and σ. $\tilde{m}$ is the median of the distribution of η relative to its range. $\tilde{m}$ close to one indicates that the mass of the distribution is located near the upper dividend paying boundary. $\tilde{m}$ close to zero indicates that the mass is located near the lower liquidation or recapitalisation boundary. $\sigma_2 = 0.0$ (the baseline value from earlier figures) so $\sigma_1 = \sigma$. Other parameters were set to baseline values; $\rho = 0.06$, $r = 0.05$, $a = 0.1$, $\bar{a} = 0.04$, $\delta = 0.02$ and $\theta = 15$. Contours are plotted at level values of $\tilde{m}$ and are spaced at intervals of 0.1.

One further finding concerns induced aversion to cash flow risk $-\frac{W''}{W'}$. This induced risk aversion is, like ergodic instability, strongly parameter and model structure dependent. In the model with renting, when firms do not recapitalise they become extremely risk-averse close to the lower boundary $\bar{\eta}$. This is revealed by an analysis of power-law behaviour of W at the lower boundary $\bar{\eta}$ (see Proposition 7). This extreme risk aversion does not arise in the model with renting or if recapitalisation is not costly.

This finding is illustrated in Figure 5 which compares induced risk-aversion for the two versions of the model, with and without the option to rent. The parameters here are the same as in Figures 1 and 2. For relatively large values of η close to η^* the option to rent provides protection against cash flow risk and induced risk aversion $-\frac{W''}{W'}$ is lower for the model with renting; but as η falls down towards $\bar{\eta}$, the model with renting induced risk aversion $-\frac{W''}{W'}$ diverges upwards—rising increasingly rapidly as η approaches $\bar{\eta}$, whereas it rises only slightly in the model without renting.

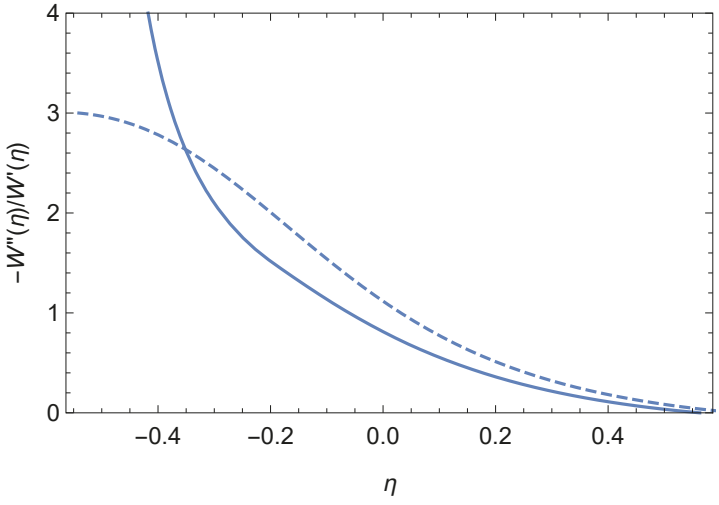


Figure 5. Induced risk aversion $-W''/W'$ as a function of η. Parameters were set to baseline. Solid curve: model with option to rent; dashed curve: model without option to rent. Significantly, the risk aversion diverges strongly near $\bar{\eta}$ in the model without renting, in contrast to the model with renting.

5. Conclusions

This paper investigated the impact of financing constraints on corporate output and investment in a simple continuous time setting with linear production and preferences. Firms face liquidation or costly recapitalisation, if their net worth (relative to capital) falls to a minimum level following a shock to cash flows or productivity. The boundary conditions resulting from these financing constraints generate potentially large and long-lasting non-linearities in the response of firm output and investment to external shocks, even in this otherwise linear setting. A fall of net worth leads to a decline of investment below normal levels. Moreover, if firms can rent out or mothball their capital stock, output also declines along with investment, and this may continue for an extended period of time: the "net worth trap".

Several further insights emerged. One is the importance of "corporate prudential saving", analogous to the household prudential saving extensively discussed in the literature on the consumption function. As their net worth declines, firms invest less and less (a fall in marginal q). The second is "induced risk aversion": firms with sufficiently high net worth have the same attitude to risk as their share holders (assumed for simplicity to be risk-neutral); but as net worth declines then firms behave increasingly as if they were averse to risk in order to reduce the probability of future liquidation or costly recapitalisation, here by mothballing or renting out more and more of their capital. Figures 4 and 5 show how resulting behaviour can vary markedly (both qualitatively and quantitatively) with parametrisation and model specification.

The finding that financing constraints mean that firm decisions can be highly non-linear functions of their indebtedness means that both the size and direction of shocks matter. This in turn helps to clarify when a linearisation—of the kind routinely employed in new Keynesian DSGE macroeconomic models—provides a reasonable approximation to the fully dynamic optimal behaviour

In normal times when shocks are comparatively small, aggregate behaviour can be sufficiently well captured in standard linearised specifications. This is illustrated by Figures 1a and 2a. In normal times most firms are located near the right-hand sides of these figures, close to the upper dividend payment boundary η^* where the value function $W(\eta)$ is approximately linear. The reserve of net worth provides and adequate hedge against both aggregate and idiosyncratic risk. Moreover, in normal times, external equity capital can, if necessary, be raised at relatively low cost; i.e., the cost of

recapitalisation parameter χ is low. As a result, impulse responses, expressed as a percentage of initial shock are then approximately the same regardless of the size or direction of the initial disturbance; linearisation of impulse responses based on past data provides a convenient and reliable summary of macroeconomic behaviour.

In times of extreme financial and economic stress, such as the Great Depression of the 1930s or the Global Financial Crisis of 2007–2008, the situation can be quite different. Uncertainty σ and the cost of recapitalisation χ rises. Large shocks push many firms towards the lower minimum level of net worth $\bar{\eta}$, where as illustrated in Figures 1a and 2a, the value function $W(\eta)$ is concave not linear and both corporate prudential saving and induced risk aversion emerge. Impulse responses based on past data are no longer a reliable guide to the response to shocks. As our Figure 4 illustrates, such an adverse change in the economic environment can lead to a "phase change," with a shift to a regime where the net worth trap emerges. The response to the Covid-19 pandemic may provide another such episode.

Most striking here, as illustrated in our Figure 4, is that relatively small parameter changes can lead to this phase change. A small increase in perceived uncertainty (our parameter σ) leads to a large change in behaviour, from relatively rapid rebuilding of physical and financial capacity following the emergence of financial distress to a slow rebuilding with a long lasting period of reduced output and investment. Small changes in the external environment faced by firms can lead to the emergence of the persistent "net worth trap." Policy makers and regulators need to be aware that whilst linear approximations may provide a good description of usual events, they can give misleading insights in more turbulent times.

Author Contributions: The mathematical modelling, proofs and the selection of illustrative figures reported in the paper are the joint work of all the authors. J.I. conducted the numerical solution and wrote most of the supporting Mathematica code, including the free standing numerical solver. A.M. and D.R. wrote the review of the literature and the economic interpretation of the model's results. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors are grateful for comments from Jussi Keppo, Feo Kusmartsev, Tassos Malliaris, Jean-Charles Rochet and Javier Suarez; and for feedback from audiences at the Bank of England, the Bank of Finland, the IBEFA January 2013 meetings, IFABS 2014 meeting, the University of Durham, Bristol University, the London School of Economics, the Bank of Japan, the National University Singapore, the University of Tasmania and internal seminar and conference presentations at Loughborough University. Remaining shortcomings are our responsibility alone.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Solution in the Absence of the Non-Negativity Constraint on Dividends

This appendix considers the solution to the model of this paper in the baseline case wherein dividend payments can be negative, or equivalently, there is no uncertainty. This provides a benchmark for studying and solving the case of the constrained firm for which there is uncertainty and dividends are required to be non-negative. It also yields a convenient formula for the maximum amount of borrowing provided by households to firms.

A crucial intuition emerges from this benchmark model, relevant to the model with a non-negativity constraint on dividend payments. The rate of growth preferred by firms is an increasing function of the ratio of debt to capital (this is because debt increases at the same rate of growth of capital, creating an additional cash flow that can be used for investment, and the higher the ratio of debt to capital, the greater this cash flow). If the financing constraint is sufficiently lax then it is possible for firms to achieve a growth rate equal to their own rate of discount while still being able to pay dividends. In this case the objective of this firm (expected discounted dividend payments)

is unbounded and the solution is no longer meaningful. Therefore, some financing constraint is required in order for the model to have a meaningful solution.

To solve this benchmark, note that since firm owners can freely transfer funds into or out of the firm, optimal policy is to maintain the ratio of cash balances $\eta = c/k$ at whatever rate is preferred by borrowers, subject to the highest level of indebtedness allowed by lenders $\eta \geq \bar{\eta}$. If the initial time $t = 0$ ratio η_0 differs from the desired ratio η then an instantaneous dividend payment of $(\eta_0 - \eta)\, k$ is immediately made to bring the cash to capital ratio to the desired value of η.

There is therefore now only a single state variable k. The value function (the value of the objective function under optimal policy) is linearly homogeneous in k and so can be written as $V = kW$, where W is a constant that depends on the parameters representing preferences and the evolution of the state variable k. This in turn implies that $V_k = W$ and $V_{kk} = 0$. Expected dividend payments will be determined by the expected net cash flow of the firm plus any additional borrowing possible because k and hence c are growing. The remaining policy decision is to choose a rate of investment i and hence expected growth of the capital stock $g = i - \delta$ to maximise Ω, Equation (2)

The solution can be summarised in the following proposition.

Proposition A1. *Assuming $\rho > r$, then an optimal policy yielding positive pay-offs for the owners of the firm can be found provided that:*

$$-2\frac{\rho - (a-\delta) + (\rho - r)\bar{\eta}}{\rho^2} < \theta < \begin{cases} \infty & \text{if } a + r\bar{\eta} \geq \delta \\ \dfrac{1}{2}\dfrac{(1+\bar{\eta})^2}{\delta - r\bar{\eta} - a} & \text{if } a + r\bar{\eta} < \delta \end{cases} \tag{A1}$$

in which case an instantaneous dividend payment of $\eta_0 - \bar{\eta}$ is made so that $\eta = \bar{\eta}$, the growth rate of the capital stock is constant (state independent) and is given by:

$$g = \rho - \sqrt{\rho^2 - 2\theta^{-1}[a - \delta - \rho + (r - \rho)\bar{\eta}]} < \rho, \tag{A2}$$

while the value of the maximised objective is given by:

$$V(\eta_0, k) = (\eta_0 - \bar{\eta})\, k + \frac{(a-\delta) + (r-g)\bar{\eta} - g - \frac{1}{2}\theta g^2}{\rho - g} k = (1 + \eta_0 + g\theta)\, k \tag{A3}$$

where $\left[(a-\delta) + (r-g)\bar{\eta} - g - \frac{1}{2}\theta g^2\right] k$ is the expected flow of dividends per period of time paid to shareholders.

Proof. The firm has two choice variables, η and g (with investment expenditure given by $ik = (g + \delta)\, k$ and associated quadratic adjustment costs of $\frac{1}{2}\theta\,(i - \delta)^2 = \frac{1}{2}\theta g^2$). The equations of motion (1) still apply and dividends are paid according to:

$$\lambda\, \mathrm{d}t = \left[(a-\delta) + (r-g)\,\eta - g - \frac{1}{2}\theta g^2\right] k\, \mathrm{d}t + \sigma k\, \mathrm{d}z$$

Substituting for λ the discounted objective can be written as:

$$\Omega = \max_{\eta, g} \left\{ \mathbb{E} \int_0^\infty \mathrm{e}^{-\rho t} \left[(a-\delta) + (r-g)\,\eta - g - \frac{1}{2}\theta g^2\right] k\ \mathrm{d}t + (\eta_0 - \eta)\, k_0 + \int \mathrm{e}^{-\rho t} \sigma k\ \mathrm{d}z \right\} \tag{A4}$$

yielding, since $\mathbb{E}[k] = k(0)\exp(gt)$ and $\mathrm{e}^{-\rho t}\sigma k = 0$:

$$\Omega = k(0) \max_{\eta, g} \left[\eta_0 - \eta + \frac{(a-\delta) + (r-g)\,\eta - g - \frac{1}{2}\theta g^2}{\rho - g}\right]. \tag{A5}$$

The growth rate g that maximises the right-hand side of this expression is determined by the first-order condition with respect to g:

$$\frac{1}{2}g^2 - \rho g - \theta^{-1}\left[\rho - (a-\delta) + (\rho - r)\eta\right] = 0 \tag{A6}$$

yielding the solution (the positive root of the quadratic can be ruled out because $g < \rho$ to ensure that the value function is finite and that the second-order condition for maximisation is satisfied):

$$g = \rho - \sqrt{\rho^2 + 2\theta^{-1}[\rho - (a-\delta) + (\rho - r)\eta]}. \tag{A7}$$

Writing $\rho - g = \sqrt{\rho^2 + 2\theta^{-1}[\rho - (a-\delta) + (\rho - r)\eta]} = R$, implying $g^2 = \rho^2 - 2\rho R + R^2$, and substitution into Equation (A5) then yields: (A3).

The indebtedness is determined by the first-order condition in (A5) with respect to η:

$$\frac{r-g}{\rho - g} - 1 = \frac{r-\rho}{\rho - g} < 0$$

establishing that the firm will seek to borrow as much as it possibly can. Hence, the firm will make an instantaneous dividend at time $t = 0$ to reduce η as far as possible, until the borrowing constraint binds so $\eta = \bar{\eta}$. The first inequality onθ in the proposition ensures that the borrowing constraint does indeed bind at a level of borrowing at which Equation (A6) has real roots.

The remaining inequality conditions on θ ensure that it is possible to achieve positive dividends per unit of capital (these are relatively weak conditions since normally $a > \delta$ in which case a policy of zero growth $g = 0$ and no indebtedness will always yield positive dividends; but if depreciation is larger than the productivity of capital then a further restriction on θ is required). To establish these further conditions note that expected dividends per unit of capital $a - \delta + r\bar{\eta} - (1 + \bar{\eta})\, g - \frac{1}{2}\theta g^2$ are maximised by choosing $g = -(1 + \bar{\eta})\,\theta^{-1}$ resulting in expected dividend payments of $\lambda = a - \delta + r\bar{\eta} + \frac{1}{2}\theta^{-1}(1 + \bar{\eta})^2$. This is always greater than zero if $a > \delta$; otherwise this requires that $\theta < (1 + \bar{\eta})^2/2(\delta - a - r\bar{\eta})$. □

This proof also shows that the Modigliani–Miller [16] proposition on the irrelevance of capital structure to the value of the firm applies, in this case when negative dividends are allowed. It does so in the sense that any net worth in excess of the minimum level $\bar{\eta}k$ is immediately paid out to shareholders and the firm always operates with maximum leverage. The value function is additive in $\bar{\eta}k$.

Finally, note that the fundamental valuation of a firm's capital by outside investors can be obtained by substituting $\rho = r$, $a = \bar{a}$ and $\bar{\eta} = 0$ into this solution. A finite positive valuation is obtained provided the parameters satisfy:

$$2\frac{\bar{a} - \delta - r}{r^2} < \theta < \begin{cases} \infty & \text{if } \bar{a} \geq \delta \\ \frac{1}{2}\frac{1}{\delta - \bar{a}} & \text{if } \bar{a} < \delta \end{cases}$$

in which case the growth rate (when held by outside investors) is given by

$$\bar{g} = r - \sqrt{r^2 - 2\theta^{-1}[\bar{a} - \delta - r]},$$

and the value of the maximised objective by

$$V = \frac{\bar{a} - \bar{g} - \delta - \frac{1}{2}\theta\bar{g}^2}{r - \bar{g}}k = (1 + \theta\bar{g})\,k.$$

This provides an immediate proof of Proposition 1 in Section 3.

Proof of Proposition 1. This valuation of the firm's assets by outside investors is also the maximum amount of debt that it can borrow from these investors, implying that the lower boundary for η is given by Equation (3). □

Appendix B. Proofs of Propositions in Sections 3 and 4

Proof of Proposition 5. (Proposition 3 requires no separate proof, since it is the special case when $\sigma_2 = 0$ and $\psi = 1$). While uniqueness of solution can be established using standard arguments based on the non-convexity of the optimisation program, the proof provided here is geometric proof, offering some additional insights into both the existence of solution and its numerical calculation.

Applying standard methods of stochastic dynamic programming, with two state variables k and c, the optimal policy by firms, at times when there is no recapitalisation ($\epsilon_t = 0$), satisfies the Hamilton–Jacobi–Bellman equation:

$$\rho V = \max_{i,\lambda,\psi} \left\{ \lambda + \left[-\lambda + (\bar{a} + (a - \bar{a})\psi) k + rc - ik - \frac{1}{2}\theta(i-\delta)^2 k \right] V_c \right. \tag{A8}$$

$$\left. + (i-\delta)kV_k + \frac{1}{2}\sigma_1^2\psi^2k^2V_{cc} + \frac{1}{2}\sigma_2^2\psi^2k^2V_{kk} \right\}, \tag{A9}$$

with three first-order conditions for maximisation. The first is:

$$\begin{cases} \lambda \geq 0 \text{ of unbounded magnitude}, & V_c = 1 \\ \lambda = 0, & V_c > 1 \end{cases},$$

there is "bang-bang" control with two distinct regions of dividend behaviour: one when $c \geq c^*(k)$ with $V_c = 1$ in which case the policy is to payout a discrete dividend to reduce cash holdings immediately to the dividend paying boundary c^*; the other when $c < c^*(k)$ wherein there is no payment of dividends and $V_c > 1$. The second first-order condition is:

$$(1 + \theta(i - \delta)) V_c = V_k$$

yielding the investment rule:

$$i = \delta + \theta^{-1}\left(\frac{V_k}{V_c} - 1\right). \tag{A10}$$

The third first-order condition for maximisation (subject to the constraint $0 \leq \psi \leq 1$) is:

$$(a - \bar{a})kV_c + \psi k^2\left(\sigma_1^2 V_{cc} + \sigma_2^2 V_{kk}\right) = 0$$

yielding the final control rule:

$$\psi = \max\left\{\min\left\{(a-\bar{a})\left[-k\frac{\sigma_1^2 V_{cc} + \sigma_2^2 V_{kk}}{V_c}\right]^{-1}, 1\right\}, 0\right\}. \tag{A11}$$

Due to the linearity of production the value function is linearly homogeneous in k and so value can be expressed as a function W of a single state variable $\eta = c/k$:

$$W(\eta) = k^{-1}V(c,k) = V(\eta, 1) \tag{A12}$$

implying the substitutions $V = kW$, $V_c = W'$, $V_k = W - \eta W'$, $q = V_k/V_c = W/W' - \eta$, $V_{cc} = k^{-1}W''$, $V_{ck} = -k^{-1}\eta W''$ and $V_{kk} = k^{-1}\eta^2 W''$ Substituting for both optimal policy and for V and its derivatives yields Equation (14). The maximisation in this second boundary condition reflects the choice available to the firm when η falls to $\bar{\eta}$; it may choose either to liquidate, in which case $W(\bar{\eta}) = 0$, or to recapitalise, which is worth doing if it can achieve a higher valuation by paying the fixed cost of recapitalisation χk and increasing η to η^*. The firm will never choose to recapitalise to a value of $\eta < \eta^*$. This is because when $\eta < \eta^*$, $V_c = W' > 1$, so the maximum possible value of $W(\bar{\eta})$ in Equation (7) is achieved by a full recapitalisation up to η^*. Turning to the uniqueness of this solution, note that as discussed in Appendix C solution of the upper boundary η^* is characterised by Equation (10) (itself obtained from

Equation (6) using $q = W/W' - \eta$ which yields $q' = -WW''/(W')^2)$. Equation (10) can be written (allowing for dependencies of σ on η) as:

$$q' = \frac{2}{\sigma_1^2 + \eta^2\sigma_2^2} Q(q,\eta)(q+\eta)$$

from which, since $\sigma_1^2 + \eta^2\sigma_2^2 > 0$ and $q^* + \eta^* > \bar{q} + \bar{\eta} \geq 0$ and $Q(q,\eta) = a - \delta - (\rho - r)\eta - \rho q + \frac{1}{2}\theta^{-1}(q-1)^2$ is a quadratic function of q and linear function of η. This in turn implies that the possible locations of q^* are given by $Q(q,\eta) = 0$, i.e., a parabola in (q,η) space, which when solved yields the location of η on the dividend paying boundaries as a function of q^*:

$$\eta^* = \frac{a - \delta - \rho q^* + \frac{1}{2}\theta^{-1}(q^* - 1)^2}{\rho - r} \tag{A13}$$

Inverting this equation to solve for $q = q^*$ on the dividend paying boundary yields:

$$q^* = 1 + \theta\left(\rho \pm \sqrt{\rho^2 - 2\theta^{-1}\{a - \delta - \rho - \eta^*(\rho - r)\}}\right) \tag{A14}$$

The uniqueness of the solution then follows (assuming continuity of $q(\eta)$) from noting that the value of $q = q^*$ is a function of the value of $q(\bar{\eta})$=$\bar{q}$ on the lower boundary. Given any starting value $\bar{q}$ the ODE characterising the solution can be computed (with $q' > 0$) until it meets $Q(q,\eta) = 0$. There can only be one such intersection. Having crossed $Q(q,\eta) = 0$, $q' < 0$ until there is another intersection, and this means any potential second intersection can only take place on the lower branch of $Q(q,\eta) = 0$. However, in order for there to be an intersection on this lower branch it is necessary that the $q-$curve falls faster than the lower branch, i.e., that on the point of intersection:

$$q' < \left.\frac{\partial q}{\partial \eta}\right|_{Q(q,\eta)=0} < 0$$

which contradicts the requirement that $q' = 0$ on $Q(q,\eta) = 0$. This contradiction shows that any solution of the ODE has at most one unique intersection with $Q(q,\eta) = 0$. □

This proof does not establish existence. While there can only be one solution to the ODE for W satisfying the boundary conditions of Proposition 5, the existence of this solution is dependent on parameter values. Proposition 2 gave sufficient conditions for a solution to exist.

Proof of Proposition 2. First note that Equation (4a) is equivalent to

$$\bar{\eta} > \overset{*}{\eta}_{\min} = -\frac{\rho - (a - \delta) + \frac{1}{2}\theta\rho^2}{\rho - r}, \tag{A15}$$

where $\eta^*_{\min}$ is the minimum value of η on the dividend paying boundary. This condition can be described as the "no-Ponzi" condition because, as stated in Proposition A1 in Appendix A when $\bar{\eta} > \overset{*}{\eta}_{\min}$ is satisfied, then the solution to the problem in the deterministic limit $\lim_{\sigma\downarrow 0}$ exists, in which the growth of the fixed capital stock is less than the discount rate of firm shareholders $g < \rho$, and the value to shareholders comes from both growth of the capital stock and dividend payments.

The idea of proof is illustrated in Figure A1. Consider possible solutions of the ODE for $q(\eta)$. In the case of no recapitalisation, $\bar{q} = -\bar{\eta}$ and the lower intersection of $Q(q,\eta) = 0$ with $\eta = \bar{\eta}$ is at $q = q^*_- = 1 + \theta(\rho - \sqrt{\rho^2 - 2\theta^{-1}\{a - \delta - \rho - \bar{\eta}(\rho - r)\}})$. This implies (using Equation (3)) that:

$$q^*_- - \bar{q} = 1 + \theta\left(\rho - \sqrt{\rho^2 - 2\theta^{-1}\{a - \delta - \rho - \bar{\eta}(\rho - r)\}}\right) + \bar{\eta} > 0$$

This shows that a solution with no recapitalisation exists, because the ODE begins at a point strictly below q^*_-, and since $q' > 0$ must eventually intersect with $Q(q,\eta) = 0$. This in turn implies the existence of solutions with recapitalisation, since these are associated with higher values of $\bar{q}$ satisfying

$-\bar{\eta} < \bar{q} < q^*_-$, in all cases with the ODE eventually intersecting with the lower branch of $Q(q, \eta) = 0$; and with values of $\chi > 0$. Eventually in the limit $\lim_{\chi \downarrow 0} \bar{q} = q^*_-$. □

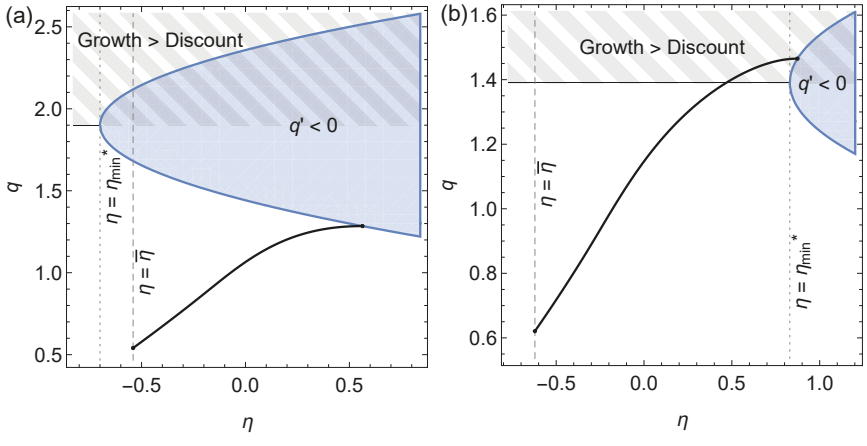


Figure A1. Illustration of Proof of Proposition 2. In subfigure (**a**), the "no-Ponzi" condition, Equation (A15), holds: the initial η is below the dividend paying boundary, and so the solution is guaranteed to hit it. Subfigure (**b**) shows a scenario wherein the condition does not hold: The solution starts from a point to the left of $\eta^*_{\min}$, and grows fast enough to miss the lower branch of $q' = 0$ curve, entering a region where growth exceeds the discount rate. In (**a**), parameters are set to baseline, $\rho = 0.06$, $r = 0.05$, $\sigma_1 = 0.2$, $\sigma_2 = 0.0$, $a = 0.1$, $\bar{a} = 0.04$, $\delta = 0.02$, $\theta = 15$ and $\chi = 0.75$; in (**b**), parameters are the same, except that $\theta = 6.5$.

Some additional intuition into the factors that determine whether the "no-Ponzi" condition is satisfied or not can be obtained by re-expressing Equation (A15) as

$$\bar{g}^* < \left[\frac{1}{2}(\rho - g^*)^2 / (\rho - r) - \theta^{-1}\right]$$

where

$$\bar{g}^* = \left(r - \sqrt{r^2 - 2\theta^{-1}[\bar{a} - \delta - r]}\right) < r$$

is the rate of growth when capital stock is owned by external investors and

$$g^* = \left(\rho - \sqrt{\rho^2 - 2\theta^{-1}[\bar{a} - \delta - \rho]}\right) < \rho$$

the rate of growth of the capital stock in the situation where firms can costlessly issued equity ($\chi = 0$) but are unable to borrow (see Appendix A). This expression indicates that in order for the "no-Ponzi" to be satisfied requires that the difference between the discount rates of firms and outside investors $\rho - r$ is comparatively small, or the net productivity of capital either in the hands of firms or investors ($a - \delta$, $\bar{a} - \delta$) relative to the maximum values given by the constraints of Equations (4a) and (4b) is comparatively small, or the costs of adjustment of capital θ are comparatively high.

What about solution in the stochastic case if the "no-Ponzi" condition is not satisfied? A numerical solution can still be obtained and indicate that an optimal policy for choice of $\{i_t\}, \{\lambda_t\}, \{\epsilon_\tau\}$ satisfying the conditions of Proposition 3, i.e., with future dividend payments after any initial dividend payment to reduce η to the desired target level η^*, may still exist, provided that g^* is not too close to ρ.

Finally there are the propositions about the ergodic density.

Proof of Proposition 4. (Proposition of Section 3 can again be obtained by imposing appropriate parameter restrictions.) Denote the density function for the location of firms across the possible values of η at the moment t by $f(t,\eta)$, with the corresponding cumulative density:

$$F(t,\eta) = \int_{\bar{\eta}}^{\eta} f(t,\eta')\,\mathrm{d}\eta', \quad F(t,\eta^*) = 1$$

The evolution of $f(t,\eta)$ is then determined by the Kolmogorov forward or Fokker-Planck equation:

$$\frac{\partial f}{\partial t}(t,\eta) = -\frac{\partial}{\partial \eta}\left[\mu^{\eta}(\eta) f(t,\eta)\right] + \frac{1}{2}\frac{\partial^2}{\partial \eta^2}\left[\sigma^{\eta}(\eta)^2 f(t,\eta)\right], \tag{A16}$$

where η follows the equation of motion

$$\mathrm{d}\eta = \mu^{\eta}\,\mathrm{d}t + \sigma^{\eta}\,\mathrm{d}z.$$

with coefficients obtained simply by Itô differentiating $\eta = c/k$:

$$\mathrm{d}\eta = \frac{1}{k}\left[\mu^c - \eta\mu^k + \frac{\eta}{k}(\sigma^k)^2\right]\mathrm{d}t + \frac{\sigma^c}{k}\,\mathrm{d}z_1 - \eta\frac{\sigma^k}{k}\,\mathrm{d}z_2, \tag{A17}$$

where $\mu^{c,k}$ and $\sigma^{c,k}$ are respectively the drift and diffusion terms for c and k: $\mathrm{d}c = \mu^c \mathrm{d}t + \sigma^c \mathrm{d}z_1$, $\mathrm{d}k = \mu^k \mathrm{d}t + \sigma^k \mathrm{d}z_2$; i.e.,

$$\begin{gathered}\mu^c = [\psi a + (1-\psi)\bar{a} - \delta + r\eta - \theta^{-1}(q-1) - \frac{1}{2}\theta^{-1}(q-1)^2]k,\\ \mu^k = \theta^{-1}(q-1)k,\\ \sigma^c = \sigma_1\psi k, \quad \sigma^k = \sigma_2\psi k.\end{gathered}$$

The increments $\mathrm{d}z_1$, $\mathrm{d}z_2$ are independent and normally distributed, and so the noise sources in $\mathrm{d}\eta$ can be combined into a single term,

$$k^{-1}\sigma^c \mathrm{d}z_1 - k^{-1}\eta\sigma^k \mathrm{d}z_2 = k^{-1}\sqrt{(\sigma^c)^2 + \eta^2(\sigma^k)^2}\,\mathrm{d}z.$$

Substituting in the expressions for $\mu^{c,k}$ and $\sigma^{c,k}$ yields

$$\begin{aligned}\mathrm{d}\eta = \Big[&\bar{a} + (a-\bar{a})\psi - \delta + r\eta - (1+\eta)\theta^{-1}(q-1)\\ &-\frac{1}{2}\theta^{-1}(q-1)^2 + \eta\sigma_2^2\psi^2\Big]\mathrm{d}t + \sqrt{\sigma_1^2 + \eta^2\sigma_2^2}\psi\,\mathrm{d}z.\end{aligned} \tag{A18}$$

The ergodic probability density is then the stationary, $\partial f/\partial t = 0$, solution of Equation (A16), also denoted by $f(\eta)$. Integration of the Kolmogorov forward equation in η yields

$$d = \mu^{\eta}(\eta) f(\eta) - \frac{1}{2}\frac{\partial}{\partial \eta}\left[\sigma^{\eta}(\eta)^2 f(\eta)\right]. \tag{A19}$$

It is convenient to write this in terms of ϕ,

$$\phi = \frac{(\sigma^{\eta})^2}{2} f,$$

so this becomes:

$$d = \left[\frac{(\sigma^{\eta})^2}{2}\right]^{-1}\mu^{\eta}(\eta)\phi(\eta) - \frac{1}{2}\frac{\partial}{\partial \eta}\phi(\eta). \tag{A20}$$

and this yields Equation (15). □

Appendix C. Numerical Solution

Appendix C.1. Preliminary Considerations and Some Economic Intuition

The ordinary differential equation governing q (Equation (17) for $\eta \leq \tilde{\eta}$ and Equation (10) for $\eta \geq \tilde{\eta}$) can be solved by forward integration using standard methods starting from any given initial condition $q(\bar{\eta}) = \bar{q}$. The solution is completed by finding an intersection with $Q(q, \eta) = 0$ on which $q' = 0$, if one exists, or establishing that there is no such intersection (moreover, any solution with an intersection with $Q(q, \eta) = 0$). Any initial value $\bar{q} \geq 1 + \theta\rho$ can be ruled out, since it implies that $g > \rho$ for all η.

The inequality in Equation (A15), which is required if the special case of the model with no uncertainty ($\sigma = 0$) is to a be one with no "Ponzi-borrowing" and also ensures the existence of solution, leads to extremely straightforward numerical solution, since intersection with the lower branch of $Q(q, \eta) = 0$ is guaranteed.

If however this inequality is not satisfied then for some values of $\bar{q}$, a value of q^* whereat the ODE interacts with $Q(q, \eta) = 0$ may be located on the upper boundary (if the ODE "misses" the lower branch, in which case it may or may not hit the upper branch). This considerably complicates the search for numerical solution because it is no longer possible to restrict the initial values $\bar{q}$ to a range of values for which intersection with $Q(q, \eta) = 0$ is guaranteed.

Such solutions with upper branch intersections are of less economic interest than those where intersection is on the lower branch. It is possible that investment close to η^* is so high that the firm has negative cash flow. This can be seen by substituting Equation (A13), $\psi = 1$, $q = q^*$and $\eta = \eta^*$ into Equation (A18), yielding the following expression for cash flow on the dividend paying boundary:

$$\mu^{\eta} = \frac{(a-\delta)\rho - \rho r + \left[(r-(a-\delta))\theta^{-1} - \rho r\right](q^*-1)}{\rho - r} + \frac{\frac{1}{2}(2r+\rho)\theta^{-1} - \frac{1}{2}\theta^{-2}(q^*-1)}{\rho - r}(q^*-1)^2 + \eta^*\sigma_2^2 \quad \text{(A21)}$$

which, for sufficiently high q^*, is negative. The economic intuition in this case is similar to that applicable to the "Ponzi" solution of the model with no non-negativity constraint on dividends in Appendix A. The firm creates the most value not by dividend payments but from growing the capital stock at a rate close to and often above the shareholder's rate of discount, and this very high rate of investment can generate the negative expected cash flows.

The function $W(\eta^*)$ is obtained by substitution into Equation (A14) on the boundary η^*, using the boundary conditions $W' = 1$, to yield:

$$W^* = (q^* + \eta^*)W' = 1 + \eta^* + \rho\theta \pm \sqrt{2\theta\left\{(\eta^* - \eta^*_{\min})(\rho - r)\right\}} \quad \text{(A22)}$$

with the positive root applying on the upper branch of q^* and the negative root on the lower branch.

This in turn results in some useful insights into the solution. In the case of recapitalisation, Equation (7) can be written:

$$0 < W^* - (\eta^* - \bar{\eta} + \chi) = 1 + \rho\theta + \bar{\eta} - \chi \pm \sqrt{2\theta\left\{(\eta^* - \eta^*_{\min})(\rho - r)\right\}} < 1 + \rho\theta + \bar{\eta} \quad \text{(A23)}$$

where the first inequality is required by the maximisation in Equation (7) and the second because the presence of financing constraints must lower value $W(\bar{\eta})$ relative to the valuation for the case of no non-negativity constraint on dividend payments given by Equation (A3)). This establishes the following further proposition:

Proposition A2. *A solution with recapitalisation (for some sufficiently low value of χ) exists. Let $\chi = \chi_0$ be the critical value of χ at which the firm is indifferent between recapitalisation and liquidation. Then: (i) If $\eta^*_{min} \leq \bar{\eta}$ a solution exists with η^* on the lower branch of $Q(q,\eta) = 0$, $\bar{q}$ satisfies:*

$$-\bar{\eta} \leq \bar{q} < \bar{q}max = 1 + \theta\left(\rho - \sqrt{\rho^2 - 2\theta^{-1}\left\{a - \delta - \rho - \bar{\eta}\left(\rho - r\right)\right\}}\right) > \bar{q} \tag{A24}$$

and the maximum possible value of η^* satisfies:

$$\eta^*_{min} \leq \eta^* < \eta^*_{min} + \frac{\left(\left(\rho - r\right) + \sqrt{r^2 - 2\theta^{-1}[\bar{a} - \delta - r]} - \theta^{-1}\chi_0\right)^2}{2\left(\rho - r\right)} \tag{A25}$$

(ii) If instead $\eta^*_{min} > \bar{\eta}$ then $-\bar{\eta} \leq \bar{q} < 1 + \rho\theta$; a solution may or may not exist solution may be on the upper branch of $Q(q,\eta) = 0$, in which case η^* satisfies:

$$\eta^*_{min} < \eta^* \leq \eta^*_{min} + \frac{\chi^2}{2\theta\left(\rho - r\right)} \tag{A26}$$

Proof. The existence of a solution with recapitalisation is guaranteed because $1 + \rho\theta + \bar{\eta} = \left(\rho - r\right)\theta + \theta\sqrt{r^2 - 2\theta^{-1}[\bar{a} - \delta - r]} > 0$. As noted above, solutions for which $\bar{q} > 1 + \theta\rho$ can be ruled out, and hence all possible solutions, with recapitalisation or without, are with an intersection of the ODE for q on the lower branch of $Q\left(q,\eta\right) = 0$; and (the value that applies when cost of recapitalisation χ=0 and hence the maximum possible value of $\bar{q}$) is given by the intersection of this lower branch of Equation (A14) with $\eta = \bar{\eta}$. If solution is on the lower branch then the largest possible value of η^* and the smallest value of $\bar{q}$ ($\bar{q} = -\bar{\eta}$) arise when $\chi = \chi_0$ (the same solution also applies if χ is higher than this critical value and no-recapitalisation takes place). If $\chi = \chi_0$ then the first inequality in Equation (A23) binds and this implies the second inequality in Equation (A25). If instead solution is on the upper branch, then the largest possible value is when $\chi < \chi_0$, so that Equation (A23) binds and this implies the second inequality in Equation (A26). □

Proposition A2 helps guide the numerical solution. If $\eta^*_{\min} \leq \bar{\eta}$ then a solution with a bounded value function exists and an intersection is guaranteed on the lower branch of $Q(q,\eta) = 0$. A first calculation of the case with no recapitalisation determines χ_0 and this can then be used to limit the scope of iteration on $\bar{q}$ (using Equation (A24)) in the search for solution in the case of recapitalisation. If instead $\eta^*_{\min} > \bar{\eta}$ then a solution with a bounded value function and finite η^* may not exist. The existence of a solution for the case of no-recapitalisation can be established by computing the ODE Equation (10) upwards. If η exceeds the upper bound given by Equation (A26) then there is no intersection and no solution) is not satisfied, then while there is an intersection there is no finite solution to the value function. A solution with recapitalisation will exist for at least some values of χ if there is a solution for no-recapitalisation. The proposition then provides a slightly different limits on the scope of iteration on $\bar{q}$ and the same criteria can be applied to establish if there is an intersection with $Q(q,\eta) = 0$, and if so whether this represents a finite value for the value function.

Appendix C.2. Model without Option to Rent

For any given $\bar{q}$, the right-hand side boundary at $\eta = \eta^*$ wherein $q'(\eta^*) = 0$ is found by evaluating the function $q'(\eta)$ during the integration. After a single integration step is found to bracket a root of $q'(\eta)$, the critical value of η is pin-pointed using standard root finding methods, here the Brent's method.

The value function W can be solved from $W' = W/(\eta + q)$ parallel to integrating the equation for q. The boundary condition $W''(\eta^*) = 0$ will be satisfied since the q variable integration is stopped at $q' = 0$. In order to also satisfy the boundary condition $W'(\eta^*) = 1$, solve W for an arbitrary initial value at $\bar{\eta}$. Let the resulting solution be $\tilde{W}$. Since the ODE for W is linear and homogeneous, simply multiply $\tilde{W}$ ex post by $[\tilde{W}'(\eta^*)]^{-1}$ to get a solution for which $W'(\eta^*) = 1$.

In the case of liquidation, the lower boundary is $\bar{\eta} = -\bar{q}$, and consequently, the derivative of W, $W' = W/(\eta + q)$ cannot be evaluated. Appendix D shows that $W \propto \eta - \bar{\eta}$, and so $W'(\bar{\eta})$ is finite. Thus, if $\bar{\eta}$ is indeed liquidating, we simply set $\tilde{W}'(\bar{\eta}) = 1$ and $W(\bar{\eta}) = 0$.

Solving for the ergodic density with an absorbing boundary requires determining the constant of integration (the rate of flow across the boundary) d and this requires two boundary conditions. These conditions are that the absorbing boundary must have a zero density, i.e., $f(\bar{\eta}) = 0$, and the cumulative density must satisfy $F(\eta^*) = 1$.

The following method enforces these conditions. Solve two independent differential equations for two densities f_0 and f_1 satisfying:

$$f_0'(\eta) = \frac{2\mu(\eta)}{\sigma^2} f_0(\eta), \quad f_1'(\eta) = \frac{2\mu(\eta)}{\sigma^2} f_1(\eta) + 1. \tag{A27}$$

These are obtained by integration starting from arbitrary non-zero initial conditions. Let F_0 and F_1 be the resulting corresponding cumulative functions, with $F_0' = f_0$, $F_1' = f_1$ and $F_0(\bar{\eta}) = F_1(\bar{\eta}) = 0$. This determines values for $F(\eta^*)$ and $F_1(\eta^*)$.

Then find the ergodic density by choosing appropriate constants a_0 and a_1 in the following function f:

$$f(\eta) = a_0 f_0(\eta) + a_1 f_1(\eta), \tag{A28}$$

These coefficients a_0, a_1 are determined by the conditions $f(\bar{\eta}) = 0$ and $F(\eta^*) = 1$ as follows. Upon substituting the trial solution (A28), one obtains

$$a_0 f_0(\bar{\eta}) + a_1 f_1(\bar{\eta}) = 0, \quad a_0 F_0(\eta^*) + a_1 F_1(\eta^*) = 1.$$

yielding a pair of linear equations that can be solved for a_0 and a_1. To obtain d differentiate (A28) and use Equation (A27), to get:

$$f'(\eta) = \frac{2\mu}{\sigma^2} f(\eta) + a_1,$$

so $a_1 = -2d/\sigma^2$ (cf. Equation (A19)).

The possibility for recapitalisation is tested by finding roots of

$$G(\bar{q}) = W[\bar{q}, \eta^*(\bar{q})] - W(\bar{q}, \bar{\eta}) - [\eta^*(\bar{q}) - \bar{\eta}] - \chi,$$

making explicit the dependence of the location of the upper dividend paying boundary $\eta = \eta^*$ and the function $W(\eta)$ on the value of q on the lower boundary $q(\bar{\eta}) = \bar{q}$. Clearly $G = 0$ is equivalent to achieving Equation (7). Functions η^*, q and W are all obtained using the same method outlined above (i.e., jointly computing the two odes for q and W using $\bar{q}$ and an arbitrary value of W on $\bar{\eta}$, locating η^* from $q' = 0$, and rescaling W to enforce $W' = 1$).

The task then is to iterate on the starting value $\bar{q}$ to find the root of $G(\bar{q})$. First a coarse root bracketing is attempted by evaluating G at $\bar{q}_i = -\bar{\eta} + (q_1 + \bar{\eta})i/n_q$, where $i = 0 \ldots n_q$ and n_q an integer (using $n_q = 10$), and q_1 is q as given by Equation (A14) if that value is real, or $1 + \theta\rho$ if it is not. If sign of G changes across a bracketing interval $(\bar{q}_i, \bar{q}_{i+1})$, the root is pin-pointed using standard root finding algorithms. This locates a recapitalisation solution. If no roots are found, or a root is found with $\bar{q} < -\bar{\eta}$ or $q^* > 1 + \theta\rho$ then the solution is identified as liquidation with $\bar{q} = -\bar{\eta}$.

Appendix C.3. Model with Option to Rent

The algorithm outline is same as in the model without the rental option. However, the solution near the lower boundary is more involved when recapitalisation is not undertaken, and so $\psi(\bar{\eta}) = 0$.

The differential equations for q can again be solved by simple forward integration starting from $q(\bar{\eta}) = \bar{q}$. If recapitalisation is available ($\bar{q} > -\bar{\eta}$), no singularities are present, and the equation for q, Equation (17), can be integrated directly to obtain $q(\eta)$, η^* and now also $\tilde{\eta}$. The point $\tilde{\eta}$ is found in the

same way as η^*, *i.e.*, by monitoring the function $\psi - 1$ as integration advances and polishing the root after a coarse approximation is found. Initial $\bar{q}$ is found the same way as for the model without renting (but with q, W computed slightly differently as described below).

If $\bar{q} = -\bar{\eta}$, then $\psi = 0$ and singularities appear. As is shown in Appendix D, the derivative $q'(\bar{\eta})$ is finite. In order to evaluate it numerically, use Equation (A32) since Equation (17) is indeterminate at $\bar{\eta}$ (in practice, numerical round-off would cause significant error in $\bar{q}$). Otherwise the solution of q proceeds the same way as with a recapitalising lower boundary.

Using ϕ, and expanding the resulting equation in the renting ($0 < \psi < 1$) and not renting regimes ($\psi = 1$), yields

$$\phi' = \begin{cases} \dfrac{\bar{a} + (a - \bar{a})\psi - \delta + r\eta + \sigma_2^2\psi^2\eta - \theta^{-1}(q-1)[\eta + \frac{1}{2}(q+1)]}{\frac{1}{2}(\sigma_1^2 + \eta^2\sigma_2^2)\psi^2}\phi - d, \\ \qquad \text{when } \psi \in (0,1), \\ \dfrac{a - \delta + r\eta + \sigma_2^2\eta - \theta^{-1}(q-1)[\eta + \frac{1}{2}(q+1)]}{\frac{1}{2}(\sigma_1^2 + \eta^2\sigma_2^2)}\phi - d, \\ \qquad \text{when } \psi = 1. \end{cases} \tag{A29}$$

When there are no recapitalisation, equations for f' and W', unlike that for q', do not tend to finite values at $\bar{\eta}$, since $\psi(\bar{\eta}) = 0$ if $q(\bar{\eta}) = -\bar{\eta}$, $q'(\bar{\eta}) > 0$. Due to this divergence, the point $\bar{\eta}$ cannot be reached by directly integrating the model equations, which in principle could be done backwards from, say, $\tilde{\eta}$ down to $\bar{\eta} + \epsilon$, $0 < \epsilon \ll 1$. Cutting the integration short in this way would lead to severe underestimation of the probability mass near $\bar{\eta}$ if f diverges fast enough at this edge.

This issue is resolved using the analytically obtained power-law solutions, $f_\mathrm{a} \propto (\eta - \bar{\eta})^{\alpha-2}$ (Equation (21)) and $W_\mathrm{a} \propto (\eta - \bar{\eta})^{\beta}$ (Equation (18)), from $\bar{\eta}$ up to a cross-over value $\eta_\times$. Numerical solutions are matched to the analytic ones so that the resulting functions are continuous. The cross-over point can determined by requiring that

$$\left| \frac{f_\mathrm{a}'(\eta_\times)}{f_\mathrm{a}(\eta_\times)} \right| = \varepsilon^{-1}, \tag{A30}$$

where $0 < \epsilon \ll 1$, implying that the divergent terms dominate the expression for the derivative of f. However, since W' also tends to infinity, the same condition applies to W_a as well. This gives two different cross-over values; the smallest is chosen:

$$\eta_\times = \varepsilon \min(|\alpha|, \beta) + \bar{\eta}, \tag{A31}$$

where α is given by Equation (A34) and $\beta = 1/(1 + q'(\bar{\eta}))$, with $q'(\bar{\eta})$ from Equation (A32). The results reported here use the value $\varepsilon = 1.0 \times 10^{-3}$ and the analytic solution for f to obtain the cumulative density F below $\eta_\times$.

If the lower boundary is at $\bar{q} = -\bar{\eta}$, then directly integrate Equation (A29) with $d = 0$ from $\eta_\times$ to η^*. The obtained solution can then be multiplied by a constant to make the cumulative distribution satisfy $F(\eta^*) = 1$. If $\bar{\eta}$ is absorbing (recapitalisation), use the same trick as in the model without rent: solve for ϕ_0 and ϕ_1 satisfy Equation (A29) with $d = 0$ and $d = 1$, respectively. The final ϕ is then constructed as a superposition of these two, $\phi = a_0\phi_0 + a_1\phi_1$. Coefficients a_0 and a_1 are determined from

$$\phi(\bar{\eta}) = 0, \quad \int_{\bar{\eta}}^{\eta^*} \frac{2}{(\sigma^\eta(\eta))^2}\phi(\eta)\,\mathrm{d}\eta = 1.$$

When needed, the same analytic solution, Equation (21), can be used for both ϕ_0 and ϕ_1 ($\phi_{0,1} \propto (\eta - \bar{\eta})^\alpha/(\sigma^\eta)^2$), since d term is negligible near $\bar{\eta}$.

Note that reverting to the analytic solution for f is equivalent to using a truncated integration range with an additional correction term coming from the analytical solution near $\bar{\eta}$. Numerical simulations confirm that this approach is sound: (i) the analytical and numerical solutions are in very good agreement across a wide range of η, (ii) the obtained solutions are independent of ϵ, provided it is small enough while keeping the numerical solution from reaching the singularity, and (iii) qualitative features of the solution do not change if the analytical correction is omitted.

Appendix D. Behaviour of Solutions Near Boundaries

This Appendix provides the derivation of the asymptotic approximations summarised in Proposition 7.

Appendix D.1. Model without Option to Rent

While no singularities emerge in the model with no option to rent, it is still useful to begin with this simple case. The evolution of the value function W is given by $W'/W = 1/(q+\eta)$, which in the case of liquidation tends to infinity as the point of maximum borrowing where $q(\bar{\eta}) = -\bar{\eta}$ is approached. This means there is a potential singularity in W at $\bar{\eta}$. It can be shows that in the model without the option to rent this does not occur and W is linear close to $\eta = \bar{\eta}$.

Suppose now that q is of the form $q(\eta) = -\bar{\eta} + q'(\bar{\eta})(\eta - \bar{\eta}) + \mathcal{O}((\eta - \bar{\eta})^2)$. Near the boundary, W follows

$$W' = \frac{1}{1 + q'(\bar{\eta})} \frac{W}{\eta - \bar{\eta}} + \mathcal{O}(\eta - \bar{\eta}).$$

The solution is then given by Equation (18) in the main text. In the case of the model without renting, it is clear from Equation (10) that $q'(\bar{\eta}) = 0$ and so W is linear near $\bar{\eta}$.

Appendix D.2. Model with Option to Rent

Turning to the model with option to rent, again a singularity can occur only on the lower boundary and only when there is no recapitalisation, i.e., when $q(\bar{\eta}) = -\bar{\eta}$ and $\psi(\bar{\eta}) = 0$.

Note now that in the equation for q', Equation (17), both the numerator and the denominator vanish. Applying the l'Hopital's rule, the derivative can be solved as:

$$q'(\bar{\eta}) = \frac{-(\rho - r - \gamma) \pm \sqrt{(\rho - r - \gamma)^2 + 4\gamma\theta^{-1}[1 + \theta\rho - \bar{q}]}}{2\theta^{-1}[1 + \theta\rho - \bar{q}]}, \tag{A32}$$

where $\gamma = (a - \bar{a})^2/2(\sigma_1^2 + \bar{\eta}^2\sigma_2^2)$. Above, only the plus sign applies. This can be seen by recalling that $\bar{q} < q_{\max} = 1 + \rho\theta$ must apply (see above for the reasoning), in which case only the plus sign gives a positive q'. Thus, the solution near $\bar{\eta}$ is given by Equation (19) in the main text.

The power-law form of W given in Equation (18) holds here as well. Since now $q'(\bar{\eta}) > 0$, the exponent $\beta = 1/(1 + q'(\bar{\eta}))$ is always less than one, in contrast to the model without option to rent, implying that $\lim_{\eta \downarrow \bar{\eta}} W' = \lim_{\eta \downarrow \bar{\eta}} (-W''/W') = +\infty$.

To find the behaviour of the ergodic density near $\bar{\eta}, \bar{q}$, requires ψ. This time $\eta - \bar{\eta}$ is not negligible compared to $q - \bar{q}$. A straight-forward calculation gives:

$$\psi = \psi'(\bar{\eta})(\eta - \bar{\eta}) \tag{A33}$$

where

$$\psi'(\bar{\eta}) = \frac{2}{a - \bar{a}} \left\{ \rho - r + \theta^{-1} [1 + \theta\rho - \bar{q}] \, q'(\bar{\eta}) \right\}.$$

Next, the $\eta \to \bar{\eta}$ limiting forms of q and ψ are substituted into Equation (A20), and only terms up to $\mathcal{O}(\eta - \bar{\eta})$ are kept. Notice that the numerator vanishes in the leading order, and hence $\phi' \propto (\eta - \bar{\eta})^{-1}$ and not $\propto (\eta - \bar{\eta})^{-2}$:

$$\phi' = \alpha \frac{\phi}{\eta - \bar{\eta}},$$

where

$$\alpha = \frac{(a-\bar{a})\psi'(\bar{\eta}) + r - \frac{1}{2}\theta^{-1}q'(\bar{\eta})(\bar{\eta}+1) + \theta^{-1}(\bar{\eta}+1)(1+q'(\bar{\eta})/2)}{\frac{1}{2}(\sigma_1^2 + \bar{\eta}^2\sigma_2^2)\psi'(\bar{\eta})^2}. \tag{A34}$$

This gives the power-law solution Equation (20) in the main text. Finally using Equation (A33) yields Equation (21) of the main text.

References

1. Milne, A.; Robertson, D. Firm Behaviour Under the Threat of Liquidation. *J. Econ. Dyn. Control* **1996**, *20*, 1427–1449. [CrossRef]
2. Brunnermeier, M.K.; Sannikov, Y. A Macroeconomic Model with a Financial Sector. *Am. Econ. Rev.* **2014**, *104*, 379–421. [CrossRef]
3. Whittle, P. *Optimization over Time: Dynamic Programming and Stochastic Control*; Wiley-Blackwell: New York, NY, USA, 1982; Volume 1.
4. Carroll, C.D. A theory of the consumption function, with and without liquidity constraints. *J. Econ. Perspect.* **2001**, *15*, 23–45. [CrossRef]
5. Arrow, K.J.; Harris, T.; Marschak, J. Optimal inventory policy. *Econometrica* **1951**, *19*, 250–272. [CrossRef]
6. Scarf, H. *The Optimality of (sS) Policies in the Dynamic Inventory Problem. Mathematical Methods in the Social Science*; Arrow, K.J., Karlin, S., Suppes, P., Eds.; Stanford University Press: Stanford, CA, USA, 1960; Chapter 22.
7. Jorgenson, D.W. Capital Theory and Investment Behavior. *Am. Econ. Rev.* **1963**, *53*, 247–259. [CrossRef]
8. Lucas, R.E., Jr.; Prescott, E.C. Investment under Uncertainty; Princeton University Press: Princeton, NJ, USA, 1971; Volume 39, pp. 659–681.
9. Miller, M.H.; Orr, D. A Model of the Demand for Money by Firms. *Q. J. Econ.* **1966**, *80*, 413–435. [CrossRef]
10. Constantinides, G.M. Stochastic Cash Management with Fixed and Proportional Transaction Costs. *Manag. Sci.* **1976**, *22*, 1320–1331. [CrossRef]
11. Frenkel, J.A.; Jovanovic, B. On Transactions and Precautionary Demand for Money. *Q. J. Econ.* **1980**, *95*, 25–43. [CrossRef]
12. Merton, R.C. On the Pricing of Corporate Debt: The Risk Structure of Interest Rates. *J. Financ.* **1974**, *29*, 449–470.
13. Anderson, R.W.; Sundaresan, S. Design and valuation of debt contracts. *Rev. Financ. Stud.* **1996**, *9*, 37–68. [CrossRef]
14. Mella-Barral, P.; Perraudin, W. Strategic debt service. *J. Financ.* **1997**, *52*, 531–556. [CrossRef]
15. Leland, H.E. Agency Costs, Risk Management, and Capital Structure. *J. Financ.* **1998**, *53*, 1213–1243. [CrossRef]
16. Modigliani, F.; Miller, M.H. The Cost of Capital, Corporation Finance and the Theory of Investment. *Am. Econ. Rev.* **1958**, *48*, 261–297. [CrossRef]
17. Tirole, J. *The Theory of Corporate Finance*; Princeton University Press: Princeton, NJ, USA, 2006.
18. Myers, S.C. The Capital Structure Puzzle. *J. Financ.* **1984**, *39*, 574–592. [CrossRef]
19. Myers, S.C.; Majluf, N.S. Corporate financing and investment decisions when firms have information that investors do not have. *J. Financ. Econ.* **1984**, *13*, 187–221. [CrossRef]
20. Froot, K.; Scharfstein, D.; Stein, J.C. Risk Management: Coordinating Corporate Investment and Financing Policies. *J. Financ.* **1993**, *48*, 1629–1658. [CrossRef]
21. Gertler, M. Financial Capacity and Output Fluctuations in an Economy with Multi-Period Financial Relationships. *Rev. Econ. Stud.* **1992**, *59*, 455. [CrossRef]
22. Townsend, R.M. Optimal contracts and competitive markets with costly state verification. *J. Econ. Theory* **1979**, *21*, 265 – 293. [CrossRef]
23. Sannikov, Y. A Continuous-Time Version of the Principal—Agent Problem. *Rev. Econ. Stud.* **2008**, *75*, 957–984. [CrossRef]
24. De Marzo, P.M.; Sannikov, Y. Optimal Security Design and Dynamic Capital Structure in a Continuous-Time Agency Model. *J. Financ.* **2006**, *61*, 2681–2724. [CrossRef]

25. Mauer, D.C.; Triantis, A.J. Interactions of corporate financing and investment decisions: A dynamic framework. *J. Financ.* **1994**, *49*, 1253–1277. [CrossRef]
26. Radner, R.; Shepp, L. Risk vs. profit potential: A model for corporate strategy. *J. Econ. Dyn. Control* **1996**, *20*, 1373–1393. [CrossRef]
27. Jeanblanc-Picqué, M.; Shiryaev, A.N. Optimization of the flow of dividends. *Russ. Math. Surv.* **1995**, *50*, 257–277. [CrossRef]
28. Asmussen, S.; Taksar, M. Controlled diffusion models for optimal dividend pay-out. *Insur. Math. Econ.* **1997**, *20*, 1–15. [CrossRef]
29. Taksar, M.I.; Zhou, X.Y. Optimal risk and dividend control for a company with a debt liability. *Insur. Math. Econ.* **1998**, *22*, 105–122. [CrossRef]
30. Jgaard, B.H.; Taksar, M. Controlling risk exposure and dividends payout schemes: Insurance company example. *Math. Financ.* **1999**, *9*, 153–182. [CrossRef]
31. Asmussen, S.; Højgaard, B.; Taksar, M. Optimal risk control and dividend distribution policies. Example of excess-of loss reinsurance for an insurance corporation. *Financ. Stochastics* **2000**, *4*, 299–324. [CrossRef]
32. Milne, A.; Whalley, A.E. Bank Capital and Risk Taking; Bank of England Working Paper: London, UK, 1999.
33. Milne, A.; Whalley, A.E. Bank Capital Regulation and Incentives for Risk Taking; Cass Business School Research Paper: London, UK, 2002.
34. Milne, A. The Inventory Perspective on Bank Capital; Cass Business School Research Paper: London, UK, 2004.
35. Peura, S.; Keppo, J. Optimal Bank Capital With Costly Recapitalization. *J. Bus.* **2006**, *79*, 2163–2201. [CrossRef]
36. Krugman, P.R. Target Zones and Exchange Rate Dynamics. *Q. J. Econ.* **1991**, *106*, 669–682. [CrossRef]
37. Mundaca, G.; Øksendal, B. Optimal stochastic intervention control with application to the exchange rate. *J. Math. Econ.* **1998**, *29*, 225–243. [CrossRef]
38. Korn, R. Some applications of impulse control in mathematical finance. *Math. Methods Oper. Res. (ZOR)* **1999**, *50*, 493–518. [CrossRef]
39. Rochet, J.C.; Villeneuve, S. Liquidity management and corporate demand for hedging and insurance. *J. Financ. Intermediation* **2011**, *20*, 303–323. [CrossRef]
40. Bolton, P.; Chen, H.; Wang, N. A Unified Theory of Tobin's q, Corporate Investment, Financing, and Risk Management. *J. Financ.* **2011**, *66*, 1545–1578. [CrossRef]
41. Bolton, P.; Chen, H.; Wang, N. Market timing, investment, and risk management. *J. Financ. Econ.* **2013**, *109*, 40–62. [CrossRef]
42. Rampini, A.A.; Viswanathan, S. Collateral and capital structure. *J. Financ. Econ.* **2013**, *109*, 466–492. [CrossRef]
43. Palazzo, B. Cash holdings, risk, and expected returns. *J. Financ. Econ.* **2012**, *104*, 162–185. [CrossRef]
44. Anderson, R.W.; Carverhill, A. Corporate Liquidity and Capital Structure. *Rev. Financ. Stud.* **2011**, *25*, 797–837. [CrossRef]
45. Gamba, A.; Triantis, A. The Value of Financial Flexibility. *J. Financ.* **2008**, *63*, 2263–2296. [CrossRef]
46. Gamba, A.; Triantis, A.J. Corporate Risk Management: Integrating Liquidity, Hedging, and Operating Policies. *Manag. Sci.* **2014**, *60*, 246–264. [CrossRef]
47. Greenwald, B.; Stiglitz, J.E.; Weiss, A. Informational imperfections in the capital market and macroeconomic fluctuations. *Am. Econ. Rev.* **1984**, pp. 194–199.
48. Kiyotaki, N.; Moore, J. Credit Cycles. *J. Political Econ.* **1997**, *105*, 211–248. [CrossRef]
49. Bernanke, B.S.; Gertler, M. Agency Costs, Net Worth, and Business Fluctuations. *Am. Econ. Rev.* **1989**, *79*, 14–31.
50. Bernanke, B.; Gertler, M.; Gilchrist, S. The Financial Accelerator in a Quantitative Business Cycle Framework. In *Handbook of Macroeconomics, Volume 1C*; Taylor, J.B.; Woodford, M., Eds.; Elsevier Science: North-Holland, The Netherlands, 1999; pp. 1341–1393.
51. Danielsson, J.; Shin, H.S.; Zigrand, J.P. The impact of risk regulation on price dynamics. *J. Bank. Financ.* **2004**, *28*, 1069–1087. [CrossRef]
52. Brunnermeier, M.K.; Pedersen, L.H. Market Liquidity and Funding Liquidity. *Rev. Financ. Stud.* **2008**, *22*, 2201–2238. [CrossRef]
53. Adrian, T.; Boyarchenko, N. Intermediary Leverage Cycles and Financial Stability; Federal Reserve Bank of New York Staff Reports No 567; Federal Reserve Bank of New York: New York, NY, USA, 2013.
54. Isohätälä, J.; Klimenko, N.; Milne, A. *Post-Crisis Macrofinancial Modeling: Continuous Time Approaches*; Palgrave Macmillan UK: London, UK, 2016. [CrossRef]

55. He, Z.; Krishnamurthy, A. A Model of Capital and Crises. *Rev. Econ. Stud.* **2012**, *79*, 735–777. [CrossRef]
56. He, Z.; Krishnamurthy, A. Intermediary Asset Pricing. *Am. Econ. Rev.* **2013**, *103*, 732–770. [CrossRef]

mathematics

MDPI

Article

Financial Distress Prediction and Feature Selection in Multiple Periods by Lassoing Unconstrained Distributed Lag Non-linear Models

Dawen Yan [1], Guotai Chi [2] and Kin Keung Lai [3,*]

[1] School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China; dawenyan@dlut.edu.cn
[2] School of Economics and Management, Dalian University of Technology, Dalian 116024, China; Chigt@dlut.edu.cn
[3] College of Economics, Shenzhen University, Shenzhen 518060, China
* Correspondence: mskklai@outlook.com

Received: 16 July 2020; Accepted: 29 July 2020; Published: 3 August 2020

Abstract: In this paper, we propose a new framework of a financial early warning system through combining the unconstrained distributed lag model (DLM) and widely used financial distress prediction models such as the logistic model and the support vector machine (SVM) for the purpose of improving the performance of an early warning system for listed companies in China. We introduce simultaneously the 3~5-period-lagged financial ratios and macroeconomic factors in the consecutive time windows $t-3$, $t-4$ and $t-5$ to the prediction models; thus, the influence of the early continued changes within and outside the company on its financial condition is detected. Further, by introducing lasso penalty into the logistic-distributed lag and SVM-distributed lag frameworks, we implement feature selection and exclude the potentially redundant factors, considering that an original long list of accounting ratios is used in the financial distress prediction context. We conduct a series of comparison analyses to test the predicting performance of the models proposed by this study. The results show that our models outperform logistic, SVM, decision tree and neural network (NN) models in a single time window, which implies that the models incorporating indicator data in multiple time windows convey more information in terms of financial distress prediction when compared with the existing singe time window models.

Keywords: financial distress prediction; unconstrained distributed lag model; multiple periods; Chinese listed companies

1. Introduction

Over the last four decades, models and methods for the prediction of corporate financial distress have attracted considerable interest among academics as well as practitioners. Financial distress prediction models can be used for many purposes including: monitoring of the solvency of regulated companies, assessment of loan default risk and the pricing of bonds, credit derivatives, and other securities exposed to credit risk (see [1–4]).

Different countries have different accounting procedures and rules; thus, the definition of financial distress put forward by different scholars is not always the same (see [4–7]). Bankruptcy is one of the most commonly used outcomes of financial distress of a company [5]. The nature of a bankrupt firm is that the owners can abandon the firm and transfer ownership to the debt holders, and bankruptcy occurs whenever the realized cash flow is less than the debt obligations [8]. It is generally agreed on that financial failure leads to substantive weakening of profitability of the company over time, but it is also feasible that a financially distressed firm may not change its formal status to bankrupt [9].

Therefore, in this paper, as done by [3] and [4], we identify a financially distressed company as the one at risk of failing, but which remains a viable entity at the present time. More specifically, "special treatment" (ST) is used to measure the financial distress status of a listed company. Further, in this paper, we provide a group of financial distress prediction models that incorporate the panel data of financial and macroeconomic indicators to implement early financial distress prediction.

In the existing studies, many classical statistical methods, such as discriminant analysis [1]; logistic regression and multinomial logit models (see [2–4,10]); heuristic algorithm methods such as the genetic algorithm and particle swarm optimization [11]; currently popular machine learning techniques, such as the support vector machine, decision tree and neural networks (see [7,12–16]), have been widely applied to develop financial distress prediction models. The relevant studies also realize that a set of indicators can be used to predict the financial distress, including financial indicators (e.g., see [7,17–20]) and macroeconomic indicators (e.g., see [4]). For example, accounting models such as Altman's 5-variable (Z-score) model (see [1]), 7-variable model (see [21]) etc. have gained popularity in both academic and industrial fields due to their discriminating ability and predictive power.

A few extant studies have predicted financial distress by using the accounting ratios from one or more years prior to its observation, given that early change in financial indicators may provide the warning sign of deterioration of financial conditions. The authors of [1] provide the evidence that bankruptcy can be predicted two years prior to the event, while those of [4] construct respectively two groups of financial distress prediction models for two periods: one year and two years before the observation of the financial distress event. They found that the models in the both time-windows have good predictive performance. Other similar examples can be found in [7,20,22].

In the relevant literatures, financial indicators in the different time windows have proven their contribution to the performance of the distress prediction models, in spite of the fact that the degree of their impact tends to change over time. However, the procedures are performed for all the different lag periods separately, i.e., only using information of one specific year prior to the date of the distress event. To the best of the authors' knowledge, no previous study, which solves the listed companies' financial distress prediction problem, takes into account the impacts of relevant indicators in the different and consecutive lag periods.

In this study, we take the form of the distributed lag model (DLM), in addition to classical classification techniques, into the financial distress prediction problem and propose a group of distress prediction models including the logistic-distributed lag model and the SVM-distributed lag model that can be treated as generalized distributed lag model. We construct the linkage between multiple lagged values of financial ratios and macroeconomic indicators and current financial status in order to capture the dynamic natures of the relevant data. Further, we propose to implement the penalized logistic-distributed lag financial distress model with the least absolute shrinkage and selection operator (lasso) penalty via the algorithm framework of the alternating direction method of multipliers (ADMM) that yields the global optimum for convex and the non-smooth optimization problem. Lasso-type penalty was applied for three purposes: to avoid the collinearity problem in applying the distributed lag models directly, to simultaneously select significant variables and estimate parameters, and to address the problem of over-fitting. We conduct a series of empirical studies to illustrate the application of our distributed lag financial distress models, including a comparison of predictive performances of the two distributed lag financial distress models proposed in this paper as well as the comparisons of predictive performances of our models with a group of widely used classification models in the different time windows. The results show that all distributed lag financial distress models aggregating the data in three consecutive time windows outperform the ones that incorporate the data in any single-period: 3 years, 4 years or 5 years before the observation year of financial distress, when the financial and macroeconomic factors are included. This paper may provide a means of improving the predictive performance of financial distress model by incorporating data of financial and macroeconomic indicators in consecutive and multiple periods before the observation of financial distress.

The rest of this paper is organized as follows. Section 2 briefly reviews the previous financial distress prediction literature and distributed lag modeling. Section 3 constructs a group of generalized distributed lag models composed of lagged explanatory variables and l_1 regularization, the logistic regression-distributed lag model and the lasso–SVM model with lags, and proposes the ADMM algorithm framework for coefficient estimations and variable selection at same time. Section 4 provides a description of the data. Section 5 presents the empirical results and compares the predictive performance of our models reflecting the extended lag effects of used indicators with the existing financial distress prediction models. Section 6 concludes the paper.

2. Background

2.1. Literature on Financial Distress Prediction

2.1.1. Financial Factors and Variable Selection

There is a large amount of theoretical and empirical microeconomic literature pointing to the importance of financial indicators on financial distress forecasting. The authors of [1] selected five financial indicators of strong predictive ability from the initial set of 22 financial indicators using stepwise discriminant analysis: earnings before interest and taxes/total assets, retained earnings/total assets, working capital/total assets, market value equity/book value of total debt and sales/total assets, which measures the productivity of assets of a firm. The other studies that concern the similar accounting ratios in financial distress prediction can also be found in [23,24]. Furthermore, the current-liabilities-to-current-assets ratio used to measure liquidity (see [22,25]) and the total-liabilities-to-total-assets ratio used to measure the degree of indebtedness of a firm (see [22,25,26]) and cash flow (see [18]) have been incorporated into the distress prediction models because of their predictive performance. The authors of [20] considered nine financial indicators and use them to predict the regulatory financial distress in Brazilian electricity distributors. The authors of [7] introduced 31 financial indicators and found that the most important financial indicators may be related to net profit, earnings before income tax, cash flow and net assets. Along this line, with machine learning methods developing, more diversified financial indicators have been considered in very recent studies of Hosaka [27], Korol et al. [28], and Gregova et al. [29].

Another very important recent research line in the area of financial distress prediction is the suitability problem of these financial ratios used as explanatory variables. For example, Kovacova et al. [30] discussed the dependence between explanatory variables and the Visegrad Group (V4) and found that enterprises of each country in V4 prefer different explanatory variables. Kliestik [31] chose eleven explanatory financial variables and proposed a bankruptcy prediction model based on local law in Slovakia and business aspects.

In this paper, we construct an original financial dataset including 43 financial ratios. The ratios are selected on the basis of their popularity in the previous literature (see [1,7,27]) and potential relevancy to this study. Then, like many relevant studies [32–34], we use the lasso method and conduct feature selection in order to exclude the potentially redundant factors.

2.1.2. Macroeconomic Conditions

Macroeconomic conditions are relevant for the business environment in which firms are operating; thus, the deterioration of macroeconomic conditions may induce the occurrence of the financial distress. Macroeconomic variables have been found to impact corporate default and bankruptcy risk significantly, and good examples can be found in [35–39]. In the aspect of financial distress of listed companies, [4] consider the macroeconomic indicators of the retail price index and the short-term bill rate adjusted for inflation in addition to the accounting variables. The results in their studies suggest that all macroeconomic indicators have significant impact on the likelihood of a firm's financial distress. In this paper, we control for macroeconomic conditions, GDP growth, inflation, unemployment rate in the

urban area and consumption level growth over the sample period. GDP growth is widely understood to be an important variable to measure economic strength and prosperity, and the increase in GDP growth may decrease the likelihood of distress. The authors of [22,40] have pointed out that the decline in GDP is significantly linked to the tightening of a firm's financial conditions, especially during the financial crisis period. The unemployment rate and inflation are two broadly used measures of overall health of the economy. High unemployment and high inflation that reflect a weaker economy may increase the likelihood of financial distress. Their impacts on financial distress have been examined in [16,37].

Different from the existing relevant studies that consider the lagged effect of macroeconomic variables only in a fixed window, such as 3 years prior to financial distress [16], this paper imposes a distributed lag structure of macroeconomic data, in addition to financial ratio data, and considers the lagged effects of the factors in the multi-periods. Particular attention is devoted to the lag structure and whether the predicting performance can be improved after introducing a series of lagged macroeconomic variables. The theoretical and empirical investigations in this study may complement the literature on financial distress prediction concerned with applying dynamic macro and financial data.

2.1.3. Related Literature on Chinese Listed Companies

The Chinese stock market has grown to over 55 trillion in market capitalization as of February 2020, and the number of listed firms has surpassed 3200, becoming the world's second largest market. Its ongoing development and the parallel evolution of regulations have made China's stock market an important subject for mainstream research in financial economics [41]. In April 1998, the Shanghai and Shenzhen stock exchanges implemented a special treatment (ST) system for stock transactions of listed companies with abnormal financial conditions or other abnormal conditions. According to the regulations, there are three main reasons for designation of a ***ST*** company: (1) a listed company has negative net profits for two consecutive years; (2) the shareholders' equity of the company is lower than the registered capital; (3) a firm's operations have stopped and there is no hope of restoring operations in the next 3 months due to natural disasters, serious accidents, or lawsuits and arbitration [7]. ***ST*** status is then usually applied as a proxy of financial distress (e.g., [7,16,42–45]).

Researchers regard the topic of financial distress prediction of Chinese listed companies as data-mining tasks, and use data mining, machine learning or statistical methods to construct a series of prediction models incorporating financial data (see [7,42,43]) or financial plus macroeconomic data [16] in one-time-period, but not in multiple periods of time. In the very recent study of [45], the authors proposed a financial distress forecast model combined with multi-period forecast results. First, with the commonly used classifiers such as the support vector machine (SVM), decision tree (DT) etc., the two to five-year-ahead financial distress forecast models are established one by one and denoted as T-2 to T-5 models, respectively. Then, through combining the forecast results of these one-time-period models, the multi-period forecast results, as a weighted average over a fixed window, with exponentially declining weights, are provided. This is obviously different from our model, as we introduce the multi-period lagged explanatory variables and detect simultaneously the effects of the variables in different prior periods on financial distress in the process of modeling.

2.2. Distributed Lag Models

Sometimes the effect of an explanatory variable on a specific outcome, such as the changes in mortality risk, is not limited to the period when it is observed, but it is delayed in time [46,47]. This introduces the problem of modeling the relationship between a future outcome and a sequence of lags of explanatory variables, specifying the distribution of the effects at different times before the outcome. Among the various methods that have been proposed to deal with delayed effects, as a major econometric approach, distributed lag models (DLMs) have been used to diverse research fields including assessing the distributed lag effects of air pollutants on children's health [48], hospital admission scheduling [49], and economical and financial time series analysis [50,51].

DLMs model the response Y_t, observed at time t in terms of past values of the independent variable X, and have a general representation given by

$$g(Y_t) = \alpha_0 + \sum_{l=0}^{L} s_l(x_{t-l}; \alpha_l) + \varepsilon_t, t = L, +1, \ldots, T \quad (1a)$$

where Y_t is the response at time t and g is a monotonic link function; the functions s_l denote a smoothed relationship between the explanatory vector x_{t-l} and the parameter vector α_l; α_0 and ε_t denote the intercept term and error term with a zero mean and a constant variance σ^2; l, L are the lag number and the maximum lag. The form that link function g takes presents a distributed lag linear model or a non-linear model. For example, linear g plus the continuous variable Y_t present distributed lag linear models, while the logit function g plus the binary variable Y_t present a distributed lag non-linear model. In model (1a), the parametric function s_l is applied to model the shape of the lag structure, usually polynomials (see [49,52]) or less often regression splines [53] or more complicated smoothing techniques of penalized splines within generalized additive models [47,48]. In fact, the introduction of s_l is originally used to solve the problem that these successive past observations may regard as collinear. If L, the number of relevant values of X, is small, as may well be the case for some problems if annual data are involved, then model (1a) degrades to an unconstrained distributed lag model given by the following general representation

$$g(Y_t) = \alpha_0 + \sum_{l=0}^{L} {\alpha_l}^T x_{t-l} + \varepsilon_t, \; t = L, +1, \ldots, T \quad (1b)$$

In (1b), the definitions of variables are the same as those in (1). Correspondingly, the coefficients in model (1b) can be estimated directly by pooled least squares for the linear case or pooled maximum likelihood for the non-linear case, e.g., logit link function g under the assumption that x_{t-l} is strictly exogenous [54].

In this paper, logistic regression with an unconstrained distributed lag structure is used to identify the relationship between financial and macroeconomic indicators and future outcome of financial distress. The logistic regression may be the most frequently used technique in the financial distress prediction field ([4,20]) because logistic regression relies on fewer assumptions due to the absence of the need for multivariate normality and homogeneity in the variance–covariance matrices of the explanatory variables [23]. Further, a lasso penalty is introduced to conduct simultaneous parameter estimation and variable selection, considering that the lasso penalty method has good performance for solving the overfitting problem caused by the introduction of factors in adjacent windows and selecting the features and the corresponding exposures with relatively significant influence on the response. In fact, the lasso method has been applied for linear-distributed lag modelling (e.g., [55]).

3. Methodology

In this section, by combining the logistic regression method and unconstrained distributed lag model, we seek to estimate which indicators and in which period prior to the distress event best predicts financial distress. First, we construct Model 1 that represents the "accounting-only" model and incorporates the financial ratios. We introduce the 3-period-lagged financial ratios as independent variables into Model 1 and use the model to predict the financial distress event in year t by using the data of relevant indicator of the consecutive years, $t-3$, $t-4$, $t-5$, simultaneously. Note that t refers to the current year in this paper. Then, we construct Model 2, which represents the 'accounting plus macroeconomic indicators' model, and includes, in addition to the accounting variables, 3-period-lagged macroeconomic indicators. Then, we introduce lasso penalty to the models and implement the coefficient estimation and feature selection. Further, we provide the algorithm framework of alternating direction method of multipliers (ADMM) that yields the global optimum for convex and the non-smooth optimization problem to obtain the optimal estimation for the coefficients. Finally, we propose a support vector machine model that includes the lagged variables of the accounting

ratios and macroeconomic factors. This model is used for comparison of the predictive performance of the logistic model with a distributed lag of variables.

3.1. Logistic Regression Framework with Distributed Lags

3.1.1. The Logistic Regression-Distributed Lag Model with Accounting Ratios Only

The logistic regression may be the most frequently used technique in the financial distress prediction field and has been widely recognized ([11,20]). We propose a logistic model composed of lagged explanatory variables. Similar to the distributed lag linear model, the model has the following general form:

$$P\left(Y_{i,t}=1\middle|X_{i,t-l}\right)=(1+\exp(-(\alpha_0+\sum\nolimits_{l=0}^{L}\alpha_{t-l}^{T}X_{i,t-l})))^{-1},\ t=t_0+L,t_0+L+1,\ \ldots,\ t_0+L+d \quad (2)$$

In (2), $Y_{i,t}$ is a binary variable, and if $Y_{i,t}$ =1, then it means that firm I at time t is a financially distressed company, otherwise, firm i (i = 1, 2, … , n) is a financially healthy company, corresponding the case of $Y_{i,t}$ = 0; α_0 is intercept, and $\alpha_{t-l}=(\alpha_{t-l,1},\alpha_{t-l,2},\ldots,\alpha_{t-l,p})^{\mathrm{T}}$ is the coefficient vector for the explanatory variable vector $X_{i,t-l}$ at time $t-l$; $X_{i,t-l}$ is the p-dimension accounting ratio vector for firm i at time $t-l$, l = 0, 1, 2, … , L; l, L are the lag number and the maximum lag; t_0 is the beginning of the observation period and d is the duration of observation. The idea in (2) is that the likelihood of occurrence of the financial distress at time t for a listed company may depend on X measured not only in the current time t, but also in the previous time windows $t-1$ through $t-L$.

In Formula (2), we assume a five-year effect and set the maximum lag L = 5, given that (1) the effect of the explanatory variable on the response variable may decline to zero in the time series data scenario; (2) the considered length of lag is not more than 5 years in most of the previous studies of financial distress prediction (see [1,7,37]). Besides, we set directly the coefficient α_{t-0}, α_{t-1}, α_{t-2} for the variables in the current year and the previous two year before ST to be 0 vectors, because (1) the financial statement in the current year (year t) is not available for financial distressed companies labeled as in financial distress in year t, since the financial statement is published at the end of the year, but special treatment probably occurs before the publication; (2) designation of an ST company depends on the financial and operating situations of the previous year before the label of ST. Put simply, it is not meaningful to forecast ST risk 0, 1 or 2 years ahead (see [7,45]). Therefore, the logistic model containing the 3~5-period-lagged financial indicators, defined as Model 1, is presented as follows:

$$P(Y_{i,t}=1)=(1+\exp(-\alpha_0-{\alpha_{t-3}}^{\mathrm{T}}X_{i,t-3}-{\alpha_{t-4}}^{\mathrm{T}}X_{i,t-4}-{\alpha_{t-5}}^{\mathrm{T}}X_{i,t-5})^{-1} \quad (3)$$

In Equation (3), $Y_{i,t}$ is binary response and is defined the same in (2); $X_{i,t-3}$, $X_{i,t-4}$ and $X_{i,t-5}$ are the p-dimensional financial indicator vectors of firm i observed in year $t-3$, $t-4$ and $t-5$; α_0, α_{t-3}, α_{t-4}, α_{t-5} are intercept terms and the coefficient vectors for the explanatory vectors $X_{i,t-3}$, $X_{i,t-4}$ and $X_{i,t-5}$, respectively, and α_{t-l} (l = 3, 4, 5) stands for the average effect of increasing by one unit in $X_{i,t-l}$ on the log of the odd of the financial distress event holding others constants. Of course, in Model 1, we consider the effect of changes in financial ratios on financial distress probability during three consecutive years ($t-3$, $t-4$, $t-5$).

3.1.2. The Logistic Regression-Distributed Lag Model with Accounting Plus Macroeconomic Variables

We further add the macro-economic factors into Equation (3) to detect the influence of macroeconomic conditions, in addition to financial indicators. Model 2, including both accounting variables and macroeconomic variables, takes the following form:

$$P(Y_{i,t}=1)=(1+\exp(-\alpha_0-{\alpha_{t-3}}^{\mathrm{T}}X_{i,t-3}-{\alpha_{t-4}}^{\mathrm{T}}X_{i,t-4}-{\alpha_{t-5}}^{\mathrm{T}}X_{i,t-5}-{\boldsymbol{\eta}_{t-3}}^{\mathrm{T}}Z_{i,t-3}-{\boldsymbol{\eta}_{t-4}}^{\mathrm{T}}Z_{i,t-4}-{\boldsymbol{\eta}_{t-5}}^{\mathrm{T}}Z_{i,t-5}))^{-1} \quad (4)$$

In Equation (4), Z_{t-l} (l = 3, 4, 5) represents the m-dimensional macroeconomic factor vector of year $t - l$; $\boldsymbol{\eta}_{t-l}$ (l = 3, 4, 5) is the coefficient vector for $Z_{i,t-l}$; the others are defined as in Equation (3). Similarly, $\boldsymbol{\eta}_{j,t-3} + \boldsymbol{\eta}_{j,t-4} + \boldsymbol{\eta}_{j,t-4}$ represent the cumulative effects on log odd of the distress event of the j-th (j = 1, 2, . . . , m) macroeconomic factor.

Models 1 and 2, marked as Equations (3) and (4), can reflect the continued influence of the financial statement and macroeconomic conditions for multi-periods on the response; however, a considerable amount of potentially helpful financial ratios, macroeconomic factors and their lags may bring redundant information, thus decreasing the models' forecast performances. In the following section, we implement feature selection by introducing lasso penalty into the financial distress forecast logistic models. Further, we provide an ADMM algorithm framework to obtain the optimal estimation for the coefficients.

3.2. The Lasso–Logistic Regression-Distributed Lag Model

There is currently much discussion about the lasso method. Lasso, as an l_1-norm penalization approach, has been actively studied. In particular, lasso has been used on the distributed lag linear model, and lasso estimators for coefficients are obtained through minimizing the residual sum of squares and the l_1-norm of coefficients simultaneously (e.g., [55]). For the logistic model with lagged financial variables (3), we can extend to logistic–lasso as follows in Equation (5):

$$(\hat{\alpha}_0, \hat{\alpha}) = \underset{\alpha_0,\alpha}{argmin} f\left(\alpha_0, \alpha \middle| X_{i,t-3}, X_{i,t-4}, X_{i,t-5}, Y_{i,t}\right) + \lambda\|\alpha\|_1 \tag{5}$$

where

$$f\left(\alpha_0, \alpha \middle| X_{it}, Y_{i,t}\right) = \sum_{t=t_0+5}^{t_0+d} \sum_{i=1}^{n} (-Y_{i,t}\left(\alpha_0 + \alpha^T X_{it}\right) + \ln(1 + \exp\{\alpha_0 + \alpha^T X_{it}\}))$$

and $\hat{\alpha}_0$ and $\hat{\alpha}$ denote the maximum likelihood estimations for intercept α_0 and coefficient vector α; f denotes the minus log-likelihood function of Model 1 and can be regarded as the loss function of the observations; $\alpha = ({\alpha_{t-3}}^{\mathrm{T}}, {\alpha_{t-4}}^{\mathrm{T}}, {\alpha_{t-5}}^{\mathrm{T}})^{\mathrm{T}}$ are the unknown coefficients for explanatory variables; $X_{it} = ({X_{i,t-3}}^{\mathrm{T}}, {X_{i,t-4}}^{\mathrm{T}}, {X_{i,t-5}}^{\mathrm{T}})^{\mathrm{T}}$, $Y_{i,t}$ are known training observations and defined as above; λ is the turning parameter; $\|\cdot\|_1$ denotes l_1-norm of a vector, i.e., the addition of absolute values of each element of a vector; t_0 and d are defined as before; n is the number of observed company samples.

Introducing the auxiliary variable $\beta \in \boldsymbol{R^{3p}}$, the lasso–logistic model (5) can be explicitly rewritten as follows:

$$\underset{\alpha_0,\alpha,\beta}{min} f\left(\alpha_0, \alpha \middle| X_{i,t-3}, X_{i,t-4}, X_{i,t-5}, Y_{i,t}\right) + \lambda\|\beta\|_1 \; s.t. \alpha = \beta \tag{6}$$

In this paper, we solve the optimization problem (6) by using alternating direction method of multipliers (ADMM) algorithm that was first introduced by [56]. ADMM is a simple but powerful algorithm and can be viewed as an attempt to blend the benefits of dual decomposition [57] and augmented Lagrangian methods for constrained optimization [58]. Now, the ADMM algorithm becomes a benchmark first-order solver, especially for convex and non-smooth minimization models with separable objective functions (see [59,60]), thus, it is applicable for the problem (6).

The augmented Lagrangian function of the optimization problem (6) can be defined as

$$L_\rho(\alpha_0, \alpha, \beta, \theta) = f(\alpha_0, \alpha | X_{it}, Y_{i,t}) + \lambda\|\beta\|_1 - \theta^T(\alpha - \beta) + \frac{\rho}{2}\|\alpha - \beta\|_2^2 \tag{7}$$

where L_ρ is the Lagrange function; θ is a Lagrange multiplier vector and ρ(>0) is an augmented Lagrange multiplier variable. In this paper, ρ is predetermined to be 1 for simplicity. Then, the iterative scheme of ADMM for the optimization problem (6) reads as

$$\left(\alpha_0^{k+1}, \alpha^{k+1}\right) = \underset{\alpha_0,\alpha}{argmin} L_\rho\left((\alpha_0, \alpha), \beta^k, \theta^k\right) \tag{8a}$$

$$\beta^{k+1} = \underset{\beta}{argmin} L_\rho\left(\alpha^{k+1}, \beta, \theta^k\right) \tag{8b}$$

$$\theta^{k+1} = \theta^k - \rho\left(\alpha^{k+1} - \beta^{k+1}\right) \tag{8c}$$

In (8a)–(8c), ${\alpha_0}^{k+1}$, α^{k+1}, β^{k+1}, and θ^k are the values of α_0, α, β, and θ the k-th iterative step of the ADMM algorithm, respectively. Further, the ADMM scheme (8a)–(8c) can be specified as

$$\left(\alpha_0^{k+1}, \alpha^{k+1}\right) = \underset{\alpha_0, \alpha}{argmin} f(\alpha_0, \alpha) - \left(\theta^k\right)^T\left(\alpha - \beta^k\right) + \frac{\rho}{2}\|\alpha - \beta^k\|_2^2 \tag{9a}$$

$$\beta^{k+1} = \underset{\beta}{argmin} \lambda\|\beta\|_1 - \left(\theta^k\right)^T\left(\alpha^{k+1} - \beta\right) + \frac{\rho}{2}\|\alpha^{k+1} - \beta\|_2^2 \tag{9b}$$

$$\theta^{k+1} = \theta^k - \rho\left(\alpha^{k+1} - \beta^{k+1}\right) \tag{9c}$$

The sub-problem in (9a), that is, the convex and smooth optimization problem, can be fast solved by the Newton method [61], after setting the initial θ, β to be arbitrary constants. More specifically, let ${\alpha_*}^{k+1} = ({\alpha_0}^{k+1}; \alpha^{k+1})$ and ${\alpha_*}^{k+1}$ be calculated via the following process:

$$\alpha_*^{k+1} = \alpha_*^k - \left(\nabla^2 l\right)^{-1} \nabla l \tag{10}$$

where

$$l(\alpha_*) = l(\alpha_0; \alpha) = f(\alpha_0, \alpha) - \left(\theta^k\right)^T\left(\alpha - \beta^k\right) + \frac{\rho}{2}\|\alpha - \beta^k\|_2^2$$

and $\nabla^2 l \in R^{(3p+1)\times(3p+1)}$, $\nabla l \in R^{3p+1}$ are the hessian matrix and the derivative of differentiable function l with respect to α_*, respectively. For sub-problem (9b), its solution is analytically given by

$$\beta_r^{k+1} = \begin{cases} \alpha_r^{k+1} - \frac{\lambda + \theta_r^k}{\rho}, \alpha_r^{k+1} > \frac{\lambda + \theta_r^k}{\rho} \\ 0, \frac{-\lambda + \theta_r^k}{\rho} < \alpha_r^{k+1} \leq \frac{\lambda + \theta_r^k}{\rho} \\ —, \ \alpha_r^{k+1} \leq \frac{-\lambda + \theta_r^k}{\rho} \end{cases} \tag{11}$$

where ${\beta_r}^{k+1}$, ${\alpha_r}^{k+1}$ and ${\theta_r}^k$ are the r-th components of β^{k+1}, ${\alpha_r}^{k+1}$ and θ^k, respectively, for the k-th iterative step and $r = 1, 2, \ldots, 3p$.

The choice of tuning parameters is important. In this study, we find an optimal tuning parameter λ by the 10-fold cross validation method. We then compare the forecast accuracy of each method based on the mean area under the curve (*MAUC*) given as follows:

$$MAUC(\lambda) = \sum_{j=1}^{10} AUC^j(\lambda)/10 \tag{12}$$

where $AUC^j(\lambda)$ denotes the area under the receiver operating characteristic (ROC) curve on j-th validation set for each tuning parameter λ.

So far, the lasso estimators for the logistic model (5) including 3~5-period-lagged financial ratios have been obtained by following the above procedures. For the convenience of readers, we summarize the whole optimization procedures in training the lasso–logistic with lagged variables and describe them in Algorithm 1.

Algorithm 1. An alternating direction method of multipliers (ADMM) algorithm framework for lasso–logistic with lagged variables (5). [1]: Dual residual and prime residua denote $\|\beta^{k+1} - \beta^k\|_2$ and $\|\alpha^{k+1} - \beta^k{+}1\|_2$ respectively. [2]: N denotes the maximum iterative number of the ADMM algorithm.

Require:

1. Training data $\{X_{i,t-3}, X_{i,t-4}, X_{i,t-5}, Y_{i,t}\}$, where $X_{i,t-l} \in R^P$, $l = 3, 4, 5$ and $Y_{i,t} \in \{0,1\}$, $i = 1, 2, \ldots, n$, $t = t_0 + 5$, $t_0 + 6, \ldots, t_0 + d$
2. Turning parameter λ
3. Choose augmented Lagrange multiplier $\rho = 1$. Set initial $(\theta^0, \beta^0) \in R \times R^P$, $(\alpha_0{}^0, \alpha^0) \in R \times R^P$ and stopping criterion $\varepsilon = 10^{-6}$.

Ensure:

4. **While** not converging (i.e., dual residual and prime residual [1] are greater than stopping criterion of 10^{-6}) **do**
5. **For** $k = 0, 1, \ldots, N$[2] **do**
6. Calculate α^{k+1} following the Newton algorithm (10)
7. Calculate β^{k+1} following (11)
8. Update $\theta^{k+1} \leftarrow \theta^k - \rho(\alpha^{k+1} - \beta^{k+1})$
9. **End for**
10. **End while**.

For the logistic model (4) with lag variables of the financial ratio and macroeconomic indicators, we can also extend the lasso as follows in (13):

$$(\hat{\alpha}_0, \hat{\gamma}) = \underset{\alpha_0, \gamma}{argmin} f\Big(\alpha_0, \gamma \Big| X_{i,t-3}, X_{i,t-4}, X_{i,t-5}, Z_{i,t-3}, Z_{i,t-4}, Z_{i,t-5}, Y_{i,t}\Big) + \lambda \|\gamma\|_1 \tag{13}$$

where $\hat{\gamma} = (\hat{\alpha}, \hat{\eta})$ is the lasso estimator vector for coefficients of lagged financial ratios and macroeconomic indicators; $\gamma = (\alpha^{\mathrm{T}}, \eta^{\mathrm{T}})^{\mathrm{T}}$ represents the unknown coefficients for explanatory variables; α, $\eta = (\eta_{t-3}{}^{\mathrm{T}}, \eta_{t-4}{}^{\mathrm{T}}, \eta_{t-5}{}^{\mathrm{T}})^{\mathrm{T}}$ and the others are defined as in Equations (4) and (5). The lasso estimator for model (13) can also be found by using the ADMM algorithm presented above.

3.3. The Lasso–SVM Model with Lags for Comparison

The support vector machine (SVM) is a widely used linear classifier with high interpretability. In this sub-section, we construct a lasso–SVM model that includes the 3-period-lagged financial indicators for comparison with the lasso–logistic-distributed lag model. The SVM formulation combing the original soft-margin SVM model [62] and a 3~5–period-lagged financial ratio variable vector is as follows:

$$\begin{cases} \underset{\alpha_0, \alpha, \xi}{min} \ \frac{1}{2}\|\alpha\|_2^2 + C \sum\limits_{t=t_0+5}^{t_0+d} \sum\limits_{i=1}^{n} \xi_{i,t} \\ s.t.\ Y_{i,t}\Big(\alpha_0 + \alpha_{t-3}{}^T X_{i,t-3} + \alpha_{t-4}{}^T X_{i,t-4} + \alpha_{t-5}{}^T X_{i,t-5}\Big) \geq 1 - \xi_{i,t},\ \xi_{i,t} \geq 0, \\ \qquad\qquad i = 1, 2, \ldots, n,\ t = t_0 + 5, \ldots, Id \end{cases} \tag{14}$$

In (14), α_0 (intercept) and $\alpha = (\alpha_{t-3}; \alpha_{t-4}; \alpha_{t-5})$ (normal vector) are the unknown coefficients of hyper-plane $f(X_{it}) = \alpha_0 + \alpha^T X_{it}$; $\|\cdot\|_2$ denotes l_2-norm of a vector; C is the penalty parameter and a predetermined positive value; $\xi_{i,t}$ is the unknown slack variable; $Y_{i,t}$ is a binary variable and $Y_{i,t} = 1$, when firm i is a financially disIressed company in year t, otherwise $Y_{i,t} = -1$; $X_{it} = (1; X_{i,t-3}; X_{i,t-4}; X_{i,t-5})$ denotes the observation vector of 3~5-period-lagged financial indicators for firm i; n represents the number of observations; t_0 and d denote the beginning and length of the observation period, respectively.

By introducing the hinge loss function, the optimization problem (14) has the equivalent form as follows [63]:

$$\underset{\alpha_*}{min} \sum_{t=t_0+5}^{t_0+d} \sum_{i=1}^{n} \left[1 - Y_{i,t}\Big(\alpha_*{}^T X_{it}\Big)\right]_+ + \lambda \|\alpha_*\|_2^2 \tag{15}$$

where $\alpha_* = (\alpha_0; \alpha)$, $[\cdot]_+$ indicates the positive part, i.e., $[x]_+ = \max\{x,0\}$, and the turning parameter $\lambda = 1/2C$.

Considering that it is regularized by l_2-norm, the SVM forces all nonzero coefficient estimates, which leads to the problem of its inability to select significant features. Thus, to prevent the influence of noise features, we replace l_2-norm in the optimization problem (15) with l_1-norm, which is able to simultaneously conduct feature selection and classification. Furthermore, for computational convenience, we replace the hinge loss function in (15) with the form of the sum of square, and present the optimization problem combining the SVM model and the lasso method (l_1 regularization) as follows:

$$\hat{\alpha}_* = \underset{\alpha_*}{argmin} \sum_{t=t_0+5}^{t_0+d} \sum_{i=1}^{n} \left([1 - Y_{i,t} {\alpha_*}^T X_{it}]_+ \right)^2 + \lambda \|\alpha_*\|_1 \tag{16}$$

In (16), $\hat{\alpha}_*$ is the optimal estimated value for the coefficients of the SVM model, and the others are defined as above. Similarly with the process of the solution to the problem (5) as presented previously, first by introducing an auxiliary variable $\beta \in \boldsymbol{R^{3p+1}}$, the lasso–SVM model (16) can be explicitly rewritten as follows:

$$\underset{\alpha_*, \beta}{min} \sum_{t=t_0+5}^{t_0+d} \sum_{i=1}^{n} \left([1 - Y_{i,t} {\alpha_*}^T X_{it}]_+ \right)^2 + \lambda \|\beta_*\|_1 \ \ s.t.\, \alpha_* = \beta_* \tag{17}$$

Then, the augmented Lagrangian function of the optimization problem (17) can be accordingly specified as

$$L_\rho(\alpha_*, \beta_*, \theta_*) = \sum_{t=t_0+5}^{t_0+d} \sum_{i=1}^{n} \left([1 - Y_{i,t} {\alpha_*}^T X_{it}]_+ \right)^2 + \lambda \|\beta_*\|_1 - {\theta_*}^T (\alpha_* - \beta_*) + \frac{\rho}{2} \|\alpha_* - \beta_*\|_2^2 \tag{18}$$

where $\theta \in \boldsymbol{R^{3p+1}}$ and $\rho \in \boldsymbol{R}$ are the Lagrange and the augmented Lagrange multipliers, respectively. Then, the iterative scheme of ADMM for the optimization problem (18) is similar with (8a)–(8c) and can be accordingly specified as

$${\alpha_*}^{k+1} = \underset{\alpha_*}{argmin} \sum_{t=t_0+5}^{t_0+d} \sum_{i=1}^{n} \left(\left[1 - Y_{i,t} \left({\alpha_*}^T X_{it} \right) \right]_+ \right)^2 - \left({\theta_*}^k \right)^T \left(\alpha_* - {\beta_*}^k \right) + \frac{\rho}{2} \|\alpha_* - {\beta_*}^k\|_2^2 \tag{19a}$$

$${\beta_*}^{k+1} = \underset{\beta_*}{argmin} \lambda \|\beta_*\|_1 - \theta^k \left({\alpha_*}^{k+1} - \beta_* \right) + \frac{\rho}{2} \|{\alpha_*}^{k+1} - \beta_*\|_2^2 \tag{19b}$$

$${\theta_*}^{k+1} = {\theta_*}^k - \rho \left({\alpha_*}^{k+1} - {\beta_*}^{k+1} \right) \tag{19c}$$

The finite Armijo–Newton algorithm [61] is applied for solving the α-sub-problem (19a), which is a convex piecewise quadratic optimization problem. Its objective function is first-order differentiable but not twice-differentiable with respect to α_*, which precludes the use of a regular Newton method. $F(\alpha_*)$ is the objective function of the sub-optimization problem (19a) and its gradient and generalized Hessian matrix are presented as follows Equations (20) and (21):

$$\nabla F(\alpha_*) = -2 \sum_{t=t_0+5}^{t_0+d} \sum_{i=1}^{n} Y_{i,t} X_{it} (1 - Y_{i,t} {\alpha_*}^T X_{it})_+ - \theta^k + \rho \left(\alpha_* - {\beta_*}^k \right) \tag{20}$$

$$\partial^2 F(\alpha_*) = 2 \sum_{t=t_0+5}^{t_0+d} \sum_{i=1}^{n} diag(1 - Y_{i,t} {\alpha_*}^T X_{it})_* X_{it} X_{it}^T + \rho I \tag{21}$$

where $\mathbf{I} \in \boldsymbol{R^{3p+1}}$ is identity matrix and $diag(1 - Y_{i,t}\ {\alpha_*}^T X_{it})_*$ is a diagonal matrix in that the j-th (j = 1, 2, … , $3p$ + 1) diagonal entry is a sub-gradient of the step function $(\cdot)_+$ as

$$(diag(1 - Y_{i,t}{\alpha_*}^T X_{it})_*)_{jj} \begin{cases} = 1 & if\ 1 - Y_{i,t}{\alpha_*}^T X_{it} > 0, \\ \in [0,1] & if\ 1 - Y_{i,t}\alpha_*^T X_{it} = 0, \\ = 0 & if\ 1 - Y_{i,t}\alpha_*^T X_{it} < 0. \end{cases} \quad (22)$$

The whole optimization procedure applied to solve the α-sub-problem (19a) is described in Algorithm 2.

Algorithm 2. A finite Armijo–Newton algorithm for the sub-problem (19a). [1]: δ is the parameter associated with finite Armijo Newton algorithm and between 0 and 1.

Require:

1. Training data $\{X_{i,t-3}, X_{i,t-4}, X_{i,t-5}, Y_{i,t}\}$, where $X_{i,t-l} \in R^P$, $l = 3, 4, 5$ and $Y_{i,t} \in \{1,-1\}$, $I = 1,2, \ldots, n$, $t = t_0 + 5, t_0 + 6, \ldots, t_0 + d$
2. Turining parameter λ
3. Choose augmented Lagrange multiplier $\rho = 1$. Set initial $({\theta_*}^0, {\beta_*}^0, {\alpha_*}^0) \in R \times R^{3P+1} \times R^{3P+1}$ and stopping criterion $\varepsilon = 10^{-6}$

Ensure:

4. **While** not converging (i.e., $\|\nabla F[({\alpha_*}^i) - \partial^2 F({\alpha_*}^i)^{-1} \nabla F({\alpha_*}^i)]\|_2 \geq \varepsilon$) **do**
5. Calculate the Newton direction $d^i = -\partial^2 F({\alpha_*}^i)^{-1} \nabla F({\alpha_*}^i)$ following (20)–(22)
6. Choose $\delta^1 = 0.4$ and find stepsize $\tau_i = \max\{1, 1/2, 1/4, \ldots\}$ such that $F({\alpha_*}^i) - F({\alpha_*}^i + \tau_i\, d^i) \geq -\delta\tau_i \nabla F({\alpha_*}^i)^T d^i$ is satisfied
7. Update $(\alpha_*)^{i+1} \leftarrow (\alpha_*)^i + \tau_i\, d^i$, $i \leftarrow I + 1$
8. **End while**
9. Output ${\alpha_*}^{k+1} = {\alpha_*}^i$

The finite Armijo–Newton algorithm can guarantee the unique global minimum solution in a finite number of iterations. The details of proof of the global convergence of the sequence to the unique solution can be found in [61]. For the sub-problem (19b), its solution can be also analytically given by (11) presented above, after replacing α, β and θ with α_*, β_* and θ_*.

So far, the lasso estimators for the SVM model (16), including 3~5-period-lagged financial ratios, have been obtained by following the above procedures. For the convenience of readers, we summarize the whole optimization procedures in training the lasso–SVM with lagged variables and describe them in Algorithm 3. It is worth to note that the estimators for the lasso–SVM model that contain 3~5-period-lagged financial ratios and macro-economic indicators can be also obtained by the following algorithm similarly.

Algorithm 3. An ADMM algorithm framework for lasso–support vector machine (SVM) with lagged variables (16)

Require:

1. Training data $\{X_{i,t-3}, X_{i,t-4}, X_{i,t-5}, Y_{i,t}\}$, where $X_{i,t-l} \in R^P$, $l = 3, 4, 5$ and $Y_{i,t} \in \{1,-1\}$, $i = 1, 2, \ldots, n$, $t = t_0 + 5, t_0 + 6, \ldots, t_0 + d$
2. Turining parameter λ
3. Choose augmented Lagrange multiplier $\rho = 1$. Set initial $({\theta_*}^0, {\beta_*}^0, {\alpha_*}^0) \in R \times R^{3P+1} \times R^{3P+1}$ and stopping criterion $\varepsilon = 10^{-6}$

Ensure:

4. **While** not converging (i.e., either $\|\beta^{k+1} - \beta^k\|_2$ or $\|\alpha^{k+1} - \beta^{k+1}\|_2$ is greater than stopping criterion of 10^{-6}) **do**
5. **For** $k = 0, 1, \ldots, N$ do
6. Calculate ${\alpha_*}^{k+1}$ following finte Armijo–Newton algorithm displayed in Algorithm 2
7. Calculate ${\beta_*}^{k+1}$ following (11)
8. Update ${\theta_*}^{k+1} \leftarrow {\theta_*}^k - \rho({\alpha_*}^{k+1} - {\beta_*}^{k+1})$
9. **End for**
10. **End while**

4. Data

4.1. Sample Description

The data used in the study are limited to manufacturing corporations. The manufacturing sector plays an important role in contributing to the economic growth of a country, especially a developing country [64]. According to the data released by the State Statistical Bureau of China, manufacturing accounts for 30% of the country's GDP. China's manufacturing sector has the largest number of listed companies as well as the largest number of ST companies each year. On the other hand, according to the data disclosed by the China Banking Regulatory Commission, in the Chinese manufacturing sector, the non-performing loan ratio has been increasing. For example, there was a jump in the non-performing loan ratio from 3.81% in December of 2017 to 6.5% in June of 2018. Therefore, it is quite important to establish an effective early warning system aiming to assess financial stress and prevent potential financial fraud of a listed manufacturing company for market participants, including investors, creditors and regulators.

In this paper, we selected 234 listed manufacturing companies from the Wind database. Among these, 117 companies are financially healthy and 117 are financially distressed, i.e., the companies being labeled as "special treatment". The samples were selected from 2007 to 2017, since the Ministry of Finance of the People's Republic of China issued the new "Accounting Standards for Business Enterprises" (new guidelines), which required that all listed companies be fully implemented from January 1, 2007. Similar to [7], [16] and [45], all 117 financially distressed companies receive ST due to negative net profit for two consecutive years. There were respectively 10, 9, 17, 24, 26 and 31 companies labeled as ST or *ST in each year from 2012 to 2017. The same number of financially healthy companies were selected in each year. Considering the regulatory requirement and qualified data of listed companies, our data sample enforces the use of 2007 (t_0) as the earliest estimation window available in forecasting a listed company's financial distress. Meanwhile, the maximum order lag used in our models is as long as 5 (years); that is, the maximum horizon is 5 years, so the number of special-treated (ST) companies was counted since 2012 ($t_0 + 5$). Furthermore, we divided the whole sample group into two groups: the training sample and the testing sample. The training sample is from 2012 to 2016, includes the data of 172 companies and is used to construct the models and estimate the coefficients. Correspondingly, the testing sample is from 2017, includes the data of 62 companies and is used to evaluate the predicting performance of the models.

4.2. Covariate

In this paper, we use the factors measured in consecutive time windows $t-3$, $t-4$ and $t-5$ to predict a listed company's financial status at time t (t = 2012, 2013, ... , 2017). Therefore, we define response y as whether a Chinese manufacturing listed company was labeled as "special treatment" by China Securities Regulatory Commission at time t (t = 2012, 2013, ... , 2017) and input explanatory variables as their corresponding financial indicators based on financial statements reported at $t-3$, $t-4$ and $t-5$. For example, we define response y as whether a Chinese manufacturing listed company was labeled as "special treatment" during the period of from January 1, 2017 to December 31, 2017 (denoted as year t) and (1) input explanatory variables as their corresponding financial indicators based on financial statements reported on December 31, 2014 (denoted as year $t-3$), in December, 2013 (denoted as year $t-4$) and in December, 2012 (denoted as year $t-5$); through this way, the time lags of the considered financial indicators and the responses are between 3 to 5 years; (2) input explanatory variables as macroeconomic indicators based on the statements reported on December 31, 2014, 2013 and 2012 by the Chinese National Bureau of Statistics; through this way, the time lags of the considered macroeconomic indicators and the response are also between 3 to 5 years. The effect of time lags of 3 to 5 years of financial indicators on the likelihood of occurrence of financial distress is separately suggested by some previous research of early warnings of listed companies' financial

distress, but the varying effects of these time lags that occur in one prediction model are not yet considered in the existing studies.

4.2.1. Firm-Idiosyncratic Financial Indicator

An original list of 43 potentially helpful ratios is compiled for prediction and provided in Table 1 because of the large number of financial ratios found to be significant indicators of corporate problems in past studies. These indicators are classified into five categories, including solvency, operational capability, profitability, structural soundness and business development and capital expansion capacity. All variables used for calculation of financial ratios are obtained from the balance sheet, income statements or cash flow statements of the listing companies. These financial data for financially distressed companies are collected in year 3, 4 and 5 before the companies receive the ST label. For example, the considered year when the selected financially distressed companies receive ST is 2017; the financial data are obtained in 2014, 2013 and 2012. Similarly, the data for financially healthy companies are also collected in 2014, 2013 and 2012. Model 1 (the accounting-only model) will be constructed using all the data in the following context. The model is used to predict whether a company is labeled in year t, incorporating the financial data of three consecutive time windows, $t-3$, $t-4$ and $t-5$ (t = 2012, 2014, . . . , 2017).

Table 1. List of financial indicators.

Solvency	**Operational Capabilities**
1 Total liabilities/total assets (TL/TA)	9 Sales revenue/average net account receivable (SR/ANAR)
2 Current assets/current liabilities (CA/CL)	10 Sales revenue/average current assets (SR/ACA)
3 (Current assets–inventory)/current liabilities (CA-I)/CL	11 Sales revenue/average total assets (AR/ATA)
4 Net cash flow from operating activities/current liabilities (CF/CL)	12 Sales cost/average payable accounts (SC/APA)
5 Current liabilities/total assets (CL/TA)	13 Sales cost/sales revenue (SC/SR)
6 Current liabilities/shareholders' equity (CL/SE)	14 Impairment losses/sales profit (IL/SP)
7 Net cash flow from operating and investing activities/total liabilities (NCL/TL)	15 Sales cost/average net inventory (SC/ANI)
8 Total liabilities/total shareholders' equity (TSE/TL)	16 Sales revenue/average fixed assets (SR/AFA)
Profitability	Structural Soundness
17 Net profit/average total assets (NP/ATA)	27 Net asset/total asset (NA/TA)
18 Shareholder equity/net profit (SE/NP)	28 Fixed assets/total assets (FA/TA)
19 (Sales revenue–sales cost)/sales revenue (SR-SC)/SR	29 Shareholders' equity/fixed assets (SE/FA)
20 Earnings before interest and tax/average total assets (EIA/ATA)	30 Current liabilities/total liabilities (CL/TL)
21 Net profit/sales revenue (NP/SR)	31 Current assets/total assets (CA/TA)
22 Net profit/average fixed assets (NP/AFA)	32 Long-term liabilities/total liabilities (LL/TL)
23 Net profit attributable to shareholders of parent company/sales revenue (NPTPC/SR)	33 Main business profit/net income from main business (MBP/NIMB)
24 Net cash flow from operating activities/sales revenue (NCFO/SR)	34 Total profit/sales revenue (TP/SR)
25 Net profit/total profit (NP/TP)	35 Net profit attributable to shareholders of the parent company/net profit (NPTPC/NP)
26 Net cash flow from operating activities/total assets at the end of the period (NCFO/TAEP)	36 Operating capital/total assets (OC/TA)
	37 Retained earnings/total assets (RE/TA)
Business Development and Capital Expansion Capacity	
38 Main sales revenue of this year/main sales revenue of last year (MSR(t)/MSR(t-1))	41 Net assets/number of ordinary shares at the end of year (NA/NOS)
39 Total assets of this year/total assets of last year (TA(t)/TA(t-1))	42 Net cash flow from operating activities/number of ordinary shares at the end of year (NCFO/NOS)
40 Net profit of this year/net profit of last year (NP(t)/NP(t-1))	43 Net increase in cash and cash equivalents at the end of year/number of ordinary shares at the end of year (NICCE/NOS)

4.2.2. Macroeconomic Indicator

Besides considering three consecutive period-lagged financial ratios for the prediction of financial distress of Chinese listed manufacturing companies, we also investigated the associations between macro-economic conditions and the possibility of falling into financial distress of these companies. The macro-economic factors include GDP growth, inflation, unemployment rate in urban areas and consumption level growth, as described in Table 2. GDP growth is widely understood to be an important variable to measure economic strength and prosperity; the increase in GDP growth may decrease the likelihood of distress. High inflation and high unemployment that reflect a weaker economy

may increase the likelihood of financial distress. Consumption level growth reflects the change in consumption level and its increase may reduce the likelihood of financial distress.

Table 2. List of macroeconomic factors.

Figure [1]	Description
1 Real GDP growth (%)	Growth in the Chinese real gross domestic product (GDP) compared to the corresponding period of previous year (GDP growth is documented yearly and by province).
2 Inflation rate (%)	Percentage changes in urban consumer price compared to the corresponding period of the previous year (inflation rate is documented regionally).
3 Unemployment rate (%)	The data derived from the Labor Force Survey (population between 16 years old and retirement age, unemployment rate is documented yearly and regionally).
4 Consumption level growth (%)	Growth in the Chinese consumption level index compared to the corresponding period of the previous year (consumption level growth is documented yearly and regionally).

[1]: All data of the macro-economic covariates are collected from the National Bureau of Statistics of China.

In the following empirical part, Model 2 represents the "accounting plus macroeconomic indicators" model and includes, in addition to the accounting variables, 3-period-lagged macroeconomic indicators. We collected the corresponding macroeconomic data in each year from 2007 to 2012 for all 234 company samples and the raw macroeconomic data are from the database of the Chinese National Bureau of Statistics.

4.3. Data Processing

The results in the existing studies suggest that the predicting models of standardized data yield better results in general [65]. Therefore, before the construction of the models, a standardization processing is implemented based on the following linear transformations:

$$x_{ij}(t) = \frac{u_{ij}(t) - \min\limits_{1\le i\le 234}\left\{u_{ij}(t)\right\}}{\max\limits_{1\le i\le 234}\left\{u_{ij}(t)\right\} - \min\limits_{1\le i\le 234}\left\{u_{ij}(t)\right\}} \tag{23}$$

where $x_{ij}(t)$ denotes the standardized value of the j-th financial indicator for the i-th firm in year t, and $j = 1, 2, \ldots, 43$, $i = 1, 2, \ldots, 234$, and $t = 2007, 2008, \ldots, 2012$; $u_{ij}(t)$ denotes the original value of the j-th indicator of the i-th company in year t. Linear transformation scales each variable into the interval [0, 1]. Similarly, the following formula is used for data standardization of the macro-economic factor:

$$z_{ij}(t) = \frac{v_{ij}(t) - \min\limits_{1\le i\le 234}\left\{v_{ij}(t)\right\}}{\max\limits_{1\le i\le 234}\left\{v_{ij}(t)\right\} - \min\limits_{1\le i\le 234}\left\{v_{ij}(t)\right\}} \tag{24}$$

In formula (24), $z_{ij}(t)$ denotes the standardized value of the j-th macro-economic factor in year t; $v_{ij}(t)$ denotes the original value of the j-th indicator of the i-th company in year t, where $j = 1, 2, 3, 4$, $i = 1, 2, \ldots, 234$, and $t = 2007, 2008, \ldots, 2012$. It is worth noting that the assignment to $v_{ij}(t)$ for each company is based on the data of the macroeconomic condition of the province where the company operates (registration location).

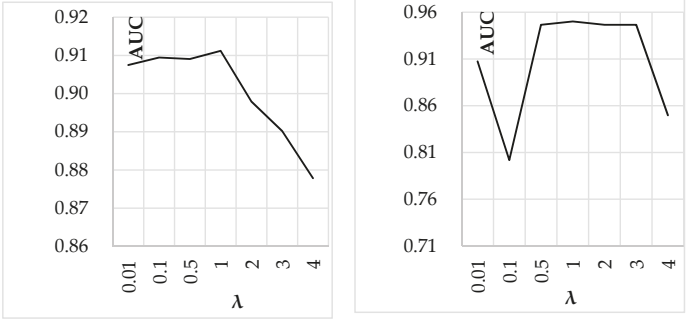
5. Empirical Results and Discussion

In this chapter, we establish a financial earning prediction system for Chinese listed manufacturing companies by using two groups of lasso-generalized distributed lag models, i.e., a logistic model and an SVM model including 3~5-period-lagged explanatory variables, and implement financial distress

prediction and feature selection simultaneously. For the selected sample set, the sample data from 2007 to 2016 were used as the training sample and the sample from 2017 as the test sample. The tuning parameter was identified from cross-validation in the training set, and the performance of the chosen method was evaluated on the testing set by the area under the receiver operating characteristics curve (AUC), G-mean and Kolmogorov–Smirnov (KS) statistics.

5.1. Preparatory Work

It is necessary to choose a suitable value for the tuning parameter λ that controls the trade-off of the bias and variance. As mentioned before, 10-fold cross-validation is used on the training dataset in order to obtain the optimal tuning parameter, λ. First, we compare prediction performance of the lasso–logistic-distributed lag model (5) including only 43 firm-level financial indicators (the accounting-only model) when the turning parameter λ changes. The results show that the mean AUCs of validation data are 0.9075, 0.9095, 0.9091, 0.9112, 0.8979, 0.8902 and 0.8779, respectively, corresponding to λ = 0.01, 0.1, 0.5, 1, 2, 3, 4. Second, we compare the prediction performance of the logistic-distributed lag model (4) incorporating lasso penalty with 43 firm-level financial indicators and 4 macro-economic factors (the model of accounting plus macroeconomic variables). The results show that the mean AUCs of validation data are 0.9074, 0.8018, 0.9466, 0.9502, 0.9466, 0.9466 and 0.8498, respectively, corresponding to λ = 0.01, 0.1, 0.5, 1, 2, 3, 4. Panel (a) and (b) in Figure 1 also show the average predictive accuracy of cross-validation that results from using seven different values of the tuning parameter λ in the accounting model and the model of accounting plus macroeconomic variables.



(**a**) (**b**)

Figure 1. (**a**,**b**) are the Cross-validation performances that result from applying lasso–logistic-distributed lag regression to the listed manufacturing companies' data with various values of λ.

Generally speaking, the two kinds of models yield the best performance when λ = 1. Therefore, in the following, we fit and evaluate the lasso–logistic-distributed lag models by using the tuning parameter of 1.

5.2. Analyses of Results

This study develops a group of ex-ante models for estimating financial distress likelihood in the time window of t to test the contribution of financial ratios and macroeconomic indicators in the consecutive time windows of $t-3$, $t-4$ and $t-5$. In the followings, Table 3 presents the results from lasso–logistic-distributed lag (LLDL) regressions of the financial distress indicator on the predictor variables and Table 4 presents the results from the lasso–SVM-distributed lag model. Furthermore, we compare predictive performance of the existing widely used ex-ante models, including neural networks (NN), decision trees (DT), SVM, and logistic models estimated in a time period from $t-3$ to $t-5$ with our models. The comparative results are shown in Table 5, Table 6 as well as Figure 2.

Table 3. The indicator selection and the estimates for lasso–logistic-distributed lag models.

Selected Indicator	Model 1 (Financial Ratios Only)			Model 2 (Financial Plus Macroeconomic Factor)		
	$t-3$	$t-4$	$t-5$ [1]	$t-3$	$t-4$	$t-5$
1 Total liabilities/total assets	× [2]	×	3.1671 (0.07) ***,[3]	×	×	3.5519 (0.07) ***
2 Current liabilities/total assets	×	−3.356 (0.27) ***	×	×	−1.2292 (0.26) ***	×
3 Sales revenue/average current assets	×	−0.5988 (0.19).	×	×	−3.887 (0.19) ***	×
4 Sales revenue/average total assets	−0.4367 (0.14) ***	−5.7393 (0.24) ***	−1.8312 (0.14) **	−0.5428 (0.14) **	−0.3907 (0.23) **	−3.5193 (0.14) ***
5 Sales cost/sales revenue	5.1892 (0.08) **	×	×	3.9211 (17.57)	×	×
6 Impairment losses/sales profit	−0.4496 (0.08) ***	×	×	−0.5777 (0.08) ***	×	×
7 Sales cost/average net inventory	−1.3265 (0.12) ***	×	×	−1.1143 (0.12) *	×	×
8 Net profit/average total assets	×	−1.1919 (0.14)	×	×	−3.3509 (0.14) **	×
9 Shareholders' equity/net profit	5.4466 (0.17) ***	×	×	4.0804 (0.18) ***	×	×
10 (Sales revenue-sales cost)/sales revenue (net income/revenue)	×	×	×	−1.2209 (17.7)	×	×
11 Net profit/average fixed assets	1.3912 (0.31) **	×	×	×	×	×
12 Net profit/total profit	×	0.2856 (0.14) ***	0.0422 (0.09) *	×	0.0371 (0.14) ***	×
13 Net cash flow from operating activities/total assets	−4.8561 (0.11) ***	−2.6798 (0.11)**	−1.0999 (0.12)	−3.005 (0.1) *	−2.7581 (0.1) **	−0.0304 (0.12)
14 Fixed assets/total assets	1.6395 (0.1)	0.9142 (0.11) **	×	1.5972 (0.09)	×	0.5416 (0.09)

Table 3. *Cont.*

Selected Indicator	Model 1 (Financial Ratios Only)			Model 2 (Financial Plus Macroeconomic Factor)		
	$t-3$	$t-4$	$t-5$ [1]	$t-3$	$t-4$	$t-5$
15 Shareholders' equity/fixed assets	×	1.0914 (0.09) ***	×	×	0.0472 (0.09) **	×
16 Current liabilities/total liabilities	2.2516 (0.06) *	×	2.2987 (0.07) ***	3.1472 (0.06) ***	×	2.1535 (0.07) ***
17 Current assets/total assets	−1.5197 (0.11)	×	×	−3.081 (0.11) *	×	×
18 Long−term liabilities/total liabilities	×	1.6 (0.07) *	×	×	1.8855 (0.06) **	×
19 Main business profit/net income from main business	−3.3814 (0.1) ***	−1.0212 (0.1) ***	5.2777 (0.1) ***	−4.0263 (0.1) ***	−0.9392 (0.1) ***	5.876 (0.1) ***
20 Net profit attributable to shareholders of the parent company/net profit	−3.5409 (0.13) ***	×	×	−2.159 (0.13) ***	×	×
21 Operating capital/total assets	−2.1682 (0.16) ***	×	×	−0.328 (0.16) *	×	×
22 Main sales revenue of this year/main sales revenue of last year	×	×	2.9534 (0.1) **	×	×	2.5545 (0.11) **
23 Net assets/number of ordinary shares at the end of year	×	−6.255 (0.07) ***	×	×	−5.8881 (0.07) ***	×
24 Real Consumer Price Index (CPI) growth (%)	×	×	×	×	−0.2536 (0.06)	0.7531 (0.05)
25 Real GDP growth (%)	×	×	×	−2.4867 (0.09) ***	−1.6404 (0.11)	−0.9319 (0.1)
26 Consumption level growth (%)	×	×	×	−0.9931 (0.08) **	−1.8625 (0.06)	×
27 Unemployment rate (%)	×	×	×	2.7262 (0.07) ***	×	×

[1]: "t − 3, t − 4 and t − 5" represent the estimates for the coefficient vectors of financial and macroeconomic indicators with lag length of 3–5 in the lasso–logistic-distributed lag model, respectively, when λ = 1. [2]: "×" in the table means that the corresponding factor cannot be selected. [3]: The values in brackets are standard error for the estimated coefficients. "*", "**" and "***" indicate that the corresponding variable being significant is accepted at significance levels of 0.1, 0.05 and 0.01, respectively.

Table 4. The indicator selection and the estimates for the lasso–SVM-distributed lag models.

Selected Indicator	Model 1			Model 2		
	$t-3$	$t-4$	$t-5$	$t-3$	$t-4$	$t-5$
1 Total liabilities/total assets	32.1469	7.7613	×	10.7109	7.8039	×
2 Current assets/current liabilities	×	13.7838	−23.2184	×	27.1895	−14.1113
3 Current liabilities/total assets	×	−28.6108	−25.7518	×	−11.9697	×
4 Net cash flow from operating and investing activities/total liabilities	−11.7710	−9.2197	−4.7334	−10.8928	−4.2075	−4.0695
5 Sales revenue/average current assets	−13.2180	−5.5190	−2.3310	×	−3.5302	−2.3478
6 Impairment losses/sales profit	−3.8743	−0.8378	−0.1417	−5.0528	×	−2.1980
7 Sales cost/average net inventory	−4.2927	×	×	−8.5432	×	×
8 Sales revenue/average fixed assets	7.5395	4.0548	4.3177	×	×	×
9 (Sales revenue–sales cost)/sales revenue	×	−4.1270	5.6701	−1.3797	−4.4148	×
10 Net profit attributable to shareholders of the parent company/sales revenue	3.8270	×	×	6.1038	×	×
11 Net cash flow from operating activities/sales revenue	×	−15.5682	×	×	−6.9168	×
12 Net profit/total profit	−1.6296	5.6995	1.6336	−2.4791	2.4036	−0.4364
13 Net cash flow from operating activities/total assets at the end of the period	−12.3631	−19.5928	−4.4420	−2.1426	×	−0.1701
14 Fixed assets/total assets	0.2474	3.0676	8.7795	5.4054	×	1.7320
15 Current liabilities/total liabilities	×	×	9.6303	5.1482	4.3620	10.0672
16 Current assets/total assets	−12.3003	−6.7341	−0.4332	−9.4098	×	−0.5753
17 Long-term liabilities/total liabilities	−5.7781	0.8473	2.0308	×	6.4801	4.2084
18 Main business profit/net income from main business	−7.3785	−7.8631	−9.5525	−2.7997	−4.3300	−4.3107
19 Net profit attributable to shareholders of the parent company/net profit	−10.7914	−6.3596	9.1739	−5.9123	1.7203	1.9460
20 Operating capital/total assets	×	19.8833	×	×	×	8.6408
21 Retained earnings/total assets	×	×	30.8895	×	×	0.9384
22 Main sales revenue of this year/main sales revenue of last year	×	×	25.2376	0.7510	0.0517	13.5974

Table 4. *Cont.*

Selected Indicator	Model 1			Model 2		
	$t-3$	$t-4$	$t-5$	$t-3$	$t-4$	$t-5$
23 Net profit of this year/net profit of last year	−16.5430	17.9966	−7.8113	×	5.9123	−6.9035
24 Net increase in cash and cash equivalents at the end of year/number of ordinary shares	14.4968	0.2728	11.3146	7.6436	×	0.2771
25 Real CPI growth (%)	×	×	×	−2.4001	−0.7880	2.6728
26 Real GDP growth (%)	×	×	×	−1.0391	−4.5598	×
27 Consumption level growth (%)	×	×	×	−1.0196	−1.1848	−2.8019
28 Unemployment rate (%)	×	×	×	16.3215	×	14.3059

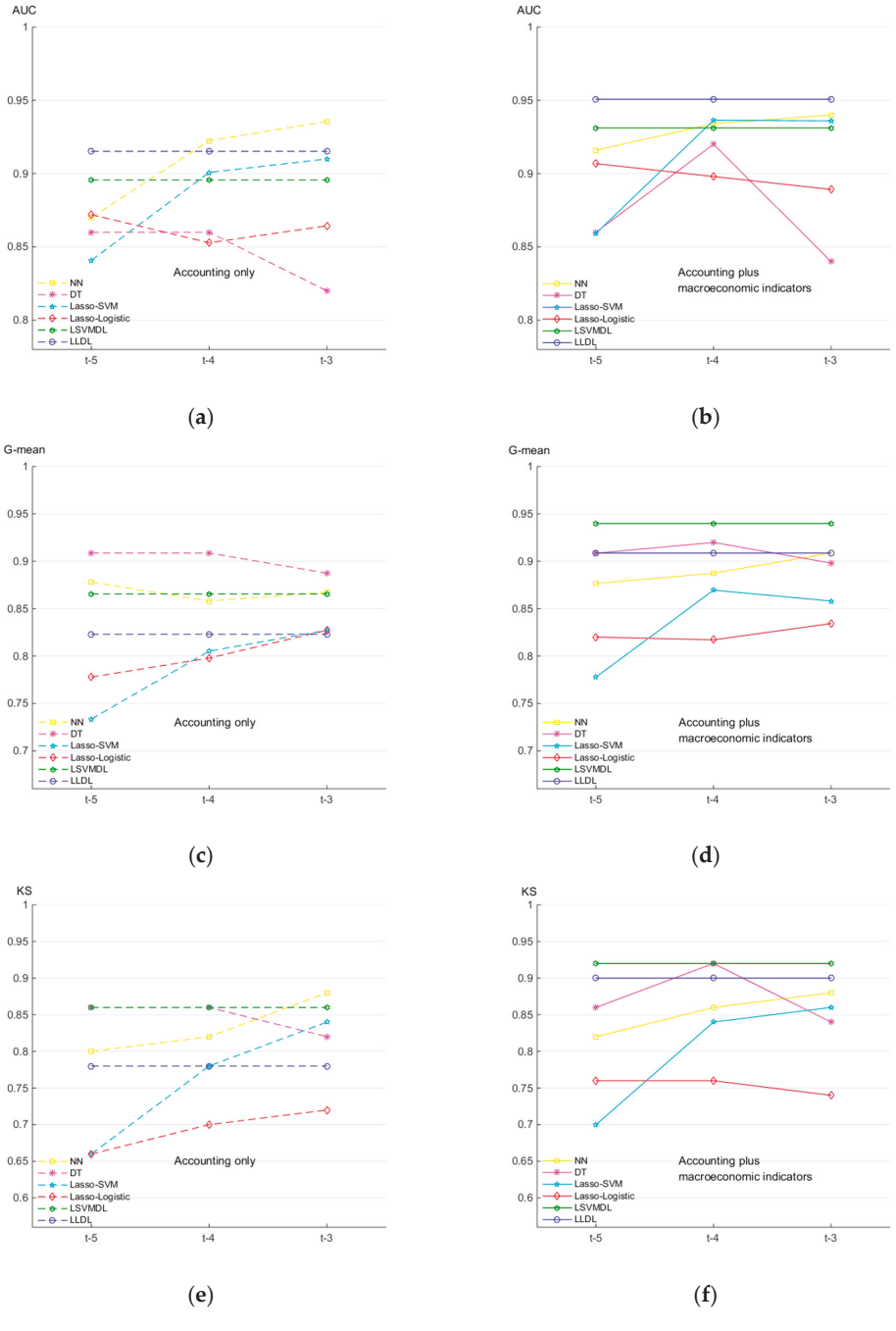


Figure 2. Predictive performance of NN, DT, lasso–SVM, and lasso–logistic models on three different time window datasets, respectively, and our models on three consecutive time window datasets, evaluated by AUC for (**a**) and (**b**), G-Mean for (**c**) and (**d**), and KS for (**e**) and (**f**).

Table 5. Prediction results of the neural network (NN), decision tree (DT), lasso–SVM and lasso–logistic in the single year time window versus the lasso–SVM-distributed lag (LSVMDL) and lasso–logistic-distributed lag (LLDL) models (financial ratios only).

	NN	DT	Lasso–SVM	Lasso–Logistic	LSVMDL	LLDL
Panel A: prediction performance of the existing models in time period $t-3$						
AUC	0.9356	0.8200	0.9100	0.8644	0.8956	0.9152
G-mean	0.8673	0.8874	0.8272	0.8272	0.8655	0.8230
KS	0.8800	0.8200	0.8400	0.7200	0.8600	0.7800
Panel B: prediction performance of the existing models in time period $t-4$						
AUC	0.9224	0.8600	0.9008	0.8528	0.8956	0.9152
G-mean	0.8580	0.9087	0.8052	0.7979	0.8655	0.8230
KS	0.8200	0.8600	0.7800	0.7000	0.8600	0.7800
Panel C: prediction performance of the existing models in time period $t-5$						
AUC	0.8700	0.8600	0.8408	0.8720	0.8956	0.9152
G-mean	0.8780	0.9087	0.7336	0.7778	0.8655	0.8230
KS	0.8000	0.8600	0.6600	0.6600	0.8600	0.7800

Table 6. Prediction results of NN, DT, lasso–SVM and lasso–logistic models in the single year time window versus the lasso–SVM-distributed lag (LSVMDL) model and the lasso–logistic-distributed lag (LLDL) model (financial ratios plus macroeconomic indicators).

	NN	DT	Lasso–SVM	Lasso–Logistic	LSVMDL	LLDL
Panel A: prediction performance of the existing models in time period $t-3$						
AUC	0.9400	0.8400	0.9360	0.8892	0.9312	0.9508
G-mean	0.9087	0.8981	0.8580	0.8343	0.9398	0.9087
KS	0.8900	0.8400	0.8600	0.7400	0.9200	0.9000
Panel B: prediction performance of the existing models in time period $t-4$						
AUC	0.9340	0.9200	0.9364	0.8980	0.9312	0.9508
G-mean	0.8874	0.9198	0.8695	0.8171	0.9398	0.9087
KS	0.8600	0.9200	0.8400	0.7600	0.9200	0.9000
Panel C: prediction performance of the existing models in time period $t-5$						
AUC	0.9160	0.8600	0.8592	0.9068	0.9312	0.9508
G-mean	0.8765	0.9085	0.7778	0.8200	0.9398	0.9087
KS	0.8200	0.8600	0.7000	0.7600	0.9200	0.9000

5.2.1. The Results of the Accounting-Only Model and Analyses

In Table 3, Model 1 represents the "accounting-only" lasso–logistic-distributed lag (LLDL) regression model including the 43 financial statement ratios in 3 adjacent years; the results of financial indicator selection and the estimations for the coefficients are listed in the first three columns. By using Algorithm 1, 23 indicators are in total chosen from the original indicator set. More specifically, two indicators, i.e., indicator number 1 and 2, are selected from the solvency category, five indicators (number 3 to 7) are selected from the operational capability category; six indicators (8-13) from operational capability, eight indicators (13–21) from profitability and two indicators (21-23) from structural soundness and business development and capital expansion capacity. It also can be found that nine financial indicators, namely, sales revenue/average total assets(1), impairment losses/sales profit(2), sales cost/average net inventory(3), shareholders' equity/net profit(4), net profit/total profit(5), net cash flow from operating activities/total assets(6), main business profit/net income from main business(7), net profit attributable to shareholders of the parent company/net profit(8) and operating capital/total assets(9), not used in the paper of [7] have quite significant influence on the future financial distress risk.

The potentially helpful ratios, such as the leverage ratio (total liabilities/total assets), shareholders' equity/net profit (ROE), net profit/average total assets (ROA), current liabilities/total liabilities etc., have significant effects on the occurrence of financial distress of Chinese listed manufacturing companies. For example, as shown in Table 3, the indicator of the leverage ratio in year $t-3$—a very early time period—is selected as a significant predictor, and the estimated value for the coefficient is 3.1671. This implies that the increase in value of the Leverage ratio in the fifth previous *ST* year increases the financial risk of the listed manufacturing companies. The indicator of ROA for year $t-4$ is selected, and the estimated value of the coefficient of the indicator is −1.1919, which implies the probability of falling into financial distress for a company will decrease with the company's ROA value, i.e., net profit/average total assets increasing.

Besides, the results in Table 3 also show that all changes in the indicator of sales revenue/average total assets for three consecutive time periods have significant effects on the future financial distress risk. It can be found that different weights are assigned to the variables of sales revenue/average total assets with different time lags, and the coefficient estimates for the indicator in the time windows of $t-3$, $t-4$ and $t-5$ are −0.4367, −5.7393 and −1.8312, respectively. This implies that increases in sale revenue in different time windows have positive and significant (but different) effects on the future financial status of a listed company. The result for the indicator of "net cash flow from operating activities/total assets" presented in row 13 and the first 3 columns of Table 3 illustrate that changes in this indicator in different time windows have different effects on the future occurrence of financial distress at a significance level and magnitude of influence. The estimated coefficients for the variable measured in the previous time windows, $t-3$, $t-4$ and $t-5$, are −4.8561, −2.6798 and −1.0999, at the significance level of 0.01, 0.05 and (>) 0.1, respectively. This indicates that (1) the higher the ratio of net cash flow from operating activities to total assets for a listed manufacturing company, the lower the likelihood of the firm's financial distress; (2) the changes in net cash flow from operating activities/total assets in the time windows $t-3$ and $t-4$ have significant influence on the risk of financial distress, and the magnitude of influence increases as the length of lag time decreases; (3) the influence of this indicator declines over time and change in this indicator in the 5 years before the observation of the financial distress event has no significant effect on financial risk when compared with relatively recent changes.

5.2.2. The Results and Analyses of the Model of Accounting Plus Macroeconomic Variables

In Table 3, Model 2 represents the "accounting plus macroeconomic factor" model, including the original 43 financial ratios and 4 macroeconomic indicators in 3 adjacent years, and the results of indicator selection and the coefficient estimates are listed in the last three columns. It can be found that for Model 2, the same group of financial variables is selected and included in the final model. Time lags of the selected financial variables and the signs (but not magnitudes) of the estimated coefficients for the variables are almost consistent for Model 1 and 2.

In addition to the accounting ratios, three macroeconomic factors are selected as significant predictors and included in the final model: GDP growth, consumption level growth and unemployment rate in time window of $t-3$. The estimate for the coefficients of the selected GDP growth and unemployment rate are −2.4867 and 2.7262, respectively, which means that high GDP growth should decrease the financial distress risk, but high unemployment will deteriorate the financial condition of a listed manufacturing company. These results are consistent, which was expected. The estimate for the coefficient of consumption level growth is −0.9931, which implies that the high consumption level growth should decrease the possibility of financial deterioration of a listed company. Finally, it cannot be found that Consumer Price Index (CPI) growth has a significant influence on the financial distress risk.

The 4 year-lagged and 5 year-lagged GDP growth and 4 year-lagged consumption level growth are also selected and included in the final model but not as very significant predictors, which implies the following: (1) the changes in macroeconomic conditions have a continuous influence on the financial

distress risk; (2) however, the effect of the macroeconomic condition' changes on the financial distress risk declines with the length of the lag window increasing.

5.2.3. The Results of Lasso–SVM-Distributed Lag (LSVMDL) Models and Analyses

We introduce 3-period lags of financial indicators presented in Table 1, i.e., TL/TA_{t-3}, TL/TA_{t-4} and TL/TA_{t-5}, CA/CL_{t-3}, CA/CL_{t-4} and CA/CL_{t-5} . . . , $NICCE/NOS_{t-3}$, $NICCE/NOS_{t-4}$ and $NICCE/NOS_{t-5}$ into the model (16) and implement the indicator selection and the coefficient estimates by using Algorithm 3. The corresponding results are presented in first three columns of Table 4. Then, we introduce 3-period lags of financial and macroeconomic indicators presented in Tables 1 and 2 into the model (16) and the coefficient estimate of selected indicators are presented in the last three columns of Table 4.

Twenty-four financial indicators are selected and included in the final SVM-distributed lag model, denoted as Model 1 in Table 4; 17 indicators among them are also included in the final logistic-distributed lag model. For convenience of comparison, the 17 indicators, such as total liabilities/total assets, current liabilities/total assets and sales revenue/average current assets etc., are italicized and shown in the "selected indicator" column of Table 4.

According to the relation between response variables and predictors in the SVM model, as mentioned before, the increase (decrease) in the factors should increase (decrease) the financial distress risk when the coefficient estimates are positive. Therefore, let us take the estimated results in the first three rows and columns as an example: (1) the increase in the total liabilities to total assets ratio should increase the financial distress risk of a listed manufacturing company; (2) the increase in current liabilities to total assets ratio should decrease the financial distress risk; (3) the changes in the indicators in the period closer to the time of obtaining ST have a more significant effect on the likelihood of financial distress in terms of magnitudes of estimates of the coefficients.

Four macroeconomic factors, in addition to 24 financial indicators, are selected and included in the final SVM-distributed lag model, denoted as Model 2 in Table 4. The results show that (1) the effects of the selected financial ratios on the response, i.e., the financial status of a company, is consistent with the results in the SVM-distributed lag model including only financial ratios, i.e., Model 1, in terms of time lags of the selected financial variables and the signs of the estimated coefficients for the explanatory variables; (2) high GDP growth and high consumption level growth should decrease the financial distress risk, but high unemployment will deteriorate the financial condition of a listed manufacturing company.

From Table 4, it can be found that different indicators have different influence on the financial status of a company. The effects of some indicators on financial distress risk increase with the decrease in the time lag, e.g., total liabilities to total assets ratio, current liabilities/total assets and net cash flow from operating and investing activities/total liabilities etc., while the effects of some other indicators should decrease with the decrease in the time lag, e.g., fixed assets/total assets, GDP growth and consumption level growth etc. However, for some indicators, the effects of different time windows on financial status change. For example, the coefficients for current assets/current liabilities (current ratio) in Model 1 are 13.7838 for time window $t-4$ and −23.2184 for time window $t-5$, which implies that a high current ratio in time window $t-5$ should decrease the financial distress risk; this, however, would be not the case in time $t-4$. Similar case can be found for CPI growth in Model 2. Thus, SVM-distributed lag models may not interpret well; therefore, it would be inferior to the logistic-distributed lag models in terms of in terms of interpretability.

5.2.4. Comparison with Other Models

For the purpose of comparison, the prediction performances of the ex-ante models for the estimation of financial distress likelihood developed by the existing studies are shown in Tables 5 and 6. The existing widely used ex-ante models include the neural network (NN), decision tree (DT), SVM, and logistic models estimated in different time periods of $t-3$, $t-4$, and $t-5$, called $t-3$ models,

$t-4$ models and $t-5$ models. The construction of these three groups of models is similar to [7]. Let us take the construction of $t-5$ model as example. For 10 financially distressed companies that received ST in 2012 and the selected 10 healthy companies until 2012 as a control group, their financial and macroeconomic data in 2007 (5 years before 2012) were collected. For 9 financially distressed companies that received ST in 2013 and the selected 9 healthy companies, their financial and macroeconomic data in 2008 (5 years before 2013) were collected. Similarly, for 17, 24, 26 financial distressed companies that receive the ST label respectively in 2014, 2015 and 2016 and the non-financial companies randomly selected at a 1:1 ratio in each year for matching with the ST companies, their data in 2009 (5 years before 2014), 2010 (5 years before 2015) and in 2011 (5 years before 2016) were collected. By using the labels of 172 companies and the data that were obtained 5 years prior to the year when the companies received the ST label, we construct $t-5$ financial distress forecast models combined with a neural network (NN), decision tree (DT), SVM, and logistic regression. Similarly, $t-3$ models and $t-4$ models can be built. The data of financially distressed companies that received ST in 2017 and non-financial distressed companies were used to evaluate these models' predicting performance.

As mentioned in the beginning of this section, three measures of prediction performances are reported in these two tables, namely, AUC, G-mean, and Kolmogorov–Smirnov statistics. In the above scenarios based on different time periods as well as division of the whole dataset, we compare respectively the predicting performance of those one-time window models ($t-3$ models, $t-4$ models and $t-5$ models) including financial ratios only and financial ratios plus macroeconomic factors with our lasso–SVM-distributed lag (LSVMDL) model and lasso–logistic-distributed lag (LLDL). The prediction results are presented in Table 5 for the case of "financial ratios only" and Table 6 for the case of "financial ratio plus macroeconomic factors".

In Table 5, panel A presents the predictive performances of NN, DT, lasso–SVM and lasso–logistic models including the original 43 financial ratios shown in Table 1 in the period $t-3$ as predictors of financial distress status in period t, while the results in the last two columns are the performances of the two groups of distributed lag financial distress predicting models including the same original 43 financial ratios but in periods $t-3$, $t-4$ and $t-5$, i.e., our models. Panel B and C of Table 5 present the prediction performance of the models used for comparison purposes estimated in $t-4$ and $t-5$, respectively. The results for our models retain the same values because these models include simultaneously the 3-year-, 4-year- and 5-year-lagged financial ratios.

The only difference between Tables 5 and 6 is that all models, in addition to the 43 original accounting rations, incorporate 4 macroeconomic indicators in different time windows. For example, for time window $t-3$, the NN, DT, lasso–SVM and lasso–logistic models include 3-year-lagged macroeconomic indicators shown in Table 2 in addition to the financial statement ratios shown in Table 1. The cases of time windows $t-4$ and $t-5$ are similar for these models. As for the LSVMDL and LLDL models, i.e., our models, they include 3-periods-lagged macroeconomic indicators in the time windows $t-3$, $t-4$ and $t-5$ in addition to the accounting ratios.

From Table 5, the prediction accuracy of NN or DT is highest in the time windows $t-3$ and $t-4$; our models outperform the others in time window $t-5$ for predicting accuracy. Generally speaking, the accuracy for time period $t-3$ is relatively higher than the other two time periods for the NN, lasso–SVM and lasso–logistic models. Furthermore, the prediction results based on time period $t-3$ are the most precise for NN when compared with other models in a single time period and even our models, which implies that the selected financial ratios in the period closer to the time of obtaining ST may contain more useful information for the prediction of financial distress, and may be applicable to NN. The AUC of 91.52% of the lasso–logistic-distributed lag model (LLDL) ranked second, close to the accuracy of 93.56% obtained by using NN. Therefore, the LLDL model should be competitive in terms of interpretability and accuracy in the case of "accounting ratio only".

From Table 6, the prediction accuracy of all used models is higher than the results in Table 5. For example, the AUC, G-mean and KS of the NN model in time window $t-3$ increases from 93.56%, 86.73% and 88.00% in Table 5 to 94.00%, 90.87% and 89.00% in Table 6, respectively. The changing

tendency of the prediction accuracy is retained for the other models, including macroeconomic indicators in addition to the accounting ratios. All results in Table 6 indicate that the introduction of the macroeconomic variables can improve predictive performance of all used models for the purpose of comparison; the changes in macroeconomic conditions do affect the likelihood of financial distress risk. On the other hand, the LLDL model performs best with the AUC of over 95% when compared with the best NN (in time period $t-3$, 94%), the best DT (in time period $t-4$, 92.24%), the best lasso–SVM (in time period $t-4$, 93.64%), the best lasso–logistic (in time period $t-5$, 90.68%) and LSVMDL (93.12%). The LSVMDL model is the best performing model in terms of G-mean and KS statistics.

Figure 2 also shows the comparative results of the accuracy of the six models. The predictive performances of all the models including accounting ratios only, indicated by the dotted lines (a), (c) and (e) in Figure 2, are worse than the models including macroeconomic indicators as well as accounting ratios, which are illustrated by the solid lines (b), (d) and (f) in Figure 2. Figures (a) and (b), G-Mean for (c) and (d), and KS for (e) and (f) present AUC, G-mean and KS for all of the examined models, respectively. The models used for comparison, namely, NN, DT, lasso–SVM and lasso–logistic models, were those that yielded the highest accuracy based on the different time window dataset. For example, based on the results of panel (b), AUC of NN (the yellow solid line), DT (the pink solid line), and lasso–logistic (the red one) models are highest in time window $t-3$, $t-4$ and $t-5$, respectively. We cannot conclude that the prediction results based on financial and macroeconomic data of one specific time window, e.g., $t-3$ (see [7]), are the most accurate. However, from the results in (b), (d) and (f), our models, the LLDL or LSVMDL model incorporating financial and macroeconomic data in three consecutive time-windows, yielded relatively robust and higher prediction performances.

Put simply, the two groups of generalized distributed lag financial distress predicting models proposed by this paper outperform the other models in each time period, especially when the accounting ratios and macroeconomic factors were introduced into the models. We demonstrated that our models provide an effective way to deal with multiple time period information obtained from changes in accounting and macroeconomic conditions.

5.2.5. Discussion

Logistic regression and multivariate discriminant methods should be the most popular statistical techniques used in financial distress risk prediction modelling for different countries' enterprise, e.g., American enterprises [1] and European enterprises [4,30,31], because of their simplicity, good predictive performance and interpretability. The main statistical approach involved in this study is logistic regression, but rather multivariate discriminant analysis, given that strict assumptions regarding normal distribution of explanatory variables are used in multivariate discriminant analysis. The results in this study conform that logistic regression models still perform well for predicting Chinese listed enterprises' financial distress risks.

The major contribution to financial distress prediction literature made by this paper is that an optimally distributed lag structure of macroeconomic data in the multi-periods, in addition to financial ratio data, are imposed on the logistic regression model through minimizing loss function, and the heterogenous lagged effects of the factors in the different period are presented. The results unveil that financial indicators, such as total liabilities/total assets, sales revenue/total assets, and net cash flow from operating activities/total assets, tend to have a significant impact over relatively longer periods, e.g., 5 years before the financial crisis of a Chinese listed manufacturing company. This finding is in accordance with the recent research of [30,31] in that the authors claim the process of going bankrupt is not a sudden phenomenon; it may take as long as 5–6 years. In the very recent study of Korol et al. [30], the authors built 10 group models comprising 10 periods: from 1 year to 10 years prior to bankruptcy. The results in [30] indicate that a bankruptcy prediction model such as the fuzzy set model maintained an effectiveness level above 70% until the eighth year prior to bankruptcy. Therefore, our model can be extended through introducing more lagged explanatory variables, e.g., 6- to 8-year-lagged financial

variables, which may bring a better distributive lag structure of explanatory variables and predicting ability of the models.

The findings of this study allow managers and corporate analysts to prevent financial crisis of a company by monitoring early changes in a few sensitive financial indicators and taking actions, such as optimizing the corporate's asset structure, increasing cash flow and sales revenue, etc. They are also helpful for investors to make investment decision by tracking continuous changes in accounting conditions of a company of interest and predicting its risk of financial distress.

Another major contribution of this study is the confirmation of the importance of macroeconomic variables in predicting the financial distress of a Chinese manufacturing company, although scholars still argue about the significance of macro variables. For example, Kacer et al. [66] did not recommend the use of macro variables in the financial distress prediction for Slovak Enterprises, while Hernandez Tinoco et al. [4] confirmed the utilization of macro variables in the financial distress prediction for listed enterprises of the United Kingdom. The results in Section 5.2.4 of this study show that the prediction performance of all models (including both the models used for comparison and our own models) was increased when the macro variables were included in each model. The findings of this study allow regulators to tighten the supervision of Chinese listed companies when macroeconomic conditions change, especially in an economic downturn.

One of the main limitations of this study is that we limited the research only to the listed manufacturing companies. Both Korol et al. [28] and Kovacova [30] emphasized that the type of industry affects the risk of deterioration in the financial situation of companies. More specifically, distinguished by factors such as intensity of competition, life cycle of products, demand, changes in consumer preferences, technological change, reducing entry barriers into the industry and susceptibility of the industry to business cycles, different industries are at different levels of risk [28]. The manufacturing sector, which includes the metal, mining, automotive, aerospace and housing industries, is highly susceptible to demands, technological changes and macroeconomic conditions, thus making it at a high level of risk, while agriculture may be at a relatively low risk level. The risk parameter assigned to the service sector, including restaurants, tourism, transport and entertainment etc., has seen significant changes following the outbreak of the Coronavirus. Therefore, applicability and critique to our models for predicting financial distress risk of the companies operating in other industry and even other countries need to be further detected.

6. Conclusions

In this paper, we propose a new framework of a financial early warning system through introducing a distributed lag structure to be widely used in financial distress prediction models such as the logistic regression and SVM models. Our models are competitive when compared with the conventional financial distress forecast models, which incorporates data from only one-period of $t-3$ or $t-4$ or $t-5$, in terms of predictive performance. Furthermore, our models are superior to the conventional one-time window financial distress forecast models, in which macroeconomic indicators of GDP growth, consumption level growth and unemployment rate, in addition to accounting factors, are incorporated. The empirical findings of this study indicate that the changes in macroeconomic conditions do have significant and continuous influence on the financial distress risk of a listed manufacturing company. This paper may provide an approach of examining the impacts of macroeconomic information from multiple periods and improving the predictive performance of financial distress models.

We implement feature selection to remove redundant factors from the original list of 43 potentially helpful ratios and their lags by introducing lasso penalty into the financial distress forecast logistic models with lags and SVM models with lags. Furthermore, we provide an ADMM algorithm framework that yields the global optimum for convex and the non-smooth optimization problem to obtain the optimal estimation for the coefficients of these financial distress forecast models with financial and macroeconomic factors and their lags. Results from the empirical study show that not only widely used financial indicators (calculated from accounting data), such as leverage ratio, ROE, ROA,

and current liabilities/total liabilities, have significant influence on the financial distress risk of a listed manufacturing company, but also the indicators that are rarely seen in the existing literature, such as net profit attributable to shareholders of the parent company and net cash flow from operating activities/total assets, may play very important roles in financial distress prediction. The closer to the time of financial crisis, the more net profit attributable to shareholders of the parent company and net cash flow from operating activities may considerably decrease the financial distress risk. These research findings may provide more evidence for company managers and investors in terms of corporate governance or risk control.

The main limitation of this research is that we limited the research only to listed manufacturing companies. Sensitivity of financial distress models and suitability of both financial and macroeconomic variables to the enterprises that operate in other industries, e.g., service companies, need to be further discussed. On the other hand, given that the utilization of financial and macroeconomic variables in predicting the risk of financial distress of Chinese listed manufacturing companies is confirmed, we intend to continue the research toward the use of interaction terms of financial and macroeconomic variables in the context of the multiple period. Furthermore, the heterogeneous effect of changes in macroeconomic conditions on the financial distress risk of a company under different financial conditions can be discovered.

Author Contributions: We attest that all authors contributed significantly to the creation of this manuscript. The conceptualization and the methodology were formulated by D.Y., data curation was completed by G.C., and the formal analysis was finished by K.K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China under grant numbers 71731003, 71301017 and by the Fundamental Research Funds for the Central Universities under grant numbers DUT19LK50 and QYWKC2018015. The authors wish to thank the organizations mentioned above.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Altman, E.I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **1968**, *23*, 589–609. [CrossRef]
2. Lau, A.H.L. A five-state financial distress prediction model. *J. Account. Res.* **1987**, *25*, 127–138. [CrossRef]
3. Jones, S.; Hensher, D.A. Predicting firm financial distress: A mixed logit model. *Account. Rev.* **2004**, *79*, 1011–1038. [CrossRef]
4. Hernandez Tinoco, M.; Holmes, P.; Wilson, N. Polytomous response financial distress models: The role of accounting, market and macroeconomic variables. *Int. Rev. Financ. Anal.* **2018**, *59*, 276–289. [CrossRef]
5. Zmijewski, M.E. Methodological issues related to the estimation of financial distress prediction models. *J. Account. Res.* **1984**, *22*, 59–82. [CrossRef]
6. Ross, S.; Westerfield, R.; Jaffe, J. *Corporate Finance*; McGraw-Hill Irwin: New York, NY, USA, 2000.
7. Geng, R.; Bose, I.; Chen, X. Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *Eur. J. Oper. Res.* **2015**, *241*, 236–247. [CrossRef]
8. Westgaard, S.; Van Der Wijst, N. Default probabilities in a corporate bank portfolio: A logistic model approach. *Eur. J. Oper. Res.* **2001**, *135*, 338–349. [CrossRef]
9. Balcaen, S.; Ooghe, H. 35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems. *Br. Account. Rev.* **2006**, *38*, 63–93. [CrossRef]
10. Martin, D. Early warnings of bank failure: A logit regression approach. *J. Bank. Financ.* **1977**, *1*, 249–276. [CrossRef]
11. Liang, D.; Tsai, C.F.; Wu, H.T. The effect of feature selection on financial distress prediction. *Knowl. Based Syst.* **2015**, *73*, 289–297. [CrossRef]
12. Frydman, H.; Altman, E.I.; Kao, D.L. Introducing recursive partitioning for financial classification: The case of financial distress. *J. Financ.* **1985**, *40*, 269–291. [CrossRef]
13. Leshno, M.; Spector, Y. Neural network prediction analysis: The bankruptcy case. *Neurocomputing* **1996**, *10*, 125–147. [CrossRef]
14. Shin, K.S.; Lee, T.S.; Kim, H.J. An application of support vector machines in bankruptcy prediction model. *Expert Syst. Appl.* **2005**, *28*, 127–135. [CrossRef]

15. Sun, J.; Li, H. Data mining method for listed companies' financial distress prediction. *Knowl. Based Syst.* **2008**, *21*, 1–5. [CrossRef]
16. Jiang, Y.; Jones, S. Corporate distress prediction in China: A machine learning approach. *Account. Financ.* **2018**, *58*, 1063–1109. [CrossRef]
17. Purnanandam, A. Financial distress and corporate risk management: Theory and evidence. *J. Financ. Econ.* **2008**, *87*, 706–739. [CrossRef]
18. Almamy, J.; Aston, J.; Ngwa, L.N. An evaluation of Altman's Z-score using cash flow ratio to predict corporate failure amid the recent financial crisis: Evidence from the UK. *J. Corp. Financ.* **2016**, *36*, 278–285. [CrossRef]
19. Liang, D.; Lu, C.C.; Tsai, C.F.; Shih, G.A. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *Eur. J. Oper. Res.* **2016**, *252*, 561–572. [CrossRef]
20. Scalzer, R.S.; Rodrigues, A.; Macedo, M.Á.S.; Wanke, P. Financial distress in electricity distributors from the perspective of Brazilian regulation. *Energy Policy* **2019**, *125*, 250–259. [CrossRef]
21. Altman, I.E.; Haldeman, G.R.; Narayanan, P. ZETA^{TM} analysis A new model to identify bankruptcy risk of corporations. *J. Bank. Financ.* **1977**, *1*, 29–54. [CrossRef]
22. Inekwe, J.N.; Jin, Y.; Valenzuela, M.R. The effects of financial distress: Evidence from US GDP growth. *Econ. Model.* **2018**, *72*, 8–21. [CrossRef]
23. Ohlson, J. Financial ratios and the probabilistic prediction of bankruptcy. *J. Account. Res.* **1980**, *18*, 109–131. [CrossRef]
24. Hillegeist, S.; Keating, E.; Cram, D.; Lundstedt, K. Assessing the probability of bankruptcy. *Rev. Account. Stud.* **2004**, *9*, 5–34. [CrossRef]
25. Teresa, A.J. Accounting measures of corporate liquidity, leverage, and costs of financial distress. *Financ. Manag.* **1993**, *22*, 91–100.
26. Shumway, T. Forecasting bankruptcy more accurately: A simple hazard model. *J. Bus.* **2001**, *74*, 101–124. [CrossRef]
27. Hosaka, T. Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert Syst. Appl.* **2019**, *117*, 287–299. [CrossRef]
28. Korol, T. Dynamic Bankruptcy Prediction Models for European Enterprises. *J. Risk Financ. Manag.* **2019**, *12*, 185. [CrossRef]
29. Gregova, E.; Valaskova, K.; Adamko, P.; Tumpach, M.; Jaros, J. Predicting Financial Distress of Slovak Enterprises: Comparison of Selected Traditional and Learning Algorithms Methods. *Sustainability* **2020**, *12*, 3954. [CrossRef]
30. Kovacova, M.; Kliestik, T.; Valaskova, K.; Durana, P.; Juhaszova, Z. Systematic review of variables applied in bankruptcy prediction models of Visegrad group countries. *Oeconomia Copernic.* **2019**, *10*, 743–772. [CrossRef]
31. Kliestik, T.; Misankova, M.; Valaskova, K.; Svabova, L. Bankruptcy Prevention: New Effort to Reflect on Legal and Social Changes. *Sci. Eng. Ethics* **2018**, *24*, 791–803. [CrossRef]
32. López, J.; Maldonado, S. Profit-based credit scoring based on robust optimization and feature selection. *Inf. Sci.* **2019**, *500*, 190–202. [CrossRef]
33. Maldonado, S.; Pérez, J.; Bravo, C. Cost-based feature selection for Support Vector Machines: An application in credit scoring. *Eur. J. Oper. Res.* **2017**, *261*, 656–665. [CrossRef]
34. Li, J.; Qin, Y.; Yi, D. Feature selection for Support Vector Machine in the study of financial early warning system. *Qual. Reliab. Eng. Int.* **2014**, *30*, 867–877. [CrossRef]
35. Duffie, D.; Saita, L.; Wang, K. Multi-Period Corporate Default Prediction with Stochastic Covariates. *J. Financ. Econ.* **2004**, *83*, 635–665. [CrossRef]
36. Greene, W.H.; Hensher, D.A.; Jones, S. An Error Component Logit Analysis of Corporate Bankruptcy and Insolvency Risk in Australia. *Econ. Rec.* **2007**, *83*, 86–103.
37. Figlewski, S.; Frydman, H.; Liang, W.J. Modeling the effect of macroeconomic factors on corporate default and credit rating transitions. *Int. Rev. Econ. Financ.* **2012**, *21*, 87–105. [CrossRef]
38. Tang, D.Y.; Yan, H. Market conditions, default risk and credit spreads. *J. Bank. Financ.* **2010**, *34*, 743–753. [CrossRef]
39. Chen, C.; Kieschnick, R. Bank credit and corporate working capital management. *J. Corp. Financ.* **2016**, *48*, 579–596. [CrossRef]
40. Jermann, U.; Quadrini, V. Macroeconomic effects of financial shocks. *Am. Econ. Rev.* **2012**, *102*, 238–271. [CrossRef]

41. Carpenter, J.N.; Whitelaw, R.F. The development of China's stock market and stakes for the global economy. *Annu. Rev. Financ. Econ.* **2017**, *9*, 233–257. [CrossRef]
42. Hua, Z.; Wang, Y.; Xu, X.; Xu, X.; Zhang, B.; Liang, L. Predicting corporate financial distress based on integration of support vector machine and logistic regression. *Expert Syst. Appl.* **2007**, *33*, 434–440. [CrossRef]
43. Li, H.; Sun, J. Hybridizing principles of the Electre method with case-based reasoning for data mining: Electre-CBR-I and Electre-CBR-II. *Eur. J. Oper. Res.* **2009**, *197*, 214–224. [CrossRef]
44. Cao, Y. MCELCCh-FDP: Financial distress prediction with classifier ensembles based on firm life cycle and Choquet integral. *Expert. Syst. Appl.* **2012**, *39*, 7041–7049. [CrossRef]
45. Shen, F.; Liu, Y.; Wang, R. A dynamic financial distress forecast model with multiple forecast results under unbalanced data environment. *Knowl. Based Syst.* **2020**, *192*, 1–16. [CrossRef]
46. Gasparrini, A.; Armstrong, B.; Kenward, M.G. Distributed lag non-linear models. *Stat. Med.* **2010**, *29*, 2224–2234. [CrossRef]
47. Gasparrini, A.; Scheipl, B.; Armstrong, B.; Kenward, G.M. Penalized Framework for Distributed Lag Non-Linear Models. *Biometrics* **2017**, *73*, 938–948. [CrossRef] [PubMed]
48. Wilson, A.; Hsu, H.H.L.; Chiu, Y.H.M. Kernel machine and distributed lag models for assessing windows of susceptibility to mixtures of time-varying environmental exposures in children's health studies. *arXiv* **2019**, arXiv:1904.12417.
49. Nelson, C.R.; Schwert, G.W. Estimating the Parameters of a Distributed Lag Model from Cross-Section Data: The Case of Hospital Admissions and Discharges. *J. Am. Stat. Assoc.* **1974**, *69*, 627–633. [CrossRef]
50. Hammoudeh, S.; Sari, R. Financial CDS, stock market and interest rates: Which drives which? *N. Am. J. Econ. Financ.* **2011**, *22*, 257–276. [CrossRef]
51. Lahiani, A.; Hammoudeh, S.; Gupta, R. Linkages between financial sector CDS spreads and macroeconomic influence in a nonlinear setting. *Int. Rev. Econ. Financ.* **2016**, *43*, 443–456. [CrossRef]
52. Almon, S. The distributed lag between capital appropriations and expenditures. *Econometrica* **1965**, *33*, 178–196. [CrossRef]
53. Dominici, F.; Daniels, M.S.L.; Samet, Z.J. Air pollution sand mortality: Estimating regional and national dose—Response relationships. *J. Am. Stat. Assoc.* **2002**, *97*, 100–111. [CrossRef]
54. Wooldridge, J.M. *Econometric Analysis of Cross Section and Panel Data*; MIT Press: Cambridge, MA, USA, 2010.
55. Park, H.; Sakaori, F. Lag weighted lasso for time series model. *Comput. Stat.* **2013**, *28*, 493–504. [CrossRef]
56. Glowinski, R.; Marroco, A. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. *ESAIM Math. Model. Numer.* **1975**, *9*, 41–76. [CrossRef]
57. Dantzig, G.; Wolfe, J. Decomposition principle for linear programs. *Oper. Res.* **1960**, *8*, 101–111. [CrossRef]
58. Hestenes, M.R. Multiplier and gradient methods. *J. Optim. Theory. Appl.* **1969**, *4*, 302–320. [CrossRef]
59. Chambolle, A.; Pock, T. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.* **2011**, *40*, 120–145. [CrossRef]
60. Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found. Trends. Mach. Learn.* **2011**, *3*, 1–122. [CrossRef]
61. Mangasarian, O.L. A finite Newton method for classification. *Optim. Methods Softw.* **2002**, *17*, 913–929. [CrossRef]
62. Shon, T.; Moon, J. A hybrid machine learning approach to network anomaly detection. *Inf. Sci.* **2007**, *177*, 3799–3821. [CrossRef]
63. Liu, D.; Qian, H.; Dai, G.; Zhang, Z. An iterative SVM approach to feature selection and classification in high-dimensional datasets. *Pattern Recognit.* **2013**, *46*, 2531–2537. [CrossRef]
64. Tiwari, R. Intrinsic value estimates and its accuracy: Evidence from Indian manufacturing industry. *Future Bus. J.* **2016**, *2*, 138–151. [CrossRef]
65. Shanker, M.; Hu, M.Y.; Hung, M.S. Effect of data standardization on neural network training. *Omega Int. J. Manag. Sci.* **1996**, *24*, 385–397. [CrossRef]
66. Kacer, M.; Ochotnicky, P.; Alexy, M. The Altman's revised Z'-Score model, non-financial information and macroeconomic variables: Case of Slovak SMEs. *Ekon. Cas.* **2019**, *67*, 335–366.

mathematics

MDPI

Article

Deep Learning Methods for Modeling Bitcoin Price

Prosper Lamothe-Fernández [1], David Alaminos [2,*], Prosper Lamothe-López [3] and Manuel A. Fernández-Gámez [4]

[1] Department of Financing and Commercial Research, UDI of Financing, Calle Francisco Tomás y Valiente, 5, Universidad Autónoma de Madrid, 28049 Madrid, Spain; prosper.lamothe@uam.es
[2] Department of Economic Theory and Economic History, Campus El Ejido s/n, University of Malaga, 29071 Malaga, Spain; mangel@uma.es
[3] Rho Finanzas Partner, Calle de Zorrilla, 21, 28014 Madrid, Spain; pll@rhofinanzas.com
[4] Department of Finance and Accounting, Campus El Ejido s/n, University of Malaga, 29071 Malaga, Spain
* Correspondence: alaminos@uma.es

Received: 25 June 2020; Accepted: 28 July 2020; Published: 30 July 2020

Abstract: A precise prediction of Bitcoin price is an important aspect of digital financial markets because it improves the valuation of an asset belonging to a decentralized control market. Numerous studies have studied the accuracy of models from a set of factors. Hence, previous literature shows how models for the prediction of Bitcoin suffer from poor performance capacity and, therefore, more progress is needed on predictive models, and they do not select the most significant variables. This paper presents a comparison of deep learning methodologies for forecasting Bitcoin price and, therefore, a new prediction model with the ability to estimate accurately. A sample of 29 initial factors was used, which has made possible the application of explanatory factors of different aspects related to the formation of the price of Bitcoin. To the sample under study, different methods have been applied to achieve a robust model, namely, deep recurrent convolutional neural networks, which have shown the importance of transaction costs and difficulty in Bitcoin price, among others. Our results have a great potential impact on the adequacy of asset pricing against the uncertainties derived from digital currencies, providing tools that help to achieve stability in cryptocurrency markets. Our models offer high and stable success results for a future prediction horizon, something useful for asset valuation of cryptocurrencies like Bitcoin.

Keywords: bitcoin; deep learning; deep recurrent convolutional neural networks; forecasting; asset pricing

1. Introduction

Bitcoin is a cryptocurrency built by free software based on peer-to-peer networks as an irreversible private payment platform. Bitcoin lacks a physical form, is not backed by any public body, and therefore any intervention by a government agency or other agent is not necessary to transact [1]. These transactions are made from the blockchain system. Blockchain is an open accounting book, which records transactions between two parties efficiently, leaving such a mark permanently and impossible to erase, making this tool a decentralized validation protocol that is difficult to manipulate, and with low risk of fraud. The blockchain system is not subject to any individual entity [2].

For Bitcoin, the concept originated from the concept of cryptocurrency, or virtual currency [3]. Cryptocurrencies are a monetary medium that is not affected by public regulation, nor is it subject to a regulatory body. It only affects the activity and rules developed by the developers. Cryptocurrencies are virtual currencies that can be created and stored only electronically [4]. The cryptocurrency is designed to serve as a medium of exchange and for this, it uses cryptography systems to secure the transaction and control the subsequent creation of the cryptocurrency. Cryptocurrency is a subset of a

digital currency designed to function as a medium of exchange and cryptography is used to secure the transaction and control the future creation of the cryptocurrency.

Forecasting Bitcoin price is vitally important for both asset managers and independent investors. Although Bitcoin is a currency, it cannot be studied as another traditional currency where economic theories about uncovered interest rate parity, future cash-flows model, and purchasing power parity matter, since different standard factors of the relationship between supply and demand cannot be applied in the digital currency market like Bitcoin [5]. On the one hand, Bitcoin has different characteristics that make it useful for those agents who invest in Bitcoin, such as transaction speed, dissemination, decentrality, and the large virtual community of people interested in talking and providing relevant information about digital currencies, mainly Bitcoin [6].

Velankar and colleagues [7] attempted to predict the daily price change sign as accurately as possible using Bayesian regression and generalized linear model. To do this, they considered the daily trends of the Bitcoin market and focused on the characteristics of Bitcoin transactions, reaching an accuracy of 51% with the generalized linear model. McNally and co-workers [8] studied the precision with which the direction of the Bitcoin price in United States Dollar (USD) can be predicted. They used a recurrent neural network (RNN), a long short-term memory (LSTM) network, and the autoregressive integrated moving average (ARIMA) method. The LSTM network obtains the highest classification accuracy of 52% and a root mean square error (RMSE) of 8%. As expected, non-linear deep learning methods exceeded the ARIMA method's prognosis. For their part, Yogeshwaran and co-workers [9] applied convolutional and recurrent neural networks to predict the price of Bitcoin using data from a time interval of 5 min to 2 h, with convolutional neural networks showing a lower level of error, at around 5%. Demir and colleagues [10] predicted the price of Bitcoin using methods such as long short-term memory networks, naïve Bayes, and the nearest neighbor algorithm. These methods achieved accuracy rates between 97.2% and 81.2%. Rizwan, Narejo, and Javed [11] continued with the application of deep learning methods with the techniques of RNN and LSTM. Their results showed an accuracy of 52% and an 8% RMSE by the LSTM. Linardatos and Kotsiantis [12] had the same results, after using eXtreme Gradient Boosting (XGBoost) and LSTM; they concluded that this last technique yielded a lower RMSE of 0.999. Despite the superiority of computational techniques, Felizardo and colleagues [13] showed that ARIMA had a lower error rate than methods, such as random forest (RF), support vector machine (SVM), LSTM, and WaveNets, to predict the future price of Bitcoin. Finally, other works showed new deep learning methods, such as Dutta, Kumar, and Basu [14], who applied both LSTM and the gated recurring unit (GRU) model; the latter showed the best error result, with an RMSE of 0.019. Ji and co-workers [15] predicted the price of Bitcoin with different methodologies such as deep neural network (DNN), the LSTM model, and convolutional neural network. They obtained a precision of 60%, leaving the improvement of precision with deep learning techniques and a greater definition of significant variables as a future line of research. These authors show the need for stable prediction models, not only with data in and out of the sample, but also in forecasts of future results.

To contribute to the robustness of the Bitcoin price prediction models, in the present study a comparison of deep learning methodologies to predict and model the Bitcoin price is developed and, as a consequence, a new model that generates better forecasts of the Bitcoin price and its behavior in the future. This model can predict achieving accuracy levels above 95%. This model was constructed from a sample of 29 variables. Different methods were applied in the construction of the Bitcoin price prediction model to build a reliable model, which is contrasted with various methodologies used in previous works to check with which technique a high predictive capacity is achieved; specifically, the methods of deep recurrent neural networks, deep neural decision trees, and deep support vector machines, were used. Furthermore, this work attempts to obtain high accuracy, but it is also robust and stable in the future horizon to predict new observations, something that has not yet been reported by previous works [7–15], but which some authors demand for the development of these models and their real contribution [9,12].

We make two main contributions to the literature. First, we consider new explanatory variables for modeling the Bitcoin price, testing the importance of these variables which have not been considered so far. It has important implications for investors, who will know which indicators provide reliable, accurate, and potential forecasts of the Bitcoin price. Second, we improve the prediction accuracy concerning that obtained in previous studies with innovative methodologies.

This study is structured as follows: Section 2 explains the theory of methods applied. Section 3 offers details of the data and the variables used in this study. Section 4 develops the results obtained. Section 5 provides conclusions of the study and the purposes of the models obtained.

2. Deep Learning Methods

As previously stated, different deep learning methods have been applied for the development of Bitcoin price prediction models. We use this type of methodology thanks to its high predictive capacity obtained in the previous literature on asset pricing to meet one of the objectives of this study, which is to achieve a robust model. Specifically, deep recurrent convolution neural network, deep neural decision trees, and deep learning linear support vector machines have been used. The characteristics of each classification technique used are detailed below. In addition, the method of analysis of the sensitivity of variables used in the present study, in particular, the method of Sobol [16], which is necessary to determine the level of significance of the variables used in the prediction of Bitcoin price is recorded, fulfilling the need presented by the previous literature in the realization of the task of feature selection [15].

2.1. Deep Recurrent Convolution Neural Network (DRCNN)

Recurrent neural networks (RNN) have been applied in different fields for prediction due to its huge prediction performance. The previous calculations made are those that form the result within the structure of the RNN [17]. Having an input sequence vector x, the hidden nodes of a layer s, and the output of a hidden layer y, can be estimated as explained in Equations (1) and (2).

$$s_t = \sigma(W_{xs}x_t + W_{ss}s_{t-1} + b_s) \tag{1}$$

$$y_t = o(W_{so}s_t + b_y) \tag{2}$$

where W_{xs}, W_{ss}, and W_{so} define the weights from the input layer x to the hidden layer s, by the biases of the hidden layer and output layer. Equation (3) points out σ and o as the activation functions.

$$STFT\{z(t)\}(\tau,\omega) \equiv T(\tau,\omega) = \int_{-\infty}^{+\infty} z(t)\omega(t-\tau)e^{-j\omega t}\,dt \tag{3}$$

where $z(t)$ is the vibration signals, and $\omega(t)$ is the Gaussian window function focused around 0. $T(\tau, \omega)$ is the function that expresses the vibration signals. To calculate the hidden layers with the convolutional operation, Equations (4) and (5) are applied.

$$S_t = \sigma(W_{TS} * T_t + W_{ss} * S_{t-1} + B_s) \tag{4}$$

$$Y_t = o(W_{YS} * S_t + B_y) \tag{5}$$

where W indicates the convolution kernels.

Recurrent convolutional neural network (RCNN) can be heaped to establish a deep architecture, called the deep recurrent convolutional neural network (DRCNN) [18,19]. To use the DRCNN method in the predictive task, Equation (6) determines how the last phase of the model serves as a supervised learning layer.

$$\hat{r} = \sigma(W_h * h + b_h) \tag{6}$$

where W_h is the weight and b_h is the bias. The model calculates the residuals caused by the difference between the predicted and the actual observations in the training stage [20]. Stochastic gradient descent is applied for optimization to learn the parameters. Considering that the data at time t is r, the loss function is determined as shown in Equation (7).

$$\mathbf{L}(\mathbf{r}, \hat{\mathbf{r}}) = \frac{1}{2}\|\mathbf{r} - \hat{\mathbf{r}}\|_2^2 \tag{7}$$

2.2. Deep Neural Decision Trees (DNDT)

Deep neural decision trees are decision tree (DT) models performed by deep learning neural networks, where a weight division corresponding to the DNDT belongs to a specific decision tree and, therefore, it is possible to interpret its information [21]. Stochastic gradient descent (SGD) is used to optimize the parameters at the same time; this partitions the learning processing in mini-batches and can be attached to a larger standard neural network (NN) model for end-to-end learning with backward propagation. In addition, standard DTs gain experience through a greedy and recursive factor division. This can make a selection of functions more efficient [22]. The method starts by performing a soft binning function to compute the residual rate for each node, making it possible to make decisions divided into DNDTs [23]. The input of a binning function is a real scalar x which makes an index of the containers to which x belongs.

The activation function of the DNDT algorithm is carried out based on the NN represented in Equation (8).

$$\pi = \text{fw,b,}\tau\ (x) = \text{softmax}((wx + b)/\tau) \tag{8}$$

where w is a constant with value w = [1, 2, ..., n + 1], $\tau > 0$ is a temperature factor, and b is defined in Equation (9).

$$b = [0, -\beta 1, -\beta 1, -\beta 2, ..., -\beta 1 - \beta 2 - \cdots - \beta n] \tag{9}$$

The coding of the binning function x is given by the NN according the expression of Equation (9) [24]. The key idea is to build the DT with the applied Kronecker product from the binning function defined above. Connecting every feature x_d with its NN $f_d\ (x_d)$, we can determine all the final nodes of the DT as appears in Equation (10).

$$z = \text{f1(x1)} \otimes \text{f2(x2)} \otimes \cdots \otimes \text{fD(xD)} \tag{10}$$

where z expresses the leaf node index obtained by instance x in vector form. The complexity parameter of the model is determined by the number of cut points of each node. There may be inactive points since the values of the cut points are usually not limited.

2.3. Deep Learning Linear Support Vector Machines (DSVR)

Support vector machines (SVMs) were created for binary classification. Training data are denoted by its labels $(x_n, y_n), n = 1, \ldots, N, x_n \in \mathbb{R}^D, t_n \in \{-1, +1\}$; SVMs are optimized according to Equation (11).

$$\begin{gathered} \min_{w\xi_n} \frac{1}{2} W^T W + C \sum_{n=1}^{N} \xi_n \\ s.t.\ W^T x_n t_n \geq 1 - \xi_n > \forall n \\ \xi_n \geq 0\ \forall n \end{gathered} \tag{11}$$

where ξ_n are features that punish observations that do not meet the margin requirements [25]. The optimization problem is defined as appears in Equation (12).

$$\min_{w} \frac{1}{2} W^T W + C \sum_{n=1}^{N} max(1 - W^T x_n t_n, 0) \tag{12}$$

Usually the Softmax or 1-of-*K* encoding method is applied in the classification task of deep learning algorithms. In the case of working with 10 classes, the Softmax layer is composed of 10 nodes and expressed by p_i, where $i = 1, ..., 10$; p_i specifies a discrete probability distribution, $\sum_i^{10} p_i = 1$.

Equation (13) is defined by h as the activation of the penultimate layer nodes, W as the weight linked by the penultimate layer to the Softmax layer, and the total input into a Softmax layer. The next expression is the result.

$$a_i = \sum_k h_k W_{ki} \tag{13}$$

$$p_i = \frac{exp(a_i)}{\sum_j^{10} exp(a_j)} \tag{14}$$

The predicted class î would be as follows in Equation (15).

$$\hat{\imath} = arg\max_i p_i = arg\max_i a_i \tag{15}$$

Since linear-SVM is not differentiable, a popular variation is known as the DSVR, which minimizes the squared hinge loss as indicated in Equation (16).

$$\min_w \frac{1}{2} W^T W + C \sum_{n=1}^{N} max(1 - W^T x_n t_n, 0)^2 \tag{16}$$

The target of the DSVR is to train deep neural networks for prediction [24,25]. Equation (17) expresses the differentiation of the activation concerning the penultimate layer, where *l (w)* is said differentiation, changing the input x for the activation h.

$$\frac{\partial l(w)}{\partial h_n} = -Ct_n w(\mathbb{I}\{1 > w^T h_t t_n\}) \tag{17}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. Likewise, for the DSVR, we have Equation (18).

$$\frac{\partial l(w)}{\partial h_n} = -2Ct_n w(max(1 - W^T h_n t_n, 0)) \tag{18}$$

2.4. Sensitivity Analysis

Data mining methods have the virtue of offering a great amount of explanation to the authors' studied problem. To know what the degree is, sensitivity analysis is performed. This analysis tries to quantify the relative importance of the independent variables concerning the dependent variable [26,27]. To do this, the search for the reduction of the set of initial variables continues, leaving only the most significant ones. The variance limit follows, where one variable is significant if its variance increases concerning the rest of the variables as a whole. The Sobol method [16] is applied to decompose the variance of the total output V (Y) offered by the set of equations expressed in Equation (19).

$$V(Y) = \sum_i V_i + \sum_i \sum_{j>1} V_{ij} + \ldots + V_{1,2,\ldots k} \tag{19}$$

where $V_i = VE(Y|X_i)$ and $V_{ij} = VE(Y|X_i, X_j)) - V_i - V_j$.

$S_i = Vi/V$ and $S_{ij} = V_{ij}/V$ define the sensitivity indexes, with S_{ij} being the effect of interaction between two variables. The Sobol decomposition allows the estimation of a total sensitivity index, *STi*, which measures the sum of all the sensitivity effects involved in the independent variables.

3. Data and Variables

The sample period selected is from 2011 to 2019, with a quarterly frequency of data. To obtain the information of the independent variables, data from the IMF's International Financial Statistics (IFS), the World Bank, FRED Sant Louis, Google Trends, Quandl, and Blockchain.info were used.

The dependent variable used in this study is the Bitcoin price and is defined as the value of Bitcoin in USD. In addition, we used 29 independent variables, classified into demand and supply variables, attractiveness, and macroeconomic and financial variables, as possible predictors of the Bitcoin future price (Table 1). These variables were used throughout the previous literature [1,3,4,14].

Table 1. Independent variables.

Variables	Description
(a) Demand and Supply	
Transaction value	Value of daily transactions
Number of Bitcoins	Number of mined Bitcoins currently circulating on the network
Bitcoins addresses	Number of unique Bitcoin addresses used per day
Transaction volume	Number of transactions per day
Unspent transactions	Number of valid unspent transactions
Blockchain transactions	Number of transactions on blockchain
Blockchain addresses	Number of unique addresses used in blockchain
Block size	Average block size expressed in megabytes
Miners reward	Block rewards paid to miners
Mining commissions	Average transaction fees (in USD)
Cost per transaction	Miners' income divided by the number of transactions
Difficulty	Difficulty mining a new blockchain block
Hash	Times a hash function can be calculated per second
Halving	Process of reducing the emission rate of new units
(b) Attractive	
Forum posts	Number of new members in online Bitcoin forums
Forum members	New posts in online Bitcoin forums
(c) Macroeconomic and Financial	
Texas oil	Oil Price (West Texas)
Brent oil	Oil Price (Brent, London)
Dollar exchange rate	Exchange rate between the US dollar and the euro
Dow Jones	Dow Jones Index of the New York Stock Exchange
Gold	Gold price in US dollars per troy ounce

The sample is fragmented into three mutually exclusive parts, one for training (70% of the data), one for validation (10% of the data), and the third group for testing (20% of the data). The training data are used to build the intended models, while the validation data attempt to assess whether there is overtraining of those models. As for the test data, they serve to evaluate the built model and measure the predictive capacity. The percentage of correctly classified cases is the precision results and RMSE measures the level of errors made. Furthermore, for the distribution of the sample data in these three phases, cross-validation 10 times with 500 iterations was used [28,29].

4. Results

4.1. Descriptive Statistics

Table 2 shows a statistical summary of the independent variables for predicting Bitcoin price. It is observed that all the variables obtain a standard deviation not higher than each value of the mean. Therefore, the data show initial stability. On the other hand, there is a greater difference between the minimum and maximum values. Variables like mining commissions and cost per transaction show a small minimum value compared to their mean value. The same fact happens with the hash

variable. Despite these extremes, they do not affect the values of the standard deviations of the respective variables.

Table 2. Summary statistics.

Variables	Obs	Mean	SD	Min	Max
Transaction value	112	342,460,106,866,711.0000	143,084,554,727,531.0000	59,238,547,391,199.6000	735,905,260,141,564.0000
Number of bitcoins	112	13,634,297.4824	3,709,010.0736	5,235,454.5455	18,311,982.5000
Bitcoins addresses	112	285,034.2515	219,406.3874	1576.8333	849,668.1000
Transaction volume	112	154,548.8041	117,104.3686	1105.5000	373,845.6000
Unspent transactions	112	28,581,914.9054	22,987,595.3012	78,469.7273	66,688,779.9000
Blockchain transactions	112	156,444,312.9120	161,252,448.1997	237,174.8889	520,792,976.5000
Blockchain addresses	112	4,812,692.05	13,735,245.35	−14,437,299.03	117,863,226.2
Block size	112	0.4956	0.3638	0.0022	0.9875
Miners reward	112	420,160,582,581,028.0000	174,396,895,338,462.0000	101,244,436,734,897.0000	796,533,076,376,536.0000
Mining commissions	112	9,581,973,325,205.4400	42,699,799,790,392.8000	0.2591	315,387,506,596,395.0000
Cost per transaction	112	155,354,364,458,705.0000	156,696,788,525,225.0000	0.1179	757,049,771,708,905.0000
Difficulty	112	187,513,499,336,866.0000	195,421,886,528,251.0000	212,295,141,771.2000	836,728,509,520,663.0000
Hash	112	110,434,372.2765	154,717,725.3881	0.5705	516,395,703.4338
Halving	112	279,853,454,485,387.0000	162,806,469,642,875.0000	6,473,142,955,255.1700	804,437,327,302,638.0000
Forum posts	112	9279.8844	8585.0583	455.0000	53132.0000
Forum members	112	2432.2545	3394.4635	30.6364	14,833.3409
Texas Oil	112	72.4878	23.7311	21.1230	135.6700
Brent Oil	112	78.4964	26.5819	19.1900	139.3800
Dollar exchange rate	112	1.3767	0.9604	1.0494	8.7912
Dow Jones	112	15,926.7161	3324.8875	11,602.5212	22,044.8627
Gold	112	1329.400847	244.4099259	739.15	1846.75

4.2. Empirical Results

Table 3 and Figures 1–3 show the level of accuracy, the root mean square error (RMSE), and the mean absolute percentage error (MAPE). In all models, the level of accuracy always exceeds 92.61% for testing data. For its part, the RMSE and MAPE levels are adequate. The model with the highest accuracy is that of deep recurrent convolution neural network (DRCNN) with 97.34%, followed by the model of deep neural decision trees (DNDT) method with 96.94% on average by regions. Taken together, these results provide a level of accuracy far superior to that of previous studies. Thus, in the work of Ji and co-workers [15], an accuracy of around 60% is revealed; in the case of McNally and co-workers [8], it is close to 52%; and in the study of Rizwan, Narejo, and Javed [11], it approaches 52%. Finally, Table 4 shows the most significative variables by methods after applying the Sobol method for the sensitivity analysis.

Table 3. Results of accuracy evaluation: classification (%).

Sample	DRCNN			DNDT			DSVR		
	Acc. (%)	RMSE	MAPE	Acc. (%)	RMSE	MAPE	Acc. (%)	RMSE	MAPE
Training	97.34	0.66	0.29	95.86	0.70	0.33	94.49	0.75	0.38
Validation	96.18	0.71	0.34	95.07	0.74	0.37	93.18	0.81	0.43
Testing	95.27	0.77	0.40	94.42	0.79	0.42	92.61	0.84	0.47

DRCNN: deep recurrent convolution neural network; DNDT: deep neural decision trees; DSVR: deep learning linear support vector machines; Acc: accuracy; RMSE: root mean square error; MAPE: mean absolute percentage error.

Table 4. Results of accuracy evaluation: greater sensitivity variables.

DRCNN	DNDT	DSVR
Transaction value	Transaction volume	Transaction value
Transaction volume	Block size	Block size
Block size	Blockchain transactions	Blockchain transactions
Cost per transaction	Cost per transaction	Cost per transaction
Difficulty	Difficulty	Difficulty
Dollar exchange rate	Forum posts	Forum posts
Dow Jones	Dow Jones	Dollar exchange rate
Gold	Gold	Dow Jones
		Gold

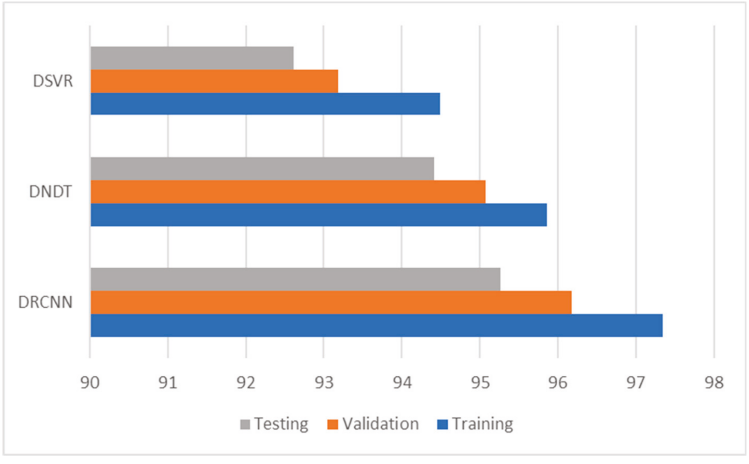


Figure 1. Results of accuracy evaluation: classification (%).

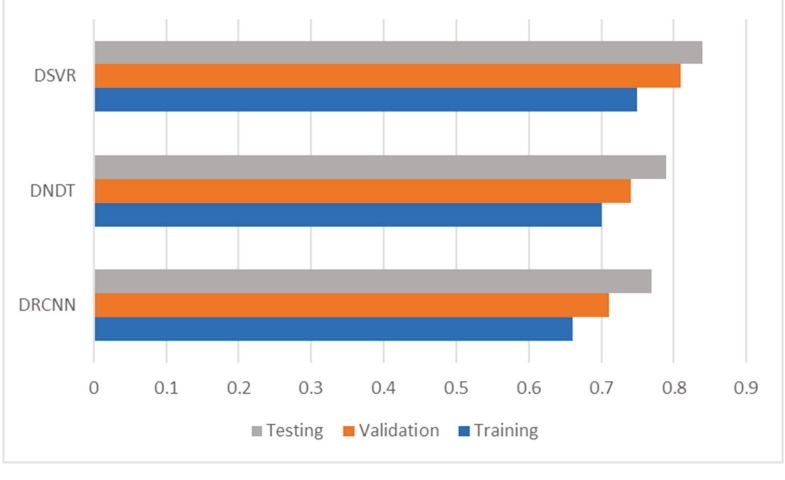


Figure 2. Results of accuracy evaluation: RMSE.

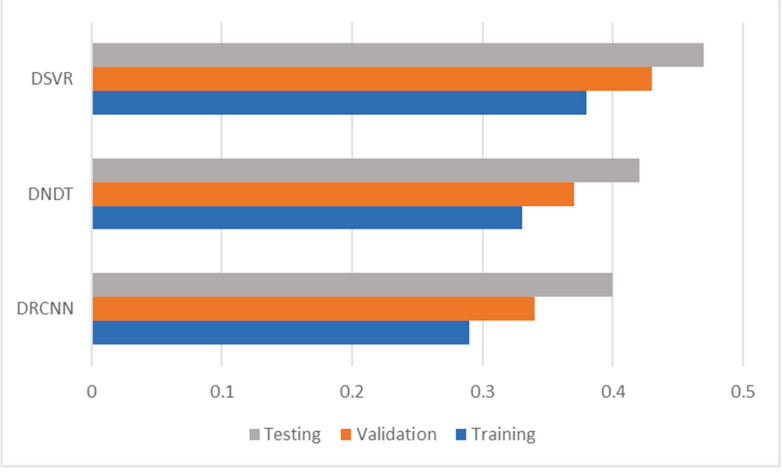


Figure 3. Results of accuracy evaluation: MAPE.

Table 4 shows additional information on the significant variables. Block size, cost per transaction, and difficulty were significant in the three models for each method applied. This demonstrates the importance of the cost to carry out the Bitcoin transaction, of the block of Bitcoins to buy, as well as the difficulty of the miners to find new Bitcoins, as the main factors in the task of determining the price of Bitcoin. This contrasts with the results shown in previous studies, where these variables are not significant or are not used by the initial set of variables [5,7,8]. The best results were obtained by the DRCNN method, where in addition to the aforementioned variables, the transaction value, transaction volume, block size, dollar exchange rate, Dow Jones, and gold were also significant. This shows that the demand and supply variables of the Bitcoin market are essential to predict its price, something that has been shown by some previous works [1,30]. Yet significant macroeconomic and financial variables have not been observed as important factors by other recent works [30,31], since they were shown as variables that did not influence Bitcoin price fluctuations. In our results, the macroeconomic variables of Dow Jones and gold have been significant in all methods.

On the other hand, the models built by the DNDT and DSVR methods show high levels of precision, although lower than those obtained by the DRCNN. Furthermore, these methods show some different significant variables. Such is the case of the variables of forum posts, a variable popularly used as a proxy for the level of future demand that Bitcoin could have, although with divergences in previous works regarding its significance to predict the price of Bitcoin, where some works show that this variable is not significant [11,14]. Finally, these methods show another macroeconomic variable that is more significant, in the case of the dollar exchange rate. This represents the importance that changes in the price of the USD with Bitcoin can be decisive in estimating the possible demand and, therefore, a change in price. This variable, like the rest of the macroeconomic variables, has not been shown as a significant variable [5,31].

This set of variables observed as significant represents a group of novel factors that determine the price of Bitcoin and therefore, is different from that shown in the previous literature.

4.3. Post-Estimations

In this section, we try to perform estimations of models to generate forecasts in a future horizon. For this, we used the framework of multiple-step ahead prediction, applying the iterative strategy and models built to predict one step forward are trained [32]. At time t, a prediction is made for moment $t + 1$, and this prediction is used to predict for moment $t + 2$ and so on. This means that the predicted data for $t + 1$ are considered real data and are added to the end of the available data [33]. Table 5

and Figures 4–6 show the accuracy and error results for $t + 1$ and $t + 2$ forecasting horizons. For $t + 1$, the range of precision for the three methods is 88.34–94.19% on average, where the percentage of accuracy is higher in the DRCNN (94.19%). For $t + 2$, this range of precision is 85.76–91.37%, where the percentage of accuracy is once again higher in the DRCNN (91.37%). These results show the high precision and great robustness of the models.

Table 5. Multiple-step ahead forecasts in forecast horizon = $t + 1$ and $t + 2$.

Horizon	DRCNN			DNDT			DSVR		
	Acc. (%)	RMSE	MAPE	Acc. (%)	RMSE	MAPE	Acc. (%)	RMSE	MAPE
$t + 1$	94.19	0.81	0.52	92.35	0.87	0.59	88.34	0.97	0.65
$t + 2$	91.37	0.92	0.63	89.41	1.03	0.67	85.76	1.10	0.78

Acc: accuracy.

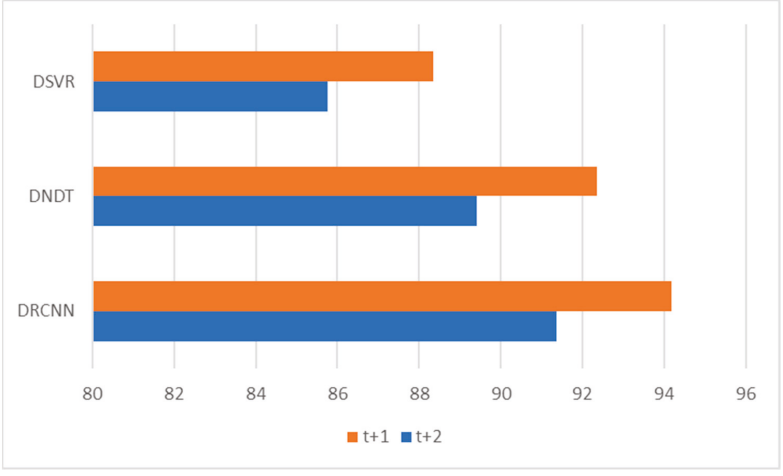


Figure 4. Multiple-step ahead forecasts in forecast horizon: accuracy.

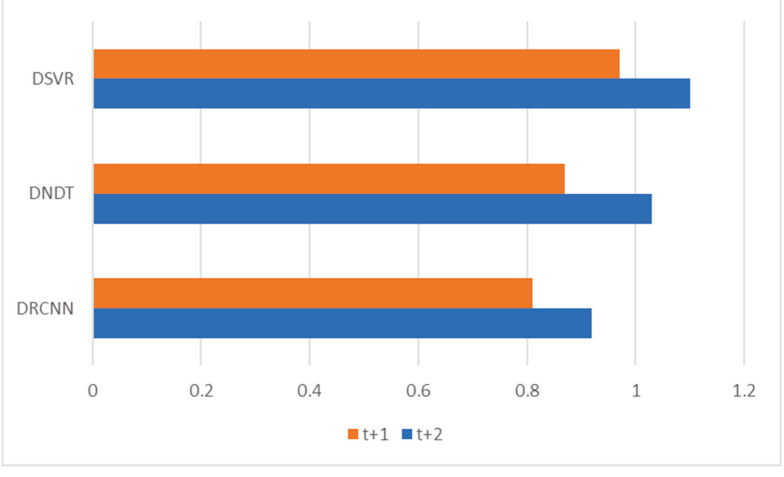


Figure 5. Multiple-step ahead forecasts in forecast horizon: RMSE.

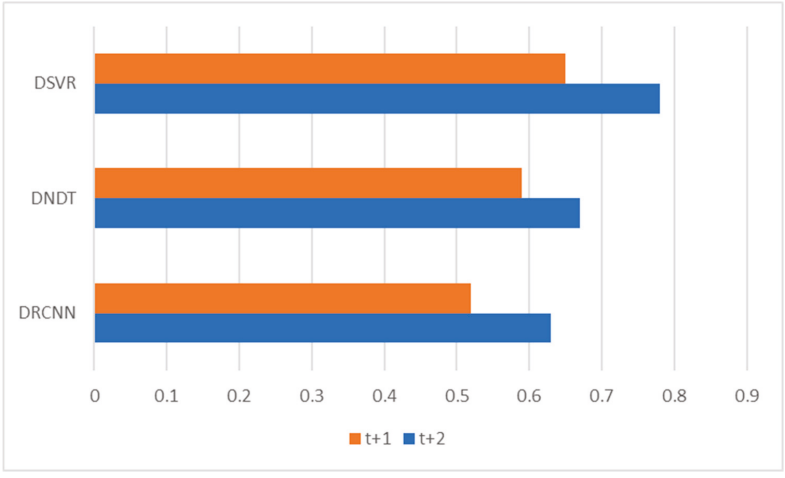


Figure 6. Multiple-step ahead forecasts in forecast horizon: MAPE.

5. Conclusions

This study developed a comparison of methodologies to predict Bitcoin price and, therefore, a new model was created to forecast this price. The period selected was from 2011 to 2019. We applied different deep learning methods in the construction of the Bitcoin price prediction model to achieve a robust model, such as deep recurrent convolutional neural network, deep neural decision trees and deep support vector machines. The DRCNN model obtained the highest levels of precision. We propose to increase the level of performance of the models to predict the price of Bitcoin compared to previous literature. This research has shown significantly higher precision results than those shown in previous works, achieving a precision hit range of 92.61–95.27%. Likewise, it was possible to identify a new set of significant variables for the prediction of the price of Bitcoin, offering great stability in the models developed predicting in the future horizons of one and two years.

This research allows us to increase the results and conclusions on the price of Bitcoin concerning previous works, both in matters of precision and error, but also on significant variables. A set of significant variables for each methodology applied has been selected analyzing our results, but some of these variables are recurrent in the three methods. This supposes an important addition to the field of cryptocurrency pricing. The conclusions are relevant to central bankers, investors, asset managers, private forecasters, and business professionals for the cryptocurrencies market, who are generally interested in knowing which indicators provide reliable, accurate, and potential forecasts of price changes. Our study suggests new and significant explanatory variables to allow these agents to predict the Bitcoin price phenomenon. These results have provided a new Bitcoin price forecasting model developed using three methods, with the DCRNN model as the most accurate, thus contributing to existing knowledge in the field of machine learning, and especially, deep learning. This new model can be used as a reference for setting asset pricing and improved investment decision-making.

In summary, this study provides a significant opportunity to contribute to the field of finance, since the results obtained have significant implications for the future decisions of asset managers, making it possible to avoid big change events of the price and the potential associated costs. It also helps these agents send warning signals to financial markets and avoid massive losses derived from an increase of volatility in the price.

Opportunities for further research in this field include developing predictive models considering volatility correlation of the other new alternative assets and also safe-haven assets such as gold or stable currencies, that evaluate the different scenarios of portfolio choice and optimization.

Author Contributions: Conceptualization, P.L.-F., D.A., P.L.-L. and M.A.F.-G.; Data curation, D.A. and M.A.F.-G.; Formal analysis, P.L.-F., D.A. and P.L.-L.; Funding acquisition, P.L.-F., P.L.-L. and M.A.F.-G.; Investigation, D.A. and M.A.F.-G.; Methodology, D.A.; Project administration, P.L.-F. and M.A.F.-G.; Resources, P.L.-F. and M.A.F.-G.; Software, D.A.; Supervision, D.A.; Validation, D.A. and P.L.-L.; Visualization, P.L.-F. and D.A.; Writing—original draft, P.L.-F. and D.A.; Writing—review & editing, P.L.-F., D.A., P.L.-L. and M.A.F.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Cátedra de Economía y Finanzas Sostenibles, University of Malaga, Spain.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kristoufek, L. What Are the Main Drivers of the Bitcoin Price? Evidence from Wavelet Coherence Analysis. *PLoS ONE* **2015**, *10*, e0123923. [CrossRef]
2. Wamba, S.F.; Kamdjoug, J.R.K.; Bawack, R.E.; Keogh, J.G. Bitcoin, Blockchain and Fintech: A systematic review and case studies in the supply chain. *Prod. Plan. Control Manag. Oper.* **2019**, *31*, 115–142. [CrossRef]
3. Chen, W.; Zheng, Z.; Ma, M.; Wu, J.; Zhou, Y.; Yao, J. Dependence structure between bitcoin price and its influence factors. *Int. J. Comput. Sci. Eng.* **2020**, *21*, 334–345. [CrossRef]
4. Balcilar, M.; Bouri, E.; Gupta, R.; Roubaud, D. Can volume predict bitcoin returns and volatility? A quantiles-based approach. *Econ. Model.* **2017**, *64*, 74–81. [CrossRef]
5. Ciaian, P.; Rajcaniova, M.; Artis Kancs, D. The economics of BitCoin price formation. *Appl. Econ.* **2016**, *48*, 1799–1815. [CrossRef]
6. Schmidt, R.; Möhring, M.; Glück, D.; Haerting, R.; Keller, B.; Reichstein, C. Benefits from Using Bitcoin: Empirical Evidence from a European Country. *Int. J. Serv. Sci. Manag. Eng. Technol.* **2016**, *7*, 48–62. [CrossRef]
7. Velankar, S.; Valecha, S.; Maji, S. Bitcoin Price Prediction using Machine Learning. In Proceedings of the 20th International Conference on Advanced Communications Technology (ICACT), Chuncheon-si, Korea, 11–14 February 2018.
8. McNally, S.; Roche, J.; Caton, S. Predicting the Price of Bitcoin Using Machine Learning. In Proceedings of the 26th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, Cambridge, UK, 21–23 March 2018.
9. Yogeshwaran, S.; Kaur, M.J.; Maheshwari, P. Project Based Learning: Predicting Bitcoin Prices using Deep Learning. In Proceedings of the 2019 IEEE Global Engineering Education Conference (EDUCON), Dubai, UAE, 9–11 April 2019.
10. Demir, A.; Akılotu, B.N.; Kadiroğlu, Z.; Şengür, A. Bitcoin Price Prediction Using Machine Learning Methods. In Proceedings of the 2019 1st International Informatics and Software Engineering Conference (UBMYK), Ankara, Turkey, 6–7 November 2019.
11. Rizwan, M.; Narejo, S.; Javed, M. Bitcoin price prediction using Deep Learning Algorithm. In Proceedings of the 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS), Karachi, Pakistan, 14–15 December 2019.
12. Linardatos, P.; Kotsiantis, S. Bitcoin Price Prediction Combining Data and Text Mining. In *Advances in Integrations of Intelligent Methods. Smart Innovation, Systems and Technologies*; Hatzilygeroudis, I., Perikos, I., Grivokostopoulou, F., Eds.; Springer: Singapore, 2020.
13. Felizardo, L.; Oliveira, R.; Del-Moral-Hernández, E.; Cozman, F. Comparative study of Bitcoin price prediction using WaveNets, Recurrent Neural Networks and other Machine Learning Methods. In Proceedings of the 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), Beijing, China, 28–30 October 2019.
14. Dutta, A.; Kumar, S.; Basu, M. A Gated Recurrent Unit Approach to Bitcoin Price Prediction. *J. Risk Financ. Manag.* **2020**, *13*, 23. [CrossRef]
15. Ji, S.; Kim, J.; Im, H. A Comparative Study of Bitcoin Price Prediction Using Deep Learning. *Mathematics* **2019**, *7*, 898. [CrossRef]
16. Saltelli, A. Making best use of model evaluations to compute sensitivity indices. *Comput. Phys. Commun.* **2002**, *145*, 280–297. [CrossRef]
17. Wang, S.; Chen, X.; Tong, C.; Zhao, Z. Matching Synchrosqueezing Wavelet Transform and Application to Aeroengine Vibration Monitoring. *IEEE Trans. Instrum. Meas.* **2017**, *66*, 360–372. [CrossRef]

18. Huang, C.-W.; Narayanan, S.S. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo, Hong Kong, China, 10–14 July 2017; pp. 583–588.
19. Ran, X.; Xue, L.; Zhang, Y.; Liu, Z.; Sang, X.; Xe, J. Rock Classification from Field Image Patches Analyzed Using a Deep Convolutional Neural Network. *Mathematics* **2019**, *7*, 755. [CrossRef]
20. Ma, M.; Mao, Z. Deep Recurrent Convolutional Neural Network for Remaining Useful Life Prediction. In Proceedings of the 2019 IEEE International Conference on Prognostics and Health Management (ICPHM), San Francisco, CA, USA, 17–20 June 2019; pp. 1–4.
21. Yang, Y.; Garcia-Morillo, I.; Hospedales, T.M. Deep Neural Decision Trees. In Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden, 14 July 2018.
22. Norouzi, M.; Collins, M.D.; Johnson, M.; Fleet, D.J.; Kohli, P. Efficient non-greedy optimization of decision trees. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2015; pp. 1729–1737.
23. Dougherty, J.; Kohavi, R.; Sahami, M. Supervised and unsupervised discretization of continuous features. In Proceedings of the 12th International Conference on Machine Learning (ICML), Tahoe City, CA, USA, 9–12 July 1995.
24. Jang, E.; Gu, S.; Poole, B. Categorical reparameterization with Gumbel-Softmax. *arXiv* **2017**, arXiv:1611.01144.
25. Tang, Y. Categorical reparameterization with Gumbel-Softmax. *arXiv* **2013**, arXiv:1306.0239.
26. Delen, D.; Kuzey, C.; Uyar, A. Measuring firm performance using financial ratios: A decision tree approach. *Expert Syst. Appl.* **2013**, *40*, 3970–3983. [CrossRef]
27. Efimov, D.; Sulieman, H. Sobol Sensitivity: A Strategy for Feature Selection. In *Mathematics Across Contemporary Sciences. AUS-ICMS 2015*; Springer Proceedings in Mathematics & Statistics: Cham, Switzerland, 2017; Volume 190.
28. Alaminos, D.; Fernández, S.M.; García, F.; Fernández, M.A. Data Mining for Municipal Financial Distress Prediction, Advances in Data Mining, Applications and Theoretical Aspects. *Lect. Notes Comput. Sci.* **2018**, *10933*, 296–308.
29. Zhang, G.P.; Qi, M. Neural network forecasting for seasonal and trend time series. *Eur. J. Oper. Res.* **2005**, *160*, 501–514. [CrossRef]
30. Polasik, M.; Piotrowska, A.I.; Wisniewski, T.P.; Kotkowski, R.; Lightfoot, G. Price fluctuations and the use of Bitcoin: An empirical inquiry. *Int. J. Electron. Commer.* **2015**, *20*, 9–49. [CrossRef]
31. Al-Khazali, O.; Bouri, E.; Roubaud, D. The impact of positive and negative macroeconomic news surprises: Gold versus Bitcoin. *Econ. Bull.* **2018**, *38*, 373–382.
32. Koprinska, I.; Rana, M.; Rahman, A. Dynamic ensemble using previous and predicted future performance for Multi-step-ahead solar power forecasting. In Proceedings of the ICANN 2019: Artificial Neural Networks and Machine Learning, Munich, Germany, 17–19 September 2019; pp. 436–449.
33. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLoS ONE* **2018**, *13*, e0194889. [CrossRef] [PubMed]

mathematics

MDPI

Article

A Novel Methodology to Calculate the Probability of Volatility Clusters in Financial Series: An Application to Cryptocurrency Markets

Venelina Nikolova [1,†], Juan E. Trinidad Segovia [1,*,†], Manuel Fernández-Martínez [2,†] and Miguel Angel Sánchez-Granero [3,†]

1 Department of Accounting and Finance, Faculty of Economics and Business, Universidad de Almería, 04120 Almería, Spain; vdn088@inlumine.ual.es
2 University Centre of Defence at the Spanish Air Force Academy, MDE-UPCT, 30720 Santiago de la Ribera, Región de Murcia, Spain; manuel.fernandez-martinez@cud.upct.es
3 Department of Mathematics, Faculty of Science, Universidad de Almería, 04120 Almería, Spain; misanche@ual.es
* Correspondence: jetrini@ual.es; Tel.: +34-950-015-817
† These authors contributed equally to this work.

Received: 25 May 2020; Accepted: 22 July 2020; Published: 24 July 2020

Abstract: One of the main characteristics of cryptocurrencies is the high volatility of their exchange rates. In a previous work, the authors found that a process with volatility clusters displays a volatility series with a high Hurst exponent. In this paper, we provide a novel methodology to calculate the probability of volatility clusters with a special emphasis on cryptocurrencies. With this aim, we calculate the Hurst exponent of a volatility series by means of the FD4 approach. An explicit criterion to computationally determine whether there exist volatility clusters of a fixed size is described. We found that the probabilities of volatility clusters of an index (S&P500) and a stock (Apple) showed a similar profile, whereas the probability of volatility clusters of a forex pair (Euro/USD) became quite lower. On the other hand, a similar profile appeared for Bitcoin/USD, Ethereum/USD, and Ripple/USD cryptocurrencies, with the probabilities of volatility clusters of all such cryptocurrencies being much greater than the ones of the three traditional assets. Our results suggest that the volatility in cryptocurrencies changes faster than in traditional assets, and much faster than in forex pairs.

Keywords: volatility cluster; Hurst exponent; FD4 approach; volatility series; probability of volatility cluster; S&P500; Bitcoin; Ethereum; Ripple

1. Introduction

It is easy to observe that large fluctuations in stock market prices are followed by large ones, whereas small fluctuations in prices are more likely to be followed by small ones. This property is known as volatility clustering. Recent works, such as [1,2], have shown that while large fluctuations tend to be more clustered than small ones, large losses tend to lump together more severely than large gains. The financial literature is interested in modeling volatility clustering since the latter is considered as a key indicator of market risk. In fact, the trading volume of some assets, such as derivatives, increases over time, making volatility their most important pricing factor.

It is worth mentioning that both, high and low volatilities, seem to be a relevant factor for stock market crises according to Danielsson et al. [3]. They also found that the relation between unexpected volatility and the incidence of crises became stronger in the last few decades. In the same line, Valentine et al. [4] showed that market instability is not only the result of large volatility, but also of small volatility.

The classical approach for volatility clusters lies in nonlinear models, based on heteroskedastic conditionally variance. They include ARCH [5], GARCH [6–8], IGARCH [9], and FIGARCH [10,11] models.

On the other hand, agent-based models allow reproducing and explaining some stylized facts of financial markets [12]. Interestingly, several works have recently been appeared in the literature analyzing a complete order book by real-time simulation [13–15]. Regarding the volatility clustering, it is worth mentioning that Lux et al. [16] highlighted that volatility is explained by market instability. Later, Raberto et al. [17] introduced an agent-based artificial market whose heterogeneous agents exchange only one asset, which exhibits some key stylized facts of financial markets. They found that the volatility clustering effect is sensitive to the model size, i.e., when the number of operators increases, the volatility clustering effect tends to disappear. That result is in accordance with the concept of market efficiency.

Krawiecki et al. [18] introduced a microscopic model consisting of many agents with random interactions. Then the volatility clustering phenomenon appears as a result of attractor bubbling. Szabolcs and Farmer [19] empirically developed a behavioral model for order placement to study endogenous dynamics of liquidity and price formation in the order book. They were able to describe volatility through the order flow parameters.

Alfarano et al. [20] contributed a simple model of an agent-based artificial market with volatility clustering generated due to interaction between traders. Similar conclusions were obtained by Cont [21], Chen [22], He et al. [23], and Schmitt and Westerhoff [24].

Other findings on the possible causes of volatility clusters are summarized below. Cont [21] showed that volatility is explained by agent behavior; Chen [22] stated that return volatility correlations arise from asymmetric trading and investors' herding behavior; He et al. [23] concluded that trade between fundamental noise and noise traders causes the volatility clustering; and Schmitt and Westerhoff [24] highlighted that volatility clustering arises due to the herding behavior of speculators.

Chen et al. [25] proposed an agent-based model with multi-level herding to reproduce the volatilities of New York and Hong Kong stocks. Shi et al. [26] explained volatility clustering through a model of security price dynamics with two kind of participants, namely speculators and fundamental investors. They considered that information arrives randomly to the market, which leads to changes in the viewpoint of the market participants according to a certain ratio. Verma et al. [27] used a factor model to analyze how market volatility could be explained by assets' volatility.

An interesting contribution was made by Barde in [28], where the author compared the performance of this kind of model with respect to the ARCH/GARCH models. In fact, the author remarked that the performance of three kinds of agent-based models for financial markets is better in key events. Population switching was found also as a crucial factor to explain volatility clustering and fat tails.

On the other hand, the concept of a volatility series was introduced in [2] to study the volatility clusters in the S&P500 series. Moreover, it was shown that the higher the self-similarity exponent of the volatility series of the S&P500, the more frequent the volatility changes and, therefore, the more likely that the volatility clusters appear. In the current article, we provide a novel methodology to calculate the probability of volatility clusters of a given size in a series with special emphasis on cryptocurrencies.

Since the introduction of Bitcoin in 2008, the cryptocurrency market has experienced a constant growth, just like the use of crypto assets as an investment or medium of exchange day to day. As of June 2020, there are 5624 cryptocurrencies, and their market capitalization exceeds 255 billion USD according to the website CoinMarketCap [29]. However, one of the main characteristics of cryptocurrencies is the high volatility of their exchange rates, and consequently, the high risk associated with their use.

Lately, Bitcoin has received more and more attention by researchers. Compared to the traditional financial markets, the cryptocurrency market is very young, and because of this, there are relatively few research works on their characteristics, and all of them quite recent. Some of the authors analyzed the Bitcoin market efficiency by applying different approaches, including the Hurst exponent (cf. [30] for a

detailed review), whereas others investigated its volatility using other methods. For instance, Letra [31] used a GARCH model for Bitcoin daily data; Bouoiyour and Selmi [32] carried out many extensions of GARCH models to estimate Bitcoin price dynamics; Bouri, Azzi, and Dyhberg [33] analyzed the relation between volatility changes and price returns of Bitcoin based on an asymmetric GARCH model; Balcilar et al. [34] analyzed the relation between the trading volume of Bitcoin and its returns and volatility by employing, in contrast, a non-parametric causality in quantiles test; and Baur et al. [35] studied the statistical properties of Bitcoin and its relations with traditional asset classes.

Meanwhile, in 2017, Bariviera et al. [36] used the Hurst exponent to compare Bitcoin dynamics with standard currencies' dynamics and detected evidence of persistent volatility and long memory, facts that justify the GARCH-type models' application to Bitcoin prices. Shortly after that, Phillip et al. [37] provided evidence of slight leverage effects, volatility clustering, and varied kurtosis. Furthermore, Zhang et al. [38] analyzed the first eight cryptocurrencies that represent almost 70% of cryptocurrency market capitalization and pointed out that the returns of cryptocurrencies exhibit leverage effects and strong volatility clustering.

Later, in 2019, Kancs et al. [39], based on the GARCH model, estimated factors that affect Bitcoin price. For it, they used hourly data for the period between 2013 and 2018. After plotting the data graphically, they suggested that periods of high volatility follow periods of high volatility, and periods of low volatility follow periods of low volatility, so in the series, large returns follow large returns and small returns small returns. All these facts indicate evidence of volatility clustering and, therefore, that the residue is conditionally heteroscedastic.

The structure of this article is as follows. Firstly, Section 2 contains some mathematical basic concepts on measure theory and probability (Section 2.1), the FD4 approach Section 2.2), and the volatility series (Section 2.3). The core of the current paper is provided in Section 3, where we explain in detail how to calculate the probability of volatility clusters of a given size. A study of volatility clusters in several cryptocurrencies, as well as in traditional exchanges is carried out in Section 4. Finally, Section 5 summarizes the main conclusions of this work.

2. Methods

This section contains some mathematical tools of both measure and probability theories (cf. Section 2.1) that allow us to mathematically describe the FD4 algorithm applied in this article (cf. Section 2.2) to calculate the self-similarity index of time series. On the other hand, the concept of a volatility series is addressed in Section 2.3.

2.1. Random Functions, Their Increments, and Self-Affinity Properties

Let $t \geq 0$ denote time and $(X, \mathcal{A}, P)$ be a probability space. We shall understand that $\mathbf{X} = \{X_t \equiv X(t, \omega) : t \geq 0\}$ is a random process (also a random function) from $[0, \infty) \times \Omega$ to $\mathbb{R}$, if X_t is a random variable for all $t \geq 0$ and $\omega \in \Omega$, where Ω denotes a sample space. As such, we may think of $\mathbf{X}$ as defining a sample function $t \mapsto X_t$ for all $\omega \in \Omega$. Hence, the points in Ω do parameterize the functions $\mathbf{X} : [0, \infty) \to \mathbb{R}$ with P being a measure of probability in the class of such functions.

Let X_t and Y_t be two random functions. The notation $X_t \sim Y_t$ means that the finite joint distribution functions of such random functions are the same. A random process $\mathbf{X} = \{X_t : t \geq 0\}$ is said to be self-similar if there exists a parameter $H > 0$ such that the following power law holds:

$$X_{at} \sim a^H X_t \tag{1}$$

for each $a > 0$ and $t \geq 0$. If Equation (1) is fulfilled, then H is named the self-similarity exponent (also index) of the process $\mathbf{X}$. On the other hand, the increments of a random function X_t are said to be stationary as long as $X_{a+t} - X_a \sim X_t - X_0$ for all $t \geq 0$ and $a > 0$. We shall understand that the

increments of a random function are self-affine of the parameter $H \geq 0$ if the next power law stands for all $h > 0$ and $t_0 \geq 0$:

$$X_{t_0+\tau} - X_{t_0} \sim h^{-H} \left(X_{t_0+h\tau} - X_{t_0} \right).$$

Let X_t be a random function with self-affine increments of the parameter H. Then, the following T^H-law holds:

$$\mathcal{M}_T \sim T^H \, \mathcal{M}_1,$$

where its (T-period) cumulative range is defined as:

$$\mathcal{M}_{t,T} := \sup \left\{ X(s,\omega) - X(t,\omega) : s \in [t, t+T] \right\} - \inf \left\{ X(s,\omega) - X(t,\omega) : s \in [t, t+T] \right\},$$

and $\mathcal{M}_T := \mathcal{M}_{0,T}$ (cf. Corollary 3.6 in [40]).

2.2. The FD4 Approach

The FD4 approach was first contributed in [41] to deal with calculations concerning the self-similarity exponent of random processes. It was proven that the FD4 generalizes the GM2 procedure (cf. [42,43]), as well as the fractal dimension algorithms (cf. [44]) to calculate the Hurst exponent of any process with stationary and self-affine increments (cf. Theorem 3.1 in [41]). Moreover, the accuracy of such an algorithm was analyzed for samples of (fractional) Brownian motions and Lévy stable processes with lengths ranging from 2^5 to 2^{10} points (cf. Section 5 in [41]).

Next, we mathematically show how that parameter could be calculated by the FD4 procedure. First of all, let $\mathbf{X} = \{X_t : t \geq 0\}$ be a random process with stationary increments. Let $q > 0$, and assume that for each $X_t \in \mathbf{X}$, there exists $m_q(X_t) := \mathrm{E}\left[|X_t|^q\right]$, its (absolute) q-order moment. Suppose, in addition, that there exists a parameter $H > 0$ for which the next relation, which involves (τ-period) cumulative ranges of $\mathbf{X}$, holds:

$$\mathcal{M}_\tau \sim \tau^H \, \mathcal{M}_1. \tag{2}$$

Recall that this power law stands for the class of (H-)self-similar processes with self-affine increments (of parameter H; see Section 2.1), which, roughly speaking, is equivalent to the class of processes with stationary increments (cf. Lemma 1.7.2 in [45]). Let us discretize the period by $\tau_n = 2^{-n} : n \in \mathbb{N}$ and take q-powers on both sides of Equation (2). Thus, we have:

$$\mathcal{M}_{\tau_n}^q \sim \tau_n^{q\,H} \, \mathcal{M}_1^q \text{ for all } n \in \mathbb{N}. \tag{3}$$

Clearly, the expression in Equation (3) could be rewritten in the following terms:

$$X_n^q \sim \tau_n^{q\,H} \, X_0^q = 2^{-n\,q\,H} \, X_0 \tag{4}$$

where, for short, the notation $X_n := \mathcal{M}_{\tau_n} = \mathcal{M}_{2^{-n}}$ is used for all $n \in \mathbb{N}$. Since the two random variables in Equation (4) are equally distributed, their means must be the same, i.e.,

$$m_q(X_n) = \mathrm{E}\left[X_n^q\right] = 2^{-n\,q\,H} \, \mathrm{E}\left[X_0^q\right] = 2^{-nqH} \, m_q(X_0). \tag{5}$$

Taking (2-base) logarithms on both sides of Equation (5), the parameter H could be obtained by carrying out a linear regression of:

$$H = \frac{1}{nq} \log_2 \frac{m_q(X_0)}{m_q(X_n)}. \tag{6}$$

vs. q. Alternatively, observe that the expression in Equation (4) also provides a relation between cumulative ranges of consecutive periods of $\mathbf{X}$, i.e.,

$$X_n^q \sim 2^{q\,H} \, X_{n+1}^q. \tag{7}$$

Since the random variables on each side of Equation (7) have the same (joint) distribution function, their means must be equal, namely,

$$m_q(X_n) = \mathrm{E}\,[X_n^q] = 2^{q\,H}\,\mathrm{E}\,[X_{n+1}^q] = 2^{q\,H}\,m_q(X_{n+1}) \text{ for all } n \in \mathbb{N}, \tag{8}$$

which provides a strong connection between consecutive moments of order q of $\mathbf{X}$. If (two-base) logarithms are taken on both sides of Equation (8), a linear regression of the expression appearing in Equation (9) vs. q allows calculating the self-similarity exponent of $\mathbf{X}$ (whenever self-similar patterns do exist for such a process):

$$H = \frac{1}{q}\,\log_2\,\frac{m_q(X_n)}{m_q(X_{n+1})}. \tag{9}$$

Hence, the FD algorithm is defined as the approach whose running is based on the expressions appearing in either Equation (5) or Equation (8). The main restriction underlying the FD algorithm consists of the assumption regarding the existence of the q-order moments of the random process $\mathbf{X}$. At first glance, any non-zero value could be assigned to q to calculate the self-similarity exponent (provided that the existence of that sample moment could be guaranteed). In the case of Lévy stable motions, for example, given q_0, it may occur that $m_q(X_n)$ does not exist for any $q > q_0$. As such, we shall select $q = 0.01$ to calculate the self-similarity index of a time series by the FD algorithm, thus leading to the so-called FD4 algorithm. Equivalently, the FD4 approach denotes the FD algorithm for $q = 0.01$. In this paper, the self-similarity exponent of a series by the FD4 approach is calculated according to the expression in Equation (6). Indeed, since it is equivalent to:

$$\log_2 m_q(X_n) = \log_2 m_q(X_0) - nqH,$$

the Hurst exponent of the series is obtained as the slope of a linear regression, which compares $\log_2 m_q(X_n)$ with respect to n. In addition, notice that a regression coefficient close to one means that the expression in Equation (5) is fulfilled. As such, the calculation of $m_q(X_n)$ becomes necessary to deal with the procedure described above, and for each n, it depends on a given sample of the random variable $X_n \in \mathbf{X}$. For computational purposes, the length of any sample of X_n is chosen to be equal to 2^n. Accordingly, the greater n, the more accurate the value of $m_q(X_n)$ is. Next, we explain how to calculate $m_q(X_n)$. Let a log-price series be given, and divide it into 2^n non-overlapping blocks, $\mathcal{B}_i : i = 1, \ldots, 2^n$. The length of each block is $k := 2^{-n} \cdot \text{length(series)}$, so for each $i = 1, \ldots, 2^n$, we can write $\mathcal{B}_i = \{B_1, \ldots, B_k\}$. Then:

1. Determine the range of each block $\mathcal{B}_i$, i.e., calculate $R_i = \max\{B_j : j = 1, \ldots, k\} - \min\{B_j : j = 1, \ldots, k\}$ for each $i = 1, \ldots, 2^n$.
2. The (q-order) sample moment is given by $m_q(X_n) = 2^{-n}\,\sum_{i=1}^{2^n} R_i^q$.

According to the step (1), both the minimum and the maximum values of each period are required to calculate each range R_i. In this way, notice that such values are usually known for each trading period in the context of financial series. It is also worth noting that when n takes the value $\log_2(\text{length(series)})$, then each block only consists of a single element. In this case, though, each range R_i can be still computed.

2.3. The Volatility Series

The concept of a volatility series was first contributed in Section 2.2 of [2] as an alternative to classical (G)ARCH models with the aim to detect volatility clusters in series of asset returns from the S&P500 index. It was found, interestingly, that whether clusters of high (resp., low) volatility appear in the series, then the self-similarity exponent of the associated volatility series increases (resp., decreases).

Let r_n denote the log-return series of a (index/stock) series. In financial series, the autocorrelation function of the r_n's is almost null, though the $|r_n|$ series is not. The associated volatility series is defined

as $s_n = |r_n| + s_{n-1} - m$, where $|\cdot|$ refers to the absolute value function, m is a constant, and $s_0 = 0$. For practical purposes, we set $m = \text{mean}\,|r_n|$.

Next, we explain how the Hurst exponent of the volatility series, s_n, could provide a useful tool to detect volatility clusters in a series of asset returns. Firstly, assume that the volatility of the series is constant. Then, the values of the associated volatility series would be similar to those from a sample of a Brownian motion. Hence, the self-similarity exponent of that volatility series would become close to 0.5. On the contrary, suppose that there exist some clusters of high (resp., low) volatility in the series. Thus, the graph of its associated volatility series becomes smoother, as illustrated in Figure 1, which also depicts the concept of a volatility series. Hence, almost all the values of the volatility series are greater (resp., lower) than the mean of the series. Accordingly, the volatility series turns out to be increasing (resp., decreasing), so its self-similarity exponent also increases (resp., decreases).

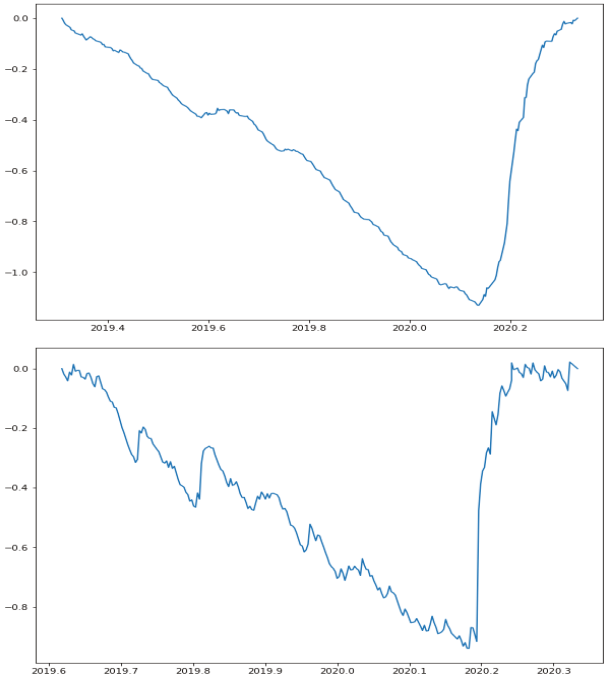


Figure 1. The picture at the top depicts the volatility series of the S&P500 index in the period ranging from March 2019 to March 2020. A self-similarity exponent equal to 0.94 is obtained by the FD4 approach. The graph below illustrates the volatility series of the Bitcoin/USD index in a similar period (both series contain 250 data (one year of trading), but recall that the Bitcoin/USD currency remains active also on weekends). In that case, a self-similarity exponent equal to 0.65 is obtained.

Following the above, the Hurst exponent of the volatility series of an index or asset provides a novel approach to explore the presence of volatility clusters in series of asset returns.

3. Calculating the Probability of Volatility Clusters of a Given Size

In this section, we explore how to estimate the probability of the existence of volatility clusters for blocks of a given size. Equivalently, we shall address the next question: What is the probability that a volatility cluster appears in a period of a given size? Next, we show that the Hurst exponent of a volatility series (see Sections 2.2 and 2.3) for blocks of that size plays a key role.

We know that the Hurst exponent of the volatility series is high when there are volatility clusters in the series [2]. However, how high should it be?

To deal with this, we shall assume that the series of (log-)returns follows a Gaussian distribution. However, it cannot be an i.i.d. process since the standard deviation of the Gaussian distribution is allowed to change. This hypothesis is more general than an ARCH or GARCH model, for example. Since we are interested in the real possibility that the volatility changes and, in fact, there exist volatility clusters, a static fixed distribution cannot be assumed. In this way, it is worth noting that the return distribution of these kinds of processes (generated from Gaussian distributions with different standard deviations) is not Gaussian, and it is flexible enough to allow very different kinds of distributions.

As such, let us assume that the series of the log-returns, r_n, follows a normal distribution, $N(0, \sigma(n))$, where its standard deviation varies over time via the function $\sigma(n)$. In fact, some classical models such as ARCH, GARCH, etc., stand as particular cases of that model. As such, we shall analyze the existence of volatility clusters in the following terms. We consider that there exist volatility clusters as long as there are, at least, both, a period of high volatility and a period of low volatility. Figure 2 illustrates that condition. Indeed, two broad periods could be observed concerning the volatility series of the S&P500 index. The first one has a low volatility (and hence, a decreasing volatility series) and the second one a high volatility (and hence, an increasing volatility series). In this case, the effect of the higher volatility (due to the COVID-19 crisis) is evident, thus being confirmed by a very high Hurst exponent of the corresponding volatility series (equal to 0.94).

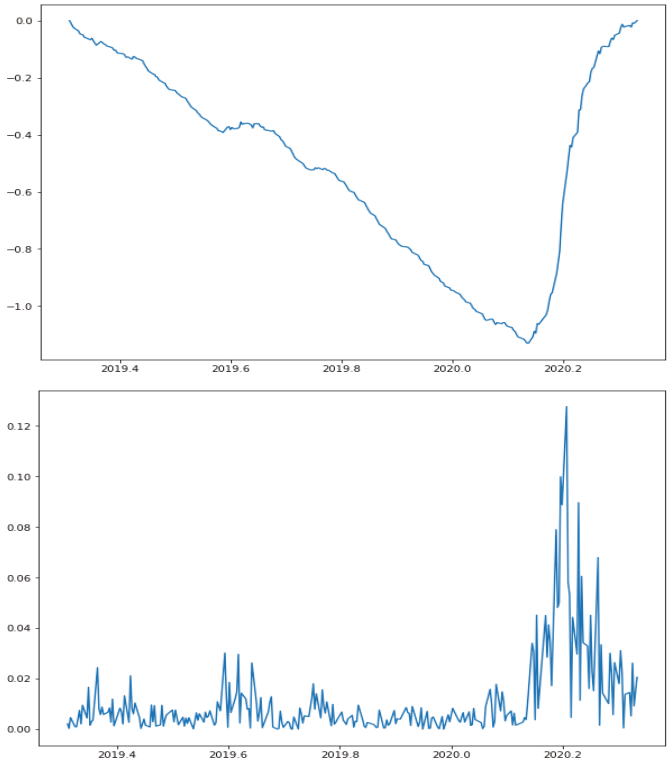


Figure 2. The graph at the top depicts the volatility series of the S&P500 index in the period ranging from March 2019 to March 2020. On the other hand, the chart at the bottom shows the series of absolute values of the log-returns of the S&P500 in the same period. That period, the self-similarity index of the volatility series of the S&P500, was found to be equal to 0.94 by the FD4 algorithm.

On the other hand, Figure 3 depicts the volatility series of the S&P500 index in the period ranging from January 2017 to January 2018. A self-similarity index equal to 0.55 was found by the FD4 algorithm. In this case, though, it is not so clear that there are volatility clusters, which is in accordance with the low Hurst exponent of that volatility series.

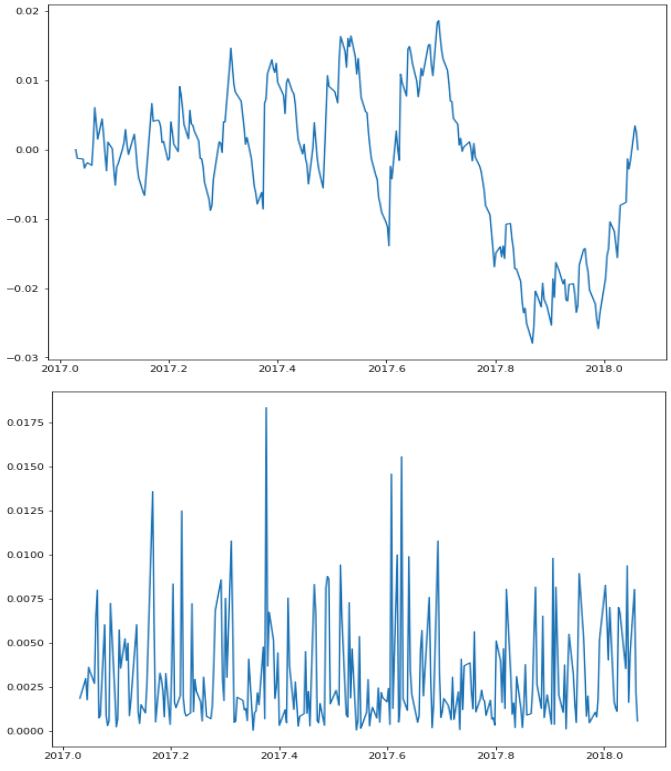


Figure 3. The plot at the top illustrates the volatility series of the S&P500 index in the period ranging from January 2017 to January 2018, whereas the graph at the bottom depicts the series of absolute values of the log-returns of the S&P500 index in the same period. In this case, the self-similarity exponent of the volatility series was found to be equal to 0.55 by the FD4 approach.

As such, the Hurst exponent of the volatility series of a Brownian motion will be considered as a benchmark in order to decide whether there are volatility clusters in the series. More precisely, first, by Monte Carlo simulation, a collection of Brownian motions was generated. For each Brownian motion, the Hurst exponents (by FD4 approach) of their corresponding volatility series were calculated. Hence, we denote by $H_{\text{lim}}(n)$ the value that becomes greater than 90% of those Hurst exponents. Observe that $H_{\text{lim}}(n)$ depends on n, the length of the Brownian motion sample. In fact, for a short series, the accuracy of the FD4 algorithm to calculate the Hurst exponent is lower. Accordingly, the value of $H_{\text{lim}}(n)$ will be higher for a lower value of n. Figure 4 illustrates (for the 90th percentile) how the benchmark given by $H_{\text{lim}}(n)$ becomes lower as the length of the Brownian motion series increases.

Therefore, we will use the following criteria. We say that there are volatility clusters in the series provided that the Hurst exponent of the corresponding volatility series is greater than H_{lim}. Then, we will measure the probability of volatility clusters for subseries of a given length as the ratio between the number of subseries with volatility clusters to the total amount of subseries of the given length.

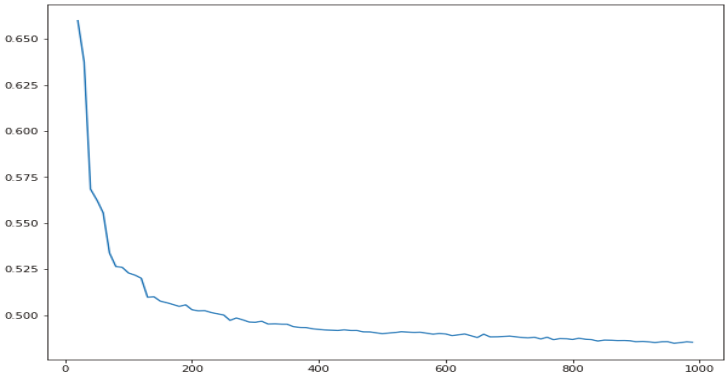


Figure 4. The 90th percentile, $H_{\text{lim}}(n)$, of the Hurst exponents of the volatility series of Brownian motions for several values of the length of the series, n.

In order to check that measure of the probability of volatility clusters, we will test it by artificial processes with volatility clusters of a fixed length (equal to 200 data). A sample from that process is generated as follows. For the first 200 data, generate a sample from a normal distribution $N(0, 0.01)$; for the next 200 data, generate a sample from a normal distribution $N(0, 0.03)$; for the next 200 data, generate a sample from a normal distribution $N(0, 0.01)$, and so on. It is worth pointing out that a mixture of (samples from) normal distributions with distinct standard deviations can lead to (a sample from) a heavy-tailed distribution. Following that example, Figure 5 depicts the distribution of that artificial process with volatility clusters compared to the one from a Gaussian distribution and also to the S&P500 return distribution (rescaled). It is clear that the process is far from Gaussian even in that easy example.

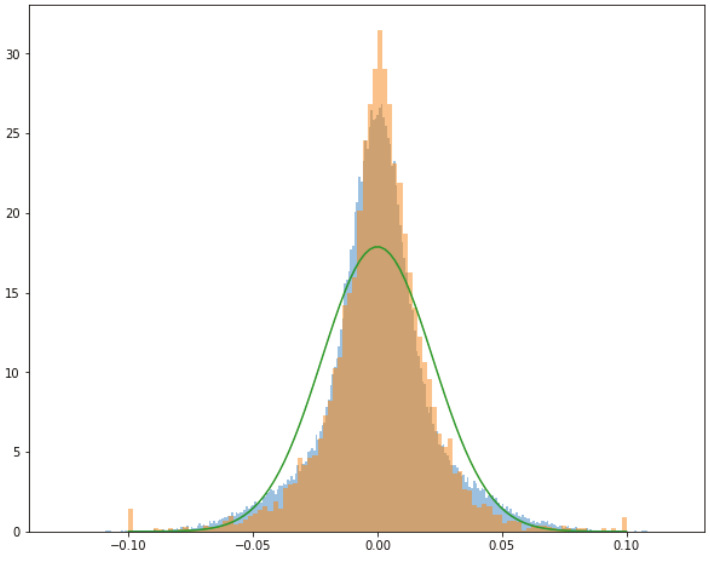


Figure 5. Histogram or density function of: (orange) return distribution of the S&P500 (rescaled and clipped to the interval $[-0.1, 0.1]$); (blue) process with volatility clusters of a fixed length (200 data); (green) normal distribution.

For that process, consider one random block of length 50. It may happen that such a block fully lies in a 200 block of fixed volatility. In this case, there will be no volatility clusters. However, if the first 20 data lie in a block of volatility equal to 0.01, with the remaining 30 data lying in a block of volatility equal to 0.03, then such a block will have volatility clusters. On the other hand, it is clear that if we have one block of length 50 with the first 49 data lying in a block of volatility equal to 0.01, whereas the remaining one datum lies in a block of 0.03 volatility, we cannot say that there are volatility clusters in such a block. Therefore, we shall consider that there are volatility clusters if there are at least 10 data in blocks with distinct volatilities. In other words, we shall assume that we cannot detect clusters with less that 10 data.

On the other hand, note that we are using a confidence level of 90%, and hence, if we get a probability of volatility clusters of, say, x%, that means that there are no volatility clusters regarding the $(100 - x)$% of the blocks of the given size. However, for that confidence level of 90%, we are missing 10% of that $(100 - x)$%, and hence, we will have the following theoretical estimates.

- Theoretical probability of volatility clusters considering clusters of at least 10 data: $(x - 20)/200$.
- Theoretical probability of volatility clusters considering clusters of at least 10 data detected at a confidence level of 90%: $(x - 20)/200 + (1 - (x - 20)/200) \cdot 0.1)$.

Figure 6 graphically shows that the proposed model for estimating the probability of volatility clusters could provide a fair approximation to the actual probability of volatility clusters for such an artificial process.

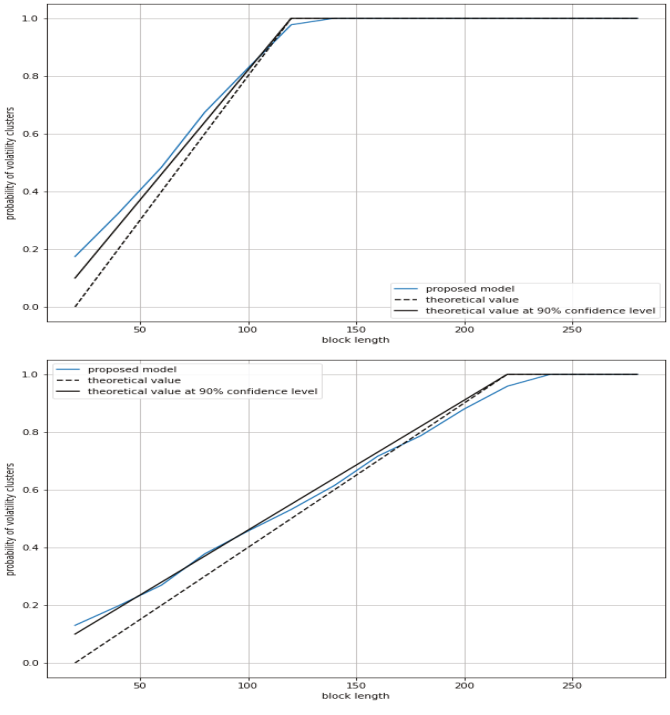


Figure 6. *Cont.*

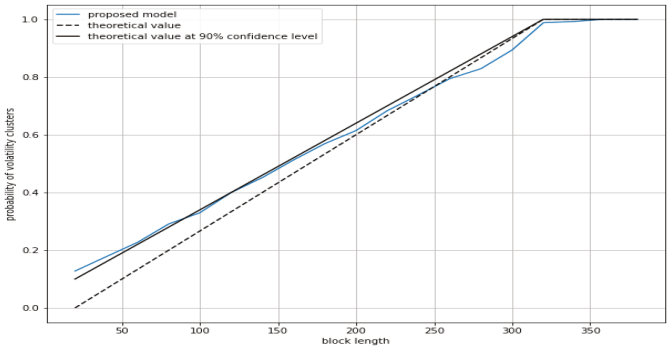


Figure 6. Probability of the existence of volatility clusters for an artificial process with volatility clusters of a fixed length (equal to 100, 200, and 300, from top to bottom).

4. Volatility Clusters in Cryptocurrencies

One of the main characteristics of cryptocurrencies is the high volatility of their exchange rates, and consequently, the high risk associated with their use.

In this section, the methodology provided in Section 3 to calculate the probability of volatility clusters is applied to different financial assets, with a special interest in cryptocurrency markets.

First, Figure 7 shows a similar profile in regard to the probabilities of volatility clusters of an index (S&P500) and a stock (Apple). On the other hand, the probability of volatility clusters of the Euro/USD exchange results in being quite lower.

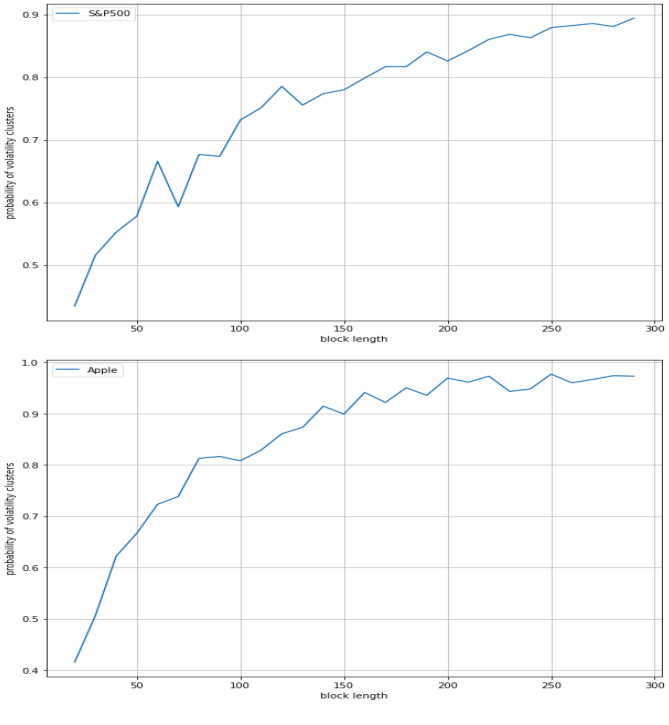


Figure 7. *Cont.*

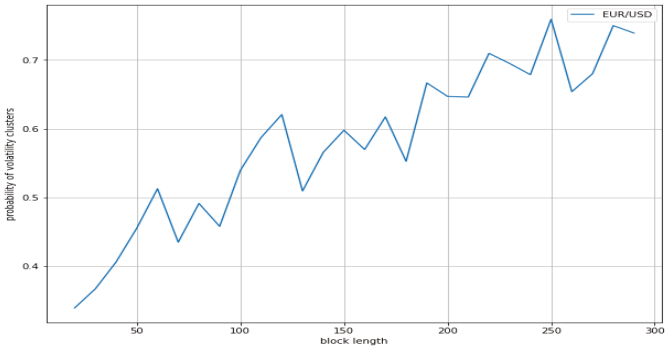


Figure 7. Probabilities of volatility clusters of the next assets. From top to bottom, an index (S&P500), a stock (Apple), and a forex pair (Euro/USD).

On the other hand, Figure 8 depicts the probability of volatility clusters of the three main cryptocurrencies, namely Bitcoin/USD, Ethereum/USD, and Ripple/USD. A similar profile appears for all such cryptocurrencies with the probabilities of their volatility clusters much greater than the ones for the three asset classes displayed in Figure 7.

These results suggest that the volatility in cryptocurrencies changes faster than in traditional assets, and much faster than in forex pairs.

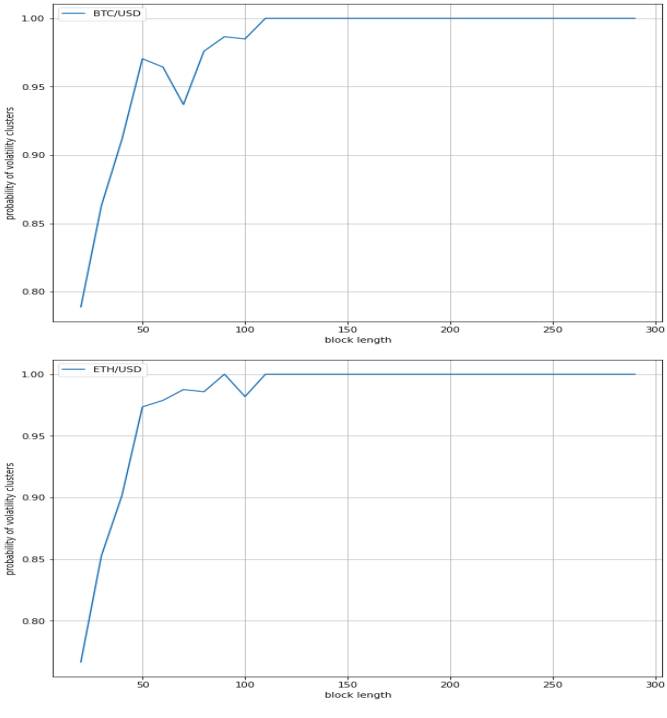


Figure 8. *Cont.*

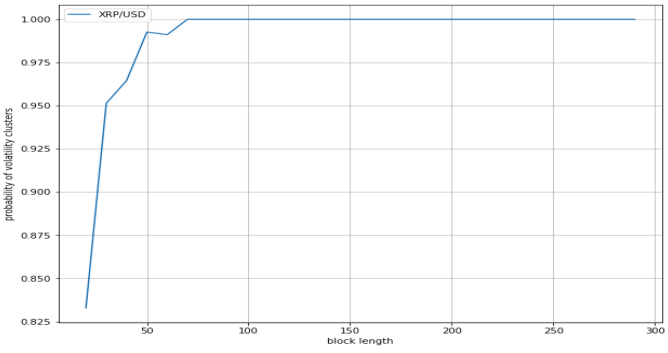


Figure 8. Probabilities of volatility clusters of the following cryptocurrencies. From top to bottom, Bitcoin/USD, Ethereum/USD, and Ripple/USD.

5. Conclusions

One of the main characteristics of cryptocurrencies is the high volatility of their exchange rates. In a previous work, the authors found that a process with volatility clusters displays a volatility series with a high Hurst exponent [2].

In this paper, we provide a novel methodology to calculate the probability of the volatility clusters of a series using the Hurst exponent of its associated volatility series. Our approach, which generalizes the (G)ARCH models, was tested for a class of processes artificially generated with volatility clusters of a given size. In addition, we provided an explicit criterion to computationally determine whether there exist volatility clusters of a fixed size. Interestingly, this criterion is in line with the behavior of the Hurst exponent (calculated by the FD4 approach) of the corresponding volatility series.

We found that the probabilities of volatility clusters of an index (S&P500) and a stock (Apple) show a similar profile, whereas the probability of volatility clusters of a forex pair (Euro/USD) results in being quite lower. On the other hand, a similar profile appears for Bitcoin/USD, Ethereum/USD, and Ripple/USD cryptocurrencies, with the probabilities of volatility clusters of all such cryptocurrencies being much greater than the ones of the three traditional assets. Accordingly, our results suggest that the volatility in cryptocurrencies changes faster than in traditional assets, and much faster than in forex pairs.

Author Contributions: Conceptualization, V.N., J.E.T.S., M.F.-M., and M.A.S.-G.; methodology, V.N., J.E.T.S., M.F.-M., and M.A.S.-G.; validation, V.N., J.E.T.S., M.F.-M., and M.A.S.-G.; formal analysis, V.N., J.E.T.S., M.F.-M., and M.A.S.-G.; writing—original draft preparation, V.N., J.E.T.S., M.F.-M., and M.A.S.-G.; writing—review and editing, V.N., J.E.T.S., M.F.-M., and M.A.S.-G.; These authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: Both J.E. Trinidad Segovia and M.A. Sánchez-Granero are partially supported by Ministerio de Ciencia, Innovación y Universidades, Spain, and FEDER, Spain, Grant PGC2018-101555-B-I00, and UAL/CECEU/FEDER, Spain, Grant UAL18-FQM-B038-A. Further, M.A. Sánchez-Granero acknowledges the support of CDTIME. M. Fernández-Martínez is partially supported by Ministerio de Ciencia, Innovación y Universidades, Spain, and FEDER, Spain, Grant PGC2018-097198-B-I00, and Fundación Séneca of Región de Murcia (Murcia, Spain), Grant 20783/PI/18.

Acknowledgments: The authors would also like to express their gratitude to the anonymous reviewers whose suggestions, comments, and remarks allowed them to enhance the quality of this paper.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results.

References

1. Tseng, J.-J.; Li, S.-P. Asset returns and volatility clustering in financial time series. *Phys. A Stat. Mech. Appl.* **2011**, *390*, 1300–1314. [CrossRef]
2. Trinidad Segovia, J.E.; Fernández-Martínez, M.; Sánchez-Granero, M.A. A novel approach to detect volatility clusters in financial time series. *Phys. A Stat. Mech. Appl.* **2019**, *535*, 122452. [CrossRef]
3. Danielsson, J.; Valenzuela, M.; Zer, I. Learning from History: Volatility and Financial Crises. *Rev. Financ.* **2018**, *21*, 2774–2805. [CrossRef]
4. Valenti, D.; Fazio, G.; Spagnolo, B. The stabilizing effect of volatility in financial markets. *Phys. Rev. E* **2018**, *97*, 062307. [CrossRef]
5. Engle, R.F. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* **1982**, *50*, 987–1007. [CrossRef]
6. Bollerslev, T. Generalized autoregressive conditional heteroskedasticity. *J. Econom.* **1986**, *31*, 307–327. [CrossRef]
7. Taylor, S.J. *Modelling Financial Time Series*; John Wiley & Sons, Ltd.: Chichester, UK, 1986.
8. Kim, Y.S.; Rachev, S.T.; Bianchi, M.L.; Fabozzi, F.J. Financial market models with Lévy processes and time-varying volatility. *J. Bank. Financ.* **2008**, *32*, 1363–1378. [CrossRef]
9. Engle, R.F.; Bollerslev, T. Modelling the persistence of conditional variances. *Econom. Rev.* **1986**, *5*, 1–50. [CrossRef]
10. Baillie, R.T.; Bollerslev, T.; Mikkelsen, H.O. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *J. Econom.* **1996**, *74*, 3–30. [CrossRef]
11. Bentes, S.R. Long memory volatility of gold price returns: How strong is the evidence from distinct economic cycles? *Phys. A Stat. Mech. Appl.* **2016**, *443*, 149–160. [CrossRef]
12. Patterson, G.; Sornette, D.; Parisi, D. Properties of balanced flows with bottlenecks: Common stylized facts in finance and vibration-driven vehicles. *Phys. Rev. E* **2020**, *101*, 042302. [CrossRef]
13. Biondo, A.E. Order book microstructure and policies for financial stability. *Stud. Econ. Financ.* **2018**, *35*, 196–218. [CrossRef]
14. Biondo, A.E. Order book modeling and financial stability. *J. Econ. Interact. Coord.* **2019**, *14*, 469–489. [CrossRef]
15. Sueshige, T.; Sornette, D.; Takayasu, H.; Takayasu, M. Classification of position management strategies at the order-book level and their influences on future market-price formation. *PLoS ONE* **2019**, *14*, e0220645. [CrossRef] [PubMed]
16. Lux, T.; Marchesi, M. Volatility clustering in financial markets: A microsimulation of interacting agents. *Int. J. Theor. Appl. Financ.* **2000**, *3*, 675–702. [CrossRef]
17. Raberto, M.; Cincotti, S.; Focardi, S.M.; Marches, M. Agent-based simulation of a financial market. *Phys. A Stat. Mech. Appl.* **2001**, *299*, 319–327. [CrossRef]
18. Krawiecki, A.; Hołyst, J.A.; Helbing, D. Volatility Clustering and Scaling for Financial Time Series due to Attractor Bubbling. *Phys. Rev. Lett.* **2002**, *89*, 158701. [CrossRef]
19. Szabolcs, M.; Farmer, J.D. An empirical behavioral model of liquidity and volatility. *J. Econ. Dyn. Control* **2008**, *32*, 200–234
20. Alfarano, S.; Lux, T.; Wagner, F. Estimation of Agent-Based Models: The Case of an Asymmetric Herding Model. *Comput. Econ.* **2005**, *26*, 19–49 [CrossRef]
21. Cont, R. Volatility Clustering in Financial Markets: Empirical Facts and Agent-Based Models. In *Long Memory in Economics*; Teyssière, G., Kirman, A.P., Eds.; Springer: Berlin/Heidelberg, Germany, 2007.
22. Chen, J.J.; Zheng, B.; Tan, L. Agent-Based Model with Asymmetric Trading and Herding for Complex Financial Systems. *PLoS ONE* **2013**, *8*, e79531. [CrossRef]
23. He, X.Z.; Li, K.; Wang, C. Volatility clustering: A nonlinear theoretical approach. *J. Econ. Behav. Organ.* **2016**, *130*, 274–297. [CrossRef]
24. Schmitt, N.; Westerhoff, F. Herding behavior and volatility clustering in financial markets. *Quant. Financ.* **2017**, *17*, 1187–1203. [CrossRef]

25. Chen, J.-J.; Tan, L.; Zheng, B. Agent-based model with multi-level herding for complex financial systems. *Sci. Rep.* **2015**, *5*, 8399. [CrossRef] [PubMed]
26. Shi, Y.; Luo, Q.; Li, H. An Agent-Based Model of a Pricing Process with Power Law, Volatility Clustering, and Jumps. *Complexity* **2019**, *2019*, 3429412. [CrossRef]
27. Verma, A.; Buonocore, R.J.; di Matteo, T. A cluster driven log-volatility factor model: A deepening on the source of the volatility clustering. *Quant. Financ.* **2018**, 1–16. [CrossRef]
28. Barde, S. Direct comparison of agent-based models of herding in financial markets. *J. Econ. Dyn. Control* **2016**, *73*, 329–353 [CrossRef]
29. CoinMarketCap, 2020. Available online: https://coinmarketcap.com/all/views/all/ (accessed on 24 March 2020).
30. Dimitrova, V.; Fernández-Martínez, M.; Sánchez-Granero, M.A.; Trinidad Segovia, J.E. Some comments on Bitcoin market (in)efficiency. *PLoS ONE* **2019**, *14*, e0219243. [CrossRef]
31. Letra, I. What Drives Cryptocurrency Value? A Volatility and Predictability Analysis. 2016. Available online: https://www.repository.utl.pt/handle/10400.5/12556 (accessed on 24 March 2020).
32. Bouoiyour, J.; Selmi, R. Bitcoin: A beginning of a new phase? *Econ. Bull.* **2016**, *36*, 1430–1440.
33. Bouri, E.; Azzi, G.; Dyhrberg, A.H. On the return-volatility relationship in the Bitcoin market around the price crash of 2013. *Economics* **2017**, *11*, 1–16.
34. Balcilar, M.; Bouri, E.; Gupta, R.; Roubaud, D. Can volume predict Bitcoin returns and volatility? A quantiles-based approach. *Econ. Model.* **2017**, *64*, 74–81. [CrossRef]
35. Baur, D.G.; Hong, K.; Lee, A.D. Bitcoin: Medium of exchange or speculative assets? *J. Int. Financ. Mark. Inst. Money* **2018**. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2561183 (accessed on 1 May 2020)
36. Bariviera, A.F.; Basgall, M.J.; Hasperué, W.; Naiouf, M. Some stylized facts of the Bitcoin market. *Physics A* **2017**, *484*, 82–90. [CrossRef]
37. Phillip, A.; Chan, J.S.K.; Peiris, S. A new look at cryptocurrencies. *Econ. Lett.* **2018**, *163*, 6–9. [CrossRef]
38. Zhang, W.; Wang, P.; Li, X.; Shen, D. Some stylized facts of the cryptocurrency market. *Appl. Econ.* **2018**, *50*, 5950–5965. [CrossRef]
39. Kancs, D.; Rajcaniova, M.; Ciaian, P. *The Price of Bitcoin: GARCH Evidence from High Frequency Data*; 29598 EN; Publications Office of the European Union: Luxembourg, 2019; ISBN 978-92-7998570-6, JRC115098. [CrossRef]
40. Mandelbrot, B.B. (Ed.) *Gaussian Self-Affinity and Fractals*; Springer-Verlag: New York, NY, USA, 2002.
41. Fernández-Martínez, M.; Sánchez-Granero, M.A.; Trinidad Segovia, J.E.; Román-Sánchez, I.M. An accurate algorithm to calculate the Hurst exponent of self-similar processes. *Phys. Lett. A* **2014**, *378*, 2355–2362. [CrossRef]
42. Sánchez-Granero, M.A.; Trinidad Segovia, J.E.; García Pérez, J. Some comments on Hurst exponent and the long memory processes on capital markets. *Phys. A Stat. Mech. Appl.* **2008**, *387*, 5543–5551. [CrossRef]
43. Trinidad Segovia, M.; Fernández-Martínez, J.E.; Sánchez-Granero, M.A. A note on geometric method-based procedures to calculate the Hurst exponent, *Phys. A Stat. Mech. Appl.* **2012**, *391*, 2209–2214. [CrossRef]
44. Sánchez-Granero, M.A.; Fernández-Martínez, M.; Trinidad Segovia, J.E. Introducing fractal dimension algorithms to calculate the Hurst exponent of financial time series. *Eur. Phys. J. B* **2012**, 85–86. [CrossRef]
45. Fernaández-Martiínez, M.; Guirao, J.L.G.; Sánchez-Granero, M.A.; Trinidad Segovia, J.E. *Fractal Dimension for Fractal Structures: With Applications to Finance*, 1st ed.; Springer Nature: Cham, Switzerland, 2019; pp. 1–204.

mathematics

MDPI

Article

A Proposal to Fix the Number of Factors on Modeling the Dynamics of Futures Contracts on Commodity Prices †

Andrés García-Mirantes [1], Beatriz Larraz [2,*] and Javier Población [3]

[1] IES Juan del Enzina, 24001 Leon, Spain; andres.web.publicar@gmail.com
[2] Statistics Department/Faculty of Law and Social Sciences, Universidad de Castilla-La Mancha, 45071 Toledo, Spain
[3] Banco de España, 28014 Madrid, Spain; Javier.poblacion@bde.es
* Correspondence: beatriz.larraz@uclm.es; Tel.: +34-925-26-88-00
† This paper should not be reported as representing the views of the Banco de España (BdE) or European Central Bank (ECB). The views expressed herein are those of the authors and should not be attributed to the BdE or ECB.

Received: 8 May 2020; Accepted: 11 June 2020; Published: 14 June 2020

Abstract: In the literature on modeling commodity futures prices, we find that the stochastic behavior of the spot price is a response to between one and four factors, including both short- and long-term components. The more factors considered in modeling a spot price process, the better the fit to observed futures prices—but the more complex the procedure can be. With a view to contributing to the knowledge of how many factors should be considered, this study presents a new way of computing the best number of factors to be accounted for when modeling risk-management of energy derivatives. The new method identifies the number of factors one should consider in the model and the type of stochastic process to be followed. This study aims to add value to previous studies which consider principal components by assuming that the spot price can be modeled as a sum of several factors. When applied to four different commodities (weekly observations corresponding to futures prices traded at the NYMEX for WTI light sweet crude oil, heating oil, unleaded gasoline and Henry Hub natural gas) we find that, while crude oil and heating oil are satisfactorily well-modeled with two factors, unleaded gasoline and natural gas need a third factor to capture seasonality.

Keywords: commodity prices; futures prices; number of factors; eigenvalues

1. Introduction

Forecasting is not a highly regarded activity for economists and financiers. For some, it evokes images of speculators, chart analysts and questionable investor newsletters. For others, there are memories of the grandiose econometric forecasting failures of the 1970's. Nevertheless, there is a need for forecasting in risk management. A prudent corporate treasurer or fund manager must have some way of measuring the risk of earnings, cash flows or returns. Any measure of risk must incorporate some estimate of the probability distribution of the futures asset prices on which financial performance depends. Consequently, forecasting is an indispensable element of prudent financial management.

When a company is planning to develop a crude oil or natural gas field, the investment is significant, and production usually lasts many years. However, there must be an initial investment for there to be any return (see, for example, [1,2], among others). Assuming that futures values are not known after a certain date because there is no trade, it makes it difficult to measure the risk of these projects. Since commodities (crude oil, gas, gasoline, etc.) are physical assets, their price dynamic is much more complex than financial assets because their prices are affected by storage and transportation

cost (cost of carry). Due to such complexity, in order to model this price dynamic we need factor models such as in [3–9]. In addition, in the transport sector [10] and [11] use different factor models for modeling bulk shipping prices and freight prices.

In order to measure exposure to price risk due to a single underlying asset, it is necessary to know the dynamics of the term structure of asset prices. Specifically, the value-at-risk (VaR, [12]) of the underlying asset price, the most widely known measure of market risk [13], is characterized by knowing the stochastic dynamic of the price, the volatility of the price and the correlation of different prices at different times. For these reasons, to date, the behavior of commodity prices has been modeled under the assumption that the spot price and/or the convenience yield of the commodity follow a stochastic process.

In the literature we find that the spot price is considered as the sum of both short-term and long-term components (see, for example, [14,15]). Short-term factors account for the mean reverting components in commodity prices, while long-term factors account for the long-term dynamics of commodity prices, assuming they follow a random walk. Sometimes a deterministic seasonal component needs to be added [16].

Following this approach, some multifactor models have been proposed in the literature. Focusing on the number of factors initially considered, [17] developed a two-factor model to value oil-linked assets. Later, [14] planned a one-factor model, two-factor model and a three-factor model, adding stochastic interest rates to the previous factors. This was superseded by a new formulation which appeared in [15], enhancing the latter article and developing a short-term/long-term model. [18] added the long-term spot price return as a third risk factor. Finally, [19] offered researchers a general N-factor model.

At this point, it should be stressed that the decision regarding the number of factors to be used in the model needs to be made a priori. According to the above literature consulted, the models are usually planned with two, three or four factors. However, in this study, the need to assume a fixed number of factors in the model is discounted. We propose a new method that identifies the number of factors one should consider in the model and the type of stochastic process to be followed. This method avoids the necessity of inaccurately suggesting a concrete number of factors in the model. This is very useful for researchers and practitioners because the optimal number of factors could change, depending on the accuracy needed in each problem. Clearly, if we do not use the optimal number of factors in modeling the commodity price dynamics, the results will not be optimal.

To the best of our knowledge, there are three previous studies applying principal component analysis [20] to the modeling of commodity futures price dynamics [21–23]. However, they only model the futures prices dynamic and ignore the dynamic followed by the spot price and, consequently carrying the risk of being incoherent, since futures price are the spot price expected value under the Q measure.

This study aims to add value to previous contributions by assuming that is the spot price can be modeled as a sum of several factors (long term and short term, seasonality, etc.). Therefore, since it is widely accepted (see, for example, [24]) that the futures price is the spot price expected value under the Q measure ($F_{t,T} = E^*[S_{t+T}|I_t]$), where S_{t+T} is the spot price at time $t+T$, I_t is the information available at time t and $E^*[\,]$ is the expected value under the Q measure.), from the variance–covariance matrix of the futures prices we can deduce the best structure for modelling the spot prices dynamic.

The remainder of this study is organized as follows. Section 2 presents a general theoretical model and explains the methodology proposed to set an optimal set of factors. In Section 3, we describe the datasets used to show the methodology and these results are described. Finally, Section 4 sets out the conclusions.

2. Theoretical Model

2.1. Theoretical Model

In the main literature to date (for example, [19]), it is assumed that the commodity log spot price is the sum of several stochastic factors: $S_t = \exp(\mathbf{C}\mathbf{X}_t)$, $t = 0, \ldots, n$ where the vector of state variables $\mathbf{X}_t = (x_{1t}, \ldots, x_{Nt})$ follows the process: $d\mathbf{X}_t = \mathbf{M}dt + \mathbf{A}\mathbf{X}_t dt + \mathbf{R}d\mathbf{W}_t$, being $\mathbf{C}$, $\mathbf{M}$, $\mathbf{A}$ and $\mathbf{R}$ vectors' and matrices' parameters.

It is widely accepted that, for the model to be identifiable, some restrictions must be imposed. This means that if we assume that $\mathbf{A}$ is diagonalizable and all its eigenvalues are real (a different formula is available if some are complex), we can take $\mathbf{C} = (1, \ldots, 1)$, $\mathbf{M}' = (\mu, 0, \ldots, 0)$ and $\mathbf{A} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & k_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & k_N \end{pmatrix}$, with $k_i, i = 1, \ldots, N$ the eigenvalues and $k_1 = 0$, by simply changing the state space basis. Therefore, we already have $\mathbf{M}$, $\mathbf{A}$ and $\mathbf{C}$.

It is also easy to prove that as $d\mathbf{W}_t$ is a $N \times 1$ vector of correlated Brownian motion increments, $\mathbf{R}$ can be assumed as $\mathbf{R} = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N \end{pmatrix}$. Note that $\mathbf{R}$ is not important, but the product $\mathbf{R}\mathbf{R}'$ is what appears in all formulae. In fact, it can be proved that any factorization of $\mathbf{R}\mathbf{R}'$ corresponds to a different definition of the noise, so we can safely take $\mathbf{R}$ as any Choleski factorization of $(\mathbf{R}\mathbf{R}')$. In the Black–Scholes world (risk-neutral world), knowing the real dynamics, the risk neutral one is $d\mathbf{X}_t = \mathbf{M}^* dt + \mathbf{A}\mathbf{X}_t dt + \mathbf{R}d\mathbf{W}_t^*$ where $\mathbf{M}^* = \mathbf{M} - \boldsymbol{\lambda}$ being $\boldsymbol{\lambda}' = (\lambda_1, \lambda_2, \ldots, \lambda_N)$ the vector formed from each state variable's risk premium).

Following [25], the futures price is given by $F_{t,T} = \exp\left(g(T) + \mathbf{C}e^{\mathbf{A}T}\mathbf{X}_t\right)$, where we know explicitly $g(T) = \mathbf{C}\int_0^T e^{\mathbf{A}(T-s)}\mathbf{M}^* ds + C\left(\int_0^T e^{\mathbf{A}(T-s)}\mathbf{R}\mathbf{R}'\left(e^{\mathbf{A}(T-s)}\right)' ds\right)\mathbf{C}'$ and where both $g(T)$ and $C(T) = e^{\mathbf{A}T}$ are known deterministic functions independent of t and $\mathbf{X}_t$ is a stochastic process with known dynamics.

Defining in a more compact form, we have:

$$\begin{cases} d\mathbf{X}_t = (\mathbf{M} + \mathbf{A}\mathbf{X}_t)dt + \mathbf{R}d\mathbf{W}_t \\ F_{t,T} = \exp[\delta(T) + \phi(T)\mathbf{X}_t + \varphi(T)\mathbf{M}^* + \boldsymbol{\varepsilon}_{t,T}] \end{cases}$$

2.2. A General Procedure to Determine the Stochastic Factors

In the previous subsection, we have presented the general model for characterizing the commodity price dynamics based on the assumption that the log commodity spot price is the sum of several factors. However, to the best of the authors' knowledge, the optimal number of stochastic factors has not yet been studied, for these models.

This subsection presents a theoretical procedure to establish the optimal number of factors. It also presents a way to determine how those factors should be aligned (long-term, short-term, seasonal, etc.).

To address this problem, let us suppose that there are M futures maturities and n observations of the forward curve, that is, the matrix $\mathbf{U} = \log\left(F_{t,T_i}\right)$, $t = 0, \ldots, n$; $i = 1, \ldots, M$ has dimension $M \times (n+1)$. We further assume, as usual, that $n \gg M$. To determine the optimal number of stochastic factors needed to characterize the commodity price dynamic in the best way, first we must realize that the number of factors is equal to $rank(\mathbf{R})$ and, from the previous expression, $rank(\mathbf{R})$ has to be equal to the rank of the variance–covariance matrix of $\mathbf{U}$. If, as usual, the process $\mathbf{X}_t$ has a unit root, so it is non-stationary and the variance and covariances are infinity, we need another matrix to determine the rank of the variance–covariance matrix of $\mathbf{U}$.

If we define volatility (instantaneous variance) as $\sigma_{T_i}^2 = \lim_{h\to 0} \frac{Var\left(\log F_{t+h,T_i} - \log F_{t,T_i}\right)}{h}$, $i = 1, \ldots, M$ and cross-volatility (instantaneous covariance) as $\sigma_{T_i,T_j} = \lim_{h\to 0} \frac{Cov\left(\log F_{t+h,T_i} - \log F_{t,T_i}, \log F_{t+h,T_j} - \log F_{t,T_j}\right)}{h}$, $i, j = 1, \ldots, M$ (as expected, $\sigma_{T_i}^2 = \sigma_{T_i,T_j}$), we have the necessary matrix. Although we cannot compute the limit from the data, we can set h as the shortest time period available and estimate it directly as $\hat{\sigma}_{T_i,T_j} = \left[\frac{\hat{Cov}\left(\log F_{t+h,T_i} - \log F_{t,T_i}, \log F_{t+h,T_j} - \log F_{t,T_j}\right)}{h}\right]$, where $\hat{Cov}$ is the sample covariance.

We thus define the matrix $\mathbf{\Theta} = \left(\Theta_{ij}\right)$ (dim $M \times M$) as $\Theta_{ij} = \sigma_{T_i,T_j}$, $i, j = 1, \ldots, M$. We can estimate it directly from our database and we can also estimate its rank. Once we have this rank, as stated above $rank(\mathbf{\Theta}) = rank(\mathbf{R}) = N$, we know the number of stochastic factors (N) that define the commodity price dynamics.

From a practical point of view, however, if we follow this procedure as explained above, unless one futures maturity is a linear combination of the rest (which is not likely), we obtain $rank(\mathbf{\Theta}) = rank(\mathbf{R}) = N$. Nevertheless, the weights of these factors are going to be different and most of them will have an insignificant weight.

Fortunately, from this procedure, we can also estimate the eigenvalues $k_1, \ldots, k_N$ and, from there, determine the factor weight through the eigenvalues' relative weight. We can estimate the eigenvalues of $\mathbf{A}$ via a nonlinear search procedure by using the fact that $\sigma_{T_i,T_j} = \mathbf{C}e^{\mathbf{A}T_i}\mathbf{R}\mathbf{R}'\left(e^{\mathbf{A}T_j}\right)'\mathbf{C}'$ (see García et al. 2008) and therefore, $\mathbf{\Theta}$ can be expressed as $\mathbf{\Theta} = \mathbf{C}\begin{pmatrix} e^{\mathbf{A}T_1} & \cdots & e^{\mathbf{A}T_M} \end{pmatrix}\mathbf{R}\mathbf{R}'\begin{pmatrix} e^{\mathbf{A}T_1} \\ \vdots \\ e^{\mathbf{A}T_M} \end{pmatrix}'\mathbf{C}'$. σ_{T_i,T_j} is a linear combination of products of $e^{k_1T}, \ldots, e^{k_NT}$. In other words, if $k_1, \ldots, k_N$ are the eigenvalues of $\mathbf{A}$, $e^{k_1T}, \ldots, e^{k_NT}$ must be the eigenvalues of $\mathbf{\Theta}$.

Moreover, from the eigenvalues of matrix $\mathbf{A}$, it is also easy to determine the factors. Taking into account that factors´ Stochastic Differential Equation (SDE) is $d\mathbf{X}_t = \mathbf{M}dt + \mathbf{A}\mathbf{X}_t dt + \mathbf{R}d\mathbf{W}_t$, if, for example, the eigenvalue is $k = 0$, the factor is a long-term one because the SDE associated with this factor is a random walk (General Brownian Motion (GBM)): $dx_{it} = \mu_i dt + \sigma_i dW_{it}$. On the other hand, if the eigenvalue is $k \in (-1, 0)$, the factor is a short-term one because the SDE associated with this factor is an Ornstein–Uhlenbeck: $dx_{it} = \lambda x_{it}dt + \sigma_i dW_{it}$. If the eigenvalue is complex, the factor is a seasonal one.

From a practical point of view, when we carry out this procedure we get N eigenvalues and we need to decide how many of them to optimally choose. The way to decide this is through the relative weight of the eigenvalues. By normalizing the largest one to 1, the smallest eigenvalues represent negligible factors. This allows us to decide how many factors must be optimally chosen.

In order to clarify concepts, the following example could be useful, if we have $M = 9$ futures with maturities at times $T_1, \ldots, T_9$. The method is as follows.

1. Compute $\hat{\Theta}_{ij} = \left[\frac{\hat{Cov}\left(\log F_{t+h,T_i} - \log F_{t,T_i}, \log F_{t+h,T_j} - \log F_{t,T_j}\right)}{h}\right]$.
2. Compute the rank of $\hat{\Theta}$. Let us assume that this is 3.
3. As a result, we have three eigenvalues k_1, k_2 and k_3. It is usual to assume that $k_1 = 0$ as the futures process is not stationary, but k_1 can nevertheless be estimated. If we do assume it, however, we obtain that σ_{T_i,T_j} is a linear combination of the products of $e^{0T} = 1$, e^{k_2T} and e^{k_3T}. Therefore, we obtain the general equation $\Theta_{ij} = \alpha_{11} + \alpha_{12}e^{k_2T_j} + \alpha_{13}e^{k_3T_j} + \alpha_{21}e^{k_2T_i} + \alpha_{31}e^{k_3T_i} + \alpha_{22}e^{k_2(T_i+T_j)} + \alpha_{23}e^{k_2T_i+k_3T_j} + \alpha_{23}e^{k_2T_j+k_3T_i} + \alpha_{33}e^{k_3(T_i+T_j)}$ which can be estimated numerically as:
    a. Select an initial estimate of (k_2, k_3).
    b. Regress $\hat{\Theta}_{ij}$ and compute the error.

c. Iteratively select another estimate of (k_2, k_3) and get back to b.

To the best of the authors' knowledge, no method has combined the knowledge of this concrete specification $G = \Phi(T)$ with a nonlinear search procedure to identify factors, which is one of the contributions made by this article.

Once we have determined the optimal number and form of the stochastic factors to characterize the commodity price dynamics, we can estimate model parameters using standard techniques. The Kalman filter (see, for example, [26]) uses a complex calibration technique. Other techniques include approximations such as [18] or [27]. Finally, the recently published option by [28] presents an optimal way of estimating model parameters by avoiding the use of the Kalman filter. Model parameters are estimated in the papers and so, for the sake of brevity we do not estimate the parameters in this study.

3. Data and Main Results

3.1. Data

In this subsection, we briefly describe the datasets used in this study. The datasets include weekly observations corresponding to futures prices for four commodities: WTI light sweet crude oil, heating oil, unleaded gasoline (RBOB) and Henry Hub natural gas. These futures were taken into consideration because they are the most representative and classic among the products. They are futures with many historical series and futures at many maturities. Therefore, they are considered as ideal for studying the optimal number of factors that should be chosen.

In this study, two data sets were considered for each commodity. Data set 1 contains less futures maturities, but more years of observations considered while data set 2 contains more futures maturities, but less years of observations. For dataset 1 (Table 1A), related to WTI crude oil, it comprised contracts from 4 September 1989 to 3 June 2013 (1240 weekly observations) for futures maturities from F1 to F17, F1 being the contract for the month closest to maturity, F2 the contract for the second-closest month to maturity, etc. In the case of heating oil, it contained contracts from 21 January 1991 to 3 June 2013 (1168 weekly observations) for futures maturities from F1 to F15. Meanwhile, RBOB gasoline first data set comprised contracts from 3 October 2005 to 3 June 2013 (401 weekly observations) for futures maturities from F1 to F12 and in the case of Henry Hub natural gas, it contained contracts from 27 January 1992 to 3 June 2013 (1115 weekly observations) for futures maturities from F1 to F16.

Looking at the dataset 2 (Table 1B), in the case of WTI crude oil, it comprised contracts from 18 September 1995 to 3 June 2013 (925 weekly observations) for futures maturities from F1 to F28 and in the case of heating oil it comprised contracts from 9 September 1996 to 3 June 2013 (874 weekly observations) for futures maturities from F1 to F18. In the meantime, RBOB gasoline comprised contracts from 2 February 2007 to 3 June 2013 (330 weekly observations) for futures maturities from F1 to F36 and, to end with, in the case of Henry Hub natural gas, dataset 2 (Table 1B) contained contracts from 24 March 1997 to 3 June 2013 (856 weekly observations) for futures maturities from F1 to F36.

Table 1 shows the main descriptive statistics of the futures, particularly the mean and volatility, for each dataset. It is interesting to note that the lack of low-cost transportation and the limited storability of natural gas made its supply unresponsive to seasonal variation in demand. Thus, natural gas prices were strongly seasonal [3]. The unleaded gasoline was also seasonal.

Table 1. Descriptive statistics.

(A) Dataset 1								
	WTI Crude Oil		**Gasoline**		**Natural Gas**		**Heating Oil**	
	Mean ($/bbl)	**Volatility (%)**	**Mean ($/bbl)**	**Volatility (%)**	**Mean ($/MMBtu)**	**Volatility (%)**	**Mean ($/bbl)**	**Volatility (%)**
F1	43.1	30%	96.7	32%	4.2	45%	63.6	28%
F2	43.3	27%	96.5	30%	4.3	40%	63.8	26%
F3	43.3	25%	96.4	29%	4.3	36%	64.0	24%
F4	43.4	24%	96.3	27%	4.4	32%	64.1	23%
F5	43.3	23%	96.1	27%	4.4	29%	64.1	22%
F6	43.3	22%	95.9	26%	4.4	27%	64.2	21%
F7	43.3	21%	95.7	25%	4.5	26%	64.2	20%
F8	43.2	20%	95.6	26%	4.5	24%	64.1	19%
F9	43.2	20%	95.6	25%	4.5	24%	64.1	19%
F10	43.1	19%	95.4	26%	4.5	22%	64.1	18%
F11	43.1	19%	95.5	25%	4.5	22%	64.1	18%
F12	43.0	18%	95.5	25%	4.5	21%	64.0	17%
F13	43.0	18%			4.5	20%	63.7	17%
F14	42.9	18%			4.5	20%	63.4	17%
F15	42.9	17%			4.5	20%	63.0	17%
F16	42.8	17%			4.5	20%		
F17	42.8	17%						
(B) Dataset 2								
	WTI Crude Oil		**Gasoline**		**Natural Gas**		**Heating oil**	
	Mean ($/bbl)	**Volatility (%)**	**Mean ($/bbl)**	**Volatility (%)**	**Mean ($/MMBtu)**	**Volatility (%)**	**Mean ($/bbl)**	**Volatility (%)**
F1	50.9	30%	101.1	32%	4.9	45%	53.3	30%
F2	51.2	28%	100.6	30%	5.0	40%	53.5	28%
F3	51.3	26%	100.3	30%	5.1	38%	53.6	26%
F4	51.3	25%	100.1	28%	5.1	33%	53.7	25%
F5	51.3	24%	99.8	28%	5.2	31%	53.7	24%
F6	51.3	23%	99.6	27%	5.2	28%	53.7	23%
F7	51.3	22%	99.3	26%	5.3	27%	53.7	22%
F8	51.3	21%	99.2	26%	5.3	26%	53.7	21%
F9	51.2	21%	99.1	25%	5.3	25%	53.7	21%
F10	51.2	20%	99.1	27%	5.3	24%	53.6	20%
F11	51.1	20%	99.1	26%	5.3	22%	53.6	19%
F12	51.1	19%	99.1	26%	5.3	21%	53.5	19%
F13	51.0	19%	99.1	26%	5.3	21%	53.3	19%
F14	50.9	19%	99.0	25%	5.3	21%	53.0	19%
F15	50.8	18%	98.9	25%	5.3	21%	53.0	18%
F16	50.8	18%	98.8	23%	5.3	20%	53.5	18%
F17	50.7	18%	98.5	24%	5.3	20%	55.4	18%
F18	50.7	18%	98.3	23%	5.3	19%	58.4	18%
F19	50.6	17%	98.1	23%	5.3	20%		
F20	50.5	17%	98.0	23%	5.3	19%		
F21	50.5	17%	98.0	23%	5.3	19%		
F22	50.4	17%	97.9	24%	5.3	20%		
F23	50.4	17%	97.9	23%	5.2	18%		
F24	50.3	17%	97.9	24%	5.2	18%		
F25	50.3	16%	97.9	24%	5.2	18%		
F26	50.2	16%	97.8	23%	5.2	18%		
F27	50.2	16%	97.8	24%	5.2	18%		
F28	50.1	16%	97.7	23%	5.2	18%		
F29			97.6	23%	5.2	18%		
F30			97.5	22%	5.2	18%		
F31			97.4	22%	5.2	18%		
F32			97.4	23%	5.2	19%		
F33			97.3	22%	5.2	18%		
F34			97.1	23%	5.2	19%		
F35			97.0	23%	5.2	18%		
F36			97.0	23%	5.2	17%		

3.2. Main Results

We now present the results after applying the method proposed to the 4 commodities (2 datasets per commodity) described above in order to select the number of factors to model the behavior of commodity prices. The results correspond to the eigenvalues in decreasing order, the percentage of the overall variability that they explain and the cumulative proportion of explained variance. These are reported in Tables 2–5.

Table 2. Eigenvalues for both datasets of the WTI light sweet crude oil.

Dataset 1			Dataset 2		
Eigenvalues	**Percentage of Total Variance**	**Cumulative Variance (%)**	**Eigenvalues**	**Percentage of Total Variance**	**Cumulative Variance (%)**
100	99.6713	99.6713	100	99.5448	99.5448
0.3202	0.3191	99.9904	0.4428	0.4408	99.9855
0.0084	0.0084	99.9988	0.0126	0.0125	99.9980
0.0010	0.0010	99.9998	0.0017	0.0017	99.9997
0.0001	0.0001	99.9999	0.0002	0.0002	99.9998
2.9905×10^{-5}	2.9806×10^{-5}	100	0.0001	0.0001	99.9999
1.2209×10^{-5}	1.2169×10^{-5}	100	3.7318×10^{-5}	3.7148×10^{-5}	100
5.4907×10^{-6}	5.4727×10^{-6}	100	1.6898×10^{-5}	1.6821×10^{-5}	100
2.7838×10^{-6}	2.7746×10^{-6}	100	8.7477×10^{-6}	8.7079×10^{-6}	100
1.5250×10^{-6}	1.5200×10^{-6}	100	4.4713×10^{-6}	4.4509×10^{-6}	100
7.5290×10^{-7}	7.5043×10^{-7}	100	3.0355×10^{-6}	3.0217×10^{-6}	100
4.3460×10^{-7}	4.3317×10^{-7}	100	2.3004×10^{-6}	2.2899×10^{-6}	100
3.3010×10^{-7}	3.2901×10^{-7}	100	1.3628×10^{-6}	1.3566×10^{-6}	100
1.8100×10^{-7}	1.8041×10^{-7}	100	8.7940×10^{-7}	8.7540×10^{-7}	100
1.1490×10^{-7}	1.1452×10^{-7}	100	4.7100×10^{-7}	4.6886×10^{-7}	100
1.0350×10^{-7}	1.0316×10^{-7}	100	3.6450×10^{-7}	3.6284×10^{-7}	100
3.9300×10^{-8}	3.9171×10^{-8}	100	2.3880×10^{-7}	2.3771×10^{-7}	100
			1.8840×10^{-7}	1.8754×10^{-7}	100
			1.4180×10^{-7}	1.4115×10^{-7}	100
			1.1480×10^{-7}	1.1428×10^{-7}	100
			1.0070×10^{-7}	1.0024×10^{-7}	100
			7.7800×10^{-8}	7.7446×10^{-8}	100
			5.4200×10^{-8}	5.3953×10^{-8}	100
			4.7800×10^{-8}	4.7582×10^{-8}	100
			3.8600×10^{-8}	3.8424×10^{-8}	100
			2.3000×10^{-8}	2.2895×10^{-8}	100
			2.1600×10^{-8}	2.1502×10^{-8}	100
			8.9000×10^{-9}	8.8595×10^{-9}	100

As a general rule, we can consider that the first factor, which corresponds to the first eigenvalue, was clearly dominant in the sense that it can explain a percentage of the total variance ranging between 95.2% and 99.7%, depending on the commodity. It captures qualitative long-run effects. However, it is always necessary to consider a second factor capable of taking up short-term effects. Both the first and second factors explain a cumulative proportion of overall variance between 97.5% and 99.9%, depending on the case under study. In WTI light sweet crude oil, these two factors explain more than a 99.99% of the total variance is explained, while in heating oil case studies, these percentages were approximately 99.88% and in unleaded gasoline and Henry Hub natural gas, they were approximately 97–98%.

Consequently, in the first commodity (crude oil) it is recommended that just the first two factors are considered. The reason is that a third factor will impose a larger estimating effort and a minimum reduction in terms of error measures. The first factor will capture long-term effects, such as world economic events, which significantly impact on commodity prices. The second factor will capture the nature of short-term components such as temporary issues and unforeseen situations. The third and

following stochastic factors can be considered as seasonal factors [28] and, as we know, crude oil is a non-seasonal commodity. This matter reinforces the idea that it is suitable to consider a model with only the first two factors.

The next commodity, heating oil, presents some seasonal behavior, which could be captured by a third factor. The fact that the gain in the percentage of cumulative proportion of overall variance goes from 99.88 to 99.94 and from 99.90 to 99.94 in its respective datasets suggest the inclusion of a third factor was not necessary.

Table 3. Eigenvalues for both datasets of the heating oil.

Dataset 1			Dataset 2		
Eigenvalues	**Percentage of Total Variance**	**Cumulative Variance (%)**	**Eigenvalues**	**Percentage of Total Variance**	**Cumulative Variance (%)**
100	99.6133	99.6133	100	99.5365	99.5365
0.2698	0.2687	99.8820	0.3666	0.3649	99.9014
0.0658	0.0655	99.9475	0.0475	0.0472	99.9486
0.0474	0.0472	99.9947	0.0448	0.0446	99.9932
0.0028	0.0028	99.9975	0.0037	0.0037	99.9969
0.0013	0.0013	99.9988	0.0012	0.0012	99.9981
0.0009	0.0008	99.9997	0.0011	0.0011	99.9992
0.0001	0.0001	99.9998	0.0005	0.0005	99.9997
0.0001	0.0001	99.9999	0.0001	0.0001	99.9998
4.7937×10^{-5}	4.7752×10^{-5}	99.9999	0.0001	0.0001	99.9999
1.9734×10^{-5}	1.9658×10^{-5}	100	4.1450×10^{-5}	4.1257×10^{-5}	99.9999
1.1626×10^{-5}	1.1581×10^{-5}	100	2.8767×10^{-5}	2.8633×10^{-5}	99.9999
1.0482×10^{-5}	1.0441×10^{-5}	100	1.5784×10^{-5}	1.5711×10^{-5}	100
9.6273×10^{-6}	9.5901×10^{-6}	100	1.3478×10^{-5}	1.3416×10^{-5}	100
6.3346×10^{-6}	6.3101×10^{-6}	100	9.3425×10^{-6}	9.2992×10^{-6}	100
			7.9877×10^{-6}	7.9507×10^{-6}	100
			6.1859×10^{-6}	6.1572×10^{-6}	100
			5.7507×10^{-6}	5.7240×10^{-6}	100

Conversely, for the unleaded gasoline and Henry hub natural gas, at least a third factor seemed to be necessary. Both were seasonal commodities (see, for example, [3]). They were characterized by very limited storability and their prices were highly dependent on the commodity demand. Third and fourth factors will acknowledge this behavior. It seems necessary to capture more than long-term and short-term dynamics. Depending on the cumulative variance, if we would like to explain (98–99%), we need to consider at least a third factor or two more. In the unleaded gasoline case, the inclusion of a third factor would increase the cumulative proportion of overall variance from 98.48% to 99.73% and from 97.49% to 98.73%. However, with a fourth factor, we would reach 99.86% and 99.76%, respectively. When we apply the methodology proposed to Henry Hub natural gas datasets, we also verify the need to consider a third and even a fourth factor to explain 99.80% and 99.65% of the total variance, respectively.

Table 4. Eigenvalues for both datasets of the unleaded gasoline (RBOB).

Dataset 1			Dataset 2		
Eigenvalues	**Percentage of Total Variance**	**Cumulative Variance (%)**	**Eigenvalues**	**Percentage of Total Variance**	**Cumulative Variance (%)**
100	96.8591	96.8591	100	95.2473	95.2473
1.6748	1.6222	98.4813	2.3570	2.2450	97.4924
1.2901	1.2496	99.7308	1.3050	1.2429	98.7353
0.1334	0.1292	99.8600	1.0762	1.0250	99.7603
0.0558	0.0540	99.9140	0.0639	0.0608	99.8212
0.0386	0.0374	99.9515	0.0599	0.0570	99.8782
0.0217	0.0210	99.9724	0.0437	0.0416	99.9198
0.0156	0.0151	99.9876	0.0217	0.0206	99.9405
0.0093	0.0090	99.9966	0.0208	0.0198	99.9602
0.0022	0.0022	99.9988	0.0171	0.0163	99.9765
0.0009	0.0009	99.9997	0.0074	0.0070	99.9835
0.0003	0.0003	100	0.0049	0.0047	99.9882
			0.0030	0.0029	99.9910
			0.0025	0.0024	99.9935
			0.0019	0.0018	99.9953
			0.0012	0.0011	99.9964
			0.0008	0.0008	99.9972
			0.0006	0.0006	99.9978
			0.0004	0.0004	99.9982
			0.0004	0.0003	99.9986
			0.0003	0.0003	99.9989
			0.0003	0.0003	99.9991
			0.0002	0.0002	99.9993
			0.0001	0.0001	99.9995
			0.0001	0.0001	99.9996
			0.0001	0.0001	99.9997
			0.0001	0.0001	99.9997
			0.0001	0.0001	99.9998
			0.0001	0.0000	99.9999
			3.8439×10^{-5}	3.6612×10^{-5}	99.9999
			2.8300×10^{-5}	2.6955×10^{-5}	99.9999
			2.4205×10^{-5}	2.3055×10^{-5}	99.9999
			1.9530×10^{-5}	1.8602×10^{-5}	100
			1.5016×10^{-5}	1.4303×10^{-5}	100
			1.2694×10^{-5}	1.2091×10^{-5}	100
			9.9475×10^{-6}	9.4747×10^{-6}	100

These results are coherent with the patterns shown in the futures contracts of each commodity. By considering seasonality as a stochastic factor instead of a deterministic one, we can choose from two- to four-factor models to better model the behavior of commodity prices. It should be noted that the long-term and short-term effects, captured by the first two factors, are clearly dominant in terms of their eigenvalues' relative weight. However, the seasonality should be considered if necessary.

It is important to bear in mind that the distinction between long term and short term is not always direct. It is related to the eigenvalue of the factor, which, as we have stated, is always in the form e^k with $k \leq 0$ (a positive k would mean an explosive process, which is clearly not observed in the data).

If $k = 0$, we have a long-term effect (a unit root). The more negative k is, the shorter the effect. Therefore, $k = -1$ means a much shorter effect than $k = -0.01$, for example.

Explanation capacities of each factor are measured according to their (relative) contribution to the global variance. For example, if there is a unique factor related to eigenvalue $k = 0$ that gives 90% of variance, we would conclude that long term dynamics explain 90% of the variance.

Table 5. Eigenvalues for both datasets of the henry hub natural gas.

Dataset 1			Dataset 2		
Eigenvalues	**Percentage of Total Variance**	**Cumulative Variance (%)**	**Eigenvalues**	**Percentage of Total Variance**	**Cumulative Variance (%)**
100	97.8179	97.8179	100	95.8957	95.8957
1.1564	1.1311	98.9491	2.7972	2.6824	98.5782
0.4785	0.4681	99.4172	0.5993	0.5747	99.1529
0.3960	0.3874	99.8046	0.5221	0.5007	99.6535
0.0839	0.0821	99.8867	0.1178	0.1130	99.7665
0.0730	0.0714	99.9580	0.0993	0.0952	99.8617
0.0223	0.0218	99.9798	0.0782	0.0750	99.9367
0.0052	0.0050	99.9849	0.0166	0.0159	99.9527
0.0039	0.0038	99.9887	0.0074	0.0071	99.9598
0.0031	0.0030	99.9917	0.0070	0.0067	99.9665
0.0027	0.0027	99.9944	0.0060	0.0057	99.9722
0.0023	0.0023	99.9967	0.0051	0.0049	99.9772
0.0012	0.0012	99.9979	0.0048	0.0046	99.9818
0.0010	0.0010	99.9988	0.0038	0.0037	99.9854
0.0007	0.0007	99.9995	0.0032	0.0031	99.9886
0.0005	0.0005	100	0.0024	0.0023	99.9908
			0.0020	0.0019	99.9927
			0.0018	0.0017	99.9944
			0.0017	0.0016	99.9961
			0.0013	0.0012	99.9973
			0.0010	0.0009	99.9983
			0.0006	0.0006	99.9988
			0.0003	0.0003	99.9991
			0.0002	0.0002	99.9993
			0.0002	0.0002	99.9995
			0.0001	0.0001	99.9996
			0.0001	0.0001	99.9997
			0.0001	0.0001	99.9998
			0.0001	0.0001	99.9998
			0.0001	0.0001	99.9999
			3.2345×10^{-5}	3.1018×10^{-5}	99.9999
			2.8128×10^{-5}	2.6974×10^{-5}	100
			1.5565×10^{-5}	1.4926×10^{-5}	100
			1.2489×10^{-5}	1.1977×10^{-5}	100
			9.3473×10^{-6}	8.9637×10^{-6}	100
			7.6972×10^{-6}	7.3813×10^{-6}	100

It should be noted that this article focuses on the econometric theory and identifies the optimal number of factors to characterize the dynamics of commodity prices. Apart from this econometric approach, where each factor represents a component—long term, short term, seasonal, etc.—these factors may also capture economic forces [29–31]. In other words, there are economic forces that are being captured by these factors, such as technology effects (long term) or the functioning of the market (short term). Following [15], we argue that the long-term factor reflects expectations of the exhaustion of the existing supply, improvements in technology for the production and discovery of the commodity, inflation, as well as political and regulatory effects. The short-term factor reflects short-term changes in demand or intermittent supply disruptions. An interpretation of seasonal factors can be found in [3].

This method provides a new selection criterion for obtaining the optimal number of factors. It is always important to keep in mind the purpose of modeling such commodity prices. If we need more accuracy because, for example, we are designing investment strategies, the consideration of more factors is understandable. We could also use fewer factors in a different case.

This is important because, on one hand, if we use too many factors the model will be too complex and parameter estimation may not be accurate. On the other hand, if we use too few factors the model will not be acceptable because it will not capture all the characteristic of the price dynamics that we need to consider in order to solve our problem.

We believe our findings to be very useful for researchers and practitioners. Based on our findings, a researcher who needs to model a commodity price dynamic can use our method to identify the number and the characteristics of the factors to be included in the model. Moreover, a practitioner who is investing or measuring risk can also use our methodology in order to identify the optimal number of factors needed and their characteristics.

Finally, as stated above, we have chosen to order the factors according to their relative (joint) contribution to variance because it is a direct and simple way to interpret the results. We are aware that collinearity and, in general, correlation structures can modify the results. However, since the first eigenvalue explains around 95% of variance, it seems unlikely that results are going to change substantially by a more refined analysis.

4. Summary and Conclusions

In this article, we propose a novel methodology for choosing the optimal number of stochastic factors to be included in a model of the term structure of futures commodity prices. With this method, we add to the research related to the way we characterize commodity price dynamics.

The procedure is based on the eigenvalues of the variance–covariance matrix. Moreover, in deciding how many of them to choose, we propose using the relative weight of the eigenvalues and the percentage of the total variance explained by them and balancing this with the effort of estimating more parameters.

In this article, we applied our method to eight datasets, corresponding with four different commodities: crude oil, heating oil, unleaded gasoline and natural gas. Results indicate that to model the first two commodity prices two factors are suitable, which corresponds with the two biggest eigenvalues, since they are sufficient to account for both long-term and short-term structures. Nevertheless, in the case of unleaded gasoline and natural gas, a third or even fourth factor is needed. We think that, in accordance with the literature, this is related to their seasonal behavior.

Our results support the notion that including too many or too few factors or factors with characteristics which are not optimal in a model for commodity prices could lead to results which may not be as accurate as they should be.

Author Contributions: Conceptualization, B.L. and J.P.; methodology, A.G.-M.; software, A.G.-M.; validation, Población and A.G.-M.; formal analysis, A.G.-M.; investigation, J.P. and A.G.-M.; resources, A.G.-M., B.L. and J.P.; data curation, A.G.-M., B.L. and J.P.; writing—original draft preparation, B.L.; writing—review and editing, B.L. y Población; visualization, B.L.; supervision, Población; project administration, B.L.; funding acquisition, B.L. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge the financial support of the Spanish Ministerio de Economía, Industria y Competitividad Grant Number ECO2017-89,715-P (Javier Población).

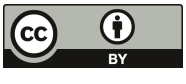
Acknowledgments: This study should not be reported as representing the views of the Banco de España (BdE) or European Central Bank (ECB). The views in this study are those of the author and do not necessarily reflect those of the Banco de España (BdE) or European Central Bank (ECB). We thank the anonymous referees. Any errors are caused by the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jahn, F.; Cook, M.; Graham, M. *Hydrocarbon Exploration and Production*; Elsevier: Aberdeen, UK, 2008.
2. Smit, H.T.J. Investment analysis of offshore concessions in the Netherlands. *Financ. Manag.* **1997**, *26*, 5–17. [CrossRef]
3. García, A.; Población, J.; Serna, G. The Stochastic Seasonal Behaviour of Natural Gas Prices. *Eur. Financ. Manag.* **2012**, *18*, 410–443.
4. García, A.; Población, J.; Serna, G. The stochastic seasonal behavior of energy commodity convenience yields. *Energy Econ.* **2013**, *40*, 155–166.
5. García, A.; Población, J.; Serna, G. Analyzing the dynamics of the refining margin: Implications for valuation and hedging. *Quant. Financ.* **2013**, *12*, 1839–1855.

6. Alquist, R.; Bhattarai, S.; Coibion, O. Commodity-price comovement and global economic activity. *J. Monet. Econ.* **2019**. [CrossRef]
7. Jacks, D.S. From boom to bust: A typology of real commodity prices in the long run. *Cliometrica* **2019**, *13*, 201–220. [CrossRef]
8. Nazlioglu, S. Oil and Agricultural Commodity Prices. In *Routledge Handbook of Energy Economics*; Soytas, U., San, R., Eds.; Routledge: London, UK, 2020; pp. 385–405.
9. Ayres, J.; Hevia, C.; Nicolini, J.P. Real exchange rates and primary commodity prices. *J. Intern. Econ.* **2020**, *122*. [CrossRef]
10. Población, J.; Serna, G. A common long-term trend for bulk shipping prices. *Marit. Econ. Logist.* **2018**, *20*, 421–432. [CrossRef]
11. García, A.; Población, J.; Serna, G. Hedging voyage charter rates on illiquid routes. *Intern. J. Shipp. Transp. Logist.* **2020**, *12*, 197–211.
12. Morgan, J.P. *Risk Metrics–Technical Document*; Reuters: New York, NY, USA, 1996.
13. Echaust, K.; Just, M. Value at Risk Estimation Using the GARCH-EVT Approach with Optimal Tail Selection. *Mathematics* **2020**, *8*, 114. [CrossRef]
14. Schwartz, E.S. The stochastic behavior of commodity prices: Implication for valuation and hedging. *J. Financ.* **1997**, *52*, 923–973. [CrossRef]
15. Schwartz, E.S.; Smith, J.E. Short-term variations and long-term dynamics in commodity prices. *Manag. Sci.* **2000**, *46*, 893–911. [CrossRef]
16. Sorensen, C. Modeling seasonality in agricultural commodity futures. *J. Futures Mark.* **2002**, *22*, 393–426. [CrossRef]
17. Gibson, R.; Schwartz, E.S. Stochastic convenience yield and the pricing of oil contingent claims. *J. Financ.* **1990**, *45*, 959–976. [CrossRef]
18. Cortazar, G.; Schwartz, E.S. Implementing a stochastic model for oil futures prices. *Energy Econ.* **2003**, *25*, 215–218. [CrossRef]
19. Cortazar, G.; Naranjo, L. An N-Factor gaussian model of oil futures prices. *J. Futures Mark.* **2006**, *26*, 209–313. [CrossRef]
20. Camiz, S.; Pillar, V.D. Identifying the Informational/Signal Dimension in Principal Component Analysis. *Mathematics* **2018**, *6*, 269. [CrossRef]
21. Cortazar, G.; Schwartz, E.S. The valuation of commodity-contingent claims. *J. Deriv.* **1994**, *1*, 27–39. [CrossRef]
22. Clewlow, L.; Strickland, C. *Energy Derivatives, Pricing and Risk Management*; Lamina Publication: London, UK, 2000.
23. Tolmasky, C.; Hindanov, D. Principal components analysis for correlated curves and seasonal commodities: The case of the petroleum market. *J. Futures Mark.* **2002**, *22*, 1019–1035. [CrossRef]
24. Hull, J. *Options, Futures and Other Derivatives*, 5th ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2003.
25. García, A.; Población, J.; Serna, G. A Note on Commodity Contingent Valuation. *J. Deriv. Hedge Fund.* **2008**, *13*, 311–320. [CrossRef]
26. Harvey, A.C. *Forecasting Structural Time Series Models and the Kalman Filter*; Cambridge University Press: Cambridge, UK, 1989.
27. Kolos, S.P.; Rohn, E.I. Estimating the commodity market price of risk for energy prices. *Energy Econ.* **2008**, *30*, 621–641. [CrossRef]
28. García, A.; Larraz, B.; Población, J. An alternative method to estimate parameters in modeling the behavior of commodity prices. *Quant. Financ.* **2016**, *16*, 1111–1127. [CrossRef]
29. Coles, J.L.; Li, Z.F. An Empirical Assessment of Empirical Corporate Finance. *SSRN* **2019**. [CrossRef]
30. Coles, J.L.; Li, Z.F. Managerial Attributes, Incentives, and Performance. *Rev. Corp. Financ. Stud.* **2019**. [CrossRef]
31. Dang, C.; Foerster, S.R.; Li, Z.F.; Tang, Z. Analyst Talent, Information, and Insider Trading. *SSRN* **2020**. [CrossRef]



mathematics

MDPI

Article

Detection of Near-Nulticollinearity through Centered and Noncentered Regression

Román Salmerón Gómez [1], Catalina García García [1,*] and José García Pérez [2]

[1] Department of Quantitative Methods for Economics and Business, University of Granada, 18010 Granada, Spain; romansg@ugr.es
[2] Department of Economy and Company, University of Almería, 04120 Almería, Spain; jgarcia@ual.es
* Correspondence: cbgarcia@ugr.es

Received: 1 May 2020; Accepted: 4 June 2020; Published: 7 June 2020

Abstract: This paper analyzes the diagnostic of near-multicollinearity in a multiple linear regression from auxiliary centered (with intercept) and noncentered (without intercept) regressions. From these auxiliary regressions, the centered and noncentered variance inflation factors (VIFs) are calculated. An expression is also presented that relates both of them. In addition, this paper analyzes why the VIF is not able to detect the relation between the intercept and the rest of the independent variables of an econometric model. At the same time, an analysis is also provided to determine how the auxiliary regression applied to calculate the VIF can be useful to detect this kind of multicollinearity.

Keywords: centered model; noncentered model; intercept; essential multicollinearity; nonessential multicollinearity

MSC: 62JXX; 62J20; 60PXX

1. Introduction

Consider the following multiple linear model with n observations and k regressors:

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times k} \cdot \boldsymbol{\beta}_{k\times 1} + \mathbf{u}_{n\times 1}, \tag{1}$$

where $\mathbf{y}$ is a vector with the observations of the dependent variable, $\mathbf{X}$ is a matrix containing the observations of regressors and $\mathbf{u}$ is a vector representing a random disturbance (that is assumed to be spherical). Generally, the first column of matrix $\mathbf{X}$ is composed of ones to denote that the model contains an intercept. Thus, $\mathbf{X} = [\mathbf{1}\ \mathbf{X}_2 \dots \mathbf{X}_k]$ where $\mathbf{1}_{n\times 1} = (1\ 1 \dots 1)^t$. This model is considered to be centered.

When this model presents worrying near-multicollinearity (hereinafter, multicollinearity), that is, when the linear relation between the regressors affects the numerical and/or statistical analysis of the model, the usual approach is to transform the regressors (see, for example, Belsley [1], Marquardt [2] or, more recently, Velilla [3]). Due to the transformations (centering, typification or standardization) implying the elimination of the intercept in the model, the transformed models are considered to be noncentered. Note that even after transforming the data, it is possible to recover the original model (centered) from the estimations of the transformed model (noncentered model). However, in this paper, we refer to the centered and noncentered model depending on whether the intercept is initially included or not. Thus, it is considered that the model is centered if $\mathbf{X} = [\mathbf{1}\ \mathbf{X}_2 \dots \mathbf{X}_k]$ and noncentered if $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2 \dots \mathbf{X}_k]$, given that $\mathbf{X}_j \neq \mathbf{1}$ with $j = 1, \dots, k$.

From the intercept is also possible to distinguish between essential and nonessential multicollinearity:

Nonessential: A near-linear relation between the intercept and at least one of the rest independent variables.

Essential: A near-linear relation between at least two of the independent variables (excluding the intercept).

A first idea of these definitions was provided by Cohen et al. [4]: Nonessential ill-conditioning results simply from the scaling of the variables, whereas essential ill-conditioning results from substantive relationships among the variables. While in some papers the idea of distinguishing between essential and nonessential collinearity is attributed to Marquardt [5], it is possible to find this concept in Marquardt and Snee [6]. These terms have been widely used not only for linear models but also, for example, for moderated models with interactions and/or with a quadratic term. However, these concepts have been analyzed fundamentally from the point of view of the solution of collinearity. Thus, as Marquardt and Snee [6] stated: In a linear model, centering removes the correlation between the constant term and all linear terms.

The variance inflation factor is one of the most applied measures to detect multicollinearity. Following O'Brien [7], commonly a VIF of 10 or even one as low as 4 have been used as rules of thumbs to indicate excessive or serious collinearity. Salmerón et al. [8] show that the VIF does not detect the nonessential multicollinearity, while this kind of multicollinearity is detected by the index of Stewart [9] (see Salmerón Gómez et al. [10]). This index has been misunderstood in the literature since its presentation by Stewart, who wrongly identified it with the VIF. Even Marquardt [11] when published a comment of the paper of Stewart [9] stated: Stewart collinearity indices are simply the square roots of the corresponding variance inflation factor. It is not clear to me whether giving a new name to the square of a VIF is a help or a hindrance to understanding. There is a long and precisely analogous history of using the term "standard error" for the square root of the corresponding "variances". Given the continuing necessity for dealing with statistical quantities on both the scale of the observable and the scale of the observable squared, there may be a place for a new term. Clearly, the essential intellectual content is identical for both terms.

However, in Salmerón Gómez et al. [12] it is shown that the VIF and the index of Stewart are not the same measure. This paper analyzes in what cases use one measure or another, focusing on the initial distinction between centered and noncentered models. Thus, the algebraic contextualization provided by Salmerón Gómez et al. [12] will be complemented from an econometric point of view. This question was also presented by Jensen and Ramirez [13], striving to commit to a clarification of the misuse given to the VIF over decades since its first use, who insinuated: To choose a model, with or without intercept, is substantive, is specific to each experimental paradigm and is beyond the scope of the present study. It was also stated that: This differs between centered and uncentered diagnostics.

This paper, focused on the differences between essential and nonessential multicollinearity in relation to its diagnostic, analyzes the behaviour of the VIF depending on whether model (1) initially includes the intercept or not. For this analysis, it will be considered that the auxiliary regression used for its calculation is centered or not since as stated by Grob [14] (p. 304): Instead of using the classical coefficient of determination in the definition of VIF, one may also apply the centered coefficient of determination. As a matter of fact, the latter definition is more common. We may call VIF uncentered or centered, depending on whether the classical or centered coefficient of determination is used. From the above considerations, a centered VIF only makes sense when the matrix $\mathbf{X}$ contains ones as a column. Additionally, although initially in the centered version of model (1) it is possible to find these two kinds of multicollinearity, and in the noncentered version, it is only possible to find essential multicollinearity, this paper shows that this statement is subject to some nuances.

On the other hand, throughout the paper the following statement of Cook [15] will be illustrated: As a matter of fact, the centered VIF requires an intercept in the model but at the same time denies the status of the intercept as an independent "variable" being possibly related to collinearity effects. Furthermore, another statement was provided by Belsley [16] (p. 29): The centered VIF has no ability to discover collinearity involving the intercept. Thus, the second part of the paper analyzes why the centered VIF is unable to detect the nonessential multicollinearity and, for this, the centered coefficient of determination of the centered auxiliary regression to calculate the centered VIF is analyzed.

This analysis will be applied to propose a methodology to detect the nonessential multicollinearity from the centered auxiliary regression.

The structure of the paper is as follows: Section 2 presents the detection of multicollinearity in noncentered models from the noncentered auxiliary regressions, Section 3 analyzes the effects of high values of the noncentered VIF on the statistical analysis of the model and Section 4 presents the detection of multicollinearity in centered models from the centered auxiliary regressions. Section 5 illustrates the contribution of the paper with two empirical applications. Finally, Section 6 summarizes the main conclusions.

2. Auxiliary Noncentered Regressions

This section presents the calculation of the VIF uncentered, VIFnc, considering that the auxiliary regression is noncentered, that is, it has no intercept. First, the method regarding how to calculate the coefficient of determination for noncentered models is presented.

2.1. Noncentered Coefficient of Determination

Given the linear regression of Equation (1) with or without the intercept, the following decomposition for the sum of squares is verified:

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \widehat{y}_i^2 + \sum_{i=1}^{n} e_i^2, \tag{2}$$

where $\widehat{\mathbf{y}}$ represents the estimation of the dependent variable of the model that is fit by employing ordinary least squares (OLS) and $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}$ are the residuals obtained from that fit. In this case, the coefficient of determination is obtained by the following expression:

$$R_{nc}^2 = \frac{\sum\limits_{i=1}^{n} \widehat{y}_i^2}{\sum\limits_{i=1}^{n} y_i^2} = 1 - \frac{\sum\limits_{i=1}^{n} e_i^2}{\sum\limits_{i=1}^{n} y_i^2}. \tag{3}$$

Comparing the decomposition of the sums of squares given by (2) with the traditionally applied method to calculate the coefficient of determination in models with the intercept, as in model (1):

$$\sum_{i=1}^{n} (y_i - \overline{\mathbf{y}})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{\mathbf{y}})^2 + \sum_{i=1}^{n} e_i^2, \tag{4}$$

it is noted that both coincide if the dependent variable has zero mean. If the mean is different from zero, both models present the same residual sum of squares but different explained and total sum of squares.

Thus, these models lead to the same value for the coefficient of determination (and, as a consequence, for the VIF) only if the dependent variable presents a mean equal to zero.

2.2. Noncentered Variance Inflation Factor

The VIFnc is obtained from the expression:

$$VIFnc(j) = \frac{1}{1 - R_{nc}^2(j)}, \quad j = 1, \ldots, k, \tag{5}$$

where $R_{nc}^2(j)$ is the coefficient of determination, calculated by following (3), of the noncentered auxiliary regression:

$$\mathbf{X}_j = \mathbf{X}_{-j}\delta + \mathbf{w}, \tag{6}$$

where $\mathbf{X}_{-j}$ is equal to the matrix $\mathbf{X}$ after eliminating the variable $\mathbf{X}_j$, for $j = 1, \ldots, k$, and it does not have a vector of ones representing the intercept.

In this case:

- $\sum_{i=1}^{n} X_{ij}^2 = \mathbf{X}_j^t \mathbf{X}_j$, and
- $\sum_{i=1}^{n} \widehat{X}_{ij}^2 = \widehat{\mathbf{X}}_j^t \widehat{\mathbf{X}}_j = \mathbf{X}_j^t \mathbf{X}_{-j} \cdot \left(\mathbf{X}_{-j}^t \mathbf{X}_{-j} \right)^{-1} \cdot \mathbf{X}_{-j}^t \mathbf{X}_j$ due to $\widehat{\mathbf{X}}_j = \mathbf{X}_{-j} \cdot \left(\mathbf{X}_{-j}^t \mathbf{X}_{-j} \right)^{-1} \cdot \mathbf{X}_{-j}^t \mathbf{X}_j$.

Then:

$$\begin{aligned} R_{nc}^2(j) &= \frac{\mathbf{X}_j^t \mathbf{X}_{-j} \cdot \left(\mathbf{X}_{-j}^t \mathbf{X}_{-j} \right)^{-1} \cdot \mathbf{X}_{-j}^t \mathbf{X}_j}{\mathbf{X}_j^t \mathbf{X}_j}, \\ 1 - R_{nc}^2(j) &= \frac{\mathbf{X}_j^t \mathbf{X}_j - \mathbf{X}_j^t \mathbf{X}_{-j} \cdot \left(\mathbf{X}_{-j}^t \mathbf{X}_{-j} \right)^{-1} \cdot \mathbf{X}_{-j}^t \mathbf{X}_j}{\mathbf{X}_j^t \mathbf{X}_j}, \\ VIFnc(j) &= \frac{\mathbf{X}_j^t \mathbf{X}_j}{\mathbf{X}_j^t \mathbf{X}_j - \mathbf{X}_j^t \mathbf{X}_{-j} \cdot \left(\mathbf{X}_{-j}^t \mathbf{X}_{-j} \right)^{-1} \cdot \mathbf{X}_{-j}^t \mathbf{X}_j}. \end{aligned} \tag{7}$$

Thus, the VIFnc coincides with the expression given by Stewart [9] for the VIF and is denoted as k_j^2, that is, $VIFnc(j) = k_j^2$.

However, recently, Salmerón Gómez et al. [12] showed that the index presented by Stewart has been misleadingly identified as the VIF, verifying the following relation between both measures:

$$k_j^2 = VIF(j) + n \cdot \frac{\overline{\mathbf{X}}_j^2}{RSS_j}, \quad j = 2, \ldots, k, \tag{8}$$

where $\overline{\mathbf{X}}_j$ is the mean of the $j-$variable of $\mathbf{X}$. This expression is also shown by Salmerón Gómez et al. [10], where it is used to quantify the proportion of essential and nonessential multicollinearity existing in a concrete independent variable.

Note that the expression:

$$VIFnc(j) = VIF(j) + n \cdot \frac{\overline{\mathbf{X}}_j^2}{RSS_j}, \tag{9}$$

is obtained by Chennamaneni et al. [17] (expression (6) page 174), although it is also limited to the particular case of the moderated regression $\mathbf{Y} = \alpha_0 \cdot \mathbf{1} + \alpha_1 \cdot \mathbf{U} + \alpha_2 \cdot \mathbf{V} + \alpha_3 \cdot \mathbf{U} \times \mathbf{V} + \nu$ where $\mathbf{U}$ and $\mathbf{V}$ are ratio-scaled explanatory variables in n-dimensional data vectors. Indeed, these authors proposed a new measure to detect multicollinearity in moderated regression models that is derived from the noncentered coefficient of determination. However, this use of the noncentered coefficient of determination lacks of the statistical contextualization provided by this paper

Finally, from expression (9), it is shown that the VIFnc and the VIF only coincide if the associated variable has zero mean, analogously to what happens in the decomposition of the sum of squares. Note that this expression also clarifies why Stewart's collinearity indices diminish when the variables are centered, which the author attributed to errors in regression variables: This phenomenon is a consequence of the fact that our definition of collinearity index compels us to work with relative errors.

Example 1. *Considering $k = 4$ in model (1), we use the noncentered coefficient of determination, R_{nc}^2, to calculate the noncentered variance inflation factor, $VIFnc$. For it, we consider the values displayed in Table 1. Note that variables $\mathbf{y}$, $\mathbf{X}_2$ and $\mathbf{X}_3$ were originally used by Belsley [1] and we have added a new variable, $\mathbf{X}_4$, that has been randomly generated (from a normal distribution with a mean equal to 4 and a variance equal to 16) to obtain a variable that is linearly independent with respect to the rest.*

Table 1. Data set applied by Belsley [1].

y	1	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$
2.69385	1	0.996926	1.00006	8.883976
2.69402	1	0.997091	0.998779	6.432483
2.70052	1	0.9973	1.00068	−1.612356
2.68559	1	0.997813	1.00242	1.781762
2.7072	1	0.997898	1.00065	2.16682
2.6955	1	0.99814	1.0005	4.045509
2.70417	1	0.998556	0.999596	4.858077
2.69699	1	0.998737	1.00262	4.9045
2.69327	1	0.999414	1.00321	8.631162
2.68999	1	0.999678	1.0013	−0.4976853
2.70003	1	0.999926	0.997579	6.828907
2.702	1	0.999995	0.998597	8.999921
2.70938	1	1.00063	0.995316	7.080689
2.70094	1	1.00095	0.995966	1.193665
2.70536	1	1.00118	0.997125	1.483312
2.70754	1	1.00177	0.998951	−1.053813
2.69519	1	1.00231	1.00102	−0.5860236
2.7017	1	1.00306	1.00186	−1.371546
2.70451	1	1.00394	1.00353	−2.445995
2.69532	1	1.00469	1.00021	5.731981

In these data, the existence of nonessential multicollinearity is intuited. This fact is confirmed by the small values of the coefficient of variation (CV) in two of the independent variables and the following conclusions obtained from the value of the condition indices and the proportions of the variance (see, for example, Belsley et al. [18] and Belsley [16] for more details) shown in Table 2:

- *Variables* $\mathbf{X}_2$ *and* $\mathbf{X}_3$ *present a CV lower than* 0.06674082 *and than* 0.1002506 *that were presented by Salmerón Gómez et al. [10] as thresholds to indicate that a variable may be related to the constant and the model will present strong and moderate nonessential multicollinearity, respectively.*
- *The second index is associated with a high proportion of the variance with the variable* $\mathbf{X}_4$*, although it is not worrisome since it does not present a high value.*
- *The third index presents a value higher than the established thresholds (20 for moderate multicollinearity and 30 for strong multicollinearity), and it is also associated with high proportions in the variables* $\mathbf{X}_2$ *and* $\mathbf{X}_3$*.*
- *The last index identified as the condition number is clearly related to the intercept, and at the same time, it includes the relation between* $\mathbf{X}_2$ *and* $\mathbf{X}_3$ *as previously commented.*
- *Finally, the condition number, 1614.829, is higher than the threshold traditionally established as indicative of worrisome multicollinearity.*

Table 2. Diagnostic of collinearity of Belsley–Kuh–Welsch and coefficient of variation of the considered variables.

		Proportion of the Variance			
Eigenvalue	**Index of Condition**	**1**	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$
3.517205	1.000	0	0	0	0.022
0.4827886	2.699	0	0	0	0.784
4.978345×10^{-6}	840.536	0	0.423	0.475	0.003
1.348791×10^{-6}	1614.829	1	0.577	0.525	0.191
Coefficients of variation			0.002	0.002	1.141

Now, other models are proposed apart from the initial model for $k = 4$*:*

- *Model 0 (**Mod0**):* $\mathbf{y} = \beta_1 \cdot \mathbf{1} + \beta_2 \cdot \mathbf{X}_2 + \beta_3 \cdot \mathbf{X}_3 + \beta_4 \cdot \mathbf{X}_4 + \mathbf{u}$.
- *Model 1 (**Mod1**):* $\mathbf{y} = \beta_1 \cdot \mathbf{1} + \beta_2 \cdot \mathbf{X}_2 + \beta_3 \cdot \mathbf{X}_3 + \mathbf{u}$.
- *Model 2 (**Mod2**):* $\mathbf{y} = \beta_1 \cdot \mathbf{1} + \beta_2 \cdot \mathbf{X}_2 + \beta_4 \cdot \mathbf{X}_4 + \mathbf{u}$.

- *Model 3 (**Mod3**):* $\mathbf{y} = \beta_1 \cdot \mathbf{1} + \beta_3 \cdot \mathbf{X}_3 + \beta_4 \cdot \mathbf{X}_4 + \mathbf{u}$.

*Table 3 presents the VIF and the VIFnc of these models. Note that by using the original variables applied by Belsley (**Mod1**), the traditional VIF (from the centered model, see Theil [19]) provides a value equal to 1 (its minimum possible value), while the VIFnc is equal to 100,032.1. If the additional variable* $\mathbf{X}_4$ *is included (**Mod0**), the traditional VIFs are also close to one while the noncentered VIFs present values higher than 100,000. The conclusion is that the VIF is not detecting the existence of nonessential multicollinearity (see Salmerón et al. [8]) while the VIFnc "does detect it". However, since the calculation of VIFnc excludes the constant term, the detected relation refers to the one between* $\mathbf{X}_2$ *and* $\mathbf{X}_3$*, and not to the relation between* $\mathbf{X}_2$ *and/or* $\mathbf{X}_3$ *with the intercept.*

*This fact is supported by the values obtained for the VIF and VIFnc of the second and fourth variables (**Mod2**) and for the third and fourth variables (**Mod3**).*

Table 3. Variance inflation factor (VIF) and VIF uncentered (VIFnc) of models proposed from Belsley [1] dataset.

		$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$
Mod0	VIF	1.155	1.084	1.239
	VIFnc	100,453.8	100,490.6	1.737
Mod1	VIF	1	1	
	VIFnc	100,032.1	100,032.1	
Mod2	VIF	1.143		1.143
	VIFnc	1.765		1.765
Mod3	VIF		1.072	1.072
	VIFnc		1.766	1.766

2.3. What Kind of Multicollinearity Detects the VIFnc?

The results of Example 1 for **Mod0** suggest a new definition of nonessential multicollinearity as the relation between at least two variables with little variability. Thus, the particular case when one of these variables is the intercept leads to the definition initially given by Marquardt and Snee [6]. Then, the initial idea that in a noncentered model, is not possible to find nonessential collinearity is of a nuanced nature.

By following Salmerón et al. [8] and Salmerón Gómez et al. [10], it can be concluded that the VIF only detects the essential multicollinearity and, with these results, the VIFnc detects the nonessential multicollinearity but in its generalized definition since the intercept is eliminated in the corresponding auxiliary regression.

This fact is contradictory to the fact that the VIFnc coincides with the index of Stewart, see expression (7), since this measure is able to detect the nonessential multicollinearity (see Salmerón Gómez et al. [10]). This is because the VIFnc could be fooled, including the constant as an independent variable in a model without the intercept, that is:

$$\mathbf{y} = \beta_1 \cdot \mathbf{X}_1 + \beta_2 \cdot \mathbf{X}_2 + \cdots + \beta_k \cdot \mathbf{X}_k + \mathbf{u},$$

where $\mathbf{X}_1$ is a column of ones but is not considered as the intercept.

Example 2. *Now, we part from model 1 in the Belsley example but include the constant as an independent variable in a model without the intercept (**Mod4**) and two additional models (**Mod5** and **Mod6**):*

- *Model 4 (**Mod4**):* $\mathbf{y} = \beta_1 \cdot \mathbf{X}_1 + \beta_2 \cdot \mathbf{X}_2 + \beta_3 \cdot \mathbf{X}_3 + \mathbf{u}$.
- *Model 5 (**Mod5**):* $\mathbf{y} = \beta_1 \cdot \mathbf{X}_1 + \beta_2 \cdot \mathbf{X}_2 + \mathbf{u}$.
- *Model 6 (**Mod6**):* $\mathbf{y} = \beta_1 \cdot \mathbf{X}_1 + \beta_3 \cdot \mathbf{X}_3 + \mathbf{u}$.

Table 4 presents the VIFnc obtained from expression (5) in Models 4–6. Results indicate that, considering the centered model and calculating the coefficient of determination of the auxiliary regressions as if the model were

noncentered, it is possible to detect the nonessential multicollinearity. Thus, the contradiction indicated at the beginning of this subsection is saved.

Table 4. VIFnc of Models 4–6 including the constant as an independent variable in a model without the intercept.

	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$
Mod4	400,031.4	199,921.7	200,158.3
Mod5	199,921.7	199,921.7	
Mod6		200,158.3	200,158.3

3. Effects of the Vifnc on the Statistical Analysis of the Model

Given the model (1), the expression obtained for the variance of the estimator is given by:

$$var(\widehat{\beta}_j) = \frac{\sigma^2}{RSS_j}, \quad j = 1, \ldots, k, \tag{10}$$

where RSS_j is the residual sum of squares of the auxiliary regression of the $j-$independent variable as a function of the rest of the independent variables (see expression (6)).

From expression (10), and considering that expression (7) can be rewritten as:

$$VIFnc(j) = \frac{\mathbf{X}_j^t \mathbf{X}_j}{RSS_j},$$

it is possible to obtain:

$$var(\widehat{\beta}_j) = \frac{\sigma^2}{RSS_j} = \frac{\sigma^2}{\mathbf{X}_j^t \mathbf{X}_j} \cdot VIFnc(j), \quad j = 1, \ldots, k. \tag{11}$$

Establishing a model as a reference is required to conclude whether the variance has been inflated (see, for example, Cook [20]). Thus, if the variables in $\mathbf{X}$ are orthogonal, it is verified that $\mathbf{X}^t\mathbf{X} = diag(d_1, \ldots, d_k)$ where $d_j = \mathbf{X}_j^t \mathbf{X}_j$. In this case, $\left(\mathbf{X}^t\mathbf{X}\right)^{-1} = diag(1/d_1, \ldots, 1/d_k)$, and consequently, the variance of the estimated coefficients in the hypothetical orthogonal case is given by the following expression:

$$var(\widehat{\beta}_{j,o}) = \frac{\sigma^2}{\mathbf{X}_j^t \mathbf{X}_j}, \quad j = 1, \ldots, k. \tag{12}$$

In this case:

$$\frac{var(\widehat{\beta}_j)}{var(\widehat{\beta}_{j,o})} = VIFnc(j), \quad j = 1, \ldots, k,$$

and it is then possible to state that the VIFnc is a factor that inflates the variance.

As consequence, high values of $VIFnc(j)$ imply high values of $var(\widehat{\beta}_j)$ and a tendency not to reject the null hypothesis in the individual significance test of model (1). Thus, the statistical analysis of the model will be affected.

Note from expression (11) that this negative effect can be offset by low values of the estimation of σ^2, that is, low values of the residual sum of squares of model (1) or high values of the number of observations, n. This is similar to what happen to the VIF (see O'Brien [7] for more details).

4. Auxiliary Centered Regressions

The use of the coefficient of determination of the auxiliary regression (6) where matrix $\mathbf{X}_{-j}$ contains a column of ones that represents the intercept is a very common approach to detect the linear relations between the independent variables of the model (1). This is motivated due to the higher relation between

$\mathbf{X}_j$ and the rest of the independent variables, that is, the higher the multicollinearity is, the higher the value of that coefficient of determination.

However, since the coefficient of determination ignores the role of the intercept, this measure is unable to detect the nonessential linear relations. The question is evident: Does another measure exist related to the auxiliary regression that allows detection of the nonessential multicollinearity?

4.1. Case When There Is Only Nonessential Multicollinearity

Example 3. *Suppose that 100 observations are simulated for variables* $\mathbf{X}$, $\mathbf{Z}$ *and* $\mathbf{W}$ *from normal distributions with a mean of 5, 4 and -4 and a standard deviation of 0.01, 4 and 0.01, respectively. Note that* $\mathbf{X}$ *and* $\mathbf{W}$ *present light variability and, for this reason, it is expected that the model presents nonessential multicollinearity.*

Then, $\mathbf{y} = \mathbf{1} + \mathbf{X} + \mathbf{Z} - \mathbf{W} + \mathbf{v}$ *is generated by simulating* $\mathbf{v}$ *as a normal distribution with a mean equal to 0 and a standard deviation equal to 2.*

The second column of Table 5 presents the results obtained after the estimation by ordinary least squares (OLS) of model $\mathbf{y} = \beta_1 \cdot \mathbf{1} + \beta_2 \cdot \mathbf{X} + \beta_3 \cdot \mathbf{Z} + \beta_4 \cdot \mathbf{W} + \mathbf{u}$. *Note that the estimations of the coefficients of the model differ substantially from the real values used to generate* $\mathbf{y}$, *except for the coefficient of the variable* $\mathbf{Z}$ *(this situation illustrates the fact that if the interest is to estimate the effect of variable* $\mathbf{Z}$ *on* $\mathbf{y}$, *the analysis will not be influenced by the linear relations between the rest of the independent variables), which is the variable free of multicollinearity (indeed, it is the unique coefficient significantly different from zero, with a 5% significance—the value used by default in this paper).*

Table 5. Estimation by ordinary least squares (OLS) of the first simulated model and its corresponding auxiliary regressions (estimated standard deviation in parenthesis and coefficients significantly different from zero in bold).

Dependent Variable	$\widehat{\mathbf{y}}$	$\widehat{\mathbf{X}}$	$\widehat{\mathbf{Z}}$	$\widehat{\mathbf{W}}$
Intercept	173.135	**4.969**	−27.63	**−3.953**
	(123.419)	(0.369)	(240.08)	(0.557)
$\mathbf{X}$	−38.308		−17.05	−0.009
	(20.035)		(38.94)	(0.111)
$\mathbf{Z}$	0.939	−0.0001		−0.0002
	(0.052)	(0.0002)		(0.0002)
$\mathbf{W}$	−7.173	−0.007	−29.34	
	(18.2309)	(0.092)	(35.34)	
R^2	0.7773	0.001	0.008	0.007
VIF		1.001	1.008	1.007

This table also shows the results obtained from the estimations of the centered auxiliary regressions. Note that the coefficients of determination are very small, and consequently, the associated VIFs do not detect the degree of multicollinearity. However, note that in the auxiliary regressions corresponding to variables $\mathbf{X}$ *and* $\mathbf{W}$*:*

- *The estimation of the coefficient of the intercept almost coincides with the mean from which each variable was generated, 5 and −4, and, at the same time, the coefficients of the rest of the independent variables are almost zero.*
- *The estimations of the coefficients of the intercept are the unique ones that are significantly different from zero.*

Thus, note that the auxiliary regressions are capturing the existence of nonessential multicollinearity. The problem is that it is not transferred to its coefficient of determination but to another characteristic.

From this finding, it is possible to propose a way to detect the nonessential multicollinearity from the centered auxiliary regression traditionally applied to calculate the VIF:

Condition 1 (C1): Quantify the contribution of the estimation of the intercept to the total sum of the estimations of the coefficients of model (6), that is, calculate:

$$\frac{|\delta_1|}{\sum_{j=1}^{k-1}|\delta_j|}\cdot 100\%.$$

Condition 2 (C2): Calculate the number of independent variables with coefficients significantly different from zero and quantify the contribution of the intercept.

A Montecarlo simulation is presented considering the model (1) where $k = 3$ and the variable $\mathbf{X}_2$ has been generated as a normal distribution with mean $\mu_2 \in \mathbf{A}$ and variance $\sigma_2^2 \in \mathbf{B}$, the variable $\mathbf{X}_3$ has been generated as normal distribution with mean $\mu_3 \in \mathbf{A}$ and variance $\sigma_3^2 \in \mathbf{C}$ being $\mathbf{A} = \{0, 1, 2, 3, 4, 5, 10, 15, 20\}$, $\mathbf{B} = \{0.00001, 0.0001, 0.001, 0.1, \mathbf{C}\}$ and $\mathbf{C} = \{1, 2, 3, 4, 5, 10, 15, 20\}$. The results are presented in Table 6. Taking into account that the sample size has varied within the set $\{15, 20, 25, \ldots, 140, 145, 150\}$, 235872 iterations have been performed.

Table 6. Values of condition **C1** depending on the coefficient of variation (CV).

	P_5	P_{95}	Mean	Typical Deviation
$CV < 0.06674082$	99.402%	99.999%	99.512%	3.786%
$CV > 0.06674082$	52.678%	99.876%	89.941%	16.837%
$CV < 0.1002506$	95.485%	99.999%	98.741%	6.352%
$CV > 0.1002506$	51.434%	99.842%	89.462%	17.114%

Considering the thresholds established by Salmerón Gómez et al. [10], 90% of the simulations present values for condition **C1** between 99.402% and 99.999% if $CV < 0.06674082$ and between 95.485% and 99.999% if $CV < 0.1002506$. Thus, we can consider that values of condition **C1** higher than 95.485% will indicate that the auxiliary centered regressions are detecting the presence of nonessential multicollinearity.

Table 7 shows that a high value is obtained for the condition **C1**, even if any estimated coefficient is significantly different from zero (**C2** = NA).

Thus, the previous threshold, 95.485%, will be considered as valid if it is accompanied by a high value in the second condition.

Table 7. Values of condition **C1** depending on condition **C2**.

	C2	**NA**	**50%**	**100%**
C1	P_5	39.251%	67.861%	89.514%
	P_{95}	98.751%	99.984%	99.997%
	Mean	81.378%	91.524%	96.965%
	Typical Deviation	19.622%	13.598%	9.972%

Example 4. *Applying these criteria to the data of the Example 1 for* ***Mod1****, it is obtained that:*

- *In the auxiliary regression* $\mathbf{X}_2 = \delta_1 \cdot \mathbf{1} + \delta_3 \cdot \mathbf{X}_3 + \mathbf{w}$*, the estimation of the intercept is equal to 99.988% of the total, and the individual significance of the intercept corresponds to 100% of the significant estimated coefficients.*
- *In the auxiliary regression* $\mathbf{X}_3 = \delta_1 \cdot \mathbf{1} + \delta_2 \cdot \mathbf{X}_2 + \mathbf{w}$*, the estimation of the intercept is equal to 99.988% of the total, and the individual significance of the intercept corresponds to 100% of the significant estimated coefficients.*

Thus, the symptoms shown in the previous simulation also appear, and consequently, in both situations, the nonessential multicollinearity will be detected.

Replicating both situations where the VIFnc was not able to detect the nonessential multicollinearity, it is obtained that:

- *For **Mod2** it is obtained that:*
  - *In the auxiliary regression $\mathbf{X}_2 = \delta_1 \cdot \mathbf{1} + \delta_4 \cdot \mathbf{X}_4 + \mathbf{w}$, the estimation of the intercept is equal to the 99.978% of the total, and the individual significance of the intercept corresponds to 100% of the significant estimated coefficients.*
  - *In the auxiliary regression $\mathbf{X}_4 = \delta_1 \cdot \mathbf{1} + \delta_2 \cdot \mathbf{X}_2 + \mathbf{w}$, the estimation of the intercept is equal to 50.138% of the total, and none of the estimated coefficients are significantly different from zero.*
- *For **Mod3** it is obtained that:*
  - *In the auxiliary regression $\mathbf{X}_3 = \delta_1 \cdot \mathbf{1} + \delta_4 \cdot \mathbf{X}_4 + \mathbf{w}$, the estimation of the intercept is equal to 99.984% of the total, and the individual significance of the intercept corresponds to 100% of the significant estimated coefficients.*
  - *In the auxiliary regression $\mathbf{X}_4 = \delta_1 \cdot \mathbf{1} + \delta_3 \cdot \mathbf{X}_3 + \mathbf{w}$, the estimation of the intercept is equal to 50.187% of the total, and none of the estimated coefficients are significantly different from zero.*

Once again, it was shown that with this procedure, it is possible to detect the nonessential multicollinearity and the variables that are causing it.

4.2. Relevance of a Variable in a Regression Model

Note that the conditions **C1** and **C2** are focused on measuring the relevance of one of the variables, in this case, the intercept, within the multiple linear regression model. It is interesting to analyze the behavior of other measures with this same goal as, for example, the index ι_j of Stewart [9]. Given model (1), Stewart defined the relevance of the $j-$variable as the number:

$$\iota_j = \frac{|\beta_j| \cdot ||\mathbf{X}_j||}{||\mathbf{y}||}, \quad j = 1, \ldots, p,$$

where $||\cdot||$ is the usual Euclidean norm. Stewart considered that a variable with a relevance higher than 0.5 should not be ignored.

Example 5. *Table 8 presents the calculation of ι_j for situations shown in Example 1. Note that in all cases, the intercept will be considered relevant, even when the variable $\mathbf{X}_4$ is analyzed as a function of $\mathbf{X}_2$ or $\mathbf{X}_3$, despite that it was previously shown that the intercept was not relevant in these situations (at least in relation to nonessential multicollinearity).*

Table 8. Calculation of ι_j for situations **Mod1**, **Mod2** and **Mod3** shown in Example 1.

	Auxiliary Regression	ι_1	ι_2
Mod1	$\mathbf{X}_2 = \delta_1 \cdot \mathbf{1} + \delta_3 \cdot \mathbf{X}_3 + \mathbf{w}$	0.999	0.0001
	$\mathbf{X}_3 = \delta_1 \cdot \mathbf{1} + \delta_2 \cdot \mathbf{X}_2 + \mathbf{w}$	0.999	0.0001
Mod2	$\mathbf{X}_2 = \delta_1 \cdot \mathbf{1} + \delta_4 \cdot \mathbf{X}_4 + \mathbf{w}$	1.0006	0.001
	$\mathbf{X}_4 = \delta_1 \cdot \mathbf{1} + \delta_2 \cdot \mathbf{X}_2 + \mathbf{w}$	119.715	119.056
Mod3	$\mathbf{X}_3 = \delta_1 \cdot \mathbf{1} + \delta_4 \cdot \mathbf{X}_4 + \mathbf{w}$	1.0005	0.0007
	$\mathbf{X}_4 = \delta_1 \cdot \mathbf{1} + \delta_3 \cdot \mathbf{X}_3 + \mathbf{w}$	88.346	87.687

*Thus, the application of ι_j seems not to be appropriate contrarily to what happens with conditions **C1** and **C2**.*

4.3. Case When There Is Generalized Nonessential Multicollinearity

Example 6. *Suppose that the previous simulation is repeated, except for the generation of the variable $\mathbf{Z}$, which, in this case, is considered to be given by $Z_i = 2 \cdot X_i - a_i$, for $i = 1, \ldots, 100$, where a_i is generated from a normal distribution with a mean equal to 2 and a standard deviation equal to 0.01.*

Table 9 presents the results of the estimation by OLS of the model $\mathbf{y} = \beta_1 \cdot \mathbf{1} + \beta_2 \cdot \mathbf{X} + \beta_3 \cdot \mathbf{Z} + \beta_4 \cdot \mathbf{W} + \mathbf{u}$ and its possible auxiliary regressions.

In this case, none of the coefficients are significantly different from zero and the coefficients are very far from the real values used in the simulation.

Table 9. Estimation by OLS of the second simulated model and its corresponding auxiliary regressions (estimated standard deviation in parenthesis and coefficients significantly different from zero in bold).

Dependent Variable	$\hat{\mathbf{y}}$	$\widehat{\mathbf{X}}$	$\widehat{\mathbf{Z}}$	$\widehat{\mathbf{W}}$
Constant	−233.37	**1.977**	**−2.638**	**−4.959**
	(167.33)	(0.2203)	(0.673)	(0.715)
X	12.02		**2.213**	−0.059
	(56.98)		(0.102)	(0.298)
Z	8.89	**0.374**		0.156
	(23.44)	(0.017)		(0.121)
W	−29.96	−0.006	0.107	
	(19.41)	(0.034)	(0.107)	
R^2	0.034	0.838	0.841	0.073
VIF		6.172	6.289	1.078

In relation to the auxiliary regression, it is possible to conclude that:

- *When the dependent variable is* **X**, *the coefficients that are significantly different from zero are the ones of the intercept and the variable* **Z**. *At the same time, the estimation of the coefficient of the intercept differs from the mean from which the variable* **X** *was generated. In this case, the contribution of the estimation of the intercept is equal to 83.837% of the total and represents 50% of the coefficients significantly different from zero.*
- *When the dependent variable is* **Z**, *the coefficients significantly different from zero are the ones of the intercept and the variable* **X**. *In this case, the contribution of the estimation of the intercept is equal to 53.196% of the total and represents 50% of the coefficients significantly different from zero.*
- *When the dependent variable is* **W**, *the signs shown in the previous section are maintained. In this case, the contribution of the intercept is equal to 95.829% of the total and represents 100% of the coefficients significantly different from zero.*
- *Finally, although it will require a deeper analysis, the last results indicate that the estimated coefficient that is significantly different from zero in the auxiliary regression represents the variables responsible for the existing linear relation (intercept included).*

Note that the existence of generalized nonessential multicollinearity distorts the symptoms previously detected. Thus, the fact that in a centered auxiliary regression, the contribution (in absolute terms) of the estimation of the intercept to the total sum (in absolute value) of all estimations will be close to 100%, and the estimation of the intercept will be uniquely significantly different from zero, are indications of nonessential multicollinearity. However, it is possible that these symptoms are not manifested but there exists worrisome nonessential multicollinearity. Thus, these conditions are sufficient but not required.

However, in situations shown in Example 6 where conditions **C1** and **C2** are not verified, the VIFnc will be equal to 1109,259.3, 758,927.7 and 100,912.7. Thus, note that these results complement the results presented in the previous section in relation to the VIFnc. Thus, VIFnc detects generalized nonessential multicollinearity while conditions **C1** and **C2** detect the traditional nonessential multicollinearity given by Marquardt and Snee [6].

5. Empirical Applications

In order to illustrate the contribution of this study, this section presents two empirical applications with financial and economic real data. Note that in a financial prediction model, a financial variable with low variance means low risk and a better prediction, because the standard deviation and volatility are lower. However, as discussed above, a lower variance of the independent variable may mean

greater nonessential multicollinearity in a GLR model. Thus, the existence of worrisome nonessential collinearity may be relatively common in financial econometric models and this idea can be extended in general to economic applications. Note that the objective is to diagnose the type of multicollinearity existing in the model and indicate the most appropriate treatment (without applying it).

5.1. Financial Empirical Application

The following model of Euribor (**100%**) is specified from the data set composed by 47 Eurozone observations for the period January 2002 to July 2013 (quarterly and seasonally adjusted data) and previously applied by Salmerón Gómez et al. [10]:

$$\mathbf{Euribor} = \beta_1 + \beta_2 \cdot \mathbf{HICP} + \beta_3 \cdot \mathbf{BC} + \mathbf{u}, \tag{13}$$

where **HICP** is the Harmonized Index of Consumer Prices (**100%**), **BC** is the Balance of Payments to net current account (millions of euros) and **u** is a random disturbance (centered, homoscedastic, and uncorrelated).

Table 10 presents the analysis of model (13) and its corresponding auxiliary regressions. The values of the VIFs which are very close to one will indicate that there is not essential multicollinearity. The correlation coefficient between **HICP** and **BC** is 0.231 and the determinant of the correlation matrix is 0.946. Both values indicate that there is no essential multicollinearity, see García García et al. [21] and Salmerón Gómez et al. [22].

However, the condition number is higher than 30 indicating a strong multicollinearity associated, see conditions **C1** and **C2**, with variable **HICP**. The values of conditions **C1** and **C2** are conclusive in the case of variable **HICP**. In the case of variable **BC**, although condition **C1** presents a high value, none of the coefficients of the auxiliary regression is significatively different from zero (condition **C2**). By following the simulation presented in subsection, this indicate that the variable **BC** is not related to the intercept. This conclusion is in line with the value of the coefficient of variation of variable **HICP** that is lower than 0.1002506, the threshold established by Salmerón Gómez et al. [10] for moderate nonessential multicollinearity.

Table 11 presents the calculation of the VIFnc. Note that it is not detecting the non-essential multicollinearity. As previously commented, the VIFnc only detects the essential and the generalized nonessential multicollinearity. This table also presents the VIFnc calculated in a model without intercept but including the constant as an independent variable (see Section 2.3). In this case, the VIFnc is able to detect the nonessential multicollinearity between the intercept and the variable **HIPC**.

In conclusion, this model will present nonessential multicollinearity caused by the variable **HICP**. This problem can be mitigated by centering that variable (see, for example, Marquardt and Snee [6] and Salmerón Gómez et al. [10]).

Table 10. Estimations by OLS of model (13) and its corresponding auxiliary regressions (estimated standard deviation in parenthesis and coefficients significantly different from zero in bold).

	Euribor	**HICP**	**BC**
Intercept	**8.442**	**104.8**	−64,955
	(1.963)	(1.09)	(43,868)
HICP	**−0.054**		663.3
	(0.018)		(415.9)
BC	$\mathbf{-3.493 \times 10^{-5}}$	8.065×10^{-5}	
	(6.513×10^{-6})	(5.057×10^{-5})	
R^2	0.517	0.053	0.053
VIF		1.055	1.055
CN	30.246		
Condition 1 (**C1**)		99.999%	98.98%
Condition 2 (**C2**)		100%	NA
Coefficients of variation		0.069	4.3403

Table 11. VIFnc of auxiliary regressions associated to model (13).

	$\mathbf{X_1}$	**HICP**	**BC**
VIFnc		1.0609	1.0609
VIFnc	217.672	219.291	1.112

5.2. Economic Empirical Application

From French economy data from Chatterjee and Hadi [23], also analyzed by Malinvaud [24], Zhang and Liu [25] and Kibria and Lukman [26], among others, the following model is analyzed:

$$\mathbf{I} = \beta_1 + \beta_2 \cdot \mathbf{DP} + \beta_3 \cdot \mathbf{SF} + \beta_4 \cdot \mathbf{DC} + \mathbf{u}, \tag{14}$$

for years 1949 through 1966 where imports (**I**), domestic production (**DP**), stock formation (**SF**) and domestic consumption (**DC**), all are measured in billions of French francs and **u** is a random disturbance (centered, homoscedastic, and uncorrelated).

Table 12 presents the analysis of model (14) and its corresponding auxiliary regressions. The values of the VIFs of variables **DP** and **DC** indicate strong essential multicollinearity. The condition number is higher than 30 also indicating a strong multicollinearity.

Note that the values of condition **C1** for variables **DP** and **DC** are lower than threshold shown in the simulation. Only the variable **SF** presents a higher value but, in this case, condition **C2** indicates that none of the estimated coefficients of the auxiliary regression are significatively different from zero. This conclusion is in line with the coefficients of variation that are higher than the threshold established by Salmerón Gómez et al. [10] indicating that there is no nonessential multicollinearity.

Table 13 presents the calculation of the VIFnc. Note that it is detecting the essential multicollinearity. This table also presents the VIFnc calculated in a model without intercept but including the constant as an independent variable. In this case, the VIFnc is also detecting the essential multicollinearity between the variables **DP** and **DC**. From thresholds established by Salmerón Gómez et al. [10] for simple linear regression ($k = 2$), the value 60.0706 will not be worrisome and, consequently, the nonessential multicollinearity will not be worrisome.

Table 12. Estimations by OLS of Model (14) and its corresponding auxiliary regressions (estimated standard deviation in parenthesis and coefficients significantly different from zero in bold).

	I	**DP**	**SF**	**DC**
Intercept	−**19.725**	−**18.052**	2.635	**12.141**
	(4.125)	(3.28)	(3.234)	(2.026)
DP	0.032		0.025	**0.654**
	(0.186)		(0.149)	(0.007)
SF	0.414	0.075		−0.038
	(0.322)	(0.444)		(0.291)
DC	0.242	**1.525**	−0.029	
	(0.285)	(0.018)	(0.228)	
R^2	0.973	0.997	0.047	0.997
VIF		333.333	1.049	333.333
CN	247.331			
Condition 1 (**C1**)		91.85%	97.94%	94.6%
Condition 2 (**C2**)		50%	NA	50%
Coefficients of variation		0.267	0.473	0.248

Table 13. VIFnc of auxiliary regressions associated to Model (14).

	$\mathbf{X_1}$	**DP**	**SF**	**DC**
VIFnc		2457.002	5.753	2512.562
VIFnc	60.0706	7424.705	6.008	8522.1308

To conclude, this model presents essential multicollinearity caused by the variables **DP** and **DC**. In this case, the problem will be mitigated by applying estimation methods other than OLS such as ridge regression (see, for example, Hoerl and Kennard [27], Hoerl et al. [28], Marquardt [29]), LASSO regression (see Tibshirani [30]), raise regression (see, for example, García et al. [31], Salmerón et al. [32], García and Ramírez [33], Salmerón et al. [34]), residualization (see, for example, York [35], García et al. [36]) or the elastic net regularization (see Zou and Hastie [37]).

6. Conclusions

The distinction between essential and nonessential multicollinearity and its diagnosis has not been not been adequately treated in either the scientific literature or in statistical software and this lack of information has led to mistakes in some relevant papers, for example Velilla [3] or Jensen and Ramirez [13]. This paper analyzes the detection of essential and nonessential multicollinearity from auxiliary centered and noncentered regressions, obtaining two complementary measures between them that are able to detect both kinds of multicollinearity. The relevance of the results is that they are obtained within an econometric context, encompassing the distinction between centered and noncentered models that is not only accomplished from a numerical perspective, as was the case presented, for example, in Salmerón Gómez et al. [12] or Salmerón Gómez et al. [10]. An undoubtedly interesting point of view of this situation is the one presented by Spanos [38] that stated: It is argued that many confusions in the collinearity literature arise from erroneously attributing symptoms of statistical misspecification to the presence of collinearity when the latter is misdiagnosed using unreliable statistical measures. That is, the distinction related to the econometric model provides confidence to the measures of detection and avoids the problems commented by Spanos.

From a computational point of view, this debate clarifies what is calculated when the VIF is obtained for centered and noncentered models. It also clarifies, see Section 2.3, what type of multicollinearity is detected (and why) when the uncentered VIF is calculated in a centered model. At the same time, a definition of nonessential multicollinearity is presented that generalizes the definition given by Marquardt and Snee [6]. Note that this generalization can be understood as a particular kind of essential multicollinearity:

A near-linear relation between two independent variables with light variability. However, it is shown that this kind of multicollinearity is not detected by the VIF, and for this reason, we consider it more appropriate to include it within the nonessential multicollinearity.

In relation to the application of the VIFnc, this paper shows that the VIFnc detects the essential and the generalized nonessential multicollinearity and even the traditional nonessential multicollinearity if it is calculated in a regression without the intercept but including the constant as an independent variable. Note that the VIF, although widely applied in many different fields, only detects the essential multicollinearity. This paper has also analyzed why the VIF is unable to detect the nonessential multicollinearity, and two conditions are presented as sufficient (but not required) to establish the existence of nonessential multicollinearity. Since these conditions, **C1** and **C2**, are based on the relevance of the intercept within the centered auxiliary regression to calculate the VIF, this scenario was compared to the measure proposed by Stewart [9], ι_j, to measure the relative importance of a variable within a multiple linear regression. It is shown that conditions **C1** and **C2** are preferable to the calculation of ι_j.

To summarize:

- A centered model can present essential, generalized nonessential and traditional nonessential collinearity (given by Marquardt and Snee [6]) while in a noncentered model only it is only possible to find the essential and the generalized nonessential collinearity.
- The VIF only detects the essential collinearity, the VIFnc detects the generalized nonessential and essential collinearity and the conditions **C1** and **C2** the traditional nonessential collinearity.
- When there is generalized nonessential collinearity it is understood that there is also traditional nonessential collinearity, but this is not detected by the conditions **C1** and **C2**. Thus, in this case it is necessary to use other alternative measures as the coefficient of variation of the condition number.

To conclude, in order to detect the kind of multicollinearity and its degree, the greatest number of measures must be used (variance inflation factors, condition number, correlation matrix and its determinant, coefficient of variation, conditions **C1** and **C2**, etc.) as in Section 5, and it is inefficient to limit oneself to the management of only a few. Similarly, it is necessary to know what kind of multicollinearity is capable of detecting each one of them.

Finally, the following will be interesting as future lines of inquiry:

- to establish the threshold for the VIFnc,
- to extend the Montecarlo simulation of Section 4.1 for models with $\boldsymbol{k > 3}$ regressors,
- a deeper analysis to conclude if the variable responsible for the existing linear relation can be identified as the one whose estimated coefficient is significantly different from zero in the auxiliary regression (see Example 6) and
- the development of a specific package in R Core Team [39] to perform the calculation of VIFnc and conditions **C1** and **C2**.

Author Contributions: Conceptualization, R.S.G. and C.G.G.; methodology, R.S.G.; software, R.S.G.; validation, R.S.G., C.G.G. and J.G.P.; formal analysis, R.S.G. and C.G.G.; investigation, R.S.G., C.G.G. and J.G.P.; resources, R.S.G., C.G.G. and J.G.P.; writing—original draft preparation, R.S.G. and C.G.G. ; writing—review and editing, R.S.G. and C.G.G.; supervision, J.G.P.; project administration, R.S.G.; funding acquisition, J.G.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by University of Almería.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Belsley, D.A. Demeaning conditioning diagnostics through centering. *Am. Stat.* **1984**, *38*, 73–77.
2. Marquardt, D.W. A critique of some ridge regression methods: Comment. *J. Am. Stat. Assoc.* **1980**, *75*, 87–91. [CrossRef]
3. Velilla, S. A note on collinearity diagnostics and centering. *Am. Stat.* **2018**, *72*, 140–146. [CrossRef]

4. Cohen, P.; West, S.G.; Aiken, L.S. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*; Psychology Press: London, UK, 2014.
5. Marquardt, D. You should standardize the predictor variables in your regression models. Discussion of: A critique of some ridge regression methods. *J. Am. Stat. Assoc.* **1980**, *75*, 87–91.
6. Marquardt, D.W.; Snee, R.D. Ridge regression in practice. *Am. Stat.* **1975**, *29*, 3–20.
7. O'Brien, R. A caution regarding rules of thumb for variance inflation factors. *Qual. Quant.* **2007**, *41*, 673–690. [CrossRef]
8. Salmerón, R.; García, C.; García, J. Variance Inflation Factor and Condition Number in multiple linear regression. *J. Stat. Comput. Simul.* **2018**, *88*, 2365–2384. [CrossRef]
9. Stewart, G. Collinearity and least squares regression. *Stat. Sci.* **1987**, *2*, 68–84. [CrossRef]
10. Salmerón Gómez, R.; Rodríguez, A.; García García, C. Diagnosis and quantification of the non-essential collinearity. *Comput. Stat.* **2020**, *35*, 647–666. [CrossRef]
11. Marquardt, D.W. [Collinearity and Least Squares Regression]: Comment. *Stat. Sci.* **1987**, *2*, 84–85. [CrossRef]
12. Salmerón Gómez, R.; García García, C.; García Pérez, J. Comment on A Note on Collinearity Diagnostics and Centering by Velilla (2018). *Am. Stat.* **2019**, 114–117. [CrossRef]
13. Jensen, D.R.; Ramirez, D.E. Revision: Variance inflation in regression. *Adv. Decis. Sci.* **2013**, *2013*, 671204. [CrossRef]
14. Grob, J. *Linear Regression*; Springer: Berlin, Germany, 2003.
15. Cook, R. Variance Inflations Factors. *R News Newsl. R Proj.* **2003**, *3*, 13–15.
16. Belsley, D.A. A guide to using the collinearity diagnostics. *Comput. Sci. Econ. Manag.* **1991**, *4*, 33–50.
17. Chennamaneni, P.R.; Echambadi, R.; Hess, J.D.; Syam, N. Diagnosing harmful collinearity in moderated regressions: A roadmap. *Int. J. Res. Mark.* **2016**, *33*, 172–182. [CrossRef]
18. Belsley, D.A.; Kuh, E.; Welsch, R.E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; Wiley: New York, NY, USA, 1980.
19. Theil, H. *Principles of Econometrics*; Wiley: New York, NY, USA, 1971; Volume 4.
20. Cook, R. [Demeaning Conditioning Diagnostics through Centering]: Comment. *Am. Stat.* **1984**, *38*, 78–79. [CrossRef]
21. García García, C.; Salmerón Gómez, R.; García García, C. Choice of the ridge factor from the correlation matrix determinant. *J. Stat. Comput. Simul.* **2019**, *89*, 211–231. [CrossRef]
22. Salmerón Gómez, R.; García García, C.; García García, J. A Guide to Using the R Package "multiColl" for Detecting Multicollinearity. *Comput. Econ.* **2020**. [CrossRef]
23. Chatterjee, S.; Hadi, A.S. *Regression Analysis by Example*; John Wiley & Sons: Hoboken, NY, USA, 2015.
24. Malinvaud, E. *Statistical Methods of Econometrics*; North Holland: New York, NY, USA, 1980.
25. Zhang, W.; Liu, L. A New Class of Biased Estimate in the Linear Regression Model. *J. Wuhan Univ. Nat. Sci. Ed.* **2006**, *52*, 281.
26. Kibria, B.; Lukman, A.F. A New Ridge-Type Estimator for the Linear Regression Model: Simulations and Applications. *Scientifica* **2020**, *2020*, 9758378 . [CrossRef]
27. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
28. Hoerl, A.; Kannard, R.; Baldwin, K. Ridge regression: Some simulations. *Commun. Stat. Theory Methods* **1975**, *4*, 105–123. [CrossRef]
29. Marquardt, D. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics* **1970**, *12*, 591–612. [CrossRef]
30. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
31. García, C.G.; Pérez, J.G.; Liria, J.S. The raise method. An alternative procedure to estimate the parameters in presence of collinearity. *Qual. Quant.* **2011**, *45*, 403–423. [CrossRef]
32. Salmerón, R.; García, C.; García, J.; López, M.d.M. The raise estimator estimation, inference, and properties. *Commun. Stat. Theory Methods* **2017**, *46*, 6446–6462. [CrossRef]
33. García, J.; Ramírez, D. The successive raising estimator and its relation with the ridge estimator. *Commun. Stat. Theory Methods* **2017**, *46*, 11123–11142. [CrossRef]
34. Salmerón, R.; Rodríguez, A.; García, C.; García, J. The VIF and MSE in Raise Regression. *Mathematics* **2020**, *8*, 605. [CrossRef]

35. York, R. Residualization is not the answer: Rethinking how to address multicollinearity. *Soc. Sci. Res.* **2012**, *41*, 1379–1386. [CrossRef]
36. García, C.; Salmerón, R.; García, C.; García, J. Residualization: Justification, properties and application. *J. Appl. Stat.* **2017**. [CrossRef]
37. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, *67*, 301–320. [CrossRef]
38. Spanos, A. Near-collinearity in linear regression revisited: The numerical vs. the statistical perspective. *Commun. Stat. Theory Methods* **2019**, 48, 5492–5516. [CrossRef]
39. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.

mathematics

MDPI

Article

Market Volatility of the Three Most Powerful Military Countries during Their Intervention in the Syrian War

Viviane Naimy [1], José-María Montero [2], Rim El Khoury [1,*] and Nisrine Maalouf [3]

[1] Faculty of Business Administration and Economics, Notre Dame University—Louaize, Zouk Mikayel, Zouk Mosbeh 72, Lebanon; vnaimy@ndu.edu.lb
[2] Department of Political Economy and Public Finance, Economic and Business Statistics, and Economic Policy, Faculty of Law and Social Sciences, University of Castilla-La Mancha, 45071 Toledo, Spain; Jose.mlorenzo@uclm.es
[3] Financial Risk Management—Faculty of Business Administration and Economics, Notre Dame University—Louaize, Zouk Mikayel, Zouk Mosbeh 72, Lebanon; nisrinemaalouf1@gmail.com
* Correspondence: rkhoury@ndu.edu.lb

Received: 8 April 2020; Accepted: 17 May 2020; Published: 21 May 2020

Abstract: This paper analyzes the volatility dynamics in the financial markets of the (three) most powerful countries from a military perspective, namely, the U.S., Russia, and China, during the period 2015–2018 that corresponds to their intervention in the Syrian war. As far as we know, there is no literature studying this topic during such an important distress period, which has had very serious economic, social, and humanitarian consequences. The Generalized Autoregressive Conditional Heteroscedasticity (GARCH (1, 1)) model yielded the best volatility results for the in-sample period. The weighted historical simulation produced an accurate value at risk (VaR) for a period of one month at the three considered confidence levels. For the out-of-sample period, the Monte Carlo simulation method, based on student t-copula and peaks-over-threshold (POT) extreme value theory (EVT) under the Gaussian kernel and the generalized Pareto (GP) distribution, overstated the risk for the three countries. The comparison of the POT-EVT VaR of the three countries to a portfolio of stock indices pertaining to non-military countries, namely Finland, Sweden, and Ecuador, for the same out-of-sample period, revealed that the intervention in the Syrian war may be one of the pertinent reasons that significantly affected the volatility of the stock markets of the three most powerful military countries. This paper is of great interest for policy makers, central bank leaders, participants involved in these markets, and all practitioners given the economic and financial consequences derived from such dynamics.

Keywords: GARCH; EGARCH; VaR; historical simulation approach; peaks-over-threshold; EVT; student t-copula; generalized Pareto distribution

1. Introduction

Political uncertainty occurs due to many factors like elections and changes in the government or parliament, changes in policies, strikes, minority disdain, foreign intervention in national affairs, and others. In many cases, these uncertainties lead to further complications affecting the economy and the financial market of the concerned country. Accordingly, the currency could devaluate, prices of assets, commodities, and stocks could fluctuate, and the growth of the economy could be hindered. From this perspective, countries strive to keep political risks controlled to be able to endure the cost or consequence of any sudden political unrest. This is one of the main reasons behind the intervention of powerful countries in the political and military affairs of less powerful countries, which is usually done at a high cost. This paper studies the impact of the intervention of the three most powerful military

countries in the world, namely, the United States, Russia, and China (Figure 1), in the Syrian war on their market volatility.

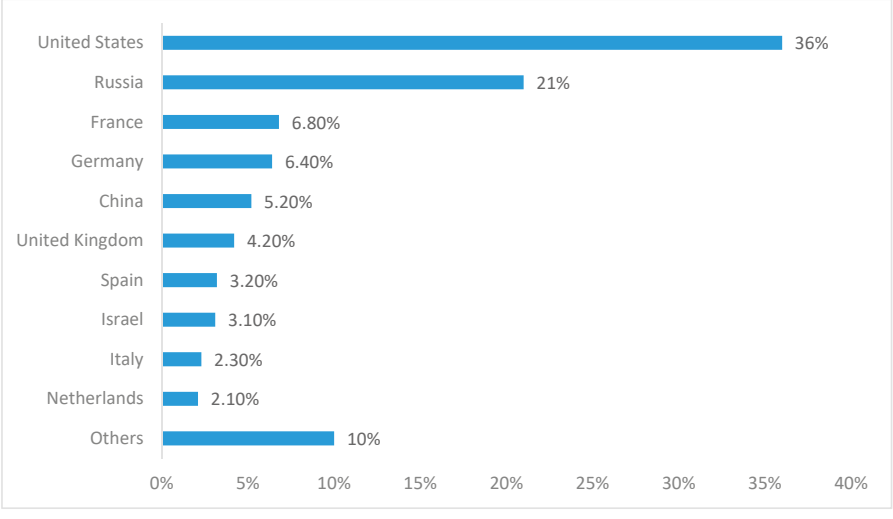


Figure 1. Global share of major arms exports by the 10 largest exporters, 2014–2018.

In March 2011, large peaceful protests broke in Syria to call for economic and political reforms with few armed protesters, leading to man arrests. Events evolved into violent acts using artillery and aircrafts, antigovernment rebels, terrorist and extremist attacks, suicide attacks, explosive operations, the intervention of foreign countries, chemical weapons, and others leading to a humanitarian crisis. In 2015, Russia started supporting the Syrian president through financial aid and military support [1]. In the meantime, the United States was providing support for the local Syrians. Later on, the United States and Russia increased their intervention in the war mainly through arms and aircrafts, each supporting their own political interests and allies. By the same token, China's involvement was shifting from humanitarian assistance and weapon exports [2] to armed forces and increased weapon exports to support its allies' objectives during this war [3].

Table 1 shows countries with the highest military spending in the world for 2016, 2017, and 2018. The U.S. spends the highest budget in the world on defense forces. This expenditure rounded up to USD 649 billion during 2018 based on information from the Stockholm International Peace Research Institute [4]. In fact, the defense spending of the United States alone is higher than the sum of that of the next eight countries in the ranking. These countries include China, Russia, Saudi Arabia, India, France, UK, Japan, and Germany. The country with the second highest defense expenditure is China with USD 250 billion in 2018 compared to USD 228 billion in 2017. As for Russia, its expenditure reached USD 61.4 billion in 2018 compared to USD 66.5 billion in 2017. Figure 1 represents the 10 largest arms exporters in the world between 2013 and 2017 [5]. Besides having the highest budgets for defense, the U.S. and Russia are also the top exporters of weapons, and China is among the top five worldwide countries. Based on these facts, the importance of the U.S., China, and Russia among military countries is highly reinforced. For this reason, we opted to study the dynamics of their financial markets to comprehend the risks and opportunities they might face, which would affect their worldwide exposure.

Table 1. Countries with the highest military spending worldwide in 2016–2018 (In Billion USD).

In USD Billion	2016	2017	2018
USA	600.1	605.8	648.8
China	216.0	227.8	250.0
Russia	69.2	66.5	61.4
Saudi Arabia	63.7	70.4	67.6
India	56.6	64.6	66.5
France	57.4	60.4	63.8

Source: Stockholm International Peace Research Institute (SIPRI), 2019.

To this end, measuring the effect of their intervention in the Syrian war on their financial market volatility is of great importance for policy makers, central bank leaders, analysts, and practitioners because there is a complete absence in the literature of studies that involve the volatility of the financial markets of the U.S., China, and Russia together. Many studies, however, explored the volatility of these countries during different periods and using different volatility models.

In his paper, Wei [6] forecasted the Chinese stock market volatility using non-linear Generalized Autoregressive Conditional Heteroscedasticity (GARCH) models such as the quadratic GARCH (QGARCH) and the Glosten, Jagannathan, and Runkle GARCH (GJR GRACH) models. The author studied seven-year data for the Shanghai Stock Exchange Composite (HSEC) and the Shenzhen Stock Exchange Component (ZSEC). The QGARCH outperformed the linear GARCH model. Furthermore, Lin and Fei [7] concluded that the nonlinear asymmetric power GARCH (APGARCH) model outperformed other GARCH models on different time scales in estimation of the "long memory property of the Shanghai and Shenzhen stock markets". Recently, Lin [8] studied the volatility of the SSE Composite Index using GARCH models during the period 2013–2017. The asymmetric exponential GARCH (EGARCH (1, 1)) model outperformed the symmetric ones in the forecasting results.

Value at risk (VaR), extreme value theory (EVT), and expected shortfall (ES) models were also used by Wang et al. [9], who implemented an EVT based VaR and ES to estimate the exchange rate risk of the Chinese currency (CNY). They found that the EVT-based VaR estimation produces accurate results for the currency exchange rate risks of EUR/CNY and JPY/CNY. However, EVT underestimated this risk for both exchange rates. Chen et al. [10] estimated VaR and ES by applying EVT on 13 worldwide stock indices. They concluded that China ranks first for VaR and ES with negative returns and ranks third for positive returns with high levels of risk.

A new strategy to estimate daily VaR based on the autoregressive fractionally integrated moving average model (ARFIMA), the multifractal volatility (MFV) model, and EVT was implemented by Wei et al. [11] for the Chinese stock market using high-frequency intraday quotes of the Shanghai Stock Exchange Component (SSEC). This hybrid ARFIMA-MFV-EVT strategy was compared to a number of popular linear and nonlinear GARCH-type-EVT models, i.e., the RiskMetrics, GARCH, IGARCH, and EGARCH models. Although GARCH-type models showed a good performance, VaR results obtained from the ARFIMA-MFV-EVT method outperformed several of them widely used in the literature. Furthermore, Hussain and Li [12] focused on the effect of extreme returns in stock markets on risk management by studying the SSEC index and by using the block maxima (Minima) method (BMM), instead of the popular peaks-over-threshold (POT) method, with various time intervals of extreme daily returns. Three well-known distributions in extreme value theory, i.e., generalized extreme value (GEV), generalized logistic (GL), and generalized Pareto distributions (GP), were employed to model the SSEC index returns. Results showed that GEV and GL distributions are found to be appropriate for the modeling of the extreme upward and downward market movements for China.

Another comparative study, conducted by Hou and Li [13], investigated the transmission of information between the U.S. and China's index futures markets using an asymmetric dynamic conditional correlation GARCH (DCC GARCH) approach. They found that the correlation between U.S. and Chinese index futures markets increases with the rise of negative shocks in these markets, and that the U.S. index futures market is more efficient in terms of price adjustment, since it is older

and more mature. On the other hand, Awartani and Corradi [14] focused on the role of asymmetries in the prediction of the volatility of the S&P 500 Composite Price Index. They examined the relative out-of-sample predictive ability of different GARCH-type models. First, they performed pairwise comparisons of various models against GARCH (1, 1). Then, they carried out a joint comparison of all models. They found that for the case of the one-step ahead pairwise comparison, GARCH (1, 1) is beaten by the asymmetric GARCH models. A similar finding applies to different longer forecast horizons. In the multiple comparison case, GARCH (1, 1) is only beaten when compared against the class of asymmetric GARCH. Another interesting finding is that the RiskMetrics exponential smoothing seems to be the worst model in terms of predictive ability. Furio and Climent [15] studied extreme movements in the return of S&P 500, FTSE 100, and NIKKEI 225 using GARCH-type models and EVT estimates. Results pointed out that more accurate estimates are derived from EVT calculations in both the in-sample and out-of-sample, when compared to less accurate estimates using the GARCH models.

As can be deduced from the review of the above literature, the question of how devastating wars, with indirect consequences all around the world, affect the volatility of financial markets of countries supporting them (and others, of course) might be of core importance for those directly or indirectly involved in such markets. Therefore, the main research question is the following: are the volatility dynamics of those countries affected by an event of the importance of the Syrian war? This paper fills this gap through evaluating the results of a number of traditional volatility models of the GARCH-type family and using EVT and historical simulation (HS) to estimate the VaR of these markets during the Syrian war period.

S&P 500 (Standard & Poor's), SSEC (Shanghai Stock Exchange Composite) and MICEX (Moscow Interbank Currency Exchange) are used to assess the financial markets' volatility of the U.S., China, and Russia, respectively. The period of study extends from 2015 to 2018. The in-sample period extends from 5 January 2015 until 30 December 2016 as it refers to the beginning of the direct and indirect intervention of the chosen countries in the war in Syria [1]; the out-of-sample interval is 3 January 2017–31 May 2018.

The paper is structured as follows: Section 2 reviews the methodology and the specificities of the applied econometric models, and Section 3 shows the estimated GARCH-type models considered and the selection process. This section also depicts the results related to the calculation of VaR using HS volatility and the "peaks-over-threshold" (POT) EVT model under the GP distribution. Section 4 concludes and discusses the empirical findings.

2. Econometric Models

As previously outlined, we use the GARCH (1, 1) and EGARCH (1, 1) as competing models to measure the volatility of the financial markets of the U.S., Russia, and China. GARCH models are commonly used by financial institutions to obtain volatility and correlation forecasts of asset and risk factor return. We use the symmetric normal GARCH given its strength to provide short- and medium-term volatility forecasts. We also use EGARCH, the asymmetric GARCH model, which is widely recognized in providing a better in-sample fit than other types of GARCH processes and avoids the need for any parameter constraints (see [16,17] for details on other GARCH-type models). The exponentially weighted moving average (EWMA) model is not used because it does not account for mean reversion and overvalue volatility after severe price fluctuation [18]. As said in the introductory section, for VaR estimation with high confidence intervals, we apply EVT [19], and more specifically GEV, GL, and GP distributions. We decided to use EVT because of its ability to provide good estimates and serve of help in situations where high confidence levels are needed, since EVT has proven to be a robust way of smoothing and extrapolating the tails of an empirical distribution [20]. The EVT implementation in this paper is based on a multivariate analysis to accurately measure the VaR of the portfolio composed of the U.S., Russia, and China stock markets. We also estimate the VaR of the portfolio using HS for comparison.

2.1. GARCH Model

The pioneering work of Engle [21], where the Autoregressive Conditional Heteroscedasticity (ARCH) model (that relates the current level of volatility to p past squared error terms) was introduced, constitutes the main pillar of modern financial econometrics. However, the ARCH strategy has some limitations, including the typically required 5–8 lagged error terms to adequately model conditional variance. That was the reason for this model to be generalized by Bollerslev [22], giving rise to the generalized ARCH (GARCH) model, by adding lagged conditional variance, which acts as a smoothing term. In practical terms, the GARCH (p, q) model builds on the ARCH (p) by including q lags of the conditional variance. Therefore, a GARCH specification uses the weighted average of long-run variance, the predicted variance for the current period, and any new information in this period, as captured by the squared residuals, to forecast a future variance. More specifically, the general GARCH (p, q) model is as shown in Equation (1):

$$\sigma_t^2 = \gamma V_L + \sum_{i=1}^{p} \alpha_i u_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \tag{1}$$

where σ_t^2 is the time $t-1$ conditional variance, V_L is the long run average variance, σ_{t-j}^2 are the lags of the conditional variance, and u_{t-i}^2 are the lagged squared error terms. $u_t = \sigma_t e_t$ with e_t *i.i.d.* $N(0, 1)$. Coefficients γ, α_i and β_j are the weights for V_L and the lags of the conditional variance and the squared error terms, respectively, and their estimates are obtained by Maximum Likelihood.

GARCH (1, 1) is the most used model of all GARCH models. It can be written as follows:

$$\sigma_t^2 = \gamma V_L + \alpha_1 u_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{2}$$

or, alternatively,

$$\sigma_t^2 = \omega + \alpha_1 u_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{3}$$

where $\omega = \gamma V_L$. Coefficients in the GARCH specification sum up to the unity and have to be restricted for the conditional variances to be uniformly positive. In the case of the GARCH (1, 1) such restrictions are: $\omega > 0$, $\alpha_1 \geq 0$ and $\beta_1 \geq 0$. In addition, the requirement for stationarity is $1 - \alpha_1 - \beta_1 > 0$. The unconditional variance can be shown to be $E\left(\sigma_t^2\right) = \omega/(1 - \alpha_1 - \beta_1)$.

2.2. EGARCH Model

The EGARCH model was proposed by Nelson [23] to capture the leverage effects observed in financial series and represents a major shift from the ARCH and GARCH models. The EGARCH specification does not model the variance directly, but its natural logarithm. This way, there is no need to impose sign restrictions on the model parameters to guarantee that the conditional variance is positive. In addition, EGARCH is an asymmetric model in the sense that the conditional variance depends not only on the magnitude of the lagged innovations but also on their sign. This is how the model accounts for the different response of volatility to the upwards and downwards movement of the series of the same magnitude. More specifically, EGARCH implements a function $g(e_t)$ of the innovations e_t, which are *i.i.d.* variables with zero mean, so that the innovation values are captured by the expression $|e_t| - E|e_t|$.

An EGARCH (p, q) is defined as:

$$\log \sigma_t^2 = \omega + \sum_{j=1}^{q} \beta_j \log \sigma_{t-j}^2 + \sum_{j=1}^{p} \theta_i g\left(e_{t-j}\right) \tag{4}$$

where $g(e_t) = \delta e_t + \alpha(|e_t| - E|e_t|)$ are variables *i.i.d.* with zero mean and constant variance. It is through this function that depends on both the sign and magnitude of e_t, that the EGARCH model captures

the asymmetric response of the volatility to innovations of different sign, thus allowing the modeling of a stylized fact of the financial series: negative returns provoke a greater increase in volatility than positive returns do.

The innovation (standardized error divided by the conditional standard deviation) is normally used in this formulation. In such a case, $E|e| = \sqrt{2/\pi}$ and the sequence $g(e_t)$ is time independent with zero mean and constant variance, if finite. In the case of Gaussianity, the equation for the variance in the model EGARCH (1, 1) is:

$$\log \sigma_t^2 = \omega + \beta \log \sigma_{t-1}^2 + \delta e_{t-1} + \alpha\left(|e_{t-1}| - \sqrt{\frac{2}{\pi}}\right). \tag{5}$$

Stationarity requires $\left|\beta\right| < 1$, the persistence in volatility is indicated by β, and δ indicates the magnitude of the leverage effect. δ is expected to be negative, which implies that negative innovations have a greater effect on volatility than positive innovations of the same magnitude. As in the case of the standard GARCH specification, maximum likelihood is used for the estimation of the model.

2.3. EVT

EVT deals with the stochastic behavior of extreme events found in the tails of probability distributions, and, in practice, it has two approaches. The first one relies on deriving block maxima (minima) series as a preliminary step and is linked to the GEV distribution. The second, referred to as the peaks over threshold (POT) approach, relies on extracting, from a continuous record, the peak values reached for any period during which values exceed a certain threshold and is linked to the GP distribution [24]. The latter is the approach used in this paper.

The generalized Pareto distribution was developed as a distribution that can model tails of a wide variety of distributions. It is based on the POT method which consists in the modelling of the extreme values that exceed a particular threshold. Obviously, in such a framework there are some important decisions to take: (i) the threshold, μ; (ii) the cumulative function that best fits the exceedances over the threshold; and (iii) the survival function, that is, the complementary of the cumulative function.

The choice of the threshold implies a trade-off bias-variance. A low threshold means more observations, which probably diminishes the fitting variance but probably increases the fitting bias, because observations that do not belong to the tail could be included. On the other hand, a high threshold means a fewer number of observations and, maybe, an increment in the fitting variance and a decrement in the fitting bias.

As for the distribution function that best fits the exceedances over the threshold, let us suppose that $F(x)$ is the distribution function for a random variable X, and that threshold μ is a value of X in the right tail of the distribution; let y denote the value of the exceedance over the threshold μ. Therefore, the probability that X lies between μ and $\mu + y$ $(y > 0)$ is $F(\mu + y) - F(\mu)$ and the probability for X greater than μ is $1 - F(\mu)$. Writing the exceedances (over a threshold μ) distribution function $F^{\mu}(y)$ as the probability that X lies between μ and $\mu + y$ conditional on $X > \mu$, and taking into account the identity linking the extreme and the exceedance: $X = Y + \mu$, it follows that:

$$F^{\mu}(y) = P(Y \leq y \big| X > \mu) = P(\mu < X \leq \mu + y \big| X > \mu) = \frac{F(x) - F(\mu)}{1 - F(\mu)} \tag{6}$$

and that

$$1 - F^{\mu}(y) = 1 - \frac{F(x) - F(\mu)}{1 - F(\mu)} = \frac{1 - F(x)}{1 - F(\mu)} \tag{7}$$

In the case that the parent distribution F is known, the distribution of threshold exceedances also would be known. However, this is not the practical situation, and approximations that are broadly applicable for high values of the threshold are sough. Here is where Pickands–Balkema–de Haan

theorem ([25,26]) comes into play. Once the threshold has been estimated, the conditional distribution $F^{\mu}(y)$ converges to the GP distribution. It is known that $F^{\mu}(y) \to G_{\xi,\sigma}(y)$ as $\mu \to \infty$, with

$$G_{\xi,\sigma}(y) = \begin{cases} 1 - \left(1 + \xi \frac{y}{\sigma}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ 1 - e^{-\frac{y}{\sigma}} & \text{if } \xi = 0 \end{cases} \tag{8}$$

where $\sigma > 0$ and $y \geq 0$ if $\xi \geq 0$ and $0 \leq y \leq -\sigma/\xi$ if $\xi < 0$. ξ is a shape parameter that determines the heaviness of the tail of the distribution, and σ is a scale parameter. When $\xi = 0$, $G_{\xi,\sigma}(y)$ reduces to the exponential distribution with expectation $\exp(\sigma)$; in the case that $\xi < 0$, it becomes a Uniform $(0, \sigma)$; finally, $\xi > 0$ leads to the Pareto distribution of the second kind [27]. In general, ξ has a positive value between 0.1 and 0.4. The GP distribution parameters are estimated via maximum likelihood.

Once the maximum likelihood estimates are available, a specific GP distribution function is selected, and an analytical expression for *VaR* with a confidence level q can be defined as a function of the GP distribution parameters:

$$VaR_{\hat{q}} = \mu + \frac{\hat{\sigma}(\mu)}{\hat{\xi}}\left(\frac{N}{N_{\mu}}(1-q)^{-\hat{\xi}} - 1\right) \tag{9}$$

where N is the number of observations in the left tail and N_{μ} is the number of excesses beyond the threshold μ.

$$VaR_{\hat{q}} = \mu + \frac{\hat{\sigma}(\mu)}{\hat{\xi}}\left(\frac{N}{N_{\mu}}(1-q)^{-\hat{\xi}} - 1\right) \tag{10}$$

3. Results

3.1. Descriptive Statistics

Data for 3 years were extracted from the Bloomberg platform for the three selected stock market indices and were manipulated to derive the return from the closing prices corresponding to each index. For the in-sample period, 458 daily observations were studied compared to 315 for the out-of-sample forecast period. It is important to note that in November 2017, the name of the MICEX index (composed of Russian stocks of the top 50 largest issues in the Moscow Exchange) was officially changed to the MOEX Russia Index, representing the "Russian stock market benchmark" [28]. Table 2 lists the descriptive statistics of S&P 500, SSEC, and MICEX for the in- and out-of-sample periods. Surprisingly, S&P 500 and the MICEX behaved similarly in terms of return during the in-sample and out-of-sample periods.

Table 2. Descriptive Statistics of S&P 500, SSEC, and MICEX: 5 January 2015–30 December 2016 and 3 January 2017–31 May 2018.

Stock Markets	S&P 500 (In-Sample)	SSEC (In-Sample)	MICEX (In-Sample)	S&P 500 (Out-of-Sample)	SSEC (Out-of-Sample)	MICEX (Out-of-Sample)
Mean	0.022%	−0.017%	0.017%	0.060%	−0.001%	0.054%
Standard Deviation	1.0%	2.10%	1.2%	0.711%	0.774%	1.025%
Skewness	1.0%	2.10%	1.2%	0.711%	0.774%	1.025%
Kurtosis	3.33	4.62	0.97	7.93	4.83	12.45
Median	0.01%	0.10%	0.04%	0.06%	0.08%	−0.05%
Minimum	−4.0%	−10.8%	−4.4%	−4.18%	−4.14%	−8.03%
Maximum	4.7%	6.0%	4.4%	2.76%	2.15%	3.89%
1st Quartile	−0.41%	−0.66%	−0.65%	−0.18%	−0.35%	−0.51%
3rd Quartile	0.49%	0.87%	0.85%	0.34%	0.40%	0.54%

In reference to the two chosen time periods, the skewness of the returns of the three indices is close to 0 and the returns display excess in kurtosis. This implies that the distributions of returns are

not normal as confirmed by Jarque-Bera normality tests (Table 3). The distributions of returns are stationary according to the augmented Dicky–Fuller (ADF) test applied to the three indices (Table 4).

Table 3. Jarque-Bera Normality Test of S&P 500, SSEC, and MICEX.

Stock Markets	In-Sample		Out-of-Sample	
	Score	***p*-Value**	**Score**	***p*-Value**
S&P 500	198.66	0.000	865.78	0.000
SSEC	483.75	0.000	352.63	0.000
MICEX	18.12	0.000	2027.03	0.000

Note: *p*-value refers to Jarque–Bera normality test, Ho: the index return is normally distributed.

Table 4. ADF Stationarity Test.

	Critical Values at 5%	STAT	*p*-Value	STAT	*p*-Value	STAT	*p*-Value
		S&P 500		**SSEC**		**MICEX**	
No Constant	−1.9	−28.4	0.001	−12.8	0.001	−26.9	0.001
Constant Only	−2.9	−28.4	0.001	−12.8	0.001	−11.6	0.001
Constant and Trend	−1.6	−28.4	0.000	−12.7	0.000	−11.6	0.000
Constant, Trend, and Trend^2	−1.6	−28.4	0.000	−12.7	0.000	−11.6	0.000

Note: ADF *p*-value refers to the augmented Dickey–Fuller unit root test, Ho: the index return has a unit root.

3.2. GARCH (1, 1) and EGARCH (1, 1) Results

GARCH (1, 1) and EGARCH (1, 1) parameters were estimated using the daily returns of each index. Results from the normal distribution, the student's t-distribution and the Generalized Error Distribution (GED) were derived. The goodness of fit test and residual analysis were then performed to ensure that the assumptions of the applied models were all met. The model parameters were estimated by maximum likelihood. Table 5 presents a summary of such estimates for GARCH (1, 1) and EGARCH (1,1).

Table 5. Estimates of the Parameters of GARCH (1,1) and EGARCH (1,1) (with GED).

	S&P 500	**SSEC**	**MICEX**
		GARCH (1, 1)	
Long-run mean (μ)	0.000249	−0.00009	0.00080
Omega (ω)	8.5218×10^{-6}	1.1035×10^{-6}	2.9655×10^{-6}
ARCH component (α)	0.1972	0.0569	0.0653
GARCH component (β)	0.7105	0.9331	0.9065
		EGARCH (1, 1)	
Long-run mean (μ)	0.00030	−0.00009	0.00142
Omega (ω)	−0.72879	−0.08598	−0.87318
ARCH component (α)	0.04766	0.20530	0.30477
Leverage coefficient (δ)	−0.29149	−0.01362	0.05345
GARCH component (β)	0.92471	0.98689	0.90237

Note: The GED was selected after checking the average, standard deviation, skewness, kurtosis, the noise, and ARCH tests corresponding to the three distributions.

3.3. Best Volatility Model Selection for the In- and Out-of-Sample Periods

The root mean square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE) are implemented to choose the best volatility model. For S&P 500, we compared the estimated volatility to the implied volatility. However, this was not possible for SSEC and MICEX due the absence of data. Results depicted in Table 6 reveal the superiority of GARCH (1, 1) in estimating the volatility of the three countries in the in-sample period, which coincides with the peak period of the Syrian war. As for the out-of-sample period, while GARCH (1, 1) ranks first for S&P 500, EGARCH (1, 1) ranks first for the SSEC and MICEX with a difference in RMSE of around 0.002 and 0.01 units respectively as compared to GARCH (1, 1). Volatilities estimated with the superior volatility models in comparison to the realized volatilities for the out-of-sample period are depicted in Figure 2.

Table 6. In-Sample Period Error Statistics.

	RMSE	**Rating**	**MAE**	**Rating**	**MAPE**	**Rating**
S&P 500 during In-Sample Period (from 5 January 2015 till 30 December 2016)						
Implied Vol.	0.047694	2	0.039028	2	0.00377	3
GARCH (1, 1)	0.040995	1	0.028954	1	0.002376	1
EGARCH (1, 1)	0.049947	3	0.039066	3	0.003261	2
SSEC during In-Sample Period (from 5 January 2015 till 30 December 2016)						
GARCH (1, 1)	0.09567	1	0.080424	1	0.002879	1
EGARCH (1, 1)	0.11588	2	0.090075	2	0.003134	2
MICEX during In-Sample Period (from 5 January 2015 till 30 December 2016)						
GARCH (1, 1)	0.184631	1	0.173261	1	0.004775	1
EGARCH (1, 1)	0.188566	2	0.179723	2	0.004997	2
S&P 500 during Out-of-Sample (from 3 January 2017 till 31 May 2018)						
Implied Vol.	0.04809	3	0.04325	3	0.006092	3
GARCH (1, 1)	0.040024	1	0.035135	1	0.004896	1
EGARCH (1, 1)	0.047078	2	0.041034	2	0.005689	2
SSEC during Out-of-Sample (from 3 January 2017 till 31 May 2018)						
GARCH (1, 1)	0.08478	2	0.080437	2	0.004033	2
EGARCH (1, 1)	0.07113	1	0.063535	1	0.00322	1
MICEX during Out-of-Sample (from 3 January 2017 till 31 May 2018)						
GARCH (1, 1)	0.071894	2	0.064616	2	0.002909	2
EGARCH (1, 1)	0.069381	1	0.061135	1	0.00271	1

Note: $RME = \sqrt{\sum_{t=1}^{n} (f-Y)^2/n}$; $MAE = \sum_{t=1}^{n} |f-Y|/n$; $MAPE = 100 \sum_{t=1}^{n} \left|\frac{f-Y}{Y}\right|/n$, where n is the number of periods, Y is the true value and f is the prediction value. The best model is the one that has a minimum value of *RMSE*, *MAE* and *MAPE*.

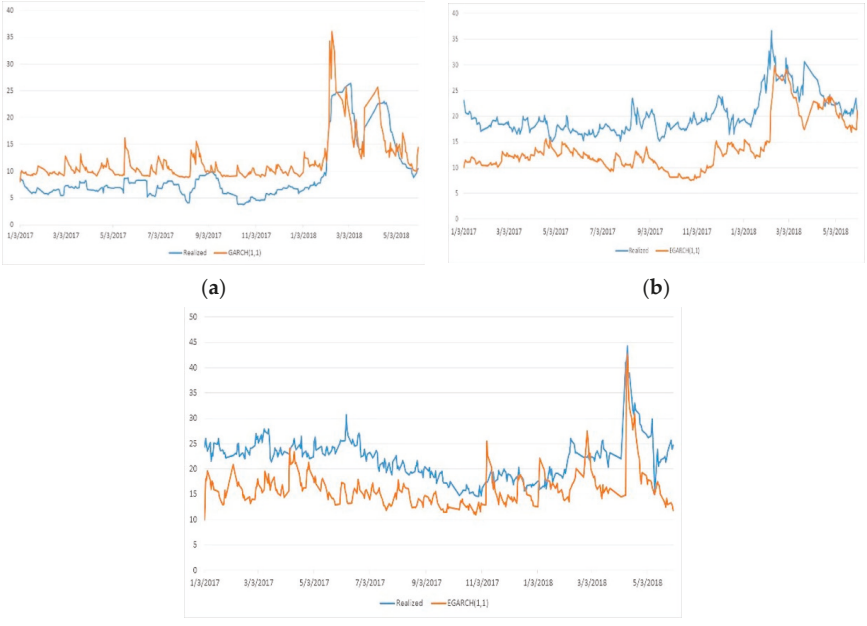


Figure 2. Realized Volatility vs. Volatilities Estimated with the Superior Volatility Models–Out-of-Sample Period: (**a**) for S&P 500. (**b**) for SSEC. (**c**) for MICEX.

3.4. VaR Output

Based on the superior model corresponding to each index, a portfolio of volatility updates was established for each sample period. First, the historical simulation approach was implemented. It involved incorporating volatility in updating the historical return. Because the volatility of a market variable may vary over time, we modified the historical data to reflect the variation in volatility. This approach uses the variation in volatility in a spontaneous way to estimate VaR by including more recent information. Second, a Monte Carlo simulation method including student t-copula and EVT was applied to the created portfolio composed of the three markets to estimate VaR with different confidence levels. The filtered residuals of each return series were extracted using EGARCH. The Gaussian kernel estimate was used for the interior marginal cumulative distribution function (CDF) and the generalized Pareto distribution (GP) was applied to estimate the upper and lower tails. The student t-copula was also applied to the portfolio's data in order to reveal the correlation among the residuals of each index. This process led to the estimation of the portfolio's VaR over a horizon of one month and confidence levels of 90%, 95%, and 99%. Table 7 summarizes all the VaR estimates calculated for the in-sample and out-of-sample periods using HS and EVT compared to the Real VaR. The visual illustrations of the relevant outcomes related to the logarithmic returns of the selected stock indices, the auto-correlation function (ACF) of returns and of the squared returns, the filtered residuals and the filtered conditional standard deviation, the ACF of standardized residuals, and the upper tail of standardized residuals for both periods, are presented in Appendix A (Figures A1 and A2).

Table 7. VaR Summary Results.

Outcome	HS (Volatility Weighted)	EVT	Real VaR
	In-Sample		
90% VaR	0.68%	2.93%	1.31%
95% VaR	1.03%	4.83%	1.68%
99% VaR	2.31%	8.39%	2.38%
	Out-of-Sample		
90% VaR	0.48%	3.76%	0.73%
95% VaR	0.71%	5.47%	0.94%
99% VaR	1.65%	9.60%	1.33%

It is apparent that VaR with a confidence level of 99%, using HS and EVT, overrates the risk for the three countries during both periods. Furthermore, the HS VaR results are closer to the Real VaR results compared to those of the EVT VaR. This is not altogether surprising since the EVT method is concerned with studying the behavior of extremes within these markets rather than simply fitting the curve. Therefore, the above output represents a benchmark that can be extrapolated beyond the data during stress periods.

4. Discussion

This paper revealed original common points among the most powerful military countries in the world regarding the behavior of their financial markets during the period 2015–2018, which corresponds to their intervention in the Syrian war. First, the returns of S&P 500 and MICEX were quite similar during the in-sample and out-of-sample periods. Second, the GARCH (1, 1) was found to be the best volatility model for the in-sample period for S&P 500, MICEX, and SSEC, outperforming EGARCH (1, 1). The incorporation of the GARCH (1, 1) specification to the HS produced an accurate VaR for a period of one month, at the three confidence levels, compared to the real VaR.

EVT VaR results are consistent with those found by Furio and Climent [15] and Wang et al. [9], who highlighted the accuracy of studying the tails of loss severity distribution of several stock markets. Furthermore, part of our results corroborates the work of Peng et al. [29], who showed that EVT GP distribution is superior to certain GARCH models implemented on the Shanghai Stock Exchange Index.

The GP distribution highlighted the behavior of "extremes" for the U.S., Russian, and Chinese financial markets, which is of great importance since it emphasizes the risks and opportunities inherent to the dynamics of their markets and also underlines the uncertainty corresponding to their worldwide exposure.

Expected EVT VaR values of 3.76%, 5.47%, and 9.60%, at 90%, 95%, and 99% confidence levels, respectively (for the out-of-sample period), might appear overstated. However, uncertainty layers are all the way inherent and our results are, naturally, subject to standard error. For comparison purposes, and in order to make a relevant interpretation of the tail distribution of returns corresponding to these markets, we opted to derive the EVT VaR of a portfolio of stock indices pertaining to non-military countries, namely Finland, Sweden, and Ecuador, for the same out-of-sample period of study (January 2017 to May 2018). These countries were chosen randomly based on the similarities in income groups when compared to the selected military countries. While Finland and Sweden are both classified as high income like the U.S., Ecuador is classified as upper middle income like Russia and China [30]. Also, the selection of these non-military countries follows the same structure of the capital market development found at the military countries which, when combined, find their average ratio of stock market capitalization to GDP is 77.23%, compared to 76.46% for Finland, Sweden, and Ecuador [30,31]. Finally, when comparing the ratio of private credit ratio to the stock market capitalization ratio corresponding to each country, we notice that the former is higher than the latter for Ecuador, Sweden, Finland, Russia, and China [32]. Only the U.S. depends mostly on its stock market to finance its

economy. The portfolio is composed of the OMX Helsinki index, the ECU Ecuador General index, and the OMX 30 Sweden index. The GP distribution was used to estimate the upper and lower tails. Remarkably, EVT VaR results were 2.23%, 3.49%, and 6.45% at the 90%, 95%, and 99% confidence levels, respectively, well below the estimates found for the U.S., Russian, and Chinese stock markets. Consequently, it can be concluded that the intervention in the Syrian war may have been one of the latent and relevant factors that affected the volatility of the stock markets of the selected military countries. This conclusion is reinforced by the fact that the EVT VaR was higher by 40%, 26%, and 32%, at the 90%, 95%, and 99% confidence levels, respectively, compared to the VaR of the portfolio constituted of the selected non-military countries. However, it can neither be deduced nor confirmed that the intervention in the Syrian war is the sole source, and more specifically, the trigger of the significant increase in the volatility of the American, Russian, and Chinese stock markets. Answering this question requires further lines of future research that involves incorporating a number of control covariates and using a different modeling methodology.

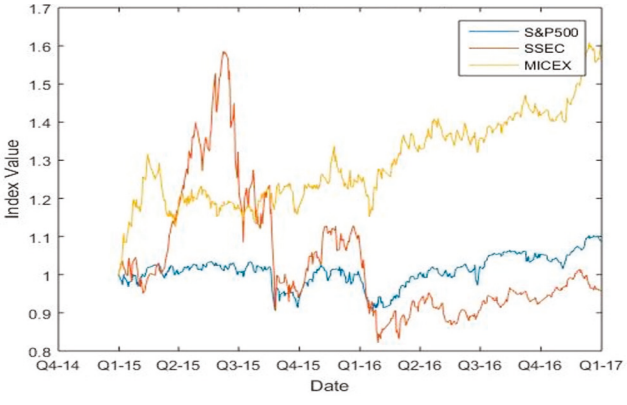
Although we covered a significant number of observations, our results are subject to errors; it is never possible to have enough data when implementing the extreme value analysis, since the tail distribution inference remains less certain. Introducing hypothetical losses to our historical data to generate stress scenarios is of no interest to this study and falls outside its main objective. It would be interesting to repeat the same study with the same selected three military countries during the period 2018–2020, which is expected to be the last phase of the Syrian war given that the Syrian army reached back to the border frontier with Turkey and that the Syrian Constitutional Committee Delegates launched meetings in Geneva to hold talks on the amendment of Syria's constitution.

Author Contributions: Conceptualization, V.N.; methodology, V.N., N.M., and J.-M.M.; formal analysis, J.-M.M.; resources, N.M.; writing—original draft preparation, V.N., N.M., and J.-M.M.; writing—review and editing, V.N., J.-M.M., and R.E.K.; visualization, R.E.K.; supervision, V.N.; project administration, J.-M.M. and V.N.; funding acquisition, J.-M.M. All authors have read and agreed to the published version of the manuscript.

Funding: José-María Montero benefited from the co-funding by the University of Castilla-La Mancha (UCLM) and the European Fund for Regional Development (EFRD) to the Research Group "Applied Economics and Quantitative Methods" (ECOAPP&QM): Grant 2019-GRIN-26913 2019.

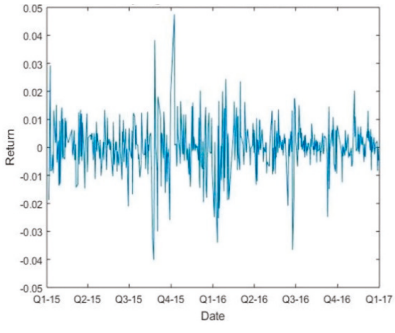
Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

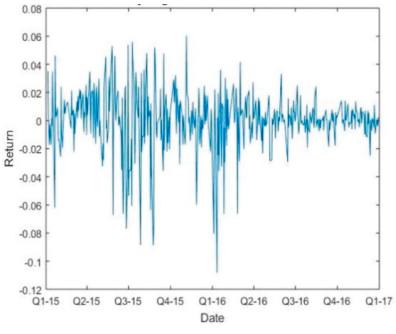


Relative Daily Index Closings of the In-Sample Portfolio

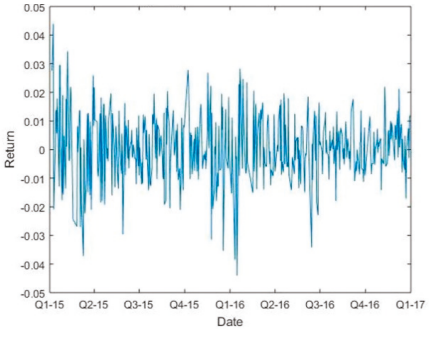
Figure A1. *Cont.*



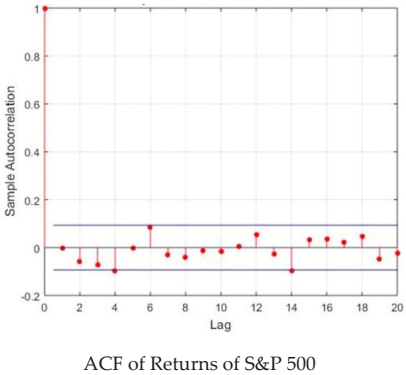
Daily Logarithmic Returns of S&P 500



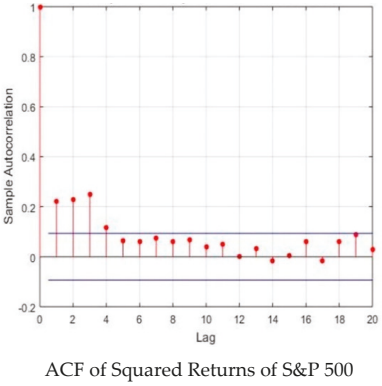
Daily Logarithmic Returns of SSEC



Daily Logarithmic Returns of MICEX

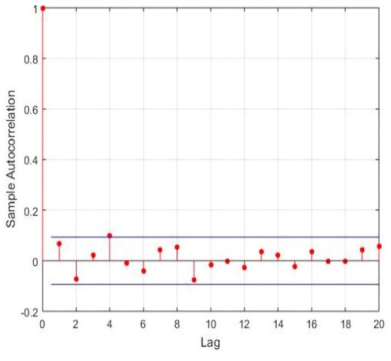


ACF of Returns of S&P 500

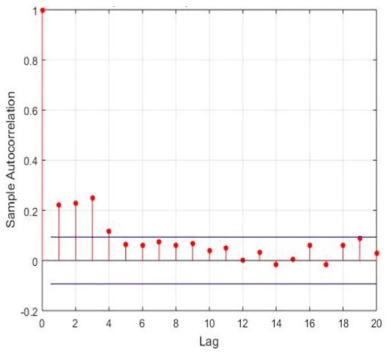


ACF of Squared Returns of S&P 500

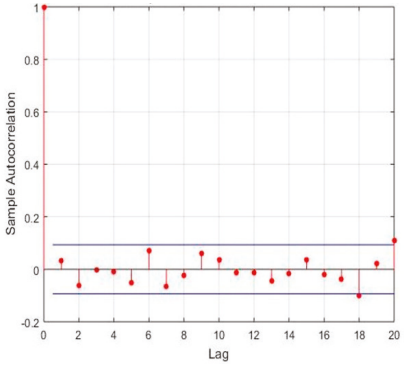
Figure A1. *Cont.*



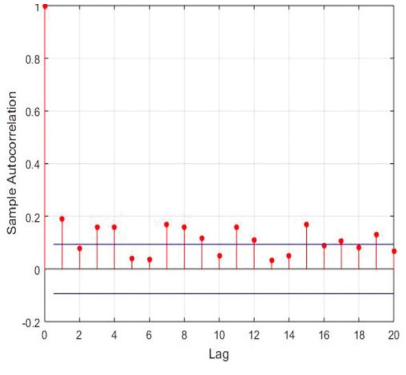
ACF of Returns of SSEC



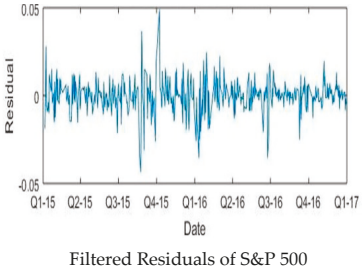
ACF of Squared Returns of SSEC



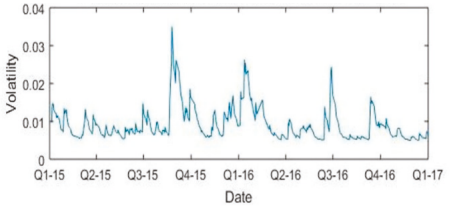
ACF of Returns of MICEX



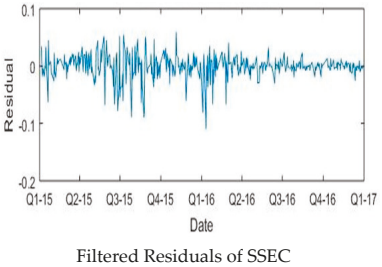
ACF of Squared Returns of MICEX



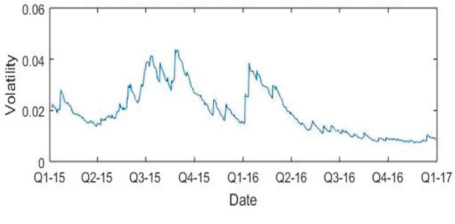
Filtered Residuals of S&P 500



Filtered Conditional Standard Deviation of S&P 500

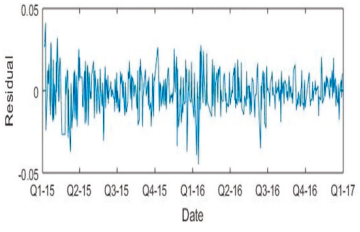


Filtered Residuals of SSEC

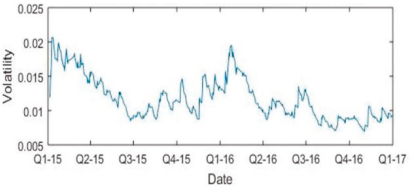


Filtered Conditional Standard Deviation of SSEC

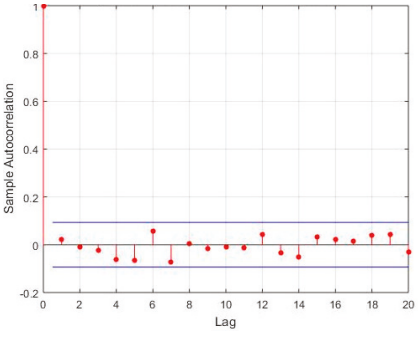
Figure A1. *Cont.*



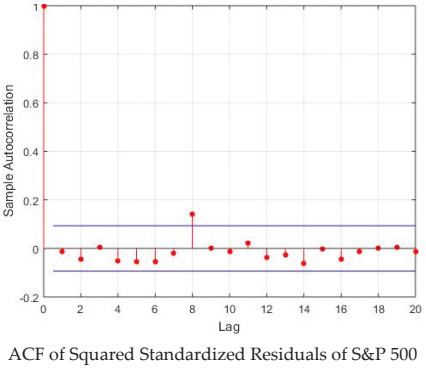
Filtered Residuals of MICEX



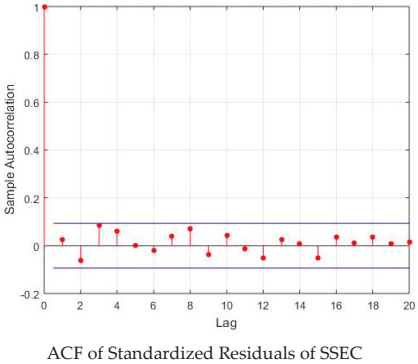
Filtered Conditional Standard Deviation of MICEX



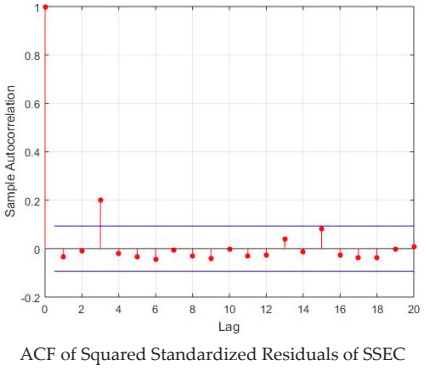
ACF of Standardized Residuals of S&P 500



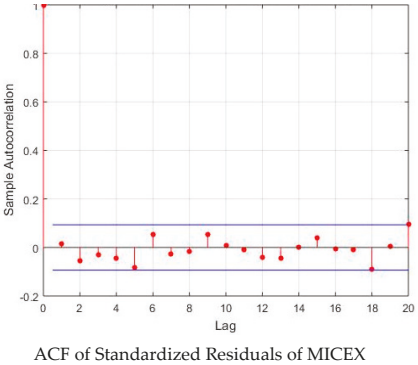
ACF of Squared Standardized Residuals of S&P 500



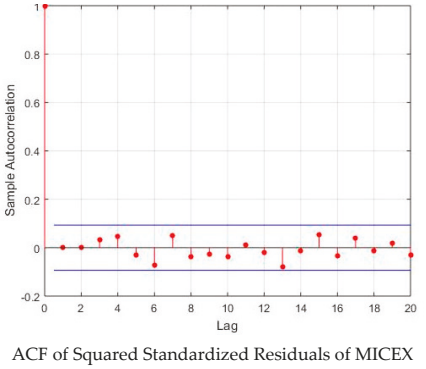
ACF of Standardized Residuals of SSEC



ACF of Squared Standardized Residuals of SSEC

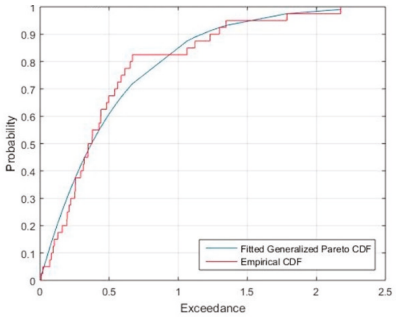


ACF of Standardized Residuals of MICEX

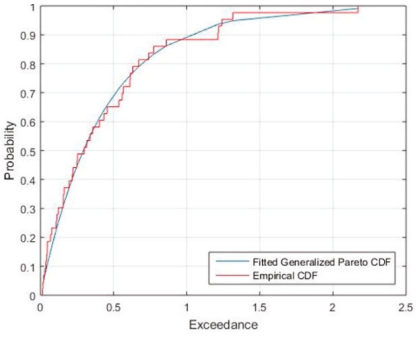


ACF of Squared Standardized Residuals of MICEX

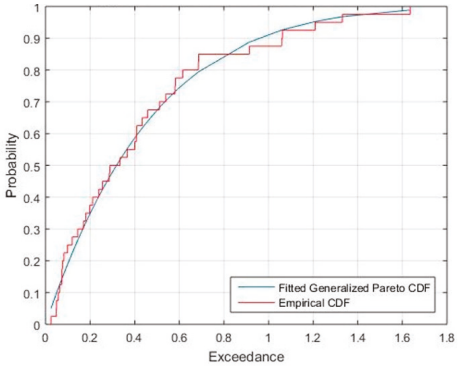
Figure A1. *Cont.*



S&P 500 Upper Tail of Standardized Residuals

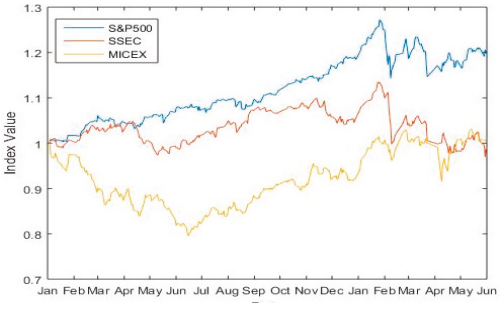


SSEC Upper Tail of Standardized Residuals



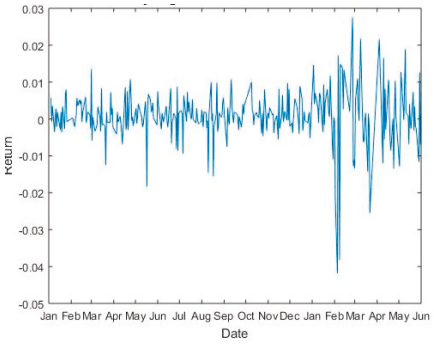
MICEX Upper Tail of Standardized Residuals

Figure A1. In-Sample VaR Figures.

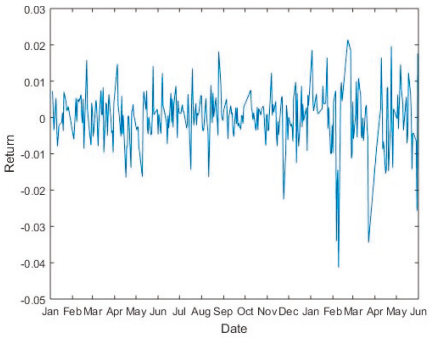


Relative Daily Index Closings of the Out-of-Sample Portfolio

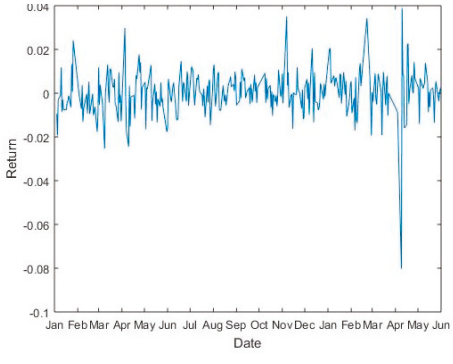
Figure A2. *Cont.*



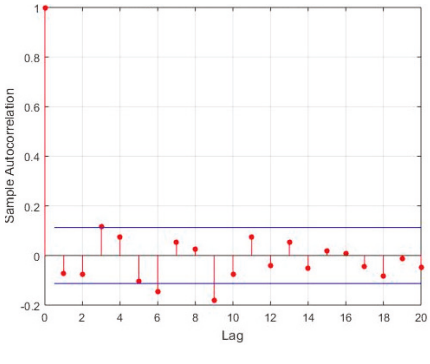
Daily Logarithmic Returns of S&P 500



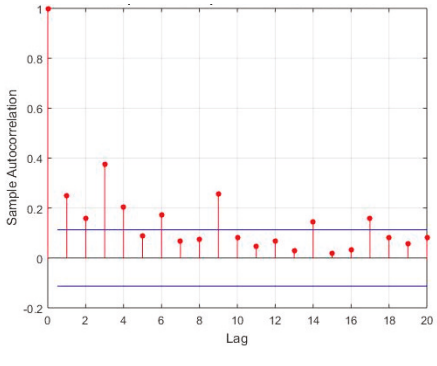
Daily Logarithmic Returns of SSEC



Daily Logarithmic Returns of MICEX

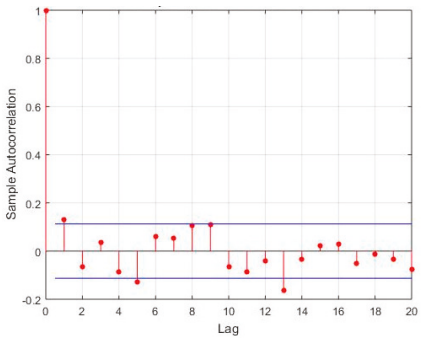


ACF of Returns of S&P 500

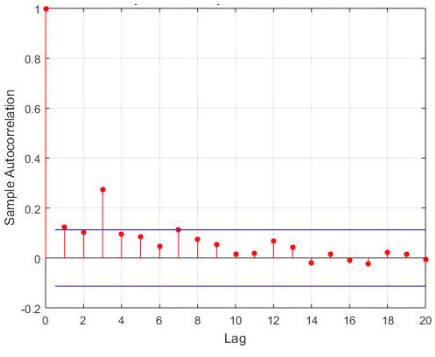


ACF of Squared Returns of S&P 500

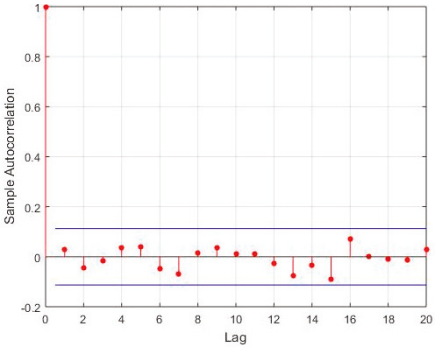
Figure A2. *Cont.*



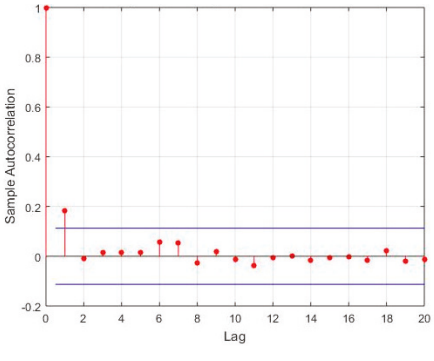
ACF of Returns of SSEC



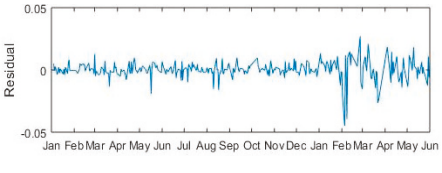
ACF of Squared Returns of SSE



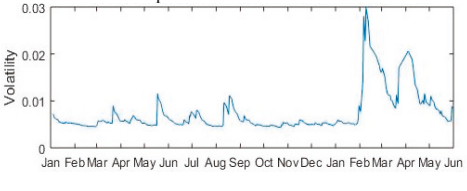
ACF of Returns of MICEX



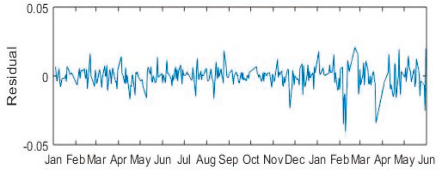
ACF of Squared Returns of MICEX



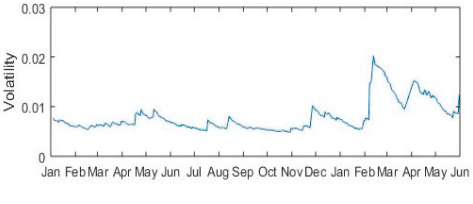
Filtered Residuals of S&P 500



Filtered Conditional Standard Deviation of S&P 500

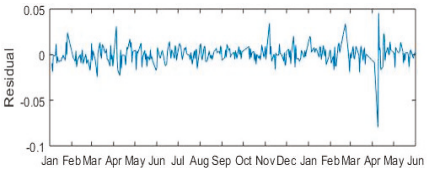


Filtered Residuals of SSEC

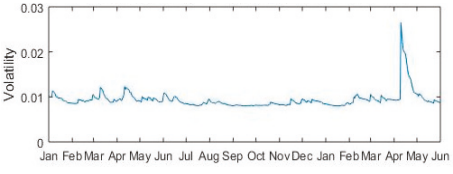


Filtered Conditional Standard Deviation of SSEC

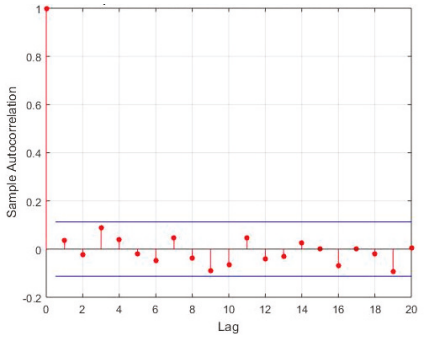
Figure A2. *Cont.*



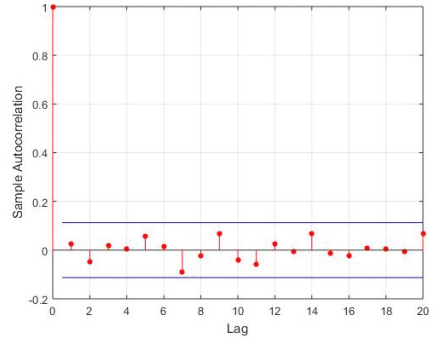
Filtered Residuals of MICEX



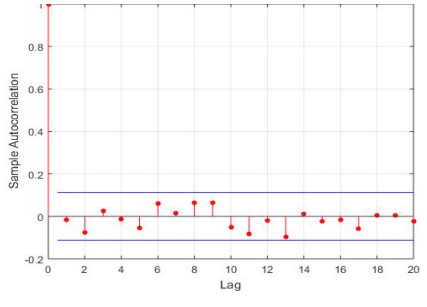
Filtered Conditional Standard Deviation of MICEX



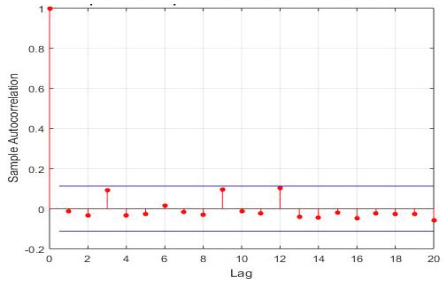
ACF of Standardized Residuals of S&P 500



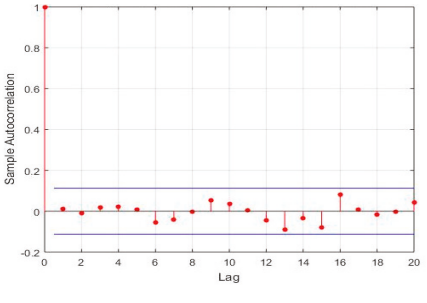
ACF of Squared Standardized Residuals of S&P 500



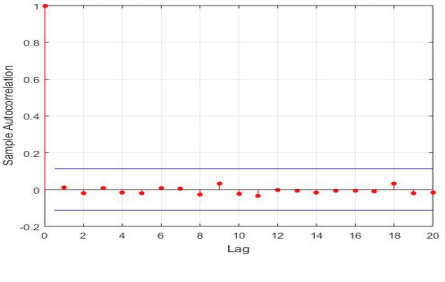
ACF of Standardized Residuals of SSEC



ACF of Squared Standardized Residuals of SSEC

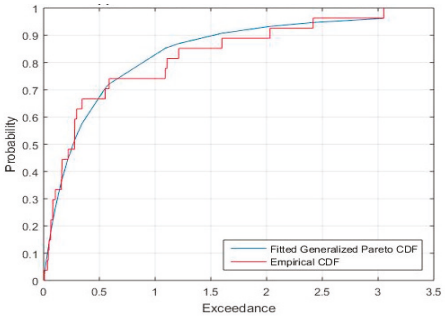


ACF of Standardized Residuals of MICEX

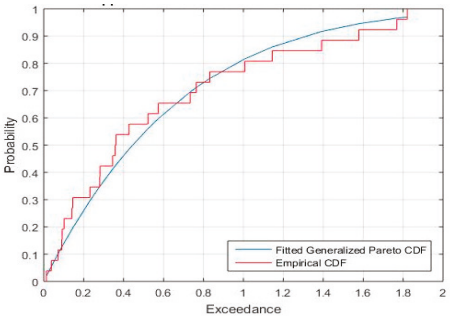


ACF of Squared Standardized Residuals of MICEX

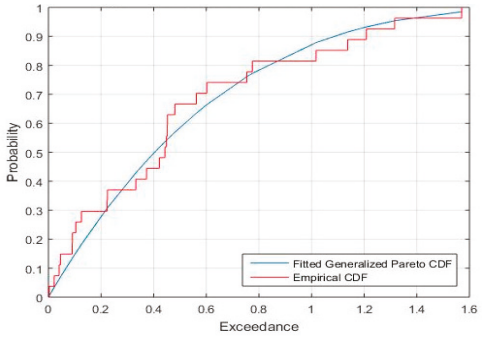
Figure A2. *Cont.*



S&P 500 Upper Tail of Standardized Residuals



SSEC Upper Tail of Standardized Residuals



MICEX Upper Tail of Standardized Residuals

Figure A2. Out-of-Sample VaR Figures.

References

1. Humud, C.E.; Blanchar, C.M.; Nikitin, M.B.D. *Armed Conflict in Syria: Overview and U.S. Response*; CRS: Washington, DC, USA, 2017.
2. Swaine, M. *Chinese Views of the Syrian Conflict*; Carnegie Endowment for International Peace: Washington, DC, USA, 2012.
3. O'Conor, T. China May Be the Biggest Winner of All If Assad Takes over Syria. *Newsweek*, 19 January 2018.
4. SIPRI. *Trends in Military Expenditures, 2018*; SIPRI: Solna, Sweden; Stockholm, Sweden, 2019.
5. SIPRI. *Trends in International Arms Transfers, 2018*; SIPRI: Solna, Sweden; Stockholm, Sweden, 2019.
6. Wei, W. Forecasting Stock Market Volatility with Non-Linear GARCH Models: A Case for China. *Appl. Econ. Lett.* **2002**, *9*, 163–166. [CrossRef]
7. Lin, X.; Fei, F. Long Memory Revisit in Chinese Stock Markets: Based on GARCH-Class Models and Multiscale Analysis. *Econ. Model.* **2013**, *31*, 265–275. [CrossRef]
8. Lin, Z. Modelling and Forecasting the Stock Market Volatility of SSE Composite Index Using GARCH Models. *Future Gener. Comput. Syst.* **2018**, *79*, 960–972. [CrossRef]
9. Wang, Z.; Wu, W.; Chen, C.; Zhou, Y. The Exchange Rate Risk of Chinese Yuan: Using VaR and ES Based on Extreme Value Theory. *J. Appl. Stat.* **2010**, *37*, 265–282. [CrossRef]
10. Chen, Q.; Giles, D.E.; Feng, H. The Extreme-Value Dependence between the Chinese and Other International Stock Markets. *Appl. Financ. Econ.* **2012**, *22*, 1147–1160. [CrossRef]
11. Wei, Y.; Chen, W.; Lin, Y. Measuring Daily Value-at-Risk of SSEC Index: A New Approach Based on Multifractal Analysis and Extreme Value Theory. *Phys. A Stat. Mech. Appl.* **2013**, *392*, 2163–2174. [CrossRef]

12. Hussain, S.I.; Li, S. Modeling the Distribution of Extreme Returns in the Chinese Stock Market. *J. Int. Financ. Mark. Inst. Money* **2015**, *34*, 263–276. [CrossRef]
13. Hou, Y.; Li, S. Information Transmission between U.S. and China Index Futures Markets: An Asymmetric DCC GARCH Approach. *Econ. Model.* **2016**, *52*, 884–897. [CrossRef]
14. Awartani, B.M.A.; Corradi, V. Predicting the Volatility of the S&P-500 Stock Index via GARCH Models: The Role of Asymmetries. *Int. J. Forecast.* **2005**, *21*, 167–183. [CrossRef]
15. Furió, D.; Climent, F.J. Extreme Value Theory versus Traditional GARCH Approaches Applied to Financial Data: A Comparative Evaluation. *Quant. Financ.* **2013**, *13*, 45–63. [CrossRef]
16. Trinidad Segovia, J.E.; Fernández-Martínez, M.; Sánchez-Granero, M.A. A Novel Approach to Detect Volatility Clusters in Financial Time Series. *Phys. A Stat. Mech. Appl.* **2019**, *535*, 122452. [CrossRef]
17. Ramos-requena, J.P.; Trinidad-segovia, J.E.; Sánchez-granero, M.Á. An Alternative Approach to Measure Co-Movement between Two Time Series. *Mathematics* **2020**, *8*, 261. [CrossRef]
18. Naimy, V.Y.; Hayek, M.R. Modelling and Predicting the Bitcoin Volatility Using GARCH Models. *Int. J. Math. Model. Numer. Optim.* **2018**, *8*, 197–215. [CrossRef]
19. Embrechts, P.; Resnick, S.I.; Samorodnitsky, G. Extreme Value Theory as a Risk Management Tool. *N. Am. Actuar. J.* **1999**, *3*, 30–41. [CrossRef]
20. Jorion, P. *Value at Risk: The New Benchmark for Managing Financial Risk*, 3rd ed.; McGraw-Hill: New York, NY, USA, 2007. [CrossRef]
21. Engle, R.F. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* **1982**, *50*, 987–1007. [CrossRef]
22. Bollerslev, T. Generalized Autoregressive Conditional Heteroskedasticity. *J. Econom.* **1986**, *31*, 307–327. [CrossRef]
23. Nelson, D. Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica* **1991**, *59*, 347–370. [CrossRef]
24. McNeil, A. Extreme Value Theory for Risk Managers. In *Internal Modelling and CAD II*; Risk Waters Books: London, UK, 1999; pp. 93–113.
25. Pickands, J. Statistical Inference Using Extreme Order Statistics. *Ann. Stat.* **1975**, *3*, 119–131.
26. Balkema, A.A.; de Haan, L. Residual Life Time at Great Age. *Ann. Probab.* **1974**, *2*, 792–804. [CrossRef]
27. Lee, W. *Applying Generalized Pareto Distribution to the Risk Management of Commerce Fire Insurance*; Working Paper; Tamkang University: New Taipei, Taiwan, 2009.
28. Russian Benchmark Officially Renamed the MOEX Russia Index. Available online: https://www.moex.com/n17810 (accessed on 3 April 2020).
29. Peng, Z.X.; Li, S.; Pang, H. *Comparison of Extreme Value Theory and GARCH Models on Estimating and Predicting of Value-at-Risk*; Working Paper; Wang Yanan Institute for Studies in Economics, Xiamen University: Xiamen, China, 2006.
30. Beck, T.; Demirguc-Kunt, A.; Levine, R.E.; Cihak, M.; Feyen, E. *Financial Development and Structure Dataset*. (updated September 2014). Available online: https://www.worldbank.org/en/publication/gfdr/data/financial-structure-database (accessed on 2 April 2020).
31. Market Capitalziation: % of GDP. Available online: https://www.ceicdata.com/en/indicator/market-capitalization--nominal-gdp (accessed on 2 April 2020).
32. Beck, T.; Demirguc-Kunt, A.; Levine, R.E.; Cihak, M.; Feyen, E. *Financial Development and Structure Dataset*. (updated September 2019). Available online: https://www.worldbank.org/en/publication/gfdr/data/financial-structure-database (accessed on 2 April 2020).

mathematics

MDPI

Article

An Extension of the Concept of Derivative: Its Application to Intertemporal Choice

Salvador Cruz Rambaud [1,*] and Blas Torrecillas Jover [2]

[1] Departamento de Economía y Empresa, Universidad de Almería, La Cañada de San Urbano, s/n, 04120 Almería, Spain

[2] Departamento de Matemáticas, Universidad de Almería, La Cañada de San Urbano, s/n, 04120 Almería, Spain; btorreci@ual.es

* Correspondence: scruz@ual.es; Tel.: +34-950-015-184

Received: 26 March 2020; Accepted: 26 April 2020; Published: 2 May 2020

Abstract: The framework of this paper is the concept of derivative from the point of view of abstract algebra and differential calculus. The objective of this paper is to introduce a novel concept of derivative which arises in certain economic problems, specifically in intertemporal choice when trying to characterize moderately and strongly decreasing impatience. To do this, we have employed the usual tools and magnitudes of financial mathematics with an algebraic nomenclature. The main contribution of this paper is twofold. On the one hand, we have proposed a novel framework and a different approach to the concept of relative derivation which satisfies the so-called generalized Leibniz's rule. On the other hand, in spite of the fact that this peculiar approach can be applied to other disciplines, we have presented the mathematical characterization of the two main types of decreasing impatience in the ambit of behavioral finance, based on a previous characterization involving the proportional increasing of the variable "time". Finally, this paper points out other patterns of variation which could be applied in economics and other scientific disciplines.

Keywords: derivation; intertemporal choice; decreasing impatience; elasticity

MSC: 16W25

JEL Classification: G41

1. Introduction and Preliminaries

In most social and experimental sciences, such as economics, psychology, sociology, biology, chemistry, physics, epidemiology, etc., researchers are interested in finding, *ceteris paribus*, the relationship between the explained variable and one or more explaining variables. This relationship has not to be linear, that is to say, linear increments in the value of an independent variable does not necessarily lead to linear variations of the dependent variable. This is logical by taking into account the non-linearity of most physical or chemical laws. These circumstances motivate the necessity of introducing a new concept of derivative which, of course, generalizes the concepts of classical and directional derivatives. Consequently, the frequent search for new patterns of variations in the aforementioned disciplines justifies a new framework and a different approach to the concept of derivative, able to help in modelling the decision-making process. Let us start with some general concepts.

Let A be an arbitrary K-algebra (non necessarily commutative or associative), where K is a field. A *derivation* over A is a K-linear map $D : A \to A$ satisfying Leibniz's identity:

$$D(ab) = D(a)b + aD(b).$$

It can be easily demonstrated that the sum, difference, scalar product and composition of derivations are derivations. In general, the product is not a derivation, but the so-called *commutator*, defined as $[D_1, D_2] = D_1D_2 - D_2D_1$, is a derivation. We are going to denote by $\mathrm{Der}_K(A)$ the set of all derivations on A. This set has the structure of a K-module and, with the commutator operation, becomes a Lie Algebra. This algebraic notion includes the classical partial derivations of real functions of several variables and the Lie derivative with respect to a vector field in differential geometry. Derivations and differentials are important tools in algebraic geometry and commutative algebra (see [1–3]).

The notion of derivation was extended to the so-called (σ, τ)-derivation [4] for an associative $\mathbb{C}$-algebra A, where σ and τ are two different algebra endomorphisms of A ($\sigma, \tau \in \mathrm{End}_{\mathrm{Alg}}(A)$), as a $\mathbb{C}$-linear map $D : A \to A$ satisfying:

$$D(ab) = D(a)\tau(b) + \sigma(a)D(b).$$

If $\tau = \mathrm{id}_A$, we obtain $D(ab) = D(a)b + \sigma(a)D(b)$. In this case, we will say that D is a σ-derivation. There are many interesting examples of these generalized derivations. The q-derivation, as a q-differential operator, was introduced by Jackson [5]. In effect, let A be a $\mathbb{C}$-algebra (which could be $\mathbb{C}[z, z^{-1}]$ or various functions spaces). The two important generalizations of derivation are D_q and $M_q : A \to A$, defined as:

$$(D_q(f))(z) = \frac{f(qz) - f(z)}{qz - z} = \frac{f(qz) - f(z)}{(q-1)z}$$

and

$$M_q(f)(z) = \frac{f(qz) - f(z)}{q - 1}.$$

These operations satisfy the q-deformed Leibniz's rule, $D(fg) = D(f)g + \sigma_q(f)g$, where $\sigma_q(f)(z) = f(qz)$, i.e., the q-derivation is a σ-derivation. Observe that this formula is not symmetric as the usual one.

Now, we can compare this q-derivation with the classical h-derivation, defined by:

$$(D_h(f))(z) = \frac{f(z+h) - f(z)}{h}.$$

In effect,

$$\lim_{q \to 1} D_q(f(z)) = \lim_{h \to 0} D_h(f(z)) = \frac{\mathrm{d}f(x)}{\mathrm{d}x},$$

provided that f is differentiable.

The q-derivation is the key notion of the quantum calculus which allows us to study those functions which are not differentiable. This theory has been developed by many authors and has found several applications in quantum groups, orthogonal polynomials, basic hypergeometric functions, combinatorics, calculus of variations and arithmetics [6,7].

In general, the q-derivation is more difficult to be computed. For instance, there is not a chain formula for this kind of derivation. We refer the interested reader to the books by Kac and Cheung [8], Ernst [9], and Annaby and Mansour [10] for more information about the q-derivations and q-integrals.

The q-derivation has been generalized in many directions within the existing literature. Recently, the β-derivative was introduced by Auch in his thesis [11], where $\beta(z) = az + b$, with $a \geq 1$, $b \geq 0$ and $a + b > 1$, the general case being considered in [12] and continued by several scholars [13]:

$$D_\beta(f)(z) = \frac{f(\beta(z)) - f(z)}{\beta(z) - z},$$

where $\beta \neq z$ and $\beta : I \to I$ is a strictly increasing continuous function, $I \subset \mathbb{R}$. Thus, a-derivation is a

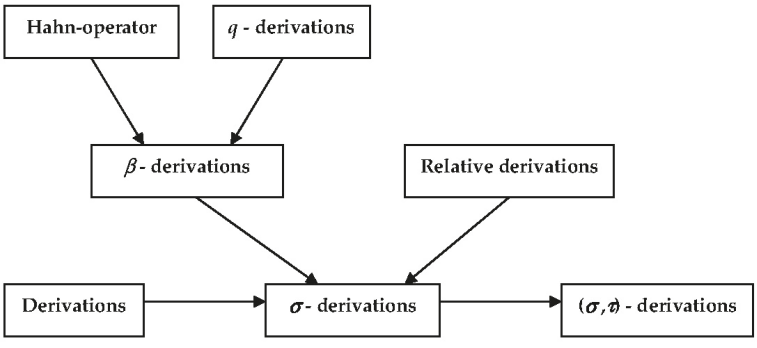


Figure 1. Chart of the different revised derivations.

This paper has been organized as follows. After this Introduction, Section 2 presents the novel concept of derivation relative to a given function f. It is shown that this derivative can be embodied in the ambit of relative derivations and satisfies Leibniz's rule. In Section 3, this new algebraic tool is used to characterize those discount functions exhibiting moderately and strongly decreasing impatience. In Section 4, the obtained results are discussed in the context of other variation patterns (quadratic, logarithmic, etc.) present in economics, finance and other scientific fields. Finally, Section 5 summarizes and concludes.

2. An Extension of the Concept of Derivative

2.1. General Concepts

Let A be a K-algebra and M be an A-module, where K is a field. Let $\sigma : A \to A$ an endomorphism of A. A *σ-derivation* D on M is an K-linear map

$$D : A \longrightarrow M,$$

such that

$$D(ab) = D(a)b + \sigma(a)D(b),$$

for every a and $b \in A$. From a structural point of view, let us denote by $\mathrm{Der}_K^\sigma(A, M)$ the set of all K-derivations on M. Obviously, $\mathrm{Der}_K^\sigma(A, M)$ is an A-module. In effect, if $D, D_1, D_2 \in \mathrm{Der}_K^\sigma(A, M)$ and $a \in A$, then $D_1 + D_2 \in \mathrm{Der}_K^\sigma(A, M)$ and $aD \in \mathrm{Der}_K^\sigma(A, M)$.

In the particular case where $M = A$, D will be called a *K-derivation on A*, and

$$\mathrm{Der}_K^\sigma(A, A) := \mathrm{Der}_K^\sigma(A)$$

will be called the *module of σ-derivations on A*.

2.2. Relative Derivation

Let us consider the module of K-derivations on the algebra A, $\mathrm{Der}_K^{\sigma}(A)$. For every D_0 and $D \in \mathrm{Der}_K^{\sigma}(A)$ and $a \in A$, we can define the *derivation relative to* D_0 *and* a as

$$D_a(\cdot) := D_0(a)D(\cdot).$$

Lemma 1. *If A is commutative, then D_a is a σ-derivation.*

Proof. In effect, clearly D is K-linear and, moreover, satisfies the generalized Leibniz's condition:

$$\begin{aligned} D_a(xy) &= D_0(a)D(xy) \\ &= D_0(a)[D(x)y + \sigma(x)D(y)] \\ &= D_0(a)D(x)y + \sigma(x)D_0(a)D(y) \\ &= D_a(x)y + \sigma(x)D_a(y). \end{aligned}$$

This completes the proof. □

Given two σ-derivations, D_0 and $D \in Der_K^{\sigma}(A)$, we can define a map

$$\mathcal{D} : A \to Der_K^{\sigma}(A)$$

such that

$$\mathcal{D}(a) = D_a := D_0(a)D.$$

Example 1. *Consider the polynomial ring $A := K[x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n]$, $D_0 = \partial_{x_1} + \cdots + \partial_{x_n}$ and $D = \partial_{y_1} + \cdots + \partial_{y_n}$. Then, for $a = x_1y_1 + \cdots + x_ny_n$ and every $f \in \mathcal{D}$, one has:*

$$\begin{aligned} D_a(f) &= D_0(a)D(f) \\ &= D_0(x_1y_1 + \cdots + x_ny_n)D(f) \\ &= (y_1 + \cdots + y_n)[\partial_{y_1}(f) + \cdots + \partial_{y_n}(f))]. \end{aligned}$$

Proposition 1. *For a commutative ring A, $\mathcal{D}$ is a σ-derivation.*

Proof. Firstly, let us see that $\mathcal{D}$ is K-linear. In effect, for every $a, b \in A$ and $k \in K$, one has:

$$\begin{aligned} \mathcal{D}(a+b) &= D_{a+b} = D_0(a+b)D \\ &= D_0(a)D + D_0(b)D = D_a + D_b = \mathcal{D}(a) + \mathcal{D}(b) \end{aligned}$$

and

$$\begin{aligned} \mathcal{D}(ka) &= D_{ka} = D_0(ka)D \\ &= kD_0(a)D = kD_a = k\mathcal{D}(a). \end{aligned}$$

Secondly, we are going to show that $\mathcal{D}$ satisfies the generalized Leibniz condition. In effect, for every $a, b \in A$, one has:

$$\begin{aligned} \mathcal{D}(ab) &= D_{ab} = D_0(ab)D = [D_0(a)b + \sigma(a)D_0(b)]D \\ &= [D_0(a)D]b + \sigma(a)[D_0(b)D] = D_ab + \sigma(a)D_b = \mathcal{D}_ab + \sigma(a)\mathcal{D}(b). \end{aligned}$$

Therefore, $\mathcal{D}$ is a σ-derivation. □

Now, we can compute the bracket of two relative σ-derivations D_a and D_b, for every $a, b \in A$:

$$\begin{aligned}
[D_a, D_b] &= D_a \circ D_b - D_b \circ D_a \\
&= D_a[D_0(b)D] - D_b[D_0(a)D] \\
&= D_0(a)D(D_0(b)D) - D_0(b)D(D_0(a)D) \\
&= D_0(a)[DD_0(b)D + \sigma(D_0(b))D^2] - D_0(b)[DD_0(a)D + \sigma(D_0(a))D^2] \\
&= [D_0(a)DD_0(b) - D_0(b)DD_0(a)]D + [D_0(a)\sigma(D_0(b)) - D_0(b)\sigma(D_0(a))]D^2.
\end{aligned}$$

Observe that, although the bracket of two derivations is a derivation, in general it is not a relative derivation. However, if A is commutative and σ is the identity, then the former bracket could be simplified as follows:

$$\begin{aligned}
[D_a, D_b] &= [D_0(a)DD_0(b) - D_0(b)DD_0(a)]D + [D_0(a), D_0(b)]D^2 \\
&= [D_0(a)DD_0(b) - D_0(b)DD_0(a)]D.
\end{aligned}$$

Moreover, if A is commutative and σ is the identity, the double bracket of three derivations D_a, D_b and D_c, for every $a, b, c \in A$, is:

$$\begin{aligned}
[[D_a, D_b], D_c] &= [D_0(a)DD_0(b) - D_0(b)DD_0(a)]DD_c - D_c[D_0(a)DD_0(b) - D_0(b)DD_0(a)]D \\
&= [D_0(a)DD_0(b) - D_0(b)DD_0(a)]DD_0(c)D - D_0(c)D[D_0(a)DD_0(b) - D_0(b)DD_0(a)]D \\
&= [D_0(a)DD_0(b) - D_0(b)DD_0(a)]DD_0(c)D + [D_0(a)DD_0(b) - D_0(b)DD_0(a))]D_0(c)D^2 \\
&= D_0(c)\{D[D_0(a)DD_0(b) - D_0(b)DD_0(a)]\}D - D_0(c)\{D[D_0(a)DD_0(b) - D_0(b)DD_0(a)]D.
\end{aligned}$$

Since the relative derivations are true derivations, they satisfy Jacobi's identity, i.e., for every $a, b, c \in A$, the following identity holds:

$$[D_a, [D_b, D_c]] + [D_b, [D_c, D_a]] + [D_b, [D_a, D_b]] = 0.$$

For σ derivation one could modify the definition of the bracket as in [17] and then a Jacobi-like identity is obtained [17, Theorem 5]. We left the details to the reader.

Given another derivation D_1, we can define a new relative σ-derivation $D'_a := D_1(a)D$. In this case, the new bracket is:

$$\begin{aligned}
[D_a, D'_a] &= D_a D'_a - D'_a D_a \\
&= D_0(a)D(D_1(a)D) - D_1(a)D(D_0(a)D) \\
&= D_0(a)[DD_1(a)D - \sigma(D_1(a))D^2] - D_1(a)[DD_0(a)D - \sigma(D_0(a))D^2] \\
&= [D_0(a)DD_1(a) - D_1(a)DD_0(a)]D + [D_1(a)\sigma(D_0(a)) - D_0(a)\sigma(D_1(a))]D^2.
\end{aligned}$$

If A is commutative and σ is the identity, then:

$$[D_a, D'_a] = [D_0(a)DD_1(a) - D_1(a)DD_0(a)]D.$$

The chain rule is also satisfied for relative derivations when A is commutative. In effect, assume that, for every $f, g \in A$, the composition, $f \circ g$, is defined. Then

$$\begin{aligned}
D_a(f \circ g) &= D_0(f \circ g)D(f \circ g) \\
&= [D_0(f) \circ g]D_0(g)[D(f) \circ g]D(g) \\
&= [D_0(f) \circ g][D(f) \circ g]D_0(g)D(g) \\
&= [D_0(f)D(f) \circ g]D_0(g)D(g) \\
&= [D_a(f) \circ g]D_a(g).
\end{aligned}$$

2.3. Derivation Relative to a Function

Let $f(x, \Delta)$ be a real function of two variables x and Δ such that, for every a:

$$\lim_{\Delta \to 0} f(a, \Delta) = a. \tag{1}$$

Let $F(x)$ be a real function differentiable at $x = a$. The *derivative of F relative to f, at* $x = a$, denoted by $D_f(F)(a)$, is defined as the following limit:

$$D_f(F)(a) := \lim_{\Delta \to 0} \frac{F[f(a, \Delta)] - F(a)}{\Delta}. \tag{2}$$

In this setting, we can define $g(x, \Delta)$ as the new function, also of two variables x and Δ, satisfying the following identity:

$$f(x, \Delta) := x + g(x, \Delta). \tag{3}$$

Observe that $g(x, \Delta) = f(x, \Delta) - x$ and, consequently,

$$\lim_{\Delta \to 0} g(a, \Delta) = 0.$$

The set of functions $g(x, \Delta)$, denoted by $\mathcal{S}$, is a subalgebra of the algebra of real-valued functions of two variables x and Δ, represented by $\mathcal{A}$. In effect, it is obvious to check that, if $g, g_1, g_2 \in \mathcal{S}$ and $\lambda \in \mathbb{R}$, then $g_1 + g_2$, $g_1 \cdot g_2$, and λg belong to $\mathcal{S}$. Therefore, if $\mathcal{F}$ denotes the set of the so-defined functions $f(x, \Delta)$, we can write:

$$\mathcal{F} = \mathrm{id} + \mathcal{S}, \tag{4}$$

where $\mathrm{id}(x) = x$ is the identity function of one variable. In $\mathcal{S}$, we can define the following binary relation:

$$g_1 \sim g_2 \text{ if, and only if, } \lim_{\Delta \to 0} \frac{g_1(x, \Delta)}{\Delta} = \lim_{\Delta \to 0} \frac{g_2(x, \Delta)}{\Delta}. \tag{5}$$

Obviously, $\sim$ is an equivalence relation. Now, we can define $h(x, \Delta)$ as the new function also of two variables x and Δ, satisfying the following identity:

$$g(x, \Delta) := h(x, \Delta)\Delta. \tag{6}$$

Thus,

$$h(x, \Delta) = \frac{f(x, \Delta) - x}{\Delta}. \tag{7}$$

Therefore,

$$D_f(F)(a) = \lim_{\Delta \to 0} \frac{F[a + h(a, \Delta)\Delta] - F(a)}{h(a, \Delta)\Delta} \lim_{\Delta \to 0} h(a, \Delta). \tag{8}$$

Observe that now $\lim_{\Delta \to 0} h(a, \Delta)$ only depends on a, whereby it can be simply denoted as $h(a)$:

$$h(a) := \lim_{\Delta \to 0} h(a, \Delta). \tag{9}$$

Thus, Equation (8) results in:

$$D_f(F)(a) = D(F)(a) \cdot h(a). \tag{10}$$

If $f(x, \Delta)$ is derivable at $\Delta = 0$, then $f(x, \Delta)$ is continuous at $\Delta = 0$, whereby

$$f(a, 0) = \lim_{\Delta \to 0} f(a, \Delta) = a \tag{11}$$

and, consequently,

$$h(a) = \lim_{\Delta \to 0} h(a, \Delta) = \lim_{\Delta \to 0} \frac{f(a, \Delta) - f(a, 0)}{\Delta} = \left. \frac{\partial f(a, \Delta)}{\partial \Delta} \right|_{\Delta = 0}. \tag{12}$$

Therefore,

$$D_f(F)(a) = \left. \frac{\partial f(a, \Delta)}{\partial \Delta} \right|_{\Delta = 0} D(F)(a). \tag{13}$$

Observe that we are representing by $D : \mathcal{C}^1(\mathbb{R}) \to \mathcal{C}^1(\mathbb{R})$ the operator $D = \frac{\mathrm{d}}{\mathrm{d}x}$. If, additionally, the partial derivative $\left. \frac{\partial f(a,\Delta)}{\partial \Delta} \right|_{\Delta=0}$ is simply denoted by $\partial_{y=0}(f)(a)$, expression (13) remains as:

$$D_f(F)(a) = \partial_{y=0}(f)(a) D(F)(a) = \partial_{y=0}(g)(a) D(F)(a)$$

or, globally,

$$D_f(F) = \partial_{y=0}(f) D(F) = \partial_{y=0}(g) D(F).$$

Thus, $D_f : \mathcal{C}^1(\mathbb{R}) \to \mathcal{C}^1(\mathbb{R})$ is really a derivation:

$$\begin{array}{rcl} D_f(FG) & = & \partial_{y=0}(f) D(FG) \\ & = & \partial_{y=0}(f)[D(F)G + FD(G)] \\ & = & [\partial_{y=0}(f) D(F)]G + F[\partial_{y=0}(f) D(G)] \\ & = & D_f(F)G + FD_f(G). \end{array}$$

Observe that $h(a)$ or $\partial_{y=0}(f)(a)$ represents the equivalence class including the function $g(a, \Delta)$. Moreover, the set of all suitable values of $D_f(F)(a)$ is restricted to the set

$$D(F)(a)(\mathcal{S}/\sim),$$

where $\mathcal{S}/\sim$ is the quotient set derived from the equivalence relation $\sim$.

The name assigned to this derivative can be justified as follows. Observe that the graphic representation of $g(a, \Delta)$ is a surface which describes a kind of "valley" over the a-axis (that is to say, $\Delta = 0$) (see Figure 2). Therefore, for every value of a, a path can be obtained by intersecting the surface with the vertical plane crossing the point $(a, 0)$, giving rise to the function $g(a, \Delta)$, which represents the increment of a. As previously indicated, $g(a, \Delta)$ tend to zero as Δ approaches to zero (represented by the red arrow).

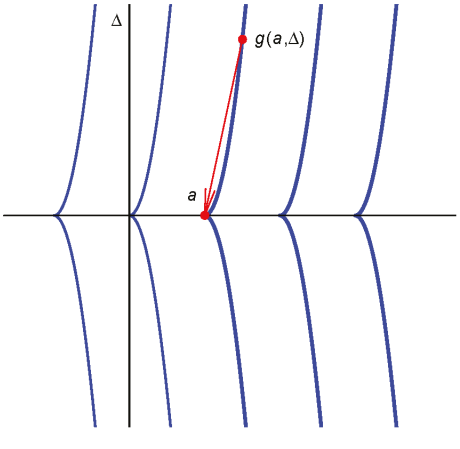


Figure 2. Plotting function $g(a, \Delta)$.

In the particular case in which

$$f(x,\Delta) = x + \Delta, \tag{14}$$

obviously, one has:

$$D_f(F)(a) = D(F)(a), \tag{15}$$

that is to say, the derivative relative to the function $f(x,\Delta) = x + \Delta$ (absolute increments) coincides with the usual derivative.

Example 2. *Assume that (percentage increments of the variable):*

$$f(x,\Delta) = x + \frac{\Delta}{x}.$$

In this case,

$$D_f(F)(a) = \frac{1}{a} D(F)(a).$$

In the context of certain scientific problems, it is interesting to characterize the variation (increase or decrease) of the so-defined derivative relative to a given function. In this case, the sign of the usual derivative of this relative derivative will be useful:

$$D[D_f(F)] = D[\partial_{y=0}(f) D(F)] = D[\partial_{y=0}(f)] D(F) + \partial_{y=0}(f) D^2(F). \tag{16}$$

Thus, if $D_f(F)$ must be increasing (resp. decreasing), then

$$D^2(F) > -\frac{D[\partial_{y=0}(f)] D(F)}{\partial_{y=0}(f)} \tag{17}$$

(resp. $D^2(F) < -\frac{D[\partial_{y=0}(f)] D(F)}{\partial_{y=0}(f)}$).

Example 3. *Assume that* $f(x,\Delta) = \ln(\exp\{a\} + \Delta)$. *In this case,*

$$D_f(F)(a) = \frac{1}{\exp\{a\}} D(F)(a).$$

The condition of increase of $D_f(F)$ *leads to*

$$D^2(F) > D(F).$$

3. An Application to Intertemporal Choice: Proportional Increments

This section is going to apply this new methodology to a well-known economic problem, more specifically to intertemporal choice. In effect, we are going to describe a noteworthy particular case of our derivative when the change in the variable is due to proportional instead to absolute increments. To do this, let us start with the description of the setting in which the new derivative will be applied (see [18,19]).

Let X be set $\mathbb{R}^+$ of non-negative real numbers and T a non-degenerate closed interval of $[0, +\infty)$. A *dated reward* is a couple $(x,t) \in X \times T$. In what follows, we will refer to x as the *amount* and t as the *time of availability* of the reward. Assume that a decision maker exhibits a continuous weak order on $X \times T$, denoted by $\preceq$, satisfying the following conditions (the relations $\prec$, $\succeq$, $\succ$ and $\sim$ can be defined as usual):

1. For every $s \in T$ and $t \in T$, $(0,s) \sim (0,t)$ holds.
2. Monotonicity: For every $t \in T$, $x \in X$ and $y \in X$, such that $x < y$, then $(x,t) \prec (y,t)$.
3. Impatience: For every $s \in T$ and $t \in T$, such that $s < t$, and $x \in X$ then $(x,s) \succ (x,t)$.

The most famous representation theorem of preferences is due to Fishburn and Rubinstein [20]: If order, monotonicity, continuity, impatience, and separability hold, and the set of rewards X is an interval, then there are continuous real-valued functions u on X and F on the time interval T such that

$$(x,s) \preceq (y,t) \text{ if, and only if, } u(x)F(s) \leq u(y)F(t).$$

Additionally, function u, called the *utility*, is increasing and satisfies $u(0) = 0$. On the other hand, function F, called the *discount function*, is decreasing, positive and satisfies $F(0) = 1$.

Assume that, for the decision maker, the rewards (x,s) and (y,t), with $s < t$, are indifferent, that is to say, $(x,s) \sim (y,t)$. Observe that, necessarily, $u(x) < u(y)$. The *impatience* in the interval $[s,t]$, denoted by $I(s,t)$, can be defined as the difference $u(y) - u(x)$ which is the amount that the agent is willing to loss in exchange for an earlier receipt of the reward. However, in economics the magnitudes should be defined in relative, better than absolute, terms. Thus, the impatience corresponding to the interval $[s,t]$, relatively to time and amount, should be:

$$I(s,t) := \frac{u(y) - u(x)}{(t-s)u(y)}.$$

According to [20], the following equation holds:

$$u(x)F(s) = u(y)F(t),$$

whereby

$$I(s,t) = \frac{F(s) - F(t)}{(t-s)F(s)}.$$

Observe that, in algebraic terms, $I(s,t)$ is the classical "logarithmic" derivative, with minus sign, of F at time s:

$$I(s,t) = -\left(\frac{D_{t-s}(F)}{F}\right)(s),$$

where D_{t-s} is the classical h-derivation, with $h = t - s$. However, in finance, the most employed measure of impatience is given by the limit of $I(s,t)$ when t tends to s, giving rise to the well-known concept of *instantaneous discount rate*, denoted by $\delta(s)$:

$$\delta(s) := -\lim_{t \to s}\left(\frac{D_{t-s}(F)}{F}\right)(s) = -D(\ln F)(s).$$

For a detailed information about the different concepts of impatience in intertemporal choice, see [21]. The following definition introduces a central concept to analyze the evolution of impatience with the passage of time.

Definition 1 ([19]). *A decision-maker exhibiting preferences $\preceq$ has decreasing impatience (DI) if, for every $s < t$, $k > 0$ and $0 < x < y$, $(x,s) \sim (y,t)$ implies $(x, s+k) \preceq (y, t+k)$.*

A consequence is that, under the conditions of Definition 1, given $\sigma > 0$, there exists $\tau = \tau(\sigma) > \sigma$ such that

$$(x, s+\sigma) \sim (y, t+\tau).$$

The existence of τ is guaranteed when, as usual, the discount function is regular, i.e., satisfies $\lim_{t\to\infty} F(t) = 0$. A specific case of DI is given by the following definition.

Definition 2 ([19]). *A decision-maker exhibiting decreasing impatience has strongly decreasing impatience if $s\tau \geq t\sigma$.*

The following proposition provides a nice characterization of strongly decreasing impatience.

Proposition 2 ([22]). *A decision-maker exhibiting preferences $\preceq$ has strongly decreasing impatience if, and only if, for every $s < t$, $\lambda > 1$ and $0 < x < y$, $(x, s) \sim (y, t)$ implies $(x, \lambda s) \prec (y, \lambda t)$.*

Definition 3. *Let $F(t)$ be a discount function differentiable in its domain. The elasticity of $F(t)$ is defined as:*

$$\epsilon_F(t) := t\frac{D(F)(t)}{F(t)} = tD(\ln F)(t) = -t\delta(t).$$

Theorem 1. *A decision-maker exhibiting preferences $\preceq$ has strongly decreasing impatience if, and only if, $D^2(F) > -\frac{D(F)}{id}$.*

Proof. In effect, for every $s < t$, $\lambda > 1$ and $0 < x < y$, by Proposition 1, $(x, s) \sim (y, t)$ implies $(x, \lambda s) \prec (y, \lambda t)$. Consequently,

$$u(x)F(s) = u(y)F(t)$$

and

$$u(x)F(\lambda s) < u(y)F(\lambda t).$$

By dividing the left-hand sides and the right-hand sides of the former inequality and equality, one has:

$$\frac{F(\lambda s)}{F(s)} < \frac{F(\lambda t)}{F(t)},$$

from where:

$$\ln F(\lambda s) - \ln F(s) < \ln F(\lambda t) - \ln F(t).$$

As $\lambda > 1$, we can write $\lambda := 1 + \Delta$, with $\Delta > 0$, and so:

$$\ln F((1 + \Delta)s) - \ln F(s) < \ln F((1 + \Delta)t) - \ln F(t).$$

By dividing both member of the former inequality by Δ and letting $\Delta \to 0$, one has:

$$D_f(F)(s) \leq D_f(F)(t),$$

where $f(x, \Delta) := (1 + \Delta)x$. Therefore, the function $D_f(F)$ is increasing, whereby:

$$D[D_f(F)] \geq 0.$$

In order to calculate $D_f(F)$, take into account that now there is a proportional increment of the variable, that is to say:

$$f(x, \Delta) = (1 + \Delta)x.$$

Thus,

$$D_f(F)(a) = aD(F)(a)$$

or, globally,

$$D_f(F) = \mathrm{id}D(F).$$

Consequently, $\mathrm{id}D(F)$ is increasing, whereby:

$$D[ID(F)] = D(F) + \mathrm{id}D^2(F) > 0,$$

from where:

$$D^2(F) > -\frac{D(F)}{\mathrm{id}}.$$

The proof of the converse implication is obvious. □

Example 4. *The discount function* $F(t) = \exp\{-\arctan(t)\}$ *exhibits strongly decreasing impatience. In effect, simple calculation shows that:*

- $D(F)(a) = -\frac{1}{1+a^2}\exp\{-\arctan(t)\}$.
- $D^2(F)(a) = \frac{1+2a}{1+a^2}\exp\{-\arctan(t)\}$.

In this case, the inequality $D^2(F) > -\frac{D(F)}{id}$ *results in* $a^2 + a - 1 > 0$ *which holds for* $a > \frac{-1+\sqrt{5}}{2}$.

The following result can be derived from Theorem 1 [22].

Corollary 1. *A decision-maker exhibiting preferences* $\preceq$ *has strongly decreasing impatience if, and only if,* ϵ_F *is decreasing.*

Proof. It is immediate, taking into account that $\mathrm{id}D(F)$ is increasing. As F is decreasing, then $\mathrm{id}\frac{D(F)}{F}$ is increasing and

$$\epsilon = -\mathrm{id}\delta$$

is decreasing. The proof of the converse implication is obvious. □

Another specific case of DI is given by the following definition.

Definition 4 ([19])**.** *A decision-maker exhibiting decreasing impatience has moderately decreasing impatience if* $s\tau < t\sigma$.

The following corollary provides a characterization of moderately decreasing impatience.

Corollary 2 ([22])**.** *A decision-maker exhibiting preferences* $\preceq$ *has moderately decreasing impatience if, and only if, for every* $s < t$, $k > 0$, $\lambda > 1$ *and* $0 < x < y$, $(x, s) \sim (y, t)$ *implies* $(x, s + k) \preceq (y, t + k)$ *but* $(x, \lambda s) \succeq (y, \lambda t)$.

Corollary 3. *A decision-maker exhibiting preferences* $\preceq$ *has moderately decreasing impatience if, and only if,* $\frac{[D(F)]^2}{F} < D^2(F) \leq -\frac{D(F)}{id}$.

Proof. It is an immediate consequence of Theorem 1 and of the fact that, in this case, $\delta = -\frac{D(F)}{F}$ is decreasing. □

The following result can be derived from Corollary 3 [22].

Corollary 4. *A decision-maker exhibiting preferences* $\preceq$ *has moderately decreasing impatience f, and only if,* ϵ_F *is increasing but* δ *is decreasing.*

4. Discussion

In this paper, we have introduced a new modality of relative derivation, specifically the so-called derivation of $F(x)$ relative to a function $f(x, \Delta) := x + g(x, \Delta)$, where $g(x, \Delta)$ represents the increments in the variable x. Obviously, this novel concept generalizes the two most important derivatives used in differential calculus:

- The classical derivative, whose increments are defined as $g(x, \Delta) = \Delta$.
- The directional derivative, characterized by linear increments: $g(x, \Delta) = k\Delta$, $k \in \mathbb{R}$.

It is easy to show that, in the former cases, Equation (13) leads to the well-known expressions of these two derivatives. In this paper, we have gone a step further and have considered proportional

variations of the independent variable. These increments appear in the so-called *sensitivity analysis* which is a financial methodology which determines how changes of a variable can affect variations of another variable. This method, also called simulation analysis, is usually employed in financial problems under uncertainty contexts and also in econometric regressions.

In effect, in some economic contexts, percentage variations of the independent variable are analyzed. For example, the *elasticity* is the ratio of the percentage variations of two economic magnitudes. In linear regression, if the explanatory and the explained variables are affected by the natural logarithm, it is noteworthy to analyze the percentage variation of the dependent variable compared to percentage changes in the value of an independent variable. In this case, we would be interested in analyzing the ratio:

$$\frac{f(x+\Delta x)-f(x)}{f(x)},$$

when $\frac{\Delta x}{x}=\lambda$, for a given λ. Thus, the former ratio remains as:

$$\frac{f(x+\lambda x)-f(x)}{f(x)}.$$

In another economic context, Karpoff [23], when searching the relationship between the price and the volume of transactions of an asset in a stock market, suggests quadratic and logarithmic increments:

- Quadratic increments aim to determine the variation of the volume when quadratic changes in the price of an asset have been considered. In this case,

  $$g(x,\Delta)=(x+\Delta)^2-x^2$$

  and so

  $$\partial_{y=0}(f)(a)=2a.$$

- Logarithmic increments aim to find the variation of the volume when considering quadratic changes in the price of an asset. In this case

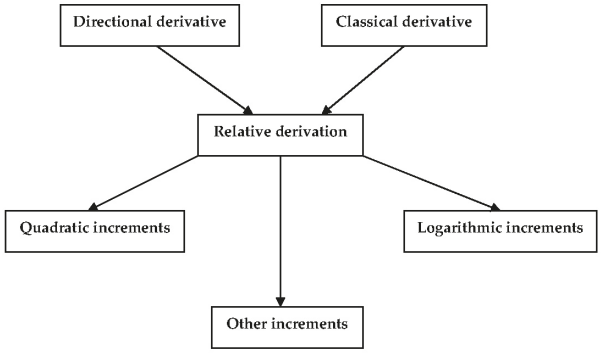


Figure 3. Chart of the different types of increments.

Indeed, some other variation models could be mentioned here. Take into account that some disciplines, such as biology, physics or economics, might be interested in explaining the increments in the dependent variable by using alternative patterns of variation. For example, think about a particle which is moving by following a given trajectory. In this context, researchers may be interested in knowing the behavior of the explained variable when the particle is continuously changing its position according to a given function.

5. Conclusions

This paper has introduced the novel concept of derivative of a function relative to another given function. The manuscript has been divided into two parts. The first part is devoted to the algebraic treatment of this concept and its basic properties in the framework of other relative derivatives. Moreover, this new derivative has been put in relation with the main variants of derivation in the field of abstract algebra. Given two σ-derivations over a K-algebra A, where K is a field, a relative σ-derivation has been associated to any function. This construction is, in fact, a derivation from A to the A-module of σ-derivations. Specifically, if σ is the identity of the algebra, these derivations can be applied to the theory of intertemporal choice.

The second part deals with the mathematical characterization of the so-called "strongly" and "moderately decreasing impatience" based on previous characterizations involving the proportional increasing of the variable "time". In effect, a specific situation, the case of proportional increments, plays a noteworthy role in economics, namely in intertemporal choice, where the analysis of decreasing impatience is a topic of fundamental relevance. In effect, the proportional increment of time is linked to the concept of strongly and moderately decreasing impatience. Therefore, the calculation of derivatives relatively to this class of increments will allow us to characterize these important modalities of decreasing impatience.

Moreover, after providing a geometric interpretation of this concept, this derivative has been calculated relatively to certain functions which represent different patterns of variability of the main variable involved in the problem.

Observe that, according to Fishburn and Rubinstein [20], the continuity of the order relation implies that functions F and u are continuous but not necessarily derivable. Indeed, this is a limitation of the approach presented in this paper which affects both function F and the variation pattern g. A further research could be to analyze the case of functions which are differentiable except at possibly a finite number of points in its domain.

Finally, apart from this financial application, another future research line is the characterization of other financial problems with specific models of variability. In this way, we can point out the proportional variability of reward amounts [24].

Author Contributions: Conceptualization, S.C.R. and B.T.J.; Formal analysis, S.C.R. and B.T.J.; Funding acquisition, S.C.R. and B.T.J.; Supervision, S.C.R. and B.T.J.; Writing – original draft, S.C.R. and B.T.J. All authors have read and agreed to the published version of the manuscript.

Funding: The authors gratefully acknowledge financial support from the Spanish Ministry of Economy and Competitiveness [National R&D Project "La sostenibilidad del Sistema Nacional de Salud: reformas, estrategias y propuestas", reference: DER2016-76053-R] and [National R&D Project "Anillos, módulos y álgebras de Hopf, reference: MTM2017-86987-P].

Acknowledgments: We are very grateful for the valuable comments and suggestions offered by three anonymous referees.

Conflicts of Interest: The authors declare no conflict of interest.

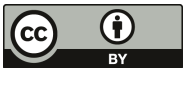
Abbreviations

The following abbreviations are used in this manuscript:

DI Decreasing Impatience

References

1. Kunz, E. *Kähler Differential*; Springer: Berlin, Germany, 1986.
2. Eisenbud, D. *Commutative Algebra with a View toward Algebraic Geometry*, 3rd ed.; Springer: Berlin, Germany, 1999.
3. Matsumura, H. *Commutative Algebra*; Mathematics Lecture Note Series; W. A. Benjamin: New York, NY, USA, 1970.
4. Jacobson, N. *Structure of Rings*; Am. Math. Soc. Coll. Pub. 37; Amer. Math. Soc: Providence, RI, USA, 1956.
5. Jackson, F.H. q-difference equations. *Amer. J. Math* **1910**, *32*, 305–314. [CrossRef]
6. Chakrabarti, R.; Jagannathan, R.;Vasudevan, R. A new look at the q-deformed calculus. *Mod. Phys. Lett. A* **1993**, *8*, 2695–2701. [CrossRef]
7. Haven, E. Itô's lemma with quantum calculus (q-calculus): Some implications. *Found. Phys.* **2011**, *41*, 529–537. [CrossRef]
8. Kac, V.; Cheung, P. *Quantum Calculus*; Springer: Berlin, Germany, 2002.
9. Ernst, T. *A Comprehensive Treatment of q-Calculus*; Birkhäuser: New York, NY, USA, 2012.
10. Annaby, M.; Mansour, Z.S. *q-Fractional Calculus and Equations*; Lecture Notes in Mathematics, 2056.; Springer: New York, NY, USA, 2012.
11. Auch, T. Development and Application of Difference and Fractional Calculus on Discrete Time Scales. Ph.D. Thesis, University of Nebraska-Lincoln, Lincoln, NE, USA, 2013.
12. Hamza, A.; Sarhan, A.; Shehata, E.; Aldwoah, K.A. General quantum difference calculus. *Adv. Differ. Equ.* **2015**, *2015*, 1–19. [CrossRef]
13. Faried, N.; Shehata, E.M.; El Zafarani, R.M. On homogeneous second order linear general quantum difference equations. *J. Inequalities Appl.* **2017**, *2017*, 198. [CrossRef] [PubMed]
14. Hahn, W. On orthogonal polynomials satisfying a q-difference equation. *Math. Nachr.* **1949**, *2*, 4–34. [CrossRef]
15. Hahn, W. Ein Beitrag zur Theorie der Orthogonalpolynome. *Monatshefte Math.* **1983**, *95*, 19–24. [CrossRef]
16. Gupta, V.; Rassias T.M.; Agrawal, P.N.; Acu, A.M. Basics of Post-Quantum Calculus. In *Recent Advances in Constructive Approximation Theory. Springer Optimization and Its Applications*; Springer: Berlin, Germany, 2018; Volume 138.
17. Hartwig, J.T.; Larsson, D.; Silvestrov, S.D. Deformations of Lie algebras using σ-derivations. *J. Algebra* **2006**, *295*, 314–361. [CrossRef]
18. Baucells, M.; Heukamp, F.H. Probability and time trade-off. *Manag. Sci.* **2012**, *58*, 831–842. [CrossRef]
19. Rohde, K.I.M. Measuring decreasing and increasing impatience. *Manag. Sci.* **2018**, *65*, 1455–1947. [CrossRef]
20. Fishburn, P.C.; Rubinstein, A. Time preference. *Int. Econ. Rev.* **1982**, *23*, 677–694. [CrossRef]
21. Cruz Rambaud, S.; Muñoz Torrecillas, M.J. Measuring impatience in intertemporal choice. *PLoS ONE* **2016**, *11*, e0149256. [CrossRef] [PubMed]
22. Cruz Rambaud, S.; González Fernández, I. A measure of inconsistencies in intertemporal choice. *PLoS ONE* **2019**, *14*, e0224242. [CrossRef] [PubMed]
23. Karpoff, J.M. The relation between price changes and trading volume: A survey. *J. Financ. Quant. Anal.* **1987**, *22*, 109–126. [CrossRef]
24. Anchugina, N.; Matthew, R.; Slinko, A. *Aggregating Time Preference with Decreasing Impatience*; Working Paper; University of Auckland: Auckland, Australia, 2016.



mathematics

MDPI

Article

The VIF and MSE in Raise Regression

Román Salmerón Gómez [1], Ainara Rodríguez Sánchez [2] and Catalina García García [1,*] and José García Pérez [3]

[1] Department of Quantitative Methods for Economics and Business, University of Granada, 18010 Granada, Spain; romansg@ugr.es
[2] Department of Economic Theory and History, University of Granada, 18010 Granada, Spain; arsanchez@ugr.es
[3] Department of Economy and Company, University of Almería, 04120 Almería, Spain; jgarcia@ual.es
* Correspondence: cbgarcia@ugr.es; Tel.: +34-958248790

Received: 1 April 2020; Accepted: 13 April 2020; Published: 16 April 2020

Abstract: The raise regression has been proposed as an alternative to ordinary least squares estimation when a model presents collinearity. In order to analyze whether the problem has been mitigated, it is necessary to develop measures to detect collinearity after the application of the raise regression. This paper extends the concept of the variance inflation factor to be applied in a raise regression. The relevance of this extension is that it can be applied to determine the raising factor which allows an optimal application of this technique. The mean square error is also calculated since the raise regression provides a biased estimator. The results are illustrated by two empirical examples where the application of the raise estimator is compared to the application of the ridge and Lasso estimators that are commonly applied to estimate models with multicollinearity as an alternative to ordinary least squares.

Keywords: detection; mean square error; multicollinearity; raise regression; variance inflation factor

1. Introduction

In the last fifty years, different methods have been developed to avoid the instability of estimates derived from collinearity (see, for example, Kiers and Smilde [1]). Some of these methods can be grouped within a general denomination known as penalized regression.

In general terms, the penalized regression parts from the linear model (with p variables and n observations), $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, and obtains the regularization of the estimated parameters, minimizing the following objective function:

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + P(\boldsymbol{\beta}),$$

where $P(\boldsymbol{\beta})$ is a penalty term that can take different forms. One of the most common penalty terms is the bridge penalty term ([2,3]) is given by

$$P(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j|^{\alpha},\ \alpha > 0,$$

where λ is a tuning parameter. Note that the ridge ([4]) and the Lasso ([5]) regressions are obtained when $\alpha = 2$ and $\alpha = 1$, respectively. Penalties have also been called soft thresholding ([6,7]).

These methods are applied not only for the treatment of multicollinearity but also for the selection of variables (see, for example, Dupuis and Victoria-Feser [8], Li and Yang [9] Liu et al. [10], or Uematsu and Tanaka [11]), which is a crucial issue in many areas of science when the number of variables

exceeds the sample size. Zou and Hastie [12] proposed elastic net regularization by using the penalty terms λ_1 and λ_2 that combine the Lasso and ridge regressions:

$$P(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2.$$

Thus, the Lasso regression usually selects one of the regressors from among all those that are highly correlated, while the elastic net regression selects several of them. In the words of Tutz and Ulbricht [13] "the elastic net catches all the big fish", meaning that it selects the whole group.

From a different point of view, other authors have also presented different techniques and methods well suited for dealing with the collinearity problems: continuum regression ([14]), least angle regression ([15]), generalized maximum entropy ([16–18]), the principal component analysis (PCA) regression ([19,20]), the principal correlation components estimator ([21]), penalized splines ([22]), partial least squares (PLS) regression ([23,24]), or the surrogate estimator focused on the solution of the normal equations presented by Jensen and Ramirez [25].

Focusing on collinearity, the ridge regression is one of the more commonly applied methodologies and it is estimated by the following expression:

$$\widehat{\boldsymbol{\beta}}(K) = \left(\mathbf{X}^t\mathbf{X} + K \cdot \mathbf{I}\right)^{-1} \mathbf{X}^t\mathbf{Y} \tag{1}$$

where $\mathbf{I}$ is the identity matrix with adequate dimensions and K is the ridge factor (ordinary least squares (OLS) estimators are obtained when $K = 0$). Although ridge regression has been widely applied, it presents some problems with current practice in the presence of multicollinearity and the estimators derived from the penalty come into these same problems whenever $n > p$:

- In relation to the calculation of the variance inflation factors (VIF), measures that quantify the degree of multicollinearity existing in a model from the coefficient of determination of the regression between the independent variables (for more details, see Section 2), García et al. [26] showed that the application of the original data when working with the ridge estimate leads to non-monotone VIF values by considering the VIF as a function of the penalty term. Logically, the Lasso and the elastic net regression inherit this property.
- By following Marquardt [27]: "The least squares objective function is mathematically independent of the scaling of the predictor variables (while the objective function in ridge regression is mathematically dependent on the scaling of the predictor variables)". That is to say, the penalized objective function will bring problems derived from the standardization of the variables. This fact has to be taken into account both for obtaining the estimators of the regressors and for the application of measures that detect if the collinearity has been mitigated. Other penalized regressions (such as Lasso and elastic net regressions) are not scale invariant and hence yield different results depending on the predictor scaling used.
- Some of the properties of the OLS estimator that are deduced from the normal equations are not verified by the ridge estimator and, among others, the estimated values for the endogenous variable are not orthogonal to the residuals. As a result, the following decomposition is verified

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i(K) - \bar{Y})^2 + \sum_{i=1}^{n} e_i(K)^2 + 2\sum_{i=1}^{n} (\hat{Y}_i(K) - \bar{Y}) \cdot e_i(K).$$

  When the OLS estimators are obtained ($K = 0$), the third term is null. However, this term is not null when K is not zero. Consequently, the relationship $TSS(K) = ESS(K) + RSS(K)$ is not satisfied in ridge regression, and the definition of the coefficient of determination may not be suitable. This fact not only limits the analysis of the goodness of fit but also affects the global significance since the critical coefficient of determination is also questioned. Rodríguez et al. [28]

showed that the estimators obtained from the penalties mentioned above inherit the problem of the ridge regression in relation to the goodness of fit.

In order to overcome these problems, this paper is focused on the raise regression (García et al. [29] and Salmerón et al. [30]) based on the treatment of collinearity from a geometrical point of view. It consists in separating the independent variables by using the residuals (weighted by the raising factor) of the auxiliary regression traditionally used to obtain the VIF. Salmerón et al. [30] showed that the raise regression presents better conditions than ridge regression and, more recently, García et al. [31] showed, among other questions, that the ridge regression is a particular case of the raise regression.

This paper presents the extension of the VIF to the raise regression showing that, although García et al. [31] showed that the application of the raise regression guarantees a diminishing of the VIF, it is not guaranteed that its value is lower the threshold traditionally established as troubling. Thus, it will be concluded that an unique application of the raise regression does not guarantee the mitigation of the multicollinearity. Consequently, this extension complements the results presented by García et al. [31] and determines, on the one hand, whether it is necessary to apply a successive raise regression (see García et al. [31] for more details) and, on the other hand, the most adequate variable for raising and the most optimal value for the raising factor in order to guarantee the mitigation of the multicollinearity.

On the other hand, the transformation of variables is common when strong collinearity exists in a linear model. The transformation to unit length (see Belsley et al. [32]) or standardization (see Marquardt [27]) is typical. Although the VIF is invariant to these transformations when it is calculated after estimation by OLS (see García et al. [26]), it is not guaranteed either in the case of the raise regression or in ridge regression as showed by García et al. [26]. The analysis of this fact is one of the goals of this paper.

Finally, since the raise estimator is biased, it is interesting to calculate its mean square error (MSE). It is studied whether the MSE of the raise regression is less than the one obtained by OLS. In this case, this study could be used to select an adequate raising factor similar to what is proposed by Hoerl et al. [33] in the case of the ridge regression. Note that estimators with MSE less than the one from OLS estimators are traditionally preferred (see, for example, Stein [34], James and Stein [35], Hoerl and Kennard [4], Ohtani [36], or Hubert et al. [37]). In addition, this measure allows us to conclude whether the raise regression is preferable, in terms of MSE, to other alternative techniques.

The structure of the paper is as follows: Section 2 briefly describes the VIF and the raise regression, and Section 3 extends the VIF to this methodology. Some desirable properties of the VIF are analyzed, and its asymptotic behavior is studied. It is also concluded that the VIF is invariant to data transformation. Section 4 calculates the MSE of the raise estimator, showing that there is a minimum value that is less than the MSE of the OLS estimator. Section 5 illustrates the contribution of this paper with two numerical examples. Finally, Section 6 summarizes the main conclusions of this paper.

2. Preliminaries

2.1. Variance Inflation Factor

The following model for p independent variables and n observations is considered:

$$\mathbf{Y} = \beta_1 + \beta_2\mathbf{X}_2 + \cdots + \beta_i\mathbf{X}_i + \cdots + \beta_p\mathbf{X}_p + \mathbf{u} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{2}$$

where $\mathbf{Y}$ is a vector $n \times 1$ that contains the observations of the dependent variable, $\mathbf{X} = [\mathbf{1}\ \mathbf{X}_2 \dots \mathbf{X}_i \dots \mathbf{X}_p]$ (with $\mathbf{1}$ being a vector of ones with dimension $n \times 1$) is a matrix with order $n \times p$ that contains (by columns) the observations of the independent variables, $\boldsymbol{\beta}$ is a vector $p \times 1$ that contains the coefficients of the independent variables, and $\mathbf{u}$ is a vector $n \times 1$ that represents the random disturbance that is supposed to be spherical ($E[\mathbf{u}] = \mathbf{0}$ and $Var(\mathbf{u}) = \sigma^2\mathbf{I}$, where $\mathbf{0}$ is a vector with zeros with dimension $n \times 1$ and $\mathbf{I}$ the identity matrix with adequate dimensions, in this case $p \times p$).

Given the model in Equation (2), the variance inflation factor (VIF) is obtained as follows:

$$VIF(k) = \frac{1}{1 - R_k^2}, \quad k = 2, \ldots, p, \tag{3}$$

where R_k^2 is the coefficient of determination of the regression of the variable $\mathbf{X}_k$ as a function of the rest of the independent variables of the model in Equation (2):

$$\mathbf{X}_k = \alpha_1 + \alpha_2 \mathbf{X}_2 + \cdots + \alpha_{k-1}\mathbf{X}_{k-1} + \alpha_{k+1}\mathbf{X}_{k+1} + \cdots + \alpha_p \mathbf{X}_p + \mathbf{v} = \mathbf{X}_{-k}\boldsymbol{\alpha} + \mathbf{v}, \tag{4}$$

where $\mathbf{X}_{-k}$ corresponds to the matrix $\mathbf{X}$ after the elimination of the column k (variable $\mathbf{X}_k$).

If the variable $\mathbf{X}_k$ has no linear relationship (i.e., is orthogonal) with the rest of the independent variables, the coefficient of determination will be zero ($R_k^2 = 0$) and the $VIF(k) = 1$. As the linear relationship increases, the coefficient of determination (R_k^2) and consequently $VIF(k)$ will also increase. Thus, the higher the VIF associated with the variable $\mathbf{X}_k$, the greater the linear relationship between this variable and the rest of the independent variables in the model in Equation (2). It is considered that the collinearity is troubling for values of VIF higher than 10. Note that the VIF ignores the role of the constant term (see, for example, Salmerón et al. [38] or Salmerón et al. [39]), and consequently, this extension will be useful when the multicollinearity is essential; that is to say, when there is a linear relationship between at least two independent variables of the model of regression without considering the constant term (see, for example, Marquandt and Snee [40] for the definitions of essential and nonessential multicollinearity).

2.2. Raise Regression

Raise regression, presented by García et al. [29] and more developed further by Salmerón et al. [30], uses the residuals of the model in Equation (4), $\mathbf{e}_k$, to raise the variable k as $\widetilde{\mathbf{X}}_k = \mathbf{X}_k + \lambda \mathbf{e}_k$ with $\lambda \geq 0$ (called the raising factor) and to verify that $\mathbf{e}_k^t \mathbf{X}_{-k} = \mathbf{0}$, where $\mathbf{0}$ is a vector of zeros with adequate dimensions. In that case, the raise regression consists in the estimation by OLS of the following model:

$$\mathbf{Y} = \beta_1(\lambda) + \beta_2(\lambda)\mathbf{X}_2 + \cdots + \beta_k(\lambda)\widetilde{\mathbf{X}}_k + \cdots + \beta_p(\lambda)\mathbf{X}_p + \widetilde{\mathbf{u}} = \widetilde{\mathbf{X}}\boldsymbol{\beta}(\lambda) + \widetilde{\mathbf{u}}, \tag{5}$$

where $\widetilde{\mathbf{X}} = [\mathbf{1}\ \mathbf{X}_2 \ldots \widetilde{\mathbf{X}}_k \ldots \mathbf{X}_p] = [\mathbf{X}_{-k}\ \widetilde{\mathbf{X}}_k]$. García et al. [29] showed (Theorem 3.3) that this technique does not alter the global characteristics of the initial model. That is to say, the models in Equations (2) and (5) have the same coefficient of determination and experimental statistics for the global significance test.

Figure 1 illustrates the raise regression for two independent variables being geometrically separated by using the residuals weighted by the raising factor λ. Thus, the selection of an adequate value for λ is essential, analogously to what occurs with the ridge factor K. A preliminary proposal about how to select the raising factor in a model with two independent standardized variables can be found in García et al. [41]. Other recently published papers introduce and highlight the various advantages of raise estimators for statistical analysis: Salmerón et al. [30] presented the raise regression for $p = 3$ standardized variables and showed that it presents better properties than the ridge regression and that the individual inference of the raised variable is not altered, García et al. [31] showed that it is guaranteed that all the VIFs associated with the model in Equation (5) diminish but that it is not possible to quantify the decrease, García and Ramírez [42] presented the successive raise regression, and García et al. [31] showed, among other questions, that ridge regression is a particular case of raise regression.

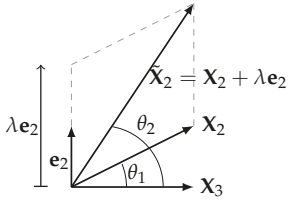


Figure 1. Representation of the raise method.

The following section presents the extension of the VIF to be applied after the estimation by raise regression since it will be interesting whether, after the raising of one independent variable, the VIF falls below 10. It will be also analyzed when a successive raise regression can be recommendable (see García and Ramírez [42]).

3. VIF in Raise Regression

To calculate the VIF in the raise regression, two cases have to be differentiated depending on the dependent variable, $\mathbf{X}_k$, of the auxiliary regression:

1. If it is the raised variable, $\widetilde{\mathbf{X}}_i$ with $i = 2, \ldots, p$, the coefficient of determination, $R_i^2(\lambda)$, of the following auxiliary regression has to be calculated:

$$\begin{array}{rcl} \widetilde{\mathbf{X}}_i & = & \alpha_1(\lambda) + \alpha_2(\lambda)\mathbf{X}_2 + \cdots + \alpha_{i-1}(\lambda)\mathbf{X}_{i-1} + \alpha_{i+1}(\lambda)\mathbf{X}_{i+1} + \cdots + \alpha_p(\lambda)\mathbf{X}_p + \widetilde{\mathbf{v}} \\ & = & \mathbf{X}_{-i}\boldsymbol{\alpha}(\lambda) + \widetilde{\mathbf{v}}. \end{array} \tag{6}$$

2. If it is not the raised variable, $\mathbf{X}_j$ with $j = 2, \ldots, p$ being $j \neq i$, the coefficient of determination, $R_j^2(\lambda)$, of the following auxiliary regression has to be calculated:

$$\begin{array}{rcl} \mathbf{X}_j & = & \alpha_1(\lambda) + \alpha_2(\lambda)\mathbf{X}_2 + \cdots + \alpha_i(\lambda)\widetilde{\mathbf{X}}_i + \cdots + \alpha_{j-1}(\lambda)\mathbf{X}_{j-1} + \alpha_{j+1}(\lambda)\mathbf{X}_{j+1} \\ & & + \cdots + \alpha_p(\lambda)\mathbf{X}_p + \widetilde{\mathbf{v}} \\ & = & \left(\mathbf{X}_{-i,-j}\ \widetilde{\mathbf{X}}_i\right)\begin{pmatrix} \boldsymbol{\alpha}_{-i,-j}(\lambda) \\ \alpha_i(\lambda) \end{pmatrix} + \widetilde{\mathbf{v}}, \end{array} \tag{7}$$

   where $\mathbf{X}_{-i,-j}$ corresponding to the matrix $\mathbf{X}$ after the elimination of columns i and j (variables $\mathbf{X}_i$ and $\mathbf{X}_j$). The same notation is used for $\boldsymbol{\alpha}_{-i,-j}(\lambda)$.

Once these coefficients of determination are obtained (as indicated in the following subsections), the VIF of the raise regression will be given by the following:

$$VIF(k,\lambda) = \frac{1}{1 - R_k^2(\lambda)}, \quad k = 2, \ldots, p. \tag{8}$$

3.1. VIF Associated with Raise Variable

In this case, for $i = 2, \ldots, p$, the coefficient of determination of the regression in Equation (6) is given by

$$\begin{array}{rcl} R_i^2(\lambda) & = & 1 - \frac{(1+2\lambda+\lambda^2)RSS_i^{-i}}{TSS_i^{-i}+(\lambda^2+2\lambda)RSS_i^{-i}} = \frac{ESS_i^{-i}}{TSS_i^{-i}+(\lambda^2+2\lambda)RSS_i^{-i}} \\ & = & \frac{R_i^2}{1+(\lambda^2+2\lambda)(1-R_i^2)}, \end{array} \tag{9}$$

since:

$$\begin{aligned} TSS_i^{-i}(\lambda) &= \widetilde{\mathbf{X}}_i^t\widetilde{\mathbf{X}}_i - n\cdot\overline{\widetilde{\mathbf{X}}}_i^2 = \mathbf{X}_i^t\mathbf{X}_i + (\lambda^2+2\lambda)\mathbf{e}_i^t\mathbf{e}_i - n\cdot\overline{\mathbf{X}}_i^2 \\ &= TSS_i^{-i} + (\lambda^2+2\lambda)RSS_i^{-i}, \\ RSS_i^{-i}(\lambda) &= \widetilde{\mathbf{X}}_i^t\widetilde{\mathbf{X}}_i - \widehat{\boldsymbol{\alpha}}(\lambda)^t\mathbf{X}_{-i}^t\widetilde{\mathbf{X}}_i = \mathbf{X}_i^t\mathbf{X}_i + (\lambda^2+2\lambda)\mathbf{e}_i^t\mathbf{e}_i - \widehat{\boldsymbol{\alpha}}^t\mathbf{X}_{-i}^t\mathbf{X}_i \\ &= (\lambda^2+2\lambda+1)RSS_i^{-i}, \end{aligned}$$

where TSS_i^{-i}, ESS_i^{-i} and RSS_i^{-i} are the total sum of squares, explained sum of squares, and residual sum of squares of the model in Equation (4). Note that it has been taken into account that

$$\widetilde{\mathbf{X}}_i^t\widetilde{\mathbf{X}}_i = (\mathbf{X}_i+\lambda\mathbf{e}_i)^t(\mathbf{X}_i+\lambda\mathbf{e}_i) = \mathbf{X}_i^t\mathbf{X}_i + (\lambda^2+2\lambda)\mathbf{e}_i^t\mathbf{e}_i,$$

since $\mathbf{e}_i^t\mathbf{X}_i = \mathbf{e}_i^t\mathbf{e}_i = RSS_i^{-i}$ and

$$\widehat{\boldsymbol{\alpha}}(\lambda) = \left(\mathbf{X}_{-i}^t\mathbf{X}_{-i}\right)^{-1}\mathbf{X}_{-i}^t\widetilde{\mathbf{X}}_i = \widehat{\boldsymbol{\alpha}},$$

due to $\mathbf{X}_{-i}^t\widetilde{\mathbf{X}}_i = \mathbf{X}_{-i}^t\mathbf{X}_i$.

Indeed, from Equation (9), it is evident that

1. $R_i^2(\lambda)$ decreases as λ increases.
2. $\lim_{\lambda\to+\infty} R_i^2(\lambda) = 0$.
3. $R_i^2(\lambda)$ is continuous in zero; that is to say, $R_i^2(0) = R_i^2$.

Finally, from properties 1) and 3), it is deduced that $R_i^2(\lambda) \leq R_i^2$ for all λ.

3.2. VIF Associated with Non-Raised Variables

In this case, for $j = 2,\ldots,p$, with $j \neq i$, the coefficient of determination of regression in Equation (7) is given by

$$\begin{aligned} R_j^2(\lambda) &= 1 - \frac{RSS_j^{-j}(\lambda)}{TSS_j^{-j}(\lambda)} \\ &= \frac{1}{TSS_j^{-j}}\left(TSS_j^{-j} - RSS_j^{-i,-j} + \frac{RSS_i^{-i,-j}\cdot\left(RSS_j^{-i,-j} - RSS_j^{-j}\right)}{RSS_i^{-i,-j} + (\lambda^2+2\lambda)\cdot RSS_i^{-i}}\right), \end{aligned} \tag{10}$$

Taking into account that $\widetilde{\mathbf{X}}_i^t\mathbf{X}_j = (\mathbf{X}_i+\lambda\mathbf{e}_i)^t\mathbf{X}_j = \mathbf{X}_i^t\mathbf{X}_j$ since $\mathbf{e}_i^t\mathbf{X}_j = 0$, it is verified that

$$TSS_j^{-j}(\lambda) = \mathbf{X}_j^t\mathbf{X}_j - n\cdot\overline{\mathbf{X}}_j^2 = TSS_j^{-j},$$

and, from Appendices A and B,

$$\begin{aligned}
RSS_j^{-j}(\lambda) &= \mathbf{X}_j^t\mathbf{X}_j - \widehat{\boldsymbol{\alpha}}(\lambda)^t \begin{pmatrix} \mathbf{X}_{-i,-j}^t\mathbf{X}_j \\ \widetilde{\mathbf{X}}_i^t\mathbf{X}_j \end{pmatrix} \\
&= \mathbf{X}_j^t\mathbf{X}_j - \widehat{\boldsymbol{\alpha}}_{-i,-j}(\lambda)^t\mathbf{X}_{-i,-j}^t\mathbf{X}_j - \widehat{\alpha}_i(\lambda)^t\mathbf{X}_i^t\mathbf{X}_j \\
&\underbrace{=}_{\text{Appendix A}} \mathbf{X}_j^t\mathbf{X}_j - \mathbf{X}_j^t\mathbf{X}_{-i,-j}\left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1}\mathbf{X}_{-i,-j}^t\mathbf{X}_j \\
&\quad - \frac{RSS_i^{-i,-j}}{RSS_i^{-i,-j} + (\lambda^2+2\lambda)\cdot RSS_i^{-i}} \cdot \\
&\quad \cdot \left(RSS_i^{-i,-j}\mathbf{X}_j^t\mathbf{X}_{-i,-j}\cdot B\cdot B^t\cdot\mathbf{X}_{-i,-j}^t\mathbf{X}_j \right. \\
&\quad \left. + \mathbf{X}_j^t\mathbf{X}_i\cdot B^t\cdot\mathbf{X}_{-i,-j}^t\mathbf{X}_j\right) \\
&\quad - \frac{RSS_i^{-i,-j}}{RSS_i^{-i,-j} + (\lambda^2+2\lambda)\cdot RSS_i^{-i}}\cdot\widehat{\alpha}_i^t\mathbf{X}_i^t\mathbf{X}_j \\
&= \mathbf{X}_j^t\left(\mathbf{I} - \mathbf{X}_{-i,-j}\left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1}\mathbf{X}_{-i,-j}^t\right)\mathbf{X}_j \\
&\quad - \frac{RSS_i^{-i,-j}}{RSS_i^{-i,-j} + (\lambda^2+2\lambda)\cdot RSS_i^{-i}} \cdot \\
&\quad \cdot \left(RSS_i^{-i,-j}\mathbf{X}_j^t\mathbf{X}_{-i,-j}\cdot B\cdot B^t\cdot\mathbf{X}_{-i,-j}^t\mathbf{X}_j \right. \\
&\quad \left. + \mathbf{X}_j^t\mathbf{X}_i\cdot B^t\cdot\mathbf{X}_{-i,-j}^t\mathbf{X}_j + \widehat{\alpha}_i^t\mathbf{X}_i^t\mathbf{X}_j\right) \\
&\underbrace{=}_{\text{Appendix B}} RSS_j^{-i,-j} \\
&\quad - \frac{RSS_i^{-i,-j}}{RSS_i^{-i,-j} + (\lambda^2+2\lambda)\cdot RSS_i^{-i}}\cdot\left(RSS_j^{-i,-j} - RSS_j^{-j}\right),
\end{aligned}$$

where TSS_j^{-j} and RSS_j^{-j} are the total sum of squares and residual sum of squares of the model in Equation (4) and where $RSS_i^{-i,-j}$ and $RSS_j^{-i,-j}$ are the residual sums of squares of models:

$$\mathbf{X}_i = \mathbf{X}_{-i,-j}\boldsymbol{\gamma} + \boldsymbol{\eta}, \tag{11}$$

$$\mathbf{X}_j = \mathbf{X}_{-i,-j}\boldsymbol{\delta} + \boldsymbol{\nu}. \tag{12}$$

Indeed, from Equation (10), it is evident that

1. $R_j^2(\lambda)$ decreases as λ increases.
2. $\lim_{\lambda\to+\infty} R_j^2(\lambda) = \frac{TSS_j^{-j} - RSS_j^{-i,-j}}{TSS_j^{-j}}$.
3. $R_j^2(\lambda)$ is continuous in zero. That is to say, $R_j^2(0) = \frac{TSS_j^{-j} - RSS_j^{-j}}{TSS_j^{-j}} = R_j^2$.

Finally, from properties 1) and 3), it is deduced that $R_j^2(\lambda) \leq R_j^2$ for all λ.

3.3. Properties of $VIF(k,\lambda)$

From conditions verified by the coefficient of determination in Equations (9) and (10), it is concluded that $VIF(k,\lambda)$ (see expression Equation (8)), verifies that

1. The VIF associated with the raise regression is continuous in zero because the coefficients of determination of the auxiliary regressions in Equations (6) and (7) are also continuous in zero. That is to say, for $\lambda = 0$, it coincides with the VIF obtained for the model in Equation (2) when it is estimated by OLS:

$$VIF(k,0) = \frac{1}{1-R_k^2(0)} = \frac{1}{1-R_k^2} = VIF(k), \quad k = 2,\ldots,p.$$

2. The VIF associated with the raise regression decreases as λ increases since this is the behavior of the coefficient of determination of the auxiliary regressions in Equations (6) and (7). Consequently,

$$VIF(k,\lambda) = \frac{1}{1-R_k^2(\lambda)} \leq \frac{1}{1-R_k^2} = VIF(k), \quad k = 2,\ldots,p, \quad \forall \lambda \geq 0.$$

3. The VIF associated with the raised variable is always higher than one since

$$\lim_{\lambda\to+\infty} VIF(i,\lambda) = \lim_{\lambda\to+\infty} \frac{1}{1-R_i^2(\lambda)} = \frac{1}{1-0} = 1, \quad i = 2,\ldots,p.$$

4. The VIF associated with the non-raised variables has a horizontal asymptote since

$$\begin{aligned} \lim_{\lambda\to+\infty} VIF(j,\lambda) &= \lim_{\lambda\to+\infty} \frac{1}{1-R_j^2(\lambda)} = \frac{1}{1-\frac{TSS_j^{-j}-RSS_j^{-i,-j}}{TSS_j^{-j}}} \\ &= \frac{TSS_j^{-j}}{RSS_j^{-i,-j}} = \frac{TSS_j^{-i,-j}}{RSS_j^{-i,-j}} = \frac{1}{1-R_{ij}^2} = VIF_{-i}(j), \end{aligned}$$

where R_{ij}^2 is the coefficient of determination of the regression in Equation (12) for $j = 2,\ldots,p$ and $j \neq i$. Indeed, this asymptote corresponds to the VIF, $VIF_{-i}(j)$, of the regression $\mathbf{Y} = \mathbf{X}_{-i}\boldsymbol{\xi} + \mathbf{w}$ and, consequently, will also always be equal to or higher than one.

Thus, from properties (1) to (4), $VIF(k,\lambda)$ has the very desirable properties of being continuous, monotone in the raise parameter, and higher than one, as presented in García et al. [26].

In addition, the property (4) can be applied to determine the variable to be raised only considering the one with a lower horizontal asymptote. If the asymptote is lower than 10 (the threshold established traditionally as worrying), the extension could be applied to determine the raising factor by selecting, for example, the first λ that verifies $VIF(k,\lambda) < 10$ for $k = 2,\ldots,p$. If none of the $p-1$ asymptotes is lower than the established threshold, it will not be enough to raise one independent variable and a successive raise regression will be recommended (see García and Ramírez [42] and García et al. [31] for more details). Note that, if it were necessary to raise more than one variable, it is guaranteed that there will be values of the raising parameter that mitigate multicollinearity since, in the extreme case where all the variables of the model are raised, all the VIFs associated with the raised variables tend to one.

3.4. Transformation of Variables

The transformation of data is very common when working with models where strong collinearity exists. For this reason, this section analyzes whether the transformation of the data affects the VIF obtained in the previous section.

Since the expression given by Equation (9) can be expressed with $i = 2, \ldots, p$ in the function of R_i^2:

$$R_i^2(\lambda) = \frac{R_i^2}{1 + (\lambda^2 + 2\lambda) \cdot (1 - R_i^2)},$$

it is concluded that it is invariant to origin and scale changes and, consequently, the VIF calculated from it will also be invariant.

On the other hand, the expression given by Equation (10) can be expressed for $j = 2, \ldots, p$, with $j \neq i$ as

$$\begin{aligned} R_j^2(\lambda) &= 1 - \frac{RSS_j^{-i,-j}}{TSS_j^{-j}} + \frac{1}{TSS_j^{-j}} \cdot \frac{RSS_i^{-i,-j} \cdot (RSS_j^{-i,-j} - RSS_j^{-j})}{RSS_i^{-i,-j} + (\lambda^2+2\lambda) \cdot RSS_i^{-i}} \\ &= R_{ij}^2 + \frac{RSS_i^{-i,-j}}{RSS_i^{-i,-j} + (\lambda^2+2\lambda) \cdot RSS_i^{-i}} \cdot \left(\frac{RSS_j^{-i,-j}}{TSS_j^{-i,-j}} - \frac{RSS_j^{-j}}{TSS_j^{-j}} \right) \\ &= R_{ij}^2 + \frac{R_j^2 - R_{ij}^2}{1 + (\lambda^2+2\lambda) \cdot \frac{RSS_i^{-i}}{RSS_i^{-i,-j}}}, \end{aligned} \tag{13}$$

where it was applied that $TSS_j^{-j} = TSS_j^{-i,-j}$.

In this case, by following García et al. [26], transforming the variable $\mathbf{X}_i$ as

$$\mathbf{x}_i = \frac{\mathbf{X}_i - a_i}{b_i},\ a_i \in \mathbb{R},\ b_i \in \mathbb{R} - \{0\}, \quad i = 2, \ldots, p,$$

it is obtained that $RSS_i^{-i}(T) = \frac{1}{b_i^2} RSS_i^{-i}$ and $RSS_i^{-i,-j}(T) = \frac{1}{b_i^2} RSS_i^{-i,-j}$ where $RSS_i^{-i}(T)$ and $RSS_i^{-i,-j}(T)$ are the residual sum of squares of the transformed variables.

Taking into account that $\mathbf{X}_i$ is the dependent variables in the regressions of RSS_i^{-i} and $RSS_i^{-i,-j}$, the following is obtained:

$$\frac{RSS_i^{-i}}{RSS_i^{-i,-j}} = \frac{RSS_i^{-i}(T)}{RSS_i^{-i,-j}(T)}.$$

Then, the expression given by Equation (13) is invariant to data transformations (As long as the dependent variables are transformed from the regressions of RSS_i^{-i} and $RSS_i^{-i,-j}$ in the same form. For example, (a) for considering that a_i is its mean and b_i is its standard deviation (typification), (b) for considering that a_i is its mean and b_i is its standard deviation multiplied by the square root of the number of observations (standardization), or (c) for considering that a_i is zero and b_i is the square root of the squares sum of observations (unit length).) and, consequently, the VIF calculated from it will also be invariant.

4. MSE for Raise Regression

Since the estimator $\boldsymbol{\beta}$ obtained from Equation (5) is biased, it is interesting to study its Mean Square Error (MSE).

Taking into account that, for $k = 2, \ldots, p$,

$$\begin{aligned}\widetilde{\mathbf{X}}_k &= \mathbf{X}_k + \lambda \mathbf{e}_k \\ &= (1+\lambda)\mathbf{X}_k - \lambda\left(\widehat{\alpha}_0 + \widehat{\alpha}_1\mathbf{X}_1 + \cdots + \widehat{\alpha}_{k-1}\mathbf{X}_{k-1} + \widehat{\alpha}_{k+1}\mathbf{X}_{k+1} + \cdots + \widehat{\alpha}_p\mathbf{X}_p\right),\end{aligned}$$

it is obtained that matrix $\widetilde{\mathbf{X}}$ of the expression in Equation (5) can be rewritten as $\widetilde{\mathbf{X}} = \mathbf{X} \cdot \mathbf{M}_\lambda$, where

$$\mathbf{M}_\lambda = \begin{pmatrix} 1 & 0 & \cdots & 0 & -\lambda\widehat{\alpha}_0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & -\lambda\widehat{\alpha}_1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 & -\lambda\widehat{\alpha}_{k-1} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1+\lambda & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & -\lambda\widehat{\alpha}_{k+1} & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & -\lambda\widehat{\alpha}_p & 0 & \cdots & 1 \end{pmatrix}. \tag{14}$$

Thus, we have $\widehat{\boldsymbol{\beta}}(\lambda) = (\widetilde{\mathbf{X}}^t \cdot \widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^t \cdot \mathbf{Y} = \mathbf{M}_\lambda^{-1} \cdot \widehat{\boldsymbol{\beta}}$, and then, the estimator of $\boldsymbol{\beta}$ obtained from Equation (5) is biased unless $\mathbf{M}_\lambda = \mathbf{I}$, which only occurs when $\lambda = 0$, that is to say, when the raise regression coincides with OLS. Moreover,

$$\begin{aligned} tr\left(Var\left(\widehat{\boldsymbol{\beta}}(\lambda)\right)\right) &= tr(\mathbf{M}_\lambda^{-1} \cdot Var(\widehat{\boldsymbol{\beta}}) \cdot (\mathbf{M}_\lambda^{-1})^t) = \sigma^2 tr((\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}})^{-1}), \\ (E[\widehat{\boldsymbol{\beta}}(\lambda)] - \boldsymbol{\beta})^t(E[\widehat{\boldsymbol{\beta}}(\lambda)] - \boldsymbol{\beta}) &= \boldsymbol{\beta}^t(\mathbf{M}_\lambda^{-1} - \mathbf{I})^t(\mathbf{M}_\lambda^{-1} - \mathbf{I})\boldsymbol{\beta}, \end{aligned}$$

where *tr* denotes the trace of a matrix.

In that case, the MSE for raise regression is

$$\begin{aligned} \mathrm{MSE}\left(\widehat{\boldsymbol{\beta}}(\lambda)\right) &= tr\left(Var\left(\widehat{\boldsymbol{\beta}}(\lambda)\right)\right) + (E[\widehat{\boldsymbol{\beta}}(\lambda)] - \boldsymbol{\beta})^t(E[\widehat{\boldsymbol{\beta}}(\lambda)] - \boldsymbol{\beta}) \\ &= \sigma^2 tr((\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}})^{-1}) + \boldsymbol{\beta}^t(\mathbf{M}_\lambda^{-1} - \mathbf{I})^t(\mathbf{M}_\lambda^{-1} - \mathbf{I})\boldsymbol{\beta} \\ &\underbrace{=}_{\text{Appendix C}} \sigma^2 tr\left(\left(\mathbf{X}_{-k}^t\mathbf{X}_{-k}\right)^{-1}\right) + \left(1 + \sum_{j=0, j\neq k}^{p}\widehat{\alpha}_j^2\right) \cdot \beta_k^2 \cdot \frac{\lambda^2 + h}{(1+\lambda)^2}, \end{aligned}$$

where $h = \frac{\sigma^2}{\beta_k^2 \cdot RSS_k^{-k}}$.

We can obtain the MSE from the estimated values of σ^2 and β_k from the model in Equation (2).

On the other hand, once the estimations are obtained and taking into account the Appendix C, $\lambda_{min} = \frac{\widehat{\sigma}^2}{\widehat{\beta}_k^2 \cdot RSS_k^{-k}}$ minimizes $\mathrm{MSE}\left(\widehat{\boldsymbol{\beta}}(\lambda)\right)$. Indeed, it is verified that $\mathrm{MSE}\left(\widehat{\boldsymbol{\beta}}(\lambda_{min})\right) < \mathrm{MSE}\left(\widehat{\boldsymbol{\beta}}(0)\right)$; that is to say, if the goal is exclusively to minimize the MSE (as in the work presented by Hoerl et al. [33]), λ_{min} should be selected as the raising factor.

Finally, note that, if $\lambda_{min} > 1$, then $\mathrm{MSE}\left(\widehat{\boldsymbol{\beta}}(\lambda)\right) < \mathrm{MSE}\left(\widehat{\boldsymbol{\beta}}(0)\right)$ for all $\lambda > 0$.

5. Numerical Examples

To illustrate the results of previous sections, two different set of data will be used that collect the two situations shown in the graphs of Figures A1 and A2. The second example also compares results obtained by the raise regression to results obtained by the application of ridge and Lasso regression.

5.1. Example 1: $h < 1$

The data set includes different financial variables for 15 Spanish companies for the year 2016 (consolidated account and results between €800,000 and €9,000,000) obtained from the dabase Sistema de Análisis de Balances Ibéricos (SABI) database. The relationship is studied between the number of employees, $\boldsymbol{E}$, and the fixed assets (€), $\boldsymbol{FA}$; operating income (€), $\boldsymbol{OI}$; and sales (€), $\boldsymbol{S}$. The model is expressed as

$$\boldsymbol{E} = \beta_1 + \beta_2 \boldsymbol{FA} + \beta_3 \boldsymbol{OI} + \beta_4 \boldsymbol{S} + \boldsymbol{u}. \tag{15}$$

Table 1 displays the results of the estimation by OLS of the model in Equation (15). The presence of essential collinearity in the model in Equation (15) is indicated by the determinant close to zero (0.0000919) of the correlation matrix of independent variables

$$R = \begin{pmatrix} 1 & 0.7264656 & 0.7225473 \\ 0.7264656 & 1 & 0.9998871 \\ 0.7225473 & 0.9998871 & 1 \end{pmatrix},$$

and the VIFs (2.45664, 5200.315, and 5138.535) higher than 10. Note that the collinearity is provoked fundamentally by the relationship between **OI** and **S**.

In contrast, due to the fact that the coefficients of variation of the independent variables (1.015027, 0.7469496, and 0.7452014) are higher than 0.1002506, the threshold established as troubling by Salmerón et al. [39], it is possible to conclude that the nonessential multicollinearity is not troubling. Thus, the extension of the VIF seems appropriate to check if the application of the raise regression has mitigated the multicollinearity.

Remark 1. *$\lambda^{(1)}$ and $\lambda^{(2)}$ will be the raising factor of the first and second raising, respectively.*

Table 1. Estimations of the models in Equations (15)–(18): Standard deviation is inside the parenthesis, R^2 is the coefficient of determination, $F_{3,11}$ is the experimental value of the joint significance contrast, and $\hat{\sigma}^2$ is the variance estimate of the random perturbation.

	Model (15)	***p*-Value**	**Model (16) for $\lambda_{vif}^{(1)} = 24.5$**	***p*-Value**	**Model (17) for $\lambda_{min}^{(1)} = 0.42$ and $\lambda_{vif}^{(2)} = 17.5$**	***p*-Value**	**Model (18) for $\lambda_{mse}^{(1)} = 1.43$ and $\lambda_{vif}^{(2)} = 10$**	***p*-Value**
Intercept	994.21 (17940)	0.957	4588.68 (17,773.22)	0.801	5257.84 (1744.26)	0.772	5582.29 (17740.18)	0.759
FA	−1.28 (0.55)	0.039	−1.59 (0.50)	0.009	−1.59 (0.51)	0.009	−1.58 (0.51)	0.009
OI	−81.79 (52.86)	0.150						
$\widetilde{\mathbf{OI}}_{\lambda_{vif}^{(1)}}$			−3.21 (2.07)	0.150				
$\widetilde{\mathbf{OI}}_{\lambda_{min}^{(1)}}$					1.67 (2.28)	0.478		
$\widetilde{\mathbf{OI}}_{\lambda_{mse}^{(1)}}$							1.51 (2.24)	0.517
S	87.58 (53.29)	0.129	8.38 (2.35)	0.004				
$\widetilde{\mathbf{S}}_{\lambda_{vif}^{(2)}}$					3.42 (2.03)	0.120	3.55 (1.99)	0.103
R^2	0.70		0.70		0.70		0.70	
$F_{3,11}$	8.50		8.50		8.50		8.50	
$\hat{\sigma}^2$	1,617,171,931		1,617,171,931		1,617,171,931		1,617,171,931	
MSE	321,730,738		321,790,581		336,915,567		325,478,516	

5.1.1. First Raising

A possible solution could be to apply the raise regression to try to mitigate the collinearity. To decide which variable is raised, the thresholds for the VIFs associated with the raise regression are calculated with the goal of raising the variable that the smaller horizontal asymptotes present. In addition to raising the variable that presents the lowest VIF, it would be interesting to obtain a lower mean squared error (MSE) after raising. For this, the $\lambda^{(1)}_{min}$ is calculated for each case. Results are shown in Table 2. Note that the variable to be raised should be the second or third since their asymptotes are lower than 10, although in both cases $\lambda^{(1)}_{min}$ is lower than 1 and it is not guaranteed that the MSE of the raise regression will be less than the one obtained from the estimation by the OLS of the model in Equation (15). For this reason, this table also shows the values of $\lambda^{(1)}$ that make the MSE of the raise regression coincide with the MSE of the OLS regression, $\lambda^{(1)}_{mse}$, and the minimum value of $\lambda^{(1)}$ that leads to values of VIF less than 10, $\lambda^{(1)}_{vif}$.

Table 2. Horizontal asymptotes for variance inflation factors (VIF) after raising each variable and $\lambda^{(1)}_{min}$, $\lambda^{(1)}_{mse}$, and $\lambda^{(1)}_{vif}$.

Raised	$\lim_{\lambda^{(1)}\to+\infty} VIF\,(\mathbf{FA}, \lambda^{(1)})$	$\lim_{\lambda^{(1)}\to+\infty} VIF\,(\mathbf{OI}, \lambda^{(1)})$	$\lim_{\lambda^{(1)}\to+\infty} VIF\,(\mathbf{S}, \lambda^{(1)})$
Variable 1	1	4429.22	4429.22
Variable 2	2.09	1	2.09
Variable 3	2.12	2.12	1
Raised	$\lambda^{(1)}_{min}$	$\lambda^{(1)}_{mse}$	$\lambda^{(1)}_{vif}$
Variable 1	0.18	0.45	$\nexists$
Variable 2	0.42	1.43	24.5
Variable 3	0.37	1.18	24.7

Figure 2 displays the VIF associated with the raise regression for $0 \leq \lambda^{(1)} \leq 900$ after raising the second variable. It is observed that VIFs are always higher than its corresponding horizontal asymptotes.

The model after raising the second variable will be given by

$$\boldsymbol{E} = \beta_1(\lambda) + \beta_2(\lambda)\mathbf{FA} + \beta_3(\lambda)\widetilde{\mathbf{OI}} + \beta_4(\lambda)\mathbf{S} + \widetilde{\boldsymbol{u}}, \tag{16}$$

where $\widetilde{\mathbf{OI}} = \mathbf{OI} + \lambda^{(1)} \cdot \mathbf{e_{OI}}$ with $\mathbf{e_{OI}}$ the residual of regression:

$$\mathbf{OI} = \alpha_1 + \alpha_2\mathbf{FA} + \alpha_3\mathbf{S} + \boldsymbol{v}.$$

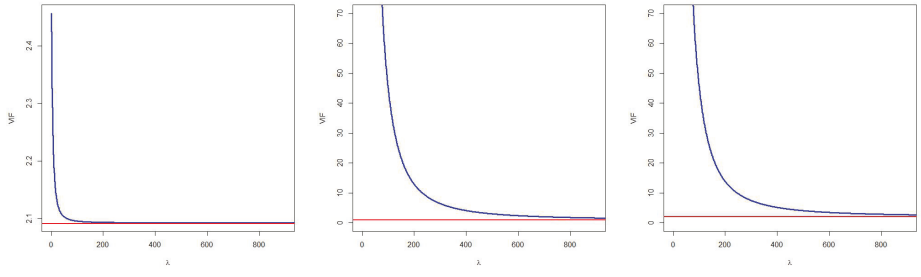


Figure 2. VIF of the variables after raising **OI**.

Remark 2. *The coefficient of variation of* $\widetilde{\mathbf{OI}}$ *for* $\lambda^{(1)} = 24.5$ *is equal to 0.7922063; that is to say, it was lightly increased.*

As can be observed from Table 3, in Equation (16), the collinearity is not mitigated by considering $\lambda^{(1)}$ equal to $\lambda^{(1)}_{min}$ and $\lambda^{(1)}_{mse}$. For this reason, Table 1 only shows the values of the model in Equation (16) for the value of $\lambda^{(1)}$ that leads to VIF lower than 10.

Table 3. VIF of regression Equation (16) for $\lambda^{(1)}$ equal to $\lambda^{(1)}_{min}$, $\lambda^{(1)}_{mse}$, and $\lambda^{(1)}_{vif}$.

	$VIF\,(\mathbf{FA}, \lambda^{(1)})$	$VIF\,(\widetilde{\mathbf{OI}}, \lambda^{(1)})$	$VIF\,(\mathbf{S}, \lambda^{(1)})$
$\lambda^{(1)}_{min}$	2.27	2587.84	2557.66
$\lambda^{(1)}_{mse}$	2.15	878.10	868.58
$\lambda^{(1)}_{vif}$	2.09	9.00	9.99

5.1.2. Transformation of Variables

After the first raising, it is interesting to verify that the VIF associated with the raise regression is invariant to data transformation. With this goal, the second variable has been raised, obtaining the $VIF(\mathbf{FA}, \lambda^{(1)})$, $VIF(\widetilde{\mathbf{OI}}, \lambda^{(1)})$, and $VIF(\mathbf{S}, \lambda^{(1)})$ for $\lambda^{(1)} \in \{0, 0.5, 1, 1.5, 2, \ldots, 9.5, 10\}$, supposing original, unit length, and standardized data. Next, the three possible differences and the average of the VIF associated with each variable are obtained. Table 4 displays the results from which it is possible to conclude that differences are almost null and that, consequently, the VIF associated with the raise regression is invariant to the most common data transformation.

Table 4. Effect of data transformations on VIF associated with raise regression.

	$VIF(\mathbf{FA}, \lambda^{(1)})$	$VIF(\widetilde{\mathbf{OE}}, \lambda^{(1)})$	$VIF(\mathbf{S}, \lambda^{(1)})$
Original–Unit length	$9.83 \cdot 10^{-16}$	$1.55 \cdot 10^{-11}$	$1.83 \cdot 10^{-10}$
Original–Standardized	$-1.80 \cdot 10^{-16}$	$-3.10 \cdot 10^{-10}$	$2.98 \cdot 10^{-10}$
Unit length–Standardized	$-1.16 \cdot 10^{-15}$	$-3.26 \cdot 10^{-10}$	$1.15 \cdot 10^{-10}$

5.1.3. Second Raising

After the first raising, we can use the results obtained from the value of λ that obtains all VIFs less than 10 or consider the results obtained for λ_{min} or λ_{mse} and continue the procedure with a second raising. By following the second option, we part from the value of $\lambda^{(1)} = \lambda^{(1)}_{min} = 0.42$ obtained after the first raising. From Table 5, the third variable is selected to be raised. Table 6 shows the VIF associated with the following model for $\lambda^{(2)}_{min}$, $\lambda^{(2)}_{mse}$, and $\lambda^{(2)}_{vif}$:

$$\boldsymbol{E} = \beta_1(\lambda) + \beta_2(\lambda)\mathbf{FA} + \beta_3(\lambda)\widetilde{\mathbf{OI}} + \beta_4(\lambda)\widetilde{\mathbf{S}} + \widetilde{\boldsymbol{u}}, \tag{17}$$

where $\widetilde{\mathbf{S}} = \mathbf{S} + \lambda^{(2)} \cdot \mathbf{e_S}$ with $\mathbf{e_S}$ the residuals or regression:

$$\mathbf{S} = \alpha_1(\lambda) + \alpha_2(\lambda)\mathbf{FA} + \alpha_3(\lambda)\widetilde{\mathbf{OI}} + \widetilde{\mathbf{v}}.$$

Remark 3. *The coefficient of variation of* $\widetilde{\mathbf{OI}}$ *for* $\lambda^{(1)} = 0.42$ *is equal to 0.7470222, and the coefficient of variation of* $\widetilde{\mathbf{S}}$ *for* $\lambda^{(2)} = 17.5$ *is equal to 0.7473472. In both cases, they were slightly increased.*

Note than it is only possible to state that collinearity has been mitigated when $\lambda^{(2)} = \lambda^{(2)}_{vif} = 17.5$. Results of this estimation are displayed in Table 1.

Table 5. Horizontal asymptote for VIFs after raising each variable in the second raising for $\lambda^{(2)}_{min}$, $\lambda^{(2)}_{mse}$ and $\lambda^{(2)}_{vif}$.

Raised	$\lim_{\lambda^{(2)}\to+\infty} VIF(\mathbf{FA},\lambda^{(2)})$	$\lim_{\lambda^{(2)}\to+\infty} VIF(\widetilde{\mathbf{OI}},\lambda^{(2)})$	$\lim_{\lambda^{(2)}\to+\infty} VIF(\mathbf{S},\lambda^{(2)})$
Variable 1	1	2381.56	2381.56
Variable 3	2.12	2.12	1
Raised	$\lambda^{(2)}_{min}$	$\lambda^{(2)}_{mse}$	$\lambda^{(2)}_{vif}$
Variable 1	0.15	0.34	∄
Variable 3	0.35	1.09	17.5

Table 6. VIFs of regression Equation (16) for $\lambda^{(2)}$ equal to $\lambda^{(2)}_{min}$, $\lambda^{(2)}_{mse}$, and $\lambda^{(2)}_{vif}$.

	$VIF(\mathbf{FA},\lambda^{(2)})$	$VIF(\widetilde{\mathbf{OI}},\lambda^{(2)})$	$VIF(\widetilde{\mathbf{S}},\lambda^{(2)})$
$\lambda^{(2)}_{min}$	2.20	1415.06	1398.05
$\lambda^{(2)}_{mse}$	2.15	593.98	586.20
$\lambda^{(2)}_{vif}$	2.12	9.67	8.47

Considering that, after the first raising, it is obtained that $\lambda^{(1)} = \lambda^{(1)}_{mse} = 1.43$, from Table 7, the third variable is selected to be raised. Table 8 shows the VIF associated with the following model for $\lambda^{(2)}_{min}$, $\lambda^{(2)}_{mse}$, and $\lambda^{(2)}_{vif}$:

$$\boldsymbol{E} = \beta_1(\lambda) + \beta_2(\lambda)\mathbf{FA} + \beta_3(\lambda)\widetilde{\mathbf{OI}} + \beta_4(\lambda)\widetilde{\mathbf{S}} + \widetilde{\boldsymbol{u}}, \tag{18}$$

where $\widetilde{\mathbf{S}} = \mathbf{S} + \lambda \cdot \mathbf{e_S}$.

Remark 4. *The coefficient of variation of* $\widetilde{\mathbf{OI}}$ *for* $\lambda^{(1)} = 1.43$ *is equal to 0.7473033, and the coefficient of variation of* $\widetilde{\mathbf{S}}$ *for* $\lambda^{(2)} = 10$ *is equal to 0.7651473. In both cases, they were lightly increased.*

Remark 5. *Observing the coefficients of variation of* $\widetilde{\mathbf{OI}}$ *for different raising factor. it is concluded that the coefficient of variation increases as the raising factor increases: 0.7470222 (*$\lambda = 0.42$*), 0.7473033 (*$\lambda = 1.43$*), and 0.7922063 (*$\lambda = 24.5$*).*

Note that it is only possible to state that collinearity has been mitigated when $\lambda^{(2)} = \lambda^{(2)}_{vif} = 10$. Results of the estimations of this model are shown in Table 1.

Table 7. Horizontal asymptote for VIFs after raising each variables in the second raising for $\lambda^{(2)}_{min}$, $\lambda^{(2)}_{mse}$, and $\lambda^{(2)}_{vif}$.

Raised	$\lim_{\lambda^{(2)}\to+\infty} VIF(\mathbf{FA},\lambda^{(2)})$	$\lim_{\lambda^{(2)}\to+\infty} VIF(\widetilde{\mathbf{OI}},\lambda^{(2)})$	$\lim_{\lambda^{(2)}\to+\infty} VIF(\mathbf{S},\lambda^{(2)})$
Variable 1	1	853.40	853.40
Variable 3	2.12	2.12	1
Raised	$\lambda^{(2)}_{min}$	$\lambda^{(2)}_{mse}$	$\lambda^{(2)}_{vif}$
Variable 1	0.12	0.27	∄
Variable 3	0.32	0.92	10

Table 8. VIFs of regression Equation (16) for $\lambda^{(2)}$ equal to $\lambda^{(2)}_{min}$, $\lambda^{(2)}_{mse}$, and $\lambda^{(2)}_{vif}$.

	$VIF(\mathbf{FA}, \lambda^{(2)})$	$VIF(\widetilde{\mathbf{OI}}, \lambda^{(2)})$	$VIF(\widetilde{\mathbf{S}}, \lambda^{(2)})$
$\lambda^{(2)}_{min}$	2.14	508.54	502.58
$\lambda^{(2)}_{mse}$	2.13	239.42	236.03
$\lambda^{(2)}_{vif}$	2.12	9.36	8.17

5.1.4. Interpretation of Results

Analyzing the results of Table 1, it is possible to conclude that

1. In the model in Equation (16) (in which the second variable is raised considering the smallest λ that makes all the VIFs less than 10, $\lambda^{(1)} = 24.5$), the variable sales have a coefficient significantly different from zero, where in the original model this was not the case. In this case, the MSE is superior to the one obtained by OLS.
2. In the model in Equation (17) (in which the second variable is raised considering the value of λ that minimizes the MSE, $\lambda^{(1)} = 0.42$, and after that, the third variable is raised considering the smallest λ that makes all the VIFs less than 10, $\lambda^{(2)} = 17.5$), there is no difference in the individual significance of the coefficient.
3. In the model in Equation (18) (in which the second variable is raised considering the value of λ that makes the MSE of the raise regression coincide with that of OLS, $\lambda^{(1)} = 1.43$, and next, the third variable is raised considering the smallest λ that makes all the VIFs less than 10, $\lambda^{(2)} = 10$), there is no difference in the individual significance of the coefficient.
4. Although the coefficient of variable **OI** is not significantly different from zero in any case, the not expected negative sign obtained in model in Equation (15) is corrected in models Equations (17) and (18).
5. In the models with one or two raisings, all the global characteristics coincide with that of the model in Equation (15). Furthermore, there is a relevant decrease in the estimation of the standard deviation for the second and third variable.
6. In models with one or two raisings, the MSE increases, with the model in Equation (16) being the one that presents the smallest MSE among the biased models.

Thus, in conclusion, the model in Equation (16) is selected as it presents the smallest MSE and there is an improvement in the individual significance of the variables.

5.2. Example 2: $h > 1$

This example uses the following model previously applied by Klein and Goldberger [43] about consumption and salaries in the United States from 1936 to 1952 (1942 to 1944 were war years, and data are not available):

$$\mathbf{C} = \beta_1 + \beta_2 \mathbf{WI} + \beta_3 \mathbf{NWI} + \beta_4 \mathbf{FI} + \mathbf{u}, \tag{19}$$

where **C** is consumption, **WI** is wage income, **NWI** is non-wage, non-farm income, and **FI** is the farm income. Its estimation by OLS is shown in Table 9.

However, this estimation is questionable since no estimated coefficient is significantly different to zero while the model is globally significant (with 5% significance level), and the VIFs associated with each variable (12.296, 9.23, and 2.97) indicate the presence of severe essential collinearity. In addition, the determinant of the matrix of correlation

$$\mathbf{R} = \begin{pmatrix} 1 & 0.9431118 & 0.8106989 \\ 0.9431118 & 1 & 0.7371272 \\ 0.8106989 & 0.7371272 & 1 \end{pmatrix},$$

is equal to 0.03713592 and, consequently, lower than the threshold recommended by García et al. [44] ($1.013 \cdot 0.1 + 0.00008626 \cdot n - 0.01384 \cdot p = 0.04714764$ being $n = 14$ and $p = 4$); it is maintained the conclusion that the near multicollinearity existing in this model is troubling.

Once again, the values of the coefficients of variation (0.2761369, 0.2597991, and 0.2976122) indicate that the nonessential multicollinearity is not troubling (see Salmerón et al. [39]). Thus, the extension of the VIF seems appropriate to check if the application of the raise regression has mitigated the near multicollinearity.

Next, it is presented the estimation of the model by raise regression and the results are compared to the estimation by ridge and Lasso regression.

5.2.1. Raise Regression

When calculating the thresholds that would be obtained for VIFs by raising each variable (see Table 10), it is observed that, in all cases, they are less than 10. However, when calculating λ_{min} in each case, a value higher than one is only obtained when raising the third variable. Figure 3 displays the MSE for $\lambda \in [0, 37)$. Note that $MSE(\widehat{\boldsymbol{\beta}}(\lambda))$ is always less than the one obtained by OLS, 49.434, and presents an asymptote in $\lim_{\lambda \to +\infty} MSE(\widehat{\boldsymbol{\beta}}(\lambda)) = 45.69422$.

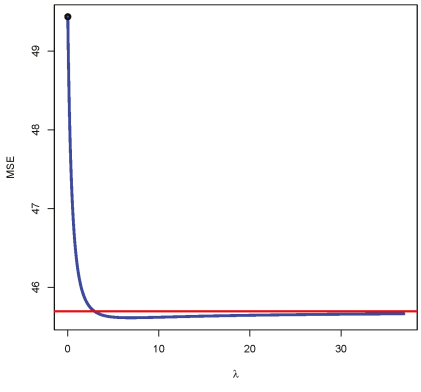


Figure 3. Mean square error (MSE) for the model in Equation (19) after raising third variable.

Table 9. Estimation of the original and raised models: Standard deviation is inside the parentheses, R^2 is the coefficient of determination, $F_{3,10}$ is the experimental value of the joint significance contrast, and $\hat{\sigma}^2$ is the variance estimate of the random perturbation.

	Model (19)	***p*-Value**	**Model (20) for $\lambda_{min} = 6.895$**	***p*-Value**	**Model (21) for $\lambda_{min} = 0.673$**	***p*-Value**
Intercept	18.7021 (6.8454)	0.021	19.21507 (6.67216)	0.016	18.2948 (6.8129)	0.023
WI	0.3803 (0.3121)	0.251	0.43365 (0.26849)	0.137		
$\widetilde{\mathbf{WI}}$					0.2273 (0.1866)	0.251
NWI	1.4186 (0.7204)	0.077	1.38479 (0.71329)	0.081	1.7269 (0.5143)	0.007
FI	0.5331 (1.3998)	0.711			0.8858 (1.2754)	0.503
$\widetilde{\mathbf{FI}}$			0.06752 (0.17730)	0.711		
R^2	0.9187		0.9187		0.9187	
$\hat{\sigma}$	6.06		6.06		6.06	
$F_{3,10}$	37.68		37.68		37.68	
MSE	49.43469		45.61387		48.7497	

Table 10. Horizontal asymptote for VIFs after raising each variable and λ_{min}.

Raised	$\lim_{\lambda \to +\infty} VIF(\mathbf{WI}, \lambda)$	$\lim_{\lambda \to +\infty} VIF(\mathbf{NWI}, \lambda)$	$\lim_{\lambda \to +\infty} VIF(\mathbf{FI}, \lambda)$	λ_{min}
Variable 1	1	2.19	2.19	0.673
Variable 2	2.92	1	2.92	0.257
Variable 3	9.05	9.05	1	6.895

The following model is obtained by raising the third variable:

$$\mathbf{C} = \beta_1(\lambda) + \beta_2(\lambda)\mathbf{WI} + \beta_3(\lambda)\mathbf{NWI} + \beta_4(\lambda)\widetilde{\mathbf{FI}} + \tilde{\mathbf{u}}, \tag{20}$$

where $\widetilde{\mathbf{FI}} = \mathbf{FI} + \lambda \cdot \mathbf{e_{FI}}$ being $\mathbf{e_{FI}}$ the residuals of regression:

$$\mathbf{FI} = \alpha_1 + \alpha_2\mathbf{WI} + \alpha_3\mathbf{NWI} + \mathbf{v}.$$

Remark 6. *The coefficient of variation* $\widetilde{\mathbf{FI}}$ *for* $\lambda^{(1)} = 6.895$ *is 1.383309. Thus, the application of the raise regression has mitigated the nonessential multicollinearity in this variable.*

Table 9 shows the results for the model in Equation (20), being $\lambda = 6.895$. In this case, the MSE is the lowest possible for every possible value of λ and lower than the one obtained by OLS for the model in Equation (19). Furthermore, in this case, the collinearity is not strong once all the VIF are lower than 10 (9.098, 9.049, and 1.031, respectively). However, the individual significance in the variable was not improved.

With the purpose of improving this situation, another variable is raised. If the first variable is selected to be raised, the following model is obtained:

$$\mathbf{C} = \beta_1(\lambda) + \beta_2(\lambda)\widetilde{\mathbf{WI}} + \beta_3(\lambda)\mathbf{NWI} + \beta_4(\lambda)\mathbf{FI} + \tilde{\mathbf{u}}, \tag{21}$$

where $\widetilde{\mathbf{WI}} = \mathbf{WI} + \lambda \cdot \mathbf{e_{WI}}$ being $\mathbf{e_{WI}}$ the residuals of regression:

$$\mathbf{WI} = \alpha_1 + \alpha_2\mathbf{NWI} + \alpha_3\mathbf{FI} + \mathbf{v}.$$

Remark 7. *The coefficient of variation of* $\widetilde{\mathbf{WI}}$ *for* $\lambda^{(1)} = 0.673$ *is 0.2956465. Thus, it is noted that the raise regression has lightly mitigated the nonessential mutlicollinearity of this variable.*

Table 9 shows the results for the model in Equation (21), being $\lambda = 0.673$. In this case, the MSE is lower than the one obtained by OLS for the model in Equation (19). Furthermore, in this case, the collinearity is not strong once all the VIF are lower than 10 (5.036024, 4.705204, and 2.470980, respectively). Note that raising this variable, the values of VIFs are lower than raising the first variable but the MSE is higher. However, this model is selected as preferable due to the individual significance being better in this model and the MSE being lower than the one obtained by OLS.

5.2.2. Ridge Regression

This subsection presents the estimation of the model in Equation (19) by ridge regression (see Hoerl and Kennard [4] or Marquardt [45]). The first step is the selection of the appropriate value of K.

The following suggestions are addressed:

- Hoerl et al. [33] proposed the value of $K_{HKB} = p \cdot \frac{\hat{\sigma}^2}{\hat{\beta}^t \hat{\beta}}$ since probability higher than 50% leads to a MSE lower than the one from OLS.
- García et al. [26] proposed the value of K, denoted as K_{VIF}, that leads to values of VIF lower than 10 (threshold traditionally established as troubling).

- García et al. [44] proposed the following values:

$$\begin{aligned} K_{exp} &= 0.006639 \cdot e^{1-det(\mathbf{R})} - 0.00001241 \cdot n + 0.005745 \cdot p, \\ K_{linear} &= 0.01837 \cdot (1 - det(\mathbf{R})) - 0.00001262 \cdot n + 0.005678 \cdot p, \\ K_{sq} &= 0.7922 \cdot (1 - det(\mathbf{R}))^2 - 0.6901 \cdot (1 - det(\mathbf{R})) - 0.000007567 \cdot n \\ &\quad -0.01081 \cdot p, \end{aligned}$$

where $det(\mathbf{R})$ denotes the determinant of the matrix of correlation, $\mathbf{R}$.

The following values are obtained $K_{HKB} = 0.417083$, $K_{VIF} = 0.013$, $K_{exp} = 0.04020704$, $K_{linear} = 0.04022313$, and $K_{sq} = 0.02663591$.

Tables 11 and 12 show (The results for K_{linear} are not considered as they are very similar to results obtained by K_{exp}.) the estimations obtained from ridge estimators (expression (1)) and the individual significance intervals obtained by bootstrap considering percentiles 5 and 95 for 5000 repeats. It is also calculated the goodness of the fit by following the results shown by Rodríguez et al. [28] and the MSE.

Note that only the constant term can be considered significatively different to zero and that, curiously, the value of K proposed by Hoerl et al. [33] leads to a value of MSE higher than the one from OLS while the values proposed by García et al. [26] and García et al. [44] lead to a value of MSE lower than the one obtained by OLS. All cases lead to values of VIF lower than 10; see García et al. [26] for its calculation:

$$\begin{aligned} 2.0529, 1.8933 \text{ and } 1.5678 \quad &\text{for} \quad K_{HKB}, \\ 9.8856, 7.5541 \text{ and } 2.7991 \quad &\text{for} \quad K_{VIF}, \\ 7.1255, 5.6191 \text{ and } 2.5473 \quad &\text{for} \quad K_{exp}, \\ 8.2528, 6.4123 \text{ and } 2.65903 \quad &\text{for} \quad K_{sq}. \end{aligned}$$

In any case, the lack of individual significance justifies the selection of the raise regression as preferable in comparison to the models obtained by ridge regression.

Table 11. Estimation of the ridge models for $K_{HKB} = 0.417083$ and $K_{VIF} = 0.013$. Confidence interval, at 10% confidence, is obtained from bootstrap inside the parentheses, and R^2 is the coefficient of determination obtained from Rodríguez et al. [28].

	Model (19) for $K_{HKB} = 0.417083$	**Model (19) for $K_{VIF} = 0.013$**
Intercept	12.2395 (6.5394, 15.9444)	18.3981 (12.1725, 24.1816)
WI	0.3495 (−0.4376, 1.2481)	0.3787 (−0.4593, 1.216)
NWI	1.6474 (−0.1453, 3.4272)	1.4295 (−0.2405, 3.2544)
FI	0.8133 (−1.5584, 3.028)	0.5467 (−1.827, 2.9238)
R^2	0.8957	0.9353
MSE	64.20028	47.99713

Table 12. Estimation of the ridge models for $K_{exp} = 0.04020704$ and $K_{sq} = 0.02663591$. Confidence interval, at 10% confidence, is obtained from bootstrap inside the parentheses, and R^2 is the coefficient of determination obtained from Rodríguez et al. [28].

	Model (19) for $K_{exp} = 0.04020704$	**Model (19) for $K_{sq} = 0.02663591$**
Intercept	17.7932 (11.4986, 22.9815)	18.0898 (11.8745, 23.8594)
WI	0.3756 (−0.4752, 1.2254)	0.3771 (−0.4653, 1.2401)
NWI	1.4512 (−0.2249, 3.288)	1.4406 (−0.2551, 3.2519)
FI	0.5737 (−1.798, 2.9337)	0.5605 (−1.6999, 2.9505)
R^2	0.918034	0.9183955
MSE	45.76226	46.75402

5.2.3. Lasso Regression

The Lasso regression (see Tibshirani [5]) is a method initially designed to select variables constraining the coefficient to zero, being specially useful in models with a high number of independent variables. However, this estimation methodology has been widely applied in situation where the model presents worrying near multicollinearity.

Table 13 shows results obtained by the application of the Lasso regression to the model in Equation (19) by using the package *glmnet* of the programming environment R Core Team [46]. Note that these estimations are obtained for the optimal value of $\lambda = 0.1258925$ obtained after a k-fold cross-validation.

Table 13. Estimation of the Lasso model for $\lambda = 0.1258925$: Confidence interval at 10% confidence (obtained from bootstrap inside the parentheses).

	Model (19) for $\lambda = 0.1258925$
Intercept	19.1444 (13.5814489, 24.586207)
WI	0.4198 (−0.2013491, 1.052905)
NWI	1.3253 (0.0000000, 2.752345)
FI	0.4675 (−1.1574169, 2.151648)

The inference obtained by bootstrap methodology (with 5000 repeats) allows us to conclude that in, at least, the 5% of the cases, the coefficient of **NWI** is constrained to zero. Thus, this variable should be eliminated from the model.

However, we consider that this situation should be avoided, and as an alternative to the elimination of variable, that is, as an alternative from the following model, the estimation by raise or ridge regression is proposed.

$$\mathbf{C} = \pi_1 + \pi_2\mathbf{WI} + \pi_3\mathbf{FI} + \epsilon, \tag{22}$$

It could be also appropriate to apply the residualization method (see, for example, York [47], Salmerón et al. [48], and García et al. [44]), which consists in the estimation of the following model:

$$\mathbf{C} = \tau_1 + \tau_2\mathbf{WI} + \tau_3\mathbf{FI} + \tau_4\mathbf{res_{NWI}} + \varepsilon, \tag{23}$$

where, for example, $\mathbf{res_{NWI}}$ represents the residuals of the regression of **NWI** as a function of **WI** that will be interpreted as the part of **NWI** not related to **WI**. In this case (see García et al. [44]), it is verified that $\widehat{\pi}_i = \widehat{\tau}_i$ for $i = 1, 2, 3$. That is to say, the model in Equation (23) estimates the same relationship between **WI** and **FI** with **C** as in the model in Equation (22) with the benefit that the variable **NWI** is not eliminated due to a part of it being considered..

6. Conclusions

The Variance Inflation Factor (VIF) is one of the most applied measures to diagnose collinearity together with the Condition Number (CN). Once the collinearity is detected, different methodologies can be applied as, for example, the raise regression, but it will be required to check if the methodology has mitigated the collinearity effectively. This paper extends the concept of VIF to be applied after the raise regression and presents an expression of the VIF that verifies the following desirable properties (see García et al. [26]):

1. continuous in zero. That is to say, when the raising factor (λ) is zero, the VIF obtained in the raise regression coincides with the one obtained by OLS;
2. decreasing as a function of the raising factor (λ). That is to say, the degree of collinearity diminishes as λ increases, and
3. always equal or higher than 1.

The paper also shows that the VIF in the raise regression is scale invariant, which is a very common transformation when working with models with collinearity. Thus, it yields identical results regardless of whether predictions are based on unstandardized or standardized predictors. Contrarily, the VIFs obtained from other penalized regressions (ridge regression, Lasso, and Elastic Net) are not scale invariant and hence yield different results depending on the predictor scaling used.

Another contribution of this paper is the analysis of the asymptotic behavior of the VIF associated with the raised variable (verifying that its limit is equal to 1) and associated with the rest of the variables (presenting an horizontal asymptote). This analysis allows to conclude that

- It is possible to know a priori how far each of the VIFs can decrease simply by calculating their horizontal asymptote. This could be used as a criterion to select the variable to be raised, the one with the lowest horizontal asymptote being chosen.
- If there is asymptote under the threshold established as worrying, the extension of the VIF can be applied to select the raising factor considering the value of λ that verifies $VIF(k, \lambda) < 10$ for $k = 2, \ldots, p$.
- It is possible that the collinearity is not mitigated with any value of λ. This can happen when at least one horizontal asymptote is greater than the threshold. In that case, a second variable has to be raised. García and Ramírez [42] and García et al. [31] show the successive raising procedure.

On the other hand, since the raise estimator is biased, the paper analyzes its Mean Square Error (MSE), showing that there is a value of λ that minimizes the possibility of the MSE being lower than the one obtained by OLS. However, it is not guaranteed that the VIF for this value of λ presents a value less than the established thresholds. The results are illustrated with two numerical examples, and in the second one, the results obtained by OLS are compared to the results obtained with the raise, ridge, and Lasso regressions that are widely applied to estimated models with worrying multicollinearity. It is showed that the raise regression can compete and even overcome these methodologies.

Finally, we propose as future lines of research the following questions:

- The examples showed that the coefficients of variation increase after raising the variables. This fact is associated with an increase in the variability of the variable and, consequently, with a decrease of the near nonessential multicollinearity. Although a deeper analysis is required, it seems that raise regression mitigates this kind of near multicollinearity.
- The value of the ridge factor traditionally applied, K_{HKB}, leads to estimators with smaller MSEs than the OLS estimators with probability greater than 0.5. In contrast, the value of the raising factor λ_{min} always leads to estimators with smaller MSEs than OLS estimators. Thus, it is deduced that the ridge regression provides estimators with MSEs higher than the MSEs of OLS estimators with probability lower than 0.5. These questions seem to indicate that, in terms of MSE, the raise regression can present better behaviour than the ridge regression. However, the confirmation of this judgment will require a more complete analysis, including other aspects such as interpretability and inference.

Author Contributions: conceptualization, J.G.P., C.G.G. and R.S.G. and A.R.S.; methodology, R.S.G. and A.R.S.; software, A.R.S.; validation, J.G.P., R.S.G. and C.G.G.; formal analysis, R.S.G. and C.G.G.; investigation, R.S.G. and A.R.S.; writing—original draft preparation, A.R.S. and C.G.G.; writing—review and editing, C.G.G.; supervision, J.G.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We thank the anonymous referees for their useful suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Given the linear model in Equation (7), it is obtained that

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}}(\lambda) &= \begin{pmatrix} \mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j} & \mathbf{X}_{-i,-j}^t\widetilde{\mathbf{X}}_i \\ \widetilde{\mathbf{X}}_i^t\mathbf{X}_{-i,-j} & \widetilde{\mathbf{X}}_i^t\widetilde{\mathbf{X}}_i \end{pmatrix}^{-1} \cdot \begin{pmatrix} \mathbf{X}_{-i,-j}^t\mathbf{X}_j \\ \widetilde{\mathbf{X}}_i^t\mathbf{X}_j \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j} & \mathbf{X}_{-i,-j}^t\mathbf{X}_i \\ \mathbf{X}_i^t\mathbf{X}_{-i,-j} & \mathbf{X}_i^t\mathbf{X}_i + (\lambda^2+2\lambda)RSS_i^{-i} \end{pmatrix}^{-1} \cdot \begin{pmatrix} \mathbf{X}_{-i,-j}^t\mathbf{X}_j \\ \mathbf{X}_i^t\mathbf{X}_j \end{pmatrix} \\
&= \begin{pmatrix} A(\lambda) & B(\lambda) \\ B(\lambda)^t & C(\lambda) \end{pmatrix} \cdot \begin{pmatrix} \mathbf{X}_{-i,-j}^t\mathbf{X}_j \\ \mathbf{X}_i^t\mathbf{X}_j \end{pmatrix} \\
&= \begin{pmatrix} A(\lambda)\cdot\mathbf{X}_{-i,-j}^t\mathbf{X}_j + B(\lambda)\cdot\mathbf{X}_i^t\mathbf{X}_j \\ B(\lambda)^t\cdot\mathbf{X}_{-i,-j}^t\mathbf{X}_j + C(\lambda)\cdot\mathbf{X}_i^t\mathbf{X}_j \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\alpha}}_{-i,-j}(\lambda) \\ \widehat{\alpha}_i(\lambda) \end{pmatrix},
\end{aligned}$$

Since it is verified that $\mathbf{e}_i^t\mathbf{X}_{-i,-j} = \mathbf{0}$, then $\widetilde{\mathbf{X}}_i^t\mathbf{X}_{-i,-j} = (\mathbf{X}_i + \lambda\mathbf{e}_i)^t\mathbf{X}_{-i,-j} = \mathbf{X}_i^t\mathbf{X}_{-i,-j}$, where

$$\begin{aligned}
C(\lambda) &= \left(\mathbf{X}_i^t\mathbf{X}_i + (\lambda^2+2\lambda)RSS_i^{-i} - \mathbf{X}_i^t\mathbf{X}_{-i,-j}\left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1}\mathbf{X}_{-i,-j}^t\mathbf{X}_i\right)^{-1} \\
&= \left(\mathbf{X}_i^t\left(\mathbf{I} - \mathbf{X}_{-i,-j}\left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1}\mathbf{X}_{-i,-j}^t\right)\mathbf{X}_i + (\lambda^2+2\lambda)RSS_i^{-i}\right)^{-1} \\
&= \left(RSS_i^{-i,-j} + (\lambda^2+2\lambda)RSS_i^{-i}\right)^{-1}, \\
B(\lambda) &= -\left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1}\mathbf{X}_{-i,-j}^t\mathbf{X}_i\cdot C(\lambda) = \frac{RSS_i^{-i,-j}}{RSS_i^{-i,-j}+(\lambda^2+2\lambda)RSS_i^{-i}}\cdot B, \\
A(\lambda) &= \left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1} + \left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1}\mathbf{X}_{-i,-j}^t\mathbf{X}_i\cdot C(\lambda)\cdot\mathbf{X}_i^t\mathbf{X}_{-i,-j}\left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1} \\
&= \left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1} + \frac{(RSS_i^{-i,-j})^2}{RSS_i^{-i,-j}+(\lambda^2+2\lambda)RSS_i^{-i}}\cdot B\cdot B^t.
\end{aligned}$$

Then,

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}}_{-i,-j}(\lambda) &= \left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1}\mathbf{X}_{-i,-j}^t\mathbf{X}_j + \frac{(RSS_i^{-i,-j})^2}{RSS_i^{-i,-j}+(\lambda^2+2\lambda)RSS_i^{-i}}\cdot B\cdot B^t\cdot\mathbf{X}_{-i,-j}^t\mathbf{X}_j \\
&\quad + \frac{RSS_i^{-i,-j}}{RSS_i^{-i,-j}+(\lambda^2+2\lambda)RSS_i^{-i}}\cdot B\cdot\mathbf{X}_i^t\mathbf{X}_j \\
&= \left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1}\mathbf{X}_{-i,-j}^t\mathbf{X}_j \\
&\quad + \frac{RSS_i^{-i,-j}\left(RSS_i^{-i,-j}\cdot B\cdot B^t\cdot\mathbf{X}_{-i,-j}^t\mathbf{X}_j + B\cdot\mathbf{X}_i^t\mathbf{X}_j\right)}{RSS_i^{-i,-j}+(\lambda^2+2\lambda)RSS_i^{-i}}, \\
\widehat{\alpha}_i(\lambda) &= \frac{RSS_i^{-i,-j}}{RSS_i^{-i,-j}+(\lambda^2+2\lambda)RSS_i^{-i}}\cdot B^t\cdot\mathbf{X}_{-i,-j}^t\mathbf{X}_j \\
&\quad + \frac{1}{RSS_i^{-i,-j}+(\lambda^2+2\lambda)RSS_i^{-i}}\cdot\mathbf{X}_i^t\mathbf{X}_j \\
&= \frac{RSS_i^{-i,-j}}{RSS_i^{-i,-j}+(\lambda^2+2\lambda)RSS_i^{-i}}\cdot\left(B^t\cdot\mathbf{X}_{-i,-j}^t\mathbf{X}_j + (RSS_i^{-i,-j})^{-1}\mathbf{X}_i^t\mathbf{X}_j\right) \\
&= \frac{RSS_i^{-i,-j}}{RSS_i^{-i,-j}+(\lambda^2+2\lambda)RSS_i^{-i}}\cdot\widehat{\alpha}_i.
\end{aligned}$$

Appendix B

Given the linear model

$$\mathbf{X}_j = \mathbf{X}_{-j}\boldsymbol{\alpha} + \mathbf{v} = \left(\mathbf{X}_{-i,-j}\ \mathbf{X}_i\right)\begin{pmatrix} \boldsymbol{\alpha}_{-i,-j} \\ \alpha_i \end{pmatrix} + \mathbf{v},$$

it is obtained that

$$\begin{aligned}\widehat{\boldsymbol{\alpha}} &= \begin{pmatrix} \mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j} & \mathbf{X}_{-i,-j}^t\mathbf{X}_i \\ \mathbf{X}_i^t\mathbf{X}_{-i,-j} & \mathbf{X}_i^t\mathbf{X}_i \end{pmatrix}^{-1} \cdot \begin{pmatrix} \mathbf{X}_{-i,-j}^t\mathbf{X}_j \\ \mathbf{X}_i^t\mathbf{X}_j \end{pmatrix} = \begin{pmatrix} A & B \\ B^t & C \end{pmatrix} \cdot \begin{pmatrix} \mathbf{X}_{-i,-j}^t\mathbf{X}_j \\ \mathbf{X}_i^t\mathbf{X}_j \end{pmatrix} \\ &= \begin{pmatrix} A \cdot \mathbf{X}_{-i,-j}^t\mathbf{X}_j + B \cdot \mathbf{X}_i^t\mathbf{X}_j \\ B^t \cdot \mathbf{X}_{-i,-j}^t\mathbf{X}_j + C \cdot \mathbf{X}_i^t\mathbf{X}_j \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\alpha}}_{-i,-j} \\ \widehat{\alpha}_i \end{pmatrix},\end{aligned}$$

where

$$\begin{aligned} C &= \left(\mathbf{X}_i^t\mathbf{X}_i - \mathbf{X}_i^t\mathbf{X}_{-i,-j}\left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1}\mathbf{X}_{-i,-j}^t\mathbf{X}_i\right)^{-1} \\ &= \left(\mathbf{X}_i^t\left(\mathbf{I} - \mathbf{X}_{-i,-j}\left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1}\mathbf{X}_{-i,-j}^t\right)\mathbf{X}_i\right)^{-1} = \left(RSS_i^{-i,-j}\right)^{-1}, \\ B &= -\left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1}\mathbf{X}_{-i,-j}^t\mathbf{X}_i \cdot C, \\ A &= \left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1} \cdot \left(\mathbf{I} + \mathbf{X}_{-i,-j}^t\mathbf{X}_i \cdot C \cdot \mathbf{X}_i^t\mathbf{X}_{-i,-j}\left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1}\right) \\ &= \left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1} + \frac{1}{C} \cdot B \cdot B^t. \end{aligned}$$

In that case, the residual sum of squares is given by

$$\begin{aligned} RSS_j^{-j} &= \mathbf{X}_j^t\mathbf{X}_j - \begin{pmatrix} A \cdot \mathbf{X}_{-i,-j}^t\mathbf{X}_j + B \cdot \mathbf{X}_i^t\mathbf{X}_j \\ B^t \cdot \mathbf{X}_{-i,-j}^t\mathbf{X}_j + C \cdot \mathbf{X}_i^t\mathbf{X}_j \end{pmatrix}^t \begin{pmatrix} \mathbf{X}_{-i,-j}^t\mathbf{X}_j \\ \mathbf{X}_i^t\mathbf{X}_j \end{pmatrix} \\ &= \mathbf{X}_j^t\mathbf{X}_j - \mathbf{X}_j^t\mathbf{X}_{-i,-j} \cdot A^t \cdot \mathbf{X}_{-i,-j}^t\mathbf{X}_j - \mathbf{X}_j^t\mathbf{X}_i \cdot B^t \cdot \mathbf{X}_{-i,-j}^t\mathbf{X}_j - \widehat{\alpha}_i^t\mathbf{X}_i^t\mathbf{X}_j \\ &= \left(\mathbf{X}_j^t\mathbf{X}_j - \mathbf{X}_j^t\mathbf{X}_{-i,-j}\left(\mathbf{X}_{-i,-j}^t\mathbf{X}_{-i,-j}\right)^{-1}\mathbf{X}_{-i,-j}^t\mathbf{X}_j\right) \\ &\quad -RSS_i^{-i,-j}\mathbf{X}_j^t\mathbf{X}_{-i,-j} \cdot B \cdot B^t \cdot \mathbf{X}_{-i,-j}^t\mathbf{X}_j - \mathbf{X}_j^t\mathbf{X}_i \cdot B^t \cdot \mathbf{X}_{-i,-j}^t\mathbf{X}_j - \widehat{\alpha}_i^t\mathbf{X}_i^t\mathbf{X}_j \\ &= RSS_j^{-i,-j} - \left(RSS_i^{-i,-j}\mathbf{X}_j^t\mathbf{X}_{-i,-j} \cdot B \cdot B^t \cdot \mathbf{X}_{-i,-j}^t\mathbf{X}_j + \mathbf{X}_j^t\mathbf{X}_i \cdot B^t \cdot \mathbf{X}_{-i,-j}^t\mathbf{X}_j\right. \\ &\quad \left. +\widehat{\alpha}_i^t\mathbf{X}_i^t\mathbf{X}_j\right), \end{aligned}$$

and consequently

$$RSS_j^{-i,-j} - RSS_j^{-j} = RSS_i^{-i,-j}\mathbf{X}_j^t\mathbf{X}_{-i,-j} \cdot B \cdot B^t \cdot \mathbf{X}_{-i,-j}^t\mathbf{X}_j + \mathbf{X}_j^t\mathbf{X}_i \cdot B^t \cdot \mathbf{X}_{-i,-j}^t\mathbf{X}_j + \widehat{\alpha}_i^t\mathbf{X}_i^t\mathbf{X}_j.$$

Appendix C

First, parting from the expression Equation (14), it is obtained that

$$\mathbf{M}_\lambda^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 & -\frac{\lambda}{1+\lambda}\widehat{\alpha}_0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \frac{\lambda}{1+\lambda}\widehat{\alpha}_1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 & (-1)^{k-1}\frac{\lambda}{1+\lambda}\widehat{\alpha}_{k-1} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{1+\lambda} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & (-1)^{k+1}\frac{\lambda}{1+\lambda}\widehat{\alpha}_{k+1} & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & (-1)^{p}\frac{\lambda}{1+\lambda}\widehat{\alpha}_{p} & 0 & \cdots & 1 \end{pmatrix},$$

and then,

$$(\mathbf{M}_\lambda^{-1}-\mathbf{I})^t(\mathbf{M}_\lambda^{-1}-\mathbf{I})=\begin{pmatrix} 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & a(\lambda) & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{pmatrix},$$

where $a(\lambda)=\frac{\lambda^2}{(1+\lambda)^2}\cdot\left(\widehat{\alpha}_0+\widehat{\alpha}_1+\cdots+\widehat{\alpha}_{k-1}^2+1+\widehat{\alpha}_{k+1}^2+\cdots+\widehat{\alpha}_p^2\right)$. In that case,

$$\boldsymbol{\beta}^t(\mathbf{M}_\lambda^{-1}-\mathbf{I})^t(\mathbf{M}_\lambda^{-1}-\mathbf{I})\boldsymbol{\beta}=a(\lambda)\cdot\beta_k^2.$$

Second, partitioning $\widetilde{\mathbf{X}}$ in the form $\widetilde{\mathbf{X}}=\left[\mathbf{X}_{-k}\ \widetilde{\mathbf{X}}_k\right]$, it is obtained that

$$\left(\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}\right)^{-1}=\begin{pmatrix} \left(\mathbf{X}_{-k}^t\mathbf{X}_{-k}\right)^{-1}+\frac{\widehat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\alpha}}^t}{(1+\lambda)^2\cdot\mathbf{e}_k^t\mathbf{e}_k} & -\frac{\widehat{\boldsymbol{\alpha}}}{(1+\lambda)^2\cdot\mathbf{e}_k^t\mathbf{e}_k} \\ -\frac{\widehat{\boldsymbol{\alpha}}^t}{(1+\lambda)^2\cdot\mathbf{e}_k^t\mathbf{e}_k} & \frac{1}{(1+\lambda)^2\cdot\mathbf{e}_k^t\mathbf{e}_k} \end{pmatrix},$$

and then,

$$tr((\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}})^{-1})=tr\left(\left(\mathbf{X}_{-k}^t\mathbf{X}_{-k}\right)^{-1}\right)+\frac{1}{(1+\lambda)^2\cdot\mathbf{e}_k^t\mathbf{e}_k}\cdot\left(tr\left(\widehat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\alpha}}^t\right)+1\right).$$

Consequently, it is obtained that

$$\text{MSE}\left(\widehat{\boldsymbol{\beta}}(\lambda)\right)=\sigma^2tr\left(\left(\mathbf{X}_{-k}^t\mathbf{X}_{-k}\right)^{-1}\right)+\left(1+\sum_{j=0,j\neq k}^{p}\widehat{\alpha}_j^2\right)\cdot\beta_k^2\cdot\frac{\lambda^2+h}{(1+\lambda)^2},\tag{A1}$$

where $h=\frac{\sigma^2}{\beta_k^2\cdot RSS_k^{-k}}$.

Third, taking into account that the first and second derivatives of expression Equation (A1) are, respectively,

$$\begin{aligned}\frac{\partial}{\partial\lambda}\text{MSE}\left(\widehat{\boldsymbol{\beta}}(\lambda)\right) &= \left(1+\sum_{j=0,j\neq k}^{p}\widehat{\alpha}_j^2\right)\cdot\beta_k^2\cdot\frac{2(\lambda-h)}{(1+\lambda)^3},\\ \frac{\partial^2}{\partial\lambda^2}\text{MSE}\left(\widehat{\boldsymbol{\beta}}(\lambda)\right) &= -2\left(1+\sum_{j=0,j\neq k}^{p}\widehat{\alpha}_j^2\right)\cdot\beta_k^2\cdot\frac{2\lambda-(1+3h)}{(1+\lambda)^4}.\end{aligned}$$

Since $\lambda\geq 0$, it is obtained that $\text{MSE}\left(\widehat{\boldsymbol{\beta}}(\lambda)\right)$ is decreasing if $\lambda<h$ and increasing if $\lambda>h$, and it is concave if $\lambda>\frac{1+3h}{2}$ and convex if $\lambda<\frac{1+3h}{2}$.

Indeed, given that

$$\begin{aligned}\lim_{\lambda\to+\infty}\text{MSE}\left(\widehat{\boldsymbol{\beta}}(\lambda)\right) &= \sigma^2tr\left(\left(\mathbf{X}_{-k}^t\mathbf{X}_{-k}\right)^{-1}\right)+\left(1+\sum_{j=0,j\neq k}^{p}\widehat{\alpha}_j^2\right)\cdot\beta_k^2,\\ \text{MSE}\left(\widehat{\boldsymbol{\beta}}(0)\right) &= \sigma^2tr\left(\left(\mathbf{X}_{-k}^t\mathbf{X}_{-k}\right)^{-1}\right)+\left(1+\sum_{j=0,j\neq k}^{p}\widehat{\alpha}_j^2\right)\cdot\beta_k^2\cdot h,\end{aligned}\tag{A2}$$

if $h > 1$, then $\text{MSE}\left(\widehat{\boldsymbol{\beta}}(0)\right) > \lim_{\lambda\to+\infty}\text{MSE}\left(\widehat{\boldsymbol{\beta}}(\lambda)\right)$, and if $h < 1$, then $\text{MSE}\left(\widehat{\boldsymbol{\beta}}(0)\right) < \lim_{\lambda\to+\infty}\text{MSE}\left(\widehat{\boldsymbol{\beta}}(\lambda)\right)$. That is to say, if $h > 1$, then the raise estimator presents always a lower MSE than the one obtained by OLS for all λ, and comparing expressions Equations (A1) and (A2) when $h < 1$, $\text{MSE}\left(\widehat{\boldsymbol{\beta}}(\lambda)\right) \leq \text{MSE}\left(\widehat{\boldsymbol{\beta}}(0)\right)$ if $\lambda \leq \frac{2\cdot h}{1-h}$.

From this information, the behavior of the MSE is represented in Figures A1 and A2. Note that the MSE presents a minimum value for $\lambda = h$.

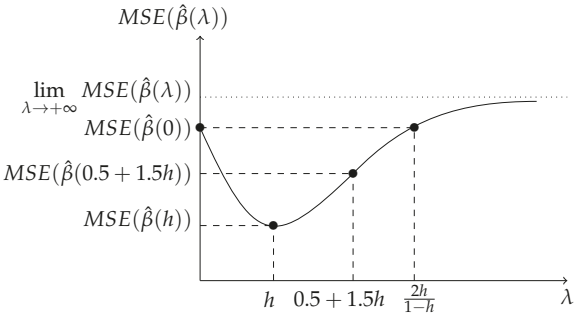


Figure A1. $MSE(\hat{\beta}(\lambda))$ representation for $h = \frac{\sigma^2}{(e_k^t e_k)\cdot\beta_k^2} < 1$.

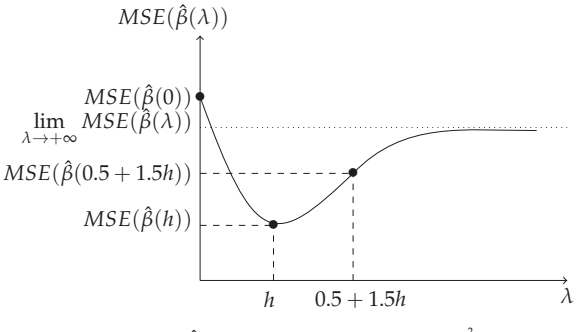


Figure A2. $MSE(\hat{\beta}(\lambda))$ representation for $h = \frac{\sigma^2}{(e_k^t e_k)\cdot\beta_k^2} > 1$.

References

1. Kiers, H.; Smilde, A. A comparison of various methods for multivariate regression with highly collinear variables. *Stat. Methods Appl.* **2007**, *16*, 193–228. [CrossRef]
2. Frank, L.E.; Friedman, J.H. A statistical view of some chemometrics regression tools. *Technometrics* **1993**, *35*, 109–135. [CrossRef]
3. Fu, W.J. Penalized regressions: the bridge versus the lasso. *J. Comput. Graph. Stat.* **1998**, *7*, 397–416.
4. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
5. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
6. Donoho, D.L.; Johnstone, I.M. Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.* **1995**, *90*, 1200–1224. [CrossRef]
7. Klinger, A. Inference in high dimensional generalized linear models based on soft thresholding. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2001**, *63*, 377–392. [CrossRef]
8. Dupuis, D.; Victoria-Feser, M. Robust VIF regression with application to variable selection in large data sets. *Ann. Appl. Stat.* **2013**, *7*, 319–341. [CrossRef]

9. Li, Y.; Yang, H. A new Liu-type estimator in linear regression model. *Stat. Pap.* **2012**, *53*, 427–437. [CrossRef]
10. Liu, Y.; Wang, Y.; Feng, Y.; Wall, M. Variable selection and prediction with incomplete high-dimensional data. *Ann. Appl. Stat.* **2016**, *10*, 418–450. [CrossRef]
11. Uematsu, Y.; Tanaka, S. High-dimensional macroeconomic forecasting and variable selection via penalized regression. *Econom. J.* **2019**, *22*, 34–56. [CrossRef]
12. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, *67*, 301–320. [CrossRef]
13. Tutz, G.; Ulbricht, J. Penalized regression with correlation-based penalty. *Stat. Comput.* **2009**, *19*, 239–253. [CrossRef]
14. Stone, M.; Brooks, R.J. Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *J. R. Stat. Soc. Ser. B (Methodol.)* **1990**, *52*, 237–269. [CrossRef]
15. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, *32*, 407–499.
16. Golan, A.; Judge, G.; Miller, D. *Maximum Entropy Econometrics: Robust Estimation With Limited Data*; John Wiley and Sons: Chichester, UK, 1997.
17. Golan, A. Information and entropy econometrics review and synthesis. *Found. Trends Econom.* **2008**, *2*, 1–145. [CrossRef]
18. Macedo, P. Ridge Regression and Generalized Maximum Entropy: An improved version of the Ridge–GME parameter estimator. *Commun. Stat.-Simul. Comput.* **2017**, *46*, 3527–3539. [CrossRef]
19. Batah, F.S.M.; Özkale, M.R.; Gore, S. Combining unbiased ridge and principal component regression estimators. *Commun. Stat. Theory Methods* **2009**, *38*, 2201–2209. [CrossRef]
20. Massy, W.F. Principal components regression in exploratory statistical research. *J. Am. Stat. Assoc.* **1965**, *60*, 234–256. [CrossRef]
21. Guo, W.; Liu, X.; Zhang, S. The principal correlation components estimator and its optimality. *Stat. Pap.* **2016**, *57*, 755–779. [CrossRef]
22. Aguilera-Morillo, M.; Aguilera, A.; Escabias, M.; Valderrama, M. Penalized spline approaches for functional logit regression. *Test* **2013**, *22*, 251–277. [CrossRef]
23. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. [CrossRef]
24. De Jong, S. SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251–263. [CrossRef]
25. Jensen, D.; Ramirez, D. Surrogate models in ill-conditioned systems. *J. Stat. Plan. Inference* **2010**, *140*, 2069–2077. [CrossRef]
26. García, J.; Salmerón, R.; García, C.; López Martín, M.D.M. Standardization of variables and collinearity diagnostic in ridge regression. *Int. Stat. Rev.* **2016**, *84*, 245–266. [CrossRef]
27. Marquardt, D. You should standardize the predictor variables in your regression models. Discussion of: A critique of some ridge regression methods. *J. Am. Stat. Assoc.* **1980**, *75*, 87–91.
28. Rodríguez, A.; Salmerón, R.; García, C. The coefficient of determination in the ridge regression. *Commun. Stat. Simul. Comput.* **2019**. [CrossRef]
29. García, C.G.; Pérez, J.G.; Liria, J.S. The raise method. An alternative procedure to estimate the parameters in presence of collinearity. *Qual. Quant.* **2011**, *45*, 403–423. [CrossRef]
30. Salmerón, R.; García, C.; García, J.; López, M.D.M. The raise estimator estimation, inference, and properties. *Commun. Stat. Theory Methods* **2017**, *46*, 6446–6462. [CrossRef]
31. García, J.; López-Martín, M.; García, C.; Salmerón, R. A geometrical interpretation of collinearity: A natural way to justify ridge regression and its anomalies. *Int. Stat. Rev.* **2020**. [CrossRef]
32. Belsley, D.A.; Kuh, E.; Welsch, R.E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; John Wiley & Sons: Hoboken, NJ, USA, 2005; Volume 571.
33. Hoerl, A.; Kannard, R.; Baldwin, K. Ridge regression: some simulations. *Commun. Stat. Theory Methods* **1975**, *4*, 105–123. [CrossRef]
34. Stein, C. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*; University of California Press: Berkeley, CA, USA, 1956; pp. 197–206.

35. James, W.; Stein, C. Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*; University of California Press: Berkeley, CA, USA, 1961; pp. 361–379.
36. Ohtani, K. An MSE comparison of the restricted Stein-rule and minimum mean squared error estimators in regression. *Test* **1998**, *7*, 361–376. [CrossRef]
37. Hubert, M.; Gijbels, I.; Vanpaemel, D. Reducing the mean squared error of quantile-based estimators by smoothing. *Test* **2013**, *22*, 448–465. [CrossRef]
38. Salmerón, R.; García, C.; García, J. Variance Inflation Factor and Condition Number in multiple linear regression. *J. Stat. Comput. Simul.* **2018**, *88*, 2365–2384. [CrossRef]
39. Salmerón, R.; Rodríguez, A.; García, C. Diagnosis and quantification of the non-essential collinearity. *Comput. Stat.* **2019**. [CrossRef]
40. Marquandt, D.; Snee, R. Ridge regression in practice. *Am. Stat.* **1975**, *29*, 3–20.
41. García, C.B.; Garcí, J.; Salmerón, R.; López, M.M. Raise regression: Selection of the raise parameter. In Proceedings of the International Conference on Data Mining, Vancouver, BC, Canada, 30 April–2 May 2015.
42. García, J.; Ramírez, D. The successive raising estimator and its relation with the ridge estimator. *Commun. Stat. Simul. Comput.* **2016**, *46*, 11123–11142. [CrossRef]
43. Klein, L.; Goldberger, A. *An Economic Model of the United States, 1929–1952*; North Holland Publishing Company: Amsterdan, The Netherlands, 1964.
44. García, C.; Salmerón, R.; García, C.; García, J. Residualization: Justification, properties and application. *J. Appl. Stat.* **2019**. [CrossRef]
45. Marquardt, D. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics* **1970**, *12*, 591–612. [CrossRef]
46. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.
47. York, R. Residualization is not the answer: Rethinking how to address multicollinearity. *Soc. Sci. Res.* **2012**, *41*, 1379–1386. [CrossRef]
48. Salmerón, R.; García, J.; García, C.; García, C. Treatment of collinearity through orthogonal regression: An economic application. *Boletín Estadística Investig. Oper.* **2016**, *32*, 184–202.

mathematics

MDPI

Article

Discounted and Expected Utility from the Probability and Time Trade-Off Model

Salvador Cruz Rambaud *,† **and Ana María Sánchez Pérez** †

Departamento de Economía y Empresa, Universidad de Almería, La Cañada de San Urbano, s/n, 04120 Almería, Spain; amsanchez@ual.es

* Correspondence: scruz@ual.es; Tel.: +34-950-015-184

† La Cañada de San Urbano, s/n, 04120 Almería, Spain.

Received: 12 March 2020; Accepted: 13 April 2020; Published: 15 April 2020

Abstract: This paper shows the interaction between probabilistic and delayed rewards. In decision-making processes, the Expected Utility (EU) model has been employed to assess risky choices whereas the Discounted Utility (DU) model has been applied to intertemporal choices. Despite both models being different, they are based on the same theoretical principle: the rewards are assessed by taking into account the sum of their utilities and some similar anomalies have been revealed in both models. The aim of this paper is to characterize and consider particular cases of the Time Trade-Off (PPT) model and show that they correspond to the EU and DU models. Additionally, we will try to build a PTT model starting from a discounted and an expected utility model able to overcome the limitations pointed out by Baucells and Heukamp.

Keywords: risk; delay; decision-making process; probability; discount

JEL Classification: G12; D81; D9

1. Introduction

The main objective of this paper is to present the Probability and Time Trade-Off Model [1] as an accurate framework where risk and intertemporal decisions can be separately considered.

The methodology used in this paper consists of considering some particular cases of the PTT model and show that they correspond to EU and DU models. Moreover, the possibility of reversing the process is provided, i.e., obtaining a PTT model starting from an EU and a DU model.

The decision-making process is relatively simple when the alternatives differ only in one dimension (e.g., the amount, the probability of occurrence or the delay) while the rest of the variables remain constant [2]. However, decision-making problems frequently involve alternatives which differ in more than one dimension [3]. Along these lines, there are two traditional models to assess choices which differ in risk or time, in addition to the amount of their reward. Despite risk and delay initially appearing quite different, the individual behavior when facing risky and delayed outcomes is analogous [4]. Currently, due to the fact that most real-world decisions are made on alternatives which are both uncertain and delayed [4], there is a growing interest in understanding and modelling how risk and delay interact in the individual behavior. In this way, there are some scholars as Luckman [5] who show the attempts to explain the complex individual behaviors when facing alternatives which differ in more than two dimensions.

The decision-making process has been analyzed in psychophysics, neuroscience and social and behavioral sciences, such as economics or psychology. From an academic point of view, one of the most studied processes in decision-making is intertemporal choice which concerns those alternatives which differ in maturities (the choice between a smaller, sooner outcome, and a larger, later one) [3]. On the other hand, the decision-making under uncertainty involves alternatives whose rewards differ

in relation to the probability of being received (the choice between "a smaller reward, to be received with greater probability, and a larger one, but less likely") [3].

Both kinds of decisions have been traditionally analyzed by using two main systems of calculation: the Discounted Utility model and the Expected Utility theory which describe the present value of delayed rewards and the actuarial value of risky rewards, respectively. Both models are simple, widely accepted, and with a similar structure, since they use the same theoretical principle: the rewards are assessed by the sum of their utilities [6,7]. In the decision-making process, individuals choose the alternative which maximizes their current utility value (denoted by U_0).

As seen in Table 1, DU and EU are the basic models employed in the decision-making process which represent the rational choice over time and under risk, respectively.

Table 1. Classical models to obtain the utility value. Source: Own elaboration.

	Discounted Utility (DU)	**Expected Utility (EU)**
Pioneer work	Samuelson [8]	Von Neumann and Morgenstern [9]
Result	The present value of delayed rewards	The expected value of risky rewards
Formula	$U_0 = \sum_{t=0}^{T} \delta^t u_t$	$U_0 = \sum_{k=0}^{n} p_k u_k$
Parameters	u_t: utility from the reward at t t: reward maturity ($t = 0, 1, \ldots, T$) δ: discount factor ($0 < \delta < 1$)	u_k: utility from the k-th reward p_k: probability of the k-th reward ($\sum_{k=0}^{n} p_k = 1$)

On the other hand, the Discounted Expected Utility (DEU) model (see Table 2) is employed in decision environments involving both intertemporal and risky decisions (Schoemaker [10] points out nine variants of the Expected Utility model). DEU model has been deeply analyzed in several recent works by Coble and Lusk [11] and Andreoni and Sprenger [12].

Table 2. DEU model. Source: Own elaboration.

Pioneer work	Jamison [13]
Result	The value of risky and delayed rewards
Formula	$V(c_0, \ldots, c_r) = \sum_{t=0}^{T} D(t) u(c_t)$
Parameters	V: valuation of consumption in different periods t: reward maturity ($t = 0, 1, \ldots, T$) $D(t)$: discount function $u(c_t)$: utility of consumption at t (c_t)

The accuracy of DU and EU models has been questioned to explain actual behaviors given that they are a simplification of reality [14]. In the same way, [15,16] show that DEU model fails as a predictor of intertemporal-risky choices. From an experimental point of view, Coble and Lusk [11] prove that the data does not support the DEU model assumption of a unique parameter that explain risk and time preferences. Usually, individuals' preferences cannot be so easily determined and then several anomalies of these models must be taken into account when a real decision is analyzed. Some of these effects or anomalies are the consequence of psychophysical properties of time, probability and pay-out dimensions [17].

Indeed, several studies, such as [6,18], compare the analogies and the anomalies present in DU (intertemporal choices) and EU (risky choices) models, by observing that some risky choice inconsistencies are parallel to delay choice inconsistencies. In this line, it should be stressed the contribution by Prelec and Loewenstein [19] where a one-to-one correspondence is shown between the

behavioral anomalies of Expected and Discounted Utility models. Nevertheless, Green, Myerson and Ostaszewski [20] findings suggest that processes involving delayed and probabilistic rewards, despite being similar, are not identical.

The main contribution of this paper is the demonstration of that both the Expected Utility (EU) and the Discounted Utility (DU) models can be embedded in the Probability and Time Trade-Off (PTT) model. Moreover, preliminary thoughts are provided on the possibility of reversing the process, i.e., obtaining a PTT starting from an EU or a DU model.

The relevance of this contribution is that the existence of the proved equivalence can be used to explain and relate behavioral inconsistencies in real choices. In effect, most existing literature presents the anomalies in intertemporal and probabilistic choices as separate inconsistencies [21] and, indeed, this paper can be used to clarify the equivalence between certain anomalies analyzed in the context of the DU and the EU models. A precedent can be found in the paper by Cruz Rambaud and Sánchez Pérez about the so-called peanuts effects [7]. This methodology could be applied to other anomalies present in both aforementioned models.

This paper has been organized as follows. After this Introduction, Section 2 presents an extensive revision of the existing literature on this topic, divided into three subsections to facilitate its reading. Section 3 has been structured into four subsections. Section 3.1 provides some basic definitions, properties and examples related to reward choices where both time and probability are involved. Sections 3.2 and 3.3 are devoted to obtaining the EU and DU from the PTT model, respectively. Section 3.4 is an essay to generate PTT starting from DU and EU models, and represents the framework in which further research is needed. In Section 4, the main obtained results are discussed regarding other works which aim to relate the properties of decisions under risk and under time. Finally, Section 5 summarizes and concludes.

2. Literature Review

Most of the previous research analyzing delay and risk parameters do so separately. Only a few papers study the individual preferences in decisions involving time-delayed and risky rewards. Initially, Prelec and Loewenstein [19] study uncertainty and delay from a common approach. In this sense, Weber and Chapman [4] and Gollier [22] analyze the relationship between choice under risk and time from a global perspective.

There have been several attempts to introduce a new model which explains individual preferences in decisions involving time-delayed and risky rewards [5,21]. Specifically, in some of these models, delayed rewards are studied from the perspective of the subjective probability [23,24].

These descriptive psychophysics models need to be empirically implemented. However, from a practical point of view, few publications, such as [25,26], have analyzed the risk and discount attitudes simultaneously. Specifically, some of them [11,27] explain the relationship between the risk and time in the decision-making process from an experimental point of view. From this perspective, it is necessary to highlight that uncertainty underlying time delays may be undermined when data is obtained through laboratory experiments [28].

Despite risk and delay being perceived as concepts psychologically distinct, in some contexts, decision-makers may identify both concepts as psychologically interchangeable since they influence preferences [17]. This is because behavioral patterns facing risk and delay are based on a common underlying dimension [17,23]. Even though time delay and uncertainty are interrelated, their interaction is controversial. Below, we can find some examples of studies that have analyzed the relation between time and risk preferences to assess delayed and probabilistic choices. In Section 2.1, uncertainty is treated as the fundamental concept and delays are transformed into risky terms. Next, in Section 2.2, by assuming that uncertainty is linked to delayed rewards, the reward probability is expressed as additional waiting time. Finally, in Section 2.3, the new models to assess risky intertemporal choices are pointed out.

2.1. Uncertainty as Central Concept. Transformation of Delays into Probability Terms

Keren and Roelofsma [28] defend that uncertainty is the psychologically central concept, given that decisions are affected by delay only if this delay entails uncertainty. They propose that the delay effect is actually the effect of the uncertainty inherent to delay. Indeed, they defend that the immediacy effect and the certainty effect are the same effects given that decision-makers are based on the implicit uncertainty, not the delay itself. In intertemporal choice, the immediate outcome entails no uncertainty while the delayed outcome is perceived as uncertain. In this way, "if delay implies risk, then risk should have the same effects as delay—the two are interchangeable" [4].

In this sense, Rachlin [23] translates reward delay into the probability of receiving it. Delays act like less-than-unit probabilities; longer delays correspond to lower probabilities, given that uncertainty increases as delay increases. In this way, intertemporal discount models may be translated into probabilistic discount functions (called "odds against" Θ, i.e., the average number of trials until a win). They are calculated as follows:

$$\Theta = \frac{1}{p} - 1,$$

where p is the probability of receiving an uncertain outcome.

2.2. Uncertainty is Inherent to Intertemporal Choice. Interpretation of Reward Probability as Waiting Time

When applying the DU model, sometimes future outcomes are modelled as though non-stochastic by ignoring its uncertainty. However, as stated by Fisher [29], "future income is always subject to some uncertainty, and this uncertainty must naturally have an influence on the rate of time preference, or degree of impatience, of its possessor". In this vein, some recent studies consider the relation between behavior under risk and over time with the premise that uncertainty is inherent to intertemporal choices, given that any event may interfere in the process of acquiring the reward between the current and the promised date [30].

This implies that the "decision-maker's valuation of delayed outcomes not only depends on her pure time preference, i.e., her preference for immediate utility over delayed utility, but also on her perception of the uncertainty and, consequently, on her risk preferences" [31]. Under the assumption that only present consumption is certain while any future consumption may be considered to be uncertain, risk preferences could influence intertemporal choice patterns. In this way, Takahashi [32] studies the aversion to subjective uncertainty associated with delay.

Soares dos Santos et al. [3] propose a generalized function for the probabilistic discount process by using time preferences. Specifically, probabilistic rewards are transformed into delayed rewards as follows: instead of using the probability of occurrence, they use the mean waiting time before a successful draw of the corresponding reward.

Probabilities are converted into comparable delays according to the constant of proportionality by [23] and through the examination of the indifference points of hypothetical rewards which are both delayed and risky. Once the probabilities are transformed into delays, this delay is added to the explicit delay, being this delay/probability combination the total delay used to assess delayed and risky rewards. The similar structure of time and risk shows that if the risk is interpreted as waiting time, both magnitudes may be combined into a single metric which is consistent with the hyperbolic discount function (better than exponential function given that as delay increases, there is a hyperbolic decay of probabilities of obtaining it). Furthermore, this metric may explain some of the observed behavior in choice under risk, such as the certainty effect [33].

2.3. Models to Assess Risky Intertemporal Choices

Traditionally, risky intertemporal rewards have been evaluated by applying the EU and DU models separately. First, risky rewards are assessed by the EU model, and then their expected value at maturity is discounted by using a constant discount rate. More flexibility in the assessment of risky

delayed rewards is requested from a descriptive point of view by [34]. This flexibility is necessary given that the consequences of delayed rewards do not only affect the present utility but also the future utility. In the same way, the probability implementation red has an influence on the discount rates [35].

Luckman [5] has shown that there are three specific unifying models to deal with delayed and probabilistic choices: the Probability and Time Trade-off model [36], the Multiplicative Hyperboloid Discounting model [37], and the Hyperbolic Discounting model [38]. These three models have a common feature as they consider two special risky intertemporal choices: pure risky choices and pure intertemporal choices. These special risky intertemporal choices are consistent with the results from traditional models.

In the assessment of delayed but certain rewards, the DU model is employed. However, for those alternatives whose maturity is unknown or uncertain at the beginning, i.e., when the decision-maker does not know the exact realization time of the future outcome, the DEU model may be implemented [30]. Specifically, when there are different possible delays for a reward and their probabilities are known, its "timing risk" [16] is identified. Meanwhile, the term "timing uncertainty" is employed to denote those outcomes which possible delays are vaguely known or unknown [30]. Time lotteries which pay a specific prize at uncertain future dates are a clear example of this kind of rewards [39]. Onay and Öncüler [16] and Coble and Lusk [11] demonstrate that DEU model is not accurate enough to forecast intertemporal choice behavior under timing risk.

In 2012, Han and Takahashi [40] proposed that psychophysical time commonly explains anomalies in decision both over time and under risk. Moreover, they introduce the non-linear psychophysical time into the time discount function according to the Weber-Fechner law. Green and Myerson [2] show that a single process to assess risky and intertemporal operations is inconsistent. Apart from the fact that risk and time are not equivalent parameters, "the interaction between them is complex and not easily understood" [17]. Nevertheless, recent studies show that it is necessary to introduce a common framework to understand people's perception of risk and delay when making decisions [41]. In this way, there is still room to improve the methodology and results.

Summarizing the introduction and the literature review, Figure 1 shows the existing methodologies to assess the delayed and uncertain rewards which differ in one, two or more dimensions.

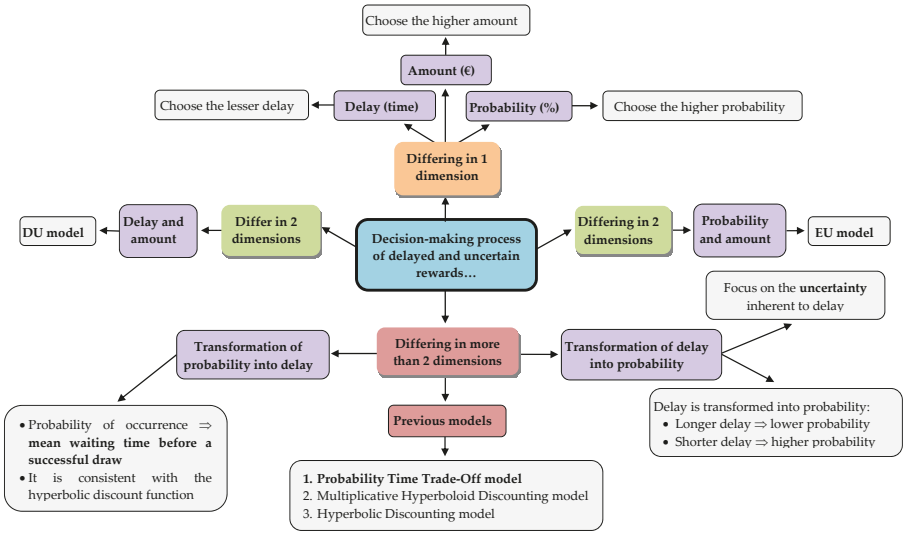


Figure 1. Decision-making process of delayed and uncertain rewards. Source: Own elaboration.

3. Deriving DU and EU from the Probability and Time Trade-Off Model

3.1. Introduction

It will prove useful to begin with the following definition in the ambit of the Probability and Time Trade-off (PTT) model [1].

Definition 1. *Let $\mathcal{M}$ be the set $X \times P \times T$, where $X = [0, +\infty)$, $P = [0, 1]$, and $T = [0, +\infty]$. A discount function in the context of the PTT model is a continuous real-valued map $V(x, p, t)$, defined on $\mathcal{M}$, which is strictly increasing with respect to the first and second components, and strictly decreasing according to the third. Moreover, it satisfies that $V(x, 1, 0) = x$, for every $x \in X$.*

Example 1. *$V(x, p, t) = (xp + 1)^{\exp\{-kt\}} - 1$, $k > 0$, is a discount function in the context of the PTT model.*

Definition 1 guarantees that $V(x, p, t) \leq x$. In effect, $x = V(x, 1, 0) \geq V(x, p, 0) \geq V(x, p, t)$. The triple (x, p, t) denotes the prospect of receiving a reward x at time t with probability p. Obviously,

1. If $p = 1$, then the concept of a discount function *in the context of the DU model* arises: $F(x, t) := V(x, 1, t)$ is a continuous real-valued map defined on $X \times T$, which is strictly increasing in the first component, strictly decreasing in the second component, and satisfies $F(x, 0) = x$, for every $x \in X$.
2. If $t = t_0$, then the concept of a discount function *in the context of the EU model* arises: $V(x, p) := V(x, p, t_0)$ is a continuous real-valued map defined on $X \times P$, which is strictly increasing in the two components, and satisfies $V(x, 1) = x$, for every $x \in X$.

Definition 2. *Given a discount function $V(x, p, t)$ in the context of the PTT model, the domain of V is the maximum subset, $\mathcal{D}$, of $\mathcal{M}$ where V satisfies all conditions of Definition 1.*

Example 2. *The domain of the discount function $V(x, p, t) = (xp + 1)^{\exp\{-kt\}} - 1$, $k > 0$, is $\mathcal{D} = \mathcal{D}$. On the other hand, the domain of the discount function*

$$V(x, p, t) = \frac{xp + it}{1 + jt},$$

where $i > 0$ and $j > 0$, is:

$$\mathcal{D} = \left\{ (x, p) \in X \times P : xp > \frac{i}{j} \right\} \times T.$$

Baucells and Heukamp [1] require that V converges to zero when xpe^{-t} converges to zero. However, in the present paper this restriction will be removed, and we will allow that V tends to zero only when $x \to 0$, or $p \to 0$:

$$\lim_{t \to +\infty} V(x, p, t) := L(x, p) > 0,$$

provided that $T \subseteq \text{proy}_3(\mathcal{D})$.

Definition 3. *A discount function in the context of the PTT model, V, is said to be regular if $L(x, p) = 0$, while V is said to be singular if $L(x, p) > 0$.*

Example 3. *The discount function $V(x, p, t) = (xp + 1)^{\exp\{-kt\}} - 1$ is regular since*

$$L(x, p) = (xp + 1)^0 - 1 = 0.$$

On the other hand, the discount functions $V(x,p,t) = \frac{xp+it}{1+jt}$ *and* $V(x,p,t) = \frac{xp+x^2p^2it}{1+jt}$, *where* $i > 0$ *and* $j > 0$, *are singular as*

$$L(x,p) = \frac{i}{j} > 0$$

and

$$L(x,p) = \frac{x^2p^2i}{j} > 0,$$

respectively.

Thus, the paper by Baucells and Heukamp [1] implicitly assumes the regularity of V. However, in this work, this requirement will not necessarily hold.

Now, we are going to provide an interpretation of probability in the context of a prospect (x,p,t). In effect, let $V(x,p,t)$ be a discount function in the context of the PTT model and consider a given value, x, of the amount. Therefore, $V_x(p,t) := V(x,p,t)$ is now a two-variable function. Consider the indifference line given by $V_x(p,t) = k$ $(0 < k \leq x)$ (observe that $k = V_x(p_0,0)$, for some p_0), which eventually can give rise to an explicit function, denoted by $p_{k,x}(t)$ (see Figure 2). Obviously, $p_x(t)$ is an increasing function which has 1 as upper bound and 0 as lower bound. Moreover, $p_{k,x}(0) = p_0 = k$. Let

$$t_{k,x}^{\min} = \min\{t : V_x(p,t) = k, \text{ for some } p\}$$

and

$$t_{k,x}^{\max} = \max\{t : V_x(p,t) = k, \text{ for some } p\}.$$

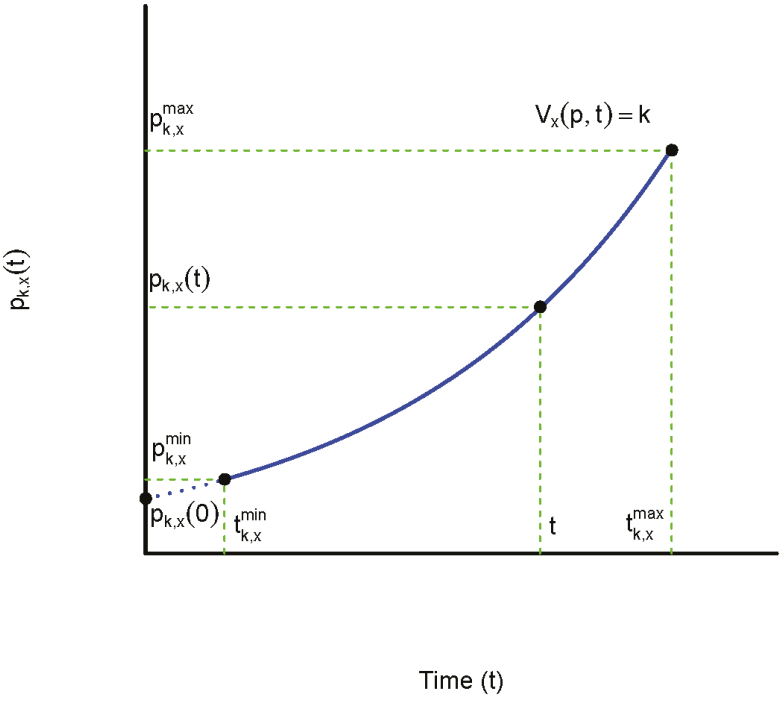


Figure 2. Indifference line $V_x(p,t) = k$ $(0 < k \leq x)$. Source: Own elaboration.

Lemma 1. *The following statements hold:*

(i) $t_{k,x}^{\min}$ *is the solution of* $V_x(0,t) = k$ *if, and only if,* $0 \in proy_2(\mathcal{D})$.
(ii) $t_{k,x}^{\max}$ *is the solution of* $V_x(1,t) = k$ *if, and only if,* $1 \in proy_2(\mathcal{D})$.

Proof. Take into account that $V_x(0,t)$ and $V_x(1,t)$ are lower and upper bounded functions in time. This guarantees the existence of both a minimum and a maximum time. □

Proposition 1. *If* V *is regular or* V *is singular and* $0 < L(x,p) < k$, *then* $1 \in proy_2(\mathcal{D})$.

Proof. In effect, by the definition of a discount function in the context of the PTT model, one has $V_x(1,0) = x$. Moreover, if V is regular or V is singular and $0 < L(x,p) < k$, the following chain of inequalities holds:

$$0 \leq L(x,1) < k < x.$$

Therefore, as V is continuous and decreasing with respect to time, by the Intermediate Value Theorem, there exists t_1, such that $V_x(1,t_1) = k$. Consequently, $1 \in \text{proy}_2(\mathcal{D})$. □

Example 4. *Let us consider the following discount function in the context of the PTT model:*

$$V(x,p,t) = \frac{xp}{1+ixpt},$$

where $i > 0$. *If* $V(x,p,t) = k$ $(0 < k \leq x)$, *then*

$$p_{k,x}(t) = \frac{k}{x - ikxt} = \frac{p_{k,x}(0)}{1 - ip_{k,x}(0)xt},$$

which is obviously increasing with respect to t. *The minimum value of* t *is* $t_{k,x}^{\min} = 0$, *and the maximum value of* t *is*

$$t_{k,x}^{\max} = \frac{x-k}{ikx} = \frac{1 - p_{k,x}(0)}{ip_{k,x}(0)x}.$$

The corresponding minimum and maximum values of p *are* $p_{k,x}^{\min} = p_{k,x}(0) = k$ *and* $p_{k,x}^{\max} = 1$, *respectively.*

Finally, the map of indifference curves $V_x(p,t) = k$, $0 < k \leq x$, gives rise to a family of distribution functions, denoted by $p_{k,x}(t)$, corresponding to a stochastic process, satisfying:

$$V(x, p_{k,x}(t), t) = k.$$

More, specifically, Figure 2 represents all couples (t,p) such that $(x,p,t) \in X \times P \times T$, for given values of x and $0 < k \leq x$. Moreover, it shows that there exists a functional relationship between p and t, given by $p_{k,x}(t)$, with $0 < k \leq x$. Indeed, $p_{k,x}(t)$ is a distribution function. In effect, let us consider the random variable $T_{k,x}$: "Time period in which the reward x can be delivered, at a discounting level k". In other words, each random variable $T_{k,x}$ is a stopping time depending on k and x. Specifically, $p_{k,x}(t)$ is the distribution function of $T_{k,x}$:

$$p_{k,x}(t) = \Pr(T_{k,x} \leq t).$$

Consequently, starting from X, a continuous stochastic process has been obtained for every $x \in X$, giving rise to the following family of stochastic processes:

$$\left\{ \{T_{k,x}\}_{0<k\leq x} \right\}_{x \in X}.$$

Assume that $x_1 < x_2$. Then $T_{k,x_1} < T_{k,x_2}$, $0 < k \leq x_1 < x_2$, i.e., the stochastic process $\{T_{k,x}\}_{x\in X}$ is increasing with respect to x. Moreover, if $0 < k_1 < k_2 \leq x$, then $T_{k_1,x} < T_{k_2,x}$ and so $\{T_{k,x}\}_{0<k\leq x}$ is increasing with respect to k. In this case, we can define a discount function where time is stochastic in the following way.

Definition 4. *A discount function with stochastic time is a real function*

$$F : \left\{\{(x, T_{k,x})\}_{0<k\leq x}\right\}_{x\in X} \longrightarrow \mathbb{R}$$

such that

$$(x, T_{k,x}) \mapsto F(x, T_{k,x}),$$

defined by:

$$F(x, T_{k,x}) = V(x, p_0, t_0),$$

such that $p_{k,x}(t_0) = p_0$.

Obviously, function F is well defined. Therefore, we can state the following theorem.

Theorem 1. *A discount function $V(x, p, t)$ is equivalent to the discount function with stochastic time $F(x, T_{k,x})$, where $T_{k,x}$ is the random variable: "Time period in which the reward x can be delivered, at a discounting level k".*

Remark 1. *Observe that for every $x \in X$, the random variables $T_{k,x}$ could be indexed, instead of by k, by another parameter in one-to-one correspondence with k. For example, for a given value p of probability, there is a biunivocal correspondence between k-values and calendar times. Thus, by denoting the calendar time as τ, the random variable $T_{k,x}$ becomes $T_{\tau,x}$, in whose case we will say that the discount function of Definition 4 is time-dependent (in the particular case in which time is age, the discount function is said to be age-dependent [42]).*

In the same way, for a given value pt of time, there is a bijective correspondence between k-values and probabilities. Thus, by denoting the probability as q, the random variable $T_{k,x}$ becomes $T_{q,x}$, in whose case we will say that the discount function of Definition 4 is risk-dependent.

On this question, we will return later in Section 3.3. Observe that now there is an extra force of discount given by the probability p. This statement can be shown in the following proposition.

Proposition 2. *Assume that V is differentiable. Then, the instantaneous discount rate of V, denoted by δ_V, is greater than the instantaneous discount, δ_F, of the discount function in the context of the DU model, $F(x, t) := V(x, p, t)$, where p is constant.*

Proof. In effect, the derivative of the implicit function $V(x, p_{k,x}(t), t) = k$ $(0 < k \leq x)$, results in:

$$\mathrm{d}V = \frac{\partial V}{\partial x}\mathrm{d}x + \frac{\partial V}{\partial p}\frac{\partial p_{k,x}(t)}{\partial x}\mathrm{d}x + \frac{\partial V}{\partial p}\frac{\partial p_{k,x}(t)}{\partial t}\mathrm{d}t + \frac{\partial V}{\partial t}\mathrm{d}t = 0,$$

from where the instantaneous discount rate of V at (x, p, t) is (see [43]):

$$\delta_V(x, p, t) := \frac{\mathrm{d}x}{x\mathrm{d}t} = -\frac{1}{x}\frac{\frac{\partial V}{\partial t} + \frac{\partial V}{\partial p}\frac{\partial p_{k,x}(t)}{\partial t}}{\frac{\partial V}{\partial x} + \frac{\partial V}{\partial p}\frac{\partial p_{k,x}(t)}{\partial x}}.$$

As

$$\frac{\partial V}{\partial p}\frac{\partial p_{k,x}(t)}{\partial t} > 0$$

and

$$\frac{\partial V}{\partial p}\frac{\partial p_{k,x}(t)}{\partial x} < 0,$$

then, for every $(x, p, t) \in \mathcal{M}$,

$$\delta_V(x, p, t) > -\frac{1}{x}\frac{\frac{\partial V}{\partial t}}{\frac{\partial V}{\partial x}},$$

which is the instantaneous discount rate when p is constant. Therefore,

$$\delta_V(x, p, t) > \delta_F(x, t),$$

as expected. □

3.2. Deriving EU from PTT Model

The following result has been inspired in [7].

Lemma 2. *Given $x \in X$, let us consider the real function $V_x : P \times T \to \mathbb{R}$, defined as $V_x(p, t) := V(x, p, t)$. Then, for every $(p, t) \in P \times T$ and every $s < t$, there exists $k = k(x, p, s, t)$ $(0 < k < 1)$ such that $V_x(kp, s) = V_x(p, t)$.*

Proof. In effect, given $x \in X$ and $(p, t) \in P \times T$, for every s $(s < t)$, let us consider the following real-valued function:

$$V_{x,s} : P \to \mathbb{R}$$

defined as:

$$V_{x,s}(q) := V_x(q, s).$$

By the definition of V, the inequality

$$V_x(p, t) < V_{x,s}(p) \tag{1}$$

holds. Moreover, when $q \to 0$, $V_{x,s}(q) \to 0$. Therefore, there exists q_0, small enough, such that

$$V_{x,s}(q_0) \leq V_x(p, t). \tag{2}$$

Putting together inequalities (1) and (2), one has:

$$V_{x,s}(q_0) \leq V_x(p, t) < V_{x,s}(p).$$

As V is continuous and increasing in probability, by the Intermediate Value Theorem, there exists a value $k = k(x, p, t, s)$ $(0 < k < 1)$, such that $V_{x,s}(kp) = V_x(p, t)$, i.e., $V_x(kp, s) = V_x(p, t)$.

Time s is well defined. In effect, starting now from another couple (p', t'), such that $V_x(p, t) = V_x(p', t')$, for every s under t and t', one has:

$$V_x(kp, s) = V_x(p, t)$$

and

$$V_x(k'p', s) = V_x(p', t'),$$

from where $kp = k'p'$. □

Analogously, it could be shown that under the same conditions as Lemma 2, for every $(p, t) \in P \times T$ and every $u > t$, there exists k $(k > 1)$ such that $V_x(kp, u) = V_x(p, t)$. However, it is possible that

this result is restricted for some values of p and t whereby, for every $x \in X$, we are going to consider the values $p_x^{\max}$ and $t_x^{\max}$ such that

$$V(x,p,t) = V(x,p_x^{\max},t_x^{\max}).$$

Example 5. *With the discount function of Example 1:*

$$V(x,p,t) = \frac{xp}{1+ixpt},$$

where $i > 0$, for every (p,t) and s, the equation in q:

$$V(x,p,t) = V(x,q,s)$$

gives the following solution:

$$q = \frac{p}{1+ixp(t-s)}.$$

Observe that q makes sense for every $s \leq t$, and even for every s such that

$$1+ixp(t-s) \geq p,$$

from where:

$$t < s \leq t + \frac{1-p}{ixp}.$$

However, we will assume that always $t_x^{\min} = 0$, for every $x \in X$. To derive EU model, let

$$(x_1,p_1,t_1),(x_2,p_2,t_2),\ldots,(x_n,p_n,t_n)$$

be a sequence of n outcomes in the context of the PTT model. According to Lemma 1, there exist $p_0^1, p_0^2, \ldots, p_0^n$ ($p_0^k \leq p_k, k = 1,2,\ldots,n$), such that:

$$V(x_k,p_k,t_k) = V(x_k,p_0^k,0),$$

for every $k = 1,2,\ldots,n$. Consequently, the present value of these outcomes is:

$$V_0 := \sum_{k=1}^{n} V(x_k,p_k,t_k) = \sum_{k=1}^{n} V(x_k,p_0^k,0).$$

Independently of the shape of function V, the PTT model has been transformed into an EU model because time has been removed from the prospect (x,p,t). Observe that probabilities are not linear with respect to the discounted amounts.

Remark 2 (On non-linear probabilities). *Chew and Epstein [44] proposed some alternative theories with non-linear probabilities which may explain many behavioral paradoxes while holding normatively properties; for instance, consistency with stochastic dominance and risk aversion. In this vein, Halevy [45] provided a function for the discounted utility based on the non-linear probability weighting to evaluate the diminishing impatience related with the uncertainty of delayed rewards. These theories are useful, analytical tools in the decision-making process which allow separating the risk aversion from the elasticity of substitution.*

To explain the non-linear psychological distance, Baucells and Heukamp [1] stated that it is necessary a non-linear probability weighing bonded with non-exponential discounting.

With the aim of illustrating the non-linear probability weighing, the mathematical description by Brandstätter, Kühberger and Schneider [46] is shown to reflect the elation and disappointment in probabilities.

In effect, being $1-p$ the utility after a success and $-p$ the disutility after a failure, the expected success is calculated as follows:

$$\textit{utility after a success} \times \textit{probability of a success} = p(1-p),$$

meanwhile the expected failure is:

$$\textit{disutility after a failure} \times \textit{probability of a failure} = -p(1-p),$$

To take into account the non-linearity of the utility after a success (that is, elation) and the disutility after a failure (that is, disappointment), new utilities may be implemented to obtain the expected success and failure. In this way, the utility after a success and the disutility after a failure are $c_e s(1-p)$ and $c_d sp$, respectively, where c_e and c_d are constants and s is a non-linear surprise function. Since disappointment is an aversive emotional state, c_d is expected to be negative. A steeper slope for disappointment than for elation is assumed (this means that $c_d > c_e$).

3.3. Deriving DU from PTT Model

Analogously to Lemma 2, we can show the following statement.

Lemma 3. *Given an $x \in X$, let us consider the real function $V_x : P \times T \to \mathbb{R}$, defined as $V_x(p,t) := V(x,p,t)$. If V is regular, then, for every $(p,t) \in P \times T$ and every $q < p$, there exists $k = k(x,p,q,t)$ $(0 < k < 1)$ such that $V_x(q,kt) = V_x(p,t)$.*

Analogously, it could be shown that under the same conditions as Lemma 3, for every $(p,t) \in P \times T$ and every $q > p$, there exists $k = k(x,p,q,t)$ $(k > 1)$ such that $V_x(q,kt) = V_x(p,t)$. However, as in Section 3.2, it is possible that this result is restricted for some values of p and t. In this section, we will always assume that $p_x^{\max} = 1$, for every $x \in X$.

To derive DU model, let

$$(x_1,p_1,t_1),(x_2,p_2,t_2),\ldots,(x_n,p_n,t_n)$$

be a sequence of n outcomes in the context of the PTT model. If V is regular, according to Lemma 3, there exist $t_1^1, t_1^2, \ldots, t_1^n$ $(t_1^k \geq t_k, k = 1,2,\ldots,n)$, such that:

$$V(x_k,p_k,t_k) = V(x_k,1,t_1^k),$$

for every $k = 1,2,\ldots,n$. Consequently, the present value of these outcomes is:

$$V_0 := \sum_{k=1}^{n} V(x_k,p_k,t_k) = \sum_{k=1}^{n} V(x_k,1,t_1^k).$$

Independently of the shape of function V, the PTT model has been transformed into a DU model because probability is constant and equal to 1 whereby all outcomes are sure.

Corollary 1. *A specific PTT model of the form $V(x,p,t) := V(xp,t)$ gives rise to both a DU and an EU model.*

Proof. In effect, the equation:

$$V(xp,t) = V(xq,s)$$

implies:

1. For $q = 1$, $V(xp,t) = V(x1,t_x^{\max}) = V(x,t_x^{\max})$, which is the DU model.
2. For $t = 0$, $V(xp,t) = V(xp_x^{\min},0) = xp_x^{\min}$, which is the EU model.

□

3.4. Deriving PTT from DU or EU Model

In this context, it could be interesting to think about the converse construction, i.e., to generate a function $V(x,p,t)$ starting from $V(x,t)$ and $V(x,p)$ coming from a DU and an EU model, respectively. In effect,

(A) Given a DU model, $V(x,t)$, it can be assumed that all prospects (x,t) have probability $p=1$. In this case, we can construct a PTT model which coincides with the DU model when $p=1$:

$$V(x,p,t) := V(x,t)p$$

or

$$V(x,p,t) := V(xp,t).$$

(B) Analogously, given an EU model, $V(x,p)$, it can be understood that all prospects (x,p) expire at the same instant $t=t_0$. In this case, we can construct a PTT model which coincides with the EU model when $t=t_0$:

$$V(x,p,t) := x\left[\frac{V(x,p)}{x}\right]^{\frac{t}{t_0}}.$$

Observe that for the sake of generality, neither of the two PTT proposed models are of the form $V(x,p,t)=w(p)f(t)v(x)$ nor $V(x,p,t)=g(p,t)v(x)$ pointed out by [1].

(C) Given a DU model, $V(x,t)$, and an EU model, $V(x,p)$, if $V(x,0)=x$ and $V(x,1)=x$, respectively, we can construct a PTT model which coincides with both models, at $t=0$ and $p=1$, respectively:

$$V(0,p,t)=0$$

and

$$V(x,p,t)=\frac{1}{x}V(x,t)V(x,p),$$

otherwise.

4. Discussion

In this paper, it has been mathematically demonstrated that the PTT model is general enough to explain intertemporal and risky decisions separately. The shown association between PTT model and DU and EU models allows explaining and relating the behavioral properties and inconsistencies in real choices.

It must be mentioned that most previous literature studies anomalies in intertemporal and probabilistic choices separately. In this way, the analysis of the equivalence between certain anomalies in the context of the DU and EU models may be analyzed under the wide setting provided by the PTT model.

A way to study the similarity between time delay and uncertainty in the decision-making process is through the analysis of the immediacy and certainty effects. These effects reflect the tendency of individuals to overestimate the significance of immediacy or certainty, relative to delayed or probable outcomes, respectively. The relation between both effects allows glimpsing the analogies and the influence of time and delay on the decision-making process. Some of the main findings are listed below:

- In the seminal paper by Allais [47], the disproportionate preference for present outcomes of the immediacy effect is shown, consequently not only of the intrinsic temptation but the certainty on the payment.
- Keren and Roelofsma [28] analyze the effect of risk on the immediacy effect and the effect of time delay on the certainty effect. They suggest that time distance makes outcomes seem more uncertain by eliminating the certainty advantage of the immediate outcome. Thereof, they reveal that the introduction of uncertainty reduces the importance of time delay (if two certain rewards

are transformed to be equally probable (i.e., $p = 0.50$), the delayed one is generally preferred). In the same way, time distance decreases the influence of the probability on preferences.

- Chapman and Weber [17] prove that when the delay is introduced to sure outcomes, the certainty effect is almost eliminated just as when uncertainty is added. On the other hand, in a similar way, when explicit risk is introduced to immediate rewards, the immediacy effect is almost eliminated just as if time delay is added. It is necessary to clarify that presently there is no consensus on this topic, while Pennesi [48] confirms that when the immediate payoff becomes uncertain, the immediacy effect disappears, Abdellaoui et al. [35] claim that the immediacy effect persists under risk.
- Epper et al. [31] stress that previous papers, in most cases, determine that there are interaction effects between time and risk, such as risk tolerance increases with delay.
- Andreoni and Sprenger [49] conclude that risk preference is not time preference: "subjects exhibit a preference for certainty when it is available, but behave largely as discounted expected utility maximizers away from certainty".

Another way to conclude that delay and risk choices have non-parallel decision mechanisms has been proved by Cruz Rambaud and Sánchez Pérez [7] and Chapman and Weber [17]. The effect of the reward size on the treatment of delayed and probabilistic outcomes moves in the opposite direction. On one hand, in intertemporal choices, as the reward amount increases, decision-makers prefer the delayed but greater reward (this is called the "magnitude effect"). On the other hand, in risky choices, the reward size increase implies a decreasing sensitivity; decision-makers prefer the more probable but lower reward (this is called the "peanut effect").

Despite some previous papers claiming that DU and EU models anomalies move in the opposite direction, the findings of our research are in line with most part of the previous literature linking the anomalies exhibited by both models. Given that the PTT model considers a unique framework to deal with uncertainty and delayed rewards, it may imply that risk and time anomalies can be captured by a unique model. Thus, the results of this research contribute to clearly understanding the aforementioned relationship by considering risk and time anomalies inside the same framework.

Specifically, Leland and Schneider [14] pointed that "the PTT model accounts for three systematic interaction effects between risk and time preferences when choices involve both risk and time delays as well as some other fundamental behaviors such as the common ratio and common difference effects".

In effect, Schneider [21] introduces the two following dual concepts:

- *Time interacts with risk preference* if, for every $x \in (0, z)$, $\alpha \in (0, 1)$, and $s > t$,

$$(x, p, t) \sim (z, \alpha p, t) \text{ implies } (x, p, s) \prec (z, \alpha p, s).$$

- *Risk interacts with time preference* if, for every $x \in (0, z)$, t, $\Delta > 0$, and $q < p$,

$$(x, p, t) \sim (z, p, t + \Delta) \text{ implies } (x, q, t) \prec (z, q, t + \Delta).$$

By applying Lemmas 2 and 3, it can be shown that the former definitions are equivalent in the presence of the so-called *reverse sub-endurance* (see [7]):

For every $x \in (0, z)$, t, $\Delta > 0$, and $\alpha \in (0, 1)$,

$$(z, p, t + \Delta) \sim (z, \alpha p, t) \text{ implies } (x, p, t + \Delta) \succ (x, \alpha p, t).$$

In this way, our objective will be to investigate the possible equivalence between other dual concepts in the ambit of risk and time preferences. However, this issue will be left for further research in the context of the PTT model.

5. Conclusions

In this paper, an extensive review of the previous methodologies which assess risk and delayed rewards in a unique framework has been made. It reveals that there is still room to improve the existing methodologies to reach this goal. Specifically, this paper has dealt with the classical problem of the possible relationship between the DU and the EU models but treated from the joint perspective of the PTT model.

In this paper, the equivalence of the PTT model with the DU is demonstrated and the EU models separately considered. In this sense, this paper's main findings are three-fold:

- On the one hand, the DU model has been derived from the PTT model, by taking a specific value of probability. Specifically, we have found that the PTT model is equivalent to the discount function with stochastic time $F(x, T_{k,x})$, where $T_{k,x}$ is the random variable: "Time period in which the reward x can be delivered, at a discounting level k".
- Analogously, given a concrete value of time, the EU model can be derived from the PTT model.
- Finally, this paper provides some insights into the construction of a PTT model starting from a DU and an EU model. However, more future research is needed on this topic.

Thus, this paper shows the validity of the PTT model to assess risky and delayed rewards separately. As pointed out, a limitation of the paper is the difficulty of building a complete PTT model starting from DU and EU models in a same framework. However, a solution to do this is provided: the construction of the PTT model starting from DU and EU models by using specific models able to satisfy the appropriate patterns of time and probability pointed out by Baucells and Heukamp [1].

As a future line of research, we stress that it is necessary to continue analyzing the equivalence between the PTT and DU and EU models from the perspective of the different effects or anomalies when considering delay and time individually (in DU and EU models, respectively) and jointly (PTT model). A second future research line is to check the validity of the PTT model and its relationship with DU and EU models from an experimental point of view.

Author Contributions: The individual contribution of each author has been as follows: formal analysis, funding acquisition, investigation and writing–original draft, S.C.R.; conceptualization, methodology, writing–review & editing and supervision, A.M.S.P. All authors have read and agreed to the published version of the manuscript.

Funding: The author gratefully acknowledges financial support from the Spanish Ministry of Economy and Competitiveness [National R&D Project "La sostenibilidad del Sistema Nacional de Salud: reformas, estrategias y propuestas", reference: DER2016-76053-R].

Acknowledgments: We are very grateful for the valuable comments and suggestions offered by the Academic Editor and three anonymous referees.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

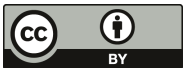
DU	Discounted Utility
EU	Expected Utility
DEU	Discounted Expected Utility
PTT	Probability and Time Trade-Off

References

1. Baucells, M.; Heukamp, F.H. Probability and time trade-off. *Manag. Sci.* **2012**, *58*, 831–842. [CrossRef]
2. Green, L.; Myerson, J. A discounting framework for choice with delayed and probabilistic rewards. *Psychol. Bull.* **2004**, *130*, 769–792. [CrossRef] [PubMed]
3. Soares dos Santos, L.; Destefano, N.; Souto Martinez, A. Decision making generalized by a cumulative probability weighting function. *Phys. Stat. Mech. Its Appl.* **2018**, *490*, 250–259. [CrossRef]

4. Weber, B.J.; Chapman, G.B. The combined effects of risk and time on choice: Does uncertainty eliminate the immediacy effect? *Organ. Behav. Hum. Decis.* **2005**, *96*, 104–118. [CrossRef]
5. Luckman, A.; Donkin, C.; Newell, B.R. Can a single model account for both risky choices and inter-temporal choices? Testing the assumptions underlying models of risky inter-temporal choice. *Psychon. Bull. Rev.* **2018**, *25*, 785–792. [CrossRef]
6. Baucells, M.; Heukamp, F.H.; Villasis, A. Risk and time preferences integrated. In Proceedings of the Foundations of Utility and Risk Conference, Rome, Italy, 30 April 2006.
7. Cruz Rambaud, S.; Sánchez Pérez, A.M. The magnitude and peanuts effects: Searching implications. *Front. Appl. Math. Stat.* **2018**, *4*, 36. [CrossRef]
8. Samuelson, P.A. A note on measurement of utility. *Rev. Econ. Stud.* **1937**, *4*, 155–161. [CrossRef]
9. Von Neumann, J.; Morgenstern, O. *Theory of Games and Economic Behavior*, 2nd ed.; Princeton University Press: Princeton, NJ, USA, 1947.
10. Schoemaker, P.J.H. The Expected Utility model: Its variants, purposes, evidence and limitations. *J. Econ. Lit.* **1982**, *20*, 529–563.
11. Coble, K.H.; Lusk, J.L. At the nexus of risk and time preferences: An experimental investigation. *J. Risk Uncertain.* **2010**, *41*, 67–79. [CrossRef]
12. Andreoni, J.; Sprenger, C. *Risk Preferences Are Not Time Preferences: Discounted Expected Utility with a Disproportionate Preference for Certainty*; Working Paper 16348; National Bureau of Economic Research: Cambridge, MA, USA, 2010.
13. Jamison, D.T. *Studies in Individual Choice Behavior*; RAND Memorandum No. P-4255; RAND Corporation: Santa Monica, CA, USA, 1970.
14. Leland J.; Schneider M. *Risk Preference, Time Preference, and Salience Perception*; ESI Working Papers 17-16; ESI: Orange, CA, USA, 2017.
15. Lanier, J.; Miao, B.; Quah, J.K.H.; Zhong, S. Intertemporal Consumption with Risk: A Revealed Preference Analysis. Available online: http://dx.doi.org/10.2139/ssrn.3168361 (accessed on 28 February 2020).
16. Onay, S.; Öncüler, A. Intertemporal choice under timing risk: An experimental approach. *J. Risk Uncertain.* **2007**, *34*, 99–121. [CrossRef]
17. Chapman, G.B.; Weber, B.J. Decision biases in intertemporal choice and choice under uncertainty: Testing a common account. *Mem. Cogn.* **2006**, *34*, 589–602. [CrossRef]
18. Quiggin, J.; Horowitz, J. Time and risk. *J. Risk Uncertain.* **1995**, *10*, 37–55. [CrossRef]
19. Prelec, D.; Loewenstein, G. Decision making over time and under uncertainty: A common approach. *Manag. Sci.* **1991**, *37*, 770–786. [CrossRef]
20. Green, L., Myerson, J.; Ostaszewski, P. Amount of reward has opposite effects on the discounting of delayed and probabilistic outcomes. *J. Exp. Psychol. Mem. Cogn.* **1999**, *2*, 418–427. [CrossRef]
21. Schneider, M. *Dual-Process Utility Theory: A Model of Decisions Under Risk and Over Time*; ESI Working Paper; ESI: Orange, CA, USA, 2016.
22. Gollier, C. *The Economics of Risk and Time*; MIT Press: Cambridge, MA, USA, 2004.
23. Rachlin, H.; Raineri, A.; Cross, D. Subjective probability and delay. *J. Exp. Anal. Behav.* **1991**, *55*, 233–244. [CrossRef] [PubMed]
24. Takahashi, T.; Ikeda, K.; Hasegawa, T. A hyperbolic decay of subjective probability of obtaining delayed rewards. *Behav. Brain Funct.* **2007**, *3*, 1–11. [CrossRef] [PubMed]
25. Issler, J.V.; Piqueira, N.S. Estimating relative risk aversion, the discount rate, and the intertemporal elasticity of substitution in consumption for Brazil using three types of utility function. *Braz. Rev. Econom.* **2000**, *20*, 201–239. [CrossRef]
26. Van Praag, B.M.S.; Booij, A.S. *Risk Aversion and the Subjective Time Preference: A Joint Approach*; *CESifo Working Paper Series* 923; CESifo Group: Munich, Germany, 2003.
27. Anderhub, V.; Güth, W.; Gneezy, U.; Sonsino, D. On the interaction of risk and time preferences: An experimental study. *Ger. Econ. Rev.* **2001**, *2*, 239–253. [CrossRef]
28. Keren, G.; Roelofsma, P. Immediacy and certainty in intertemporal choice. *Organ. Behav. Hum. Decis. Process.* **1995**, *63*, 287–297. [CrossRef]
29. Fisher, I. *The Theory of Interest*; Macmillan: New York, NY, USA, 1930.
30. Dai, J.; Pachur, T.; Pleskac, T.J.; Hertwig, R. What the future holds and when: A description-experience gap in intertemporal choice. *Psychol. Sci.* **2019**, *30*, 1218–1233. [CrossRef]

31. Epper, T.; Fehr-Duda, H.; Bruhin, A. Viewing the future through a warped lens: Why uncertainty generates hyperbolic discounting. *J. Risk Uncertain.* **2011**, *43*, 169–203. [CrossRef]
32. Takahashi, T. A comparison of intertemporal choices for oneself versus someone else based on Tsallis' statistics. *Phys. Stat. Mech. Its Applications* **2007**, *385*, 637–644. [CrossRef]
33. Ericson, K.M.; Laibson, D. Intertemporal choice (No. w25358). In *Handbook of Behavioral Economics*; Elsevier: Amsterdam, The Netherlands, 2018.
34. Loewenstein, G.; Prelec, D. Anomalies in intertemporal choice: Evidence and an interpretation. *Q. J. Econ.* **1992**, *107*, 573–597. [CrossRef]
35. Abdellaoui, M.; Kemel, E.; Panin, A.; Vieider, F.M. Measuring time and risk preferences in an integrated framework. *Games Econ. Behav.* **2019**, *115*, 459–469. [CrossRef]
36. Baucells, M.; Heukamp, F.H. Common ratio using delay. *Theory Decis.* **2010**, *68*, 149–158. [CrossRef]
37. Vanderveldt, A.; Green, L.; Myerson, J. Discounting of monetary rewards that are both delayed and probabilistic: Delay and probability combine multiplicatively, not additively. *J. Exp. Psychol. Learn. Cogn.* **2015**, *41*, 148–162. [CrossRef]
38. Yi, R.; de la Piedad, X.; Bickel, W.K. The combined effects of delay and probability in discounting. *Behav. Process.* **2006**, *73*, 149–155. [CrossRef]
39. DeJarnette, P.; Dillenberger, D.; Gottlieb, D.; Ortoleva, P. Time lotteries, second version (No. 15-026v2). *Penn Institute for Economic Research*; Department of Economics, University of Pennsylvania: Philadelphia, PA, USA, 2018.
40. Han, R.; Takahashi, T. Psychophysics of time perception and valuation in temporal discounting of gain and loss. *Phys. Stat. Mech. Its Appl.* **2012**, *391*, 6568–6576. [CrossRef]
41. Konstantinidis, E.; Van Ravenzwaaij, D.; Güney, Ş; Newell, B.R. Now for sure or later with a risk? Modeling risky intertemporal choice as accumulated preference. *Decision* **2020**, *7*, 91–120. [CrossRef]
42. Caliendo, F.N.; Findley, T.S. Discount functions and self-control problems. *Econ. Lett.* **2014**, *122*, 416–419. [CrossRef]
43. Cruz Rambaud, S.; Parra Oller, I.M.; Valls Martínez, M.C. The amount-based deformation of the *q*-exponential discount function: A joint analysis of delay and magnitude effects. *Phys. Stat. Mech. Its Applications* **2019**, *508*, 788–796. [CrossRef]
44. Chew, S.H.; Epstein, L.G. Nonexpected utility preferences in a temporal framework with an application to consumption-savings behaviour. *J. Econ. Theory* **1990**, *50*, 54–81. [CrossRef]
45. Halevy, Y. Strotz meets Allais: Diminishing impatience and the certainty effect. *Am. Econ.* **2008**, *98*, 1145–1162. [CrossRef]
46. Brandstätter, E.; Kühberger, A.; Schneider, F. A cognitive-emotional account of the shape of the probability weighting function. *J. Behav. Decis. Mak.* **2002**, *15*, 79–100. [CrossRef]
47. Allais, M. Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine. *Econometrica* **1953**, *21*, 503–546. [CrossRef]
48. Pennesi, D. Uncertain discount and hyperbolic preferences. *Theory Decis.* **2017**, *83*, 315–336. [CrossRef]
49. Andreoni, J.; Sprenger, C. Risk preferences are not time preferences. *Am. Econ. Rev.* **2012**, *102*, 3357–3376. [CrossRef]



mathematics

MDPI

Article

Some Notes on the Formation of a Pair in Pairs Trading

José Pedro Ramos-Requena [1], Juan Evangelista Trinidad-Segovia [1,*] and Miguel Ángel Sánchez-Granero [2]

[1] Department of Economics and Business, University of Almería, Ctra. Sacramento s/n, La Cañada de San Urbano, 04120 Almería, Spain; jpramosre@ual.es

[2] Department of Matematics, University of Almería, Ctra. Sacramento s/n, La Cañada de San Urbano, 04120 Almería, Spain; misanche@ual.es

* Correspondence: jetrini@ual.es

Received: 20 January 2020; Accepted: 28 February 2020; Published: 5 March 2020

Abstract: The main goal of the paper is to introduce different models to calculate the amount of money that must be allocated to each stock in a statistical arbitrage technique known as pairs trading. The traditional allocation strategy is based on an equal weight methodology. However, we will show how, with an optimal allocation, the performance of pairs trading increases significantly. Four methodologies are proposed to set up the optimal allocation. These methodologies are based on distance, correlation, cointegration and Hurst exponent (mean reversion). It is showed that the new methodologies provide an improvement in the obtained results with respect to an equal weighted strategy.

Keywords: pairs trading; hurst exponent; financial markets; long memory; co-movement; cointegration

1. Introduction

Efficient Market Hypothesis (EMH) is a well-known topic in finance. Implications of the weak form of efficiency is that information about the past is reflected in the market price of a stock and therefore, historical market data is not helpful for predicting the future. An investor in an efficient market will not be able to obtain a significant advantage over a benchmark portfolio or a market index trading based on historical data (for a review see Reference [1,2]).

On the opposite way, some researchers have shown that the use of historical data as well as trading techniques is sometimes possible due to temporal markets anomalies. Despite that most of economists consider that these anomalies are not compatible with an efficient market, recent papers have shown new perspectives called Fractal Market Hypothesis (FMH) and Adaptive Market Hypothesis (AMH), that tries to integrate market anomalies into the efficient market hypothesis.

The EMH was questioned by the mathematician Mandelbrot in 1963 and after the economist Fama showed his doubts about the Normal distribution of stock returns, essential point of the efficient hypothesis. Mandelbrot concluded that stock prices exhibit long-memory, and proposed a Fractional Brownian motion to model the market. Di Matteo [3,4] considered that investors can be distinguished by the investment horizons in which they operate. This consideration allows us to connect the idea of long memory and the efficiency hypothesis. In the context of an efficient market, the information is considered as a generic item. This means that the impact that public information has over each investor is similar. However, the FMH assumes that information and expectations affect in a different way to traders, which are only focused on short terms and long term investors [5,6].

The idea of a AMH has been recently introduced by Lo [7] to reflect an evolutionary perspective of the market. Under this new idea, markets show complex dynamics at different times which make that some arbitrage techniques perform properly in some periods and poorly in others.

In an effort of conciliation, Sanchez et al. [8] remarks that the market dynamic is the results of different investors interactions. In this way, scaling behavior patterns of a specific market can characterize it. Developed market price series usually show only short memory or no memory whereas emerging markets do exhibit long-memory properties. Following this line, in a recent contribution, Sanchez et al. [9] proved that pairs trading strategies are quite profitable in Latin American Stock Markets whereas in Nasdaq 100 stocks, it is only in high volatility periods. These results are in accordance with both markets hyphotesis. A similar result is obtained by Zhang and Urquhart [10] where authors are able to obtain a significant exceed return with a trading strategy across Mainland China and Hong kong but not when the trading is limited to one of the markets. The authors argue that this is because of the increasing in the efficiency of Mainland China stock market and the decreasing of the Hong Kong one because of the integration of Chinese stock markets and permission of short selling.

These new perspectives of market rules explain why statistical arbitrage techniques, such as pairs trading, can outperform market indexes if they are able to take advantage of market anomalies. In a previous paper, Ramos et al. [11] introduced a new pairs trading technique based on Hurst exponent which is the classic and well known indicator of market memory (for more details, References [8,12] contain an interesting review). For our purpose, the selection of the pair policy is to choose those pairs with the lowest Hurst exponent, that is, the more anti-persistent pairs. Then we use a reversion to the mean trading strategy with the more anti-persistent pairs according with the previously mentioned idea that developed market prices show short memory [3,13–15].

Pairs trading literature is extensive and mainly focused on the pair selection during the trading period as well as the developing of a trading strategy. The pioneer paper was Gatev et al. [16] where authors introduced the distance method with an application to the US market. In 2004, Vidyamurthy [17] presented the theoretical framework for pair selection using the cointegration method. Since then, different analysis have been carried out using this methodology in different markets, such us the European market [18,19], the DJIA stocks [20], the Brazilian market [21,22] or the STOXX 50 index [23]. Galenko et al. [24] made an application of the cointegration method to arbitrage in fund traded on different markets. Lin et al. [25] introduced the minimum profit condition into the trading strategy and Nath [26] used the cointegration method in intraday data. Elliott et al. [27] used Markov chains to study a mean reversion strategy based on differential predictions and calibration from market observations. The mean reversion approach has been tested in markets not considered efficient such us Asian markets [28] or Latin American stock markets [9]. A recent contribution of Ramos et al. [29] introduced a new methodology for testing the co-movement between assets and they tested it in statistical arbitrage. However, researchers did not pay attention to the amount of money invested in every asset, considering always a null dollar market exposition. This means that when one stock is sold, the same amount of the other stock is purchased. In this paper we propose a new methodology to improve pairs trading performance by developing new methods to improve the efficiency in calculating the ratio to invest in each stock that makes up the pair.

2. Pair Selection

One of the topics in pairs trading is how to find a suitable pair for pairs trading. Several methodologies have been proposed in the literature, but the more common ones are co-movement and the distance method.

2.1. Co-Movement

Baur [30] defines co-movement as the shared movement of all assets at a given time and it can be measured using *correlation* or *cointegration* techniques.

Correlation technique is quite simple, and the higher the correlation coefficient is, the greatest they move in sync. An important issue to be considered is that correlation is intrinsically a short-run measure, which implies that a correlation strategy will work better with a lower frequency trading strategy.

In this work, we will use the Spearman correlation coefficient, which is a nonparametric range statistic which measure the relationship between two variables. This coefficient is particularly useful when the relationship between the two variables is described by a monotonous function, and does not assume any particular distribution of the variables [31].

The Spearman correlation coefficient for a sample A_i, B_i of size n can be described as follows: first, consider the ranks of the samples rgA_i, rgB_i, then the Spearman correlation coefficient r_s is calculated as:

$$r_s = \rho_{rg_A, rg_B} = \frac{cov(rg_A, rg_B)}{\sigma_{rg_A} * \sigma_{rg_B}}, \tag{1}$$

where

- ρ denotes the Pearson correlation coefficient, applied to the rank variables
- $cov(rg_A, rg_B)$, is the covariance of the rank variables.
- σ_{rg_A} and σ_{rg_B}, are the standard deviations of the rank variables.

Cointegration approach was introduced by Engle and Granger [32] and it considers a different type of co-movement. In this case, cointegration refers to movements in prices, not in returns, so cointegration and correlation are related, but different concepts. In fact, cointegrated series can perfectly be low correlated.

Two stocks A and B are said to be cointegrated if there exists γ such that $P_t^A - \gamma P_B^t$ is a stationary process, where P_t^A and P_B^t are the log-prices A and B, respectively. In this case, the following model is considered:

$$P_t^A - \gamma P_B^t = \mu + \epsilon_t, \tag{2}$$

where

- μ is the mean of the cointegration model
- ϵ_t is the cointegration residual, which is a stationary, mean-reverting process
- γ is the cointegration coefficient.

We will use the ordinary least squares (OLS) method to estimate the regression parameters. Through the Augmented Dickey Fuller test, we will verify if the residual ϵ_t is stationary or not, and with it we will check if the stocks are co-integrated.

2.2. The Distance Method

This methodology was introduced by Gatev et al. [16]. It is based on minimizing the sum of squared differences between somehow normalized price series:

$$ESD = \sum_t (S_A(t) - S_B(t))^2, \tag{3}$$

where $S_A(t)$ is the cumulative return of stock A at time t and $S_B(t)$ is the cumulative return of stock B at time t.

The best pair will be the pair whose distance between its stocks is the lowest possible, since this means that the stocks moves in sync and there is a high degree of co-movement between them.

An interesting contribution to this trading system was introduced by Do and Faff [33,34]. The authors replicated this methodology for the U.S. CRSP stock universe and an extended period. The authors confirmed a declining profitability in pairs trading as well as the unprofitability of the trading strategy due to the inclusion of trading costs. Do and Faff then refined the selection method to improve the pair selection. The authors restricted the possible combinations only within the 48 Fama-French industries and they looked for pairs with a high number of zero-crossings to favor the pairs with greatest mean-reversion behavior.

2.3. Pairs Trading Strategy Based on Hurst Exponent

Hurst exponent (H from now on) was introduced by Hurst in 1951 [35] to deal with the problem of reservoir control for the Nile River Dam. Until the beginning of the 21st century, the most common methodology to estimate H was the R/S analysis [36] and the DFA [37], but due to accuracy problems remarked by several studies (see for example References [38–41]), new algorithms were developed for a more efficient estimation of the Hurst exponent, some of them with its focus on financial time series. One of the most important methodologies is the GHE algorithm, introduced in Reference [42], which is a general algorithm with good properties.

The GHE is based on the scaling behavior of the statistic

$$K_q(\tau) = \frac{< |X(t+\tau) - X(t)|^q >}{< |X(t)|^q >}$$

which is given by

$$K_q(\tau) \propto \tau^{qH}., \tag{4}$$

where τ is the scale (usually chosen between 1 and a quarter of the length of the series), H is the Hurst exponent, $< \cdot >$ denotes the sample average on time t and q is the order of the moment considered. In this paper we will always use $q = 1$.

The GHE is calculated by linear regression, taking logarithms in the expression contained in (4) for different values of τ [3,43].

The interpretation of H is as follow: when H is greater than 0.5, the process is persistent, when H is less than 0.5, it is anti persistent, while Brownian motion has a value of H equal to 0.5.

With this technique, pairs with the lowest Hurst exponent has to be chosen in order to apply reversion to the mean strategies which is also the base of correlation and cointegration strategies.

2.4. Pairs Trading Strategy

Next, we describe the pairs trading strategy, which is taken from Reference [11]. As usual, we consider two periods. The first one is the formation period (one year), which is used for the pair selection. This is done using the four methods defined in this section (distance, correlation, cointegration and Hurst exponent). The second period is the execution period (six months), in which all selected pairs are traded as follows:

- In case $s > m + \sigma$ the pair will be sold. The position will be closed if $s < m$ or $s > m + 2\sigma$.
- In case $s < m - \sigma$ the pair will be bought. The position will be closed if $s > m$ or $s < m - 2\sigma$.

where m is a moving average of the series of the pair and s is a moving standard deviation of m.

3. Forming the Pair: Some New Proposals

As we remarked previously, all works assume that the amount purchased in a stock is equal to the amount sold in the other pair component. The main contribution of this paper is to analyse if not assuming an equal weight ratio in the formation of the pair improves the performance of the different pair trading strategies. In this section different methods are proposed.

When a pair is formed, we use two stocks A and B. This two stocks have to be normalized somehow, so we introduce a constant b such that stock A is comparable to stock bB. Then, to buy an amount T of the pair AB means that we buy $\frac{1}{b+1}T$ of stock A and sell $\frac{b}{b+1}T$ of stock B, while to sell an amount T of the pair AB means that we sell $\frac{1}{b+1}T$ of stock A and buy $\frac{b}{b+1}T$ of stock B.

We will denote by $p_X(t)$ the logarithm of the price of stock X in time t minus the logarithm of the price of stock X at time $t = 0$, that is $p_X(t) = \log(price_X(t)) - \log(price_X(0))$, and by $r_X(t)$ the log-return of stock X between times $t-1$ and t, $r_X(t) = p_X(t) - p_X(t-1)$.

In this paper we discuss the following ways to calculate the weight factor b:

1. Equal weight (EW).

   In this case $b = 1$. This is the way used in most of the literature. In this case, the position in the pair is dollar neutral. This method was used in Reference [16], and since then, it has become the more popular procedure to fix b.
2. Based on volatility.

   Volatility of stock A is $std(r_A)$ and volatility of stock B is $std(r_B)$. If we want that A and bB have the same volatility then $b = std(r_A)/std(r_B)$. This approach was used in Reference [11] and it is based on the idea that both stocks are normalized if they have the same volatility.
3. Based on minimal distance of the log-prices.

   In this case we minimize the function $f(b) = \sum_t |p_A(t) - bp_B(t)|$, so we look for the weight factor b such that p_A and bp_B has the minimum distance. This approach is based on the same idea that the distance as a selection method. The closer is the evolution of the log-price of stocks A and bB, the more reverting to the mean properties the pair will have.
4. Based on correlation of returns.

   If returns are correlated then r_A is approximately equal to br_B, where b is obtained by linear regression $r_A = br_B$. In this case, if returns of stocks A and B are correlated, then the distribution of r_A and br_B will be the same, so we can use this b to normalize both stocks.
5. Based on cointegration of the prices.

   If the prices (in fact, the log-prices) of both stocks A and B are cointegrated then $p_A - bp_B$ is stationary, whence b is obtained by linear regression $p_A = bp_B$. In this case, this value of b makes the pair series stationary so we can expect reversion to the mean properties of the pair series. Even if the stocks A and B are not perfectly cointegrated, this method for the calculation of b may be still valid, since, thought $p_A - bp_B$ may be not stationary, it can be somehow close to it or still have mean-reversion properties.
6. Based on lowest Hurst exponent of the pair.

   The series of the pair is defined as $s(b)(t) = p_A(t) - bp_B(t)$. In this case, we look for the weight factor b such that the series of the pair $s(b)$ has the lowest Hurst exponent, what implies that the series is as anti-persistent as possible. So we look for b which minimizes the function $f(b) = H(s(b))$, where $H(s(b))$ is the Hurst exponent of the pair series $s(b)$. The idea here is similar to the cointegration method, but from a theoretical point of view, we do not expect $p_A - bp_B$ to be stationary (which is quite difficult with real stocks), but to be anti-persistent, which is enough for our trading strategy.

4. Experimental Results

For testing the results through the different models introduced in this paper, we will use the components of the Nasdaq 100 index technological sector (see Table A1 in Appendix A),for the period between January 1999 and December 2003, coinciding with the "dot.com" bubble crash and the period between January 2007 and December 2012, this period coincides with the financial instability caused by the "subprime" crisis. These periods are choosen based on the results showed by Sánchez et al. [9].

We use Pairs Trading traditional methods (Distance Method, Correlation and Cointegration) in addition to the method developed by Ramos et al. [11] based on the Hurst exponent.

In Appendix B, it is shown the results obtained for different selection methods and different ways to calculate b, for the two selected periods. In addition to the returns obtained for each portfolio of pairs, we include two indicators of portfolio performance and risk, the Sharpe Ratio and the maximum Drawdown.

In the first period analyzed, the *EW* method to calculate b is never the best one. The best methods to calculate b seems to be the cointegration method and the minimization of the Hurst exponent. Also note that the Spearman correlation, the cointegration and the Hurst exponent selection methods provide strategies with high Sharpe ratios for several methods to calculate b.

In the second period analyzed, the *EW* method to calculate b works fine with the cointegration selection method, but it is not so good with the other ones, while the correlation method to calculate b is often one of the best ones.

Note that, in both periods, the Sharpe ratio when we use *EW* to calculate b are usually quite low with respect to the other methods.

Figures 1–4 show the cumulative log-return of the strategy for different selection methods and different ways to calculate b.

Figure 1 shows the returns obtained for the period 1999–2003 using the co-integration approach as a selection method. We can observe that during the whole period, the best option is to choose to calculate the b factor by means of the lowest value of the Hurst exponent, while the *EW* method is the worst.

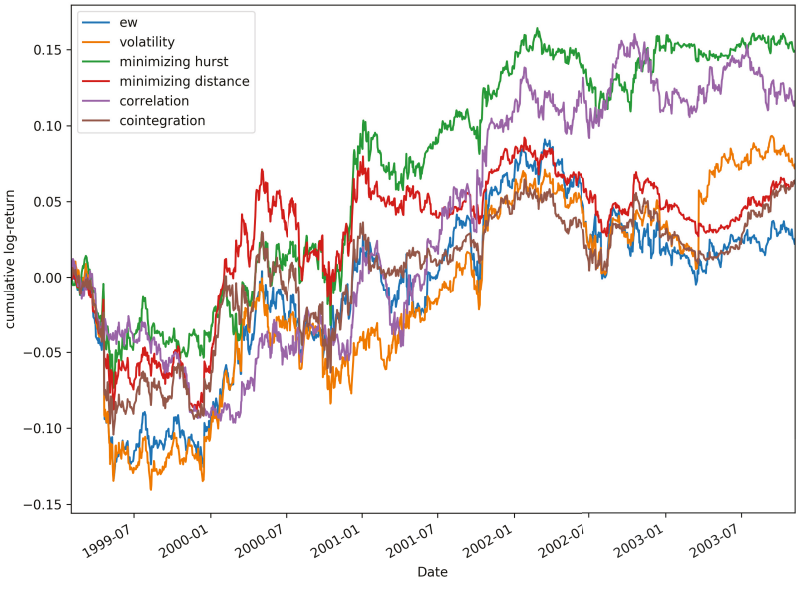


Figure 1. Comparative portfolio composed of 30 pairs using cointegration method for selection during the period 1999–2003.

Figure 2 represents the returns obtained for each of the b calculation methods for the 1999–2003 period, using the Hurst exponent method for the selection of pairs and a portfolio composed of 20 pairs. It can be observed that during the period studied, the results obtained using the *EW* method are also negative, while the Hurst exponent method is again the best option.

For the period 2007–2012, for a portfolio composed of 20 pairs selected using the distance method, Figure 3 shows the cumulative returns for the different methods proposed. In this case we can highlight

the methods of correlation, minimizing distance and cointegration, as the methods to calculate b that provide the highest returns. Again, we can observe that the worst options would be the EW method together with the volatility one.

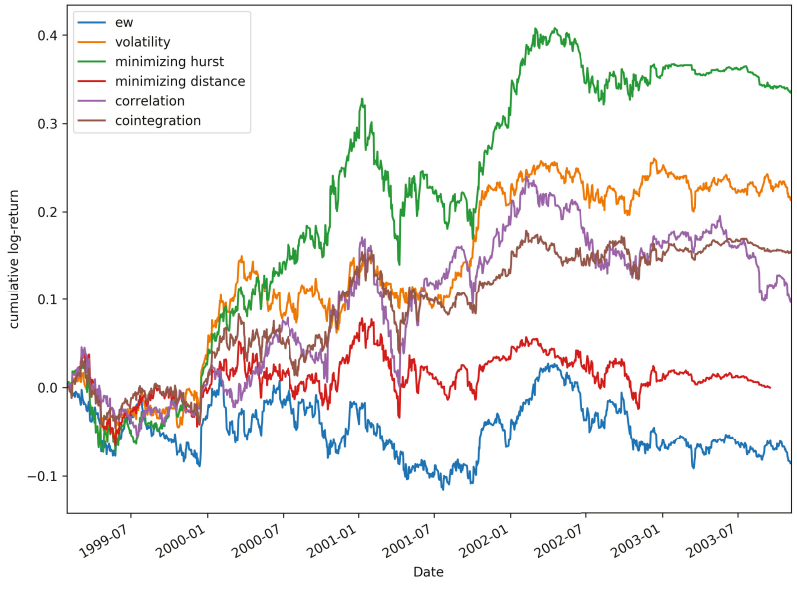


Figure 2. Comparative portfolio composed of 20 pairs using the Hurst exponent method for selection during the period 1999–2003.

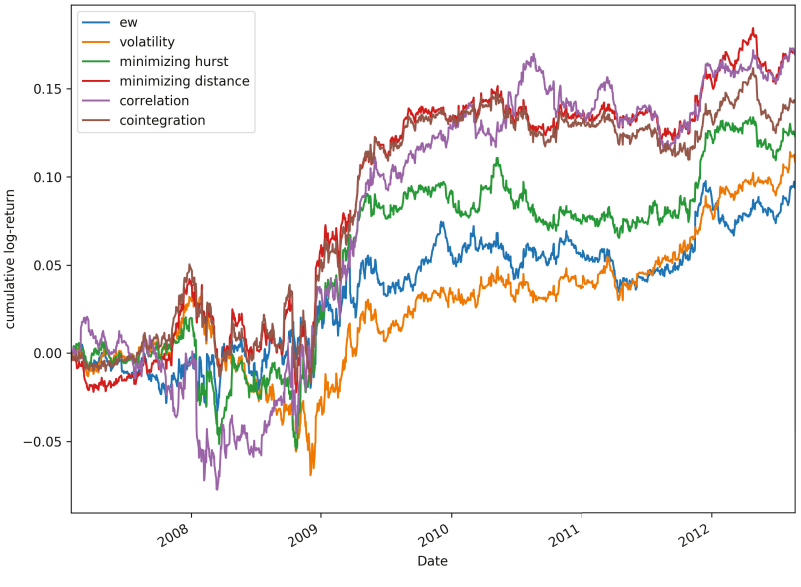


Figure 3. Comparative portfolio composed of 20 pairs using distance method for selection during the period 2007–2012.

Figure 4 shows the results obtained using the different models to calculate the b factor for a portfolio of 10 pairs by selecting them using the Spearman model. We can observe that all returns are positive throughout the period studied (2007–2012). The most outstanding are the methods of correlation, minimum distance and volatility, which move in a very similar way during this period. On the contrary, the method of the lowest value of the Hurst exponent and the *EW* one are the worst options during the whole period.

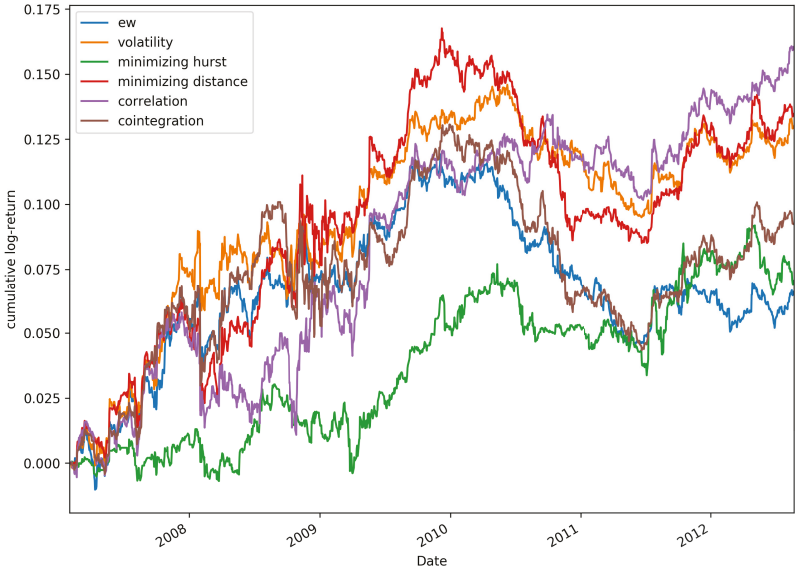


Figure 4. Comparative portfolio composed of 10 pairs using Spearman method for selection during the period 2007–2012.

Finally, we complete our sensitivity analysis by analyzing the influence of the strategy considered in Section 2.4. We consider the Hurst exponent as the selection method, 20 pairs in the portfolio and the period 1999–2003. We change the strategy by using 1 (as before), 1.5 and 2 standard deviations. That is, we modify the strategy as follows:

- In case $s > m + k\sigma$ the pair will be sold. The position will be closed if $s < m$ or $s > m + 2k\sigma$.
- In case $s < m - k\sigma$ the pair will be bought. The position will be closed if $s > m$ or $s < m - 2k\sigma$.

where $k = 1, 1.5, 2$. Table A2 shows that the *EW*, correlation and minimal distance obtain the worst results, while cointegration and the Hurst exponent obtain robust and better results for the different values of k.

Discussion of the Results

In Tables A3–A10, the results obtained with a pair trading strategy are shown. In those tables, we have consider four different methods for the pair selection (distance, correlation, cointegration and Hurst exponent), three different number of pairs (10, 20 and 30 pairs) and two periods (1999–2003 and 2007–2012). Overall, if we focus on the Sharpe ratio of the results, in 58% of the cases (14 out of 24) the *EW* method for calculating b obtains one of the three (out of seven) worst results. If we compare the *EW* method with the other methods proposed we obtain the following: minimal Hurst exponent is better than *EW* in 58% of the cases, minimal distance is better than *EW* in 58% of the cases, correlation is better than *EW* in 67% of the cases, cointegration is better than *EW* in 58% of the cases and volatility

is better than *EW* in 50% of the cases. So, in general, the proposed methods (except the volatility one) tend to be better than the *EW* one.

However, since we are considering stocks in the technology sector, if we focus in the dot.com bubble (that is, the period 1999–2003) which affected more drastically to the stocks in the portfolio, we have, considering the Sharpe ratio of the results, that in 83% of the cases (10 out of 12) the *EW* method for calculating b obtains one of the three (out of seven) worst results. In this period, if we compare the *EW* method with the other methods proposed we obtain the following: minimal Hurst exponent is better than *EW* in 75% of the cases, minimal distance is better than *EW* in 83% of the cases, correlation is better than *EW* in 83% of the cases, cointegration is better than *EW* in 83% of the cases and volatility is better than *EW* in 58% of the cases. So, in general, the proposed methods (except the volatility one) tend to be much better than the *EW* one in this period.

On the other hand, in the second period (2007–2012), the *EW* performs much better than in the first period (1999–2003) and it does similarly or slightly better than the other methods.

Results show that these novel approaches used to calculate the factor b improve the results obtained compared with the classic *EW* method for the different strategies and mainly in the first period considered (1999–2003). Therefore, it seems that the performance of pairs trading can be improved not only acting on the strategy, but also on the method for the allocation in each stock.

In this section we have tested different methods for the allocation in each stock of the pair. Though we have used the different allocation methods with all the selection methods analyzed, it is clear that some combinations make more sense than others. For example, if the selection of the pair is done by selecting the pair with a lower Hurst exponent, the allocation method based on the minimization of the Hurst exponent of the pair should work better than other allocation methods.

One of the main goal of this paper is to point out that the allocation in each stock of the pair can be improved in the pairs trading strategy and we have given some ways to make this allocation. However, further research is needed to asses which of the methods is the best for this purpose. Even better, which of the combinations of selection and allocation method is the best. Though this problem depends on many factors, and some of them changes, depending on investor preferences, a multi-criteria decision analysis (see, for example References [44–46]) seems to be a good approach to deal with it.

In fact, in future research it can be tested if the selection method can be improved if we take into account the allocation method. For example, for the distance selection method, we can use the allocation method based on the minimization of the distance to normalize the price of the stocks in a different way than in the classical distance selection method, taking into account the allocation in each stock. Not all selection methods can be improved in this way (for example, the correlation selection method will not improve), but some of them, including some methods which we have not analyzed in this paper or future selection methods, could be improved.

5. Conclusions

In pairs trading literature, researchers have focused their attention in increasing pairs trading performance proposing different methodologies for pair selection. However, in all cases it is assumed that the amount invested in each stock of a pair (b) must be equal. This technique is called *Equally Weighted* (*EW*).

This paper presents a novel approach to try to improve the performance of this statistical arbitrage technique through novel methodologies in the calculation of b. Any selection method can benefit from these new allocation methods. Depending on the selection method used, we prove that the new methodologies for calculating the factor b obtain a greater return than those used up to the present time.

Results show that the classic EW method does not performance as well as the others. Cointegration, correlation and Hurst exponent give excellent results when are used to calculate factor b.

Author Contributions: Conceptualization, J.P.R.-R., J.E.T.-S. and M.Á.S.-G.; Methodology, J.P.R.-R., J.E.T.-S. and M.Á.S.-G.; Software, J.P.R.-R., J.E.T.-S. and M.Á.S.-G.; Validation, J.P.R.-R., J.E.T.-S. and M.Á.S.-G.; Formal Analysis, J.P.R.-R., J.E.T.-S. and M.Á.S.-G.; Investigation, J.P.R.-R., J.E.T.-S. and M.Á.S.-G.; Resources, J.P.R.-R., J.E.T.-S. and M.Á.S.-G.; Data Curation, J.P.R.-R., J.E.T.-S. and M.Á.S.-G.; Writing–Original Draft Preparation, J.P.R.-R., J.E.T.-S. and M.Á.S.-G.; Writing–Review & Editing, J.P.R.-R., J.E.T.-S. and M.Á.S.-G.; Visualization, J.P.R.-R., J.E.T.-S. and M.Á.S.-G.; Supervision, J.P.R.-R., J.E.T.-S. and M.Á.S.-G.; Project Administration, J.P.R.-R., J.E.T.-S. and M.Á.S.-G.; Funding Acquisition, J.P.R.-R., J.E.T.-S. and M.Á.S.-G. All authors have read and agreed to the published version of the manuscript.

Funding: Juan Evangelista Trinidad-Segovia is supported by grant PGC2018-101555-B-I00 (Ministerio Español de Ciencia, Innovación y Universidades and FEDER) and UAL18-FQM-B038-A (UAL/CECEU/FEDER). Miguel Ángel Sánchez-Granero acknowledges the support of grants PGC2018-101555-B-I00 (Ministerio Español de Ciencia, Innovación y Universidades and FEDER) and UAL18-FQM-B038-A (UAL/CECEU/FEDER) and CDTIME.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A. Stocks Portfolio Technology Sector Nasdaq 100

Table A1. The Technology Sector Nasdaq 100.

Ticker	Company
AAPL	Apple Inc.
ADBE	Adobe Systems Incorporated
ADI	Analog Devices, Inc.
ADP	Automatic Data Processing, Inc.
ADSK	Autodesk, Inc.
AMAT	Applied Materials, Inc.
ATVI	Activision Blizzard, Inc.
AVGO	Broadcom Limited
BIDU	Baidu, Inc.
CA	CA, Inc.
CERN	Cerner Corporation
CHKP	Check Point Software Technologies Ltd.
CSCO	Cisco Systems, Inc.
CTSH	Cognizant Technology Solutions Corporation
CTXS	Citrix Systems, Inc.
EA	Electronic Arts Inc.
FB	Facebook, Inc.
FISV	Fiserv, Inc.
GOOG	Alphabet Inc.
GOOGL	Alphabet Inc.
INTC	Intel Corporation
INTU	Intuit Inc.
LRCX	Lam Research Corporation
MCHP	Microchip Technology Incorporated
MSFT	Microsoft Corporation
MU	Micron Technology, Inc.
MXIM	Maxim Integrated Products, Inc.
NVDA	NVIDIA Corporation
QCOM	QUALCOMM Incorporated
STX	Seagate Technology plc
SWKS	Skyworks Solutions, Inc.
SYMC	Symantec Corporation
TXN	Texas Instruments Incorporated
VRSK	Verisk Analytics, Inc.
WDC	Western Digital Corporation
XLNX	Xilinx, Inc.

Appendix B. Empirical Results

For each model (*Equal Weight*, *Volatility*, Minimal *Distance* of the log-prices, *Correlation* of returns, *Cointegration* of the prices, lowest *Hurst* exponent of the pair), we have considered 3 scenarios, depending on the amount of pairs included in the portfolio.

1. Number of standard deviations.

Table A2. Comparison of results using the Hurst exponent selection method for the period 1999–2003 with 20 pairs and different numbers of standard deviations.

b Calculation Method	k [1]	Sharpe [2]	Profit TC [3]
Cointegration	1.0	0.39	14.55%
Cointegration	1.5	0.60	26.00%
Cointegration	2.0	0.59	24.08%
Correlation	1.0	0.15	6.10%
Correlation	1.5	0.17	8.21%
Correlation	2.0	0.31	13.82%
EW	1.0	−0.28	−11.25%
EW	1.5	0.38	15.49%
EW	2.0	0.21	7.74%
Lowest Hurst Exponent	1.0	0.70	40.51%
Lowest Hurst Exponent	1.5	0.51	28.00%
Lowest Hurst Exponent	2.0	0.57	28.51%
Minimal Distance	1.0	0.03	0.05%
Minimal Distance	1.5	0.39	15.70%
Minimal Distance	2.0	0.31	11.48%
Volatility	1.0	0.49	18.22%
Volatility	1.5	0.41	16.37%
Volatility	2.0	0.25	9.12%

[1] number of standard deviations; [2] Sharpe Ratio; [3] Profitability with transaction costs.

2. Distance (1999–2003).

Table A3. Comparison of results using the distance selection method for the period 1999–2003.

b Calculation Method	N [1]	Oper [2]	AR [3]	%Profit TC [4]	Sharpe [5]	Max Drawdown
Cointegration	10	1375	0.40%	0.72%	0.05	13.70%
Correlation	10	1357	−0.60%	−4.36%	−0.07	18.60%
EW	10	1403	−1.30%	−7.30%	−0.15	19.60%
Minimal distancie	10	1389	−0.90%	−5.49%	−0.10	16.30%
Lowest Hurst Exponent	10	1352	−1.50%	−8.55%	−0.16	19.20%
Volatility	10	1370	−1.30%	−7.57%	−0.16	13.90%
Cointegration	20	2786	3.50%	16.31%	0.47	7.40%
Correlation	20	2630	2.80%	12.68%	0.36	9.20%
EW	20	2884	1.00%	3.36%	0.14	12.30%
Minimal distancie	20	2794	2.50%	11.00%	0.34	8.30%
Lowest Hurst Exponent	20	2685	0.60%	1.66%	0.08	12.00%
Volatility	20	2812	0.40%	0.39%	0.06	8.70%
Cointegration	30	4116	2.80%	12.83%	0.42	8.00%
Correlation	30	3830	2.00%	8.62%	0.27	12.60%
EW	30	4247	1.10%	4.18%	0.18	14.80%
Minimal distancie	30	4105	1.90%	7.93%	0.28	8.40%
Lowest Hurst Exponent	30	3861	0.30%	0.01%	0.04	11.80%
Volatility	30	4160	0,10%	-0,99%	0.01	9,20%

[1] Number of pairs; [2] Number of operations; [3] Annualised return; [4] Profitability with transaction costs; [5] Sharpe Ratio.

3. Distance (2007–2012).

Table A4. Comparison of results using the distance selection method for the period 2007–2012.

b Calculation Method	N [1]	Oper [2]	AR [3]	%Profit TC [4]	Sharpe [5]	Max Drawdown
Cointegration	10	1666	1.80%	8.73%	0.35	10.20%
Correlation	10	1594	3.50%	19.51%	0.55	12.10%
EW	10	1677	1.20%	5.42%	0.22	9.40%
Minimal distance	10	1649	2.80%	15.15%	0.56	8.80%
Lowest Hurst Exponent	10	1677	2.60%	13.42%	0.51	8.00%
Volatility	10	1684	1.20%	5.22%	0.24	11.90%
Cointegration	20	3168	2.60%	13.82%	0.60	6.50%
Correlation	20	2985	3.10%	16.91%	0.58	9.30%
EW	20	3219	1.70%	8.19%	0.36	4.20%
Minimal distance	20	3172	3.10%	17.01%	0.72	6.20%
Lowest Hurst Exponent	20	3116	2.20%	11.54%	0.51	7.20%
Volatility	20	3221	2.00%	9.89%	0.48	10.00%
Cointegration	30	4714	1.50%	7.33%	0.38	6.70%
Correlation	30	4453	1.40%	6.42%	0.29	10.90%
EW	30	4791	1.40%	6.50%	0.34	5.30%
Minimal distance	30	4709	1.70%	8.43%	0.44	5.90%
Lowest Hurst Exponent	30	4545	1.40%	6.48%	0.35	6.80%
Volatility	30	4785	1.60%	7.60%	0.43	9.00%

[1] Number of pairs; [2] Number of operations; [3] Annualised return; [4] Profitability with transaction costs; [5] Sharpe Ratio.

4. Spearman Correlation (1999–2003).

Table A5. Comparison of results using the Spearman correlation selection method for the period 1999–2003.

b Calculation Method	N [1]	Oper [2]	AR [3]	%Profit TC [4]	Sharpe [5]	Max Drawdown
Cointegration	10	1274	4.10%	19.93%	0.50	14.30%
Correlation	10	1432	3.00%	13.67%	0.36	11.20%
EW	10	1400	4.30%	20.80%	0.56	9.30%
Minimal distance	10	1219	4.00%	19.68%	0.51	12.50%
Lowest Hurst Exponent	10	1103	5.70%	29.20%	0.64	10.70%
Volatility	10	1405	3.30%	15.39%	0.45	8.40%
Cointegration	20	2583	4.70%	23.41%	0.69	12.30%
Correlation	20	2833	3.90%	18.78%	0.55	10.90%
EW	20	2814	2.80%	12.69%	0.45	8.30%
Minimal distance	20	2538	4.40%	21.63%	0.65	12.50%
Lowest Hurst Exponent	20	2176	3.50%	16.71%	0.48	10.40%
Volatility	20	2781	2.50%	11.01%	0.41	9.30%
Cointegration	30	3776	4.90%	24.54%	0.79	8.10%
Correlation	30	4196	2.80%	12.90%	0.41	8.60%
EW	30	4168	0.40%	0.71%	0.08	9.90%
Minimal distance	30	3717	4.20%	20.76%	0.69	8.30%
Lowest Hurst Exponent	30	3236	4.20%	20.72%	0.56	8.20%
Volatility	30	4125	1.20%	4.52%	0.22	9.50%

[1] Number of pairs; [2] Number of operations; [3] Annualised return; [4] Profitability with transaction costs; [5] Sharpe Ratio.

5. Spearman Correlation (2007–2012).

Table A6. Comparison of results using the Spearman correlation selection method for the period 2007–2012.

b Calculation Method	N [1]	Oper [2]	AR [3]	%Profit TC [4]	Sharpe [5]	Max Drawdown
Cointegration	10	1614	1.70%	8.09%	0.38	8.30%
Correlation	10	1653	2.90%	15.75%	0.75	4.60%
EW	10	1620	1.20%	5.28%	0.37	7.00%
Minimal distancie	10	1551	2.50%	13.05%	0.58	7.80%
Lowest Hurst Exponent	10	1117	1.30%	6.18%	0.42	4.20%
Volatility	10	1668	2.30%	12.13%	0.72	5.00%
Cointegration	20	3022	1.00%	4.09%	0.29	6.80%
Correlation	20	3268	2.60%	13.77%	0.75	4.00%
EW	20	3236	1.40%	6.78%	0.46	5.70%
Minimal distance	20	2944	1.10%	4.93%	0.34	7.80%
Lowest Hurst Exponent	20	1966	0.40%	1.12%	0.14	3.90%
Volatility	20	3282	1.30%	5.76%	0.42	4.70%
Cointegration	30	4342	0.80%	3.15%	0.27	5.80%
Correlation	30	4872	2.60%	13.58%	0.74	4.30%
EW	30	4814	1.60%	7.80%	0.57	4.90%
Minimal distance	30	4222	0.90%	3.69%	0.30	7.00%
Lowest Hurst Exponent	30	2718	0.60%	2.49%	0.26	2.80%
Volatility	30	4864	1.90%	9.28%	0.67	3.60%

[1] Number of pairs; [2] Number of operations; [3] Annualised return; [4] Profitability with transaction costs; [5] Sharpe Ratio.

6. Cointegration (1999–2003).

Table A7. Comparison of results using the cointegration selection method for the period 1999–2003.

b Calculation Method	N [1]	Oper [2]	AR [3]	%Profit TC [4]	Sharpe [5]	Max Drawdown
Cointegration	10	998	5.30%	26.80%	0.58	12.40%
Correlation	10	1015	7.30%	39.38%	0.78	9.30%
EW	10	1369	4.30%	20.83%	0.41	10.30%
Minimal distance	10	945	4.00%	19.45%	0.47	9.40%
Lowest Hurst Exponent	10	1123	7.70%	41.68%	0.79	11.90%
Volatility	10	1376	6.90%	36.62%	0.68	11.00%
Cointegration	20	1984	5.50%	28.41%	0.78	9.00%
Correlation	20	1985	5.50%	28.31%	0.76	6.40%
EW	20	2718	2.90%	13.24%	0.36	9.90%
Minimal distance	20	1876	4.50%	22.36%	0.67	8.10%
Lowest Hurst Exponent	20	2031	6.90%	36.88%	0.90	7.70%
Volatility	20	2688	4.10%	19.76%	0.50	11.50%
Cointegration	30	2957	0.90%	3.51%	0.14	11.00%
Correlation	30	3132	2.40%	11.06%	0.36	10.30%
EW	30	4064	−0.10%	−1.85%	−0.01	12.20%
Minimal distance	30	2783	0.70%	2.67%	0.12	9.40%
Lowest Hurst Exponent	30	2924	3.60%	17.23%	0.50	7.50%
Volatility	30	4040	0.90%	3.25%	0.12	13.00%

[1] Number of pairs; [2] Number of operations; [3] Annualised return; [4] Profitability with transaction costs; [5] Sharpe Ratio.

7. Cointegration (2007–2012).

Table A8. Comparison of results using the cointegration selection method for the period 2007–2012.

b Calculation Method	N [1]	Oper [2]	AR [3]	%Profit TC [4]	Sharpe [5]	Max Drawdown
Cointegration	10	1516	−1.00%	−6.82%	−0.19	15.50%
Correlation	10	1512	−0.10%	−2.11%	−0.02	14.70%
EW	10	1604	1.40%	6.80%	0.30	9.60%
Minimal distance	10	1478	0.70%	2.32%	0.12	12.50%
Lowest Hurst Exponent	10	1502	−1.60%	−9.90%	−0.32	10.90%
Volatility	10	1635	−0.10%	−2.14%	−0.02	12.90%
Cointegration	20	2884	−0.70%	−5.44%	−0.19	9.90%
Correlation	20	2955	−0.70%	−5.28%	−0.16	11.70%
EW	20	3195	1.80%	8.90%	0.48	4.40%
Minimal distance	20	2709	0.20%	−0.15%	0.06	9.50%
Lowest Hurst Exponent	20	2666	−0.90%	−6.53%	−0.26	9.00%
Volatility	20	3189	0.50%	1.31%	0.14	8.90%
Cointegration	30	4142	0.00%	−1.38%	0.00	8.80%
Correlation	30	4373	0.20%	−0.56%	0.04	9.50%
EW	30	4720	2.70%	14.63%	0.75	4.90%
Minimal distance	30	3923	1.10%	4.69%	0.28	7.90%
Lowest Hurst Exponent	30	3694	−0.30%	−2.93%	−0.09	7.60%
Volatility	30	4742	1.30%	5.82%	0.36	7.00%

[1] Number of pairs; [2] Number of operations; [3] Annualised return; [4] Profitability with transaction costs; [5] Sharpe Ratio.

8. Hurst exponent (1999–2003).

Table A9. Comparison of results using the Hurst exponent selection method for the period 1999–2003.

b Calculation Method	N [1]	Oper [2]	AR [3]	%Profit TC [4]	Sharpe [5]	Max Drawdown
Cointegration	10	1136	−0.60%	−3.94%	−0.06	15.60%
Correlation	10	1176	0.50%	1.32%	0.04	24.40%
EW	10	1334	2.80%	12.87%	0.29	12.20%
Minimal distance	10	1166	−1.20%	−6.87%	−0.12	13.50%
Lowest Hurst Exponent	10	1234	4.60%	22.77%	0.37	21.40%
Volatility	10	1401	7.40%	39.60%	0.72	14.40%
Cointegration	20	2104	3.10%	14.55%	0.39	11.10%
Correlation	20	2400	1.50%	6.10%	0.15	15.70%
EW	20	2695	-2.10%	−11.25%	−0.28	12.30%
Minimal distance	20	2093	0.20%	0.05%	0.03	10.60%
Lowest Hurst Exponent	20	2375	7.50%	40.51%	0.70	17.10%
Volatility	20	2755	3.80%	18.22%	0.49	8.90%
Cointegration	30	2984	3.10%	14.91%	0.48	7.80%
Correlation	30	3516	2.00%	8.83%	0.22	16.50%
EW	30	4066	−1.30%	−7.56%	−0.19	11.80%
Minimal distance	30	2915	2.70%	12.63%	0.41	6.50%
Lowest Hurst Exponent	30	3411	7.10%	37.86%	0.78	13.40%
Volatility	30	3994	4.40%	21.57%	0.63	6.50%

[1] Number of pairs; [2] Number of operations; [3] Annualised return; [4] Profitability with transaction costs; [5] Sharpe Ratio.

9. Hurst exponent (2007–2012).

Table A10. Comparison of results using the Hurst exponent selection method for the period 2007–2012.

b Calculation Method	N [1]	Oper [2]	AR [3]	%Profit TC [4]	Sharpe [5]	Max Drawdown
Cointegration	10	1596	3.00%	16.40%	0.55	8.00%
Correlation	10	1587	3.70%	21.51%	0.57	9.80%
EW	10	1643	3.00%	16.26%	0.59	9.10%
Minimal distancie	10	1581	2.10%	11.02%	0.41	8.70%
Lowest Hurst Exponent	10	1649	4.80%	28.15%	0.83	8.30%
Volatility	10	1724	1.30%	5.98%	0.27	8.40%
Cointegration	20	2795	0.80%	3.40%	0.21	7.10%
Correlation	20	3001	2.60%	14.10%	0.50	7.50%
EW	20	3258	1.70%	8.27%	0.40	9.10%
Minimal distancie	20	2758	−0.40%	−3.48%	−0.10	10.60%
Lowest Hurst Exponent	20	3129	1.90%	9.84%	0.39	8.30%
Volatility	20	3204	0.40%	0.90%	0.11	6.40%
Cointegration	30	4100	−0.20%	−2.27%	−0.05	9.30%
Correlation	30	4418	2.00%	10.23%	0.43	8.10%
EW	30	4666	1.90%	9.34%	0.46	7.60%
Minimal distancie	30	4049	0.00%	−1.55%	−0.01	10.50%
Lowest Hurst Exponent	30	4248	0.40%	0.78%	0.09	9.20%
Volatility	30	4790	0.60%	1.50%	0.15	8.40%

[1] Number of pairs; [2] Number of operations; [3] Annualised return; [4] Profitability with transaction costs; [5] Sharpe Ratio.

References

1. Fama, E. Efficient capital markets: II. *J. Financ.* **1991**, *46*, 1575–1617. [CrossRef]
2. Dimson, E.; Mussavian, M. A brief history of market efficiency. *Eur. Financ. Manag.* **1998**, *4*, 91–193. [CrossRef]
3. Di Matteo, T.; Aste, T.; Dacorogna, M.M. Using the scaling analysis to characterize their stage of development. *J. Bank. Financ.* **2005**, *29*, 827–851. [CrossRef]
4. Di Matteo, T. Multi-scaling in finance. *Quant. Financ.* **2007**, *7*, 21–36. [CrossRef]
5. Peters, E.E. *Chaos and Order in the Capital Markets. A New View of Cycle, Prices, and Market Volatility*; Wiley: New York, NY, USA, 1991.
6. Peters, E.E. *Fractal Market Analysis: Applying Chaos Theory to Investment and Economics*; Wiley: New York, NY, USA, 1994.
7. Lo, A.W. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *J. Portf. Manag.* **2004**, *30*, 15–29. [CrossRef]
8. Sanchez-Granero, M.A.; Trinidad Segovia, J.E.; García, J.; Fernández-Martínez, M. The effect of the underlying distribution in Hurst exponent estimation. *PLoS ONE* **2015**, *28*, e0127824.
9. Sánchez-Granero, M.A.; Balladares, K.A.; Ramos-Requena, J.P.; Trinidad-Segovia, J.E. Testing the efficient market hypothesis in Latin American stock markets. *Phys. A Stat. Mech. Its Appl.* **2020**, *540*, 123082. [CrossRef]
10. Zhang, H.; Urquhart, A. Pairs trading across Mainland China and Hong Kong stock Markets. *Int. J. Financ. Econ.* **2019**, *24*, 698–726. [CrossRef]
11. Ramos-Requena, J.P.; Trinidad-Segovia, J.E.; Sanchez-Granero, M.A. Introducing Hurst exponent in pairs trading. *Phys. A Stat. Mech. Its Appl.* **2017**, *488*, 39–45. [CrossRef]
12. Taqqu, M.S.; Teverovsky, V. Estimators for long range dependence: An empirical study. *Fractals* **1995**, *3*, 785–798. [CrossRef]
13. Kristoufek, L.; Vosvrda, M. Measuring capital market efficiency: Global and local correlations structure. *Phys. A Stat. Mech. Its Appl.* **2013**, *392*, 184–193. [CrossRef]
14. Kristoufek, L.; Vosvrda, M. Measuring capital market efficiency: Long-term memory, fractal dimension and approximate entropy. *Eur. Phys. J. B* **2014**, *87*, 162. [CrossRef]

15. Zunino, L.; Zanin, M.; Tabak, B.M.; Pérez, D.G.; Rosso, O.A. Complexity-entropy causality plane: A useful approach to quantify the stock market inefficiency. *Phys. A Stat. Mech. Its Appl.* **2010**, *389*, 1891–1901. [CrossRef]
16. Gatev, E.; Goetzmann, W.; Rouwenhorst, K. Pairs Trading: Performance of a relative value arbitrage rule. *Rev. Financ. Stud.* **2006**, *19*, 797–827. [CrossRef]
17. Vidyamurthy, G. *Pairs Trading, Quantitative Methods and Analysis*; John Wiley and Sons: Toronto, ON, Canada, 2004.
18. Dunis, L.; Ho, R. Cointegration portfolios of European equities for index tracking and market neutral strategies. *J. Asset Manag.* **2005**, *6*, 33–52. [CrossRef]
19. Figuerola Ferretti, I.; Paraskevopoulos, I.; Tang, T. Pairs trading and spread persistence in the European stock market. *J. Futur. Mark.* **2018**, *38*, 998–1023 [CrossRef]
20. Alexander, C.; Dimitriu, A. *The Cointegration Alpha: Enhanced Index Tracking and Long-Short Equity Market Neutral Strategies*; SSRN eLibrary: Rochester, NY, USA, 2002.
21. Perlin, M.S. Evaluation of Pairs Trading strategy at the Brazilian financial market. *J. Deriv. Hedge Funds* **2009**, *15*, 122–136. [CrossRef]
22. Caldeira, J.F.; Moura, G.V. *Selection of a Portfolio of Pairs Based on Cointegration: A Statistical Arbitrage Strategy*; Revista Brasileira de Financas (Online): Rio de Janeiro, Brazil, 2013; Volume 11, pp. 49–80.
23. Burgess, A.N. Using cointegration to hedge and trade international equities. In *Applied Quantitative Methods for Trading and Investment*; John Wiley and Sons: Chichester, UK, 2003; pp. 41–69.
24. Galenko, A.; Popova, E.; Popova, I. Trading in the presence of cointegration. *J. Altern. Investments* **2012**, *15*, 85–97. [CrossRef]
25. Lin, Y.X.; Mccrae, M.; Gulati, C. Loss protection in Pairs Trading through minimum profit bounds: A cointegration approach. *J. Appl. Math. Decis. Sci.* **2006**. [CrossRef]
26. Nath, P. *High frequency Pairs Trading with U.S Treasury Securities: Risks and Rewards for Hedge Funds*; SSRN eLibrary: Rochester, NY, USA, 2003.
27. Elliott, R.; van der Hoek, J.; Malcolm, W. Pairs Trading. *Quant. Financ.* **2005**, *5*, 271–276. [CrossRef]
28. Dunis, C.L.; Shannon, G. Emerging markets of South-East and Central Asia: Do they still offer a diversification benefit? *J. Asset Manag.* **2005**, *6*, 168–190. [CrossRef]
29. Ramos-Requena, J.P.; Trinidad-Segovia, J.E.; Sánchez-Granero, M.A. An Alternative Approach to Measure Co-Movement between Two Time Series. *Mathematics* **2020**, *8*, 261. [CrossRef]
30. Baur, D. What Is Co-movement? In *IPSC-Technological and Economic Risk Management*; Technical Report; European Commission, Joint Research Center: Ispra, VA, Italy, 2003.
31. Hauke, J.; Kossowski, T. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaest. Geogr.* **2011**, *30*, 87–93. [CrossRef]
32. Engle, R.F.; Granger, C.W.J. Co-integration and error correction: Representation, estimation, and testing. *Econometrica* **1987**, *5*, 251–276. [CrossRef]
33. Do, B.; Faff, R. Does simple pairs trading still work? *Financ. Anal. J.* **2010**, *66*, 83–95. [CrossRef]
34. Do, B.; Faff, R. Are pairs trading profits robust to trading costs? *J. Financ. Res.* **2012**, *35*, 261–287. [CrossRef]
35. Hurst, H. Long term storage capacity of reservoirs. *Trans. Am. Soc. Civ. Eng.* **1951**, *6*, 770–799.
36. Mandelbrot, B.; Wallis, J.R. Robustness of the rescaled range R/S in the measurement of noncyclic long-run statistical dependence. *Water Resour. Res.* **1969**, *5*, 967–988. [CrossRef]
37. Peng, C.K.; Buldyrev, S.V.; Havlin, S.; Simons, M.; Stanley, H.E.; Goldberger, A.L. Mosaic organization of DNA nucleotides. *Phys. Rev. E* **1994**, *49*, 1685–1689. [CrossRef]
38. Lo, A.W. Long-term memory in stock market prices. *Econometrica* **1991**, *59*, 1279–1313. [CrossRef]
39. Sanchez-Granero, M.A.; Trinidad-Segovia, J.E.; Garcia-Perez, J. Some comments on Hurst exponent and the long memory processes on capital markets. *Phys. A Stat. Mech. Its Appl.* **2008**, *387*, 5543–5551. [CrossRef]
40. Weron, R. Estimating long-range dependence: Finite sample properties and confidence intervals. *Phys. A Stat. Mech. Its Appl.* **2002**, *312*, 285–299. [CrossRef]
41. Willinger, W.; Taqqu, M.S.; Teverovsky, V. Stock market prices and long-range dependence. *Financ. Stochastics* **1999**, *3*, 1–13. [CrossRef]
42. Barabasi, A.L.; Vicsek, T. Multifractality of self affine fractals. *Phys. Rev. A* **1991**, *44*, 2730–2733. [CrossRef]
43. Barunik, J.; Kristoufek, L. On Hurst exponent estimation under heavy-tailed distributions. *Phys. A Stat. Mech. Its Appl.* **2010**, *389*, 3844–3855. [CrossRef]

44. Goulart Coelho, L.M.; Lange, L.C.; Coelho, H.M. Multi-criteria decision making to support waste management: A critical review of current practices and methods. *Waste Manag. Res.* **2017**, *35*, 3–28. [CrossRef]
45. Meng, K.; Cao, Y.; Peng, X.; Prybutok, V.; Gupta, V. Demand-dependent recovery decision-making of a batch of products for sustainability. *Int. J. Prod. Econ.* **2019**, 107552. [CrossRef]
46. Roth, S.; Hirschberg, S.; Bauer, C.; Burgherr, P.; Dones, R.; Heck, T.; Schenler, W. Sustainability of electricity supply technology portfolio. *Ann. Nucl. Energy* **2009**, *36*, 409–416. [CrossRef]

mathematics

MDPI

Article

Exploring the Link between Academic Dishonesty and Economic Delinquency: A Partial Least Squares Path Modeling Approach

Elena Druică [1,*], Călin Vâlsan [2], Rodica Ianole-Călin [1], Răzvan Mihail-Papuc [1] and Irena Munteanu [3]

1 Faculty of Business and Administration, University of Bucharest, 030018 Bucharest, Romania; rodica.ianole@faa.unibuc.ro (R.I.-C.) razvanmihail.papuc@faa.unibuc.ro (R.M.-P.)
2 Williams School of Business, Bishop's University, Sherbrooke, QC J1M1Z7, Canada; cvalsan@ubishops.ca
3 Faculty of Economic Sciences, Ovidius University, 900470 Constanta, Romania; irena.munteanu@365.univ-ovidius.ro
* Correspondence: elena.druica@faa.unibuc.ro

Received: 13 November 2019; Accepted: 11 December 2019; Published: 15 December 2019

Abstract: This paper advances the study of the relationship between the attitude towards academic dishonesty and other types of dishonest and even fraudulent behavior, such as tax evasion and piracy. It proposes a model in which the attitudes towards two types of cheating and fraud are systematically analyzed in connection with a complex set of latent construct determinants and control variables. It attempts to predict the tolerance towards tax evasion and social insurance fraud and piracy, using academic cheating as the main predictor. The proposed model surveys 504 student respondents, uses a partial least squares—path modeling analysis, and employs two subsets of latent constructs to account for context and disposition. The relationship between the outcome variable and the subset of predictors that account for context is mediated by yet another latent construct—Preoccupation about Money—that has been shown to strongly influence people's attitude towards a whole range of social and economic behaviors. The results show academic dishonesty is a statistically significant predictor of an entire range of unethical and fraudulent behavior acceptance, and confirm the role played by both contextual and dispositional variables; moreover, they show that dispositional and contextual variables tend to be segregated according to how they impact the outcome. They also show that money priming does not act as a mediator, in spite of its stand-alone impact on the outcome variables. The most important result, however, is that the effect size of the main predictor is large. The contribution of this paper is two-fold: it advances a line of research previously sidestepped, and it proposes a comprehensive and robust model with a view to establish a hierarchy of significance and effect size in predicting deviance and fraud. Most of all, this research highlights the central role played by academic dishonesty in predicting the acceptance of any type of dishonest behavior, be it in the workplace, at home, or when discharging one's responsibilities as a citizen. The results presented here give important clues as to where to start intervening in order to discourage the acceptance of deviance and fraud. Educators, university professors, and academic administrators should be at the forefront of targeted campaigns and policies aimed at fighting and reducing academic dishonesty.

Keywords: academic cheating; tax evasion; informality

1. Introduction

Academic cheating and workplace cheating are like two peas in a pod. Those who engage in dishonest behavior during university are more likely to lie, cheat, and steal later on during their professional career [1,2]. These findings echo the widely-held belief that early cheating in school

and university is a good predictor of dishonesty in the workplace later on. The relationship among academic dishonesty, workplace unethical behavior, and integrity standards is well documented [3].

If people cheat in school, and they cheat later on, at work, why would dishonesty, fraud, and cheating stop there? It makes sense to assume that once people become accustomed to disrespecting academic integrity, they might show lack of integrity in every aspect of their lives, be it at work, at home, or in public. However, there is very little academic research in this direction. This issue has received little attention and the evidence is still sketchy. This paper takes the relationship between the attitude towards academic dishonesty and other types of dishonest and even fraudulent behavior one step further. It proposes a study in which the attitudes towards two types of cheating and fraud are systematically analyzed in connection with a complex set of latent construct determinants and control variables. It attempts to predict the tolerance towards tax evasion and social insurance fraud, and piracy using academic cheating as the main predictor. The proposed model uses a subset of latent constructs that account for context, and another subset to account for disposition. The relationship between the outcome variable and the subset of predictors that account for context is mediated by yet another latent construct—Preoccupation about Money—that has been shown to strongly influence people's attitude towards a whole range of social and economic behaviors.

The contribution of this paper is two-fold: it advances a line of research previously sidestepped, and it proposes a comprehensive and robust model with a view to establishing a hierarchy of significance and effect size in predicting deviance and fraud. As it will be shown here, it turns out that the significance and the effect size of the main predictor, i.e., academic dishonesty, is huge compared with the rest of the other predictors.

Our results have theoretical as well as practical implications. The results confirm the role played by both contextual and dispositional variables; moreover, they show that dispositional and contextual variables tend to be segregated according to how they impact the outcome [3]. They also show that money priming does not act as a mediator in spite of its stand-alone impact on the outcome variables. Most of all, they buttress the important role played by academic dishonesty in predicting the acceptance of any type of dishonest behavior, be it in the workplace, at home, or when discharging one's responsibilities as a citizen.

The results presented here give important clues as to where to start intervening in order to discourage the acceptance of deviance and fraud. By far, the first choice should be tackling the issue of academic dishonesty. There is a respectable body of evidence showing that ethical education and moral sensitivity training tend to work in lowering the acceptance and incidence of academic cheating [2,4,5]. Educators, university professors, and academic administrators should be at the forefront of targeted campaigns and policies aimed at fighting and reducing academic dishonesty.

A word of caution is in order: the results show where to direct the intervention in order to reduce the acceptance of unethical behavior and fraud. The unambiguous course of action is to tackle academic dishonesty, but the findings do not illuminate the manner in which this should be done, which is a different matter altogether and not the subject of the current paper.

The paper is organized as follows. The next section discusses the literature, while Section three presents the sample of respondents and the methodology used to construct the latent variables. Section four presents the results of the partial least squares – path model (PLS-PM henceforth). Section five provides a brief discussion and interpretation of the findings. Section 6 concludes the paper.

2. Literature Review

It is truly worrisome that academic cheating is more prevalent than most people think. It seems that almost all university students have engaged in or witnessed academic dishonesty at some point during their studies [6,7]. There are many studies using a diverse methodology, including but not limited to surveys, factor analysis, structural equation modelling, and cross-lagged regressions, which show how truly widespread academic cheating is. They also show how academic cheating is linked to

workplace cheating. Nursing students cheat in large proportion [8–10]; engineering students cheat [11]; psychology students cheat [12]; IT students cheat [13]; and business students seem to outdo everyone else; they have the worst reputation for academic integrity standards among all other students [14–16]

More recently, other researchers [17] found that personality traits defined according to the reinforcement sensitivity theory (RST) are predictors for the extent to which individuals engage in academic dishonesty. Among these predictors, impulsivity and Fight–Flight–Freeze behaviors appear to play an important role.

Other authors [18] administered the Academic Honesty Scale, the Brief Self-Control Scale, the Social Success Index, the Normalcy Feeling Scale, the Social Comparison Scale, and the Satisfaction With Life Scale to a sample of 631 Polish respondents, and found self-regulation to be inversely related to academic cheating. The study also found a gender gap when engaging in academic cheating. Moreover, social comparison appears to be directly related to plagiarism.

Yet, in other cases [19] the investigation was based on a cross-lagged model to describe a complex dynamic between academic cheating on the one hand, and regulatory self-efficacy and moral disengagement on the other hand. The authors found a negative relationship between academic cheating and regulatory self-efficacy, and a positive relationship between academic cheating and moral disengagement. The results are not surprising; moral disengagement encourages academic dishonesty, which in turn legitimizes a shift in attitudes towards cheating. This repositioning leads to an increasing acceptance of cheating as a normal state of affairs.

Using a survey of 185 faculty and 295 students, other authors found a significant difference in perception and attitudes between students and faculty when it comes to assessing the consequences and implication of academic dishonesty [20].

There is a rich literature on the link between academic cheating and workplace deviance. However, there are almost no studies linking academic cheating to economic delinquency beyond the workplace. In [21], the authors argue that individuals who are more tolerant of academic dishonesty also tend to be less reliable, engage in risky behaviors, and accept more readily illegal behaviors. They are also among the very few who link academic dishonesty to dishonesty in politics, athletics, and even tax evasion [22]. This current paper differs from previous research because it takes over where [22] left, and concentrates on tax evasion and piracy, using a well-rounded system of predictors, among which academic cheating represents the most important element.

3. Materials and Methods

The data consists of 504 respondents aged 18–25 (mean 19.82, SD = 1.55), all of whom are students from various national university centers. The distribution is 74.2% females and 25.8% males, with 71.3% originating from urban areas and 28.7% from rural areas. The data was collected via an online questionnaire applied between March and April 2019. Participation in the study was voluntarily.

There are two versions of the same variance-based, structural equation model. There are eight latent variables, and four observed variables. The difference between the two models is in the outcome variable. The first version uses tax evasion and social insurance fraud acceptance as the outcome, while the second version uses piracy acceptance as the outcome. Both versions of the model have ten predictors and one mediator variable.

The items used in the measurement of tax evasion, social insurance fraud, and piracy are taken from a survey on attitudes and behavior towards tax evasion and compliance implemented in Ireland in 2008/2009 by the Office of the Revenue Commissioners (the government body responsible for tax administration and customs regime) [23]. The original survey question contained 14 items, but following an exploratory factor analysis, only six items were retained, as shown in Table 1.

Table 1. Six questionnaire items and two latent (outcome) variables: tax evasion & social insurance fraud acceptance, and piracy acceptance.

	Items [23]	Dimension
ATT1	To claim credits or tax/payment reliefs that you are not entitled to	Tax Evasion and Social Insurance Fraud Acceptance (TESIFA)
ATT2	To deliberately not pay the taxes you are supposed to pay	
ATT3	To deliberately claim state social benefits that you are not entitled to	
ATT4	To knowingly buy counterfeit goods (e.g., clothing, handbags)	Piracy Acceptance (PA)
ATT5	To knowingly buy pirated goods (e.g., books, CDs, DVDs)	
ATT6	To use a computer software without having a valid license for it	

The attitude toward money scale developed by Lim and Teo [24] is used to extract the first of the four observed predictors: "I believe that a person's pay is related to their ability and effort", is deemed as "fairness" (FAIR). The second observed predictor is taken from the money scale developed by Yamauchi and Templer [25]. The item is part of a broader measure of power, but here, "money makes people respect you" is assigned to a variable deemed "money as social status" (MASS). It can be argued that a stronger belief in fairness would result in a weaker acceptance of any deviant or fraudulent behavior [26]. On the other hand, one expects that a greater need to express social status through a display of wealth is associated with more cynicism, a stronger sense of entitlement, and eventually more acceptance of cheating and fraudulent behavior [27]. It could even be the case that cheating becomes a compulsive behavior when driven by social status [28].

The third and fourth predictors are two latent variables called hard work as the path to achievement (HAWPACH) and valuing leisure time (VLT), adapted from Mudrack and McHoskey [29,30]. Both predictors contain items belonging to the work attitudes construct, and are taken directly from the Protestant ethics scale [31]. In the original work, the scale used to measure valuing leisure time is reversed, in an attempt to capture negative judgment against idleness [32]. Here, a 1–7 Likert scale is used, where 1 corresponds to "total disagreement", and 7 means "total agreement," to ensure higher scores are associated with higher valuation of leisure time (Table 2). It is expected that a strong predilection for hard work would lower the acceptance of any type of cheating or fraudulent behavior [33]. There is no prior expectation about the relationship between VLT and the two outcome variables.

Table 2. The structure of two latent variable predictors: hard work as the path to achievement (HAWPACH) and valuing leisure time (VLT).

Items from the Protestant Ethic Scale [31]		
Item	**Latent Variable**	**Manifest Variable**
WORK1	Hard Work as the Path to Achievement (HAWPACH) [30]	Any person who is able and willing to work hard has a good chance of succeeding
WORK2		If one works hard enough they are likely to make a good life for themselves
WORK3	Valuing leisure time (VLT) [30]	People should have more leisure time to spend in relaxation
WORK4		Life would be more meaningful if we had more leisure time

Academic dishonesty (ACADISH), the main predictor, is a latent variable using only two items [34]: "cheating during an exam in order to obtain a better grade" (DIS1), and "cheating during an exam in order to obtain a passing grade" (DIS2). There is a difference between these two instances stemming from the nature of the consequences. Cheating in order to obtain a passing grade seems to be perceived as more acceptable because it thwarts failure [13]. On the other hand, cheating to merely get a good grade tends to be perceived as less acceptable [35].

The model adapts the Rosenberg self-esteem scale to the context of the current survey [36–38]. The scale is one-dimensional and has 10 items. The original measurement is on a 1–4 Likert scale,

where items 2, 5, 6, 8, 9 have their scores reversed, and the total score is obtained by summing up the results of each item. Here, a 1–7 Likert scale is used (1—"total disagreement", 7—"total agreement"). Higher scores are an indication of higher levels of self-esteem. Factor analysis reveals the presence of two latent variables labeled, "positive feelings" (POF), and "negative feelings" (NEF), presented in Table 3. The two latent variables show a negative Pearson's correlation coefficient of 50%. While some findings suggest that the importance of self-esteem as a determinant for a wide array of behavior types has been overstated [36], POF and NEF are included in the model as two distinct dispositional control variables [39].

Table 3. Two latent predictors derived from the self-esteem scale.

	Self-Esteem Scale [36]	**Latent Variables—Feelings**	
EST1	On the whole, I am satisfied with myself.	Positive	
EST2	At times I think I am no good at all.		Negative
EST3	I feel that I have a number of good qualities.	Positive	
EST4	I am able to do things as well as most other people.	Positive	
EST5	I feel I do not have much to be proud of.		Negative
EST6	I certainly feel useless at times.		Negative
EST7	I feel that I'm a person of worth, at least on an equal plane with others.	Positive	
EST8	I wish I could have more respect for myself.		Negative
EST9	All in all, I am inclined to feel that I am a failure.		Negative
EST10	I take a positive attitude toward myself.	Positive	

Self-efficacy (SELFEFF) is measured using the 10-item general self-efficacy scale [40]. Here, a 1–7 Likert scale is also used (1—"Not at all true", 7—"Exactly true"). There are 10 items (SELFEFF1–SELFEFF10) resulting in a single latent variable, following an exploratory parallel analysis based on maximum likelihood extraction. Cronbach's Alpha shows very good internal consistency at 0.92 and cannot be increased any further. The variance explained by this factor is 55.2%. Since self-efficacy is most often associated with an internal locus of control, one would expect to find a direct relationship between this latent construct and the acceptance of unethical and fraudulent behavior [41–43].

Preoccupation with money (PFM) is a latent variable predictor, based on the items presented in Table 4, and introduced as a mediator. It was extracted from the attitude toward money scale, and corresponds to one of the four dimensions found in the original study [44]. Money represents a powerful extrinsic motivator, and other studies have already found that it plays an important mediating role [45].

Money priming increases the acceptance of interactions based predominantly on market transactions at the expense of other types of social interaction. As such, money makes respondents endorse steeper hierarchical economic systems more readily. Because wealth and status are perceived as a reward for focusing on money and market transactions, money priming reduces the level of empathy and compassion towards more disadvantaged categories [27]. When reminded about money, individuals shift their frame of mind to a modus operandi in which efficiency and results take precedence over all other considerations [46].

It is expected that preoccupation with money is likely to increase the acceptance of cheating and fraudulent behavior for at least two reasons. When framed in terms of eliciting results and achieving performance measured in monetary terms, the focus of individuals is funneled towards obtaining the required results, while other contextual concerns fade into the background [47]. On the other hand, exposure to money and wealth makes people feel more entitled, and this is bound to increase the likelihood of engaging in, or more easily accepting unethical behavior [48,49].

Table 4. The latent variable "preoccupation with money".

Item	Latent Variable	Manifest Variable
PFM1	Preoccupation with money [44]	Compared to people I know, I believe I think about money more than they do.
PFM2		I often fantasize about money and what I can do with it.
PFM3		Most of my friends have more money than I do.
PFM4		Money is the most important thing in my life.

Age represents a commonly used control variable. The segment of young adults used in the current study is relevant when exploring the relationships between money attitudes and materialism [50], and relationships among money attitudes, credit card usage, and compulsive buying [28,51,52]. Given the relatively narrow range of this observed variable, we do not expect to find a statistically significant effect. We include it, nevertheless, for the sake of following a consecrated methodology.

Gender represents another commonly used control variable. Previous research finds that money is less important for women than for men [53]; however, this finding has to be qualified by cross-cultural research, taking into account the role played by women in the financial management of the household [54]. This qualifier notwithstanding, it is expected that men are more likely to accept dishonest and fraudulent behavior than women.

Table 5 summarizes the latent variable predictors, constructed as a weighted average of their corresponding manifest variables [55]. Figure 1 presents the research model.

Table 5. A summary of latent predictors, with abbreviations and descriptors.

Latent Structure	Observed Variables
NEF	Negative feelings. Part of the self-esteem scale, the items capture negative feelings toward oneself: EST2, EST5, EST6, EST8, EST9
POF	Positive feelings. Part of the self-esteem scale, the items capture positive feelings toward oneself: EST1, EST3, EST4, EST7, EST10
HAWPACH	Hard work as the path to achievement. Hard work provides ground for success in life: WORK1, WORK2
VLT	Valuing leisure time. Appreciation for leisure time: WORK3, WORK4
ACADISH	Academic dishonesty. Motivators for academic cheating: DIS1, DIS2
SELFEFF	Self-efficacy. The level of self-efficacy: SELFEFF1–SELFEFF10
PFM	Preoccupation with money. Importance of money and fantasies around them: MON1, MON2, MON3, MON4
TESIFA	Tax evasion and social insurance fraud acceptance. The level of acceptance of active rule bending: ATT1, ATT2, ATT3
PA	Piracy acceptance. The level of piracy acceptance: ATT4, ATT5, ATT6

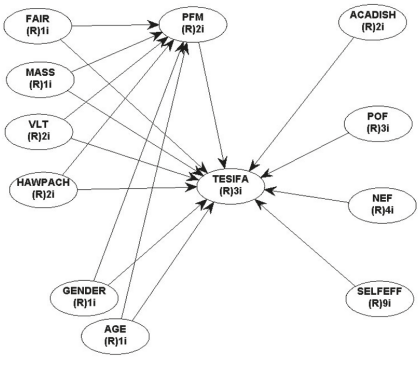


Figure 1. The research model.

Several items were subsequently dropped, either to increase internal consistency or to maintain factor loadings above 0.7. Eventually, MON3 and MON4 (preoccupation with money), EST3 and EST8 (self-esteem), and EFF2 and EFF3 (self-efficacy) were dropped from the final version of the model.

4. Results

The PLS-PM analysis is aimed at maximizing the explained variance of the dependent, endogenous latent variable [56]. The two outcome variables used here are tax evasion and social insurance fraud acceptance (TESIFA), and piracy acceptance (PA). Preoccupation with money (PFM) serves as a mediator in the relationship between contextual and dispositional constructs and the outcome variables. Academic dishonesty (ACADISH) represents the main predictor.

At its core, the estimation method is an iterative algorithm based on ordinary least squares (OLS). Any PLS-PM model consists of two parts: an outer, or measurement model, and an inner, or structural model. The outer model estimates the relationship between the latent constructs and their respective indicator manifest variables, assessed in terms of composite indices. The inner model estimates the relationships among the latent variables themselves.

One begins by estimating the model using R software version 3.4.3, with the "plspm" and the "plsdepot" packages. Subsequently, the results are cross-checked using WarpPLS software version 6.0. The statistical inference part is based on bootstrapping with 999 repetitions.

Table 6 provides the reliability results for the measurement model and shows the robustness of the measures. The composite reliability results range from 0.828 to 0.936 (self-efficacy). These values are above the threshold of 0.7 recommended in the literature [57]. Some alpha values are in the moderate range ("valuing leisure time" and "preoccupation with money"), but one might take the view that even an alpha of 0.5 or 0.6 could be acceptable, in particular when the number of scale items is small [55,58]. Also, the results are considered to be relevant when the average variance extracted (AVE) for each individual latent construct exceeds 0.50, a requirement that in this case is met across the board.

Table 6. Assessment of the measurement model.

Variable	Abbreviation	Composite Reliability (Dillon Goldstein rho)	Cronbach's Alpha	Average Variance Extracted (AVE)
Negative feelings	NEF	0.889	0.833	0.668
Positive feelings	POF	0.897	0.827	0.743
Valuing leisure time	VLT	0.867	0.692	0.765
Hard Work as the Path to Achievement	HAWPACH	0.873	0.710	0.775
Self-efficacy	SELFEFF	0.936	0.923	0,621
Preoccupation with money	PFM	0.828	0.583	0.706
Academic Dishonesty	ACADISH	0.955	0.906	0.914
Tax Evasion and Social Insurance Fraud	TESIFA	0.883	0.824	0.654
Piracy acceptance	PA	0.851	0.736	0.656

Table 7 shows the square roots of all the AVEs (the diagonal elements of the inter-correlation matrix) to be greater than the off-diagonal elements in their corresponding rows and columns. In addition, the off-diagonal correlations are all below the threshold value of 0.8 recommended by Kennedy [59].

Table 7. Discriminant validity (inter-correlations) of variable constructs.

Variable	NEF	POF	VLT	HAWPACH	SELFEFF	PFM	ACADISH	TESIFA	PA
NEF	0.817	0.593	−0.028	0.163	0.395	−0.114	−0.130	−0.272	−0.195
POF		0.862	0.081	0.288	0.583	0.030	−0.021	−0.096	−0.073
VLT			0.874	0.145	0.127	0.136	0.084	0.049	0.143
HAWPACH				0.881	0.347	0.031	−0.087	−0.197	−0.193
SELFEFF					0.788	0.112	−0.049	−0.224	−0.073
PFM						0.840	0.249	0.190	0.275
ACADISH							0.956	0.392	0.475
TESIFA								0.809	–
PA									0.810

Table 8 presents the loadings and cross-loadings of all manifest variables. All loadings are higher than 0.7, ranging from 0.708 to 0.874. It is easy to notice how clusters of indicators uniquely circumscribe each latent construct, with factor loadings of 0.7 or higher, and with high statistical significance ($p < 0.001$). At the same time, same-construct item loadings are higher than cross-construct loadings. This fact confirms the convergent validity of these indicators and suggests that they group into distinct latent constructs.

Table 8. Convergent validity (inter-correlations) of variable constructs.

Variable	NEF	POF	VLT	HAWPACH	SELFEFF	PFM	ACADISH	TESIFA	PA
EST2	0.854	−0.034	−0.059	−0.044	−0.005	−0.032	−0.010	0.066	0.008
EST5	0.722	0.016	0.049	0.041	−0.021	−0.055	0.039	−0.008	−0.037
EST6	0.851	−0.119	0.032	0.028	0.024	0.046	−0.016	0.063	0.075
EST9	0.835	0.142	−0.015	−0.019	−0.002	0.034	−0.006	−0.125	−0.053
EST1	−0.099	0.865	−0.009	−0.005	−0.008	−0.037	0.000	0.034	0.002
EST7	0.033	0.856	0.027	−0.003	−0.037	0.061	0.021	−0.052	0.045
EST10	0.066	0.866	−0.017	0.007	0.044	−0.023	−0.021	0.017	−0.047
WORK3	0.009	−0.035	0.874	0.049	0.009	−0.102	−0.009	0.016	0.003
WORK4	−0.009	0.035	0.874	−0.049	−0.009	0.102	0.009	−0.016	−0.003
WORK1	0.006	0.026	−0.011	0.881	0.016	0.055	0.013	−0.032	0.001
WORK2	−0.006	−0.026	0.011	0.881	−0.016	−0.055	−0.013	0.032	−0.001
SELFEFF1	−0.064	0.042	0.022	0.066	0.827	−0.012	0.018	−0.067	0.067
SELFEFF4	0.038	0.070	−0.004	−0.039	0.859	−0.053	0.050	0.020	−0.006
SELFEFF5	0.013	0.085	−0.019	−0.093	0.805	0.116	0.034	0.009	−0.003
SELFEFF6	−0.037	0.014	−0.000	0.162	0.788	0.029	−0.024	−0.034	−0.032
SELFEFF7	0.099	−0.077	0.005	−0.061	0.708	−0.123	−0.104	0.081	−0.095
SELFEFF8	−0.009	−0.015	0.032	−0.044	0.837	−0.066	−0.042	−0.059	−0.005
SELFEFF9	−0.059	−0.060	0.012	−0.025	0.724	0.028	−0.009	−0.085	0.037
SELFEFF10	−0.003	0.009	−0.038	−0.032	0.832	0.033	0.004	0.042	−0.016
MON1	−0.045	0.056	−0.071	−0.062	−0.010	0.840	0.013	0.006	0.039
MON2	0.045	−0.056	0.071	0.062	0.010	0.840	0.013	−0.006	−0.039
DIS1	−0.001	0.034	0.002	0.008	−0.014	0.008	0.956	0.018	−0.007
DIS2	0.001	−0.034	−0.002	−0.008	0.014	−0.008	0.956	−0.018	0.007
ATT1	0.060	−0.005	0.034	−0.059	0.032	0.054	0.107	0.823	–
ATT2	−0.025	−0.014	−0.031	0.074	0.010	−0.022	−0.054	0.834	–
ATT3	−0.008	0.011	0.015	0.049	0.002	0.037	0.008	0.780	–
ATT4	0.003	−0.001	0.006	−0.031	0.047	−0.046	0.098	–	0.839
ATT5	−0.001	−0.059	0.033	−0.038	−0.031	0.030	−0/065	–	0.739
ATT6	−0.002	0.053	−0.034	0.064	−0.019	0.019	−0.041	–	0.848

Table 9 presents the results for the first version of the structural model, using tax evasion and social insurance fraud as the endogenous variable. Table 10 presents the results for the second version of the structural model, using piracy as the endogenous variable.

Table 9. The results of the structural equations model—model 1.

	Direct Effects		Indirect Effects	Direct Effect Sizes (f2)		Total Effects (Direct Effect + Indirect Effect via Preoccupation for Money)
	Preoccupation for Money	TESIFA	TESIFA	Preoccupation for Money	TESIFA	PA
Preoccupation with money	–	0.100 * (0.011)			0.022	0.100 * (0.011)
FAIR	0.090 * (0.021)	0.048 (0.137)	0.009 (0.387)	0.015	0.005	0.057. (0.097)
MASS	0.353 *** (<0.001)	0.093 * (0.017)	0.035 (0.129)	0.141	0.018	0.129 ** (0.002)
HAWPACH	0.085 * (0.026)	−0.079 * (0.036)	0.009 (0.392)	0.007	0.006	−0.071. (0.054)
VLT	0.099 * (0.013)	0.058 (0.096)	0.010 (0.376)	0.014	0.016	0.068. (0.063)
Gender Male Female	Reference −0.128 ** (0.002)	Reference −0.083 * (0.030)	Reference −0.013 (0.341)	0.019	0.014	Reference −0.096 * (0.015)
Age	0.040 (0.181)	−0.105 ** (0.008)	0.004 (0.449)	0.003	0.015	−0.101 * (0.011)
ACADISH	–	0.296 *** (<0.001)	–	–	0.119	0.296 *** (<0.001)
NEF	–	−0.164 *** (<0.001)	–	–	0.047	−0.164 *** (<0.001)
POF	–	−0.062. (0.081)	–	–	0.008	−0.062. (0.081)
SELFEFF	–	−0.191 *** (<0.001)	–	–	0.048	−0.191 *** (<0.001)
R2/ Adjusted R2	20%/19%	31.8%/30.3%	–	–	–	–

Note: ***—p value < 0.001; **—p value < 0.01; *—p value < 0.05; .—p value < 0.10.

Table 10. The results of the structural equations model—model 2.

	Direct Effects		Indirect Effects	Direct Effect Sizes (f2)		Total Effects (Direct Effect + Indirect Effect via Preoccupation for Money)
	Preoccupation for Money	PA	PA	Preoccupation for Money	PA	PA
Preoccupation with money	–	0.139 *** (<0.001)	–		0.040	0.139 *** (<0.001)
FAIR	0.090 * (0.021)	0.017 (0.354)	0.012 (0.346)	0.015	0.001	0.029 (0.255)
MASS	0.353 *** (<0.001)	−0.015 (0.365)	0.049. (0.058)	0.141	0.003	0.034 (0.223)
HAWPACH	0.085 * (0.026)	−0.110 ** (0.003)	0.012 (0.353)	0.007	0.021	−0.099 * (0.013)
VLT	0.099 * (0.013)	0.122 ** (0.006)	0.014 (0.331)	0.014	0.022	0.136 *** (<0.001)
Gender Male Female	Reference −0.128 ** (0.002)	Reference −0.165 *** (<0.001)	Reference −0.018 (0.285)	0.019	0.046	Reference −0.183 *** (<0.001)
Age	0.040 (0.181)	0.131 ** (0.001)	0.006 (0.429)	0.003	0.023	0.137 *** (<0.001)
ACADISH	–	0.376 *** (<0.001)	–	–	0.180	0.376 *** (<0.001)
NEF	–	−0.088 * (0.023)	–	–	0.017	−0.088 * (0.023)
POF	–	0.026 (0.276)	–	–	0.003	0.026 (0.276)
SELFEFF	–	0.002 (0.486)	–	–	0.000	0.002 (0.486)
R2 /Adjusted R2	20%/19%	34.5%/33%	–	–	–	–

Note: ***—p value < 0.001; **—p value < 0.01; *—p value < 0.05; .—p value < 0.10.

These results show the estimated direct, indirect, and total effects, along with their statistical significance. The effect size for each of the direct paths is also reported. Tables 9 and 10 indicate good explanatory power with R-squared values of 19.6% (preoccupation with money), 21.1% (tax evasion and social insurance fraud), and 16.7% (piracy acceptance). The overall model fit, measured by the standardized root mean square residual (SRMR), is 0.06 for both versions of the model, well within the acceptable level. In general, it is considered that a SRMR < 0.08 indicates a very good fit [60].

Tables 11 and 12 summarize the results already presented in Tables 9 and 10 by indicating the direction and the significance of each relationship in simplified form, showing the two versions of the model side-by-side. FAIR has no significant impact—direct or indirect—on either outcome variable. It does, however, have a marginally significant total effect on TESIFA, and it is a significant predictor of the mediator variable. The belief in fairness does not appear to increase or reduce the level of tolerance towards deviant or fraudulent behavior. It does increase, however, the preoccupation with money, which makes sense if one perceives money as a means of keeping the score for effort and diligence.

Table 11. The results of the structural equations model, simplified and side-by-side (via the mediator).

Predictor	Preoccupation for Money	Direct Effects		Indirect Effects		Total Effects	
		TESIFA	PA	TESIFA	PA	TESIFA	PA
FAIR	+ (*)	None	None	None	None	+ (.)	None
MASS	+ (***)	+ (*)	None	None	+ (.)	+ (**)	None
HAWPACH	+ (*)	– (*)	– (**)	None	None	– (.)	– (*)
VLT	+ (*)	+ (.)	+ (***)	None	None	+ (.)	+ (***)
Gender Male Female	Reference – (***)	Reference – (*)	Reference – (***)	None	None	Reference – (*)	Reference – (***)
Age	None	– (**)	+ (**)	None	None	– (*)	+ (***)

Note: ***—p value < 0.001; **—p value < 0.01; *—p value < 0.05; .—p value < 0.10.

Table 12. The results of the structural equations model, simplified and side-by-side (predictors only).

Predictor	Tax Evasion and Security Insurance Fraud	Piracy Acceptance
PFM	+ (*)	+ (***)
ACADISH	+ (***)	+ (***)
NEF	– (***)	– (*)
POF	– (.)	None
SELFEFF	– (***)	None

Note: ***—p value < 0.001; **—p value < 0.01; *—p value < 0.05; .—p value < 0.10.

MASS is a positive and significant predictor of the mediator variable, and has a significant and positive total effect on TESIFA. There is a direct, significant, and positive effect on TESIFA, but there is no indirect effect. The total effect of MASS on the level of piracy acceptance is not statistically significant. However, when decomposed, the indirect effect (via the mediator, preoccupation about money) is marginally significant ($p = 0.058$), while the direct effect is not at all significant. Therefore,

preoccupation with money marginally mediates the relationship between MASS and the level of piracy acceptance. Money as social status increases the level of tolerance towards tax evasion and social insurance fraud. As expected, this relationship appears to be mediated by the preoccupation with money.

Interestingly, money as social status does not increase the level of piracy acceptance, and one can only speculate as to why this is the case. Perhaps tax evasion pays better than piracy, and the opportunity cost expressed in terms of social status is higher. On the other hand, money as social status is associated with a stronger sense of entitlement that extends to defying the authority of the government, whereas piracy is merely petty behavior relegated to penny-pinching. People who need money to enhance their social status are more likely to wear an Armani suit and use an expensive MacBook Pro while dodging taxation; they are less likely to wear a knock-off pair of shoes and use a pirated version of Windows. Expensive notebooks and pirated software do not go together well.

HAWPACH has a negative total effect on both TESIFA and PA, yet the relationship is significant only in the case of PA, and is marginally significant in the case of TESIFA. Direct effects are significant in both cases, but indirect effects are not at all significant. Although HAWPACH is a predictor of Preoccupation with money, the latter does not mediate the relationship with the outcome variables.

This result is highly expected, because one would assume that valuing hard work is at loggerheads with any type of cheating. One can attribute the difference in significance between the two models to the fact that piracy is probably a tangible and real experience that students (who make up our entire sample) can relate to in daily life. On the other hand, most of the students probably have a good mental representation of tax evasion and social insurance fraud, but not yet the experience of engaging in such behavior. Piracy is at hand, while tax evasion is a potential.

VLT has a positive total effect on both outcome variables, yet the relationship is significant only in the case of PA. There are no indirect effects and the only significant direct effect is in the relationship with PA, while this is marginally significant in the case of TESIFA. Although VLT is a predictor of preoccupation with money, the latter does not mediate the relationship with the outcome variables. People valuing leisure time appear to be more tolerant towards cheating and fraudulent behavior. This might be explained by the perceived high opportunity cost of hard work.

Gender has a statistically significant total effect on both TESIFA and PA. There are no indirect effects, only direct effects. As is the case with the previous two variables, gender predicts preoccupation with money, yet the latter does not act as a mediator in the relationship with the outcome variables. As expected, women are less inclined to tolerate tax evasion, social insurance fraud, and piracy.

A relatively similar situation appears in the case of age—the total effects are significant and there are only direct effects on TESIFA and PA. However, the effects on the two outcome variables have opposite signs—TESIFA decreases, but PA increases with age. Moreover, age is not a predictor of preoccupation with money. Older students appear less tolerant towards tax evasion and social insurance fraud, yet more tolerant towards piracy. This is hard to explain without introducing additional assumptions and variables that cannot be pursued and tested here.

The relationship between ACADISH and the two outcome variables is positive and highly significant, as expected. NEF is negatively and significantly related to both TESIFA and PA, yet POF is only marginally related to TESIFA, and not at all to PA. Self-efficacy is negatively and significantly related to TESIFA, and not at all to PA.

5. Discussion

Contextual effect and dispositional effect predictors appear segregated according to the statistical significance of their relationship with the two outcome variables. If one sets aside the case of money as social status (for reasons discussed earlier), one notices that the entire set of contextual effects predictors show a stronger statistical significance in their relationship to piracy acceptance; and a weaker statistical significance in their relationship to tax evasion and social insurance fraud acceptance. One might interpret this in light of the fact that most of the students who answered our questionnaire

are perhaps much more often exposed to, or engaged in piracy than in tax evasion and social insurance fraud at this stage in their life. Piracy represents a behavior one can easily relate to, while tax evasion is still a theoretical possibility.

On the other hand, dispositional effect predictors show a stronger statistical significance in their relationship with tax evasion and social insurance fraud acceptance than in the case of piracy acceptance. Perhaps the former variable is seen as more consequential and less socially acceptable than the latter; hence, a behavior such as tax evasion, by virtue of its perceived importance, is more likely to be at the center of thought processes associated with self-evaluation, self-cognition, and self-control.

Money priming has been shown to increase materialistic values and to make people more unscrupulous, yet contextual effects predict the outcomes without the mediation of PFM, although PFM, as expected, remains a predictor of both outcomes. Academic dishonesty is predicting cheating and fraudulent behavior well beyond the workplace, and size effects are the largest among all the predictors.

Both models have a good explanatory power. The predictors explain about 30% or more of the variation of the outcome constructs. Moreover, 20% of the variation of the mediator is explained by contextual and control variables (gender and age). From a practical perspective, the relative contribution of each individual predictor to the combined explanatory power of the model matters a lot. This contribution is usually measured in terms of effect sizes. In order for an intervention to be of any consequence, the effect size has to be above 0.02 [61]. By far, academic dishonesty displays the largest effect size of all predictors for both models. This is without any doubt one of the more important findings of this research.

In order to reduce the acceptance of tax evasion, social insurance fraud, piracy, and perhaps of other types of cheating and fraud, it helps to act on academic dishonesty first and foremost [62]. It has been shown that business students who are coached and sensitized about social responsibility and ethical management appear less inclined to engage in dishonest behavior and are less tolerant towards cheating [4]. Sometimes, simple, old-fashioned moral education and making people self-aware about cheating and deception in the context of social norms might be sufficient to effect a shift in attitudes. Short-circuiting self-deception through sensitivity training represents in itself a deterrent to cheating [5].

In the case of tax evasion and social insurance fraud, another lever appears to be self-efficacy. Boosting self-efficacy seemingly reduces the acceptance of tax evasion. However, this result is more or less trivial. In the case of piracy acceptance, gender and age also show small effect sizes, which are suitable for intervention. Both gender and age, however, represent control variables in this model, and they cannot be manipulated or acted upon in the same way as one might act upon other predictors.

In a similar vein, negative feelings also show relevant effect sizes (in the case of the first outcome variable) but cannot and should not be manipulated—without raising serious ethical concerns—merely for the sake of reducing the level of tolerance towards tax evasion.

Finally, preoccupation with money appears to have a small effect size on both outcome variables. Here, the path of intervention has to go through MASS, which also has a relevant effect size in its relationship with preoccupation with money.

6. Conclusions

An important body of literature shows that academic cheating is a good predictor of cheating and unethical behavior in the workplace. This paper takes this line of research one step further and investigates the relationship between academic dishonesty and other types of unethical and even fraudulent behaviors, well beyond the workplace; the focus is on tax evasion, social insurance fraud, and piracy. The source of data is a sample of 504 respondents, all students, aged 18–25. The PLS-PM analysis used in this research alternates the outcome variable between tax evasion and social insurance fraud acceptance, and piracy acceptance. The main predictor is academic dishonesty. The control variables are

age, gender, and several latent constructs accounting for context and disposition. Preoccupation with money is used as a mediator between contextual predictors and our outcome variables.

It is contended that the results are consequential because they show that academic dishonesty is a statistically significant predictor of an entire range of unethical and fraudulent behavior acceptance. In addition, one finds that contextual constructs are segregated in the way they impact the two outcome variables. One explanation is that the sample is composed of students, and the respondents relate differently to tax evasion and social insurance fraud acceptance than to piracy acceptance for obvious reasons. The most important result, however, is that the effect size of the main predictor is large. This should shift the focus to a broader debate about how to fight academic dishonesty most efficiently. It is a question that should be taken very seriously, because there is a lot at stake—cheating has far-reaching implications, measured in billions of dollars. Yet again, the question of ethical behavior in general does not represent an esoteric, theoretical concept. It is consequential precisely because morality relates to real costs, some direct and tangible, other indirect and harder to ascertain, but which are nevertheless significant. The findings presented here have certain limitations. There is no doubt that vigorous intervention should be undertaken to influence the attitude towards academic dishonesty, but there is no obvious indication on how to achieve this. Effective intervention should start in school and university, but the most appropriate course of action should be the subject of other research. Last but not least, the focus of this study has been on attitudes rather than on behavior. It is unclear the extent to which the attitudes measured here would translate into actual behavior, be it academic cheating, tax evasion, or piracy.

Author Contributions: Conceptualization, E.D. and R.I.-C.; methodology, E.D. and C.V.; software, E.D. validation, I.M.; formal analysis, E.D.; investigation, R.I.-C., I.M., and R.M.-P.; resources, R.M.-P.; data curation, E.D.; writing—E.D., C.V., and R.I.-C.; writing—review and editing, E.D. and C.V.; funding acquisition, R.M.-P.

Funding: This research received no external funding.

Acknowledgments: This research was supported by a Marie Curie Research and Innovation Staff Exchange scheme within the H2020 Programme (grant acronym: SHADOW, no: 778118).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. LaDuke, R.D. Academic Dishonesty Today, Unethical Practices Tomorrow? *J. Prof. Nurs.* **2013**, *29*, 402–406. [CrossRef] [PubMed]
2. Nonis, S.; Swift, C.O. An Examination of the Relationship Between Academic Dishonesty and Workplace Dishonesty: A Multicampus Investigation. *J. Educ. Bus.* **2001**, *77*, 69–77. [CrossRef]
3. Lucas, G.M.; Friedrich, J. Individual Differences in Workplace Deviance and Integrity as Predictors of Academic Dishonesty. *Ethics Behav.* **2005**, *15*, 15–35. [CrossRef]
4. Chen, Y.-J.; Tang, T.L.-P. Attitude Toward and Propensity to Engage in Unethical Behavior: Measurement Invariance across Major among University Students. *J. Bus. Ethics* **2006**, *69*, 77–93. [CrossRef]
5. Mazar, N.; Ariely, D. Dishonesty in everyday life and its policy implications. *J. Public Policy Mark.* **2006**, *25*, 117–126. [CrossRef]
6. Harper, M.G. High tech cheating. *Nurse Educ. Pract.* **2006**, *6*, 364–371. [CrossRef]
7. Schmelkin, L.P.; Gilbert, K.; Spencer, K.J.; Pincus, H.S.; Silva, R. A Multidimensional Scaling of College Students' Perceptions of Academic Dishonesty. *J. High. Educ.* **2008**, *79*, 587–607. [CrossRef]
8. Brown, D.L. Cheating Must Be Okay—Everybody Does It! *Nurse Educ.* **2002**, *27*, 6. [CrossRef]
9. Faucher, D.; Caves, S. Academic dishonesty: Innovative cheating techniques and the detection and prevention of them. *Teach. Learn. Nurs.* **2009**, *4*, 37–41. [CrossRef]
10. Gaberson, K.B. Academic Dishonesty Among Nursing Students. *Nurs. Forum* **1997**, *32*, 14–20. [CrossRef]
11. Carpenter, D.D.; Harding, T.S.; Finelli, C.J.; Passow, H.J. Does academic dishonesty relate to unethical behavior in professional practice? An exploratory study. *Sci. Eng. Ethics* **2004**, *10*, 311–324. [CrossRef] [PubMed]

12. Li-Ping Tang, T.; Chen, Y.-J.; Sutarso, T. Bad apples in bad (business) barrels: The love of money, Machiavellianism, risk tolerance, and unethical behavior. *Manag. Decis.* **2008**, *46*, 243–263. [CrossRef]
13. Sheard, J.; Markham, S.; Dick, M. Investigating Differences in Cheating Behaviours of IT Undergraduate and Graduate Students: The maturity and motivation factors. *High. Educ. Res. Dev.* **2003**, *22*, 91–108. [CrossRef]
14. Gerlach, P. The games economists play: Why economics students behave more selfishly than other students. *PLoS ONE* **2017**, *12*, e0183814. [CrossRef] [PubMed]
15. Harris, J.R. A Comparison of the Ethical Values of Business Faculty and Students: How Different Are They? *Bus. Prof. Ethics J.* **1988**, *7*, 27–49. [CrossRef]
16. Wood, J.A.; Longenecker, J.G.; McKinney, J.A.; Moore, C.W. Ethical attitudes of students and business professionals: A study of moral reasoning. *J. Bus. Ethics* **1988**, *7*, 249–257. [CrossRef]
17. Bacon, A.M.; McDaid, C.; Williams, N.; Corr, P.J. What motivates academic dishonesty in students? A reinforcement sensitivity theory explanation. *Br. J. Educ. Psychol.* **2019**. [CrossRef]
18. Błachnio, A. Don't cheat, be happy. Self-control, self-beliefs, and satisfaction with life in academic honesty: A cross-sectional study in Poland. *Scand. J. Psychol.* **2019**, *60*, 261–266. [CrossRef]
19. Fida, R.; Tramontano, C.; Paciello, M.; Ghezzi, V.; Barbaranelli, C. Understanding the Interplay Among Regulatory Self-Efficacy, Moral Disengagement, and Academic Cheating Behaviour During Vocational Education: A Three-Wave Study. *J. Bus. Ethics* **2018**, *153*, 725–740. [CrossRef]
20. Keener, T.A.; Galvez Peralta, M.; Smith, M.; Swager, L.; Ingles, J.; Wen, S.; Barbier, M. Student and faculty perceptions: Appropriate consequences of lapses in academic integrity in health sciences education. *BMC Med. Educ.* **2019**, *19*, 209. [CrossRef]
21. Blankenship, K.L.; Whitley, B.E. Relation of General Deviance to Academic Dishonesty. *Ethics Behav.* **2000**, *10*, 1–12. [CrossRef]
22. Fass, R. Cheating and plagiarism. *Ethics High. Educ.* **1990**. Available online: https://eric.ed.gov/?id=ED324727 (accessed on 12 November 2019).
23. Walsh, K. Understanding Taxpayer Behaviour—New Opportunities for Tax Administration. *Econ. Soc. Rev.* **2012**, *43*, 451–475.
24. Lim, V.K.G.; Teo, T.S.H. Sex, money and financial hardship: An empirical study of attitudes towards money among undergraduates in Singapore. *J. Econ. Psychol.* **1997**, *18*, 369–386. [CrossRef]
25. Yamauchi, K.T.; Templer, D.J. The Development of a Money Attitude Scale. *J. Personal. Assess.* **1982**, *46*, 522–528. [CrossRef] [PubMed]
26. Fox, S.; Spector, P.E.; Miles, D. Counterproductive Work Behavior (CWB) in Response to Job Stressors and Organizational Justice: Some Mediator and Moderator Tests for Autonomy and Emotions. *J. Vocat. Behav.* **2001**, *59*, 291–309. [CrossRef]
27. Caruso, E.M.; Vohs, K.D.; Baxter, B.; Waytz, A. Mere exposure to money increases endorsement of free-market systems and social inequality. *J. Exp. Psychol. Gen.* **2013**, *142*, 301–306. [CrossRef]
28. Phau, I.; Woo, C. Understanding compulsive buying tendencies among young Australians: The roles of money attitude and credit card usage. *Mark. Intell. Plan.* **2008**, *26*, 441–458. [CrossRef]
29. McHoskey, J.W. Factor structure of the protestant work ethic scale. *Personal. Individ. Differ.* **1994**, *17*, 49–52. [CrossRef]
30. Mudrack, P.E. Protestant work-ethic dimensions and work orientations. *Personal. Individ. Differ.* **1997**, *23*, 217–225. [CrossRef]
31. Mirels, H.L.; Garrett, J.B. The Protestant Ethic as a personality variable. *J. Consult. Clin. Psychol.* **1971**, *36*, 40–44. [CrossRef]
32. Hassall, S.L.; Muller, J.J.; Hassall, E.J. Comparing the Protestant work ethic in the employed and unemployed in Australia. *J. Econ. Psychol.* **2005**, *26*, 327–341. [CrossRef]
33. Amos, C.; Zhang, L.; Read, D. Hardworking as a Heuristic for Moral Character: Why We Attribute Moral Values to Those Who Work Hard and Its Implications. *J. Bus. Ethics* **2019**, *158*, 1047–1062. [CrossRef]
34. Murdock, T.B.; Stephens, J.M. 10—Is Cheating Wrong? Students' Reasoning about Academic Dishonesty. In *Psychology Academic Cheating*; Anderman, E.M., Murdock, T.B., Eds.; Academic Press: Burlington, ON, Canada, 2007; pp. 229–251. ISBN 978-0-12-372541-7.
35. Jensen, L.A.; Arnett, J.J.; Feldman, S.S.; Cauffman, E. It's Wrong, But Everybody Does It: Academic Dishonesty among High School and College Students. *Contemp. Educ. Psychol.* **2002**, *27*, 209–228. [CrossRef]

36. Ciarrochi, J.; Heaven, P.C.; Davies, F. The impact of hope, self-esteem, and attributional style on adolescents' school grades and emotional well-being: A longitudinal study. *J. Res. Personal.* **2007**, *41*, 1161–1178. [CrossRef]
37. Mullen, S.P.; Gothe, N.P.; McAuley, E. Evaluation of the factor structure of the Rosenberg Self-Esteem Scale in older adults. *Personal. Individ. Differ.* **2013**, *54*, 153–157. [CrossRef]
38. Rosenberg, M. *Society and the Adolescent Self-Image*; Princeton University Press: Princeton, NJ, USA, 2015; ISBN 978-1-4008-7613-6.
39. Fox, S.; Spector, P.E. A model of work frustration–aggression. *J. Organ. Behav.* **1999**, *20*, 915–931. [CrossRef]
40. Schwarzer, R.; Jerusalem, M. The general self-efficacy scale (GSE). *Anxiety Stress Coping* **2010**, *12*, 329–345.
41. Luszczynska, A.; Scholz, U.; Schwarzer, R. The General Self-Efficacy Scale: Multicultural Validation Studies. *J. Psychol.* **2005**, *139*, 439–457. [CrossRef]
42. Reiss, M.C.; Mitra, K. The Effects of Individual Difference Factors on the Acceptability of Ethical and Unethical Workplace Behaviors. *J. Bus. Ethics* **1998**, *17*, 1581–1593. [CrossRef]
43. Scholz, U.; Doña, B.G.; Sud, S.; Schwarzer, R. Is general self-efficacy a universal construct? Psychometric findings from 25 countries. *Eur. J. Psychol. Assess.* **2002**, *18*, 242. [CrossRef]
44. Hoon, L.S.; Lim, V.K.G. Attitudes towards money and—for Asian management style following the economic crisis. *J. Manag. Psychol.* **2001**, *16*, 159–173. [CrossRef]
45. Chitchai, N.; Senasu, K.; Sakworawich, A. The moderating effect of love of money on relationship between socioeconomic status and happiness. *Kasetsart J. Soc. Sci.* **2018**. [CrossRef]
46. Vohs, K.D. Money priming can change people's thoughts, feelings, motivations, and behaviors: An update on 10 years of experiments. *J. Exp. Psychol. Gen.* **2015**, *144*, e86–e93. [CrossRef] [PubMed]
47. Mitchell, M.S.; Baer, M.D.; Ambrose, M.L.; Folger, R.; Palmer, N.F. Cheating under pressure: A self-protection model of workplace cheating behavior. *J. Appl. Psychol.* **2018**, *103*, 54–73. [CrossRef]
48. Kouchaki, M.; Smith-Crowe, K.; Brief, A.P.; Sousa, C. Seeing green: Mere exposure to money triggers a business decision frame and unethical outcomes. *Organ. Behav. Hum. Decis. Process.* **2013**, *121*, 53–61. [CrossRef]
49. Piff, P.K. Wealth and the Inflated Self: Class, Entitlement, and Narcissism. *Personal. Soc. Psychol. Bull.* **2014**, *40*, 34–43. [CrossRef]
50. Durvasula, S.; Lysonski, S. Money, money, money—How do attitudes toward money impact vanity and materialism?—The case of young Chinese consumers. *J. Consum. Mark.* **2010**, *27*, 169–179. [CrossRef]
51. Roberts, J.A.; Jones, E. Money Attitudes, Credit Card Use, and Compulsive Buying among American College Students. *J. Consum. Aff.* **2001**, *35*, 213–240. [CrossRef]
52. Roberts, J.A.; Sepulveda, M.C.J. Demographics and money attitudes: A test of Yamauchi and Templers (1982) money attitude scale in Mexico. *Personal. Individ. Differ.* **1999**, *27*, 19–35. [CrossRef]
53. Wernimont, P.F.; Fitzpatrick, S. The meaning of money. *J. Appl. Psychol.* **1972**, *56*, 218–226. [CrossRef]
54. Roberts, J.A.; Sepulveda, M.C.J. Money Attitudes and Compulsive Buying. *J. Int. Consum. Mark.* **1999**, *11*, 53–74. [CrossRef]
55. Fornell, C.; Bookstein, F.L. Two Structural Equation Models: LISREL and PLS Applied to Consumer Exit-Voice Theory. *J. Mark. Res.* **1982**, *19*, 440–452. [CrossRef]
56. Joreskog, K.G.; Wold, H. The ML and PLS techniques for modeling with latent variables: Historical and comparative aspects. In *Systems under Indirect Observation: Part I.*; Joreskog, K.G., Wold, H., Eds.; North–Holland: Amsterdam, The Netherland, 1982; pp. 263–270.
57. Nunnally, J.C.; Bernstein, I.H. *Psychometric Theory*, 3rd ed.; McGraw-Hill: New York, NY, USA, 1994; ISBN 978-0-07-047849-7.
58. Fornell, C.; Larcker, D.F. Evaluating structural equation models with unobservable variables and measurement error. *J. Mark. Res.* **1981**, *18*, 39–50. [CrossRef]
59. Kennedy, P. *A Guide to Econometrics*, 6st ed.; Wiley-Blackwell: Malden, MA, USA, 2008; ISBN 978-1-4051-8257-7.
60. Hu, L.; Bentler, P.M. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Model. Multidiscip. J.* **1999**, *6*, 1–55. [CrossRef]

61. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Routledge: Hillsdale, MI, USA, 1988; ISBN 978-0-8058-0283-2.
62. Mirshekary, S.; Lawrence, A.D.K. Academic and Business Ethical Misconduct and Cultural Values: A Cross National Comparison. *J. Acad. Ethics* **2009**, *7*, 141–157. [CrossRef]

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Mathematics Editorial Office
E-mail: mathematics@mdpi.com
www.mdpi.com/journal/mathematics

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34
Fax: +41 61 302 89 18

www.mdpi.com

ISBN 978-3-0365-0197-0