



Journal of  
*Risk and Financial  
Management*

Special Issue Reprint

---

# Financial Data Analytics and Statistical Learning

---

Edited by  
Shuangzhe Liu, Tiefeng Ma and Seng Huat Ong

[mdpi.com/journal/jrfm](https://mdpi.com/journal/jrfm)



# **Financial Data Analytics and Statistical Learning**



# Financial Data Analytics and Statistical Learning

Editors

**Shuangzhe Liu**

**Tiefeng Ma**

**Seng Huat Ong**



Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

*Editors*

Shuangzhe Liu  
Faculty of Science and  
Technology  
University of Canberra  
Canberra  
Australia

Tiefeng Ma  
School of Statistics  
Southwestern University of  
Finance and Economics  
Chengdu  
China

Seng Huat Ong  
Institute of Actuarial Science  
and Data Analytics  
UCSI University  
Kuala Lumpur  
Malaysia

*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 *Journal of Risk and Financial Management* (ISSN 1911-8074) (available at: [www.mdpi.com/journal/jrfm/special\\_issues/Financial.Statistics.II](http://www.mdpi.com/journal/jrfm/special_issues/Financial.Statistics.II)).

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

Lastname, Firstname, Firstname Lastname, and Firstname Lastname. Article Title. <i>Journal Name</i> <b>Year</b> , Volume Number, Page Range.
--

**ISBN 978-3-7258-0482-5 (Hbk)**

**ISBN 978-3-7258-0481-8 (PDF)**

**[doi.org/10.3390/books978-3-7258-0481-8](https://doi.org/10.3390/books978-3-7258-0481-8)**

© 2024 by the authors. Articles in this book are Open Access and distributed under the Creative Commons Attribution (CC BY) license. The book as a whole is distributed by MDPI under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) license.

# Contents

<b>About the Editors</b> . . . . .	<b>vii</b>
<b>Preface</b> . . . . .	<b>ix</b>
<b>T. Sathiyaraj, T. Ambika and Ong Seng Huat</b>	
Exponential Stability of Fractional Large-Scale Neutral Stochastic Delay Systems with Fractional Brownian Motion Reprinted from: <i>J. Risk Financial Manag.</i> <b>2023</b> , <i>16</i> , 278, doi:10.3390/jrfm16050278 . . . . .	<b>1</b>
<b>Kentarou Wada and Takeshi Kurosawa</b>	
The Naive Estimator of a Poisson Regression Model with a Measurement Error Reprinted from: <i>J. Risk Financial Manag.</i> <b>2023</b> , <i>16</i> , 186, doi:10.3390/jrfm16030186 . . . . .	<b>16</b>
<b>Yeh-Ching Low and Seng-Huat Ong</b>	
Modelling of Loan Non-Payments with Count Distributions Arising from Non-Exponential Inter-Arrival Times Reprinted from: <i>J. Risk Financial Manag.</i> <b>2023</b> , <i>16</i> , 150, doi:10.3390/jrfm16030150 . . . . .	<b>31</b>
<b>Abbas Mahdavi, Omid Kharazmi and Javier E. Contreras-Reyes</b>	
On the Contaminated Weighted Exponential Distribution: Applications to Modeling Insurance Claim Data Reprinted from: <i>J. Risk Financial Manag.</i> <b>2022</b> , <i>15</i> , 500, doi:10.3390/jrfm15110500 . . . . .	<b>45</b>
<b>Ashis SenGupta and Moumita Roy</b>	
Circular-Statistics-Based Estimators and Tests for the Index Parameter $\alpha$ of Distributions for High-Volatility Financial Markets Reprinted from: <i>J. Risk Financial Manag.</i> <b>2023</b> , <i>16</i> , 405, doi:10.3390/jrfm16090405 . . . . .	<b>63</b>
<b>Haokun Dong, Rui Liu and Allan W. Tham</b>	
Accuracy Comparison between Five Machine Learning Algorithms for Financial Risk Evaluation Reprinted from: <i>J. Risk Financial Manag.</i> <b>2024</b> , <i>17</i> , 50, doi:10.3390/jrfm17020050 . . . . .	<b>77</b>
<b>Ali İhsan Çetin, Arzu Ece Çetin and Syed Ejaz Ahmed</b>	
The Impact of Non-Financial and Financial Variables on Credit Decisions for Service Companies in Turkey Reprinted from: <i>J. Risk Financial Manag.</i> <b>2023</b> , <i>16</i> , 487, doi:10.3390/jrfm16110487 . . . . .	<b>96</b>
<b>Kangyi Wang, Jing Ma, Chunxiao Xue and Jianing Zhang</b>	
Board Gender Diversity and Firm Performance: Recent Evidence from Japan Reprinted from: <i>J. Risk Financial Manag.</i> <b>2024</b> , <i>17</i> , 20, doi:10.3390/jrfm17010020 . . . . .	<b>114</b>
<b>Paul R. Dewick</b>	
On Financial Distributions Modelling Methods: Application on Regression Models for Time Series Reprinted from: <i>J. Risk Financial Manag.</i> <b>2022</b> , <i>15</i> , 461, doi:10.3390/jrfm15100461 . . . . .	<b>141</b>
<b>Indranil Ghosh, Dalton Watts and Subrata Chakraborty</b>	
Modeling Bivariate Dependency in Insurance Data via Copula: A Brief Study Reprinted from: <i>J. Risk Financial Manag.</i> <b>2022</b> , <i>15</i> , 329, doi:10.3390/jrfm15080329 . . . . .	<b>156</b>
<b>Linyu Cao, Ruili Sun, Tiefeng Ma and Conan Liu</b>	
On Asymmetric Correlations and Their Applications in Financial Markets Reprinted from: <i>J. Risk Financial Manag.</i> <b>2023</b> , <i>16</i> , 187, doi:10.3390/jrfm16030187 . . . . .	<b>176</b>



# About the Editors

## **Shuangzhe Liu**

Shuangzhe Liu currently serves as the leader of the Data Science Group within the Faculty of Science and Technology at the University of Canberra, Australia. He obtained his Ph.D. in Econometrics from the Tinbergen Institute, University of Amsterdam, the Netherlands, specializing in matrix differential calculus, multivariate analysis, and statistical learning. His extensive expertise is evidenced by his numerous publications in prestigious journals in the fields of econometrics, mathematics, statistics, and related areas. Additionally, he has co-authored a comprehensive book on time series analysis using SAS Enterprise Guide. Demonstrating a strong commitment to advancing statistical knowledge, Shuangzhe actively contributes to the field as an associate editor for multiple statistical journals and holds an editor position at *Statistical Papers*.

## **Tiefeng Ma**

Tiefeng Ma earned his Ph.D. degree in probability theory and mathematical statistics from Beijing University of Technology, China, in 2008. Presently, he holds the esteemed position of professor within the School of Statistics at Southwestern University of Finance and Economics, located in Chengdu, China. Dr. Ma's research focuses on several intriguing areas, including linear modeling, data mining, clustering, change point analysis, and power data analysis. His expertise lies in these domains, where he continuously seeks to advance knowledge and understanding. Throughout his academic career, Dr. Ma has made significant contributions to the field, with more than 60 articles published in esteemed journals and conference proceedings. His work showcases his dedication and commitment to scholarly pursuits, thereby enhancing the scientific community's understanding of these subjects.

## **Seng Huat Ong**

Seng Huat Ong joined UCSI University in 2017 and is currently the head of research and postgraduate studies at the Institute of Actuarial Science and Data Analytics (IASDA). Professor Ong is also an honorary professor at the Institute of Mathematical Sciences, University of Malaya. His research interest is in statistical and stochastic modeling and distribution theory, with a focus on models for physical and biological systems, statistical computation and simulation, data analytics, and stochastic fractional differential equations. He became an elected member of the International Statistical Institute in 2001 and was admitted as a fellow of the Academy of Science Malaysia in 2010.





# Preface

In this reprint, we gathered the latest developments in financial analysis and statistical learning, along with practical applications.

Sathiyaraj et al. delved into the exponential stability of fractional-order large-scale neutral stochastic delay systems with fractional Brownian motion, commonly used to model financial phenomena due to their long memory property.

Wada and Kurosawa generalized the naive estimator of a Poisson regression model with measurement errors, extending the assumptions beyond normal distributions for explanatory variables.

Modeling non-payment counts as a renewal process involves examining the inter-arrival times between events. Low and Ong introduced a method for numerically computing probabilities and the renewal function based on Laplace transform inversion.

Deriving loss distribution from insurance data poses a challenge due to its skewed nature with heavy tails and the presence of outliers. Mahdavi et al. made an extension of the weighted exponential family, incorporating flexible features such as bimodality and a range of skewness and kurtosis.

Stable distributions offer better modeling for high-volatility financial data. SenGupta and Roy introduced a novel estimator for the index parameter using a trigonometric moment estimator based on circular distributions.

Accurate loan default prediction is crucial for credit risk assessment. Dong et al. explored a non-parametric approach with five machine learning classifiers on large datasets.

Çetin et al. analyzed factors influencing credit decision-making in Turkey's dynamic service sector post-2000, amid accelerated economic growth.

Wang et al. examined board gender diversity's effect on firm performance using 1990 publicly listed Japanese companies from 2006 to 2023.

The financial market poses challenges in identifying the distribution and stylized facts of time series data. Dewick employed regression modeling to assess the goodness-of-fit between original and generated time series models, aiding in model selection.

Ghosh et al. used the VineCopula package in R to analyze the dependence structure of real-life insurance data.

Cao et al. reviewed recent advancements in understanding asymmetric correlations of asset returns and explored their implications for hedging, diversification, and multifractal asymmetric detrend cross-correlation analysis.

We hope this Special Volume proves valuable for graduate students and researchers in fields related to financial analytics, business statistics, econometrics, insurance studies, and other relevant areas.

Finally, we express our appreciation to Boris Buchmann, Manuel Galea, Roger Gay, Shigeyuki Hamori, Kazuhiko Kakamu, Takeaki Kariya, Victor Leiva, Yonghui Liu, Changyu Lu, Ross Maller, Gilberto Paula, Milind Sathye, Kunio Shimizu, Zari Rachev, Lei Shi, Ken Siu, Hailiang Yang, and Fukang Zhu for their strong encouragement and support. We also extend our gratitude to all the authors and reviewers for their significant contributions.

**Shuangzhe Liu, Tiefeng Ma, and Seng Huat Ong**

*Editors*



Article

# Exponential Stability of Fractional Large-Scale Neutral Stochastic Delay Systems with Fractional Brownian Motion

T. Sathiyaraj <sup>1,\*</sup>, T. Ambika <sup>2</sup> and Ong Seng Huat <sup>1,3</sup>

<sup>1</sup> Institute of Actuarial Science and Data Analytics, UCSI University, Kuala Lumpur 56000, Malaysia

<sup>2</sup> Department of Computer Science, Rev. Jacob Memorial Christian College, Dindigul 624612, India

<sup>3</sup> Institute of Mathematical Sciences, University of Malaya, Kuala Lumpur 50603, Malaysia

\* Correspondence: sathiyaraj133@gmail.com

**Abstract:** Mathematics plays an important role in many fields of finance. In particular, it presents theories and tools widely used in all areas of finance. Moreover, fractional Brownian motion (fBm) and related stochastic systems have been used to model stock prices and other phenomena in finance due to the long memory property of such systems. This manuscript provides the exponential stability of fractional-order Large-Scale neutral stochastic delay systems with fBm. Based on fractional calculus (FC),  $\mathbb{R}^n$  stochastic space and Banach fixed point theory, sufficiently useful conditions are derived for the existence of solution and exponential stability results. In this study, we tackle the nonlinear terms of the considered systems by applying local assumptions. Finally, to verify the theoretical results, a numerical simulation is provided.

**Keywords:** dynamic risk in asset pricing; exponential stability; finance modeling and derivatives; fractional calculus; fractional Brownian motion; large dimensional problems; simulation and computation in long short-term memory; time delay



**Citation:** Sathiyaraj, T., T. Ambika, and Ong Seng Huat. 2023.

Exponential Stability of Fractional Large-Scale Neutral Stochastic Delay Systems with Fractional Brownian Motion. *Journal of Risk and Financial Management* 16: 278. <https://doi.org/10.3390/jrfm16050278>

Academic Editor: Thanasis Stengos

Received: 31 January 2023

Revised: 5 March 2023

Accepted: 17 May 2023

Published: 19 May 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Knowledge of mathematics, probability, statistics, and other analytical approaches is essential to develop methods and theories in finance and to test their validity through analysis of empirical real-world data. For example, mathematics, probability, and statistics help develop pricing models for financial assets such as stocks, bonds, currencies, and derivative securities and propose financially optimal strategies to decision makers based on their preferences. Brownian motion is a mathematical process used to describe random fluctuations in the stock market. It assumes that stock prices move randomly and follow a random walk. It is a type of stochastic process which can often be seen to model the movement of particles in a fluid or gas. However, Brownian motion is widely used in finance to model the random walk of stock prices over time. To apply Brownian motion in stock market modeling, the randomness of the price movement is used, as there is no particular trend and direction. This randomness is then modeled as a series of random steps, where each step represents a small change in the stock price. The size of each step is determined by the stock volatility, which is a measure of how much the stock price tends to oscillate over time. One important feature of Brownian motion is that it is a continuous process, meaning that the stock price can take on any value within a certain range. This makes it useful for modeling the behavior of stock prices over time, as it allows us to capture the full range of possible outcomes. However, while Brownian motion can be a useful tool for understanding the behavior of stock prices, it is not a perfect model. Stock prices can be influenced by a wide range of factors, including news events, company performance, and economic conditions. These factors can cause stock prices to move in ways that are not easily captured by a simple model such as Brownian motion.

The Hurst index has recently been introduced as a useful tool for assessing the memory effect, frequently measured by the autocorrelation function Hurst (1951).  $\mathcal{H}(0 < \mathcal{H} < 1)$  is a common way to represent the Hurst index.

- (1) When  $0 < \mathcal{H} < 0.5$ , the time series exhibits a negative correlation and antipersistent behaviour, or short-dependence memory.
- (2) When  $\mathcal{H} = 0.5$ , the time series is independent.
- (3) When  $0.5 < \mathcal{H} < 1$ , the time series exhibits persistent behaviour, or long-dependence memory.

The concept of fractional derivatives is not new, and FC has a long history of up to three centuries. The number of FC-related publications increased significantly in the later decades and mid-20th century. One of explanations for the high level of curiosity in fractional differential equations (FDEs) is that they can be used to define a diverse range of physical Hilfer (2000), chemical Oldham (2010), and biological Magin (2010) processes. Fractional derivative plays an important role in memory and hereditary processes. Several studies have been conducted to examine the long memory in the financial markets, since memory effect is a significant feature in financial systems. FC can be found in a variety of applications as a new branch of applied mathematics. Leibnitz, Caputo, Liouville, Riemann, Euler, and others are credited with a significant amount of foundational mathematical theory relevant to FC analysis. Nonetheless, throughout the last few decades, increasingly compelling representations have been discovered in numerous engineering and science disciplines (see Ortigueira (2011)). It should be highlighted that the existence hypothesis of FDEs is committed to a considerable part of the recent studies (see Balachandran et al. (2012); Nieto and Samet (2017); Singh et al. (2017); Tian and Nieto (2017)).

Recently, Bhaskar and Biswajit (2023) examined the effects of the steep surge in crude oil price shock on the stock price returns and currency exchange rates of G7 countries, namely Canada, France, Germany, Italy, Japan, the United Kingdom and the United States, in the context of the Russia–Ukraine conflict. Regime switches in the empirical relation between return dynamics and implied volatility in energy markets have been discussed in Okawa (2023). Optimal combination of proportional and Stop-Loss reinsurance with dependent claim and stochastic insurance premium have been studied in Sari et al. (2023). Herding trend in working capital management practices: evidence from the non-financial sector of Pakistan is analyzed in Farooq et al. (2023). Growth of venture firms under state capitalism with Chinese characteristics: qualitative comparative analysis of fuzzy set is discussed in Yun et al. (2023). In Li et al. (2014), the authors established a fractional-order stochastic differential equation model to describe the effect of trend memory in financial pricing.

While analyzing, there must be considerations for functional structures, ambient noise, and temporal delays, which can be quite valuable when constructing further sensible scientific models Mao (1997). The solution process for a stochastic fractional partial differential equation driven by space–time white noise has been studied in Wu (2011). The controllability of fractional and Hilfer fractional dynamical systems has been studied in Kumar et al. (2022a, 2022b, 2023). The relations between a singular system of differential equations and a system with delays, and stability of fractional-order quasi-linear impulsive integro-differential systems with multiple delays have been studied in Dassios (2022); Kalidass et al. (2022).

Another type of noise exposure is continuous. This can be modeled using Levy methods. In particular, methods based on Poisson random measures, as a common non-Gaussian stochastic method, have already received a lot of attention in a variety of fields and have been used to predict when demand for supply chain systems will increase Song (2009). Mathematical modeling of one-server m-form random queuing in a network system is modeled in the stochastic environment problems Seo and Lee (2011), distribution patterns of phone users in the service area of wireless links Taheri et al. (2010), as well as other naturally occurring anomalies in a variety of areas Applebaum (2009). In Rockner and Zhang (2007) the existence, uniqueness and huge deviation principle solutions to jump

type stochastic evolution equations were investigated. Many researchers have recently turned to FDEs as a useful tool for describing a variety of steady physical processes.

However, research into nonlinear FDE stability theory is still in its early phases, and much more work in this field is possible. Recently, the theoretical notion of FDEs was thoroughly investigated, yielding several fundamental discoveries, including the stability theory. In mathematical terms, stability theory is concerned with the convergence of differential equation solutions under minor changes in the original data. The topic of stability is critical in the study of FDEs, and many writers have addressed it (see Ahmed et al. (2007); Gao and Yu (2005); Odibat (2010); Wang et al. (2012)). In any event, nonlinear FDEs are more difficult to analyze for stability than conventional integer-order differential equations. Many authors have been drawn to the study of nonlinear FDE stability theory during the last few decades, and as a result, numerous approaches have been created. However, it is important to emphasize that just a few steps have been carried out to study the durability of FDEs using fixed point theorems. Burton and Zhang (2012) began a thorough investigation of the stability properties of differential equations using fixed point theorems. Following that, several authors used the fixed point method to establish sufficient conditions for the stability of the differential systems (see Ren et al. (2017); Shen et al. (2020)). Based on the above discussions, the exponential stability of FDEs with order  $\tilde{\alpha} \in (\frac{1}{2}, 1)$  is considered through a fixed point approach. It is envisaged that FDEs with fBM will be important for modeling the chaotic behavior of stock prices and financial instruments. The exponential stability of FDEs is an important property in analysis and application in financial systems.

This paper’s main contributions are as follows:

- (i) A nonlinear fractional Large-Scale neutral stochastic delay system (NFSDS) is considered in  $\mathbb{R}^n$  stochastic settings.
- (ii) To determine the existence and uniqueness of a solution, the fixed point theorem and local assumptions on the nonlinear portion are utilized.
- (iii) The stability and exponential stability of a certain NFSDS are established by the use of Hölder inequality and Gronwall’s inequality.

The following assertions outline the paper’s innovations and challenges and future direction:

- (i) Stability and exponential stability results for NFSDS are new in  $\mathbb{R}^n$  stochastic settings.
- (ii) Study of the exponential stability of the proposed system is not easy, taking the norm estimation on nonlinear stochastic and Large-Scale neutral as the terms used in this paper.
- (iii) It is more difficult to validate the system’s weaker assumptions (1).

The following is an outline of the study: In Section 2, the model description and prelims are given. Our major findings are proved in Sections 3 and 4. Finally, Section 5 presents an illustration of the theory and Section 6 draws a conclusion.

## 2. System Description and Preliminaries

Consider the following NFSDS given by

$$\begin{aligned}
 {}^C \tilde{D}^{\tilde{\alpha}} \left[ x_l(t) - \tilde{g}_l(t, x_l(t), x_l(t - \tilde{h}(t))) \right] &= \tilde{A}_l x_l(t) + \tilde{f}_l(t, x_l(t), x_l(t - \tilde{h}(t))) \\
 &+ \int_0^t \tilde{\sigma}_l(s, x_l(s), x_l(s - \tilde{h}(s))) dw(s) \\
 &+ \int_0^t \tilde{\eta}_l(s, x_l(s), x_l(s - \tilde{h}(s))) dw_{(s)}^H, \\
 x_l(t) &= \varphi(t), \quad t \in [-h, 0],
 \end{aligned} \tag{1}$$

where  $t \in [0, T]$ ,  $\frac{1}{2} < \tilde{\alpha} < 1$ ,  $x_l(t) \in \mathbb{R}^{n_l}$  ( $l = 1$  to  $N$ ),  $\exists \sum_{l=1}^N n_l = n$  and  $\tilde{A}_l$  is  $n_l \times n_l$  continuous matrix valued functions. Define  $C^{n_l} = \mathcal{C}([-h, 0], \mathbb{R}^{n_l})$ , a Banach space of continuous functions mapping from  $[-h, 0] \rightarrow \mathbb{R}^{n_l}$ . Define  $[0, T] := J$ , Further,  $\tilde{g}_l : J \times C^{n_l} \times C^{n_l} \rightarrow \mathbb{R}^{n_l}$ ,  $\tilde{f}_l : J \times C^{n_l} \times C^{n_l} \rightarrow \mathbb{R}^{n_l}$ ,  $\tilde{\sigma}_l : J \times C^{n_l} \times C^{n_l} \rightarrow \mathbb{R}^{n_l \times n_l}$ ,  $\tilde{\eta}_l : J \times C^{n_l} \times C^{n_l} \rightarrow$

$\mathbb{R}^{n_1 \times n_1}$  are continuous functions which will be specified in the future. Moreover,  $w_{(s)}^{\mathcal{H}}$  is a fBm with  $\mathcal{H} \in (\frac{1}{2}, 1)$  which is defined by its stochastic representation

$$w_{(s)}^{\mathcal{H}} := \frac{1}{\Gamma(\mathcal{H} + \frac{1}{2})} \left( \int_{-\infty}^0 [(t-s)^{\mathcal{H}-\frac{1}{2}} - (-s)^{\mathcal{H}-\frac{1}{2}}] dw(s) + \int_0^t (t-s)^{\mathcal{H}-\frac{1}{2}} dw(s) \right)$$

here  $\Gamma$  denotes the Gamma function  $\Gamma(\alpha) := \int_0^\infty y^{\alpha-1} \exp(-y) dy$  and  $0 < \mathcal{H} < 1$  is called the Hurst parameter (one can see the connection with the Hurst parameter for self-similar processes).

Let us consider a probability space  $(\Omega, \mathcal{F}, P)$  with a probability measure  $P$  and  $w(t) = (w_1(t), w_2(t), \dots, w_n(t))^T$  be an  $n$ -dimensional Wiener process defined on  $(\Omega, \mathcal{F}, P)$ . Let  $\{\mathcal{F}_t / t \in J\}$  be the filtration generated by  $\{w(s), w_{(s)}^{\mathcal{H}} : 0 \leq s \leq t\}$  defined on  $(\Omega, \mathcal{F}, P)$ . Let  $L_2(\Omega, \mathcal{F}_t, \mathbb{R}^{n_1})$  denote the Hilbert space of all  $\mathcal{F}_t$ -measurable square integrable random variables with values in  $\mathbb{R}^{n_1}$ . Let  $L_2^{\mathcal{F}}(J, \mathbb{R}^{n_1})$  be the Hilbert space of all square integrable and  $\mathcal{F}_t$ -measurable processes with values of  $\mathbb{R}^{n_1}$ . Let  $\mathcal{B} = \{x_1(t) : x_1(t) \in C(J, L_2(\Omega, \mathcal{F}_t, \mathbb{R}^{n_1}))\}$  be a Banach space of all continuous square integrable and  $\mathcal{F}_t$ -adapted processes with norm  $\|x_1\|^2 = \sup_{t \in J} \mathbb{E} \|x_1(t)\|^2$  and  $\|\varphi\|^2 = \max\{\mathbb{E} \|\varphi(t)\|^2 : t \in [-h, 0]\}$  for any  $t \geq 0$ , any given  $\varphi \in C([-h, 0], \mathbb{R}^{n_1})$  denotes the Banach space of continuous functions mapping from  $[-h, 0]$  to  $\mathbb{R}^{n_1}$ . For more details on fractional calculus definitions, stochastic theory and fBm, one can read our published paper Balasubramaniam et al. (2020); Sathiyaraj and Balasubramaniam (2018); Sathiyaraj et al. (2019).

**Definition 1.** The Riemann–Liouville fractional operators (left sided) for  $\tilde{n} - 1 < \tilde{\alpha} < \tilde{n}$  for  $f_l : [0, \infty) \rightarrow \mathbb{R}$  are as follows:

$$(I_{0+}^{\tilde{\alpha}} f_l)(\tilde{x}_l) = \frac{1}{\Gamma(\tilde{\alpha})} \int_0^{\tilde{x}_l} (\tilde{x}_l - t)^{\tilde{\alpha}-1} f_l(t) dt.$$

$$(D_{0+}^{\tilde{\alpha}} f_l)(\tilde{x}_l) = D^{\tilde{n}} (I_{0+}^{\tilde{n}-\tilde{\alpha}} f_l)(\tilde{x}_l).$$

**Definition 2.** Podlubny (1998): The Caputo derivative for  $\tilde{n} - 1 < \tilde{\alpha} < \tilde{n}$  for  $f_l : [0, \infty) \rightarrow \mathbb{R}$  is as follows:

$${}^c D_t^{\tilde{\alpha}} f_l(t) = \frac{1}{\Gamma(\tilde{n} - \tilde{\alpha})} \int_0^t \frac{f_l^{\tilde{n}}(s)}{(t-s)^{\tilde{\alpha}-\tilde{n}+1}} ds.$$

and its Laplace transform is

$$\mathcal{L}\{{}^c D_t^{\tilde{\alpha}} f_l(t)\}(s) = s^{\tilde{\alpha}} f_l(s) - \sum_{l=0}^{\tilde{n}-1} f_l^{(l)}(0^+) s^{\tilde{\alpha}-1-l}.$$

**Definition 3.** Podlubny (1998): The two-parameter family of Mittag–Leffler function is given by

$$\mathcal{E}_{\tilde{\alpha}, \beta}(z) = \sum_{l=0}^{\infty} \frac{z^l}{\Gamma(l\tilde{\alpha} + \beta)} \quad \text{for } \tilde{\alpha}, \beta > 0.$$

The general Mittag–Leffler function satisfies the below identity

$$\int_0^\infty e^{-t} t^{\beta-1} \mathcal{E}_{\tilde{\alpha}, \beta}(t^{\tilde{\alpha}} z) dt = \frac{1}{1-z} \quad \text{for } |z| < 1.$$

The Laplace transform of two-parameter Mittag–Leffler function  $\mathcal{E}_{\tilde{\alpha},\beta}(z)$  is described using the following integral

$$\int_0^\infty e^{-st} t^{\beta-1} \mathcal{E}_{\tilde{\alpha},\beta}(\pm at^{\tilde{\alpha}}) dt = \frac{s^{\tilde{\alpha}-\beta}}{(s^{\tilde{\alpha}} \mp a)}.$$

That is,  $\mathcal{L}\{t^{\beta-1} \mathcal{E}_{\tilde{\alpha},\beta}(\pm at^{\tilde{\alpha}})\}(s) = \frac{s^{\tilde{\alpha}-\beta}}{(s^{\tilde{\alpha}} \mp a)}$ .

**Lemma 1.** Kreyszig (1978): Suppose that the bounded linear operator  $A_I : \mathbb{R}^{n_I} \rightarrow \mathbb{R}^{n_I}$  is determined on a Banach space. Take that  $\|A_I\| < 1$ . Then  $(I - A_I)^{-1}$  is linear and bounded,  $(I - A_I)^{-1} = \sum_{i=0}^\infty A_I^i$ . Then,  $\|(I - A_I)^{-1}\| \leq (1 - \|A_I\|)^{-1}$ .

**Lemma 2.** Mao (1997): Let  $\tilde{g}_I \in \mathcal{M}^2(J; \mathbb{R}^{d \times m}) \ni$

$$\mathbb{E} \int_0^T |\tilde{\sigma}_I(s)|^p ds < \infty. \text{ Then, } \mathbb{E} \left| \int_0^T \tilde{\sigma}_I(s) dB(s) \right|^p \leq \left( \frac{p(p-1)}{2} \right)^{\frac{p}{2}} T^{\frac{p-2}{2}} \mathbb{E} \int_0^T |\tilde{\sigma}_I(s)|^p ds$$

where  $p \geq 2$ .

**Lemma 3.** Applebaum (2009): For any  $p \geq 2$ , there exists  $\tilde{\mathcal{A}}_k > 0$ , such that

$$\mathbb{E} \sup_{s \in [0,t]} \left\| \int_0^s \int_{-\infty}^{+\infty} \tilde{g}_k(v,z) \hat{N}(dv,dz) \right\|^p \leq \tilde{\mathcal{A}}_k \left\{ \mathbb{E} \left[ \left( \int_0^t \int_{-\infty}^{+\infty} \|\tilde{g}_k(s,z)\|^2 \kappa(dz) ds \right)^{\frac{p}{2}} \right] + \mathbb{E} \left[ \int_0^t \int_{-\infty}^{+\infty} \|\tilde{g}_k(s,z)\|^p \kappa(dz) ds \right] \right\}.$$

**Definition 4.** A normalized fBm  $w^{\mathcal{H}} = \{w_{(t)}^{\mathcal{H}} : 0 \leq t < \infty\}$  with  $0 < \mathcal{H} < 1$  on  $(\Omega, \mathcal{F}, P)$  is uniquely characterized by the following properties:

- $w_{(t)}^{\mathcal{H}}$  has stationary increments;
- $w_{(0)}^{\mathcal{H}} = 0$ , and  $\mathbb{E}w_{(t)}^{\mathcal{H}} = 0$  for  $t \geq 0$ ;
- $w_{(t)}^{\mathcal{H}}$  has a Gaussian distribution for  $t > 0$ .

From the above three properties, it follows that the covariance function is given by

$$R_{\mathcal{H}}(s,t) = \mathbb{E}\left(w_{(s)}^{\mathcal{H}} w_{(t)}^{\mathcal{H}}\right) = \frac{1}{2} \left\{ t^{2\mathcal{H}} + s^{2\mathcal{H}} - |t-s|^{2\mathcal{H}} \right\} \text{ for } 0 < s \leq t.$$

**Definition 5.** Seemab and Rehman (2018): The solution  $x_1(t) = \varphi(t)$  of (1) is called stable, if for every  $\epsilon > 0$  and  $t_0 \geq 0$ ,  $\exists \delta = \delta(t_0, \epsilon) > 0 \ni |x_1(t, x_{10}, t_0) - \varphi(t)| < \epsilon$  for  $|x_{10} - \varphi(t_0)| \leq \delta(t_0, \epsilon)$  and all  $t \geq t_0$ .

**Definition 6.** Equation (1) is said to be exponentially stable if  $\exists \mu$  is positive,  $1 \leq M^* \ni t \geq 0$ ,

$$\mathbb{E}\|x_1(t)\|^2 \leq M^* e^{-\mu t}.$$



The solution of Equation (1) can be explained as follows

$$\begin{aligned} x_1(t) = & \mathcal{E}_{\tilde{\alpha}}(\tilde{\mathcal{A}}_I t^{\tilde{\alpha}}) \left[ \varphi(0) + \tilde{g}_1(0, \varphi(0)) \right] + \tilde{g}_1(t, x_1(t), x_1(t - \tilde{h}(t))) \\ & + \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \tilde{f}_1(s, x_1(s), x_1(s - \tilde{h}(s))) ds \\ & + \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \left[ \int_0^s \tilde{\sigma}_1(\tilde{\tau}, x_1(\tilde{\tau}), x_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) d\omega(\tilde{\tau}) \right] ds \\ & + \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \tilde{\mathcal{A}}_I \tilde{g}_1(s, x_1(s), x_1(s - \tilde{h}(s))) ds \\ & + \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \left[ \int_0^s \tilde{\eta}_1(\tilde{\tau}, x_1(\tilde{\tau}), x_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) d\omega_{\tilde{\tau}}^{\mathcal{H}} \right] ds. \end{aligned}$$

### 3. Existence and Uniqueness of Solutions

In this section, we show the existence and uniqueness of solutions and stability results. As a result, we establish the below hypothesis:

(H<sub>1</sub>) For  $\tilde{f}_1, \tilde{\sigma}_1, \tilde{g}_1 \exists q > 1$  (constant) and  $V_{\tilde{f}_1}(\cdot), V_{\tilde{\sigma}_1}(\cdot)$  and  $V_{\tilde{g}_1}(\cdot) \in L^q(J, \mathbb{R}^+)$   $\ni$

- (i)  $\mathbb{E} \|\tilde{f}_1(t, x_1(t), x_1(t - \tilde{h}(t))) - \tilde{f}_1(t, y_1(t), y_1(t - \tilde{h}(t)))\|^2 \leq V_{\tilde{f}_1}(t) \mathbb{E} \|x_1(t) - y_1(t)\|^2$
- (ii)  $\mathbb{E} \|\tilde{\sigma}_1(t, x_1(t), x_1(t - \tilde{h}(t))) - \tilde{\sigma}_1(t, y_1(t), y_1(t - \tilde{h}(t)))\|^2 \leq V_{\tilde{\sigma}_1}(t) \mathbb{E} \|x_1(t) - y_1(t)\|^2$
- (iii)  $\mathbb{E} \|\tilde{g}_1(t, x_1(t), x_1(t - \tilde{h}(t))) - \tilde{g}_1(t, y_1(t), y_1(t - \tilde{h}(t)))\|^2 \leq V_{\tilde{g}_1}(t) \mathbb{E} \|x_1(t) - y_1(t)\|^2$ .
- (iv)  $\mathbb{E} \left\| \int_0^t \tilde{\eta}_1(\tilde{\tau}, x_1(\tilde{\tau}), x_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) d\omega_{\tilde{\tau}}^{\mathcal{H}} - \int_0^t \tilde{\eta}_1(\tilde{\tau}, y_1(\tilde{\tau}), y_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) d\omega_{\tilde{\tau}}^{\mathcal{H}} \right\|^2 \leq 2\mathcal{H}t^{2\mathcal{H}-1} \int_0^s V_{\tilde{\eta}_1}(t) \mathbb{E} \|x_1(t) - y_1(t)\|_{L^2}^2 ds$ .

(H<sub>2</sub>) The below properties are true, for  $t \geq 0, N_1, N_2 \geq 1$

- (i)  $\|\mathcal{E}_{\tilde{\alpha}}(\tilde{\mathcal{A}}_I t^{\tilde{\alpha}})\| \leq N_1 e^{-\omega t}$ .
- (ii)  $\|\mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}})\| \leq N_2 e^{-\omega(t-s)}$ .

(H<sub>3</sub>)  $\exists \hat{V}_{\tilde{f}_1}, \hat{V}_{\tilde{\sigma}_1}$  (constants), and  $\hat{V}_{\tilde{g}_1} \ni$

- (i)  $\mathbb{E} \|\tilde{f}_1(t, x_1(t), x_1(t - \tilde{h}(t)))\|^2 \leq \hat{V}_{\tilde{f}_1} (1 + \mathbb{E} \|x_1(t)\|^2)$
- (ii)  $\mathbb{E} \|\tilde{\sigma}_1(t, x_1(t), x_1(t - \tilde{h}(t)))\|^2 \leq \hat{V}_{\tilde{\sigma}_1} (1 + \mathbb{E} \|x_1(t)\|^2)$
- (iii)  $\mathbb{E} \|\tilde{g}_1(t, x_1(t), x_1(t - \tilde{h}(t)))\|^2 \leq \hat{V}_{\tilde{g}_1} (1 + \mathbb{E} \|x_1(t)\|^2)$ .
- (iv)  $\mathbb{E} \left\| \int_0^t \tilde{\eta}_1(\tilde{\tau}, x_1(\tilde{\tau}), x_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) d\omega_{\tilde{\tau}}^{\mathcal{H}} \right\|^2 \leq 2\mathcal{H}t^{2\mathcal{H}-1} \int_0^t V_{\tilde{\eta}_1}(s) \mathbb{E} \|1 + x_1(s)\|_{L^2}^2 ds$ .

In addition, we set

$$\begin{aligned} Q_1 = & 5\hat{V}_{\tilde{g}_1} + 10N_2 \left( \frac{1 - e^{-2p\omega T}}{2p\omega} \right)^{\frac{1}{p}} \left[ \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} \|V_{\tilde{f}_1}\|_{L^q(J, \mathbb{R}^+)} + \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \|V_{\tilde{\sigma}_1}\|_{L^q(J, \mathbb{R}^+)} \right. \\ & \left. + \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} \tilde{\mathcal{A}}_I \|V_{\tilde{g}_1}\|_{L^q(J, \mathbb{R}^+)} + 2\mathcal{H}t^{2\mathcal{H}-1} \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \|V_{\tilde{\eta}_1}\|_{L^q(J, \mathbb{R}^+)} \right] \\ Q_2 = & 5\hat{V}_{\tilde{g}_1} + 10N_2 \left( \frac{1 - e^{-2\omega T}}{2\omega} \right) \left[ \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} R_{\tilde{f}_1} + \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} R_{\tilde{\sigma}_1} + \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} \tilde{\mathcal{A}}_I R_{\tilde{g}_1} + 2\mathcal{H}t^{2\mathcal{H}-1} \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} R_{\tilde{\eta}_1} \right]. \end{aligned}$$

Here, we take  $R_{\tilde{f}_1} = \sup_{t \in J} \mathbb{E} \|\tilde{f}_1(t, 0, 0)\|^2$ ,  $R_{\tilde{\sigma}_1} = \sup_{t \in J} \mathbb{E} \|\tilde{\sigma}_1(t, 0, 0)\|^2$ ,  $R_{\tilde{g}_1} = \sup_{t \in J} \mathbb{E} \|\tilde{g}_1(t, 0, 0)\|^2$  and  $R_{\tilde{\eta}_1} = \sup_{t \in J} \mathbb{E} \|\tilde{\eta}_1(t, 0, 0)\|^2$ .

**Theorem 1.** Consider hypothesis  $(H_1)$  and  $(H_2)$  are true; then (1) has at least one solution provided that

$$M_2 := 4V_{\tilde{g}_1} + 4N_2 \left( \frac{1 - e^{-2p\omega T}}{2p\omega} \right)^{\frac{1}{p}} \left[ \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} \|V_{\tilde{f}_1}\|_{L^q(J, \mathbb{R}^+)} + \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \|V_{\tilde{\sigma}_1}\|_{L^q(J, \mathbb{R}^+)} \right. \\ \left. + \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} \tilde{\mathcal{A}}_1 \|V_{\tilde{g}_1}\|_{L^q(J, \mathbb{R}^+)} + 2\mathcal{H}t^{2\mathcal{H}-1} \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \|V_{\tilde{\eta}_1}\|_{L^q(J, \mathbb{R}^+)} \right] < 1, \tag{2}$$

where  $\frac{1}{p} + \frac{1}{q} = 1$ ,  $p, q > 1$  and  $x_1 \equiv 0$  (the trivial solution) of Equation (1) are stable in  $\mathcal{B}$ .

**Proof.** For each  $r \geq 0$ , define  $\mathcal{B}_r = \{x_1(t) : x_1(t) \in \mathcal{B}; \mathbb{E} \|x_1(t)\|^2 \leq r\}$  and then for each  $r$ ,  $\mathcal{B}_r$  is a bounded, closed and convex subset of  $\mathcal{B}$ . Define the operator  $\Phi : \mathcal{B}_r \rightarrow \mathcal{B}_r$

$$(\Phi x_1)(t) = \mathcal{E}_{\tilde{\alpha}}(\tilde{\mathcal{A}}_1 t^{\tilde{\alpha}}) \left[ \varphi(0) + \tilde{g}_1(0, \varphi(0)) \right] + \tilde{g}_1(t, x_1(t), x_1(t - \tilde{h}(t))) \\ + \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_1(t-s)^{\tilde{\alpha}}) \tilde{f}_1(s, x_1(s), x_1(s - \tilde{h}(s))) ds \\ + \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_1(t-s)^{\tilde{\alpha}}) \left[ \int_0^s \tilde{\sigma}_1(\tilde{\tau}, x_1(\tilde{\tau}), x_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) d\omega(\tilde{\tau}) \right] ds \\ + \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_1(t-s)^{\tilde{\alpha}}) \tilde{\mathcal{A}}_1 \tilde{g}_1(s, x_1(s), x_1(s - \tilde{h}(s))) ds \\ + \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_1(t-s)^{\tilde{\alpha}}) \left[ \int_0^s \tilde{\eta}_1(\tilde{\tau}, x_1(\tilde{\tau}), x_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) d\omega_{\tilde{\tau}}^{\mathcal{H}} \right] ds.$$

**Step I:** To prove that  $\exists r \geq 0 \ni \Phi(\mathcal{B}_r) \subseteq \mathcal{B}_r$ . Based on  $(H_1)$ ,  $(H_2)$  and Hölder inequality, we get

$$\mathbb{E} \left\| \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_1(t-s)^{\tilde{\alpha}}) \tilde{f}_1(s, x_1(s), x_1(s - \tilde{h}(s))) ds \right\|^2 \\ \leq \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} N_2 \int_0^t e^{-2\omega(t-s)} \mathbb{E} \|\tilde{f}_1(s, x_1(s), x_1(s - \tilde{h}(s))) - \tilde{f}_1(s, 0, 0) + \tilde{f}_1(s, 0, 0)\|^2 ds \\ \leq 2 \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} N_2 \left\{ \int_0^t e^{-2\omega(t-s)} V_{\tilde{f}_1}(s) \mathbb{E} \|x_1(s)\|^2 ds + \int_0^t e^{-2\omega(t-s)} \mathbb{E} \|\tilde{f}_1(s, 0, 0)\|^2 ds \right\} \\ \leq 2 \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} N_2 \left\{ \left( \int_0^t e^{-2p\omega(t-s)} ds \right)^{\frac{1}{p}} \left( \int_0^t V_{\tilde{f}_1}^q(s) ds \right)^{\frac{1}{q}} \mathbb{E} \|x_1\|^2 + R_{\tilde{f}_1} \int_0^t e^{-2\omega(t-s)} ds \right\} \\ \leq 2 \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} N_2 \left\{ \left( \frac{1 - e^{-2p\omega T}}{2p\omega} \right)^{\frac{1}{p}} \|V_{\tilde{f}_1}\|_{L^q(J, \mathbb{R}^+)} r + R_{\tilde{f}_1} \left( \frac{1 - e^{-2\omega T}}{2\omega} \right) \right\}.$$

Similarly,

$$\begin{aligned} & \mathbb{E} \left\| \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \left[ \int_0^s \tilde{\sigma}_I(\tilde{\tau}, x_1(\tilde{\tau}), x_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) dw(\tilde{\tau}) \right] ds \right\|^2 \\ & \leq 2 \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} N_2 \left\{ \left( \frac{1 - e^{-2p\omega T}}{2p\omega} \right)^{\frac{1}{p}} \|V_{\tilde{\sigma}_I}\|_{L^q(J, \mathbb{R}^+)} r + R_{\tilde{\sigma}_I} \left( \frac{1 - e^{-2\omega T}}{2\omega} \right) \right\}, \end{aligned}$$

$$\begin{aligned} & \mathbb{E} \left\| \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \tilde{\mathcal{A}}_I \tilde{g}_I(s, x_1(s), x_1(s - \tilde{h}(s))) ds \right\|^2 \\ & \leq 2 \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} N_2 \tilde{\mathcal{A}}_I \left\{ \left( \frac{1 - e^{-2p\omega T}}{2p\omega} \right)^{\frac{1}{p}} \|V_{\tilde{g}_I}\|_{L^q(J, \mathbb{R}^+)} r + R_{\tilde{g}_I} \left( \frac{1 - e^{-2\omega T}}{2\omega} \right) \right\} \end{aligned}$$

and

$$\begin{aligned} & \mathbb{E} \left\| \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \left[ \int_0^s \tilde{\eta}_I(\tilde{\tau}, x_1(\tilde{\tau}), x_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) d\mathcal{W}_{(\tilde{\tau})}^{\mathcal{H}} \right] ds \right\|^2 \\ & \leq 4\mathcal{H}t^{2\mathcal{H}-1} \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} N_2 \left\{ \left( \frac{1 - e^{-2p\omega T}}{2p\omega} \right)^{\frac{1}{p}} \|V_{\tilde{\eta}_I}\|_{L^q(J, \mathbb{R}^+)} r + R_{\tilde{\eta}_I} \left( \frac{1 - e^{-2\omega T}}{2\omega} \right) \right\} \end{aligned}$$

Now,

$$\begin{aligned} \mathbb{E} \|(\Phi_{x_1})(t)\|^2 & \leq 5 \left\{ \mathbb{E} \|\mathcal{E}_{\tilde{\alpha}}(\tilde{\mathcal{A}}_I t^{\tilde{\alpha}}) [\varphi(0) + \tilde{g}_I(0, \varphi(0))]\|^2 + \mathbb{E} \|\tilde{g}_I(t, x_1(t), x_1(t - \tilde{h}(t)))\|^2 \right. \\ & + \mathbb{E} \left\| \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \tilde{f}_I(s, x_1(s), x_1(s - \tilde{h}(s))) ds \right\|^2 \\ & + \mathbb{E} \left\| \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \left[ \int_0^s \tilde{\sigma}_I(\tilde{\tau}, x_1(\tilde{\tau}), x_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) dw(\tilde{\tau}) \right] ds \right\|^2 \\ & + \mathbb{E} \left\| \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \tilde{\mathcal{A}}_I \tilde{g}_I(s, x_1(s), x_1(s - \tilde{h}(s))) ds \right\|^2 \\ & \left. + \mathbb{E} \left\| \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \left[ \int_0^s \tilde{\eta}_I(\tilde{\tau}, x_1(\tilde{\tau}), x_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) d\mathcal{W}_{(\tilde{\tau})}^{\mathcal{H}} \right] ds \right\|^2 \right\} \\ & \leq 5 \left\{ N_1 e^{-2\omega T} \mathbb{E} \|\varphi(0) + \tilde{g}_I(0, \varphi(0))\|^2 + \hat{V}_{\tilde{g}_I} (1 + \mathbb{E} \|x_1(t)\|^2) \right. \\ & + 2 \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} N_2 \left[ \left( \frac{1 - e^{-2p\omega T}}{2p\omega} \right)^{\frac{1}{p}} \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} r + R_{\tilde{f}_I} \left( \frac{1 - e^{-2\omega T}}{2\omega} \right) \right] \\ & \left. + 2 \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} N_2 \left[ \left( \frac{1 - e^{-2p\omega T}}{2p\omega} \right)^{\frac{1}{p}} \|V_{\tilde{\sigma}_I}\|_{L^q(J, \mathbb{R}^+)} r + R_{\tilde{\sigma}_I} \left( \frac{1 - e^{-2\omega T}}{2\omega} \right) \right] \right\} \end{aligned}$$

$$\begin{aligned}
 & + 2 \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} N_2 \tilde{\mathcal{A}}_I \left[ \left( \frac{1-e^{-2p\omega T}}{2p\omega} \right)^{\frac{1}{p}} \|V_{\tilde{g}_1}\|_{L^q(J, \mathbb{R}^+)} r + R_{\tilde{g}_1} \left( \frac{1-e^{-2\omega T}}{2\omega} \right) \right] \\
 & + 4\mathcal{H}t^{2\mathcal{H}-1} \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} N_2 \left\{ \left( \frac{1-e^{-2p\omega T}}{2p\omega} \right)^{\frac{1}{p}} \|V_{\tilde{\eta}_1}\|_{L^q(J, \mathbb{R}^+)} r + R_{\tilde{\eta}_1} \left( \frac{1-e^{-2\omega T}}{2\omega} \right) \right\} \\
 \leq & 5N_1 e^{-2\omega T} \mathbb{E} \|\varphi(0) + \tilde{g}_1(0, \varphi(0))\|^2 + 5\widehat{V}_{\tilde{g}_1} + 10N_2 \left\{ \left( \frac{1-e^{-2p\omega T}}{2p\omega} \right)^{\frac{1}{p}} \right. \\
 & \times \left[ \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} \|V_{\tilde{f}_1}\|_{L^q(J, \mathbb{R}^+)} + \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \|V_{\tilde{\sigma}_1}\|_{L^q(J, \mathbb{R}^+)} + \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} \tilde{\mathcal{A}}_I \|V_{\tilde{g}_1}\|_{L^q(J, \mathbb{R}^+)} \right. \\
 & \left. \left. + 2\mathcal{H}t^{2\mathcal{H}-1} \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \|V_{\tilde{\eta}_1}\|_{L^q(J, \mathbb{R}^+)} + \right\} r \\
 & + 5\widehat{V}_{\tilde{g}_1} + 10N_2 \left( \frac{1-e^{-2\omega T}}{2\omega} \right) \left[ \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} R_{\tilde{f}_1} + \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} R_{\tilde{\sigma}_1} + \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} \tilde{\mathcal{A}}_I R_{\tilde{g}_1} + 2\mathcal{H}t^{2\mathcal{H}-1} \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} R_{\tilde{\eta}_1} \right] \\
 \leq & 5N_1 e^{-2\omega T} \mathbb{E} \|\varphi(0) + \tilde{g}_1(0, \varphi(0))\|^2 + Q_2 + Q_1 r = r.
 \end{aligned}$$

For,  $r = \frac{5N_1 e^{-2\omega T} \mathbb{E} \|\varphi(0) + \tilde{g}_1(0, \varphi(0))\|^2 + Q_2}{(1-Q_1)}$ ,  $Q_1 < 1$ . Hence, we obtain  $\Phi(\mathcal{B}_r) \subseteq \mathcal{B}_r$  for such an  $r$ .

**Step II.** To prove that  $\Phi$  is a contraction.

Assume  $x_1, y_1 \in \mathcal{B}_r$ . Using,  $(H_1), (H_2)$  and Hölder inequality, for every  $t \in J$ , we get

$$\begin{aligned}
 & \mathbb{E} \|(\Phi_{x_1})(t) - (\Phi_{y_1})(t)\|^2 \\
 = & \mathbb{E} \left\{ \left\| \mathcal{E}_{\tilde{\alpha}}(\tilde{\mathcal{A}}_I t^{\tilde{\alpha}}) [\varphi(0) + \tilde{g}_1(0, \varphi(0))] + \tilde{g}_1(s, x_1(s), x_1(s - \tilde{h}(s))) \right. \right. \\
 & + \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \tilde{f}_1(s, x_1(s), x_1(s - \tilde{h}(s))) ds \\
 & + \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \left[ \int_0^s \tilde{\sigma}_1(\tilde{\tau}, x_1(\tilde{\tau}), x_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) d\omega(\tilde{\tau}) \right] ds \\
 & + \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \tilde{\mathcal{A}}_I \tilde{g}_1(s, x_1(s), x_1(s - \tilde{h}(s))) ds \\
 & + \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \left[ \int_0^s \tilde{\eta}_1(\tilde{\tau}, x_1(\tilde{\tau}), x_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) d\omega_{\mathcal{H}}^{\mathcal{H}}(\tilde{\tau}) \right] ds \\
 & - \mathcal{E}_{\tilde{\alpha}}(\tilde{\mathcal{A}}_I t^{\tilde{\alpha}}) [\varphi(0) + \tilde{g}_1(0, \varphi(0))] - \tilde{g}_1(s, y_1(s), y_1(s - \tilde{h}(s))) \\
 & - \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \tilde{f}_1(s, y_1(s), y_1(s - \tilde{h}(s))) ds \\
 & - \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \left[ \int_0^s \tilde{\sigma}_1(\tilde{\tau}, y_1(\tilde{\tau}), y_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) d\omega(\tilde{\tau}) \right] ds \\
 & \left. \left. - \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \tilde{\mathcal{A}}_I \tilde{g}_1(s, y_1(s), y_1(s - \tilde{h}(s))) ds \right\} \right.
 \end{aligned}$$

$$\begin{aligned}
 & - \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \left[ \int_0^s \tilde{\eta}_I(\tilde{\tau}, y_1(\tilde{\tau}), y_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) d\omega_{(\tilde{\tau})}^{\mathcal{H}} \right] ds \Bigg\|^2 \\
 \leq & 4 \left\{ \mathbb{E} \left\| \tilde{g}_I(s, x(s), x(s - \tilde{h}(s))) - \tilde{g}_I(s, y_1(s), y_1(s - \tilde{h}(s))) \right\|^2 \right. \\
 & + \mathbb{E} \left\| \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) [\tilde{f}_I(s, x_1(s), x_1(s - \tilde{h}(s))) - \tilde{f}_I(s, y_1(s), y_1(s - \tilde{h}(s)))] ds \right\|^2 \\
 & + \mathbb{E} \left\| \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \right. \\
 & \times \left. \left[ \int_0^s (\tilde{\sigma}_I(\tilde{\tau}, x_1(\tilde{\tau}), x_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) - \tilde{\sigma}_I(\tilde{\tau}, y_1(\tilde{\tau}), y_1(\tilde{\tau} - \tilde{h}(\tilde{\tau})))) d\omega(\tilde{\tau}) \right] ds \right\|^2 \\
 & + \mathbb{E} \left\| \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \tilde{\mathcal{A}}_I \right. \\
 & \times \left. (\tilde{g}_I(s, x_1(s), x_1(s - \tilde{h}(s))) - \tilde{g}_I(s, y_1(s), y_1(s - \tilde{h}(s)))) ds \right. \\
 & + \left. \int_0^t (t-s)^{\tilde{\alpha}-1} \mathcal{E}_{\tilde{\alpha}, \tilde{\alpha}}(\tilde{\mathcal{A}}_I(t-s)^{\tilde{\alpha}}) \right. \\
 & \times \left. \left[ \int_0^s (\tilde{\eta}_I(\tilde{\tau}, x_1(\tilde{\tau}), x_1(\tilde{\tau} - \tilde{h}(\tilde{\tau}))) - \tilde{\eta}_I(\tilde{\tau}, y_1(\tilde{\tau}), y_1(\tilde{\tau} - \tilde{h}(\tilde{\tau})))) d\omega_{(\tilde{\tau})}^{\mathcal{H}} \right] ds \right\|^2 \\
 \leq & 4 \|V_{\tilde{g}_I}\|_{L^q(J, \mathbb{R}^+)} + 4N_2 \left( \frac{1 - e^{-2p\omega T}}{2p\omega} \right)^{\frac{1}{p}} \left[ \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} \|V_{\tilde{f}_I}\|_{L^q(J, \mathbb{R}^+)} + \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \|V_{\tilde{\sigma}_I}\|_{L^q(J, \mathbb{R}^+)} \right. \\
 & \left. + \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} \tilde{\mathcal{A}}_I \|V_{\tilde{g}_I}\|_{L^q(J, \mathbb{R}^+)} + 2\mathcal{H}t^{2\mathcal{H}-1} \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \|V_{\tilde{\eta}_I}\|_{L^q(J, \mathbb{R}^+)} \right] \mathbb{E} \|x_1(t) - y_1(t)\|^2,
 \end{aligned}$$

which reveals that

$$\mathbb{E} \|(\Phi x_1)(t) - (\Phi y_1)(t)\|^2 \leq M_2 \mathbb{E} \|x_1 - y_1\|^2.$$

Using (2), we conclude that  $M_2 < 1$ , which implies  $\Phi$  is a contraction mapping with a unique fixed point  $x_1(t) \in \mathcal{B}_r$ , which is a solution of (1). Now, we prove the stability conditions of (1)

For any given  $\varepsilon > 0$ ,  $\exists \lambda = \frac{\varepsilon(1-Q_1)-Q_2}{5N_1e^{-2\omega T}} \ni \|\varphi(0) + \tilde{g}_I(0, \varphi(0))\|^2 \leq \lambda$ , which implies

$$\begin{aligned} \mathbb{E}\|x_1(t)\|^2 &\leq 5N_1 e^{-2\omega T} \mathbb{E}\|\varphi(0) + \tilde{g}_1(0, \varphi(0))\|^2 + 10N_2 \left\{ \left( \frac{1 - e^{-2p\omega T}}{2p\omega} \right)^{\frac{1}{p}} \left[ \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} \|V_{\tilde{f}_1}\|_{L^q(J, \mathbb{R}^+)} \right. \right. \\ &\quad \left. \left. + \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \|V_{\tilde{\sigma}_1}\|_{L^q(J, \mathbb{R}^+)} + \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} \tilde{\mathcal{A}}_l \|V_{\tilde{g}_1}\|_{L^q(J, \mathbb{R}^+)} + 2\mathcal{H}t^{2\mathcal{H}-1} \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \|V_{\tilde{\eta}_1}\|_{L^q(J, \mathbb{R}^+)} \right] \right\} r \\ &\quad + 10N_2 \left( \frac{1 - e^{-2\omega T}}{2\omega} \right) \left[ \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} R_{\tilde{f}_1} + \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} R_{\tilde{\sigma}_1} + \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} \tilde{\mathcal{A}}_l R_{\tilde{g}_1} + 2\mathcal{H}t^{2\mathcal{H}-1} \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} R_{\tilde{\eta}_1} \right] \\ &\leq 5N_1 e^{-2\omega T} \lambda + Q_1 r + Q_2 \\ r(1 - Q_1) &\leq 5N_1 e^{-2\omega T} \lambda + Q_2 \\ r &\leq \epsilon. \end{aligned}$$

Thus, the proof is over.  $\square$

#### 4. Exponential Stability

**Theorem 2.** *If hypotheses (H<sub>2</sub>) – (H<sub>3</sub>) are true, then (1) is exponentially stable, provided that*

$$\omega > \beta = N_2 \left[ \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} (\hat{V}_{\tilde{f}_1} + \tilde{\mathcal{A}}_l \hat{V}_{\tilde{g}_1})(1+r) + \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \hat{V}_{\tilde{\sigma}_1}(1+r) + 2\mathcal{H}t^{2\mathcal{H}-1} \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \hat{V}_{\tilde{\eta}_1}(1+r) \right]. \tag{3}$$

**Proof.**

$$\begin{aligned} \mathbb{E}\|x_1(t)\|^2 &\leq 5e^{-2\omega t} N_1 \mathbb{E}\|\varphi(0) + \tilde{g}_1(0, \varphi(0))\|^2 + 5N_2 e^{-2\omega t} \left[ \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} (1+r) [\hat{V}_{\tilde{f}_1} + \tilde{\mathcal{A}}_l \hat{V}_{\tilde{g}_1}] \right. \\ &\quad \left. + \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \hat{V}_{\tilde{\sigma}_1}(1+r) + 2\mathcal{H}t^{2\mathcal{H}-1} \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \hat{V}_{\tilde{\eta}_1}(1+r) \right] \int_0^t e^{2\omega s} ds \\ \mathbb{E}\|x_1(t)\|^2 e^{2\omega t} &\leq 5N_1 \mathbb{E}\|\varphi(0) + \tilde{g}_1(0, \varphi(0))\|^2 + 5N_2 \left[ \frac{T^{2\tilde{\alpha}-1}}{2\tilde{\alpha}-1} (1+r) [\hat{V}_{\tilde{f}_1} + \tilde{\mathcal{A}}_l \hat{V}_{\tilde{g}_1}] \right. \\ &\quad \left. + \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \hat{V}_{\tilde{\sigma}_1}(1+r) + 2\mathcal{H}t^{2\mathcal{H}-1} \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \hat{V}_{\tilde{\eta}_1}(1+r) \right] \int_0^t e^{2\omega s} ds. \end{aligned}$$

We get the result by using the Gronwall’s inequality

$$\begin{aligned} e^{2\omega t} \mathbb{E}\|x_1(t)\|^2 &\leq 5N_1 \mathbb{E}\|\varphi(0) + \tilde{g}_1(0, \varphi(0))\|^2 \\ &\quad \times \exp \left( 5N_2 \left[ (\hat{V}_{\tilde{f}_1} + \tilde{\mathcal{A}}_l \hat{V}_{\tilde{g}_1})(1+r) + \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \hat{V}_{\tilde{\sigma}_1}(1+r) + 2\mathcal{H}t^{2\mathcal{H}-1} \frac{T^{2\tilde{\alpha}}}{\tilde{\alpha}^2} \hat{V}_{\tilde{\eta}_1}(1+r) \right] t \right). \end{aligned}$$

Therefore,

$$\mathbb{E}\|x_1(t)\|^2 \leq M \mathbb{E}\|\varphi(0) + \tilde{g}_1(0, \varphi(0))\|^2 \exp((-vt)).$$

where  $v = 2\omega - 5\beta$ ,  $M = 5N_1$ . Thus, according to (3), (1) is exponentially stable in  $\mathcal{B}$ . Thus, the proof is over.  $\square$

**Remark 1.** *Existence, uniqueness, and stability of mild solutions for second-order neutral stochastic evolution equations with infinite delay and Poisson jumps by the authors in Ren and Sakthivel (2012) using successive approximation techniques. The uniqueness and existence of solutions, in*

addition to their controllability (relative), have been demonstrated using the fixed point approach in Sathiyaraj and Balasubramaniam (2016). In Wang et al. (2017), the authors investigate the controllability of a differential delay semilinear system with linear sections determined by matrices (permutable). We proposed a new real concept of stability results in finite dimensional space in this study by using weaker conditions for nonlinear terms.

### 5. Numerical Simulations

Consider the system of NFSDS described by

$${}^c \bar{D}^{0.6} [x_{11}(t) - (-t + 2)e^{-t}x_{11}(t)] = (0.1)x_{11}(t) - (3 - t) \frac{x_{11}^2(t)}{1 - t} - \int_0^t s x_{11}(s) \sigma_{11} dB_1 + \int_0^t 3s x_{11}(s) \eta_{11} dB_1^H \quad (4)$$

$${}^c \bar{D}^{0.6} [x_{12}(t) - (2 - t)x_{12}(t)e^{-t}] = -(0.1)x_{12}(t) - (3 - t) \frac{x_{12}^3(t)}{1 - t} - \int_0^t s x_{12}(s) \sigma_{12} dB_2 + \int_0^t 5s x_{12}(s) \eta_{12} dB_2^H \quad (5)$$

for  $t \in J_1 = [0, 1]$  and  $0.5 < \bar{\alpha} < 1$ . Let us take

$$\bar{A}_l = \begin{pmatrix} 0.1 & 0 \\ 0 & -0.1 \end{pmatrix}, \quad \bar{f}_l(t, x_1(t), x_1(t - \bar{h}(t))) = \begin{pmatrix} -(3 - t) \frac{x_{11}^2(t)}{1 - t} \\ -(3 - t) \frac{x_{12}^3(t)}{1 - t} \end{pmatrix},$$

$$\bar{\sigma}_l(t, x_1(t), x_1(t - \bar{h}(t))) = \begin{pmatrix} -t x_{11}(t) \sigma_{11} dB_1 \\ -t x_{12}(t) \sigma_{12} dB_2 \end{pmatrix}, \quad \bar{g}_l(t, x_1(t), x_1(t - \bar{h}(t))) = \begin{pmatrix} -(2 - t)x_{11}(t)e^{-t} \\ -(2 - t)x_{12}(t)e^{-t} \end{pmatrix},$$

$$\bar{\eta}_l(t, x_1(t), x_1(t - \bar{h}(t))) = \begin{pmatrix} 3t x_{11}(t) \eta_{11} dB_1^H \\ 5t x_{12}(t) \eta_{12} dB_2^H \end{pmatrix} \text{ where, } \bar{h} = 0.01, \sigma_{11} = 0.3, \sigma_{12} = 0.5 \text{ and } \bar{\alpha} = 0.6.$$

Furthermore, it is easy to verify that for any  $x_1(t), y_1(t) \in \mathbb{R}^2$ .

- (i).  $\mathbb{E} \|\bar{f}_l(t, x_k(t), x_1(t - \bar{h}(t))) - \bar{f}_l(t, y_1(t), y_1(t - \bar{h}(t)))\|^2 \leq -(3 - t) \mathbb{E} \|x_1(t) - y_1(t)\|^2$
- (ii).  $\mathbb{E} \|\bar{\sigma}_l(t, x_1(t), x_1(t - \bar{h}(t))) - \bar{\sigma}_l(t, y_1(t), y_1(t - \bar{h}(t)))\|^2 \leq -0.5t \mathbb{E} \|x_1(t) - y_1(t)\|^2$
- (iii).  $\mathbb{E} \|\bar{g}_l(t, x_1(t), x_1(t - \bar{h}(t))) - \bar{g}_l(t, y_1(t), y_1(t - \bar{h}(t)))\|^2 \leq -(2 - t) \mathbb{E} \|x_1(t) - y_1(t)\|^2$
- (iv).  $\mathbb{E} \|\bar{\eta}_l(t, x_1(t), x_1(t - \bar{h}(t))) - \bar{\eta}_l(t, y_1(t), y_1(t - \bar{h}(t)))\|^2 \leq 4t \mathbb{E} \|x_1(t) - y_1(t)\|^2$ .

Thus,  $\bar{f}_l, \bar{\sigma}_l$  and  $\bar{g}_l$  satisfies the assumption  $(H_1)$ , where we set  $V_{\bar{f}_l}(\cdot), V_{\bar{\sigma}_l}(\cdot), V_{\bar{g}_l}(\cdot) \in L^q(J_1, \mathbb{R}^+)$ .

Hence, all the conditions of Theorem 1 are satisfied. Hence, the fractional systems are stable for  $J_1$ . The Figures 1 and 2 show the related stability results for various values of ' $\bar{\alpha}$ '.

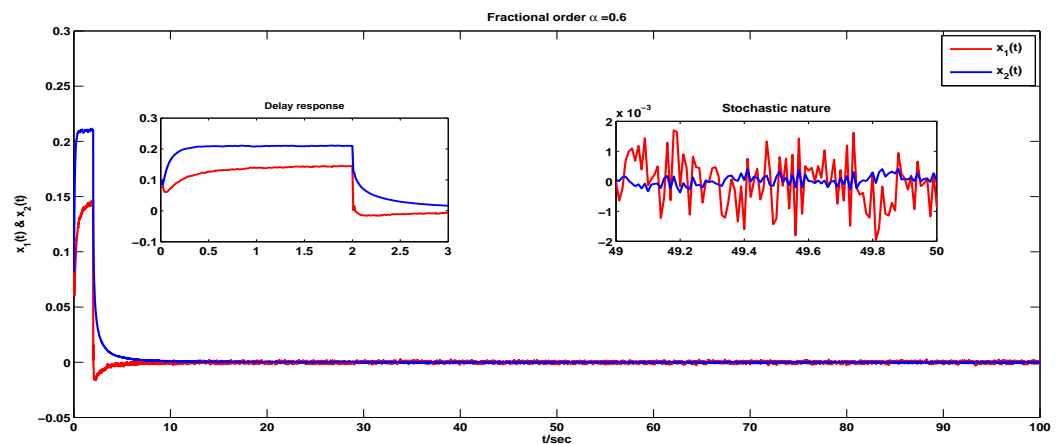
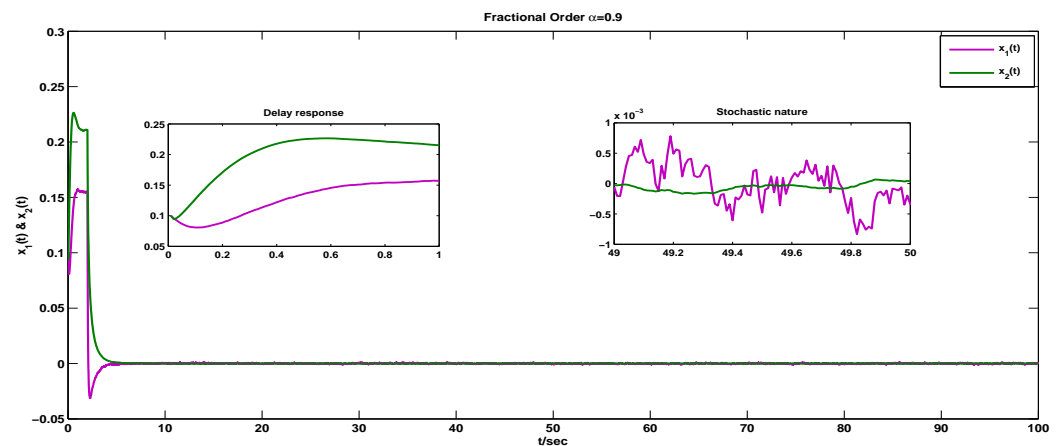


Figure 1. The systems (4)–(5) are stable at  $\bar{\alpha} = 0.6$ .



**Figure 2.** The systems (4)–(5) are stable at  $\tilde{\alpha} = 0.9$ .

Here, the delay response for the systems (4)–(5) is calculated for various values  $\tilde{\alpha} = 0.6, 0.9$  and the delay occurred at  $t = 2$ . Further, the nonlinear functions  $\tilde{f}_1, \tilde{\sigma}_1$  and  $\tilde{g}_1$  are continuous and satisfy the assumption  $(H_1)$ , and then using Theorem 1, the systems (4)–(5), they are stable on  $[0, 100]$ .

## 6. Conclusions and Future Research

In this paper, some useful and general conditions for exponential stability of NFSDS with fBm has been derived. The existence and uniqueness of fixed points, as well as the stability analysis of NFSDS, have been demonstrated. Finally, a numerical simulation was provided to demonstrate the theoretical findings. Based on the application of fractional-order stochastic financial modeling, the authors are interested in establishing the proposed model by considering the exponential stability of fractional stochastic delay systems with finance and stock price models and optimal control of stochastic insurance premium model in the near future.

**Author Contributions:** Conceptualization, O.S.H.; methodology, T.S.; software, T.A.; validation, O.S.H.; formal analysis, T.S.; investigation, T.S. and T.A.; resources, O.S.H. and T.S.; data curation, T.S.; writing—original draft preparation, T.S.; writing—review and editing, O.S.H.; visualization, T.S.; supervision, O.S.H.; project administration, T.S.; funding acquisition, O.S.H. and T.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors wish to thank the reviewers for their insightful comments which have greatly improved the paper.

**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.

## References

- Ahmed, E., A. M. A. El-Sayed, and Hala A. A. El-Saka. 2007. Equilibrium points, stability and numerical solutions of fractional-order predatorprey and rabies models. *Journal of Mathematical Analysis and Applications* 325: 542–53. [CrossRef]
- Applebaum, David. 2009. *Levy Processes and Stochastic Calculus*. Cambridge: Cambridge University Press.
- Balachandran, Krishnan, Jayakumar Kokila, and Juan J. Trujillo. 2012. Relative controllability of fractional dynamical systems with multiple delays in control. *Computers & Mathematics with Applications* 64: 3037–45.



- Balasubramaniam, P., T. Sathiyaraj, and K. Priya. 2020. Exponential stability of nonlinear fractional stochastic system with Poisson jumps. *Stochastics* 93: 945–57. [CrossRef]
- Bhaskar, Bagchi, and Paul Biswajit. 2023. Effects of crude oil price shocks on stock markets and currency exchange rates in the context of Russia-Ukraine conflict: evidence from G7 countries. *Journal of Risk and Financial Management* 16: 64.
- Burton, T. A., and Bo Zhang. 2012. Fractional equations and generalizations of Schaefer and Krasnoselskii's fixed point theorems. *Nonlinear Analysis: Theory, Methods and Applications* 75: 6485–95. [CrossRef]
- Dassios, Ioannis. 2022. On the relations between a singular system of differential equations and a system with delays. *Mathematical Modelling and Numerical Simulation with Applications* 2: 221–27. [CrossRef]
- Farooq, Umar, Mosab I. Tabash, Ahmad A. Al-Naimi, Linda Nalini Daniel, and Mohammad Ahmad Al-Omari. 2023. Herding Trend in Working Capital Management Practices: Evidence from the Non-Financial Sector of Pakistan. *Journal of Risk and Financial Management* 16: 127. [CrossRef]
- Gao, Xin, and Juebang Yu. 2005. Chaos in the fractional order periodically forced complex Duffing oscillators. *Chaos, Solitons and Fractals* 24: 1097–104. [CrossRef]
- Hilfer, Rudolf. 2000. *Applications of Fractional Calculus in Physics*. Singapore: World Scientific.
- Hurst, Harold Edwin. 1951. Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers* 116: 770–99. [CrossRef]
- Kalidass, Mathiyalagan, Shengda Zeng, and Mehmet Yavuz. 2022. Stability of fractional-order quasi-linear impulsive integro-differential systems with multiple delays. *Axioms* 11: 308. [CrossRef]
- Kreyszig, Erwin. 1978. *Introductory Functional Analysis with Applications*. New York: John Wiley and Sons Inc.
- Kumar, Vipin, Gani Stamov, and Ivanka Stamova. 2023. Controllability results for a class of piecewise nonlinear impulsive fractional dynamic systems. *Applied Mathematics and Computation* 439: 127625. [CrossRef]
- Kumar, Vipin, Malik Muslim, and Dumitru Baleanu. 2022a. Results on Hilfer fractional switched dynamical system with non-instantaneous impulses. *Pramana* 96: 172. [CrossRef]
- Kumar, Vipin, Marko Kostić, Abdessamad Tridane, and Amar Debboche. 2022b. Controllability of switched Hilfer neutral fractional dynamic systems with impulses. *IMA Journal of Mathematical Control and Information* 39: 807–36. [CrossRef]
- Li, Qing, Yanli Zhou, Xinquan Zhao, and Xiangyu Ge. 2014. Fractional order stochastic differential equation with application in European option pricing. *Discrete Dynamics in Nature and Society* 2014: 621895. [CrossRef]
- Magin, Richard L. 2010. Fractional calculus models of complex dynamics in biological tissues. *Computers & Mathematics with Applications* 59: 1586–93.
- Mao, Xuerong. 1997. *Stochastic Differential Equations and Applications*. Chichester: Horwood Publishing.
- Nieto, Juan J., and Bessem Samet. 2017. Solvability of an implicit fractional integral equation via a measure of noncompactness argument. *Acta Mathematica Scientia* 37: 195–204. [CrossRef]
- Odiat, Zaid M. 2010. Analytic study on linear systems of fractional differential equations. *Computers & Mathematics with Applications* 59: 1171–83.
- Okawa, Hiroyuki. 2023. Markov-Regime switches in oil markets: The fear factor dynamics. *Journal of Risk and Financial Management* 16: 67. [CrossRef]
- Oldham, Keith B. 2010. Fractional differential equations in electrochemistry. *Advances in Engineering Software* 41: 1171–83. [CrossRef]
- Ortigueira, Manuel Duarte. 2011. *Fractional Calculus for Scientists and Engineers*. New York: Springer Science & Business.
- Podlubny, Igor. 1998. *Fractional Differential Equations: An Introduction to Fractional Derivatives, Fractional Differential Equations, to Methods of their Solution and Some of their Applications*. Millbrae: Academic Press.
- Ren, Yong, and R. Sakthivel. 2012. Existence, uniqueness, and stability of mild solutions for second-order neutral stochastic evolution equations with infinite delay and Poisson jumps. *Journal of Mathematical Physics* 53: 073517 [CrossRef]
- Ren, Yong, Xuejuan Jia, and Lanying Hu. 2017. The  $p$ -th moment stability of solutions to impulsive stochastic differential equations driven by G-Brownian motion. *Applicable Analysis* 96: 988–1003. [CrossRef]
- Rockner, Michael, and Tusheng Zhang. 2007. Stochastic evolution equations of jump type: Existence, uniqueness and large deviation principle. *Potential Analysis* 26: 255–79. [CrossRef]
- Sari, Suci, Arief Hakim, Ikha Magdalena, and Khreshna Syuhada. 2023. Modeling the optimal combination of proportional and Stop-Loss reinsurance with dependent claim and stochastic insurance premium. *Journal of Risk and Financial Management* 16: 95. [CrossRef]
- Sathiyaraj, T., and P. Balasubramaniam. 2016. Fractional order stochastic dynamical systems with distributed delayed control and Poisson jumps. *The European Physical Journal Special Topics* 225: 83–96. [CrossRef]
- Sathiyaraj, T., and P. Balasubramaniam. 2018. Controllability of fractional higher order stochastic integrodifferential systems with fractional Brownian motion. *ISA transactions* 82: 107–19. [CrossRef]
- Sathiyaraj, T., Jinrong Wang, and P. Balasubramaniam. 2019. Ullam stability of Hilfer fractional stochastic differential systems. *The European Physical Journal Plus* 134: 605. [CrossRef]
- Seemab, Arjumand, and M. Rehman. 2018. Existence and stability analysis by fixed point theorems for a class of nonlinear Caputo fractional differential equations. *Dynamics Systems and Applications* 27: 445–56.
- Seo, Dong-Won, and Hochang Lee. 2011. Stationary waiting times in m-node tandem queues with production blocking. *IEEE Transactions on Automatic Control* 56: 958–61. [CrossRef]

- Shen, Guangjun, R. Sakthivel, Yong Ren, and Mengyu Li. 2020. Controllability and stability of fractional stochastic functional systems driven by Rosenblatt process. *Collectanea Mathematica* 71: 63–82. [CrossRef]
- Singh, Jagdev, Devendra Kumar, and Juan J. Nieto. 2017. Analysis of an el nino-southern oscillation model with a new fractional derivative. *Chaos, Solitons and Fractals* 99: 109–15. [CrossRef]
- Song, Dong-Ping. 2009. Optimal integrated ordering and production policy in a supply chain with stochastic lead-time, processing time, and demand. *IEEE Transactions on Automatic Control* 54: 2027–41. [CrossRef]
- Taheri, Mohammad M., Keivan Navaie, and Mohammad H. Bastani. 2010. On the outage probability of SIR-based power-controlled DS-CDMA networks with spatial Poisson traffic. *IEEE Transactions on Vehicular Technology* 59: 499–506. [CrossRef]
- Tian, Yu, and Juan J. Nieto. 2017. The applications of critical-point theory discontinuous fractional-order differential equations. *Proceedings of the Edinburgh Mathematical Society* 60: 1021–51. [CrossRef]
- Wang, Jinrong, Yong Zhou, and Feckan Fec. 2012. Nonlinear impulsive problems for fractional differential equations and Ulam stability. *Computers & Mathematics with Applications* 64: 3389–405.
- Wang, Jinrong, Zijian Luo, and Michal Feckan. 2017. Relative controllability of semilinear delay differential systems with linear parts defined by permutable matrices. *European Journal of Control* 30: 39–46. [CrossRef]
- Wu, Dongsheng. 2011. On the solution process for a stochastic fractional partial differential equation driven by space-time white noise. *Statistics and Probability Letters* 81: 1161–72. [CrossRef]
- Yun, Kyung Hwan, and Chenguang Hu. 2023. Growth of venture firms under state capitalism with chinese characteristics: Qualitative comparative analysis of fuzzy set. *Journal of Risk and Financial Management* 16: 138. [CrossRef]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

# The Naive Estimator of a Poisson Regression Model with a Measurement Error

Kentarou Wada \* and Takeshi Kurosawa

Department of Applied Mathematics, Tokyo University of Science, Kagurazaka 1-3, Shinjuku-ku, Tokyo 1628601, Japan

\* Correspondence: wadaken5269@gmail.com

**Abstract:** We generalize the naive estimator of a Poisson regression model with a measurement error as discussed in Kukush et al. in 2004. The explanatory variable is not always normally distributed as they assume. In this study, we assume that the explanatory variable and measurement error are not limited to a normal distribution. We clarify the requirements for the existence of the naive estimator and derive its asymptotic bias and asymptotic mean squared error (MSE). The requirements for the existence of the naive estimator can be expressed using an implicit function, which the requirements can be deduced by the characteristic of the Poisson regression models. In addition, using the implicit function obtained from the system of equations of the Poisson regression models, we propose a consistent estimator of the true parameter by correcting the bias of the naive estimator. As illustrative examples, we present simulation studies that compare the performance of the naive estimator and new estimator for a Gamma explanatory variable with a normal error or a Gamma error.

**Keywords:** Poisson regression model; error in variable; naive estimator; asymptotic bias



**Citation:** Wada, Kentarou, and Takeshi Kurosawa. 2023. The Naive Estimator of a Poisson Regression Model with a Measurement Error. *Journal of Risk and Financial Management* 16: 186. <https://doi.org/10.3390/jrfm16030186>

Academic Editors: Shuangzhe Liu, Tiefeng Ma, Seng Huat Ong and Thanasis Stengos

Received: 18 January 2023

Revised: 21 February 2023

Accepted: 6 March 2023

Published: 9 March 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

We often cannot measure explanatory variables correctly in regression models because an observation may not be performed properly. The estimation result may be distorted when we estimate the model from data with measurement errors. We call models with measurement errors in an explanatory variable Error in Variable (EIV) models. In addition, actual phenomena often cannot be explained adequately by a simple linear structure, and the estimation of non-linear models, especially generalized linear models, from data with errors is a significant problem. Various studies have focused on non-linear EIV models (see, for example, Box 1963; Geary 1953). Classical error models assume that an explanatory variable is measured with independent stochastic errors (Kukush and Schneeweiss 2000). Berkson error models assume that the explanatory variable is a controlled variable with an error and that only the controlled variable can be measured (Burr 1988; Huwang and Huang 2000). Approaches to EIV models vary according to the situation. In this paper, we consider the former EIV. The corrected score function in Nakamura (1990) has been used to estimate generalized linear models. In particular, the Poisson regression model is easy to handle analytically in generalized linear models as we see later. Thus, we focus on the Poisson regression model with measurement errors.

Approaches to a Poisson regression model with classical errors have been discussed by Kukush et al. (2004), Shklyar and Schneeweiss (2005), Jiang and Ma (2020), Guo and Li (2002), and so on. Kukush et al. (2004) described the statistical properties of the naive estimator, corrected score estimator, and structural quasi score estimator of a Poisson regression model with normally distributed explanatory variable and measurement errors. Shklyar and Schneeweiss (2005) assumed an explanatory variable and a measurement error with a multivariate normal distribution and compared the asymptotic covariance matrices of the corrected score estimator, simple structural estimator, and structural quasi score estimator of a Poisson regression model. Jiang and Ma (2020) assumed a high-dimensional

explanatory variable with a multivariate normal error and proposed a new estimator for a Poisson regression model by combining Lasso regression and the corrected score function. Guo and Li (2002) assumed a Poisson regression model with classical errors and proposed an estimator that is a generalization of the corrected score function discussed in Nakamura (1990) for generally distributed errors; they derived the asymptotic normality of the proposed estimator.

In this study, we generalize the naive estimator discussed in Kukush et al. (2004). They reported the bias of the naive estimator, however, the explanatory variable is not always normally distributed as they assume. In practice, the assumption of a normal distribution is not realistic. Here, we assume that the explanatory variable and measurement error are not limited to normal distributions. However, the naive estimator does not always exist in every situation. Therefore, we clarify the requirements for the existence of the naive estimator and derive its asymptotic bias. The constant vector to which the naive estimator converges in probability does not coincide with the unknown parameter in the model. Therefore, we propose a consistent estimator of the unknown parameter using the naive estimator. It is obtained from a system of equations that represent the relationship between the unknown parameter and constant vector. As illustrative examples, we present explicit representations of the new estimator for a Gamma explanatory variable with a normal error or a Gamma error.

In Section 2, we present the Poisson regression model with measurement errors and the definition of the naive estimator and show that the naive estimator has an asymptotic bias for the true parameter. In Section 3, we consider the requirements for the existence of the naive estimator and derive its asymptotic bias and asymptotic mean squared error (MSE) assuming that the explanatory variable and measurement error are generally distributed. In addition, we introduce application examples of a Gamma explanatory variable with a normal error or a Gamma error. In Section 4, we propose the corrected naive estimator as a consistent estimator of the true parameter under general distributions and give application examples for a Gamma explanatory variable with a normal error or a Gamma error. In Section 5, we present simulation studies that compare the performance of the naive estimator and corrected naive estimator. In Section 6, we apply the naive and corrected naive estimators to real data in two cases. Finally, discussions are presented in Section 7.

## 2. Preliminary

In this section, we state the statistical model considered in this paper and the definition of the naive estimator and show that the naive estimator has an asymptotic bias for the true parameter.

### 2.1. Poisson Regression Models with an Error

We assume a single covariate Poisson regression model between the objective variable  $Y$  and explanatory variable  $X$

$$Y|X \sim Po(\exp(\beta_0 + \beta_1 X)).$$

$X$  can typically be correctly observed. We assume here that  $X$  has a stochastic error  $U$  as

$$W = X + U,$$

where  $U$  is supposed to be independent of  $(X, Y|X)$ . We also assume that

$$(Y_i, X_i, U_i) \quad (i = 1, \dots, n) \tag{1}$$

are independent and identically distributed samples of the distributions of  $(Y|X, X, U)$ . Although we can observe  $Y|X$  and  $W$ , we assume that  $X$  and  $U$  cannot be directly observed. However, even if we know the family of the distributions of  $X$  and  $U$ , we can-not make a

statistical inference regarding  $X$  and  $U$  if we can observe only  $W$ . Because  $U$  is the error distribution, the mean of  $U$  is often zero, and we may suppose that we have empirical information about the degree of error (the variance of  $U$ ). Therefore, in this study, we assume that the mean and variance of  $U$  are known. From the above assumption,  $Y$  and  $W$  are independent for the given  $X$ .

$$\begin{aligned} f_{Y,W|X}(y, w|x) &= \frac{f_{Y,W,X}(y, w, x)}{f_X(x)} = \frac{f_{Y,W,U}(y, w, w-x)}{f_X(x)} \\ &= \frac{f_{Y,X}(y, x)f_U(w-x)}{f_X(x)} = f_{Y|X}(y|x)f_{W|X}(w|x). \end{aligned}$$

We use this conditional independence when we calculate the expectations.

### 2.2. The Naive Estimator

The naive estimator  $\hat{\boldsymbol{\beta}}^{(N)} = (\hat{\beta}_0^{(N)}, \hat{\beta}_1^{(N)})'$  for  $\boldsymbol{\beta} = (\beta_0, \beta_1)'$  is defined as the solution of the equation

$$S_n(\hat{\boldsymbol{\beta}}^{(N)}|\mathcal{X}) = \mathbf{0}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{2}$$

where

$$S_n(\mathbf{b}|\mathcal{X}) = \frac{1}{n} \sum_{i=1}^n \{Y_i - \exp(b_0 + b_1 W_i)\} (1, W_i)'$$

is a function of indeterminate  $\mathbf{b} = (b_0, b_1)'$  given  $\mathcal{X} = (X_1, \dots, X_n)'$ . The naive estimator can be interpreted as the maximum likelihood estimator if we wrongly assume that  $Y|W \sim Po(\exp(\beta_0 + \beta_1 W))$  because (2) is the log-likelihood equation for  $Y|W \sim Po(\exp(\beta_0 + \beta_1 W))$ . The correct distribution of  $Y|W$  is

$$\begin{aligned} f_{Y|W}(y|w) &= \frac{1}{f_W(w)} \int_{supp(f_U)} f_{Y|W,U}(y|w, u) f_U(u) f_X(w-u) du \\ &= \frac{1}{f_W(w)} \int_{supp(f_U)} f_{Y|X}(y|w-u) f_U(u) f_X(w-u) du \\ &= \frac{1}{f_W(w)} \int_{supp(f_U)} Po(\exp(\beta_0 + \beta_1(w-u))) f_U(u) f_X(w-u) du \end{aligned}$$

assuming that  $U$  is independent of  $(X, Y|X)$ . The right-hand side must be different from  $Po(\exp(\beta_0 + \beta_1 W))$  in general. If one ignores the error  $U$  and fits the likelihood estimation using  $W$  instead of  $X$ , a biased estimator is obtained. In fact, by the law of large numbers, we have

$$\begin{aligned} S_n(\hat{\boldsymbol{\beta}}^{(N)}|\mathcal{X}) &= \frac{1}{n} \sum_{i=1}^n \{Y_i - \exp(\hat{\beta}_0^{(N)} + \hat{\beta}_1^{(N)} W_i)\} (1, W_i)' \\ &\xrightarrow{p} \mathbf{E}_{X,W}[\mathbf{E}_{Y|(X,W)}[\{Y - \exp(\hat{\beta}_0^{(N)} + \hat{\beta}_1^{(N)} W)\} (1, W)']]. \end{aligned}$$

Thus, the naive estimator converges to  $\mathbf{b} = (b_0, b_1)'$  which is the solution of the estimating equation

$$\mathbf{E}_{X,W}[\mathbf{E}_{Y|(X,W)}[\{Y - \exp(\hat{\beta}_0^{(N)} + \hat{\beta}_1^{(N)} W)\} (1, W)']] = \mathbf{0}_2. \tag{3}$$

Equation (3) implies that for a given  $\mathcal{X}$

$$\hat{\boldsymbol{\beta}}^{(N)} \xrightarrow{p} \mathbf{b} \neq \boldsymbol{\beta}.$$

The solution  $\mathbf{b}$  of the estimating equation is generally different from  $\boldsymbol{\beta}$ .

### 3. Properties of the Naive Estimator

In this section, we consider the requirements for the existence of the naive estimator and derive its asymptotic bias and asymptotic MSE assuming that the explanatory variable and measurement error are generally distributed. In addition, we introduce application examples for a Gamma explanatory variable with a normal error or a Gamma error.

#### 3.1. The Existence of the Naive Estimator

The naive estimator does not always exist for general random variables  $X$  and  $U$ . Thus, we assume the existence of the expectation

$$E_{X,Y,W}[\{Y - \exp(b_0 + b_1 W)\}(1, W)']$$

as a requirement for the existence of the naive estimator. Consequently, the following four expectations should exist.

$$\begin{cases} E[Y] &= E_X[E[Y|X]] = E_X[\exp(\beta_0 + \beta_1 X)] = e^{\beta_0} M_X(\beta_1), \\ E[\exp(b_0 + b_1 W)] &= e^{b_0} E[e^{b_1 X + b_1 U}] = e^{b_0} M_X(b_1) M_U(b_1), \\ E[YW] &= E_X[E[Y|X]E[W|X]] = E_X[(X + E[U]) \exp(\beta_0 + \beta_1 X)] \\ &= e^{\beta_0} E[U] M_X(\beta_1) + e^{\beta_0} E[X e^{\beta_1 X}] \\ &= e^{\beta_0} E[U] M_X(\beta_1) + e^{\beta_0} \nabla M_X(\beta_1), \\ E[W \exp(b_0 + b_1 W)] &= E_X[E_U[(X + U) \exp(b_0 + b_1 X + b_1 U)]] \\ &= e^{b_0} E[X e^{b_1 X}] M_U(b_1) + e^{b_0} E[U e^{b_1 U}] M_X(b_1) \\ &= e^{b_0} M_U(b_1) \nabla M_X(b_1) + e^{b_0} M_X(b_1) \nabla M_U(b_1). \end{cases} \quad (4)$$

Therefore, these expectations require that  $M_X(\beta_1), M_X(b_1), M_U(b_1)$  exist. This condition is the requirement for the existence of the naive estimator. Here, we assume the existence of

$$M_X(\beta_1), M_X(b_1), M_U(b_1) \quad (5)$$

for the distributions of  $X$  and  $U$ .

#### 3.2. Asymptotic Bias of the Naive Estimator

The naive estimator satisfies

$$\hat{\beta}^{(N)} \xrightarrow{p} \mathbf{b}$$

and has an asymptotic bias for the true  $\beta$ . Here, we derive the asymptotic bias under general conditions. From (3), we obtain two equations:

$$\begin{cases} E[Y] &= E[\exp(b_0 + b_1 W)], \\ E[YW] &= E[W \exp(b_0 + b_1 W)]. \end{cases} \quad (6)$$

From (4) with the above equalities, we have

$$\begin{aligned} e^{\beta_0} M_X(\beta_1) &= e^{b_0} M_X(b_1) M_U(b_1), \\ e^{\beta_0} E[U] M_X(\beta_1) + e^{\beta_0} \nabla M_X(\beta_1) &= e^{b_0} (\nabla M_X(b_1)) M_U(b_1) + e^{b_0} (\nabla M_U(b_1)) M_X(b_1) \\ &= e^{b_0} \nabla (M_X(b_1) M_U(b_1)) = e^{b_0} \nabla M_W(b_1). \end{aligned}$$

Therefore, we use a transformation to obtain the following system of equations:

$$\begin{cases} b_0 &= \beta_0 + \log\left(\frac{M_X(\beta_1)}{M_W(b_1)}\right), \\ K'_W(b_1) &= \frac{1}{M_W(b_1)} \nabla M_W(b_1) = E[U] + \frac{\nabla M_X(\beta_1)}{M_X(\beta_1)}, \end{cases} \quad (7)$$

where  $K_W$  is the cumulant generating function of  $W$ . Thus,  $\mathbf{b} = (b_0, b_1)'$  is determined by the solution of this system of equations. Therefore, the equation

$$K'_W(b_1) = \mathbf{E}[U] + \frac{\nabla M_X(\beta_1)}{M_X(\beta_1)}$$

should have a solution with respect to  $b_1$ . Here, we set

$$G(\beta_1, b_1) := K'_W(b_1) - \mathbf{E}[U] - K'_X(\beta_1).$$

We assume  $G(\beta_1, b_1)$  has zero in  $\mathbb{R}^2$  and satisfies

$$\frac{\partial G(\beta_1, b_1)}{\partial b_1} = K''_W(b_1) \neq 0.$$

$G$  is continuously differentiable because we assume the existence of (5). Then, by the theorem of implicit functions, there exists a unique  $C^1$ -class function  $g$  that satisfies  $b_1 = g(\beta_1)$  in the neighborhood of the zero of  $G$ . Using this expression, we write the asymptotic bias of the naive estimator as

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbf{E}[\hat{\beta}_0^{(N)} - \beta_0] &= b_0 - \beta_0 = \log\left(\frac{M_X(\beta_1)}{M_W \circ g(\beta_1)}\right), \\ \lim_{n \rightarrow \infty} \mathbf{E}[\hat{\beta}_1^{(N)} - \beta_1] &= b_1 - \beta_1 = g(\beta_1) - \beta_1. \end{aligned}$$

We also derive the asymptotic MSE of the naive estimator. The MSE can be represented as the sum of the squared bias and variance. The asymptotic variance of the naive estimator is 0 because the naive estimator is a consistent estimator of  $\mathbf{b}$ . Thus, we obtain the asymptotic MSE of the naive estimator as

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbf{E}[(\hat{\beta}_0^{(N)} - \beta_0)^2] &= (b_0 - \beta_0)^2 = \left(\log\left(\frac{M_X(\beta_1)}{M_W \circ g(\beta_1)}\right)\right)^2, \\ \lim_{n \rightarrow \infty} \mathbf{E}[(\hat{\beta}_1^{(N)} - \beta_1)^2] &= (b_1 - \beta_1)^2 = (g(\beta_1) - \beta_1)^2. \end{aligned}$$

Therefore, the asymptotic bias is given by the following theorem assuming general distributions.

**Theorem 1.** Let  $Y|X \sim Po(\exp(\beta_0 + \beta_1 X))$ . Assume that  $W = X + U$  and  $U$  is independent of  $(X, Y|X)$ . Assume the existence of  $M_X(\beta_1), M_X(b_1), M_U(b_1)$ . Let

$$G(\beta_1, b_1) := K'_W(b_1) - \mathbf{E}[U] - K'_X(\beta_1).$$

Assume the function  $G$  has a zero in  $\mathbb{R}^2$ , namely there exist solutions with  $G(\beta_1, b_1) = 0$ , and satisfies

$$\frac{\partial G(\beta_1, b_1)}{\partial b_1} = K''_W(b_1) \neq 0.$$

Then, the asymptotic biases of the naive estimators  $\hat{\beta}_0^{(N)}$  and  $\hat{\beta}_1^{(N)}$  are given by

$$\log\left(\frac{M_X(\beta_1)}{M_W \circ g(\beta_1)}\right) \quad \text{and} \quad g(\beta_1) - \beta_1$$

respectively, where  $g$  is a  $C^1$ -class function satisfying  $b_1 = g(\beta_1)$  in the neighborhood of the zero of  $G$ . Furthermore, the asymptotic MSEs of the naive estimators  $\hat{\beta}_0^{(N)}$  and  $\hat{\beta}_1^{(N)}$  are given by their squared asymptotic biases.

### 3.3. Examples

In this section, we present two type of examples. First, we assume that a Gamma explanatory variable with a normal error. Let

$$X \sim \Gamma(k, \lambda), \quad U \sim N(0, \sigma^2),$$

where  $k > 0, \lambda > 0, 0 < \sigma^2 < \infty$ . We apply the naive estimation under this condition. From the assumptions of Theorem 1, we assume the existence of

$$M_X(\beta_1), M_X(b_1) \text{ and } M_U(b_1).$$

Therefore, we obtain the parameter conditions

$$\lambda - \beta_1 > 0, \quad \lambda - b_1 > 0.$$

Next, we derive  $\mathbf{b} = (b_0, b_1)'$ . Under this condition, we obtain

$$G(\beta_1, b_1) = K'_W(b_1) - \mathbf{E}[U] - K'_X(\beta_1) = \frac{k}{\lambda - b_1} + \sigma^2 b_1 - \frac{k}{\lambda - \beta_1}.$$

Thus, the set of zeros of  $G$  is

$$\left\{ (\beta_1, b_1) \in \mathbb{R}^2; \beta_1 = \frac{k + \lambda\sigma^2(\lambda - b_1)}{k + \sigma^2(\lambda - b_1)b_1} b_1 \right\}.$$

In addition,

$$\frac{\partial G(\beta_1, b_1)}{\partial b_1} = \frac{k}{(\lambda - b_1)^2} + \sigma^2 > 0.$$

Therefore,  $G$  has a zero in  $\mathbb{R}^2$  and satisfies  $\frac{\partial G(\beta_1, b_1)}{\partial b_1} \neq 0$ . From  $G(\beta_1, b_1) = 0$ , we obtain two implicit functions

$$b_1^{(1)} = \frac{(\lambda - \beta_1)\lambda\sigma^2 + k + \sqrt{s}}{2(\lambda - \beta_1)\sigma^2},$$

$$b_1^{(2)} = \frac{(\lambda - \beta_1)\lambda\sigma^2 + k - \sqrt{s}}{2(\lambda - \beta_1)\sigma^2},$$

where  $s = (\lambda - \beta_1)^2\lambda^2\sigma^4 + 2(\lambda - \beta_1)(\lambda - 2\beta_1)\sigma^2k + k^2 > 0$ . Then, we obtain two expressions of  $b_0$  corresponding to  $b_1$ .

$$b_0^{(1)} := \beta_0 + \log\left(\frac{M_X(\beta_1)}{M_W(b_1^{(1)})}\right)$$

$$= \beta_0 + k \log \frac{(\lambda - \beta_1)\lambda\sigma^2 - k - \sqrt{s}}{2(\lambda - \beta_1)^2\sigma^2}$$

$$- \frac{(\lambda - \beta_1)^2\lambda^2\sigma^4 + 2(\lambda - \beta_1)^2\sigma^2k + k^2 + ((\lambda - \beta_1)\lambda\sigma^2 + k)\sqrt{s}}{4(\lambda - \beta_1)^2\sigma^2},$$

$$b_0^{(2)} := \beta_0 + \log\left(\frac{M_X(\beta_1)}{M_W(b_1^{(2)})}\right)$$

$$= \beta_0 + k \log \frac{(\lambda - \beta_1)\lambda\sigma^2 - k + \sqrt{s}}{2(\lambda - \beta_1)^2\sigma^2}$$

$$- \frac{(\lambda - \beta_1)^2\lambda^2\sigma^4 + 2(\lambda - \beta_1)^2\sigma^2k + k^2 - ((\lambda - \beta_1)\lambda\sigma^2 + k)\sqrt{s}}{4(\lambda - \beta_1)^2\sigma^2}.$$



In addition,

$$s = ((\lambda - \beta_1)\lambda\sigma^2 - k)^2 + 4(\lambda - \beta_1)^2\sigma^2k;$$

therefore,  $s$  satisfies  $\sqrt{s} > |(\lambda - \beta_1)\lambda\sigma^2 - k|$ . From the antilogarithm condition,  $\mathbf{b} = (b_0^{(2)}, b_1^{(2)})'$  is a solution of the system of Equation (6) in the range of  $\mathbb{R}^2$ . Thus, the asymptotic biases are given by

$$b_0 - \beta_0 = k \log \frac{(\lambda - \beta_1)\lambda\sigma^2 - k + \sqrt{s}}{2(\lambda - \beta_1)^2\sigma^2} - \frac{(\lambda - \beta_1)^2\lambda^2\sigma^4 + 2(\lambda - \beta_1)^2\sigma^2k + k^2 - ((\lambda - \beta_1)\lambda\sigma^2 + k)\sqrt{s}}{4(\lambda - \beta_1)^2\sigma^2},$$

$$b_1 - \beta_1 = \frac{\lambda}{2} - \beta_1 + \frac{k - \sqrt{s}}{2(\lambda - \beta_1)\sigma^2}.$$

Next, we present another example, Gamma explanatory variable with a Gamma error. Let

$$X \sim \Gamma(k_1, \lambda), \quad U \sim \Gamma(k_2, \lambda),$$

where  $k_1 > 0, k_2 > 0, \lambda > 0$ . We apply the naive estimation under this condition. From the assumptions of Theorem 1, we assume the existence of

$$M_X(\beta_1), M_X(b_1) \text{ and } M_U(b_1).$$

Therefore, we obtain the parameter conditions

$$\lambda - \beta_1 > 0, \quad \lambda - b_1 > 0.$$

Next, we derive  $\mathbf{b} = (b_0, b_1)'$ . Under this condition, we obtain

$$G(\beta_1, b_1) = \frac{k_1 + k_2}{\lambda - b_1} - \frac{k_1}{\lambda - \beta_1} - \frac{k_2}{\lambda}.$$

Thus, the set of zeros of  $G$  is

$$\left\{ (\beta_1, b_1) \in \mathbb{R}^2; b_1 = \frac{k_1\lambda\beta_1}{k_1\lambda + k_2(\lambda - \beta_1)} \right\}.$$

In addition,

$$\frac{\partial G(\beta_1, b_1)}{\partial b_1} = \frac{k_1 + k_2}{(\lambda - b_1)^2} > 0.$$

Therefore,  $G$  has a zero in  $\mathbb{R}^2$  and satisfies  $\frac{\partial G(\beta_1, b_1)}{\partial b_1} \neq 0$ . From  $G(\beta_1, b_1) = 0$ , we obtain the implicit function

$$b_1 = \frac{k_1\lambda\beta_1}{k_1\lambda + k_2(\lambda - \beta_1)}.$$

Thus, by Theorem 1, the asymptotic biases are given by

$$b_0 - \beta_0 = -k_1 \log(1 - \beta_1/\lambda) + (k_1 + k_2) \log(1 - b_1/\lambda),$$

$$b_1 - \beta_1 = -\frac{k_2(\lambda - \beta_1)\beta_1}{k_1\lambda + k_2(\lambda - \beta_1)}.$$

#### 4. Corrected Naive Estimator

In this section, we propose a corrected naive estimator as a consistent estimator of  $\beta$  under general distributions and give application examples for a Gamma explanatory variable with a normal error or a Gamma error. From (7), we have the following system of equations:

$$\beta_0 = b_0 + \log\left(\frac{M_W(b_1)}{M_X(\beta_1)}\right),$$

$$G(\beta_1, b_1) = K'_W(b_1) - E[U] - K'_X(\beta_1) = 0.$$

By solving this system of equations for  $\beta_0, \beta_1$  and replacing  $\mathbf{b} = (b_0, b_1)'$  with the naive estimator  $\hat{\beta}^{(N)} = (\hat{\beta}_0^{(N)}, \hat{\beta}_1^{(N)})'$ , we obtain the consistent estimator of the true  $\beta$ . Here,

$$\hat{\beta}^{(N)} = \begin{pmatrix} \hat{\beta}_0^{(N)} \\ \hat{\beta}_1^{(N)} \end{pmatrix} \xrightarrow{p} \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}.$$

Therefore,

$$\hat{\beta}^{(CN)} \xrightarrow{p} \beta.$$

Thus,  $\hat{\beta}^{(CN)}$  is a consistent estimator of  $\beta$ . If  $G$  has zero in  $\mathbb{R}^2$  and satisfies

$$\frac{\partial G(\beta_1, b_1)}{\partial \beta_1} = -K''_X(\beta_1) \neq 0,$$

then, by the theorem of implicit functions, there exists a unique  $C^1$ -class function  $h$  that satisfies  $\beta_1 = h(b_1)$  in the neighborhood of the zero of  $G$ . We note that  $h$  is the inverse function of  $g$  in Theorem 1. We propose a corrected naive estimator that is the consistent estimator of the true  $\beta$  as follows.

**Theorem 2.** Let  $Y|X \sim Po(\exp(\beta_0 + \beta_1 X))$ . Assume that  $W = X + U$  and  $U$  is independent of  $(X, Y|X)$ . Assume the existence of  $M_X(\beta_1), M_X(b_1), M_U(b_1)$ . Let

$$G(\beta_1, b_1) := K'_W(b_1) - E[U] - K'_X(\beta_1).$$

Assume  $G$  has zero in  $\mathbb{R}^2$  and satisfies

$$\frac{\partial G(\beta_1, b_1)}{\partial \beta_1} = -K''_X(\beta_1) \neq 0.$$

Then, the corrected naive estimator  $\hat{\beta}^{(CN)} = (\hat{\beta}_0^{(CN)}, \hat{\beta}_1^{(CN)})'$ , which corrects the bias of the naive estimator  $\hat{\beta}^{(N)} = (\hat{\beta}_0^{(N)}, \hat{\beta}_1^{(N)})'$ , is given by

$$\hat{\beta}_0^{(CN)} = \hat{\beta}_0^{(N)} + \log\left(\frac{M_W(\hat{\beta}_1^{(N)})}{M_X(\hat{\beta}_1^{(CN)})}\right),$$

$$\hat{\beta}_1^{(CN)} = h(\hat{\beta}_1^{(N)}),$$

where  $h$  is a  $C^1$ -class function satisfying  $\beta_1 = h(b_1)$  in the neighborhood of the zero of  $G$ . Furthermore, the corrected naive estimator is a consistent estimator of  $\beta$ .

**Example 1.** We derive the corrected naive estimator assuming

$$X \sim \Gamma(k, \lambda), U \sim N(0, \sigma^2).$$

We obtain

$$G(\beta_1, b_1) = \frac{k}{\lambda - b_1} + \sigma^2 b_1 - \frac{k}{\lambda - \beta_1},$$

$$\frac{\partial G(\beta_1, b_1)}{\partial \beta_1} = -\frac{k}{(\lambda - \beta_1)^2} < 0.$$

$G$  has zero in  $\mathbb{R}^2$  and satisfies  $\frac{\partial G(\beta_1, b_1)}{\partial \beta_1} \neq 0$ . From  $G(\beta_1, b_1) = 0$ , we obtain the implicit function

$$\beta_1 = \frac{\sigma^2 \lambda b_1^2 - (k + \lambda^2 \sigma^2) b_1}{\sigma^2 b_1^2 - \lambda \sigma^2 b_1 - k} = h(b_1).$$

Thus, by Theorem 2, the corrected naive estimator is given by

$$\begin{aligned} \hat{\beta}_0^{(CN)} &= \hat{\beta}_0^{(N)} + \log\left(\frac{M_W(\hat{\beta}_1^{(N)})}{M_X(\hat{\beta}_1^{(CN)})}\right) \\ &= \hat{\beta}_0^{(N)} + \frac{1}{2} \hat{\beta}_1^{(N)2} \sigma^2 + k \log(1 - \hat{\beta}_1^{(CN)} / \lambda) - k \log(1 - \hat{\beta}_1^{(N)} / \lambda), \\ \hat{\beta}_1^{(CN)} &= h(\hat{\beta}_1^{(N)}) = \frac{\lambda \sigma^2 \hat{\beta}_1^{(N)2} - (k + \lambda^2 \sigma^2) \hat{\beta}_1^{(N)}}{\sigma^2 \hat{\beta}_1^{(N)2} - \lambda \sigma^2 \hat{\beta}_1^{(N)} - k}. \end{aligned}$$

**Example 2.** We derive the corrected naive estimator assuming

$$X \sim \Gamma(k_1, \lambda), U \sim \Gamma(k_2, \lambda).$$

We obtain

$$\begin{aligned} G(\beta_1, b_1) &= \frac{k_1 + k_2}{\lambda - b_1} - \frac{k_1}{\lambda - \beta_1} - \frac{k_2}{\lambda}, \\ \frac{\partial G(\beta_1, b_1)}{\partial \beta_1} &= -\frac{k_1}{(\lambda - \beta_1)^2} < 0. \end{aligned}$$

$G$  has zero in  $\mathbb{R}^2$  and satisfies  $\frac{\partial G(\beta_1, b_1)}{\partial \beta_1} \neq 0$ . From  $G(\beta_1, b_1) = 0$ , we obtain the implicit function

$$\beta_1 = \frac{(k_1 + k_2) b_1 \lambda}{k_1 \lambda + k_2 b_1} = h(b_1).$$

Thus, by Theorem 2, the corrected naive estimator is given by

$$\begin{aligned} \hat{\beta}_0^{(CN)} &= \hat{\beta}_0^{(N)} + \log\left(\frac{M_W(\hat{\beta}_1^{(N)})}{M_X(\hat{\beta}_1^{(CN)})}\right) \\ &= \hat{\beta}_0^{(N)} + k_1 \log(1 - \hat{\beta}_1^{(CN)} / \lambda) - (k_1 + k_2) \log(1 - \hat{\beta}_1^{(N)} / \lambda), \\ \hat{\beta}_1^{(CN)} &= h(\hat{\beta}_1^{(N)}) = \frac{(k_1 + k_2) \hat{\beta}_1^{(N)} \lambda}{k_1 \lambda + k_2 \hat{\beta}_1^{(N)}}. \end{aligned}$$

### 5. Simulation Studies

In this section, we present simulation studies that compare the performance of the naive estimator and corrected naive estimator. We denote the sample size by  $n$  and the number of simulations by MC. We calculate the estimated bias for  $\hat{\beta}^{(N)}$  and  $\hat{\beta}^{(CN)}$  as follows:

$$\begin{aligned} \widehat{\text{BIAS}}(\hat{\beta}^{(N)}) &= \frac{1}{MC} \sum_{i=1}^{MC} \hat{\beta}_i^{(N)} - \beta, \\ \widehat{\text{BIAS}}(\hat{\beta}^{(CN)}) &= \frac{1}{MC} \sum_{i=1}^{MC} \hat{\beta}_i^{(CN)} - \beta, \end{aligned}$$

where  $\hat{\beta}_i^{(N)}$  and  $\hat{\beta}_i^{(CN)}$  represent the naive estimator and corrected naive estimator in the  $i$ th time simulation, respectively. Similarly, we calculate the estimated MSE matrix for  $\hat{\beta}^{(N)}$  and  $\hat{\beta}^{(CN)}$  as follows:

$$\widehat{\text{MSE}}(\hat{\beta}^{(N)}) = \frac{1}{MC} \sum_{i=1}^{MC} (\hat{\beta}_i^{(N)} - \beta)(\hat{\beta}_i^{(N)} - \beta)',$$

$$\widehat{\text{MSE}}(\hat{\beta}^{(CN)}) = \frac{1}{MC} \sum_{i=1}^{MC} (\hat{\beta}_i^{(CN)} - \beta)(\hat{\beta}_i^{(CN)} - \beta)'.$$

5.1. Case 1

We assume  $X \sim \Gamma(k, \lambda), U \sim N(0, \sigma^2)$ . Let  $\beta_0 = 0.2, \beta_1 = 0.3, k = 2, \lambda = 1.2, n = 500, MC = 1000$ . We perform simulations with  $\sigma^2 = 0.05, 0.5, 2$ . Note that we assume that the true value of  $\sigma^2$  is known. We estimate  $k, \lambda$  in the formula of the corrected naive estimator by the moment method in terms of  $W$  because the value of  $X$  cannot be directly observed.

$$\hat{k} = \left( \frac{1}{n} \sum_{i=1}^n w_i \right) \hat{\lambda},$$

$$\hat{\lambda} = \frac{\frac{1}{n} \sum_{i=1}^n w_i}{\frac{1}{n} \sum_{i=1}^n (w_i - \bar{w})^2 - \sigma^2},$$

where  $w_i$  ( $i = 1, \dots, n$ ) is the samples of  $W$ .

Table 1 shows the estimated bias of the true  $\beta$ . Asy.Bias  $\hat{\beta}_0$  and Asy.Bias  $\hat{\beta}_1$  denote the theoretical asymptotic biases of  $\hat{\beta}_0^{(N)}$  and  $\hat{\beta}_1^{(N)}$ , respectively, given in Theorem 1. The bias correction of the naive estimator is performed by the corrected naive estimator. With increasing  $\sigma^2$ , the bias of the naive estimator increases. However, the bias of the corrected naive estimator is small for large  $\sigma^2$ .

**Table 1.** Estimated bias of a Gamma distribution with a Normal error.

		Asy.Bias $\hat{\beta}_0$	$\widehat{\text{BIAS}}(\hat{\beta}_0)$	Asy.Bias $\hat{\beta}_1$	$\widehat{\text{BIAS}}(\hat{\beta}_1)$
$\sigma^2 = 0.05$	Naive	0.01111	0.01139	-0.005993	-0.007199
	CN	0	0.00003532	0	0.0002603
$\sigma^2 = 0.5$	Naive	0.09912	0.1025	-0.05297	-0.05582
	CN	0	0.007817	0	0.0007142
$\sigma^2 = 2$	Naive	0.2757	0.2774	-0.1454	-0.1472
	CN	0	-0.009493	0	0.002736

Table 2 shows the estimated MSE of the true  $\beta$ . Asy.MSE  $\hat{\beta}_0$  and Asy.MSE  $\hat{\beta}_1$  denote the theoretical asymptotic MSEs of  $\hat{\beta}_0^{(N)}$  and  $\hat{\beta}_1^{(N)}$ , respectively, given in Theorem 1. The MSE of the corrected naive estimator is smaller than that of the naive estimator in all cases.

**Table 2.** Estimated MSE of a Gamma distribution with a normal error.

		Asy.MSE $\hat{\beta}_0$	$\widehat{\text{MSE}}(\hat{\beta}_0)$	Asy.MSE $\hat{\beta}_1$	$\widehat{\text{MSE}}(\hat{\beta}_1)$
$\sigma^2 = 0.05$	Naive	0.0001235	0.003003	0.00003592	0.0004536
	CN	0	0.002920	0	0.0004254
$\sigma^2 = 0.5$	Naive	0.009824	0.01362	0.002806	0.003508
	CN	0	0.003806	0	0.0006354
$\sigma^2 = 2$	Naive	0.07600	0.08124	0.02115	0.02214
	CN	0	0.01021	0	0.002160

5.2. Case 2

We assume  $X \sim \Gamma(k_1, \lambda), U \sim \Gamma(k_2, \lambda)$ . Let  $\beta_0 = 0.2, \beta_1 = 0.3, k_1 = 2, \lambda = 1.2, n = 500, MC = 1000$ . We perform simulations with  $k_2 = 0.072, 0.72, 2.88$ . Similarly, we assume that the true value of  $k_2$  is known. We estimate  $k_1, \lambda$  in the formula of the corrected naive estimator by the moment method in terms of  $W$  because the value of  $X$  cannot be directly observed.

$$\hat{k}_1 = \left( \frac{1}{n} \sum_{i=1}^n w_i \right) \hat{\lambda} - k_2,$$

$$\hat{\lambda} = \frac{\frac{1}{n} \sum_{i=1}^n w_i}{\frac{1}{n} \sum_{i=1}^n (w_i - \bar{w})^2},$$

where  $w_i (i = 1, \dots, n)$  is the samples of  $W$ .

Table 3 shows the estimated bias of the true  $\beta$ . Similarly, the bias correction of the naive estimator is performed by the corrected naive estimator. The bias of the corrected naive estimator is small when the variance of the error is large. Table 4 shows the estimated MSE of the true  $\beta$ . The MSE of the corrected naive estimator is also smaller than that of the naive estimator.

**Table 3.** Estimated bias of a Gamma distribution with a Gamma error.

		Asy.Bias $\hat{\beta}_0$	$\widehat{\text{BIAS}}(\hat{\beta}_0)$	Asy.Bias $\hat{\beta}_1$	$\widehat{\text{BIAS}}(\hat{\beta}_1)$
$k_2 = 0.072$	Naive	-0.002634	-0.005415	-0.007887	-0.008874
	CN	0	-0.0006636	0	0.0002777
$k_2 = 0.72$	Naive	-0.02090	-0.01725	-0.06378	-0.06475
	CN	0	-0.0002963	0	-0.003184
$k_2 = 2.88$	Naive	-0.04953	-0.05439	-0.1558	-0.1569
	CN	0	0.002954	0	-0.003224

**Table 4.** Estimated MSE of a Gamma distribution with a Gamma error.

		Asy.MSE $\hat{\beta}_0$	$\widehat{\text{MSE}}(\hat{\beta}_0)$	Asy.MSE $\hat{\beta}_1$	$\widehat{\text{MSE}}(\hat{\beta}_1)$
$k_2 = 0.072$	Naive	0.08533	0.003109	0.000006940	0.0005384
	CN	0	0.003074	0	0.0004743
$k_2 = 0.72$	Naive	0.05580	0.005320	0.0004368	0.004894
	CN	0	0.004457	0	0.0008818
$k_2 = 2.88$	Naive	0.02080	0.01147	0.002453	0.02553
	CN	0	0.007401	0	0.001963

6. Real Data Analysis

In this section, we apply the naive and corrected naive estimators to real data in two cases. First, we consider football data provided by Understat (2014). In this work, we focus on Goals and expected Goals (xG) in data on  $N = 24,580$  matches over 6 seasons between 2014–2015 and 2019–2020 from the Serie A, the Bundesliga, La Liga, the English Premier League, Ligue 1, and the Russian Premier League. Detail, such as the types and descriptions of the features, used in this section are provided in Table 5.

**Table 5.** Details of the variables.

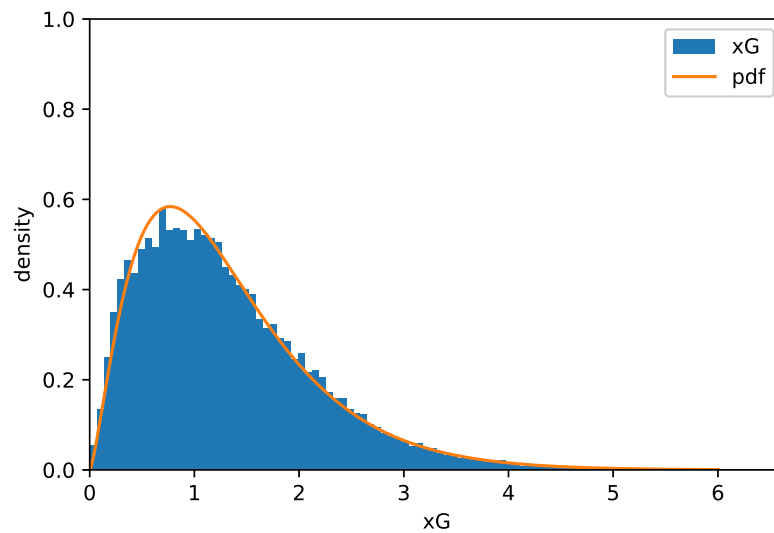
Features	Type	Description
Goals	counting	number of goals scored in the match
xG	continuous	performance metric used to evaluate football team and player performance

We use goals as an objective variable  $Y$  and  $xG$  as an explanatory variable  $X$  and assume  $Y|X \sim Po(\exp(\beta_0 + \beta_1 X))$  as the true model. Thus, this Poisson regression model refers to the extent to which expected goals ( $xG$ ) explains (true) goals. We assume that the true parameter  $\beta$  is obtained by the estimate from all  $N$  data.

As a diagnostic technique, we calculate a measure of goodness-of-fit to verify that the dataset follows a Poisson regression model. Table 6 shows estimates of  $\phi$  and  $R_{McF}$  (McFadden 1974), where  $R_{McF}$  is the ratio of the log-likelihood estimate to the initial log-likelihood.  $\phi = \mathbf{V}[Y|X] / \mathbf{E}[Y|X]$  is an overdispersion parameter. We may consider that overdispersion is not observed because  $\phi = 1$  equates to the standard Poisson regression model. The estimated value of  $\beta$  is  $(-0.5225, 0.5308)'$ . Thus, we use this estimate as a true value. We assume  $X (xG) \sim \Gamma(k_1, \lambda)$  and obtain estimates of  $k_1, \lambda$  as  $k_1 = 2.425, \lambda = 1.851$  (see Figure 1).

**Table 6.** Estimates of  $\phi$  and  $R_{McF}$ .

$\hat{\phi}$	$\widehat{R_{McF}}$
0.8907	0.1589



**Figure 1.** Distribution of  $xG$ .

Expected goals ( $xG$ ) is a performance metric used to represent the probability of a scoring opportunity that may result in a goal.  $xG$  is typically calculated from shot data. The measurer assigns a probability of scoring to a given shot and calculates the sum of the probabilities over a single game as  $xG$ . Observation error may occur in subjective evaluations. We can consider the situation that a high scorer happened to rate. Thus, we assume that  $X$  includes a stochastic error  $U$  given as

$$W = X + U.$$

Because  $W$  must be a positive value, we choose a positive error by  $U \sim \Gamma(k_2, \lambda)$  with  $k_2 = k_1/10, k_1/3, k_1$ . We sample 1000 random samples from among all  $N$  samples to obtain the values of the estimates of  $\beta$ s. We repeat the estimations  $MC = 10,000$  times to obtain the Monte Carlo mean of  $\beta$ s. The bias is calculated by the difference between the Monte Carlo mean and the true value.

Table 7 shows the estimated bias calculated by 10,000 simulations. The estimated bias of the corrected naive estimator is smaller than that of the naive estimator in all cases.

**Table 7.** Estimated bias and asymptotic bias in football data.

		Asy.Bias $\hat{\beta}_0$	$\widehat{\text{BIAS}}(\hat{\beta}_0)$	Asy.Bias $\hat{\beta}_1$	$\widehat{\text{BIAS}}(\hat{\beta}_1)$
$k_2 = k_1/10$	Naive	-0.01148	-0.01337	-0.03534	-0.03471
	CN	0	-0.001804	0	0.0006200
$k_2 = k_1/3$	Naive	-0.03263	-0.02383	-0.1020	-0.1067
	CN	0	0.008176	0	-0.005575
$k_2 = k_1$	Naive	-0.06889	-0.04692	-0.2210	-0.2291
	CN	0	0.01871	0	-0.01215

Next, we apply the naive and corrected naive estimators to financial data based on data collected in the FinAccess survey conducted in 2019, provided by Kenya National Bureau of Statistics (2019). In this study, we focus on the values labelled as finhealthscore and Normalized Household weights, with a sample size of  $N = 8669$ . Details of the features used in this section, such as their types and descriptions, are provided in Table 8.

**Table 8.** Details of the variables.

Features	Type	Description
finhealthscore	counting	Score of financial health for households
Normalized Household weights	continuous	Weighted and normalized households

We use finhealthscore as an objective variable  $Y$  and normalized household weights as an explanatory variable  $X$  and assume  $Y|X \sim Po(\exp(\beta_0 + \beta_1 X))$  as the true model. We further assume that the true parameter  $\beta$  is obtained by the estimate from all  $N$  data.

As a diagnostic technique, we calculate a measure of goodness-of-fit to verify that the dataset follows a Poisson regression model. Table 9 shows estimates of  $\phi$  and  $R_{McF}$  (McFadden 1974). Overdispersion tends to occur to some extent in this Poisson regression model because the estimate of  $\phi$  is greater than 1. The estimated value of  $\beta$  is  $(1.0442, 0.1568)'$ . As in the previous example, we regard the estimate as a true value. We assume  $X \sim \Gamma(k_1, \lambda)$  and obtain estimates of  $k_1, \lambda$  as  $k_1 = 2.0746, \lambda = 2.0746$  (see Figure 2).

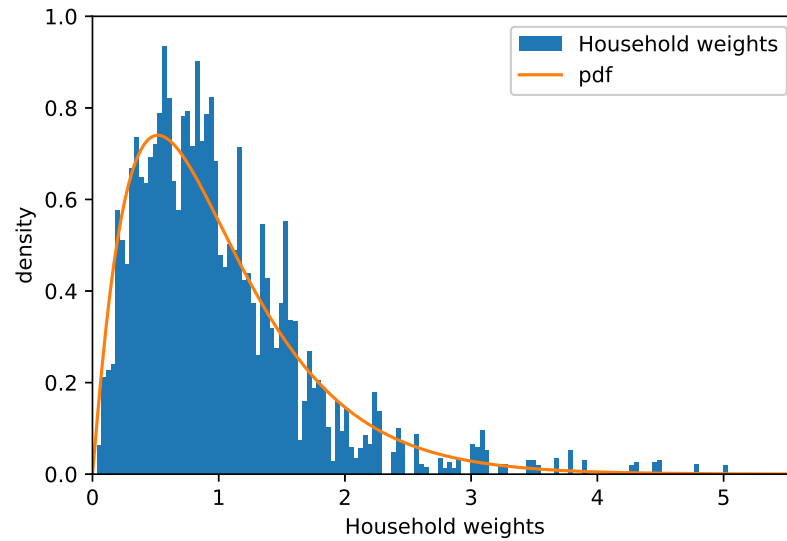
**Table 9.** Estimates of  $\phi$  and  $R_{McF}$ .

$\hat{\phi}$	$\widehat{R}_{McF}$
1.4360	0.4478

According to Kenya National Bureau of Statistics (2019), the data from the FinAccess survey were weighted and adjusted for non-responses to obtain a representative dataset at the national and county level. Thus, we may consider the situation that  $X$  exhibits a stochastic error  $U$  as

$$W = X + U.$$

We assume a positive error by  $U \sim \Gamma(k_2, \lambda)$  with  $k_2 = k_1/10, k_1/3, k_1$  because the distribution of normalized household weights is positive. We sample random 1000 samples from among all  $N$  samples to obtain the values of the estimates of  $\beta$ s. We repeat the estimations over  $MC = 10,000$  iterations to obtain the Monte Carlo mean of  $\beta$ s. The bias is calculated by the difference between the Monte Carlo mean and the true value.



**Figure 2.** Distribution of normalized household weights.

Table 10 shows estimated bias calculated by 10,000 simulations. The estimated bias of the corrected naive estimator is smaller than that of the naive estimator in all cases.

**Table 10.** Estimated bias and asymptotic bias in financial data.

		Asy.Bias $\hat{\beta}_0$	$\widehat{\text{BIAS}}(\hat{\beta}_0)$	Asy.Bias $\hat{\beta}_1$	$\widehat{\text{BIAS}}(\hat{\beta}_1)$
$k_2 = k_1/10$	Naive	-0.0005704	-0.002225	-0.01327	-0.01207
	CN	0	-0.001628	0	0.001275
$k_2 = k_1/3$	Naive	-0.001581	-0.004088	-0.03694	-0.03522
	CN	0	-0.002404	0	0.002119
$k_2 = k_1$	Naive	-0.003204	-0.008314	-0.07534	-0.07283
	CN	0	-0.004744	0	0.004338

### 7. Discussion

In this study, we have proposed a corrected naive estimator as a consistent estimator for a Poisson regression model with a measurement error. Although Kukush et al. (2004) showed that the naive estimator has an asymptotic bias, the authors did not provide a method to correct this bias. Therefore, we developed an approach to estimate a Poisson regression model with an error. In contrast, the authors of Kukush et al. (2004) also proposed a corrected score estimator and a structural quasi-score estimator for a Poisson regression model with an error. These estimators are score-based and consistent for unknown parameters. Hence, a generalization of these estimators should be considered in future research. In addition, the model considered in the present work is restricted in the univariate case. Extending the explanatory variable to the multivariate case also remains a challenge of note.

**Author Contributions:** K.W. mainly worked this study supported by the second named author. K.W.: Derivation of the formulae, Proof of propositions, Application to the specific problems, Conduct the simulation study, Real data analysis, Coding of the programs. T.K.: Basic idea, Theoretical advice of the proof, Advice at each step, Whole checking. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data were obtained from <https://understat.com/> and <https://knbs.or.ke> (accessed on 11 February 2023).



**Acknowledgments:** The authors thank to four anonymous referees for giving us valuable and insightful comments on the first draft. We missed future views of the draft and did not discuss drawbacks of our approach. Thanks to their sincere support for the draft, the revised version is now improved.

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

## References

- Box, George Edward Pelham. 1963. The Effects of Errors in the Factor Levels and Experimental Design. *Technometrics* 5: 247–62. [CrossRef]
- Burr, Deborah. 1988. On Errors-in-Variables in Binary Regression-Berkson Case. *Journal of the American Statistical Association* 83: 739–43.
- Geary, ROBERT C. 1953. Non-Linear Functional Relationship between Two Variables When One Variable is Controlled. *Journal of the American Statistical Association* 48: 94–103. [CrossRef]
- Guo, Jie Q., and Tong Li. 2002. Poisson regression models with errors-in-variables: implication and treatment. *Journal of Statistical Planning and Inference* 104: 391–401. [CrossRef]
- Huwang, Longcheen, and YH Steve Huang. 2000. On error-in-variables in polynomial regression-Berkson case. *Statistica Sinica* 10: 923–36.
- Jiang, Fei, and Yanyuan Ma. 2020. Poisson Regression with Error Corrupted High Dimensional Features. *Statistica Sinica* 32: 2023–46. [CrossRef]
- Kenya National Bureau of Statistics (KNBS). 2019. Available online: <https://knbs.or.ke> (accessed on 11 February 2023).
- Kukush, Alexander, and Hans Schneeweiss. 2000. A Comparison of Asymptotic Covariance Matrices of Adjusted Least Squares and Structural Least Squares in Error Ridden Polynomial Regression. *Sonderforschungsbereich* 386: Paper 218. [CrossRef]
- Kukush, Alexander, Hans Schneeweiss, and Roland Wolf. 2004. Three Estimators for the Poisson Regression Model with Measurement Errors. *Statistical Papers* 45: 351–68. [CrossRef]
- McFadden, Daniel. 1974. Conditional logit analysis of qualitative choice behavior. *Computations in Statistics-Theory and Methods* 47: 105–42.
- Nakamura, Tsuyoshi. 1990. Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika* 77: 127–37. [CrossRef]
- Shklyar, Schneeweiss, and Hans Schneeweiss. 2005. A comparison of asymptotic covariance matrices of three consistent estimators in the Poisson regression model with measurement errors. *Journal of Multivariate Analysis* 94: 250–70. [CrossRef]
- Understat. 2014. Available online: <https://understat.com/> (accessed on 11 February 2023).

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

# Modelling of Loan Non-Payments with Count Distributions Arising from Non-Exponential Inter-Arrival Times

Yeh-Ching Low <sup>1,\*</sup>  and Seng-Huat Ong <sup>2,3</sup>

<sup>1</sup> Department of Computing and Information Systems, Sunway University, Petaling Jaya 47500, Malaysia

<sup>2</sup> Institute of Actuarial Science and Data Analytics, UCSI University, Kuala Lumpur 56000, Malaysia

<sup>3</sup> Institute of Mathematical Sciences, University of Malaya, Kuala Lumpur 50603, Malaysia

\* Correspondence: yehchingl@sunway.edu.my

**Abstract:** The number of non-payments is an indicator of delinquent behaviour in credit scoring, hence its estimation and prediction are of interest. The modelling of the number of non-payments, as count data, can be examined as a renewal process. In a renewal process, the number of events (such as non-payments) which has occurred up to a fixed time  $t$  is intimately connected with the inter-arrival times between the events. In the context of non-payments, the inter-arrival times correspond to the time between two subsequent non-payments. The probability mass function and the renewal function of the count distribution are often complicated, with terms involving factorial and gamma functions, and thus their computation may encounter numerical difficulties. In this paper, with the motivation of modelling the number of non-payments through a renewal process, a general method for computing the probabilities and the renewal function based on numerical Laplace transform inversion is discussed. This method is applied to some count distributions which are derived given the distributions of the inter-arrival times. Parameter estimation with maximum likelihood estimation is considered, with an application to a data set on number of non-payments from the literature.

**Keywords:** birth and renewal processes; loan default; non-payments; inter-arrival times; renewal function; over and under dispersion; Laplace transform



**Citation:** Low, Yeh-Ching, and Seng-Huat Ong. 2023. Modelling of Loan Non-Payments with Count Distributions Arising from Non-Exponential Inter-Arrival Times. *Journal of Risk and Financial Management* 16: 150. <https://doi.org/10.3390/jrfm16030150>

Academic Editor: Thanasis Stengos

Received: 31 January 2023

Revised: 15 February 2023

Accepted: 20 February 2023

Published: 23 February 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In credit scoring, default probabilities are often of interest to identify and manage the risk of bad loans. However, evaluation of default probabilities alone is insufficient to assess the risk and returns of bank funding (Dionne et al. 1996). Before an accepted loan is classified as a bad loan, there would have been several non-payments which come with costs incurred by reminders, collection, and other administrative charges. Therefore, instead of classification of a loan as either good or bad, a flexible alternative approach to risk evaluation is through the modelling of the number of non-payments (Karlis and Rahmouni 2007). The number of non-payments, which is a primary indicator of delinquent behaviour, are count data. Modelling of the counts of non-payments will be useful for estimating the probability of default. The basic model for count data is the well-known Poisson model which exhibits equi-dispersion where the mean is equal to its variance. As such, the Poisson model is often found to be inadequate in the presence of over- or under-dispersion. Various approaches have been proposed to extend or generalize the Poisson distribution. Examples of such approaches are: mixture models for heterogeneity (Gupta and Ong 2005), such as the negative binomial (NB) (Greenwood and Yule 1920) and Poisson-inverse Gaussian (P-iG) (Holla 1967; Sankaran 1968), Lagrange expansion generalization of the Poisson distribution (Consul and Jain 1973), and count distributions from renewal processes where the time between events are non-exponential distributions (Winkelmann 1995). In the context of modelling number of non-payments, truncated count models (Dionne et al. 1996), Poisson finite mixtures (Karlis and Rahmouni 2007) and non-parametric models (Mestiri and Farhat 2021) have been investigated in the literature.

It is well-known that, in a renewal process, if the waiting times are exponential and independent, we obtain the Poisson distribution for the event counts. In the context of loan non-payments, the inter-arrival times refer to the duration between two subsequent non-payments. Thomas et al. (2016) used Markov chains to model the payment patterns to estimate recover rates. This renewal process approach to derive count distributions has been considered by several researchers. Winkelmann (1995) derived the count distribution when the inter-arrival time is an Erlang distribution. Other distributions which have been considered by various authors to model the inter-arrival times are the gamma distribution (Winkelmann 1995), Weibull distribution (McShane et al. 2008), which is very popular in the field of reliability studies, Mittag-Leffler (Jose and Abraham 2011), Gumbel Type II (Jose and Abraham 2013), and generalized Weibull (Ong et al. 2015); see Table 1. The count distributions were mostly obtained using extensive numerical and analytical methods. For example, McShane et al. (2008) and Jose and Abraham (2013) used the polynomial expansion method to derive the count distribution for Weibull and Gumbel inter-arrival times, respectively. A different approach by From (2004) is to use a family of generalized Poisson distributions to approximate the renewal counting processes with Weibull, truncated normal and exponentiated Weibull inter-arrival times. Baker and Kharrat (2017) proposed the use of repeated convolutions of the discretized distributions with Richard extrapolation as well as an adaptation of De Pril’s method to compute probabilities in event count distributions from renewal processes. Nadarajah and Chan (2018) derived count distributions arising from 13 different inter-arrival time distributions and studied their fit to football home goals data using the algebraic manipulation package Maple. A similar perspective in the modelling of non-life insurance claims data was discussed by Maciak et al. (2021) through infinitely stochastic processes and Lindholm and Zakrisson (2022).

**Table 1.** Some existing count distributions in renewal theory.

Inter-Arrival Time Distribution	Probability Mass Function (pmf) of Corresponding Count Distribution
Gamma	$Pr\{N(t) = n\} = G(\alpha n, \beta t) - G(\alpha n + \alpha, \beta t),$ $G(\alpha n, \beta t) = \frac{1}{\Gamma(\alpha n)} \int_0^{\beta t} u^{\alpha n - 1} e^{-u} du$
Weibull	$Pr\{N(t) = n\} = \sum_{j=n}^{\infty} \frac{(-1)^{j+n} (\lambda t^c)^j \alpha_j^n}{\Gamma(cj+1)},$ $\alpha_j^0 = \frac{\Gamma(cj+1)}{\Gamma(j+1)}, j = 0, 1, 2, \dots, \alpha_j^{n+1} = \sum_{m=n}^{j-1} \alpha_m^n \frac{\Gamma(cj - cm + 1)}{\Gamma(j - m + 1)},$ $n = 0, 1, 2, \dots, j = n + 1, n + 2, n + 3, \dots$
Mittag-Leffler	$Pr(N(t) = n) = \sum_{j=n}^{\infty} \binom{j}{n} (-1)^{j-n} t^{j\alpha} / \Gamma(1 + j\alpha)$
Gumble Type II	$Pr\{N(t) = n\} = \sum_{j=n}^{\infty} \frac{(-1)^{j+n} (bt^{-a})^j \delta_j^n}{\Gamma(-aj+1)}, a < 0$ $\delta_j^0 = \frac{\Gamma(-aj+1)}{\Gamma(j+1)}, j = 0, 1, 2, \dots, \delta_j^{n+1} = \sum_{m=n}^{j-1} \delta_m^n \frac{\Gamma(-aj+am+1)}{\Gamma(j-m+1)}$ $n = 0, 1, 2, \dots, j = n + 1, n + 2, n + 3, \dots$
Generalized Weibull	$Pr\{N(t) = n\} = (a\alpha)^n \sum_{p=0}^{\infty} \frac{(-a/\lambda)^p}{\Gamma(\alpha(p+n)+1)} t^{\alpha(p+n)} c_n(p),$ $c_n(p) = \sum_{q=0}^p \binom{\lambda - 1}{q} \Gamma(\alpha(q + 1)) c_{n-1}(p - q), n \geq 1, c_0(p) = \binom{\lambda}{p} \Gamma(\alpha p + 1)$

The objectives of this paper are to propose the modelling of number of loan non-payments through the renewal process approach and to examine the computation of the pmf. Due to the rather involved computation of the probabilities mentioned previously, a simple, general and efficient method of computing the probabilities of count distributions arising from non-exponential inter-arrival time distributions of renewal processes is discussed to facilitate the statistical modelling. We consider the generalized Weibull, inverse Gaussian and convolution of two gamma distributions due to their greater generality, as they include, among others, the Weibull and gamma distributions as special cases. These inter-arrival times’ distributions have flexible hazard functions so that the corresponding

count distributions are able to cater for under-, equi- and over dispersion. This relationship between the inter-arrival times' hazard function and the dispersion of the corresponding count distribution has been proven by Winkelmann (1995). We propose an easily implemented and efficient method to compute the probabilities of the counts and, subsequently, the renewal function (expected number of renewals), given the Laplace transform of the inter-arrival times density function. The computation of the renewal function has been extensively studied by various authors, for example, in the case of the Weibull renewal function, see Smith and Leadbetter (1963); Constantine and Robinson (1997).

In Section 2, we briefly describe the relationship between the distribution of the inter-arrival times and the count distribution, as well as some existing count distributions. We focus on the case when the sequence of inter-arrival times is independent and identically distributed, which gives rise to the renewal process. Count distributions arising from inverse Gaussian and convolution of two gamma distributions as inter-arrival times are considered. In these sections, we assume that the inter-arrival time  $X_i$  is independent and identically distributed and we drop the index  $i$  from the notation, and thus  $X$  denotes the inter-arrival time. The proposed method for the computation of the count probabilities and its renewal function is discussed in Section 3. Section 4 details the application of the distributions on a data set on number of non-payments from the literature. We perform parameter estimation using maximum likelihood estimation. Finally, a concluding discussion is given in Section 5.

## 2. Modelling of Loan Non-Payment Counts

### 2.1. Count Distribution and Inter-Arrival Times Distribution

A counting process is a stochastic point process  $\{N(t), t \geq 0\}$  where  $N(t)$  represents the total number of events that have occurred by time  $t$ . In this paper, the number of events corresponds to the number of non-payments. Let  $S_n$  denote the waiting time to (or arrival time of) the  $n$ th non-payment, and  $X_n$  denote the time between the  $(n - 1)$ st and the  $n$ -th non-payment of this process, i.e., two subsequent non-payments. In the rest of this paper,  $X_n$  will be referred to as inter-arrival times. Therefore,  $S_0 = 0$  and  $S_n = \sum_{i=1}^n X_i, n \geq 1$ . If the sequence of inter-arrival times  $\{X_1, X_2, \dots\}$  is independent and identically distributed as  $f(x)$  with cumulative distribution function (cdf)  $F(x)$ , the counting process  $\{N(t), t \geq 0\}$  is known as a renewal process. In a renewal process, the distribution function of  $S_n$  can be obtained as the  $n$ -fold convolution  $F_n(x)$  of the distribution of  $X_i$  and  $F_0(t) = 1$ . In this case, the renewal function or expected number of non-payments  $E[N(t)]$  and the distribution of  $N(t)$  can be obtained from the relationship  $N(t) \geq n \Leftrightarrow S_n \leq t$ . As such, the probability mass function (pmf) of the count distribution is

$$Pr\{N(t) = n\} = Pr\{S_n \leq t\} - Pr\{S_{n+1} \leq t\} = F_n(t) - F_{n+1}(t), \tag{1}$$

where  $n = 0, 1, \dots$ , and  $F_n(x)$  is the cdf of  $S_n$ . The renewal function is defined as

$$H(t) = E[N(t)] = \sum_{i=1}^{\infty} F_i(t). \tag{2}$$

A d example is, when the inter-arrival times are exponentially distributed, the counting process is a Poisson process with intensity  $\lambda(t) = \lambda$  with pmf

$$Pr\{N(t) = n\} = \frac{e^{-\lambda t} (\lambda t)^n}{n!}, n = 0, 1, 2, \dots$$

The Laplace transform  $\varphi(s)$  of a function  $f(x)$  is defined as  $\varphi(s) = \int_0^{\infty} e^{-sx} f(x) dx$ , where  $s$  is a complex number. The Laplace transform exists for the function  $f(x)$  defined over  $(0, \infty)$ , whenever the integral converges. Since the inter-arrival times  $X_i$ 's are independent and identically distributed, the Laplace transform of the arrival time  $S_n = \sum_{i=1}^n X_i$  is simply

the  $n$ -fold convolution of the Laplace transform of  $X_i$ . Consequently, the Laplace transform of the count distribution is derived as

$$\varphi_n(s) = L(Pr\{N(t) = n\}) = L(F_n(t) - F_{n+1}(t)) = \frac{1 - \varphi(s)}{\varphi(s)} (\varphi(s))^n, \tag{3}$$

where  $\varphi(s)$  is the Laplace transform of the inter-arrival time's probability density function (pdf)  $f(x)$ . On the other hand, the Laplace transform of (2) is  $L(E[N(t)]) = \frac{\varphi(s)}{s(1-\varphi(s))}$ ,  $|\varphi(s)| < 1$ .

In the existing literature, Poisson distribution and negative binomial distribution have been proposed for modelling non-payments (Dionne et al. 1996). In the following sections, we present alternative count distributions for modelling of non-payments examined from the perspective of their inter-arrival times.

### 2.1.1. Count Distribution for Generalized Weibull Duration

The pdf of a generalized Weibull distribution is given as

$$f(x; \alpha, \lambda) = a\alpha x^{\alpha-1} (1 - ax^\alpha/\lambda)^{\lambda-1}, \tag{4}$$

for  $a, \alpha > 0, x > 0$  if  $\lambda \leq 0$  and  $0 < x < (\lambda/a)^{1/\alpha}$  if  $\lambda > 0$  (Mudholkar et al. 1996). An important limiting case is the Weibull distribution when  $\lambda \rightarrow \infty$ , with pdf  $f(x; a, \alpha, \lambda) = a\alpha x^{\alpha-1} e^{-ax^\alpha}$ . We shall re-write the Weibull pdf as  $f(x; a, \alpha, \lambda) = \left(\frac{\lambda}{a}\right) \left(\frac{x}{\alpha}\right)^{\lambda-1} e^{-\left(\frac{x}{\alpha}\right)^\lambda}$ . The generalized Weibull distribution has a flexible and closed form hazard function.

Ong et al. (2015) applied the Laplace transform technique and a formal Taylor expansion to derive the count distribution for generalized Weibull duration. The count distribution has pmf given by

$$Pr\{N(t) = n\} = (a\alpha)^n \sum_{p=0}^{\infty} \frac{(-a/\lambda)^p}{\Gamma(\alpha(p+n) + 1)} t^{\alpha(p+n)} c_n(p), \tag{5}$$

where  $c_n(p) = \sum_{q=0}^p \binom{\lambda-1}{q} \Gamma(\alpha(q+1)) c_{n-1}(p-q), n \geq 1$  and  $c_0(p) = \binom{\lambda}{p} \Gamma(\alpha p + 1)$ .

When  $n = 0, Pr\{N(t) = 0\} = (1 - at^\alpha/\lambda)^\lambda$ . This count model is able to model under-, equi- and over-dispersion, since the generalized Weibull hazard function can be increasing, constant or decreasing. Special cases are as follows:

- When  $\lambda < 0$  and  $\alpha = 1$ , we obtain the count distribution with Lomax duration. Its pmf is given by Ong et al. (2015) as

$$Pr\{N(t) = n\} = (at)^n \sum_{p=0}^{\infty} \frac{(a/\Gamma)^p}{\Gamma(p+n+1)} t^p c_n(p). \tag{6}$$

- When  $\lambda \rightarrow \infty$ , we obtain the Weibull count distribution and Ong et al. (2015) gives its pmf as

$$Pr\{N(t) = n\} = (a\alpha)^n \sum_{p=0}^{\infty} \frac{(-a)^p}{\Gamma(\alpha(p+n) + 1)} t^{\alpha(p+n)} c_n(p), \tag{7}$$

where  $c_n(p) = \sum_{q=0}^p \frac{\Gamma(\alpha(q+1))}{\Gamma(q+1)} c_{n-1}(p-q), n \geq 1$  and  $c_0(p) = \frac{\Gamma(\alpha p + 1)}{\Gamma(p+1)}$ . When  $n = 0, Pr\{N(t) = 0\} = e^{-at^\alpha}$ . McShane et al. (2008) applied Taylor series approximation in the derivation of the Weibull count pmf which they have found to be computationally feasible.

- Furthermore, when  $\alpha = 1$ , (5) reduces to the Poisson pmf.

### 2.1.2. Count Distribution for Gamma Duration

Let  $X$  have a gamma distribution with pdf given by

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{8}$$

for  $x > 0$  and  $\alpha, \beta > 0$ . It has mean  $E(X) = \alpha/\beta$  and variance  $\text{Var}(X) = \alpha/\beta^2$ . The hazard function of the gamma distribution is not available in closed form but its behaviour is well-known as being monotonic increasing ( $\alpha > 1$ ), decreasing ( $\alpha < 1$ ) or constant ( $\alpha = 1$ ). When  $\alpha = 1$ , we obtain the exponential distribution. The Laplace transform of the gamma distribution is given as  $\varphi(s) = \left(\frac{\beta}{\beta+s}\right)^\alpha$ . The gamma distribution has the advantage of having a reproductive property, hence the arrival time  $S_n$  is also gamma distributed.

Winkelmann (1995) has studied the count process with gamma inter-arrival times and gives its pmf as

$$\text{Pr}\{N(t) = n\} = G(\alpha n, \beta t) - G(\alpha n + \alpha, \beta t) \tag{9}$$

where  $G(\alpha n, \beta t) = \frac{1}{\Gamma(\alpha n)} \int_0^{\beta t} u^{\alpha n-1} e^{-u} du$ , the integral is the lower incomplete gamma function. Since the pmf is not available in closed form, Winkelmann (1995) suggested using numerical methods for its computation. The gamma count distribution inherits the properties of the gamma distribution's hazard function; thus it is able to model over dispersion ( $\alpha < 1$ ) and under dispersion ( $\alpha > 1$ ). Its expected value is given by  $E[N(t)] = \sum_{i=1}^\infty G(\alpha i, \beta t)$ . Special cases are as follows:

- When  $\alpha = 1$ , the count distribution simplifies to the Poisson distribution.
- For integer values of  $\alpha$ , Winkelmann (1995) has derived the Erlangian count distribution with pmf given as

$$\text{Pr}\{N(t) = n\} = e^{-\beta t} \sum_{i=0}^{\alpha-1} \frac{(\beta t)^{\alpha n+i}}{(\alpha n+i)!}, n = 0, 1, 2, \dots \tag{10}$$

### 2.1.3. Count Distribution for Convolution of Two Gamma Durations

If we represent the inter-arrival time  $X$  as a sum of two independent gamma random variables, then  $X$  has a convolution of two gamma distributions. Its density function has been studied by various authors; see Johnson et al. (2005) for a brief overview. We shall adapt the density function given by Moschopoulos (1985) for the sum of  $n$  independent gamma random variables, which is derived from the  $n$ -convolutions of the moment generating function. Let  $X = X_1 + X_2$ , where  $X_i, i = 1, 2$ , are distributed as gamma with parameters  $\alpha_i$  and  $\beta_i$  respectively. We obtain the density function of  $X$  as

$$f(x; \rho, \beta_1) = \left(\frac{\beta_1}{\beta_2}\right)^{\alpha_2} \sum_{k=0}^\infty \frac{\delta_k x^{\rho+k-1} \exp\left(-\frac{x}{\beta_1}\right)}{\Gamma(\rho+k)\beta_1^{\rho+k}} \tag{11}$$

for  $x > 0, \alpha_i > 0, \beta_i > 0$  where  $\beta_1 = \min(\beta_1, \beta_2), \rho = \alpha_1 + \alpha_2, \delta_{k+1} = \frac{1}{k+1} \sum_{i=1}^{k+1} i \Gamma_i \delta_{k+1-i}$  for  $k = 0, 1, 2, \dots$ , and  $\Gamma_k = \left\{ \alpha_2 \left(1 - \frac{\beta_1}{\beta_2}\right)^k \right\}$ . The convolution of two gamma distributions has an increasing hazard function when its two component distributions have an increasing hazard function, but convolutions of two distributions, both with decreasing hazard function, may give rise to a distribution with increasing hazard function. Therefore, we expect the count distribution to be more flexible in modelling over-dispersed and under-dispersed count data. As a special case, when  $\alpha_1 = \alpha_2 = 1$ , we obtain the convolution of two exponential distributions which has an increasing hazard function.

**Proposition 1.** *If the inter-arrival time (duration) has a convolution of two gamma distributions with pdf (3.1.1), the count distribution has pmf given by*

$$Pr\{N(t) = n\} = C_n(t, \alpha_1, \alpha_2, \beta_1, \beta_2) - C_{n+1}(t, \alpha_1, \alpha_2, \beta_1, \beta_2), \tag{12}$$

where  $C_n(t, \alpha_1, \alpha_2, \beta_1, \beta_2) = (\beta_1^{\alpha_1} \beta_2^{\alpha_2})^n \left\{ \frac{t^{n(\alpha_1+\alpha_2)}}{\Gamma(1+n(\alpha_1+\alpha_2))} \Phi_2(n\alpha_1, n\alpha_2; 1+n(\alpha_1+\alpha_2); -\beta_1 t, -\beta_2 t) \right\}$  and  $\Phi_2(b, b'; c; w, z) = \sum_{k,l=0}^{\infty} \frac{(b)_k (b')_l}{(c)_{k+l}} \frac{w^k z^l}{k! l!}$ .

#### 2.1.4. Count Distribution for Inverse Gaussian Duration

The inverse Gaussian (IG) distribution is also known as the first passage time distribution of Brownian motion with positive drift. Let  $X$  have an IG distribution with pdf given by

$$f(x; \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi}} x^{-\frac{3}{2}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\}, \tag{13}$$

for  $x > 0$ , where  $\mu, \lambda > 0$  (Johnson et al. 2005, p. 261). It is a unimodal distribution and has applications in modelling survival period, service time, equipment lives, hospital stay duration, employee service times and duration of strikes. Chhikara and Folks (1977) have discussed the application of the inverse Gaussian distribution in reliability and showed that the distribution has a non-monotonic hazard function with an almost increasing failure rate. There are several parameterizations of the IG distributions, but we adopt this particular one because it is expressed in terms of its mean  $E(X) = \mu$  and  $\lambda$  is the scale parameter. The shape of the distribution is determined by the ratio  $\lambda/\mu$  and the pdf is highly skewed for moderate values of this ratio. The Laplace transform is derived by Seshadri (1999) as

$$\varphi(s) = \exp\left\{\frac{\lambda}{\mu} \left(1 - \sqrt{1 + \frac{2s\mu^2}{\lambda}}\right)\right\}, s \geq 0 \tag{14}$$

when  $\mu \rightarrow \infty$ , we obtain a one-parameter limiting form of IG, known as the distribution of the first passage time of drift-free Brownian motion. Its pdf is given as  $f(x; \lambda) = \sqrt{\frac{\lambda}{2\pi}} x^{-\frac{3}{2}} \exp\left(-\frac{\lambda}{2x}\right)$  with  $x > 0$ , where  $\lambda > 0$  (Johnson et al. 2005). The expected value and variance of this distribution are infinite. On the other hand, when  $\mu = 1$ , the distribution is also known as the Wald distribution.

The count distribution with inverse Gaussian inter-arrival times has also been proposed (Nadarajah and Chan 2018) with the probability mass function given in terms of the convolution of inter-arrival distributions  $F_n(x)$ , involving the standard normal cumulative distribution function. We derive an explicit expression for the inverse Gaussian count distribution, given in the following proposition.

**Proposition 2.** *If the inter-arrival time has an inverse Gaussian distribution with pdf (13), the count distribution has pmf given by*

$$Pr\{N(t) = n\} = \sum_{k=0}^{\infty} \sum_{l=0}^k \frac{n^{k-l}}{(l+1)!(k-l)!} \left(\frac{\lambda}{\mu}\right)^{k+1} c_k(m), \tag{15}$$

where  $c_k(m) = \sum_{m=0}^{k+1} \binom{k+1}{m} (-1)^m \left(\sum_{\nu=0}^{\infty} \binom{m}{\nu} \frac{1}{\Gamma(1-\nu)} \left(\frac{2\mu^2}{\lambda t}\right)^{\nu}\right)$ .

#### 2.2. Computation of the Probabilities of Count Distribution

The computation of the probabilities for most of the count distributions, such as the generalized Weibull count distribution (5), involves an infinite series and/or gamma functions  $\Gamma(x)$ , which tends to quickly numerically overflow. As such, we propose a

computational method whereby the probability function of the counts can be recovered by numerically inverting the Laplace transform (3). Using this method, given the inter-arrival time distribution and its Laplace transform, we will be able to compute the corresponding count probabilities.

For some common functions, the inverse Laplace transforms  $f(x)$  are readily available from existing tables (Erdelyi et al. 1953). Otherwise, there are explicit formulae for inverting a Laplace transform  $\varphi(s)$ , such as the Bromwich inversion integral formula and the Post-Widder inversion formula. In most cases, it is difficult to find an analytical expression for the inverse Laplace transform using these formula and, therefore, a numerical inversion is necessary. There are numerous methods for numerical inversion of Laplace transforms in the existing literature; for a comprehensive review, see (Abate and Valkó 2004; Dubner and Abate 1968). In our study, we use a numerical inversion algorithm which is based on the Bromwich inversion integral and gives good results for smooth functions. The algorithm was originally proposed by Dubner and Abate (1968), improved by Abate and Whitt (1992) and discussed by Abate and Whitt (1995) and Abate et al. (2000) for the numerical inversion of Laplace transforms of probability distributions. The Bromwich inversion integral formula is given as

$$f(x) = L^{-1}(\varphi(s)) = \lim_{R \rightarrow \infty} \frac{1}{2\pi i} \int_{a-iR}^{a+iR} \varphi(s)e^{sx} ds, \tag{16}$$

where  $a$  is another real number such that  $a > s_0$  and  $i = \sqrt{-1}$ . The numerical inversion algorithm is developed by first applying the trapezoidal rule to the integral in (16), and subsequently using a Fourier-series method for approximation. Based on the algorithm, we obtain the following formula for computing the count probabilities

$$Pr\{N(t) = n\} = \frac{e^{A/2}}{2s} Re\left(\varphi_n\left(\frac{A}{2s}\right)\right) + \frac{e^{A/2}}{s} \sum_{k=1}^{\infty} (-1)^k Re\left(\varphi_n\left(\frac{A + 2k\pi i}{2s}\right)\right), \tag{17}$$

where  $\varphi_n(\cdot)$  is as defined in (3).

The convergence of the infinite sum in (17) can be accelerated by applying the well-known Euler’s algorithm for alternating series. Therefore, the count probabilities are approximated using the following formula

$$Pr\{N(t) = n\} \approx \sum_{k=0}^m \binom{m}{k} 2^{-m} s_{p+k}(s), \tag{18}$$

where  $s_p(s)$  is the  $p$ th partial sum

$$s_p(s) = \frac{e^{A/2}}{2s} Re\left(\varphi_n\left(\frac{A}{2s}\right)\right) + \frac{e^{A/2}}{s} \sum_{k=1}^p (-1)^k Re\left(\varphi_n\left(\frac{A + 2k\pi i}{2s}\right)\right). \tag{19}$$

The choice of  $A$  affects the discretization error which results from using the trapezoidal rule. We use Abate and Whitt’s (1995) suggestion to set  $A = 18.4$ ,  $p = 38$  and  $m = 11$ . The value of  $p$  may be increased when necessary. The algorithm can be implemented in programming languages which provide for complex number computation, such as MATLAB©.

### 2.3. Renewal Function

There are many studies on the approximation of the renewal function. Using a generalized cubic splining algorithm which provides piecewise polynomial approximations to recursively defined convolution integrals, Baxter et al. (1982) has tabulated the renewal function and variance function for renewal processes with gamma, inverse Gaussian, lognormal, truncated normal and Weibull inter-arrival times. However, they noted that the convergence of the algorithm is slow for some of the parameter values. Chaudhry et al. (2013) took a slightly different approach by using the probability function obtained from numerically inverting the Laplace transform in rational function form to calculate the



renewal function and variance of several count distributions. They obtained the distribution function, mean and variance of  $N(t)$  using the method of roots for numerically inverting the Laplace transform when it can be expressed as a rational function. They also studied the Padè approximation method to obtain an approximate rational function for the Laplace transform when it is not a rational function. In addition, they used the Padè approximation method prior to the roots method when the Laplace transform could not be expressed as a rational function, such as in the case of gamma and inverse Gaussian distribution.

### 3. Numerical Results

#### 3.1. Count Probabilities

To illustrate the accuracy of this numerical Laplace transform inversion method, we apply it in calculating the count probabilities for generalized Weibull duration and Erlangian duration and compare the values to those obtained using Formulas (5) and (10), respectively. The formula in Equation (10) is in closed form and simple enough to compute, hence there is no need to use the method which we propose here, but it serves as a good example for this comparison. Since the Laplace transform of the generalized Weibull density function is not available in closed form, we can approximate it using Gaussian quadrature. The computed probabilities are presented in Table 2. The count probabilities for generalized Weibull duration are computed when  $a = 1, \alpha = 1$  and  $\lambda = -2, t = 0.25$  and  $t = 1$ . For the Erlangian count distribution, we compute the probabilities when  $\alpha = 2, \beta = 0.8, t = 0.25$  and  $t = 1$ . In all cases, we find that our approximation is accurate up to at least seven decimal places. To illustrate the issue of overflowing which might occur, we present the count probabilities for generalized Weibull duration when  $a = 2, \alpha = 1$  and  $\lambda = -2$  and  $t = 1$  in Table 3. It is clear that, in this case, there is a numerical error in the computation of the probabilities with Formula (5) when  $n = 1, 2$  due to instability caused by the presence of an infinite sum in Equation (5) and truncation error.

**Table 2.** Computation of probabilities for (a) generalized Weibull, and (b) Erlangian count distributions using the proposed method and pmf formula.

$n$	$Pr\{N(t) = n\}$ $t = 0.25$			$Pr\{N(t) = n\}$ $t = 1$		
	Proposed Method	Pmf Formula	Difference	Proposed Method	Pmf Formula	Difference
0	0.790123462190233	0.790123456790123	5.4001 (−9)	0.444444446077630	0.444444444444444	1.6331 (−9)
1	0.185268558281666	0.185268554955749	3.3259 (−9)	0.341447772405153	0.341447770099717	2.3054 (−9)
2	0.022624019619715	0.022624018469588	1.1501 (−9)	0.152421254574663	0.152421252253988	2.3207 (−9)
3	0.001862447034136	0.001862446759278	2.7486 (−10)	0.047632000079489	0.047631998279757	1.7997 (−9)
4	0.000115528824677	0.000115528774610	5.0067 (−11)	0.011418307350013	0.011418306220399	1.1296 (−9)
5	0.000005746921940	0.000005746914580	7.3600 (−12)	0.002217009636005	0.002217009042290	5.9371 (−10)
6	0.000000238568216	0.000000238567310	9.0600 (−13)	0.000361439244000	0.000361438976100	2.6790 (−10)
7	0.000000008496400	0.000000008496304	9.0600 (−13)	0.000050759289875	0.000050759184107	1.0577 (−10)

(a) Generalized Weibull count distribution

$n$	$Pr\{N(t) = n\}$ $t = 0.25$			$Pr\{N(t) = n\}$ $t = 1$		
	Proposed Method	Pmf Formula	Difference	Proposed Method	Pmf Formula	Difference
0	0.982476912658251	0.982476903693578	8.9647 (−9)	0.808792138560495	0.808792135410999	3.1495 (−9)
1	0.017466257275868	0.017466256065664	1.2102 (−9)	0.182128011589934	0.182128006788847	4.8011 (−9)
2	0.000056765366099	0.000056765332213	3.3886 (−11)	0.008895517173780	0.008895515278950	1.8948 (−9)
3	0.000000074855777	0.000000074855383	3.9400 (−13)	0.000182292662905	0.000182292332810	3.3009 (−10)
4	0.000000000053140	0.000000000053138	2.0000 (−15)	0.000002035889418	0.000002035857392	3.2026 (−11)
5	0.000000000000024	0.000000000000024	0.0000	0.000000014264304	0.000000014262333	1.9710 (−12)
6	0.000000000000000	0.000000000000000	0.0000	0.00000000068513	0.00000000068429	8.4000 (−14)
7	0.000000000000000	0.000000000000000	0.0000	0.00000000000241	0.00000000000239	1.9999 (−15)

(b) Erlangian count distribution

**Table 3.** Count probabilities for generalized Weibull count distribution when  $a = 2$ ,  $\alpha = 1$  and  $\lambda = -2$  and  $t = 1$ .

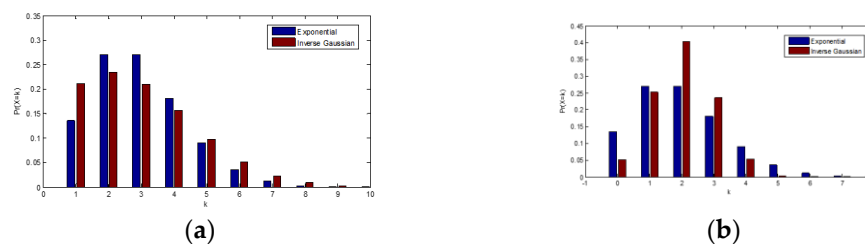
$n$	$Pr\{N(t) = n\}$	
	Formula	Proposed Inverse Laplace Transform Method
0	0.2500	0.2500
1	63.5982	0.2971
2	2.3327	0.2305
3	0.1839	0.1317
4	0.0604	0.0593
5	0.0220	0.0220
6	0.0069	0.0069
7	0.0019	0.0019

Using this proposed method, the count probabilities for convolution of two gamma and inverse Gaussian inter-arrival distributions proposed in Section 2.2 can be easily computed. Chaudhry et al. (2013) used the roots method and a Padè approximation method for computing the count probabilities for several inter-arrival times distributions. In Table 4, we compare the probability function of gamma, inverse Gaussian and Weibull count distributions with those obtained by Chaudhry et al. (2013). We note that the difference in the probabilities is at most two decimal places. In the case of Weibull count distribution, we include only the results when  $t = 0.25$ , because the algorithm could not converge for  $t = 0.60$  and  $t = 1$  when  $\lambda = 3$ , which are the other two values included by Chaudhry et al. (2013). Convergence issues with the Weibull renewal function were also discussed by Constantine and Robinson (1997) whereby they developed a convergent damped exponential series by residue calculations of the Laplace transform of the renewal integral equation for the Weibull renewal function when  $\lambda > 1$ .

**Table 4.** Computation of probabilities for (a) gamma, (b) inverse Gaussian, and (c) Weibull count distributions for selected values of  $t$  using (i) proposed method, (ii) method of Chaudhry et al. (2013).

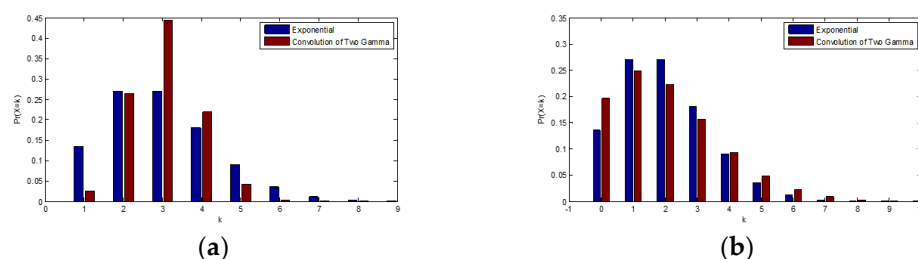
$t$	$Pr(N(t) = 0)$		$Pr(N(t) = 1)$		$Pr(N(t) = 2)$		$Pr(N(t) = 3)$		$Pr(N(t) = 4)$	
	(i)	(ii)	(i)	(ii)	(i)	(ii)	(i)	(ii)	(i)	(ii)
0.1	0.6938	0.6871	0.2341	0.2385	0.0579	0.0602	0.0117	0.0119	0.0021	0.0019
0.4	0.4061	0.4071	0.3092	0.3088	0.1683	0.1677	0.0744	0.0743	0.0283	0.0284
1.25	0.1291	0.1291	0.1952	0.1951	0.2050	0.2050	0.1730	0.1730	0.1249	0.1249
(a) Gamma count distribution										
$t$	$Pr(N(t) = 0)$		$Pr(N(t) = 1)$		$Pr(N(t) = 2)$		$Pr(N(t) = 3)$		$Pr(N(t) = 4)$	
	(i)	(ii)	(i)	(ii)	(i)	(ii)	(i)	(ii)	(i)	(ii)
0.25	0.7394	0.7445	0.2497	0.2442	0.0108	0.0112	0.0001	0.0001	0.0000	0.0000
0.7	0.3377	0.3390	0.4070	0.4042	0.2044	0.2062	0.0460	0.0457	0.0047	0.0046
1.0	0.1623	0.1623	0.2865	0.2869	0.2871	0.2867	0.1763	0.1762	0.0681	0.0683
(b) Inverse Gaussian count distribution										
$t$	$Pr(N(t) = 0)$		$Pr(N(t) = 1)$		$Pr(N(t) = 2)$		$Pr(N(t) = 3)$		$Pr(N(t) = 4)$	
	(i)	(ii)	(i)	(ii)	(i)	(ii)	(i)	(ii)	(i)	(ii)
0.25	0.9845	0.9841	0.0155	0.0159	0.0000	0.0000	0.0000	-	0.0000	-
(c) Weibull count distribution										

We compare the pmf of the two count distributions proposed in Sections 2.1.3 and 2.1.4 with the Poisson distribution. For comparison purposes, the mean for all of the distributions is set to 2, i.e.,  $E(N) = 2$ . Figure 1 compares the probability functions of the inverse Gaussian count distribution with a Poisson distribution.



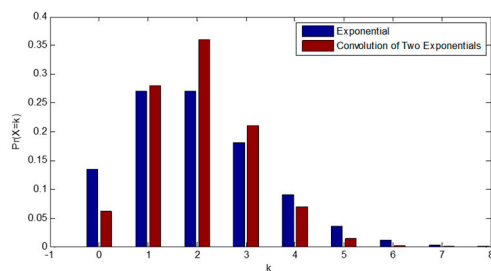
**Figure 1.** Plots of Poisson and inverse Gaussian probabilities: (a)  $\lambda = 0.17, \mu = 1$  (over dispersion); (b)  $\lambda = 1, \mu = 0.438$  (under dispersion).

Figure 2 compares the probability functions of the convolution of two gamma count distribution with a Poisson distribution. The convolution of the two gamma count model can model both over-dispersion and under-dispersion relative to the Poisson distribution.



**Figure 2.** Plots of Poisson and convolution of two gamma probabilities: (a)  $\alpha_1 = 1.5, \alpha_2 = 1.9$  (under dispersion); (b)  $\alpha_1 = 0.2, \alpha_2 = 0.5$  (over dispersion).

The convolution of two gamma distributions nests the special case of convolution of two exponential distributions, that is, when  $\alpha_1 = \alpha_2 = 1$ . This two-component hypo exponential count distribution with parameters  $\beta_1$  and  $\beta_2$  can model under-dispersion and Figure 3 compares its probability function with a Poisson distribution.



**Figure 3.** Plot of Poisson and convolution of two exponentials probabilities:  $\beta_1 = 4.2, \beta_2 = 4.85$  (under dispersion).

### 3.2. Renewal Function and Variance

Using the probability of the counts computed using our proposed method, we also computed the renewal function and variance function for comparison with those obtained by Chaudhry et al. (2013) and Baxter et al. (1982). The details are presented in Table 5. In most cases, the values computed using our proposed method are closer to those of Baxter et al. (1982). We note that Baxter et al. (1982) verified the accuracy of their extended cubic splining algorithm through comparisons with previous tabulations for the Weibull count distribution in the literature (see Baxter et al. 1982 for details) and a direct evaluation of the incomplete gamma integral for the gamma count distribution.

**Table 5.** Computation of renewal and variance functions for (a) gamma, (b) inverse Gaussian, and (c) Weibull count distributions for selected values of  $t$  using (i) proposed method, (ii) method of Baxter et al. (1982), and (iii) method of Chaudhry et al. (2013).

$t$	Renewal Function			Variance Function		
	(i)	(ii)	(iii)	(i)	(ii)	(iii)
0.1	0.3953	0.3933	0.4040	0.4580	0.4485	0.4623
0.4	1.0560	1.0550	1.0545	1.3954	1.3901	1.3970
1.25	2.6662	2.6653	2.6663	4.0491	4.0441	4.0487
(a) Gamma count distribution						
$t$	Renewal Function			Variance Function		
	(i)	(ii)	(iii)	(i)	(ii)	(iii)
0.25	0.2716	0.2715	0.2669	0.2198	0.2200	0.2188
0.7	0.9739	0.9739	0.9736	0.7717	0.7718	0.7732
1.0	1.7636	1.7638	1.7635	1.5290	1.5293	1.5294
(b) Inverse Gaussian count distribution						
$t$	Renewal Function			Variance Function		
	(i)	(ii)	(iii)	(i)	(ii)	(iii)
0.25	0.0155	0.0156	0.0159	0.0153	0.0154	0.0156
(c) Weibull count distribution						

#### 4. Real Data Analysis

Table 6 gives the distribution for the number of monthly non-payments for personal loan in a sample of 2446 clients in a Spanish bank (Dionne et al. 1996). In personal loans, small amounts of money are lent with a relatively short repayment or loan period. The repayment schedule is typically on a monthly basis with a constant amount. The empirical data has a sample mean of 1.109 and variance of 4.860, indicating presence of over dispersion, hence a simple Poisson process may not be sufficient to model the counts. The majority (68.1%) of the counts are zeroes, which correspond to clients who never missed a payment, followed by 11.1% who missed one payment and a cumulative percentage of 11.4% who missed two to four payments. The count distributions are applied to fit this data set. For the simple Poisson count process, observations with expected frequencies which are less than 1.0 are grouped in one class. We also include the log-likelihood function and Akaike information criterion (AIC) values for each fitted model in the tables.

The pmf of the count distributions is evaluated using the numerical inverse Laplace transform method discussed in Section 2.2. The maximum likelihood (ML) estimates of the parameters are obtained with numerical global optimization using the simulated annealing algorithm (Goffe et al. 1994). For numerical stability, we transform the parameters for the generalized Weibull count distributions to their corresponding reciprocals prior to performing ML estimation. The ML estimates are given in Table 7.

The count distribution with generalized Weibull as the distribution for inter-arrival times gives the best fit for the data presented in Table 6. Since the generalized Weibull distribution does not have a closed form Laplace transform, the model fitting takes up a significantly longer time. In the case of distributions with closed Laplace transform, the convolution of two gamma count distribution gives the best fit. We also verify that the convolution of the two exponentials count distribution gives the same fit as the simple Poisson distribution, implying that this distribution is not suitable for over dispersed count data. The inverse Gaussian distribution also gives a poor fit to this data set. This coincides with the characteristic of inter-arrival time distributions, which has an increasing hazard function.

**Table 6.** Number of monthly non-payments for personal loan (Dionne et al. 1996).

Count	Observed	Expected Frequencies						
		Exponential	Gamma	Convolution of Two Exponentials	Convolution of Two Gamma	Inverse Gaussian	Weibull	Generalized Weibull
0	1665	806.78	1159.28	806.78	1159.18	703.13	1156.51	1172.12
1	271	894.85	610.04	894.85	609.94	614.81	607.38	599.05
2	101	496.26	320.92	496.26	320.89	470.06	319.98	309.55
3	73	183.48	168.77	183.48	168.79	314.25	169.15	162.84
4	106	50.88	88.73	50.88	88.78	183.69	89.74	87.75
5	72	11.29	46.64	11.29	46.68	93.88	47.80	48.58
6	43	2.09	24.51	2.09	24.55	41.96	25.56	27.60
7	31	0.38	12.87	0.38	12.90	16.39	13.72	16.00
8	31		6.76		6.78	5.60	7.39	9.39
9	25		3.55		3.56	1.67	4.00	5.53
10	19		1.86		1.87	0.44	2.17	3.25
11	9		0.98		0.98	0.10	1.18	1.89
12 or more	0		1.08		1.09	0.02	1.42	2.44
Total		2446.00	2446.00	2446.00	2446.00	2446.00	2446.00	2446.00
$\chi^2$		37,242.91	1111.77	37,242.91	1108.75	4057.66	1032.59	838.51
Log-likelihood		-4954.79	-3569.93	-4954.79	-3569.49	-4231.06	-3558.13	-3511.39
AIC		9911.57	7143.85	9913.57	7146.99	8466.11	7118.27	7028.77

**Table 7.** ML estimates of the fitted distributions.

Inter-Arrival Distribution	ML Estimates of Parameters
Exponential	$\hat{\lambda} = 1.1092$
Gamma	$\hat{\alpha} = 0.0136, \hat{\beta} = 0.0000$
Convolution of two exponentials	$\hat{\beta}_1 = 1.1092, \hat{\beta}_2 \rightarrow \infty$
Convolution of two gamma	$\hat{\alpha}_1 = 0.0097, \hat{\beta}_1 = 0.0000, \hat{\alpha}_2 = 0.0000, \hat{\beta}_1 = 4.5611$
Inverse Gaussian	$\hat{\lambda} = 0.1358, \hat{\mu} \rightarrow \infty$
Weibull	$\hat{\alpha} = 18.2613, \hat{\lambda} = 3.0684$
Generalized Weibull	$\hat{a} = 40.6405; \hat{\alpha} = 1.0000, \hat{\lambda} = -0.2044$

### 5. Discussion and Conclusions

This article examines the modelling of count data commonly encountered in finance and risk management with count distributions arising from non-exponential inter-arrival time distributions in a renewal process. A specific application example on modelling of loan non-payments is presented. Since the number of non-payments and the lapsed time between payments reflect a lender’s payment behaviour, models which account for these data can assist in the development of further diagnostic techniques such as loan default prediction and tools for early warning detection. Due to the complicated calculations, computation of the probabilities arising from these distributions is investigated and discussed in this paper. The inversion of the Laplace transform is proposed as a generic method of computation, since the transforms have relatively simple forms compared to the probabilities. The proposed method is compared with some existing techniques in the literature.

When the Laplace transform of the inter-arrival time distribution is not available in closed form, other methods to approximate the Laplace transform for numerical inversion can be explored, such as the infinite series, Gaussian quadrature, Laguerre method and the continued fractions technique. This will be considered elsewhere.

**Author Contributions:** Conceptualization, S.-H.O.; methodology, Y.-C.L.; software, Y.-C.L.; validation, S.-H.O.; formal analysis, Y.-C.L.; investigation, S.-H.O. and Y.-C.L.; resources, S.-H.O. and Y.-C.L.; data curation, Y.-C.L.; writing—original draft preparation, Y.-C.L.; writing—review and editing, S.-H.O.; visualization, Y.-C.L.; supervision, S.-H.O.; project administration, Y.-C.L.; funding acquisition, S.-H.O. and Y.-C.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** The authors are supported by the Malaysia Ministry of Higher Education grant FRGS/1/2020/STG06/SYUC/02/1; S.-H.O. is supported by UCSI University grant REIG-FBM-2022/050.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors wish to thank the reviewers for their insightful comments which have greatly improved the paper.

**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.

## References

- Abate, Joseph, and Peter P. Valkó. 2004. Multi-precision Laplace transform inversion. *International Journal for Numerical Methods in Engineering* 60: 979–93. [CrossRef]
- Abate, Joseph, and Ward Whitt. 1992. The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems* 10: 5–87. [CrossRef]
- Abate, Joseph, and Ward Whitt. 1995. Numerical Inversion of Laplace Transforms of Probability Distributions. *ORSA Journal on Computing* 7: 36–43. [CrossRef]
- Abate, Joseph, Gagan L. Choudhury, and Ward Whitt. 2000. An Introduction to Numerical Transform Inversion and Its Application to Probability Models. In *International Series in Operations Research & Management Science*. Boston: Springer, pp. 257–323.
- Baker, Rose, and Tarak Kharrat. 2017. Event count distributions from renewal processes: Fast computation of probabilities. *IMA Journal of Management Mathematics* 29: 415–33. [CrossRef]
- Baxter, Laurence A., Ernest M. Scheuer, Denis J. McConalogue, and Wallace R. Blischke. 1982. On the Tabulation of the Renewal Function. *Technometrics* 24: 151. [CrossRef]
- Chaudhry, Mohan L., Xiaofeng Yang, and Boon Ong. 2013. Computing the Distribution Function of the Number of Renewals. *American Journal of Operations Research* 3: 380–86. [CrossRef]
- Chhikara, Raj S., and J. Leroy Folks. 1977. The Inverse Gaussian Distribution as a Lifetime Model. *Technometrics* 19: 461–68. [CrossRef]
- Constantine, A. Graham, and Neville I. Robinson. 1997. The Weibull renewal function for moderate to large arguments. *Computational Statistics & Data Analysis* 24: 9–27.
- Consul, Prem C., and Gaurav C. Jain. 1973. A Generalization of the Poisson Distribution. *Technometrics* 15: 791–99. [CrossRef]
- Dionne, Georges, Manuel Artis, and Montserrat Guillén. 1996. Count data models for a credit scoring system. *Journal of Empirical Finance* 3: 303–25. [CrossRef]
- Dubner, Harvey, and Joseph Abate. 1968. Numerical inversion of Laplace transforms by relating them to the finite Fourier cosine transform. *Journal of the ACM* 15: 115–23. [CrossRef]
- Erdelyi, Arthur M., Fritz Oberhettinger, and Francesco G. Tricomi. 1953. *Higher Transcendental Functions*. New York: McGraw-Hill.
- From, Steven G. 2004. Approximating the distribution of a renewal process using generalized Poisson distributions. *Journal of Statistical Computation and Simulation* 74: 667–81. [CrossRef]
- Goffe, William L., Gary D. Ferrier, and John Rogers. 1994. Global optimization of statistical functions with simulated annealing. *Journal of Econometrics* 60: 65–99. [CrossRef]
- Greenwood, Major, and G. Udny Yule. 1920. An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents. *Journal of the Royal Statistical Society* 83: 255. [CrossRef]
- Gupta, Ramesh C., and Seng-Huat Ong. 2005. Analysis of Long-Tailed Count Data by Poisson Mixtures. *Communications in Statistics—Theory and Methods* 34: 557–73. [CrossRef]
- Holla, M. S. 1967. On a Poisson-inverse Gaussian distribution. *Metrika* 11: 115–21. [CrossRef]
- Johnson, Norman L., Adrienne W. Kemp, and Samuel Kotz. 2005. *Univariate Discrete Distributions*, 3rd ed. New York: John Wiley and Sons.
- Jose, K. Kanichukattu, and Bindu Abraham. 2011. A Count Model Based on Mittag-Leffler Interarrival Times. *Statistica* LXXI: 501–14.
- Jose, K. Kanichukattu, and Bindu Abraham. 2013. A Counting Process with Gumbel Inter-arrival Times for Modeling Climate Data. *Journal of Environmental Statistics* 4: 13.
- Karlis, Dimitris, and Mohieddine Rahmouni. 2007. Analysis of defaulters' behaviour using the Poisson-mixture approach. *IMA Journal of Management Mathematics* 18: 297–311. [CrossRef]
- Lindholm, Mathias, and Henning Zakrisson. 2022. A Collective Reserving Model with Claim Openness. *ASTIN Bulletin: The Journal of the IAA* 52: 117–43. [CrossRef]
- Maciak, Matus, Ostap Okhrin, and Michal Pesta. 2021. Infinitely Stochastic Micro Reserving. *Insurance: Mathematics and Economics* 100: 30–58. [CrossRef]

- McShane, Blake, Moshe Adrian, Eric T. Bradlow, and Peter S. Fader. 2008. Count Models Based on Weibull Interarrival Times. *Journal of Business & Economic Statistics* 26: 369–78.
- Mestiri, Sami, and Abdeljelil Farhat. 2021. Using Non-parametric Count Model for Credit Scoring. *Journal of Quantitative Economics* 19: 39–49. [CrossRef]
- Moschopoulos, Peter G. 1985. The distribution of the sum of independent gamma random variables. *Annals of the Institute of Statistical Mathematics* 37: 541–44. [CrossRef]
- Mudholkar, Govind S., Deo Kumar Srivastava, and Georgia D. Kollia. 1996. A Generalization of the Weibull Distribution with Application to the Analysis of Survival Data. *Journal of the American Statistical Association* 91: 1575–83. [CrossRef]
- Nadarajah, Saralees, and Stephen Chan. 2018. Discrete distributions based on inter arrival times with application to football data. *Communications in Statistics–Theory and Methods* 47: 147–65. [CrossRef]
- Ong, Seng-Huat, Atanu Biswas, Shelton Peiris, and Yeh-Ching Low. 2015. Count distribution for generalized Weibull duration with applications. *Communications in Statistics–Theory and Methods* 44: 4203–16. [CrossRef]
- Sankaran, Munuswamy. 1968. Mixtures by the Inverse Gaussian Distribution. *Sankhyā: The Indian Journal of Statistics, Series B* 30: 455–58.
- Seshadri, Vanamamalai. 1999. *The Inverse Gaussian Distribution*. Lecture Notes in Statistics. New York: Springer.
- Smith, W. L., and M. Ross Leadbetter. 1963. On the Renewal Function for the Weibull Distribution. *Technometrics* 5: 393–96. [CrossRef]
- Thomas, Lyn C., Anna Matuszyk, Mee Chi So, Christophe Mues, and Angela Moore. 2016. Modelling repayment patterns in the collections process for unsecured consumer debt: A case study. *European Journal of Operational Research* 249: 476–86. [CrossRef]
- Winkelmann, Rainer. 1995. Duration Dependence and Dispersion in Count-Data Models. *Journal of Business & Economic Statistics* 13: 467.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

# On the Contaminated Weighted Exponential Distribution: Applications to Modeling Insurance Claim Data

Abbas Mahdavi <sup>1</sup>, Omid Kharazmi <sup>1</sup> and Javier E. Contreras-Reyes <sup>2,\*</sup>

<sup>1</sup> Department of Statistics, Vali-e-Asr University of Rafsanjan, Rafsanjan 7718897111, Iran

<sup>2</sup> Instituto de Estadística, Facultad de Ciencias, Universidad de Valparaíso, Valparaíso 2360102, Chile

\* Correspondence: jecontrr@uc.cl; Tel.: +56(32)-250-8242

**Abstract:** Deriving loss distribution from insurance data is a challenging task, as loss distribution is strongly skewed with heavy tails with some levels of outliers. This paper extends the weighted exponential (WE) family to the contaminated WE (CWE) family, which offers many flexible features, including bimodality and a wide range of skewness and kurtosis. We adopt Expectation-Maximization (EM) and Bayesian approaches to estimate the model, providing the likelihood and the priors for all unknown parameters. Finally, two sets of claims data are analyzed to illustrate the efficiency of the proposed method in detecting outliers.

**Keywords:** bayesian estimation; EM algorithm; Gibbs sampler; Mixture model; insurance claim data



**Citation:** Mahdavi, Abbas, Omid Kharazmi, and Javier E. Contreras-Reyes. 2022. On the Contaminated Weighted Exponential Distribution: Applications to Modeling Insurance Claim Data. *Journal of Risk and Financial Management* 15: 500. <https://doi.org/10.3390/jrfm15110500>

Academic Editors: Shuangzhe Liu, Tiefeng Ma and Seng Huat Ong

Received: 21 September 2022

Accepted: 25 October 2022

Published: 27 October 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In many applied areas, particularly in finance and actuarial sciences, data are usually positive, right-skewed, leptokurtic and multimodal (Cummins et al. 1990). To capture a wide range of population heterogeneity and tail behavior, one practical way is to conduct analyses over subsets of claims with distinct claim characteristics. But the approach falls short of providing a full picture of claim dynamics. Classical distributions are not flexible enough to cater to heavy-tailed datasets due to extreme values that are far from the other observed data points. These unusual observations are usually called outliers. The presence of outliers in the data may distort both the estimated model parameters and the model's goodness-of-fit. Recently, many authors have focused on a finite mixture approach that shares the efficiency of parametric modeling and the flexibility of non-parametric density estimation techniques. The flexibility of finite mixtures is accommodating various shapes of insurance and economic data (Bernardi et al. 2012; Hennig and Liao 2013; Maruotti et al. 2016; Punzo et al. 2018).

Okhli and Nooghabi (2021) introduced the contaminated exponential (CE) distribution as an alternative platform for analyzing positive-valued insurance datasets with some level of outliers. The pdf of CE distribution with scale parameter  $\lambda$  and contamination factor  $\theta$  is defined as follows:

$$f_{CE}(y; \lambda, \theta, \omega) = (1 - \omega)\lambda e^{-\lambda y} + \omega\lambda\theta e^{-\lambda\theta y}, \quad y > 0, \lambda > 0, \quad (1)$$

where  $\omega \in (0, 1)$  is the proportion of contaminated points. The Bayesian approach is developed for computing the parameter estimates. It is demonstrated that the effect of outliers is automatically reflected in the posterior distribution for any sample size. This way, an outlier observation has the highest posterior probability of outlying, but the main observations have a relatively small such probability, indicating that the CE model can detect outliers well.

Weighted distributions are used to adjust the probabilities of events as observed and recorded (Chung and Kim 2004; Gupta and Kirmani 1990; Larose and Dey 1996); (Navarro et al. 2006). Patil (1991) proceeded from applications involving statistical ecology to generate and review many useful general results concerning weighted distributions. Mild outliers, on which this paper focuses, can be dealt with by using heavy-tailed distributions



for data. Weighted distributions offer the flexibility needed for achieving mild outlier robustness, while the usual distributions like exponential, gamma and Weibull models lack sufficient fit. For more information and applications of weighted distributions see Patil and Rao (1977).

A two-parameter weighted exponential (WE) distribution (Gupta and Kundu 2009) was developed as a lifetime model which has been widely used in engineering, medicine and insurance. The sensitive skewness parameter governs essentially the shape of the probability density function (pdf) of the WE distribution. A random variable  $Y$  is said to have a weighted exponential distribution with a shape parameter  $\alpha > 0$ , and scale parameter  $\lambda > 0$ , denoted by  $WE(\alpha, \lambda)$ , if its pdf is given by

$$f_{WE}(y; \alpha, \lambda) = \left(1 + \frac{1}{\alpha}\right) \lambda e^{-\lambda y} (1 - e^{-\alpha \lambda y}), \quad y > 0. \tag{2}$$

In this paper, we introduce a class of contaminated weighted exponential (CWE) distributions to account for all possible features of insurance and economic data. Crucially, the CWE model is a two-component mixture in which one component, with a large prior probability, represents the reference distribution, and another, with small prior probability and inflated variability, represents the degree of contamination. For Bayesian inference, we consider several asymmetric and symmetric loss functions like squared error loss, modified squared error, precautionary, weighted squared error, linear exponential, general entropy, and  $K$ -loss functions to estimate the parameters of the CWE model. Further, using the independent prior distributions, Bayesian 95% credible and highest posterior density (HPD) intervals (see Chen et al. 1999) are provided for each parameter of the proposed model.

The paper is organized as follows. Section 2 presents the CWE model and some illustrations of the density, skewness and kurtosis. In Sections 3 and 4, the EM algorithm and Bayesian inference are respectively developed for CWE parameters. Section 5 illustrates several simulations of proposed estimation methods of Sections 3 and 4. Sections 6 and 7 illustrates numerical examples for insurance data fitting using proposed estimation methods of Sections 3 and 4, respectively. Finally, discussions and conclusions are presented in Section 8.

## 2. The CWE Model

The pdf of a CWE model with contamination factor  $\theta$  can be written as

$$f_{CWE}(y; \alpha, \lambda, \theta, \omega) = (1 - \omega) f_{WE}(y; \alpha, \lambda) + \omega f_{WE}(y; \alpha, \lambda \theta), \tag{3}$$

where  $\theta > 0$  and  $\omega \in [0, 1]$  denotes the proportion of outliers or unusual points and  $\Theta = (\omega, \alpha, \lambda, \theta)^\top$  contains all model parameters. The CE model given in (1) is obtained as a special case of (3) when  $\alpha \rightarrow \infty$ . The effect of varying each parameter when one varies, but keeping others fixed, is illustrated by a set of CWE densities shown in Figure 1. The plots show that the distribution is more likely to be bimodal as  $\omega$  increases, whereas flatness parameter vector  $\alpha$  controls tail behavior. This implies that the CWE model provides a component of the WE distribution to capture the vast majority of small losses, whereas the contaminated component accommodates clusters of larger losses with an enhanced tail to capture extreme losses. Furthermore, the skewness and kurtosis 3D plots of the CWE model for numerous values of  $\alpha$  and  $\theta$  with fixed  $\lambda = 1$  are depicted in the Figure 2. The fitting of this four-parameter CWE model via the likelihood approach is difficult because of the log-likelihood function's complexity. But the EM and Bayesian approaches can help.

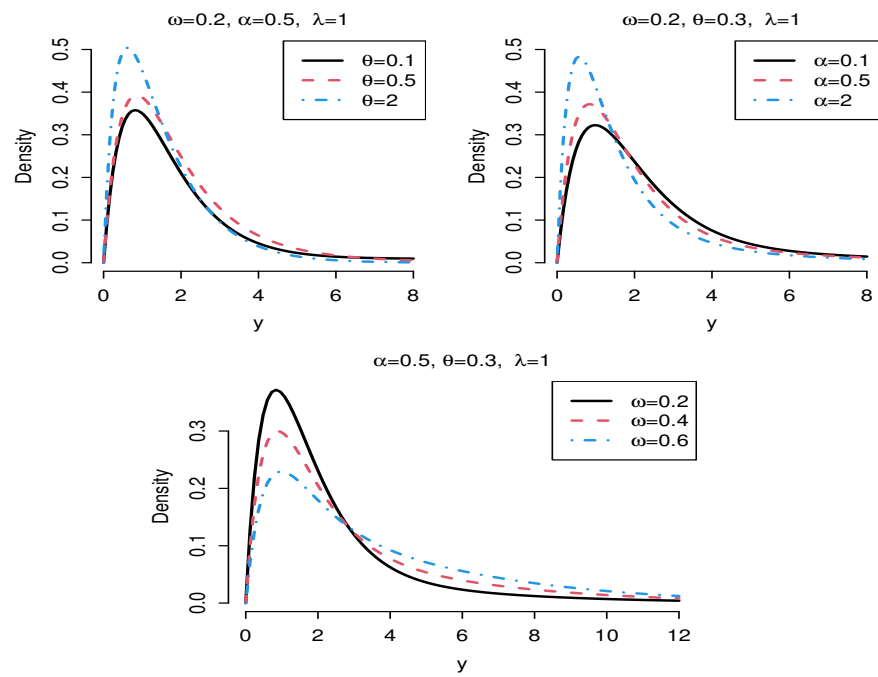


Figure 1. Density plots for different CWE distributions.

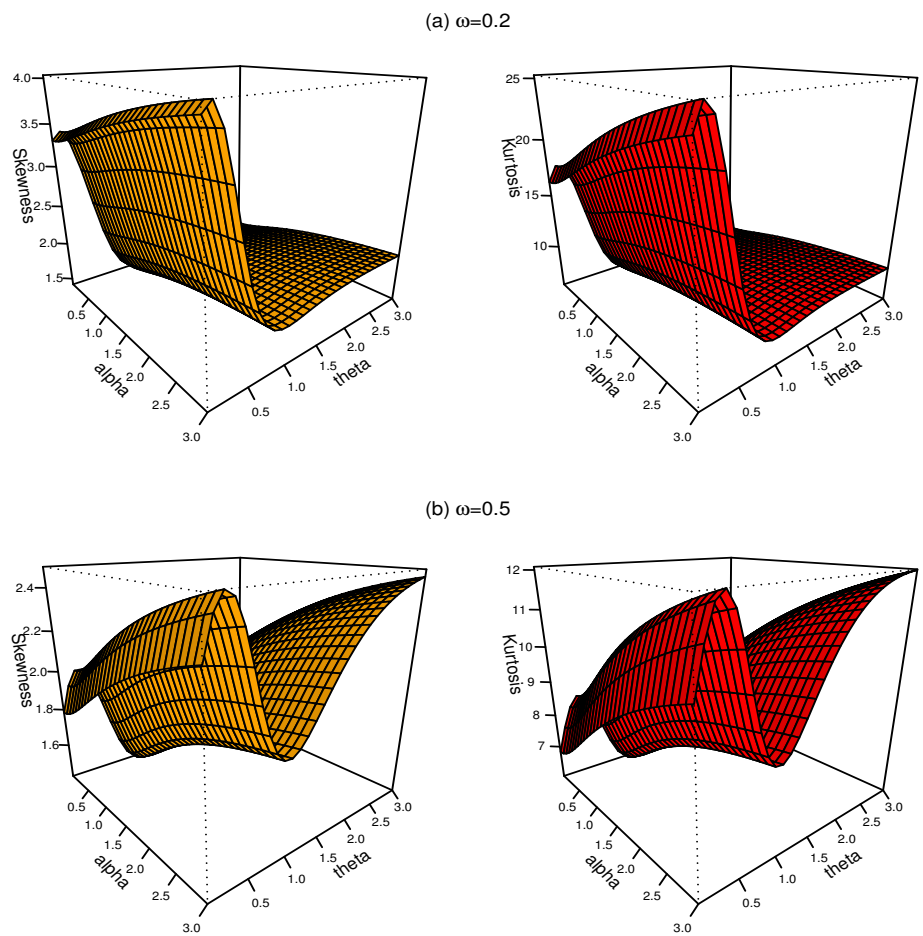


Figure 2. 3D plots of skewness and kurtosis of CWE distribution for two fixed values of  $\omega$ .

### 3. Maximum Likelihood Estimation via EM Algorithm

The EM algorithm (Dempster et al. 1977) and some of its extraordinary variants such as the expectation conditional maximization (ECM) algorithm (Meng and Rubin 1993) and the expectation-conditional maximization either (ECME) algorithm (Liu and Rubin 1994) are broadly applicable methods to carry out ML estimation for mixture distributions and variety of incomplete-data problems (Aitkin and Wilson 1980; McLachlan and Krishnan 2007; Redner and Walker 1984). Mahdavi et al. (2021a, 2021b) and Cavieres et al. (2022) developed novel EM-based procedures designed under the selection mechanism to compute the ML estimates of scale-shape mixtures of flexible generalized skew-normal and multivariate flexible skew-symmetric-normal distributions. Here, we develop a novel EM-based procedure designed under the selection mechanism to compute the ML estimates of the proposed model.

A random variable  $Y \sim WE(\alpha, \lambda)$  is said to follow WE distribution with shape parameter  $\alpha$  and scale parameter  $\lambda$  if it has the following stochastic selection representation:

$$Y \stackrel{d}{=} X_0|U < 1, \tag{4}$$

where  $U = X_1/(\alpha X_0)$  and  $X_0$  and  $X_1$  are two independent exponential random variables with mean  $1/\lambda$ . To perform an EM-type algorithm for fitting the CWE model, we introduce a latent variable  $\tau = U|U < 1$  based on (4). The joint pdf of  $(Y, \tau)^\top$  is given by

$$\begin{aligned} f_{Y,\tau}(y, \tau) &= \frac{1}{\mathbb{P}(U < 1)} f_{X_0,U}(y, \tau) = \left(1 + \frac{1}{\alpha}\right) f_{X_0}(y) f_{U|X_0}(\tau) \\ &= (\alpha + 1)\lambda^2 y e^{-\lambda y} e^{-\lambda \alpha \tau y}, \quad y > 0, \quad 0 < \tau < 1. \end{aligned} \tag{5}$$

Dividing (5) by (2) yields

$$f_{\tau|Y}(\tau) = \frac{\alpha \lambda y e^{-\alpha \lambda y \tau}}{1 - e^{-\alpha \lambda y}}, \quad 0 < \tau < 1. \tag{6}$$

Using (6), it is clear that

$$\tau|Y = y \sim TExp(\alpha \lambda y; (0, 1)), \tag{7}$$

where  $TExp(\lambda; (0, b))$  represents the truncated exponential distribution with mean  $1/\lambda$  on interval  $(0, b)$ .

Let us introduce an  $n$ -dimensional binary random variable  $\gamma = (\gamma_1, \dots, \gamma_n)^\top$  where a particular element  $\gamma_i$  is equal to 1 if  $Y_i$  belongs to unusual observations and is equal to zero otherwise. Note that,  $\gamma_i$  follows a Bernoulli random variable with success probability  $\omega$  denoted by  $\gamma_i \sim Ber(\omega)$ .

Now, consider  $n$  independent random variables  $Y_1, \dots, Y_n$ , which are taken from a mixture model (3) and latent variable  $\tau = (\tau_1, \dots, \tau_n)^\top$ , where  $\Theta = (\omega, \alpha, \lambda, \theta)^\top$  denotes the unknown vector of parameters. Clearly,

$$\begin{aligned} Y_i | (\gamma_i = 0) &\sim WE(\alpha, \lambda) & \text{and} & & Y_i | (\gamma_i = 1) &\sim WE(\alpha, \lambda \theta), \\ \tau_i | (Y_i = y_i, \gamma_i = 0) &\sim TExp(\alpha \lambda y_i; (0, 1)), \\ \tau_i | (Y_i = y_i, \gamma_i = 1) &\sim TExp(\alpha \lambda \theta y_i; (0, 1)). \end{aligned}$$

According to (3) and (5), it is clear that

$$f_{Y_i, \tau_i | \gamma_i}(y_i, \tau_i) = \{(\alpha + 1)\lambda^2 y_i e^{-\lambda y_i} e_i^{-\lambda \alpha \tau_i y_i}\}^{1-\gamma_i} \{(\alpha + 1)\lambda^2 \theta^2 y_i e^{-\lambda \theta y_i} e^{-\lambda \theta \alpha \tau_i y_i}\}^{\gamma_i}.$$

The complete log-likelihood function of  $\Theta$  given  $\mathbf{y}_c = (\mathbf{y}^\top, \boldsymbol{\tau}^\top, \dots, \boldsymbol{\gamma}^\top)^\top$  is

$$\begin{aligned} \ell_c(\Theta|\mathbf{y}_c) &= \ln \{f_{\mathbf{Y}, \boldsymbol{\gamma}, \boldsymbol{\tau}}(\mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\tau})\} = \ln \{f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma})f_{\mathbf{Y}, \boldsymbol{\tau}|\boldsymbol{\gamma}}(\mathbf{y}, \boldsymbol{\tau})\} \\ &= \sum_{i=1}^n \left\{ \gamma_i \ln \omega + (1 - \gamma_i) \ln(1 - \omega) + \ln(\alpha + 1) + 2 \ln \lambda + 2\gamma_i \ln \theta \right. \\ &\quad \left. - (1 - \gamma_i)\lambda y_i - (1 - \gamma_i)\lambda \alpha \tau_i y_i - \gamma_i \lambda \theta y_i - \gamma_i \lambda \theta \alpha \tau_i y_i \right\}. \end{aligned} \tag{8}$$

To evaluate the Q-function, the necessary conditional expectations include

$$\begin{aligned} \hat{\gamma}_i^{(k)} &= E(\gamma_i | Y_i = y_i, \hat{\Theta}^{(k)}) = \frac{\hat{\omega}^{(k)} f_{WE}(y_i; \hat{\alpha}^{(k)}, \hat{\lambda}^{(k)} \hat{\Theta}^{(k)})}{f_{CWE}(y_i; \hat{\alpha}^{(k)}, \hat{\lambda}^{(k)}, \hat{\Theta}^{(k)})}, \\ \hat{\tau}_{1i}^{(k)} &= E((1 - \gamma_i)\tau_i | Y_i = y_i, \hat{\Theta}^{(k)}) = (1 - \hat{\gamma}_i^{(k)}) \left( \frac{1}{\hat{\alpha}^{(k)} \hat{\lambda}^{(k)} y_i} - \frac{1}{e^{\hat{\alpha}^{(k)} \hat{\lambda}^{(k)} y_i} - 1} \right), \\ \hat{\tau}_{2i}^{(k)} &= E(\gamma_i \tau_i | Y_i = y_i, \hat{\Theta}^{(k)}) = \hat{\gamma}_i^{(k)} \left( \frac{1}{\hat{\alpha}^{(k)} \hat{\lambda}^{(k)} \hat{\theta}^{(k)} y_i} - \frac{1}{e^{\hat{\alpha}^{(k)} \hat{\lambda}^{(k)} \hat{\theta}^{(k)} y_i} - 1} \right). \end{aligned}$$

Therefore, the Q-function is given by

$$\begin{aligned} Q(\Theta|\hat{\Theta}^{(k)}) &= \sum_{i=1}^n \left\{ \hat{\gamma}_i^{(k)} \ln \omega + (1 - \hat{\gamma}_i^{(k)}) \ln(1 - \omega) + \ln(\hat{\alpha}^{(k)} + 1) + 2 \ln \hat{\lambda}^{(k)} \right. \\ &\quad \left. + 2\hat{\gamma}_i^{(k)} \ln \hat{\theta}^{(k)} - \hat{\lambda}^{(k)} (1 - \hat{\gamma}_i^{(k)}) y_i - \hat{\lambda}^{(k)} \hat{\alpha}^{(k)} \hat{\tau}_{1i}^{(k)} y_i \right. \\ &\quad \left. - \hat{\lambda}^{(k)} \hat{\theta}^{(k)} \hat{\gamma}_i^{(k)} y_i - \hat{\lambda}^{(k)} \hat{\theta}^{(k)} \hat{\alpha}^{(k)} \hat{\tau}_{2i}^{(k)} y_i \right\}. \end{aligned} \tag{9}$$

In summary, the implementation of the ECM algorithm proceeds as follows:

**E-step:** Given  $\Theta = \hat{\Theta}^{(k)}$ , compute  $\hat{\gamma}_i^{(k)}$ ,  $\hat{\tau}_{1i}^{(k)}$  and  $\hat{\tau}_{2i}^{(k)}$  for  $i = 1, \dots, n$ .

**CM-step 1:** Calculate

$$\hat{\omega}^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \hat{\gamma}_i^{(k)}.$$

**CM-step 2:** Fix  $\lambda = \hat{\lambda}^{(k)}$ ,  $\theta = \hat{\theta}^{(k)}$  and update  $\hat{\alpha}^{(k)}$  by maximizing (9) over  $\alpha$ , which gives

$$\hat{\alpha}^{(k+1)} = \frac{n}{\hat{\lambda}^{(k)} \sum_{i=1}^n (\hat{\tau}_{1i}^{(k)} y_i - \hat{\theta}^{(k)} \hat{\tau}_{2i}^{(k)} y_i)} - 1.$$

**CM-step 3:** Fix  $\alpha = \hat{\alpha}^{(k+1)}$ ,  $\theta = \hat{\theta}^{(k)}$  and update  $\hat{\lambda}^{(k)}$  by

$$\hat{\lambda}^{(k+1)} = \frac{2n}{\sum_{i=1}^n \{ (1 - \hat{\gamma}_i^{(k)}) y_i + \hat{\alpha}^{(k+1)} \hat{\tau}_{1i}^{(k)} y_i + \hat{\theta}^{(k)} \hat{\gamma}_i^{(k)} y_i + \hat{\alpha}^{(k+1)} \hat{\theta}^{(k)} \hat{\tau}_{2i}^{(k)} y_i \}}.$$

**CM-step 4:** Fix  $\alpha = \hat{\alpha}^{(k+1)}$ ,  $\lambda = \hat{\lambda}^{(k+1)}$  and update  $\hat{\theta}^{(k)}$  by

$$\hat{\theta}^{(k+1)} = \frac{2 \sum_{i=1}^n \hat{\gamma}_i^{(k)}}{\hat{\lambda}^{(k+1)} \sum_{i=1}^n \{ \hat{\gamma}_i^{(k)} y_i + \hat{\alpha}^{(k+1)} \hat{\tau}_{2i}^{(k)} y_i \}}.$$

This process is repeated until a suitable convergence rule is satisfied. The convergence appears when the relative difference between two successive log-likelihood values is less than tolerance ( $\epsilon$ ). In our numerical experiments,  $\epsilon = 10^{-6}$  is used. An R code about EM algorithm is available in Appendix A.

### 4. Bayesian Inference

In this section, we discuss the Bayesian estimation for the CWE distribution parameters in terms of several symmetric and asymmetric loss functions such as squared error loss function (SELF), weighted squared error loss function (WSELF), modified squared error loss function (MSELF), precautionary loss function (PLF) and K-loss function (KLF). The considered loss functions and their Bayesian estimators with corresponding posterior risks are reported in Table 1.

**Table 1.** Bayes estimator and posterior risk under several loss functions.

Loss Function $L(\psi, \delta)$	Bayes Estimator $\psi_B$	Posterior Risk $\rho_\psi$
$SELF = (\psi - d)^2$	$E(\psi x)$	$Var(\psi x)$
$WSELF = \frac{(\psi-d)^2}{\psi}$	$(E(\psi^{-1} x))^{-1}$	$E(\psi x) - (E(\psi^{-1} x))^{-1}$
$MSELF = \left(1 - \frac{d}{\psi}\right)^2$	$\frac{E(\psi^{-1} x)}{E(\psi^{-2} x)}$	$1 - \frac{E(\psi^{-1} x)^2}{E(\psi^{-2} x)}$
$PLF = \frac{(\psi-d)^2}{d}$	$\sqrt{E(\psi^2 x)}$	$2\left(\sqrt{E(\psi^2 x)} - E(\psi x)\right)$
$KLF = \left(\sqrt{\frac{d}{\psi}} - \sqrt{\frac{\psi}{d}}\right)^2$	$\sqrt{\frac{E(\psi x)}{E(\psi^{-1} x)}}$	$2\left(\sqrt{E(\psi x)E(\psi^{-1} x)} - 1\right)$

For pertinent details about these loss functions, refer to Kharazmi et al. (2021, 2022) and references therein.

#### 4.1. Joint and Marginal Posterior Distributions

Assume that the parameters of the CWE distribution have independent prior distributions as follows:  $\alpha \sim Gamma(\alpha_0, \alpha_1)$ ,  $\theta \sim Gamma(\theta_0, \theta_1)$ ,  $\lambda \sim Gamma(\lambda_0, \lambda_1)$ , and  $\omega \sim Beta(\omega_0, \omega_1)$ , where all hyper-parameters are positive. Consequently, the joint prior density is formulated as

$$\pi(\alpha, \lambda, \theta, \omega) = \frac{\omega^{\omega_0}(1-\omega)^{\omega_1}\alpha_1^{\alpha_0}\theta_1^{\theta_0}\lambda_1^{\lambda_0}}{Beta(\omega_0, \omega_1)\Gamma(\alpha_0)\Gamma(\theta_0)\Gamma(\lambda_0)}\alpha^{\alpha_0-1}\theta^{\theta_0-1}\lambda^{\lambda_0}e^{-(\alpha_1\alpha+\theta_1\theta+\lambda_1\lambda)}.$$

For simplicity, we define function  $\zeta$  as

$$\zeta(\alpha, \theta, \lambda, \omega) = \alpha^{\alpha_0-1}\beta^{\beta_0-1}\lambda^{\lambda_0}e^{-(\alpha_1\alpha+\beta_1\beta+\lambda_1\lambda)}\omega^{\omega_0}(1-\omega)^{\omega_1}.$$

From (10) and likelihood function  $L(data)$ , the joint posterior distribution is

$$\pi^*(\alpha, \theta, \lambda, \omega|data) \propto \pi(\alpha, \theta, \lambda, \omega) L(data).$$

Therefore, the exact joint posterior pdf is given by

$$\pi^*(\alpha, \theta, \lambda, \omega|x) = K\zeta(\alpha, \theta, \lambda, \omega) L(x, \Psi), \tag{10}$$

where

$$L(x; \Psi) = \left[\lambda\left(1 + \frac{1}{\alpha}\right)\right]^n \prod_{i=1}^n \left\{ (1-\omega)e^{-\lambda x_i}(1 - e^{-\alpha\lambda x_i}) + \omega\theta e^{-\theta\lambda x_i}(1 - e^{-\alpha\theta\lambda x_i}) \right\}, \tag{11}$$

$\Psi = (\alpha, \theta, \lambda, \omega)$  and  $K$  is a normalizing constant with form

$$K^{-1} = \int_0^1 \int_0^\infty \int_0^\infty \int_0^\infty \zeta(\alpha, \theta, \lambda, \omega) L(x, \xi) d\alpha d\theta d\lambda d\omega.$$

Moreover, the marginal posterior density of  $\alpha, \theta, \lambda$  and  $\omega$  (assuming  $\Psi = (\Psi_1, \Psi_2, \Psi_3, \Psi_4) = (\alpha, \theta, \lambda, \omega)$ ) can be expressed as

$$\pi(\Psi_i|\underline{x}) = \begin{cases} \int_0^1 \int_0^\infty \int_0^\infty \pi^*(\Psi|\underline{x}) \partial\Psi_j \partial\Psi_k \partial\Psi_4, & i = 1, 2, 3, \\ \int_0^\infty \int_0^\infty \int_0^\infty \pi^*(\Psi|\underline{x}) \partial\Psi_1 \partial\Psi_2 \partial\Psi_3, & i = 4, \end{cases} \tag{12}$$

where  $j, k = 1, 2, 3, j \neq k \neq i$  and  $\Psi_i$  is the  $i$ th member of vector  $\Psi$ .

#### 4.2. Bayesian Point Estimation

From the marginal posterior pdf in (12) and under framework of the loss functions listed in Table 1, the Bayesian point estimation for parameter vector  $\Psi = (\Psi_1, \Psi_2, \Psi_3, \Psi_4) = (\alpha, \theta, \lambda, \omega)$  is formulated via minimizing the expectation of loss function with respect to the marginal posterior pdf in (12) as follows:

$$\operatorname{argmin} C_\delta \int_0^\infty L(\Psi_i, \delta) \pi(\Psi_i|\underline{x}) \partial\Psi_i. \tag{13}$$

In practice, because of the intractable integral in (13), we can use the Gibbs sampler (Geman and Geman 1984) or Metropolis-Hastings algorithms (Hastings 1970; Metropolis et al. 1953) to generate posterior samples. We will argue this issue more precisely in Section 4.5.

#### 4.3. Credibility Interval

In the Bayesian framework, interval estimation is done via credibility interval conception. Consider parameter vector  $\Psi = (\Psi_1, \Psi_2, \Psi_3, \Psi_4) = (\alpha, \theta, \lambda, \omega)$ , which is associated with CWE distribution and  $\pi(\Psi_j|\underline{x})$  the marginal posterior pdf of parameter  $\Psi_j, j = 1, 2, 3, 4$ , as in (12). For a given value of  $\eta \in (0, 1)$ , the  $(1 - \eta)100\%$  credibility interval  $CI(L_{\Psi_j}, U_{\Psi_j})$  is defined as

$$\int_{L_{\Psi_j}}^\infty \pi(\Psi_j|\underline{x}) \partial\Psi_j = 1 - \frac{\eta}{2}, \tag{14}$$

$$\int_{U_{\Psi_j}}^\infty \pi(\Psi_j|\underline{x}) \partial\Psi_j = \frac{\eta}{2}. \tag{15}$$

By considering relation (14) and (15), it is not feasible to obtain the explicit marginal pdf from the joint posterior distribution. To overcome this difficulty, we use the Gibbs sampler algorithm and generate posterior samples from the CWE distribution. Let  $\Psi^1, \dots, \Psi^k$  (where  $\Psi^i = (\Psi_1^i, \Psi_2^i, \Psi_3^i, \Psi_4^i)$ ) be a posterior random sample of size  $k$  which is extracted from the joint posterior pdf in (10). Using these samples, the marginal posterior pdf of  $\Psi_j$  given  $\underline{x}$  is defined by

$$\frac{1}{K} \sum_{i=1}^K \pi^*(\Psi_j, \Psi_{-j}^i|\underline{x}), \quad j = 1, 2, 3, 4, \tag{16}$$

where  $\Psi_{-j}^i$  represents the vector of posterior samples when the  $j$ th component is removed. Inserting (16) in (15), it is possible to compute the credibility intervals for  $\Psi_j, j = 1, 2, 3, 4$ , as follows

$$\frac{1}{K} \sum_{i=1}^K \int_{L_{\Psi_j}}^\infty \pi^*(\Psi_j, \Psi_{-j}^i|\underline{x}) \partial\Psi_j = 1 - \frac{\eta}{2}, \tag{17}$$

$$\frac{1}{K} \sum_{i=1}^K \int_{U_{\Psi_j}}^\infty \pi^*(\Psi_j, \Psi_{-j}^i|\underline{x}) \partial\Psi_j = \frac{\eta}{2}. \tag{18}$$

#### 4.4. Highest Posterior Density Interval

Highest posterior density (HPD) interval is a credibility interval under a specific restriction. A  $(1 - \eta)100\%$  HPD interval for  $\Psi_j, j = 1, 2, 3, 4$  is the simultaneous solution of integral equations

$$\frac{1}{K} \sum_{i=1}^K \int_{L_{\Psi_j}}^{U_{\Psi_j}} \pi^*(\Psi_j, \Psi_{-j}^i | \underline{x}) \partial \Psi_j = 1 - \eta, \tag{19}$$

$$\sum_{i=1}^K \pi^*(L_{\Psi_j}, \Psi_{-j}^i | \underline{x}) = \sum_{i=1}^K \pi^*(U_{\Psi_j}, \Psi_{-j}^i | \underline{x}). \tag{20}$$

#### 4.5. Generating Posterior Samples

It is clear from Equations (10) and (12) that there are no explicit expressions for the Bayesian point estimators under the loss functions in Table 1. Because of intractable integrals associated with joint posterior and marginal posterior distributions, we require numerical software to solve the integral equations numerically via MCMC methods such as the Metropolis-Hastings algorithm and Gibbs sampling (Contreras-Reyes et al. 2018). Assuming general model  $f(\underline{x} | \boldsymbol{\psi})$  is associated with parameter vector  $\boldsymbol{\psi} = (\psi_1, \psi_2, \dots, \psi_p)$  and observed data  $\underline{x}$ , the joint posterior distribution is  $\pi(\psi_1, \psi_2, \dots, \psi_p | \underline{x})$ . We also assume that  $\boldsymbol{\psi}_0 = (\psi_1^{(0)}, \psi_2^{(0)}, \dots, \psi_p^{(0)})$  is the initial vector to start the Gibbs sampler (Quintero et al. 2017). The steps for any iteration, say iteration  $k$ , are as follows:

- Starting with an initial estimate  $(\psi_1^{(0)}, \psi_2^{(0)}, \dots, \psi_p^{(0)})$ ;
- draw  $\psi_1^k$  from  $\pi(\psi_1 | \psi_2^{k-1}, \psi_3^{k-1}, \dots, \psi_p^{k-1}, \underline{x})$ ;
- draw  $\psi_2^k$  from  $\pi(\psi_2 | \psi_1^k, \psi_3^{k-1}, \dots, \psi_p^{k-1}, \underline{x})$ ; and so on down to
- draw  $\psi_p^k$  from  $\pi(\psi_p | \psi_1^k, \psi_2^k, \dots, \psi_{p-1}^k, \underline{x})$ .

In the case of the CWE distribution, by considering parameter vector  $\Psi = (\alpha, \theta, \lambda, \omega)$  and initial parameter vector  $\Psi_0 = (\alpha^0, \theta^0, \lambda^0, \omega^0)$ , the posterior samples are extracted based on Gibbs sampler where the full conditional distributions are

$$\pi(\alpha | \theta^{k-1}, \lambda^{k-1}, \omega^{k-1}, \underline{x}) \propto \left(\frac{\alpha + 1}{\alpha}\right)^n \alpha^{\alpha_0} e^{-\alpha_1 \alpha} \prod_{i=1}^n Y(x_i, \Psi), \tag{21}$$

$$\pi(\theta | \alpha^{k-1}, \lambda^{k-1}, \omega^{k-1}, \underline{x}) \propto \beta^{\theta_0} e^{-\theta_1 \theta} \prod_{i=1}^n Y(x_i, \Psi), \tag{22}$$

$$\pi(\lambda | \alpha^{k-1}, \theta^{k-1}, \omega^{k-1}, \underline{x}) \propto \lambda^{\lambda_0 + n} e^{-\lambda_1 \lambda} \prod_{i=1}^n Y(x_i, \Psi), \tag{23}$$

and

$$\pi(\omega | \alpha^{k-1}, \theta^{k-1}, \lambda^{k-1}, \underline{x}) \propto \omega^{\omega_0} (1 - \omega)^{\omega_1} \prod_{i=1}^n Y(x_i, \Psi), \tag{24}$$

where  $Y(x_i, \Psi) = (1 - \omega)e^{-\lambda x_i}(1 - e^{-\alpha \lambda x_i}) + \omega \theta e^{-\theta \lambda x_i}(1 - e^{-\alpha \theta \lambda x_i})$ .

In practice, simulations related to Gibbs sampling can be done with special software WinBUGS. This software was developed in 1997 to simulate data of complex posterior distributions, where analytical or numerical integration techniques cannot be applied. Moreover, Gibbs sampling processes can be carried out via OpenBUGS software, which is an open source version of WinBUGS. Since there isn't any prior information about hyper-parameters in (10), we follow Congdon (2001) and the hyper-parameter values are set as  $\alpha_i = \theta_i = \lambda_i = \omega_i = 0.0001, i = 0, 1$ , so we can use the MCMC procedure to extract posterior samples of (10) by means of Gibbs sampling process in OpenBUGS software.

### 5. Simulation Study: Recovery of the True Underlying Parameters

An experiment intends to investigate the ability of the proposed EM algorithm to recover the true underlying parameters. We generate 5000 synthetic Monte Carlo samples

of different sample sizes  $n = 30, 70, 100$  and  $200$  from the CWE distribution and following three parameter scenarios (each scenario corresponding to density plotted as “dotdash” line in Figure 1):

**Scenario 1:**  $\alpha = 0.5, \lambda = 1, \theta = 2, \omega = 0.2$ .

**Scenario 2:**  $\alpha = 2, \lambda = 1, \theta = 0.3, \omega = 0.2$ .

**Scenario 3:**  $\alpha = 0.5, \lambda = 1, \theta = 0.3, \omega = 0.6$ .

The accuracies of the parameter estimates are measured by computing the mean absolute bias (MAB) and the root mean square error (RMSE), defined as

$$MAB = \frac{1}{5000} \sum_{i=1}^{5000} |\hat{\theta}_i - \theta_A| \quad \text{and} \quad RMSE = \sqrt{\frac{1}{5000} \sum_{i=1}^{5000} (\hat{\theta}_i - \theta_A)^2},$$

where  $\hat{\theta}_i$  denotes the prediction of a specific parameter at the  $i$ -th replication and  $\theta_A$  denotes the actual specific parameter value. Table 2 shows the simulation results for the CWE distribution. As expected, the MAB and RMSE tend toward zero when the sample size increases, showing empirically the consistency of the ML estimates obtained via the EM algorithm.

**Table 2.** Simulation results, based on 5000 replications, to evaluate the EM algorithm under three scenarios.

Sample Size	Parameter	$n = 30$		$n = 70$		$n = 100$		$n = 200$	
		MAB	RMSE	MAB	RMSE	MAB	RMSE	MAB	RMSE
Scenario 1	$\alpha$	0.357	0.419	0.270	0.329	0.232	0.284	0.171	0.213
	$\lambda$	0.204	0.259	0.139	0.176	0.118	0.150	0.083	0.104
	$\theta$	1.906	7.158	1.176	2.943	0.955	1.825	0.645	1.003
	$\omega$	0.094	0.148	0.073	0.109	0.065	0.093	0.050	0.069
Scenario 2	$\alpha$	1.749	2.892	1.165	1.626	0.959	1.286	0.697	0.898
	$\lambda$	0.308	0.419	0.197	0.259	0.163	0.213	0.115	0.148
	$\theta$	0.189	0.282	0.102	0.151	0.080	0.115	0.052	0.069
	$\omega$	0.118	0.158	0.098	0.126	0.088	0.112	0.069	0.087
Scenario 3	$\alpha$	0.449	0.665	0.359	0.508	0.306	0.412	0.236	0.310
	$\lambda$	0.366	0.564	0.229	0.323	0.187	0.252	0.132	0.172
	$\theta$	0.092	0.121	0.059	0.075	0.050	0.064	0.036	0.045
	$\omega$	0.169	0.204	0.126	0.155	0.109	0.135	0.082	0.102

## 6. Numerical Examples for Insurance Data Fitting

In this section, we evaluate the performance and various aspects of the proposed model using insurance claims data. The proposed distribution is fitted to the data by implementing the ECM algorithm described in Section 3. For the sake of comparison, the reduced WE, CE and exponential (Exp) models are also fitted as sub-models of CWE distribution. To compare how well the models fit the data, we adopt the Akaike information criterion (AIC) (Akaike 1973) and the Bayesian information criterion (BIC) (Schwarz 1978), defined as  $AIC = 2p - 2\ell_{max}$  and  $BIC = p \log n - 2\ell_{max}$ , where  $p$  is the number of free parameters in the model and  $\ell_{max}$  the maximized log-likelihood value. For both AIC and BIC, a smaller value indicates a better model fit.

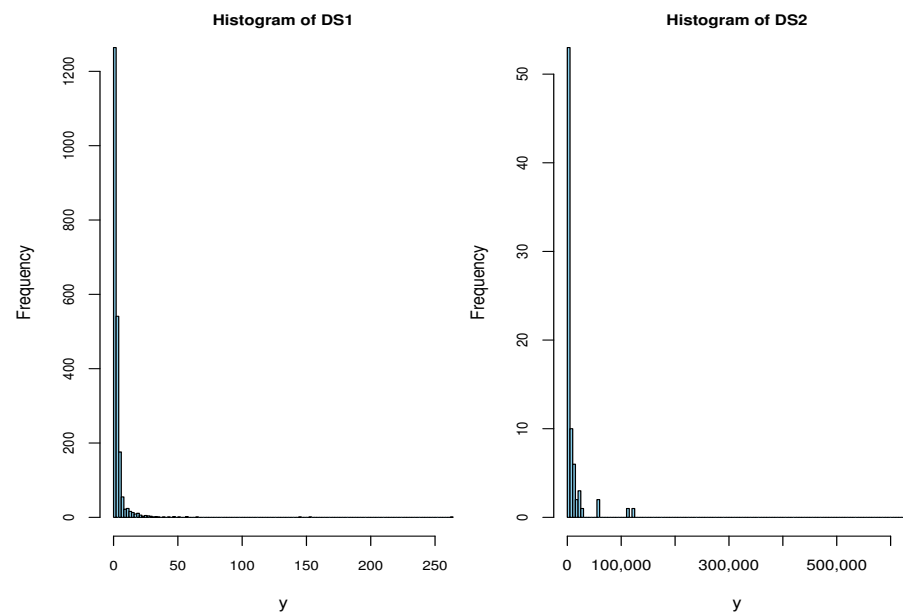
The first dataset (DS1) comprises Danish fire losses analyzed in McNeil (1997). This dataset is frequently used for comparison of methods; see Eling (2012) and references therein. These data represent Danish fire losses in million Danish Kroner and were collected by a Danish reinsurance company. The dataset contains individual losses above 1 million Danish Kroner, a total of 2167 individual losses, covering the period from 3 January 1980 to 31 December 1990. Data are adjusted for inflation to reflect 1985 values and are available in R packages `evir` and `fExtremes`.

The second dataset (DS2), analyzed by Cummins and Freifelder (1978), contains 80 fire losses from 500 buildings a large university owned from 1951 to 1973. Cummins et al.

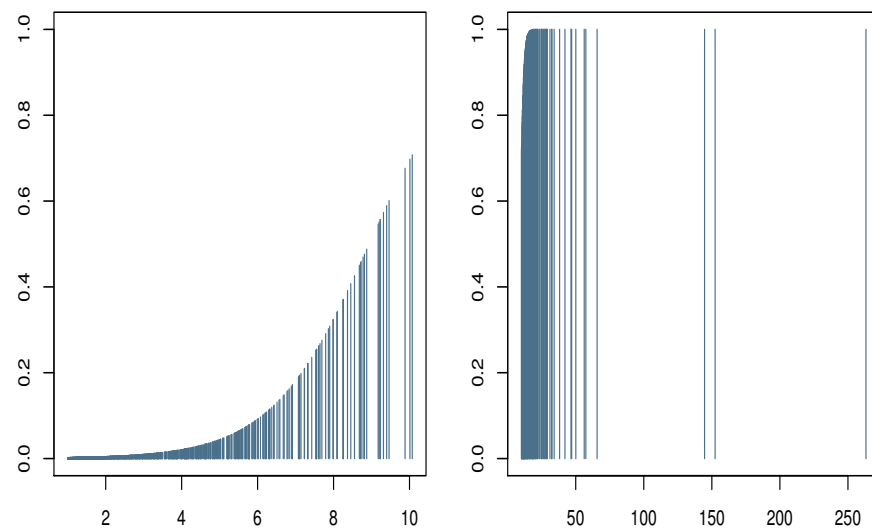


(1990) found that the log-normal and gamma distributions did not have sufficient heavy tails to model the data, so they considered the generalized beta of the second kind (GB2) distribution.

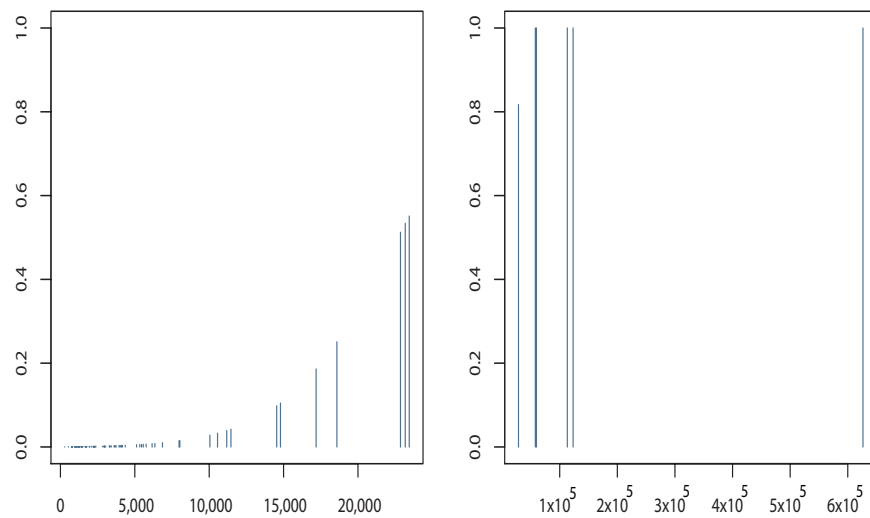
Figure 3 presents two histograms for the considered datasets. Both histograms reveal a typical feature of insurance claims data: a large number of small losses and a small number of very large losses. Table 3 reports parameter estimates, standard error and model fit criteria for all fitted models. Observing the Table 3, it is evident from the AIC and BIC values that the CWE model provides better fit than other fitted models. The posterior probability of each observation belonging to unusual observations is depicted in Figures 4 and 5, those reveal that the unusual data have the highest posterior probability and the original data have small posterior probability, showing clearly the impact of outliers.



**Figure 3.** Data histograms corresponding to DS1 and DS2 datasets.



**Figure 4.** Posterior probability that each observation is unusual, corresponding to DS1 dataset. (Left) panel is for the first 2060 observations and (right) panel for the 107 last observations.



**Figure 5.** Posterior probability that each observation is unusual, corresponding to the DS2 dataset. Left panel is for the first 74 observations and right panel for the six last observations.

**Table 3.** Summary results from fitting various models to the data. The bold entries highlight the smallest AIC and BIC values for each model.

Dataset	Model	$\hat{\lambda}$	$\hat{\theta}$	$\hat{\alpha}$	$\hat{\omega}$	$p$	$\ell_{max}$	AIC	BIC
DS1	Exp	0.295	–	–	–	1	–4809.396	9620.792	9626.474
	WE	0.350	–	4.420	–	2	–4576.327	9160.655	9183.379
	CE	0.401	0.107	–	0.043	3	–4556.646	9119.292	9136.335
	CWE	0.818	0.113	0.194	0.064	4	– <b>4119.475</b>	<b>8246.950</b>	<b>8269.675</b>
DS2	Exp	$0.590 \times 10^{-5}$	–	–	–	1	–859.0414	1720.083	1722.465
	WE	$0.596 \times 10^{-4}$	–	103.667	–	2	–858.548	1725.096	1734.624
	CE	$0.223 \times 10^{-3}$	0.033	–	0.096	3	–796.815	1599.630	1606.776
	CWE	$0.258 \times 10^{-3}$	0.035	9.112	0.104	4	– <b>793.087</b>	<b>1594.175</b>	<b>1603.703</b>

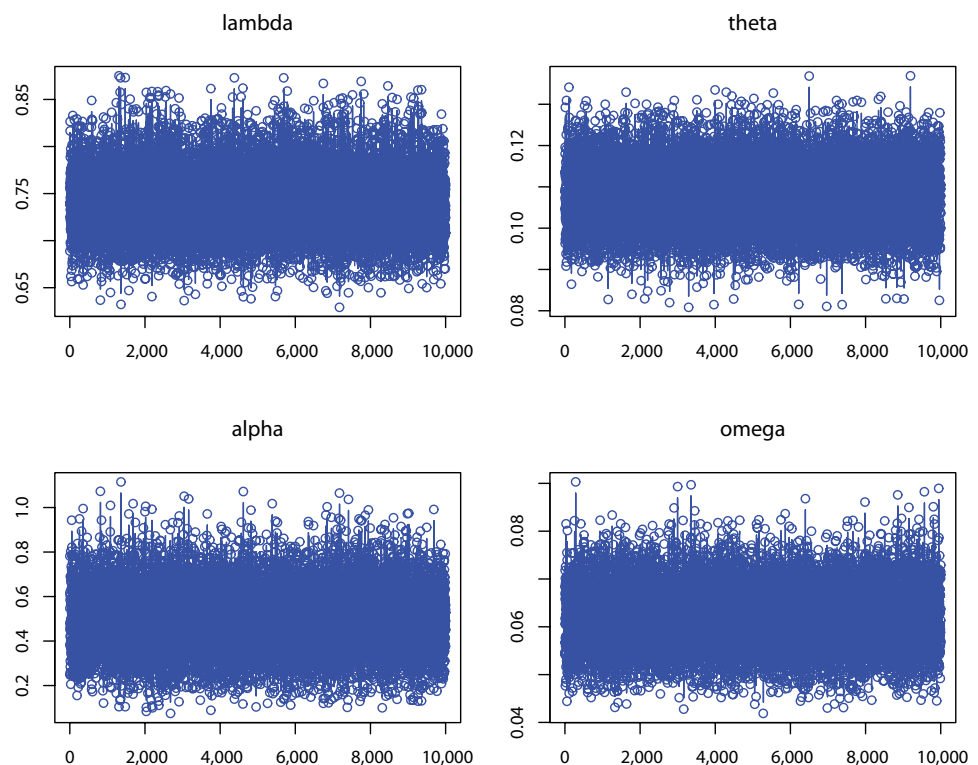
### 7. Bayesian Numerical Results

We used an MCMC procedure based on 10,000 replicates with 1000 samples discarded as burn-in to compute the Bayesian estimators. The corresponding Bayesian point estimation and posterior risk based on DS1 and DS2 datasets are provided in Table 4. It can be seen that for the both datasets, the resulting log-likelihood values ( $\ell_{max}$ ) are close to the obtained ones by the EM-algorithm given in Table 3, indicating the efficiency of the Bayesian approach to estimate the model parameters. It is noteworthy to mention that the KLF and PLF loss functions yields the highest log-likelihood values for DS1 and DS2 datasets, respectively.

Table 5 provides 95% credible and HPD intervals for the parameters of the CWE distribution. The posterior samples are extracted using Gibbs sampling technique. Moreover, we provide the posterior summary plots in Figures 6–8. These plots confirm that the convergence of the Gibbs sampling process occurred.

**Table 4.** Bayesian estimates and their posterior risks of the CWE distribution parameters under different loss functions based on DS1 and DS2 datasets. The bold entries highlight the highest  $\ell_{max}$  values for each model.

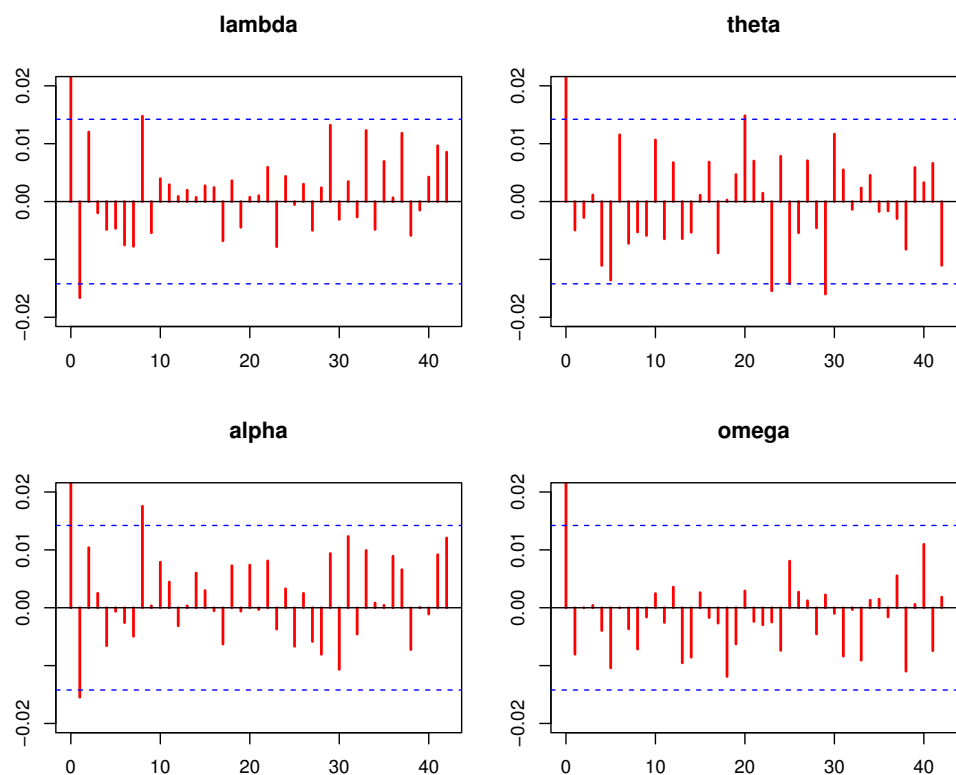
Data	DS1				
Bayesian Estimation					
Loss Function	$\hat{\lambda} (r_{\hat{\lambda}})$	$\hat{\theta} (r_{\hat{\theta}})$	$\hat{\alpha} (r_{\hat{\alpha}})$	$\hat{\omega} (r_{\hat{\omega}})$	$\ell_{max}$
SELF	0.74215 (0.00119)	0.10790 (0.00006)	0.48874 (0.02041)	0.06219 (0.00004)	-4120.942
WSELF	0.74054 (0.00160)	0.10734 (0.00056)	0.44265 (0.04609)	0.06150 (0.00068)	-4120.876
MSELF	0.73894 (0.00216)	0.10677 (0.00527)	0.38906 (0.12106)	0.06081 (0.01121)	-4121.926
PLF	0.74296 (0.00161)	0.10818 (0.00056)	0.50920 (0.04090)	0.06253 (0.00067)	-4121.245
KLF	0.74135 (0.00217)	0.10762 (0.00525)	0.46513 (0.10154)	0.06184 (0.01109)	<b>-4120.797</b>
Data	DS2				
Bayesian Estimation					
Loss Function	$\hat{\lambda} (r_{\hat{\lambda}})$	$\hat{\theta} (r_{\hat{\theta}})$	$\hat{\alpha} (r_{\hat{\alpha}})$	$\hat{\omega} (r_{\hat{\omega}})$	$\ell_{max}$
SELF	0.000275 ( $1.889 \times 10^{-9}$ )	0.0366 ( $3.6 \times 10^{-5}$ )	6.90240 (1.0208)	0.1038 (0.0019)	-793.234
WSELF	0.000268 ( $6.411 \times 10^{-6}$ )	0.0356 (0.0010)	6.75207 (0.1503)	0.0824 (0.0214)	-793.4759
MSELF	0.000262 (0.022395)	0.0345 (0.0309)	6.60050 (0.0224)	0.0605 (0.2651)	-794.104
PLF	0.000278 ( $6.821 \times 10^{-6}$ )	0.0371 (0.0009)	6.97590 (0.1471)	0.1128 (0.0179)	<b>-793.208</b>
KLF	0.000271 (0.023714)	0.0361 (0.0285)	6.82680 (0.0221)	0.0925 (0.2448)	-809.881



**Figure 6.** Plots of Bayesian analysis and performance of Gibbs sampling for DS1 dataset. Trace plots of each CWE distribution parameter.

**Table 5.** Credible and HPD intervals of parameters  $\lambda$ ,  $\theta$ ,  $\alpha$  and  $\omega$  for DS1 and DS2 datasets.

Data	DS1	
	Credible Interval	HPD Interval
$\lambda$	(0.7184, 0.7645)	(0.6740, 0.8108)
$\theta$	(0.1025, 0.1132)	(0.09299, 0.12300)
$\alpha$	(0.3901, 0.5777)	(0.2336, 0.7923)
$\omega$	(0.05767, 0.06658)	(0.04907, 0.07438)
Data	DS2	
	Credible Interval	HPD Interval
$\lambda$	(0.00024, 0.00030)	(0.00019, 0.00035)
$\theta$	(0.03269, 0.04018)	(0.02322, 0.04958)
$\alpha$	(6.16500, 7.68300)	(5.04200, 8.76500)
$\omega$	(0.07135, 0.12980)	(0.03289, 0.19750)

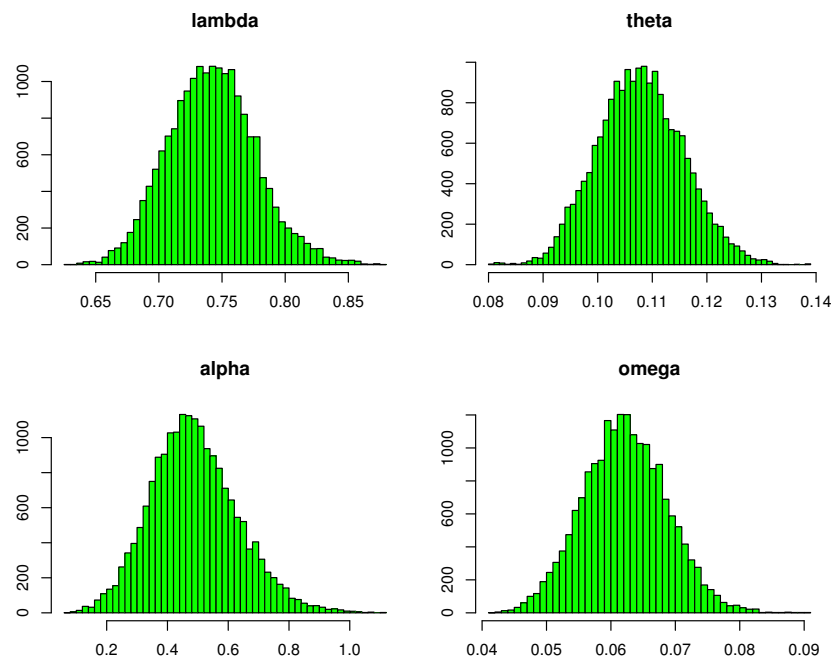


**Figure 7.** Plots of Bayesian analysis and performance of Gibbs sampling for DS1 dataset. Autocorrelation plots of each CWE distribution parameter.

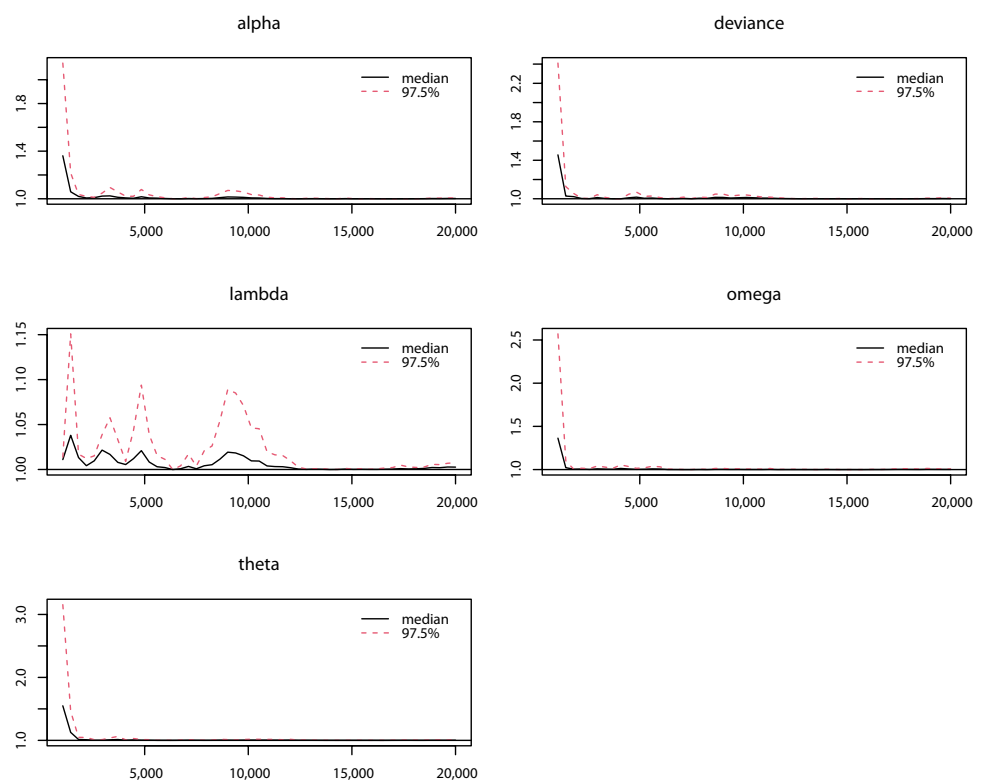
In order to avoid repetition in evaluation of the MCMC procedure in Bayesian analysis, we just reported the Gelman–Rubin and Geweke–Raftery–Lewis diagnostics measures for checking the convergence based on data set DS1 in Table 6. For more details on these indexes see Lee et al. (2014). The Gelman–Rubin diagnostic is equal to 1 for parameters  $\lambda$ ,  $\theta$ ,  $\alpha$  and  $\omega$ . Hence, the chains could be accepted, and this indicates the estimates come from a state space of the parameter, as depicted in Figure 9.

**Table 6.** Diagnostics using the Gelman-Rubin and Geweke-Raftery-Lewis methods for parameters  $\alpha$ ,  $\beta$  and  $\lambda$  based on DS1 dataset.

Parameter	Gelman-Rubin	Geweke ( $Z_{0.025} = \pm 1.96$ )	Raftery-Lewis
$\lambda$	1	-0.5880	5.1
$\theta$	1	0.3205	4.8
$\alpha$	1	0.7607	5.01
$\omega$	1	0.3679	4.632

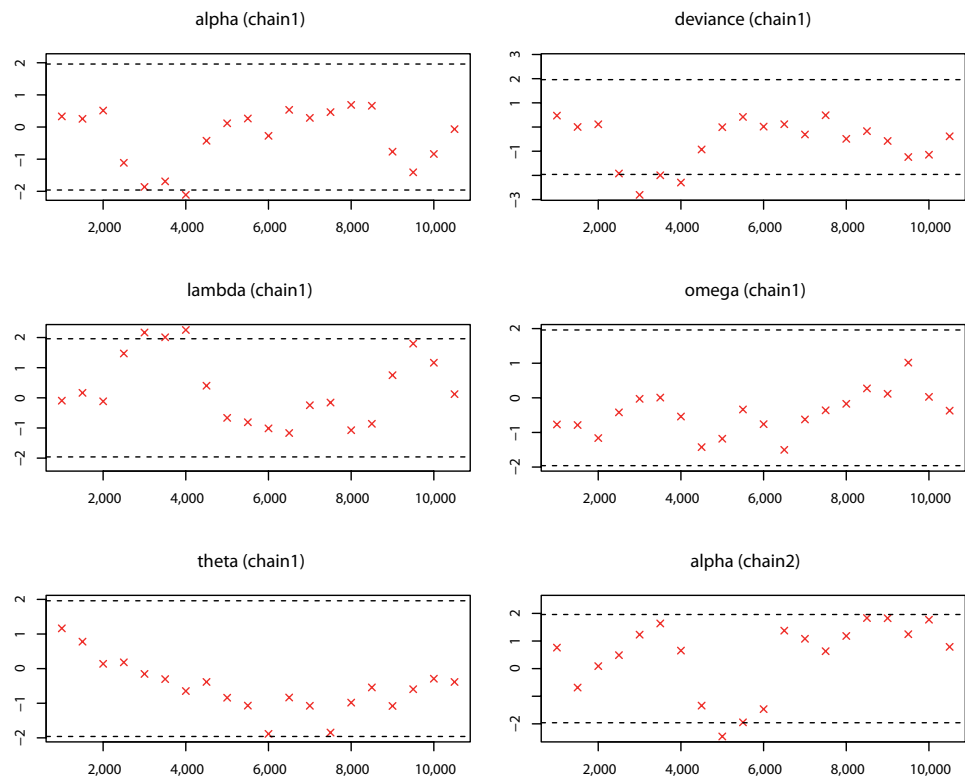


**Figure 8.** Plots of Bayesian analysis and performance of Gibbs sampling for DS1 dataset. Histogram plots of each CWE distribution parameter.

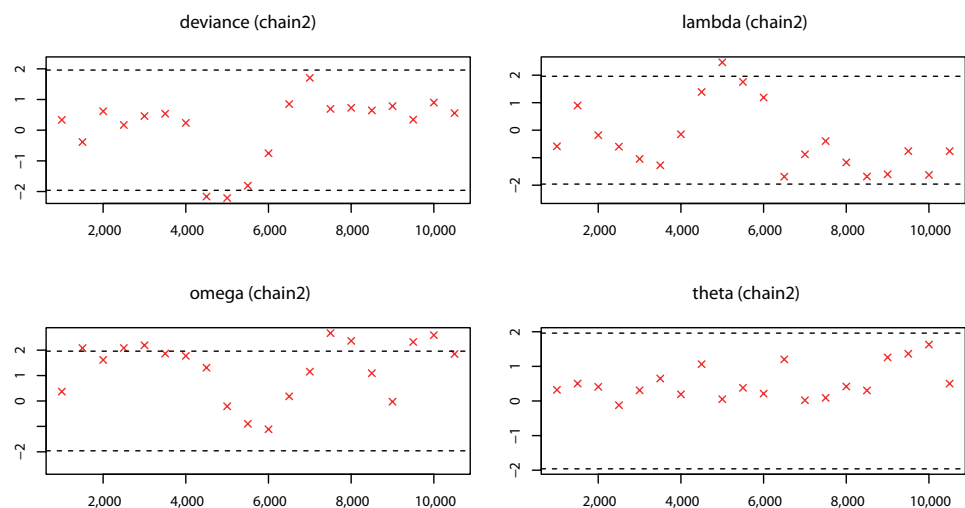


**Figure 9.** Gelman plot diagnostic for each CWE distribution parameter based on DS1 dataset.

From Table 6, Geweke–Raftery–Lewis test statistics for parameters  $\lambda$ ,  $\theta$ ,  $\alpha$  and  $\omega$  are  $-0.588$ ,  $0.320$ ,  $0.761$  and  $0.368$ , respectively. Therefore, also in this case, the chain is acceptable, as shown in Figures 10 and 11. Moreover, the reported diagnostics statistics for parameters  $\alpha$ ,  $\beta$  and  $\lambda$  based on the Geweke–Raftery–Lewis measure don't show significant correlations between estimates. Hence, the estimated values have good mixing.



**Figure 10.** Geweke plot diagnostic (chain1) for each CWE distribution parameter based on DS1 dataset.



**Figure 11.** Geweke plot diagnostic (chain2) for each CWE distribution parameter based on DS1 dataset.

### 8. Conclusions

This paper extended the WE distribution to a richer family, the CWE distribution, to deal with data displaying large and positive skewness as well as a wide right tail. This four-parameter model is a mixture of two WE distributions in which one has an enhanced

scale and hence a thicker tail to capture extreme losses. EM and Bayesian computational techniques were used to estimate parameters. The effectiveness and efficiency of the EM algorithm were evaluated by conducting one simulation study. By analyzing two real insurance claims datasets, we found that the CWE distribution outperformed the CE distribution in terms of model fit. The result show that both EM and Bayesian approaches are appropriate tools to estimate the model parameters. In addition, it is possible to consider proposed distribution to fit lifetimes, and how the suggested algorithms will be adjusted in case of truncated or censored data. Another application could be done in actuarial science context; specifically, how CWE distribution could be employed to calculate the VaR and TVaR (Bargès et al. 2009).

**Author Contributions:** Conceptualization, A.M. and O.K.; methodology, A.M. and O.K.; software, A.M. and O.K.; validation, A.M., O.K. and J.E.C.-R.; formal analysis, O.K.; investigation, A.M., O.K. and J.E.C.-R.; resources, J.E.C.-R.; data curation, A.M.; writing—original draft preparation, A.M., O.K. and J.E.C.-R.; writing—review and editing, A.M., O.K. and J.E.C.-R.; visualization, A.M. and O.K.; supervision, J.E.C.-R.; project administration, O.K.; funding acquisition, J.E.C.-R. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was fully supported by FONDECYT (Chile) grant No. 11190116.

**Data Availability Statement:** The datasets analyzed during the current study are available from the corresponding author on reasonable request.

**Acknowledgments:** The authors also thank the editor and two anonymous referees for their helpful comments and suggestions. All R and OpenBUGS codes used in this paper are available upon request from the corresponding author. The datasets analyzed during the current study are available from the corresponding author on reasonable request.

**Conflicts of Interest:** The authors declare that there is no conflict of interest in the publication of this paper.

## Appendix A. R Code to Fit the CWE Distribution Using EM-Algorithm

```
EM.CWE <- function(y, om, al, la, th, iter.max = 500, tol=10^-6){
  f.CWE <- function(y,om,al,la,th)
  (1-om)*(al+1)/al*la*exp(-la*y)*(1-exp(-la*al*y))+om*(al+1)/al*th*la
  *exp(-th*la*y)*(1-exp(-th*la*al*y))
  n <- length(y); LL <- 1 ; dif <- 1 ; count <- 1
  while ((dif > tol) & (count <= iter.max)) {
    # E steps
    gam <- om*(al+1)/al*th*la*exp(-th*la*y)*(1-exp(-th*la*al*y))/
    f.CWE(y,om,al,la,th)
    ta1 <- (1-gam)*(1/(la*al*y)-1/(exp(la*al*y)-1) )
    ta2 <- gam*(1/(th*la*al*y)-1/(exp(th*la*al*y)-1) )
    # M steps
    om <- sum(gam)/n
    al <- n/(la*sum(ta1*y+th*ta2*y))-1
    la <- 2*n/sum((1-gam)*y+al*ta1*y+th*gam*y+al*th*ta2*y)
    th <- (2*sum(gam))/(la*sum(gam*y+al*ta2*y))
    LL.new <- sum(log(f.CWE(y,om,al,la,th)))
    count <- count +1
    dif <- abs(LL.new/LL-1)
  }
  print.foo <- function(x) print(x[1:8])
  aic <- -2 * LL.new + 2 * 4
  bic <- -2 * LL.new + log(n) * 4
  Ret <-list(omega=om, alpha=al, lambda=la, theta=th, loglik=LL.new,
  AIC=aic, BIC=bic, iter=count, out.prob=gam)
  class(Ret) <- "foo"
  return(Ret)
}
```

## References

- Aitkin, Murray, and Granville Tunnicliffe Wilson. 1980. Mixture models, outliers, and the em algorithm. *Technometrics* 22: 325–31. [CrossRef]
- Akaike, Hirotogu. 1973. Information theory as an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*. Budapest: BNPBF Csaki Budapest, Akademiai Kiado, Hungary.
- Bargès, Mathieu, Hélène Cossette, and Etienne Marceau. 2009. TVaR-based capital allocation with copulas. *Insurance: Mathematics and Economics* 45: 348–61. [CrossRef]
- Bernardi, Mauro, Antonello Maruotti, and Lea Petrella. 2012. Skew mixture models for loss distributions: A bayesian approach. *Insurance: Mathematics and Economics* 51: 617–23. [CrossRef]
- Cavieres, Joaquin, German Ibacache-Pulgar, and Javier E. Contreras-Reyes. 2022. Thin plate spline model under skew-normal random errors: Estimation and diagnostic analysis for spatial data. *Journal of Statistical Computation and Simulation*, in press. [CrossRef]
- Chen, Ming-Hui, Joseph G. Ibrahim, and Debajyoti Sinha. 1999. A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association* 94: 909–19. [CrossRef]
- Chung, Younshik, and Chansoo Kim. 2004. Measuring robustness for weighted distributions: Bayesian perspective. *Statistical Papers* 45: 15–31. [CrossRef]
- Congdon, Peter. 2001. *Bayesian Statistical Modelling*. New York: John Wiley & Sons.
- Contreras-Reyes, Javier E., Freddy O. López Quintero, and Rodrigo Wiff. 2018. Bayesian modeling of individual growth variability using back-calculation: Application to pink cusk-eel (*Genypterus blacodes*) off Chile. *Ecological Modelling* 385: 145–53. [CrossRef]
- Cummins, J. David, Georges Dionne, James B. McDonald, and B. Michael Pritchett. 1990. Applications of the GB2 family of distributions in modeling insurance loss processes. *Insurance: Mathematics and Economics* 9: 257–72. [CrossRef]
- Cummins, J. David, and Leonard R. Freifelder. 1978. A comparative analysis of alternative maximum probable yearly aggregate loss estimators. *Journal of Risk and Insurance* 45: 27–52. [CrossRef]
- Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39: 1–22.
- Eling, Martin. 2012. Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models? *Insurance: Mathematics and Economics* 51: 239–48. [CrossRef]
- Geman, Stuart, and Donald Geman. 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721–41. [CrossRef] [PubMed]
- Gupta, Ramesh C., and S. N. U. A. Kirmani. 1990. The role of weighted distributions in stochastic modeling. *Communications in Statistics-Theory and Methods* 19: 3147–62. [CrossRef]
- Gupta, Rameshwar D., and Debasis Kundu. 2009. A new class of weighted exponential distributions. *Statistics* 43: 621–34. [CrossRef]
- Hastings, W. Keith. 1970. Monte carlo sampling methods using markov chains and their applications. *Biometrika* 57: 97–109. [CrossRef]
- Hennig, Christian, and Tim F. Liao. 2013. How to find an appropriate clustering for mixed-type variables with application to socioeconomic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 62: 309–69.
- Kharazmi, Omid, Ali Saadati Nik, Behrang Chaboki, and Gauss M. Cordeiro. 2021. A novel method to generating two-sided class of probability distributions. *Applied Mathematical Modelling* 95: 106–24. [CrossRef]
- Kharazmi, Omid, G. G. Hamedani, and Gauss M. Cordeiro. 2022. Log-mean distribution: Applications to medical data, survival regression, bayesian and non-bayesian discussion with MCMC algorithm. *Journal of Applied Statistics* 1–26. in press. [CrossRef]
- Larose, Daniel T., and Dipak K. Dey. 1996. Weighted distributions viewed in the context of model selection: a bayesian perspective. *Test* 5: 227–46. [CrossRef]
- Lee, Cheol-Eung, Sang Ug Kim, and Sangho Lee. 2014. Time-dependent reliability analysis using bayesian MCMC on the reduction of reservoir storage by sedimentation. *Stochastic Environmental Research and Risk Assessment* 28: 639–54. [CrossRef]
- Liu, Chuanhai, and Donald B. Rubin. 1994. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika* 81: 633–48. [CrossRef]
- Mahdavi, Abbas, Vahid Amirzadeh, Ahad Jamalizadeh, and Tsung-I. Lin. 2021a. Maximum likelihood estimation for scale-shape mixtures of flexible generalized skew normal distributions via selection representation. *Computational Statistics* 36: 2201–30. [CrossRef]
- Mahdavi, Abbas, Vahid Amirzadeh, Ahad Jamalizadeh, and Tsung-I. Lin. 2021b. A multivariate flexible skew-symmetric-normal distribution: Scale-shape mixtures and parameter estimation via selection representation. *Symmetry* 13: 1343. [CrossRef]
- Maruotti, Antonello, Valentina Raponi, and Francesco Lagona. 2016. Handling endogeneity and nonnegativity in correlated random effects models: Evidence from ambulatory expenditure. *Biometrical Journal* 58: 280–302. [CrossRef] [PubMed]
- McLachlan, Geoffrey J., and Thiriyambakam Krishnan. 2007. *The EM Algorithm and Extensions*. New York: John Wiley & Sons, vol. 382.
- McNeil, Alexander J. 1997. Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin: The Journal of the IAA* 27: 117–37. [CrossRef]
- Meng, Xiao-Li, and Donald B. Rubin. 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* 80: 267–78. [CrossRef]
- Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* 21: 1087–92. [CrossRef]
- Navarro, Jorge, Jose M. Ruiz, and Yolanda Del Aguila. 2006. Multivariate weighted distributions: A review and some extensions. *Statistics* 40: 51–64. [CrossRef]
- Okhli, Kheirolah, and Mehdi Jabbari Nooghabi. 2021. On the contaminated exponential distribution: A theoretical bayesian approach for modeling positive-valued insurance claim data with outliers. *Applied Mathematics and Computation* 392: 125712. [CrossRef]



- Patil, Ganapati P, and C. R. Rao. 1977. *The Weighted Distributions: A Survey and Their Applications*. *Applications of Statistics*. Amsterdam: North Holland, vol. 383, p. 405.
- Patil, Ganapati P. 1991. Encountered data, statistical ecology, environmental statistics, and weighted distribution methods. *Environmetrics* 2: 377–423. [CrossRef]
- Punzo, Antonio, Angelo Mazza, and Antonello Maruotti. 2018. Fitting insurance and economic data with outliers: A flexible approach based on finite mixtures of contaminated gamma distributions. *Journal of Applied Statistics* 45: 2563–84. [CrossRef]
- Lopez Quintero, Freddy Omar, Javier E. Contreras-Reyes, Rodrigo Wiff, and Reinaldo B. Arellano-Valle. 2017. Flexible bayesian analysis of the von Bertalanffy growth function with the use of a log-skew-t distribution. *Fishery Bulletin* 115: 13–26. [CrossRef]
- Redner, Richard A., and Homer F. Walker. 1984. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* 26: 195–239. [CrossRef]
- Schwarz, Gideon. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6: 461–64. [CrossRef]

Article

# Circular-Statistics-Based Estimators and Tests for the Index Parameter $\alpha$ of Distributions for High-Volatility Financial Markets

Ashis SenGupta <sup>1,2,3,\*</sup> and Moumita Roy <sup>4,\*</sup>

<sup>1</sup> Department of Mathematics, Indian Institute of Technology, Kharagpur 721302, India

<sup>2</sup> Department of Population Health Sciences, MCG, Augusta University, Augusta, GA 30912-4900, USA

<sup>3</sup> Department of Statistics, Middle East Technical University, Ankara 06800, Turkey

<sup>4</sup> Department of Statistics, Midnapore College (Autonomous), Midnapore 721101, India

\* Correspondence: amsseng@gmail.com (A.S.); mouroy.roy@gmail.com (M.R.)

**Abstract:** The distributions for highly volatile financial time-series data are playing an increasingly important role in current financial scenarios and signal analyses. An important characteristic of such a probability distribution is its tail behaviour, determined through its tail thickness. This can be achieved by estimating the index parameter of the corresponding distribution. The normal and Cauchy distributions, and, sometimes, a mixture of the normal and Cauchy distributions, are suitable for modelling such financial data. The family of stable distributions can provide better modelling for such financial data sets. Financial data in high-volatility markets may be better modelled, in many cases, by the Linnik distribution in comparison to the stable distribution. This highly flexible family of distributions is better capable of modelling the inflection points and tail behaviour compared to the other existing models. The estimation of the tail thickness of heavy-tailed financial data is important in the context of modelling. However, the new probability distributions do not admit any closed analytical form of representation. Thus, novel methods need to be developed, as only a few can be found in the literature. Here, we recall a recent novel method, developed by the authors, based on a trigonometric moment estimator using circular distributions. The linear data may be transformed to yield circular data. This transformation is solely for yielding a suitable estimator. Our aim in this paper is to provide a review of the few existing methods, discuss some of their drawbacks, and also provide a universal ( $\forall \alpha \in (0, 2]$ ), efficient, and easily implementable estimator of  $\alpha$  based on the transformation mentioned above. Novel, circular-statistics-based tests for the index parameter  $\alpha$  of the stable and Linnik distributions are introduced and also exemplified with real-life financial data. Two real-life data sets are analysed to exemplify the methods recommended and enhanced by the authors.

**Keywords:** characteristic function-based estimator; estimation; fractional moment estimator; Hill estimator; index parameter; trigonometric method of moment estimator; wrapped Linnik; wrapped stable



**Citation:** SenGupta, Ashis, and Moumita Roy. 2023.

Circular-Statistics-Based Estimators and Tests for the Index Parameter  $\alpha$  of Distributions for High-Volatility Financial Markets. *Journal of Risk and Financial Management* 16: 405.  
<https://doi.org/10.3390/jrfm16090405>

Academic Editors: Shigeyuki Hamori and Robert Brooks

Received: 12 May 2023

Revised: 1 August 2023

Accepted: 7 August 2023

Published: 11 September 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In the modern era, there is an increasing need for modelling financial markets (and engineering sciences, e.g., signal detection) with high volatility. An important characteristic of such a probability distribution is its tail behaviour, determined through its tail thickness. There is a need for modelling such financial data. High variability has also been a common characteristic of modern circular data.

Corresponding circular distributions are characterised by heavy or long tails. The normal and Cauchy distributions, and, sometimes, a mixture of the normal and Cauchy distributions, are suitable for modelling such financial data. The family of stable distributions can provide better modelling for such financial data sets. Highly volatile financial time-series data may be better modelled, in many cases, by the Linnik distribution in

comparison to the stable distribution, e.g., see Anderson and Arnold (1993). This highly flexible family of distributions is better capable of modelling the inflection points and tail behaviour compared to the existing popular flexible symmetric unimodal models.

There have been a lot of studies establishing the use of the Linnik family of distributions as a highly flexible, important, and useful family for modelling financial data. However, its implementation for real-life data seems to have been somewhat restricted, possibly because of the lack of a simple and efficient estimator of the parameter, particularly that of the index parameter  $\alpha$ . The estimation of the tail thickness of heavy-tailed financial data using the index parameter  $\alpha$  is important in the context of modelling.

Our aim in this paper is to provide a universal (for all  $\alpha \in (0, 2]$ ), efficient, and easily implementable estimator of  $\alpha$  after presenting a review of the few existing methods. The issue behind the derivation is to study the advantages and also to point out the shortcomings of some of the estimators and hence to obtain better estimators that eliminate the effects of the shortcomings of the former ones.

We have observed that the circular-statistics-based estimators can be quite useful in this context, which is enhanced in this paper. Circular statistics are obtained for circular data. In many emerging real-life situations, we not only make observations on linear variables but also on circular ones, that is, on angular propagations, orientations, directional movements, and strictly periodic occurrences. Such data are referred to as directional data, which, in two dimensions, are known as circular data. Linear data may be transformed into circular data using the method of wrapping.

Here, we recall two highly flexible families of circular distributions, e.g., the wrapped stable family, in Section 2, and the wrapped Linnik family in Section 3, and the novel, universal, efficient, and easily implementable estimators of  $\alpha$  derived from these are presented in Section 7. In Section 4, descriptions of the classical Hill estimator by Hill (1975), and its generalisation by Brilhante et al. (2013), are presented. In Section 5, the fractional moment estimator for the symmetric Linnik distribution proposed by Kozubowski (2001) is reviewed. The fractional moment estimator of the characteristic exponent used to measure the tail thickness for skewed stable distributions, proposed by Kuruoglu (2001), is particularised to obtain the same for the symmetric stable distribution in Section 5. In Section 6, the characteristic function-based estimator proposed by Anderson and Arnold (1993) is presented. In Section 7, the trigonometric method of moment estimators proposed by SenGupta (1996) and SenGupta and Roy (2023) is presented, which is further modified to obtain an improved estimator (as in SenGupta and Roy 2019, 2023) in Section 8. The trigonometric method of moment estimation is exploited here for symmetric circular distributions only. It can be used for asymmetric distributions as well. But the computations involved are complicated and time consuming and hence are not considered here. In Section 9, the performance of the estimators is discussed through extensive simulations, focusing on their estimated mean bias and estimated root-mean-square errors, as presented in Tables 1 and 2. In Section 10, the computed values of the estimators are obtained for two real-life financial data sets, which are presented in Table 3. In Section 11, novel tests for the index parameter  $\alpha$  of the stable and Linnik distributions are introduced and also illustrated with real-life financial data. Some discussions and conclusions on the different estimators are provided in Section 12. In the Acknowledgement section, the authors express their acknowledgements.

## 2. The Symmetric Stable and Wrapped Stable Family of Distributions

The regular symmetric stable distribution is defined through its characteristic function given by

$$\psi_S(t) = \exp(it\mu - |\sigma t|^\alpha), \quad (1)$$

where  $\mu$  is the location parameter,  $\sigma$  is the scale parameter, and  $\alpha$  is the index or shape parameter of the distribution.

Using Proposition 2.1 on page 31 of Jammalamadaka and SenGupta (2001), the following theorem is obtained (see SenGupta and Roy 2023).

**Theorem 1.** (a) The trigonometric moment of order  $p$  for a wrapped stable distribution corresponds to the value of the characteristic function of the linear stable random variable at the integer value  $p = 1, 2, \dots$ . (b) The characteristic function of the wrapped stable random variable  $\theta$  at the integer  $p$  is

$$\psi_{WS}(p) = E[\exp(ip(\theta - \mu))] = \exp(ip\mu - \rho^{p^\alpha}), \tag{2}$$

where  $\rho = \exp(-\sigma^\alpha)$ ,  $\mu$  is the location parameter,  $\sigma$  is the scale parameter,  $\alpha$  is the index parameter and  $i = \sqrt{-1}$ .

From the stable distribution, we can obtain the wrapped stable distribution (the process of wrapping is explained by Jammalamadaka and SenGupta (2001)). Suppose that  $\theta_1, \theta_2, \dots, \theta_m$  are a random sample of size  $m$  drawn from the wrapped stable distribution (provided by Jammalamadaka and SenGupta (2001)), whose probability density function is given by

$$f(\theta, \rho, \alpha, \mu) = \frac{1}{2\pi} [1 + 2 \sum_{p=1}^{\infty} \rho^{p^\alpha} \cos p(\theta - \mu)] \quad 0 < \rho \leq 1, 0 < \alpha \leq 2, 0 < \mu \leq 2\pi. \tag{3}$$

where  $p = 1, 2, \dots$  and the parameters explained as above.

### 3. The Symmetric Linnik and the Wrapped Linnik Family of Distributions

It was established by Pakes (1998) that the characteristic function of a symmetric ( $\alpha$ ) Linnik (linear) distribution is given by

$$\psi_L(t) = \exp(it\mu)(1 + |t\sigma|^\alpha)^{-1}. \tag{4}$$

The density function cannot be written in an analytical form except for  $\alpha = 2$ . The wrapping of this distribution yields the wrapped symmetric  $\alpha$  Linnik family of distributions. However, this circular family differs from that of the symmetric wrapped stable family and none of these families is a sub-family of the other. In particular, taking  $\alpha = 2$ , for the wrapped symmetric stable family one gets the wrapped Cauchy, while for the wrapped symmetric Linnik family it gives the wrapped Laplace (double exponential) distribution.

Using Proposition 2.1 on page 31 of Jammalamadaka and SenGupta (2001), the following theorem is obtained (see SenGupta and Roy 2023).

**Theorem 2.** (a) The trigonometric moment of order  $p$  for a wrapped Linnik distribution corresponds to the value of the characteristic function of the linear Linnik random variable at the integer value  $p$ . (b) The characteristic function of the wrapped Linnik random variable  $\theta$  at the integer  $p$  is

$$\psi_{WL}(p) = E[\exp(ip(\theta - \mu))] = \exp(ip\mu)(1 + (p\sigma)^\alpha)^{-1}.$$

The probability density function of wrapped Linnik distribution is defined as

$$f(\theta) = \frac{1}{2\pi} [1 + 2 \sum_{p=1}^{\infty} ((1 + (\sigma p)^\alpha)^{-1}) \cos p(\theta - \mu)], \tag{5}$$

where the parameter space is given by

$$\begin{aligned} \Omega &= \Omega_1 \cup \Omega_2, \\ \Omega_1 &= \{(\alpha, \sigma, \mu_0) : 1 \leq \alpha \leq 2, \sigma \geq 1, 0 \leq \mu_0 < 2\pi\} \text{ and} \\ \Omega_2 &= \{(\alpha, \sigma, \mu_0) : 1 < \alpha \leq 2, \sigma < 1, 0 \leq \mu_0 < 2\pi\}. \end{aligned}$$

We observe that these wrapped distributions preserve the parameter  $\alpha$  for the corresponding linear distributions.

Without a loss of generality, we take  $\mu = 0$  and  $\sigma = 1$  in the following. The index parameter of the circular family of distributions plays an important role in determining the thickness and hence the tail behaviour of the distribution. There are, in fact, four possible names for the parameter  $\alpha$ . Some interpret it as the tail thickness parameter or the index parameter used to measure tail thickness mainly for heavy tailed distributions. Others interpret it as the characteristic exponent when it is present in an exponential form in a characteristic function. Sometimes,  $\alpha$  is also defined as the shape parameter along with its three other companions viz. location parameter  $\mu$ , scale parameter  $\sigma$ , and skewness parameter  $\beta$ . For this paper, we assume the symmetric case that is  $\beta = 0$ . Several estimators of this parameter have been developed over time.

#### 4. Hill Estimator and Its Generalisation

The classical Hill estimator (see Hill 1975; Dufour and Kurz-Kim 2010), is a simple non-parametric estimator based on order statistics. Given a sample of  $n$  observations  $X_1, X_2, \dots, X_n$  the Hill estimator is defined as

$$\hat{\alpha}_H = \left[ \left( \frac{1}{k} \sum_{j=1}^k \ln X_{n+1-j:n} \right) - \ln X_{n-k:n} \right]^{-1}$$

with standard error

$$SD(\hat{\alpha}_H) = \frac{k\hat{\alpha}_H}{(k-1)\sqrt{k-2}}$$

where  $k$  is the number of observations which lie on the tails of the distribution of interest and is to be optimally chosen depending on the sample size,  $n$ , and tail thickness  $\alpha$ , as  $k = k(n, \alpha)$  and  $X_{j:n}$  denotes the  $j$ -order statistic of the sample of size  $n$ .

The asymptotic normality of the classical Hill estimator is provided by Goldie and Smith (1987) as

$$\sqrt{k}(\hat{\alpha}_H^{-1} - \alpha^{-1}) \xrightarrow{L} N(0, \alpha^{-2})$$

which leads to the following lemma

**Lemma 1.**

$$\hat{\alpha}_H - \alpha \xrightarrow{L} N\left(0, \frac{1}{\alpha^2 k}\right).$$

This estimator uses the linear function of the order statistics and can be used to estimate  $\alpha \in [1, 2]$  only. Further, it is also “extremely sensitive” to the choice of the optimal number of tail observations  $k$ , which itself is a function of the unknown index parameter  $\alpha$  being estimated.

The Hill estimator is scale invariant since it is defined in terms of the log of ratios but not location invariant. Therefore, centering needs to be performed in order to address the location invariance.

The classical Hill estimator is actually the logarithm of the geometric mean or the logarithm of the mean of order  $p = 0$  of a set of statistics. This estimator has been generalized to a more general mean of order  $p \geq 0$  of the same set of statistics by Brillhante et al. (2013) as follows:

$$\hat{\alpha}_{H_p} = \begin{cases} \frac{(1 - A_p^{-p}(k))}{p}, & \text{if } p > 0 \\ \log_e A_0(k) \equiv \hat{\alpha}_H, & \text{if } p = 0, \end{cases}$$

where the class of statistics  $A_p(k)$  is taken as the mean of order  $p$  of the statistics  $U_{ik}$  given by

$$U_{ik} = \frac{X_{n+1-i:n}}{X_{n-k:n}} = \frac{U(Y_{n+1-i:n})}{U(Y_{n-k:n})},$$

where  $U(\cdot)$  is the generalized inverse function of the cumulative distribution function  $F$  of  $X$  and using the distributional identity  $X = U(Y)$  with  $Y$  as a unit Pareto random variable and

$$A_p(k) = \begin{cases} \left(\frac{\sum_{i=1}^k U_{ik}^p}{k}\right)^{1/p}, & \text{if } p > 0 \\ \left(\prod_{i=1}^k U_{ik}\right)^{1/k}, & \text{if } p = 0 \end{cases} \tag{6}$$

Under the first order condition that the generalized inverse function  $U(\cdot)$  is of regular variation with index  $\alpha$ , the consistency of the generalized class of Hill estimators  $\hat{\alpha}_{H_p}$  is established, provided  $p < \frac{1}{\alpha}$ . In addition, under the assumption of the second order condition, the asymptotic normality of  $\hat{\alpha}_{H_p}$  can also be obtained (see Brillhante et al. 2013) as

$$\hat{\alpha}_{H_p} \equiv^d \alpha + \frac{\sigma_p(\alpha)Z_p(k)}{\sqrt{(k)}} + b_p(\alpha|\rho)A(n/k) + o_p\left(A(n/k)\right),$$

holds for all  $p < \frac{1}{2\alpha}$  and  $Z_p(k)$  is asymptotically standard normal and

$$\sigma_p(\alpha) = \frac{\alpha(1 - p\alpha)}{\sqrt{(1 - 2p\alpha)}} \quad \text{and} \quad b_p(\alpha|\rho) = \frac{1 - p\alpha}{1 - p\alpha - \rho},$$

with  $\rho$  being the second-order parameter, controlling the rate of convergence for the first order condition.

### 5. Fractional Moment Estimator

Another alternative estimator of the index parameter  $\alpha$  is given by Kozubowski (2001) as the usual method of moment estimator with fractional order. If  $x_1, x_2, \dots, x_n$  are realizations from the symmetric Linnik distribution with index parameter  $\alpha$  and scale parameter  $\sigma$ , then the  $p$ th absolute moment is

$$e(p) = E|Y|^p = \frac{p(1 - p)\sigma^p\pi}{\alpha\Gamma(2 - p)\sin(\pi p/\alpha)\cos(\pi p/2)},$$

where  $0 < \alpha \leq 2$  and  $0 < p < \alpha$ . As suggested in Kozubowski (2001), using suitable choices of  $p$  as  $1/2$  and  $1$  and solving the respective equations, the fractional moment estimator of the the index parameter  $\alpha$  can be obtained. This estimator is valid only for  $\alpha > 1$ . To overcome this restriction, a universal and efficient estimator for both stable and Linnik distributions will appear in our next works.

If  $x_1, x_2, \dots, x_n$  are realizations from the symmetric stable distribution with index parameter  $\alpha$ , scale parameter  $\sigma$ , and location parameter  $0$ , then the  $p$ th absolute moment given by Kuruoglu (2001) is

$$E|Y|^p = \frac{\Gamma\left(1 - \frac{p}{\alpha}\right)}{\Gamma(1 - p)} \frac{|\sigma|^\frac{p}{\alpha}}{\cos\left(\frac{p\pi}{2}\right)},$$

where  $-1 < p < \alpha, p \neq 1$  and  $\alpha \neq 1$ . Using the method of moments with the corresponding sample moment,

$$A_p = \frac{1}{n} \sum_{i=1}^n |X_i|^p$$

and applying the following property of gamma function,

$$\frac{\Gamma(p)}{\Gamma(1-p)} = \frac{\pi}{\sin(p\pi)}, \quad p \neq 1$$

the fractional moment estimator of the index parameter  $\alpha$  can be obtained.

### 6. Characteristic Function-Based Estimator

The characteristic function-based estimator of the index parameter of symmetric stable distribution (see Anderson and Arnold 1993) is obtained by the minimization of the objective function (where location parameter  $\mu = 0$  and scale parameter  $\sigma$  unknown) given by,

$$\hat{I}'_s(\alpha) = \sum_{i=1}^n w_i (\hat{\eta}(z_i) - \exp(-|\sigma z_i|^\alpha))^2, \tag{6}$$

where

$$\hat{\eta}(t) = \frac{1}{n} \sum_{j=1}^n \cos(tx_j), \quad t \in R$$

and  $x_1, x_2, \dots, x_n$  are realizations from the symmetric stable( $\alpha$ ) distribution with the theoretical characteristic function  $\exp(-|\sigma z_i|^\alpha)$ ,  $z_i$  is the  $i$ th zero of the  $m$ th degree Hermite polynomial  $H_m(z)$  and

$$w_i = \frac{2^{m-1} m! \sqrt{m}}{(m H_{m-1}(z))^2}.$$

Similarly, the characteristic function-based estimator for that of the symmetric Linnik distribution is obtained by the minimization of the objective function given by

$$I_l(\alpha, \sigma) = \sum_{i=1}^n w_i (\hat{\eta}(z_i) - (1 + |\sigma z_i|^\alpha)^{-1})^2 \tag{7}$$

subject to the constraints,  $1 < \alpha \leq 2$  and  $\sigma > 0$ , where  $x_1, x_2, \dots, x_n$  are realizations from the symmetric Linnik( $\alpha$ ) distribution with the theoretical characteristic function  $(1 + |\sigma z_i|^\alpha)^{-1}$ .

This estimator is consistent, as seen by Anderson and Arnold (1993). However, it cannot be obtained explicitly and needs to be obtained by solving the estimating equations in iterative methods such as the L-BFGS-B method used in R software (see Byrd et al. (1995)).

### 7. The Trigonometric Moment Estimator

It is known, in general, by Jammalamadaka and SenGupta (2001) that the characteristic function of  $\theta$  at the integer  $p$  is defined as,

$$\psi_\theta(p) = E[\exp(ip(\theta - \mu))] = \alpha_p + i\beta_p$$

$$\text{where } \alpha_p = E \cos p(\theta - \mu) \quad \text{and} \quad \beta_p = E \sin p(\theta - \mu).$$

Further by, Jammalamadaka and SenGupta (2001) we know that, for the p.d.f given by (3),

$$\psi_\theta(p) = \rho^{p^\alpha}.$$

$$\text{Hence, } E \cos p(\theta - \mu) = \rho^{p^\alpha} \quad \text{and} \quad E \sin p(\theta - \mu) = 0$$

Suppose  $\theta_1, \theta_2, \dots, \theta_m$  are a random sample of size  $m$  drawn from the wrapped stable density given by (3). We define

$$\begin{aligned} \bar{C}_1 &= \frac{1}{m} \sum_{i=1}^m \cos \theta_i, & \bar{C}_2 &= \frac{1}{m} \sum_{i=1}^m \cos 2\theta_i, & \bar{S}_1 &= \frac{1}{m} \sum_{i=1}^m \sin \theta_i \\ \text{and } \bar{S}_2 &= \frac{1}{m} \sum_{i=1}^m \sin 2\theta_i. \end{aligned}$$

Then, we note that  $\bar{R}_1 = \sqrt{\bar{C}_1^2 + \bar{S}_1^2}$  and  $\bar{R}_2 = \sqrt{\bar{C}_2^2 + \bar{S}_2^2}$ .

Using the method of trigonometric moments estimation, and equating  $\bar{R}_1$  and  $\bar{R}_2$  to the corresponding functions of the theoretical trigonometric moments, we get the estimator of the index parameter  $\alpha$  of wrapped stable distribution (see SenGupta 1996):

$$\alpha_{\hat{W}S} = \frac{1}{\ln 2} \ln \frac{\ln \bar{R}_2}{\ln \bar{R}_1}.$$

Now, suppose  $\theta_1, \theta_2, \dots, \theta_m$  are a random sample of size  $m$  drawn from the wrapped Linnik density given by (5). Using the method of trigonometric moments estimation, and equating the empirical trigonometric moments  $\bar{R}_1$  and  $\bar{R}_2$  to the corresponding theoretical moments, we get the estimator of index parameter  $\alpha$  of wrapped Linnik distribution (as obtained for the wrapped stable distribution by SenGupta 1996),

$$\alpha_{\hat{W}L} = \frac{\ln[(1/\bar{R}_1 - 1)/(1/\bar{R}_2 - 1)]}{\ln(1/2)},$$

where  $\bar{R}_j = \frac{1}{m} \sum_{i=1}^m \cos j(\theta_i - \bar{\theta})$ ,  $j = 1, 2$  and  $\bar{\theta}$  is the mean direction given by  $\bar{\theta} = \arctan\left(\frac{\bar{S}_1}{\bar{C}_1}\right)$ . Note that  $\bar{R}_1 \equiv \bar{R}$ .

The asymptotic normality of the estimators  $\alpha_{\hat{W}S}$  and  $\alpha_{\hat{W}L}$  have been established in the following Theorems 3 and 4 respectively (see SenGupta and Roy 2019, 2023).

**Theorem 3.**

$$\sqrt{m}(\alpha_{\hat{W}S} - \alpha) \xrightarrow{L} N(0, \gamma' \Sigma \gamma),$$

where

$$\gamma = \frac{1}{\ln 2} \left( \frac{-\cos \mu_0}{\rho \ln \rho}, \frac{\cos 2\mu_0}{\rho^{2\alpha} \ln \rho^{2\alpha}}, \frac{-\sin \mu_0}{\rho \ln \rho}, \frac{\sin 2\mu_0}{\rho^{2\alpha} \ln \rho^{2\alpha}} \right)'$$

and

$$\underline{\gamma}' \Sigma \underline{\gamma} = \frac{1}{(\ln 2)^2} \left[ \frac{1 + \rho^{2\alpha} - 2\rho^2}{2(\rho \ln \rho)^2} + \frac{1 + \rho^{4\alpha} - 2(\rho^{2\alpha})^2}{2(\rho^{2\alpha} \ln \rho^{2\alpha})^2} + \frac{2\rho^{2\alpha+1} - \rho - \rho^{3\alpha}}{\rho \ln \rho \rho^{2\alpha} \ln \rho^{2\alpha}} \right].$$

**Theorem 4.**  $\sqrt{m}(\alpha_{\hat{W}L} - \alpha) \xrightarrow{L} N(0, \gamma' \Sigma \gamma)$ , where

$$\gamma = \frac{1}{\ln(1/2)} \begin{pmatrix} \frac{-\cos \mu_0 (1 + (\sigma)^\alpha)^2}{(\sigma)^\alpha} \\ \frac{\cos 2\mu_0 (1 + (2\sigma)^\alpha)^2}{(2\sigma)^\alpha} \\ \frac{-\sin \mu_0 (1 + (\sigma)^\alpha)^2}{(\sigma)^\alpha} \\ \frac{\sin 2\mu_0 (1 + (2\sigma)^\alpha)^2}{(2\sigma)^\alpha} \end{pmatrix} \text{ and}$$



$$\begin{aligned} \underline{\gamma}'\underline{\Sigma}\underline{\gamma} = & \frac{1}{(\ln(1/2))^2} \left[ -\frac{\cos^2 2\mu_0(1 + \sigma^\alpha)(1 + (2\sigma)^\alpha)^2}{2^\alpha \sigma^{2\alpha}} + \frac{\cos^2 2\mu_0(1 + \sigma^\alpha)(1 + (2\sigma)^\alpha)}{2^\alpha \sigma^{2\alpha}} \right. \\ & + \frac{(1 + (2\sigma)^\alpha)^4}{(2\sigma)^{2\alpha}} - \frac{(1 + (2\sigma)^\alpha)^2}{(2\sigma)^{2\alpha}} - \frac{\sin^2 2\mu_0(1 + \sigma^\alpha)(1 + (2\sigma)^\alpha)^2}{2^\alpha \sigma^{2\alpha}} \\ & + \frac{\cos 3\mu_0 \sin \mu_0 \sin 2\mu_0(1 + \sigma^\alpha)^2(1 + (2\sigma)^\alpha)^2}{2^\alpha \sigma^{2\alpha}(1 + (3\sigma)^\alpha)} - \frac{(1 + (\sigma)^\alpha)^2}{(\sigma)^{2\alpha}} + \frac{(1 + \sigma^\alpha)(1 + (2\sigma)^\alpha)}{2^\alpha \sigma^{2\alpha}} \\ & + \frac{\sin^2 \mu_0(1 + \sigma^\alpha)(1 + (2\sigma)^\alpha)}{2^\alpha \sigma^{2\alpha}} - \frac{\cos \mu_0 \cos 2\mu_0 \cos 3\mu_0(1 + \sigma^\alpha)^2(1 + (2\sigma)^\alpha)^2}{2^\alpha \sigma^{2\alpha}(1 + (3\sigma)^\alpha)} \\ & - \frac{3 \cos \mu_0 \sin 2\mu_0 \sin 3\mu_0(1 + \sigma^\alpha)^2(1 + (2\sigma)^\alpha)^2}{2^{\alpha+1} \sigma^{2\alpha}(1 + (3\sigma)^\alpha)} + \frac{\cos 2\mu_0(1 + \sigma^\alpha)^4}{2\sigma^{2\alpha}} + \frac{\sin^2 2\mu_0(1 + \sigma^\alpha)^4}{2\sigma^{2\alpha}(1 + (2\sigma)^\alpha)} \\ & \left. + \frac{\cos 2\mu_0(1 + \sigma^\alpha)^4}{2\sigma^{2\alpha}(1 + (2\sigma)^\alpha)} - \frac{\sin \mu_0 \sin 3\mu_0 \cos 2\mu_0(1 + \sigma^\alpha)^2(1 + (2\sigma)^\alpha)^2}{2^\alpha \sigma^{2\alpha}(1 + (3\sigma)^\alpha)} \right]. \end{aligned}$$

Where  $m$  denotes the sample size and  $\Sigma$  denotes the dispersion matrix of  $(\bar{C}_1, \bar{S}_1, \bar{C}_2, \bar{S}_2)$  in both the above theorems.

Unlike for the previous estimators where at the most simulation results were given for the properties of the estimators, the asymptotic distributions obtained in the Theorems 3 and 4 establish rigorously the theoretical and the analytical properties of the trigonometric moment estimators. The estimators can be shown to be consistent and asymptotically normal(CAN) through the use of the theorems. Additionally, the usefulness of the theorems is to provide a methodology to rigorously test for the index parameter  $\alpha$  which is illustrated in Section 11.

### 8. The Truncated Trigonometric Moment Estimator

The moment estimators  $\alpha_{\hat{W}S}$  and  $\alpha_{\hat{W}L}$  need not always remain in the support of the true parameter  $\alpha$  (that is  $(0,2]$ ). Hence, the moment estimators proposed above need not be proper estimators of  $\alpha$ . Hence, the modified estimators for wrapped stable and wrapped Linnik distribution free from this defect are, respectively, given by

$$\alpha_{\hat{W}S}^{ftm} = \begin{cases} \alpha_{\hat{W}S} & \text{if } 0 < \alpha_{\hat{W}S} < 2 \\ 2 & \text{if } \alpha_{\hat{W}S} \geq 2 \end{cases}$$

and

$$\alpha_{\hat{W}L}^{ftm} = \begin{cases} 1 & \text{if } \alpha_{\hat{W}L} \leq 1 \\ \hat{\alpha} & \text{if } 1 < \alpha_{\hat{W}L} < 2 \\ 2 & \text{if } \alpha_{\hat{W}L} \geq 2 \end{cases}$$

(since the support of  $\alpha$  excludes non-positive values).

The asymptotic normality of the modified truncated estimators  $\alpha_{\hat{W}S}^{ftm}$  and  $\alpha_{\hat{W}L}^{ftm}$  are established, respectively, in the following theorems (see SenGupta and Roy 2019, 2023). We have

#### Theorem 5.

$$(\alpha_{\hat{W}S}^{ftm} - \alpha) \xrightarrow{L} N(0, V(\alpha_{\hat{W}S}^{ftm}))$$

where  $V(\alpha_{\hat{W}S}^{ftm}) = E(\alpha_{\hat{W}S}^{ftm 2}) - \alpha^2$

where  $E(\alpha_{\hat{W}S}^{ftm 2}) = \sigma^2 \left[ \{a^* \phi(a^*) - b^* \phi(b^*) + \Phi(b^*) - \Phi(a^*)\} \right] + \alpha^2 \{ \Phi(b^*) - \Phi(a^*) \} + 2\alpha\sigma \{ \phi(a^*) - \phi(b^*) \}$

where  $a^* = \frac{-\alpha}{\sqrt{\frac{\gamma'\Sigma\gamma}{m}}}$  and  $b^* = \frac{2-\alpha}{\sqrt{\frac{\gamma'\Sigma\gamma}{m}}}$

**Theorem 6.**

$$(\alpha_{WL}^{ftm} - \alpha) \xrightarrow{L} N(0, V(\alpha_{WL}^{ftm}))$$

where  $V(\alpha_{WL}^{ftm}) = E(\alpha_{WL}^{ftm^2}) - \alpha^2$

where  $E(\alpha_{WL}^{ftm^2}) = \Phi(a^*) + \sigma^2 \left[ \{a^* \phi(a^*) - b^* \phi(b^*) + \Phi(b^*) - \Phi(a^*)\} \right] + \alpha^2 \{ \Phi(b^*) - \Phi(a^*) \}$   
 $+ 2\alpha\sigma \{ \phi(a^*) - \phi(b^*) \} + 4 \cdot [1 - \Phi(b^*)]$

where  $a^* = \frac{1-\alpha}{\sqrt{\frac{\gamma \Sigma \gamma}{m}}}$  and  $b^* = \frac{2-\alpha}{\sqrt{\frac{\gamma \Sigma \gamma}{m}}}$

$$\sigma = \sqrt{\frac{\gamma \Sigma \gamma}{m}}$$

In both the above theorems,  $\phi(\cdot)$  and  $\Phi(\cdot)$  denote the p.d.f and c.d.f of a standard normal variable respectively.

**9. Efficiency of the Estimators**

It is naturally of interest to see how close these estimators are. Here, we briefly discuss this aspect with an empirical sample. The raw financial data can be transformed into circular data by using the method of wrapping (see, e.g., page 31 of Jammalamadaka and SenGupta (2001)). That is, for positive (linear) values, after dividing by  $2\pi$ , we take the remainder, while for negative (linear) values, we add  $2\pi$  to the remainder to produce the corresponding circular values in  $(0, 2\pi]$ . The fractional moment estimator, as suggested by Kozubowski (2001), for the Linnik distribution is valid when  $\alpha > 1$  and that, for wrapped stable distribution, as suggested by Kuruoglu (2001), needs iterative techniques. The properties of this estimator also need to be studied. The efficiency of the estimators obtained using the four methods has been carried out, as suggested by the referees, by including the estimated bias (through the mean bias) and the standard errors (through the root mean square errors) of the estimators in Tables 1a,b and 2a,b. A comparison of the performance of the truncated trigonometric moment estimator  $\alpha_{WS}^{ftm}$  is made with that of the characteristic function-based estimator  $\alpha_{WS}^{cf}$  of  $\alpha$  of wrapped stable distribution based on their mean bias and root mean square errors (RMSEs) for moderate sample sizes in Table 1a,b. In Table 1a,b, a simulation is performed for the values of  $\alpha_{WS}^{ftm}$  and  $\alpha_{WS}^{cf}$ , each with sample size  $n = 30, 50, 80$  and  $100$  when the skewness parameter  $\beta = 0$ . For each sample size  $n$ , 1000 replications are made. A similar simulation is performed in Table 2a,b for a comparison of the performance of the estimators of  $\alpha$  of the wrapped Linnik distribution. It can be observed from Tables 1a,b and 2a,b that the mean bias and the root mean square error of the truncated trigonometric moment estimator of  $\alpha$  is less than that of the characteristic function-based estimator for most sample sizes, indicating the efficiency of the former over the latter.

**Table 1.** (a) Data 1: Estimated bias (mean bias) and estimated standard error (RMSE) of the estimator of  $\alpha$  of wrapped stable distribution. (b) Data 2: Estimated bias (mean bias) and estimated standard error (RMSE) of the estimator of  $\alpha$  of wrapped stable distribution.

(a) Data 1				
Sample Size	Mean Bias ( $\alpha_{WS}^{ftm}$ )	Mean Bias ( $\alpha_{WS}^{cf}$ )	RMSE ( $\alpha_{WS}^{ftm}$ )	RMSE ( $\alpha_{WS}^{cf}$ )
30	0.175	0.383	0.498	0.6697
50	0.1215	0.429	0.4286	0.667
80	0.014	0.457	0.363	0.656
100	0.029	0.478	0.3475	0.650

**Table 1.** Cont.

<b>(b) Data 2</b>				
Sample Size	Mean Bias ( $\alpha_{WS}^{\hat{t}m}$ )	Mean Bias ( $\alpha_{WS}^{\hat{c}f}$ )	RMSE ( $\alpha_{WS}^{\hat{t}m}$ )	RMSE ( $\alpha_{WS}^{\hat{c}f}$ )
30	0.009	1.087	0.267	1.341
50	0.179	1.138	0.438	1.353
80	0.128	1.225	0.552	1.384
100	0.042	1.236	0.141	1.389

**Table 2.** (a) Data 1: Estimated bias (mean bias) and estimated standard error (RMSE) of the estimator of  $\alpha$  of wrapped Linnik distribution; (b) Data 2: Estimated bias (mean bias) and estimated standard error (RMSE) of the estimator of  $\alpha$  of wrapped Linnik distribution.

<b>(a) Data 1</b>				
Sample Size	Mean Bias ( $\alpha_{WL}^{\hat{t}m}$ )	Mean Bias ( $\alpha_{WL}^{\hat{c}f}$ )	RMSE ( $\alpha_{WL}^{\hat{t}m}$ )	RMSE ( $\alpha_{WL}^{\hat{c}f}$ )
30	0.491	0.287	0.812	0.583
50	0.058	0.215	0.058	0.482
80	0.190	0.201	0.396	0.451
100	0.191	0.188	0.392	0.425

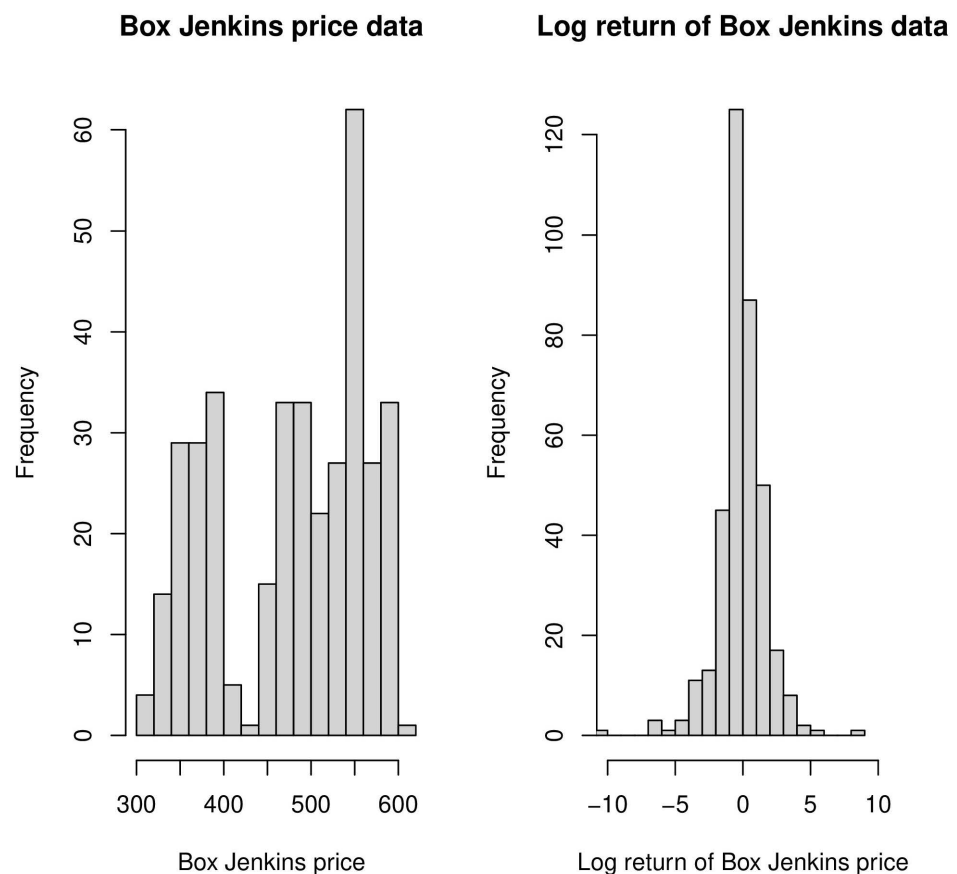
<b>(b) Data 1</b>				
Sample Size	Mean Bias ( $\alpha_{WL}^{\hat{t}m}$ )	Mean Bias ( $\alpha_{WL}^{\hat{c}f}$ )	RMSE ( $\alpha_{WL}^{\hat{t}m}$ )	RMSE ( $\alpha_{WL}^{\hat{c}f}$ )
30	0.085	0.478	0.641	0.682
50	0.034	0.483	0.565	0.664
80	0.017	0.519	0.478	0.664
100	0.013	0.552	0.428	0.666

## 10. Examples

In this section, we consider the wrapped stable and the wrapped Linnik densities as possible underlying models of the financial data, on the Box–Jenkins common stock closing price data of IBM taken from Box et al. (1976), with the characteristic function estimate and the truncated trigonometric moment estimate, respectively. Further financial data considered in this section, as an example, are the gold price data which were collected per ounce in US dollars over the years 1980–2008. Gold is an important asset to mankind and is hence important in financial market. Aggarwal and Lucey (2007) have suggested some statistical procedures which provide the existence of psychological barriers in daily gold prices and also in change of gold prices from day to day. The prices, being in round numbers, present an obstacle with important effects on the conditional mean and variance of the gold price series around psychological barriers. Mills (2004) studied the properties of the daily gold price from 1971 to 2002 and found them to be characterised by the presence of autocorrelation, volatility and 15-day scaling. The distribution of daily returns of gold is highly leptokurtic and multi-period returns attain normality only after 235 days. Byström (2020) studied the link between happiness and gold price changes. He observed that there is no significant correlation between happiness and gold price changes. However, assuming the tails of the happiness distribution to be non-normal, the gold price change seems to increase particularly on a person’s extremely unhappy days. However, the log returns (as in the analysis of stock data by Anderson and Arnold (1993)) data of the Indian gold market that we present here exhibit mild asymmetry, pronounced platykurtic and quite small first-order autocorrelation properties, which motivated us to study the symmetric Linnik distribution as an initial approximation of its distribution. The analysis of stock price data is generally carried out on a difference of order 1 in relation to the original series. So, denoting the original stock price data by  $x_t$ , they undergo transformation as

$z_t = 100(\ln(x_t) - \ln(x_{t-1}))$  which is then wrapped by the process as mentioned above. This transformation of log returns aims to achieve symmetry and reduce autocorrelation in the transformed series (for details, refer to SenGupta and Roy 2019, 2023). The Box–Jenkins data are denoted as data set 1, and the gold price data as data set 2, in the given tables. The computed estimates of  $\alpha$  are shown in Table 3. Note that the values of the estimators  $\hat{\alpha}$  by these two methods are quite different for each of the probability models. The values of the estimators are not comparable between the two families of distributions. However, within each family they determine a specific distribution. For example, an estimate of  $\alpha$  close to 1 indicates a Cauchy (wrapped Cauchy) distribution in the family of stable (wrapped stable) distributions, while an estimate of  $\alpha$  close to 2 indicates a Laplace (wrapped Laplace) in the family of Linnik (wrapped Linnik) distributions. With real life data sets, the use of these estimators can lead to quite different, possibly even contradictory, conclusions.

It can be observed from Figures 1 and 2 that the distribution of the log returns of the Box–Jenkins data is, while that of the gold price data is approximately symmetric with a certain amount of left skewness, whereas the gold price data are highly skewed in nature and the Box–Jenkins price data are bimodal. Still, we have used both the gold price and Box–Jenkins log return data sets as illustrations for our proposed estimators, as well as to explore their properties. We also note that both the methods of estimation based on trigonometric moments and characteristic function are not applicable to the two price data sets, since the underlying assumptions of the model are violated by the data sets.



**Figure 1.** Histograms of Box–Jenkins price data and their logarithm return data.

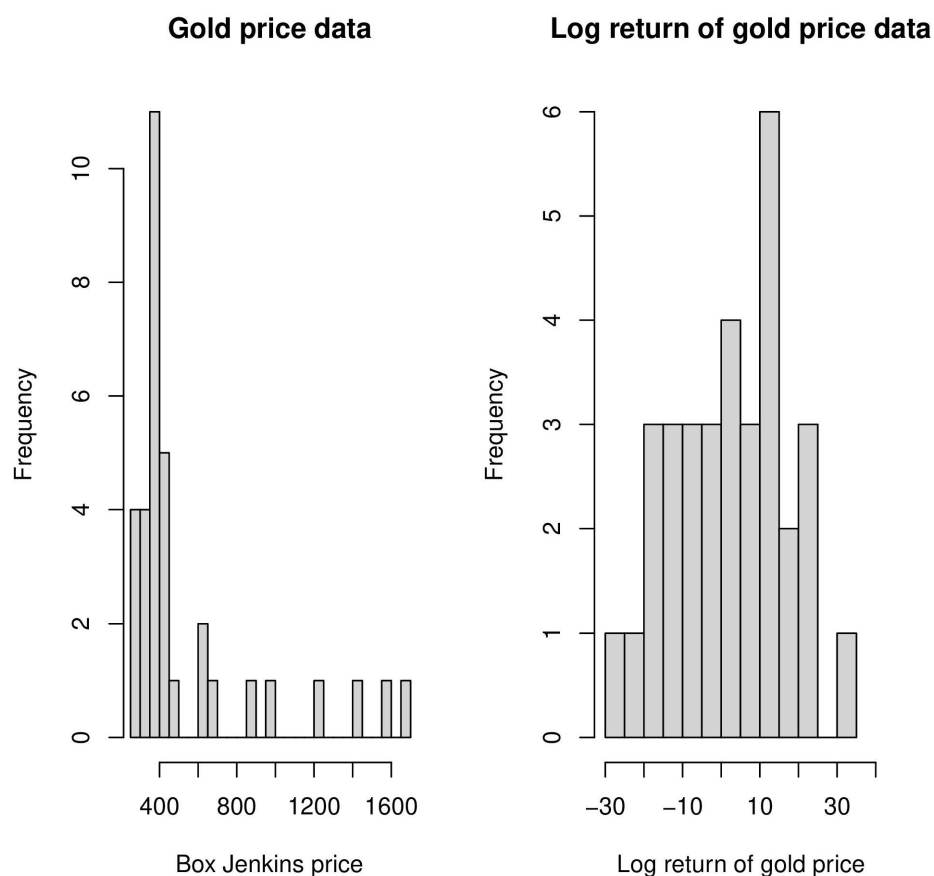


Figure 2. Histograms of gold price data and their logarithm return data.

Table 3. The estimates of  $\alpha$ .

Data	$\hat{\alpha}_{WS}^{ftm}$	$\hat{\alpha}_{WL}^{ftm}$	$\hat{\alpha}_{WS}^{cf}$	$\hat{\alpha}_{WL}^{cf}$
1	1.102854	1.941821	1.27487	2.0
2	0.3752206	1.263993	0.4149459	2.0

It can be observed from Table 3 that both the estimators are quite close to each other for the Box–Jenkins log return data, since they are symmetric in nature. The two estimators for the log return of the gold price data do not differ for wrapped stable distribution, but there seems to be an appreciable difference for wrapped Linnik distribution due to their differences in robustness against the asymmetric nature (e.g., the estimator of the location parameter by the mean and median give similar values for symmetric distribution but do not for asymmetric or skewed distribution, due to the difference in the robustness properties of the estimators). Thus, it is necessary that the assumptions of the symmetry of and independence in the data sets be verified in order to produce good estimates of the parameter by our proposed estimators as above.

### 11. Novel Tests for $\alpha$ Based on Circular Statistics

We are presenting here, to the best of our knowledge, the maiden attempt of testing for the index parameter of stable and Linnik distributions. Let  $x_1, x_2, \dots, x_n$  be realizations of symmetric stable ( $\mu = 0, \sigma = 1, \alpha$ ) distribution. The choice of  $\mu = 0$  and  $\sigma = 1$  are justified, as given in Section 3. When the sample size  $n$  is large, we can use the asymptotic distribution of  $\hat{\alpha}_{WS}^{ftm}$ , as stated in Theorem 5, to perform the test for the null hypothesis  $H_0 : \alpha = \alpha_0$ . Also, since the data have undergone logarithm ratio transformation, they are thus scale invariant and hence we can take the scale parameter  $\sigma = 1$  in the expression

of the estimator of the variance,  $\widehat{V(\alpha_{WS}^{ftm})}$  to perform the test. Thus, the test statistics are given by

$$\frac{\alpha_{WS}^{ftm} - \alpha_0}{\sqrt{\widehat{V(\alpha_{WS}^{ftm})}}} \rightarrow N(0, 1)$$

where  $\alpha_{WS}^{ftm}$  denotes the trigonometric truncated moment estimator of  $\alpha$  for the data, assuming a stable distribution.

Let  $x_1, x_2, \dots, x_n$  be realizations of symmetric Linnik ( $\mu = 0, \sigma = 1, \alpha$ ) distribution. When the sample size  $n$  is large, we can use the asymptotic distribution of  $\alpha_{WL}^{ftm}$ , as stated in Theorem 6, to perform the test for the null hypothesis  $H_0 : \alpha = \alpha_0$ . Also, since the data have undergone logarithm ratio transformation, they are thus scale invariant and hence we can take the scale parameter  $\sigma = 1$  in the expression of estimator of the variance,  $\widehat{V(\alpha_{WL}^{ftm})}$  to perform the test. Thus, the test statistics are given by

$$\frac{\alpha_{WL}^{ftm} - \alpha_0}{\sqrt{\widehat{V(\alpha_{WL}^{ftm})}}} \rightarrow N(0, 1)$$

where  $\alpha_{WL}^{ftm}$  denotes the trigonometric truncated moment estimator of  $\alpha$  for the data assuming the Linnik distribution. Depending on the alternative hypothesis, the cut-off points of the tests can be determined from standard normal distribution tables.

A similar test can also be carried out based on a Hill estimator using Lemma 1, but it is not studied here because the determination of  $k$  is complicated.

**Example:**

Anderson and Arnold (1993) have suggested the Linnik distribution for the financial data on Box–Jenkins based on their characteristic function-based method of estimation. We assume that the data come from a member of the Linnik family. In this family, the Linnik distribution is characterized by  $\alpha = 2$ . This has motivated us to rigorously verify their claim based on the corresponding test  $H_0 : \alpha = 2$  against the alternative hypothesis  $H_1 : \alpha < 2$ . As per the suggestions of the referee, we perform a test for Laplace (a.k.a. double exponential) distribution corresponding to  $\alpha = 2$  in the family of Linnik distributions. The test statistics as defined above are given by

$$\frac{\alpha_{WL}^{ftm} - 2}{\sqrt{\widehat{V(\alpha_{WL}^{ftm})}}}$$

The value of the test statistic is obtained as  $-0.3456217$ , implying that the null hypothesis of the claim of double exponential distribution is accepted both at the 5% (1.645) and 1% (2.326) levels of significance.

The Laplace distribution has been earlier used on an adhoc basis by Anderson and Arnold (1993) for the financial data on Box–Jenkins based on results of estimation. We have established it formally by providing rigorous proof through testing procedure which supports their findings.

**12. Discussions and Conclusions**

We have obtained a universal and efficient estimator of  $\alpha$  which can be easily implemented in practice. We have studied the various properties of the estimators, pointed out their drawbacks and also obtained improved estimators eliminating these drawbacks. We have also compared the efficiency of some estimators, as observed in the above Tables 1 and 2. We have also introduced a novel method of testing for the index

parameter of the stable and Linnik distributions. We thus hope that this maiden attempt will be useful for future analysis.

**Author Contributions:** Each author contributed to each section equally. All authors have read and agreed to the published version of the manuscript.

**Funding:** The research of the first author was supported, in part, by a CSIR Emeritus Scientist, Government of India, ES Grant No. 21 (1155)/22/EMR-II, from the Council of Scientific and Industrial Research. HRDG. EMR-INew Delhi Govt. of India.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The Box–Jenkins common stock closing price data of IBM were taken from Box et al. (1976). Regarding the gold price data, which were collected per ounce in US dollars over the years 1980–2008, the data from 1980–1994 were taken from Timothy Green’s Historical Gold PriceTable, World Gold Council and those from 1995–2008 were taken from the website www.kitco.com.

**Acknowledgments:** We thank the referees for their useful comments and, in particular, thank the referee who suggested we provide a statistical test for the index parameter. We also thank the Academic Editor for their useful comments and suggestions. The research of the first author was supported, in part, by a CSIR Emeritus Scientist, Government of India, ES Grant No. 21 (1155)/22/EMR-II, from the Council of Scientific and Industrial Research, HRDG, EMR-II, New Delhi, Govt. of India.

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

## References

- Aggarwal, Raj, and Brian M. Lucey. 2007. Psychological barriers in gold prices? *Review of Financial Economics* 16: 217–30. [CrossRef]
- Anderson, Dale N., and Barry C. Arnold. 1993. Linnik Distributions and Processes. *Journal of Applied Probability* 30: 330–40. [CrossRef]
- Box, George E. P., Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. 1976. *Time Series Analysis: Forecasting and Control*. Oakland: Holden-Day.
- Brilhante, M. Fátima, Maria Ivette Gomes, and Dinis Pestana. 2013. A simple generalization of the Hill estimator. *Computational Statistics and Data Analysis* 57: 518–35. [CrossRef]
- Byrd, Richard H., Peihuang Lu, Jorge Nocedal, and Ciyong Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* 16: 1190–208. [CrossRef]
- Byström, Hans. 2020. Happiness and Gold Prices. *Finance Research Letters* 35: 101599. [CrossRef]
- Dufour, Jean-Marie, and Jeong-Ryeol Kurz-Kim. 2010. Exact inference and optimal invariant estimation for the stability parameter of symmetric  $\alpha$ -stable distributions. *Journal of Empirical Finance* 17: 180–94. [CrossRef]
- Goldie, Charles M., and Richard L. Smith. 1987. Slow variation with remainder: A survey of the theory and its applications. *Quarterly Journal of Mathematics* 38: 45–71. [CrossRef]
- Hill, Bruce M. 1975. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics* 3: 1163–74. [CrossRef]
- Jammalamadaka, S. Rao, and Ashis SenGupta. 2001. *Topics in Circular Statistics*. Hackensack: World Scientific Publishers. [CrossRef]
- Kozubowski, Tomasz J. 2001. Fractional Moment Estimation of Linnik and Mittag-Leffler Parameters. *Mathematical and Computer Modelling* 34: 1023–35.
- Kuruoglu, Ercan E. 2001. Density Parameter Estimation of Skewed  $\alpha$  Stable Distributions. *IEEE Transactions on Signal Processing* 49: 2192–201. [CrossRef]
- Mills, Terence. 2004. Statistical analysis of daily gold price data. *Physica A: Statistical Mechanics and its Applications* 338: 559–66. [CrossRef]
- Pakes, Anthony G. 1998. Mixture representations for symmetric generalized Linnik laws. *Statistics & Probability Letters* 37: 213–21. [CrossRef]
- SenGupta, Ashis. 1996. Analysis of Directional Data in Agricultural Research using DDSTAP. In Proceedings of the Golden Jubilee International Conference of the Indian Society of Agricultural Statistics, New Delhi, India, December 19–21. pp. 43–59.
- SenGupta, Ashis, and Moumita Roy. 2019. An Universal, Simple, Circular Statistics-Based Estimator of  $\alpha$  for Symmetric Stable Family. *Journal of Risk and Financial Management* 12: 171.
- SenGupta, Ashis, and Moumita Roy. 2023. A Characteristic function based circular distribution family and its goodness of fit: The flexible wrapped Linnik. *Journal of Applied Statistics, revised paper submitted*. [CrossRef]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

# Accuracy Comparison between Five Machine Learning Algorithms for Financial Risk Evaluation

Haokun Dong, Rui Liu and Allan W. Tham \*

Faculty of Science and Technology, University of Canberra, Canberra 2617, Australia;  
haokun.dong@canberra.edu.au (H.D.); drruiiu@yeah.net (R.L.)

\* Correspondence: allan.tham@canberra.edu.au

**Abstract:** An accurate prediction of loan default is crucial in credit risk evaluation. A slight deviation from true accuracy can often cause financial losses to lending institutes. This study describes the non-parametric approach that compares five different machine learning classifiers combined with a focus on sufficiently large datasets. It presents the findings on various standard performance measures such as accuracy, precision, recall and F1 scores in addition to Receiver Operating Curve-Area Under Curve (ROC-AUC). In this study, various data pre-processing techniques including normalization and standardization, imputation of missing values and the handling of imbalanced data using SMOTE will be discussed and implemented. Also, the study examines the use of hyper-parameters in various classifiers. During the model construction phase, various pipelines feed data to the five machine learning classifiers, and the performance results obtained from the five machine learning classifiers are based on sampling with SMOTE or hyper-parameters versus without SMOTE and hyper-parameters. Each classifier is compared to another in terms of accuracy during training and prediction phase based on out-of-sample data. The 2 data sets used for this experiment contain 1000 and 30,000 observations, respectively, of which the training/testing ratio is 80:20. The comparative results show that random forest outperforms the other four classifiers both in training and actual prediction.

**Keywords:** financial data analysis; machine learning algorithms; loan default assessment; classification



**Citation:** Dong, Haokun, Rui Liu, and Allan W. Tham. 2024. Accuracy Comparison between Five Machine Learning Algorithms for Financial Risk Evaluation. *Journal of Risk and Financial Management* 17: 50. <https://doi.org/10.3390/jrfm17020050>

Academic Editor: Thanasis Stengos

Received: 1 October 2023

Revised: 19 January 2024

Accepted: 22 January 2024

Published: 29 January 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Financial institutions are facing increasing challenges in mitigating various kinds of risks. In his “taxonomy of risks”, Christoffersen (2011) defines risks as market volatility, liquidity, operational, credit and business risks. Due to uncertainties, financial risk evaluation (FRE) is increasingly playing a pivotal role in ensuring organizations maximize their profitability by minimizing losses due to a failure to mitigate risks. Noor and Abdalla (2014) argue that there is a direct negative impact on profitability in proportion to unmitigated risks. Hence, the primary approach of FRE is to identify risks in advance to allow for an appropriate course of action before any investments or decisions can be made. As financial risks evolve over time due to factors such as economic fluctuations, market conditions and other factors beyond control, the evaluation process requires constant update to keep up with market conditions.

Credit risk analysis undertaken in recent years mostly involves financial risk prediction. For example, loan default analysis, which often comes in the form of binary classification problems, has become an integral part of FRE. As financial institutions today are dealing with millions of customers, the traditional human approach for loan approval processes are no longer feasible. Moreover, with the advent of computing power today coupled with the advancement of machine learning algorithms and the availability of large volume of information, the world has entered the renaissance of computational modeling with non-parametric classification methods and machine learning for loan default prediction becoming widely adopted. It is worth noting that machine learning classification achieves



an accuracy that directly increases the bottom line whilst providing instantaneous approval decision through real-time decisions.

The use of advanced computing power and machine learning algorithms to predict whether a loan is performing or non-performing (NPL) is increasingly essential for the longevity of any lending institute. As the pool of consumers enlarges with proportional increases in spending, the ability to provide early warnings by accurately predicting the probability of defaulting a loan has become even more crucial. Deploying advanced machine learning algorithms to identify patterns from large features in high-volume data has become a mandatory process by banks to minimize the NPLs, and thus to increase their profitability and consumers' confidence.

In this study, we aim our attention at loan default detection as an element of credit risk analysis through models built in k-nearest neighbour, naïve-bayes, decision tree, logistic regression, and random forest. The intention is to answer the following questions:

- Is there a significant performance difference between the five machine learning algorithms piping through Scikit-learn data transformation steps?
- Can the steps be repeated with the same level of consistency using different data sets but similar analytics pipelines with data transformation?

The study is arranged in the following way:

- In Section 2, we briefly cover the background of the rise of statistical methods used from the 60s to the 80s, primarily in the form of parametric approaches in predicting bankruptcy. This section also covers how modern predictive techniques were born in the 90s and beyond in conjunction with the availability of computing power, resulting in the advancements of this field.
- In Section 3, we conduct a case study to illustrate the use of the five machine learning algorithms to predict the loan default based on University of California at Irvine (UCI) data set. In this section, we present the analytics life-cycle methodology with emphasis on data pre-processing. It highlights the repeatability and validity of the methodology in conducting research.
- In Section 4, we construct various models and measure them using various tools, of which ROC-AUC is the main measurement. Other measurements are accuracy, precision, F1 score, etc. In this section, we draw comparisons between five classifiers and present the results neatly in various tables. We also present the out-of-sample prediction results to validate model accuracy.

## 2. Related Work

It is imperative for financial institutes to detect NPLs in advance and segregate them for further treatments. Unlike today, however, the ability to predict NPLs in the 60s was not commonplace due to the fact that data mining and predictive capabilities were in their embryonic state. During that era, financial analysis using a quantitative approach was in its nascent form. Mathematical models and statistical methods were basic compared to modern quantitative techniques. Apart from relying on studying a company's financial statements, most financial risks analysis primarily relied on fundamental analysis which involves studying external factors such as market trends and economic indicators.

Beaver (1966) laid a foundation of groundbreaking work in accounting, earning himself management using financial ratios. "Beaver's Model" involved seminal univariate analysis to predict corporate failure. Altman (1968) devised the "Altman Z-Score" to predict the probability of whether a company will undergo bankruptcy. Beaver and Altman's work pioneered approaches to financial risk analysis for the next decade.

Finance-related prediction in the 1970s hinged on Altman's Z-Score, which had garnered popularity since the late 1960s. Although Altman's work primarily involved predicting bankruptcy, academics and researchers adapted the underlying principles to perform prediction of risks to maintain financial health. The 1970s marked the emergence of modern risk management concepts with financial institutes becoming aware of the importance of identifying and managing various risk portfolios. The 1970s laid the groundwork for iden-

tifying and understanding financial risk prediction and management. This development was the beginning of the evolution of risk assessment methodologies and the adoption of risk management practices together. The regulatory frameworks aimed to enhance stability and resilience were set up by regulatory bodies.

Black and Scholes (1973) developed the Black–Scholes–Merton (BSM) model in 1973 which aimed to calculate the theoretical price of European-style options. The model uses complex mathematical formulas and assumes standard normal distribution including logarithms, standard deviations (precursor to Z-Score) and cumulative distribution functions. The Black–Scholes–Merton model remains a foundation of today’s market risk assessment and serves as a fundamental tool for pricing options. Although specific research publications in the 1970s may not be common enough to be readily cited, many ideas, concepts and methodologies established during that timeframe set the stage for subsequent developments. Most notable is the gaining of traction of the quantitative approach to credit risk modeling and scoring. The rise of algorithms such as regression, discriminant analysis and logistic regression dominated the 1970s. The duo’s empirical results also demonstrated how efficient regulatory policy should be formulated from the regression outcomes. Deakin (1972), standing on the shoulders of Beaver and Altman, brought the analysis one notch higher using a more complex, albeit discriminatory, analysis to improve on the 20% error in misclassification of bankruptcy for the year prior. Deakin’s model of an early warning system assumed a random draw of samples and used various financial ratios and indicators including profitability ratios, efficiency ratios and liquidity ratios (amongst others) to distinguish between troubled and healthy firms. Martin (1977) leveraged the logit regression approach to predict the likelihood of banks experiencing financial distress.

The 1980s saw an increased focus on credit risk measurement within banking industries. Managing creditworthiness, credit exposures and the probabilities to default were key research topics by researchers and practitioners. Ohlson took interest of White and Turnbull’s unpublished work on systematically developed probabilistic estimates of failures. Ohlson (1980) used the maximum likelihood estimation methodology, which is a form of conditional logit model (logistic regression), to avoid the pitfall of well-known issues associated with multivariate discriminant analysis (MDA) deployed in previous studies. Ohlson’s model, primarily a parametric one (as most models were in that era), provided advantages in that no assumptions must be made to account for prior probabilities regarding bankruptcy and the distribution of predictors. Ohlson argued that Moody’s manual, as relied on by previous works, could be flawed due to the fact that numerous studies that derived financial ratios from the manual did not account for the timing of data availability and the complexity in reconstructing balance sheet information from the highly condensed report. In his concluding remark, Ohlson stated that the prediction power of any model depends upon when the financial information is assumed to be available. West (1985) combined the traditional parameter approach using a logit algorithm with factor analysis. West’s work was promising, as the empirical results show the combination of the two techniques closely matched the CAMEL rating system widely used by bank examiners in that era.

The 1990s and 2000s saw the birth of some exciting machine learning algorithms. Up until this point, most statistical methods used for credit assessment were related to the parametric approach. The parametric algorithms mandate that the assumptions of linearity, independence, or constant variance are met before meaningful analysis can be derived. The birth of Adaptive Boosting can be indebted to the work of Freund and Schapire (1997). The duo proposed that a strong classifier can be obtained by combining multiple weak classifiers iteratively. Friedman (2001) devised a method to improve the predictive accuracy by optimizing a loss function through iterative processes. Friedman’s gradient boosting machine (GBM) builds the trees sequentially, with each tree correcting by fitting the residuals of the previous trees. Friedman’s work was influential and subsequently gave rise to other boosting variations, including XGBoost by Chen and Guestrin (2016) and LightGBM by Ke et al. (2017). Breiman and Cutler (1993), however, proposed a way

to construct multiple independent decision trees during training, with each tree deriving from a subset of training data and available features. Breiman's (2001) random forest model ensures that each tree is trained on a bootstrap sample of data (random sample with replacement). The final prediction is made from aggregating the prediction from an ensemble of diverse decision trees. Vapnik and Chervonenkis' early work dated as far back as the early 1960s in theory of pattern recognition, and statistical learning laid the groundwork for their support vector machine (SVM). Vapnik's (1999) algorithm is known for the ability to classify both linear and non-linear data by finding the optimal hyperplane that best separates various classes whilst maximizing the margin between them.

Contemporary literature works in predicting financial risk has mushroomed over the past decade. Peng et al. (2011) suggest that a unique classification algorithm that could achieve the best accuracy given different measures under various circumstances does not exist. In their early attempts, Desai et al. (1996), and later West (2000), both proposed that the performance of generic models such as linear discriminant were not a better performer than customized models, except for a customized neural network. However, further studies by Yobas et al. (2000) using linear discriminant, neural network, genetic algorithms and decision tree concluded that the best performer was linear discriminant analysis. Due to the inconsistencies of previous studies, Peng et al. (2011) suggested multiple criteria decision making (MCDM), whereby a process to allow systematic ranking and selecting of an appropriate classifier or cluster of classifiers should be at the forefront of classification research. In the first ever academic study of Israeli mortgage, Feldman and Gross (2005) applied the simple yet powerful classification and regression tree (CART) to 3035 mortgage borrowers in Israel, including 33 features such as asset value, asset age, mortgage size, number of applicants, income, etc. The goal was to classify between potential defaulters and those unlikely to default. The distinct feature of CART that resulted in it being chosen over its primary competitors is its ability to manage missing data. Khandani et al. (2010) predicted the binary outcome that indicates whether an account is delinquent by 90 days by including the time dimension of 3-, 6- or 12-month windows. Using a proprietary dataset from a major bank, Khandani and others combined customer banking transactions (expenditures, savings and debt repayments), debt-to-income ratios and credit bureau data to improve the classification rates of credit card holders' delinquencies and defaults. CART was chosen as the non-parametric approach due to its ability to manage the non-linearity nature of data and inherent explainability of the algorithm. Their work proved that the time series properties of the machine learning prediction commensurate with realized delinquency rates, with  $R^2$  of 85%. He suggested assigning weight in training data as adaptive boosting to manage imbalanced class.

The rise of data gathering exercises made available hundreds or thousands of features compounded with imbalanced data, posing an issue for traditional approaches. The non-parametric approach burst onto the scene to manage the ever-increasing dimension, imbalanced data and the non-linear nature of models. The 2000s saw a rise of applying multi-layer neural networks and support vector machines (SVM) to financial prediction. Atiya (2001) proposed a non-parametric approach using a novel neural network model and was able to achieve accuracy of 3-year-ahead out-of-sample predictions between 81–85% accuracy. Zhang et al. (1999) suggested that artificial neural networks outperformed logistic regression. Huang et al. (2004) deployed backpropagation neural networks (BNN) and SVMs to achieve an accuracy of 80%.

Although the majority of datasets used for the studies are propriety in nature, there was little mention regarding the engagement of various data preparation techniques except from the recent study of the importance of data pre-processing effects on machine learning by Zelaya (2019) using the contemporary machine learning package such as Scikit-learn popularized by Pedregosa et al. (2011). The modern machine learning packages with full pipeline feature as shown by Varoquaux et al. (2015) are worth exploring. Equally omitted is the implementation of techniques such as SMOTE to manage imbalanced class, as proposed by Fernández et al. (2018), which is also worth further study.

In this study, we aim our attention at loan default detection as an element of credit risk analysis.

### 3. Case Study—Advanced Machine Learnings for Financial Risk Mitigation

#### 3.1. Methodology—Computational Approach

In this study, the machine learning analytics cycle use Scikit-learn packages to implement an analytics pipeline that includes data collection, data pre-processing, model constructions and model performance comparisons. Matplotlib supplies graphing capability to allow for the visual analysis of data.

Scikit-learn allows for the full analytics pipeline to specifically unravel the underlying pattern in data sets, therefore resulting in the best fitting for various classifiers. The pipeline contains end-to-end processes that performs these tasks: (i) ingest the data sets and perform preliminary data analysis to identify missing values, outliers and imbalanced class—any missing values will be imputed and imbalanced data is identified; (ii) standardize data which includes scaling and normalization to ensure consistent model performance; (iii) encode categorical (nominal and ordinal) and one-hot-encode for predictors and label for target variable; (iv) select top N most influential predictors and reduce total dimension to the influential ones; (v) cross validate using k-fold stratified to ensure the ratio of imbalance remains intact and subsequently treated by SMOTE as suggested by Chawla et al. (2002); (vi) train and fit data using various distance- and tree-based classifiers; (vii) compare the final performance measurements and report the most effective hyper-parameters.

Figure 1 illustrates the machine learning analytics life cycle implemented as an end-to-end analytics pipeline using Scikit-learn’s pipeline capability.

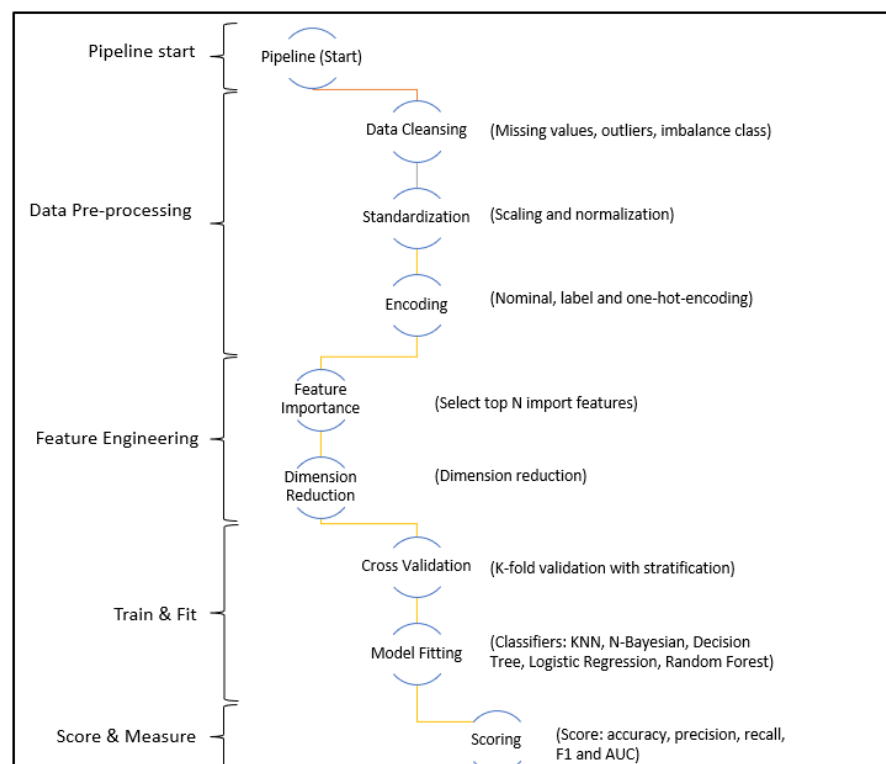


Figure 1. End-to-end analytics pipeline for machine learning classification.

The machine learning lifecycle is implemented as Scikit-learn’s pipeline, easing the foremost data pre-processing in missing value imputation with either the most frequent value (categorical features) or standard mean/median (numeric variable). The analytics pipeline detects outliers and imbalanced class, as well as manages the treatment of detection further down the pipeline right before the actual model fitting. Next is to apply data

standardization, which includes transforming the data into a common scale using Z-Score, and normalize the data to a range between zero and one. The goal is to ensure data consistency across various classifiers which will result in comparisons at similar scales, thus improving model performance.

Subsequently, the analytics pipeline automatically detects the champion model (winner of the best classifier) and reports the top N predictors that are most influential to the model. The analytics pipeline finds the least influential predictors which subsequently truncated to reduce the dimension whilst not affecting the performance of the models. The analytics pipelines split the data into two sections with training and testing data segregated by a ratio of 80:20. The analytics pipeline implements k-fold with stratification to ensure that the imbalanced class stays intact. It also ensures a full data split throughout with little-to-no possibility of a data leak. Finally, it trains and fits the data through the five classifiers. At the end, it obtains the performance scores for final comparisons.

Apart from its stochastic nature, the research method is sound and repeatable, and researchers can refer to it for further studies with various data sets applied to different classifiers.

### 3.2. Data Collection

This study uses two credit card client data sets obtained from UCI repository.

The first set is the payment data set obtained from one major bank in Taiwan from 2005, donated by Yeh and Lien (2009) and Yeh (2016) to the UCI data repository. The data set holds 30,000 observations, of which 6636 are default payment (showed by variable id, x24 as 1) whilst healthy payment occupies the remaining 23,364 observations. The data set holds no duplicate and missing values. The Taiwan credit card payment data set shows a strong skew (healthy:default ratio) due to imbalanced class of 77.88% to 22.12%. The variable id, x24 is the target whilst it uses the remaining features (x1 to x23) as predictors (Table 1).

**Table 1.** Dataset 1: Taiwan credit card client data set features and types.

Total Missing Values	Taiwan Credit Data Set Features		
	Feature ID/Name	Description	Numeric/ Nominal/Ordinal
0	x1 (limit_bal)	Amount of the given credit (NT dollar): includes both the individual consumer credit and his/her family (supplementary) credit	Numeric
0	x2 (sex)	Gender (1 = male, 2 = female).	Numeric
0	x3 (education)	Education (1 = graduate school, 2 = university, 3 = high school, 4 = others)	Numeric
0	x4 (marriage)	Marital status (1 = married, 2 = single, 3 = others)	Numeric
0	x5 (age)	Age (year)	Numeric
0	x6–x11 (pay_1 to pay_6)	History of past payment. We tracked the past monthly payment records (from April to September 2005) as follows: X6 = the repayment status in September 2005, X7 = the repayment status in August 2005,; X11 = the repayment status in April 2005. The measurement scale for the repayment status is: −1 = pay duly, 1 = payment delay for one month, 2 = payment delay for two months,; 8 = payment delay for eight months, 9 = payment delay for nine months and above.	Numeric

Table 1. Cont.

Total Missing Values	Taiwan Credit Data Set Features		
	Feature ID/Name	Description	Numeric/Nominal/Ordinal
0	x12–x17 (bill_amt1 to bill_amt6)	Amount of bill statement (NT dollar). X12 = amount of bill statement in September 2005, X13 = amount of bill statement in August 2005,; X17 = amount of bill statement in April, 2005	Numeric
0	x18–x23 (pay_amt1 to pay_amt6)	Amount of previous payment (NT dollar). X18 = amount paid in September 2005, X19 = amount paid in August 2005,; X23 = amount paid in April 2005	Numeric
0	x24 (default_payment_next_month)	Default or not (default = 1, health = 0)	Numeric

This data set contains only numeric features. It is used as a control data set for the analytics pipeline due to its larger set of observations. It will be used to validate the analytics pipeline that includes data transformations.

The second set is a German credit card client data set obtained from UCI data repository, Hofmann (1994). It contains 1000 observations. The data set contains one target variable with an imbalanced class ratio of 70% to 30% (no:yes ratio). The data set is void of missing values and duplicates. Table 2 shows the data set features, description and data types.

Table 2. Data Set 2: German credit card client data set features and types.

Total Missing Values	German Credit Data Set Features		
	Feature	Description	Numeric/Nominal/Ordinal
0	checking_balance	Status of existing checking account	Ordinal
0	months_loan_duration	Duration in months	Numeric
0	credit_history	Credit history	Ordinal
0	purpose	Purpose of loan	Nominal
0	amount	Credit amount	Numeric
0	savings_balance	Saving accounts/bonds	Ordinal
0	employment_duration	Present employment since	Ordinal
0	percent_of_income	Install rate (% of disposable income)	Numeric
0	years_at_residence	Present residence since	Numeric
0	age	Age in years	Numeric
0	other_credit	Other installment plans	Nominal
0	housing	Housing Situation	Nominal
0	existing_loans_count	Number of existing credits	Numeric
0	job	Job skill level	Ordinal
0	dependents	Number of dependents	Numeric
0	phone	Holding Telephone or not	Nominal
0	default	Default or not	Nominal

This data set contains both numerical and categorical data and is used to train and test various classifiers initially.

### 3.3. Visual Data Exploration

Either a classifier is parametric or non-parametric. Visual data exploration aids in understanding data structure and nature, which includes the data distributions, correlations, multi-collinearity and other patterns. Visual data exploration helps to identify anomalies and outliers in the data set that can skew analysis and model accuracy. In particular, the

involvement of logistic regression and naïve-bayes necessitate a thorough analysis of data structure and patterns as these classifiers assume independence and linearity, amongst other things. Table 3 reveals the correlation between numerical features.

**Table 3.** Correlation between numerical features—German credit data set.

		Correction Matrix					
	C1	C2	C3	C4	C5	C6	C7
C1	1.0000	0.6250	0.0747	0.0341	−0.0361	−0.0113	−0.0238
C2	0.6250	1.0000	−0.2713	0.0289	0.0327	0.0208	0.0171
C3	0.0747	−0.2713	1.0000	0.0493	0.0583	0.0217	−0.0712
C4	0.0341	0.0289	0.0493	1.0000	0.2664	0.0896	0.0426
C5	−0.0361	0.0327	0.0583	0.2664	1.0000	0.1493	0.1182
C6	−0.0113	0.0208	0.0217	0.0896	0.1493	1.0000	0.1097
C7	−0.0238	0.0171	−0.0712	0.0426	0.1182	0.1097	1.0000

Note: C1—months\_loan\_duration, C2—amount, C3—percent\_of\_income, C4—years\_at\_residence, C5—age, C6—existing\_loans\_count, C7—dependents.

As indicated in Table 3, the German credit data set has a low correlation between features. The highest correlation is 0.6250 between “months loan duration” and “amount”. It can be said that the correlation amongst other features is non-existent as indicated by scatterplots in Figure 2. The only correlation of 0.6250 shows a positively trending linear relationship. However, “age” and “percent of income” do not show a visual pattern and therefore do not indicate a relationship with “months loan duration.”



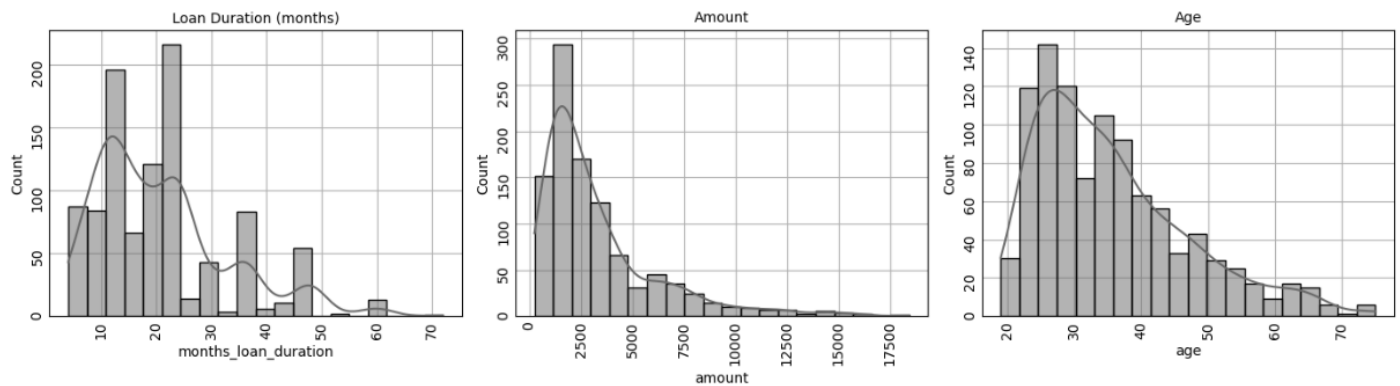
**Figure 2.** Scatterplots for selected numeric features.

Further investigation into multicollinearity of the German credit data set (Table 4) shows that the variance inflation factor (VIF) values between predictors are reasonable and do not cause alarm.

**Table 4.** VIF—German credit data set.

Feature	VIF
months_loan_duration	7.3588
amount	4.5758
percent_of_income	7.9587
years_at_residence	7.7354
age	10.9257
existing_loans_count	6.6793
dependents	8.7906

The distribution of data can be considered partial-normal or normal (right skewed) for only three predictors, as seen in Figure 3. The remaining numerical predictors are dichotomous in nature.



**Figure 3.** Data distributions—German credit data set.

The final examination confirms that the German credit data set does not contain outliers or missing values.

### 3.4. Data Pre-Processing

The analytics pipelines correspond in a 1:1 ratio with the permutations of the test scenarios. The first analytics pipeline generates Permutation-1, the second generates Permutation-2, etc. The goal of having various permutations is to achieve the best model performance for the classifiers.

The analytics pipelines apply standardization, including scaling and normalization to all analytics pipelines in this study. For example, the k-nearest neighbour being a distance-based classifier requires that the features contribute more equally to distance calculation, therefore enhancing model performance. The pipelines ensure that there is no unintentional data leakage between the training and testing data sets.

After standardization, the analytics pipelines perform encoding (ordinal, one-hot and label encodings) for categorical features and class feature followed by imbalanced data treatment using SMOTE as investigated by Alam et al. (2020). Subsequently, the analytics pipelines reduce the dimensions of features to the system default and a preset number, respectively.

Prior to model training and fitting, the analytics pipelines implement a manual split of training/testing (80:20) data sets with stratification to ensure the imbalanced data ratio is intact. In search for the optimal hyper-parameters, the grid search function performs the 10-fold cross validations where data is split and internally evaluated for each fold.

### 3.5. Model Construction and Evaluation

During the model construction phase, the analytics pipeline includes the five classifiers (k-nearest neighbour, naïve-bayes, decision tree, logistic regression and random forest). The final prepared and split data, after being fully cleansed, standardized and encoded, has become a training source to fit the models. The only distance-based classifier used in this study is k-nearest neighbour. K-nearest neighbour is a simple, non-parametric classifier that is not subservient to the Gaussian distributions and is robust to outliers. The curse of dimension takes effect with k-nearest neighbour in that it poses two challenges: (i) it increases computational challenges when high dimensions and large data sets are involved and (ii) it degrades model accuracy when including irrelevant features.

The two tree-based methods are decision tree and random forest. Similar to k-nearest neighbour, decision tree is a tree-based classifier and can manage non-linearity and outliers well. Decision tree has an inherent ability to be unaffected by non-related features. However,



the downside is that it tends to overfit. In this study, the analytics pipeline considers the tree-depth hyper-parameter to ensure that the decision tree classifier does not overfit. Random forest as an ensemble method inherits the strength from decision tree. Additionally, unlike decision tree, it aggregates the prediction of multiple decision trees and offsets the tendency to overfit.

Logistic regression and naïve-bayes are the only two parametric approaches used in this study. That said, they are susceptible to independence assumptions, non-linearity and outliers. In addition, imbalanced data affects naïve-bayes predictions. The analytics pipeline includes the data pre-processing to ensure the training is conducive to fit using the five classifiers, in particular, the parametric ones.

Table 5 illustrates the characteristics of the classifiers implemented in this study.

**Table 5.** High level characteristics of classifiers.

Classifier	Type	Dependence (H/L) and Tolerance (H/L) for Various Characteristics								
		Dependence					Tolerance			
		C1	C2	C3	C4	C5	C6	C7	C8	C9
k-nearest neighbour (knn)	NP—DB	L	L	L	L	L	L	L	L	B
naïve-bayes (nb)	P—PB	L	H	L	L	H	H	L	H	S
decision tree (dt)	NP—TB	L	L	L	L	H	H	H	L	S
logistic reg. (lr)	P—PB	L	H	L	H	L	L	L	L	B
random forest (rf)	NP—TB	L	L	L	L	H	H	H	L	S

Note: L—low, H—high. P—parametric, NP—non-parametric. DB—distance-based, TB—tree-based, PB—probability-based. B—bigger size data, S—smaller size data. C1—normality, C2—independence, C3—homoscedasticity, C4—linearity. C5—outliers, C6—multicollinearity, C7—irrelevant features. C8—imbalanced class, C9—minimum sample size required for stable estimates.

Whilst these dependency and tolerance level are common in the statistical and machine learning techniques, not all analyses require these assumptions to be met. Under certain conditions, there are methods to relax these dependencies and increase tolerance for the classifiers. Blatant ignorance of the requirements based on each classifier’s characteristic will result in poor and unreliable models.

As far as model measurement is concerned, Han et al. (2022) outlined the limitations of relying only on the rate of error as the default measurement as suggested by Jain et al. (2000) and Nelson et al. (2003). Since most of dataset one (Taiwan credit card client) is made up of non-risky value (77.88%), the error rate measurement is not appropriate as it is insensitive to the classification accuracy. The main measurement in this study is AUC despite the fact that Lobo et al. (2008) asserted that area under the receiver operating characteristic (ROC) curve, known as AUC, has its own limitations. Furthermore, the study includes error rate measurement which includes accuracy, precision, recall (sensitivity) and F1 score for the sake of completeness.

Table 2 shows the four error rate-related measurements in model evaluations:

- Accuracy provides the proportion of correctly classified instances from the total instances.
- Precision provides the ratio of true positive predictions versus the total number of positive predictions made.
- Recall provides the proportion of actual positives correctly predicted by the model.
- F1 provides a balance mean of precision and recall that deals with imbalanced class.

Table 6 depicts the performance metrics used throughout the model comparisons.

**Table 6.** Performance measurements.

Evaluation Metrics	Formula
Accuracy Score	$\frac{TP+TN}{TP+FN+TN+FP}$
Precision Score	$\frac{TP}{TP+FP}$
Recall Score (Sensitivity)	$\frac{TP}{TP+FN}$
F1 Score	$\frac{2*(Precision*Recall)}{(Precision+Recall)}$

Note: *N*—sample size, *TP*—true positive, *FN*—false negative, *TN*—true negative, *FP*—false positive.

All the scores used in this study are based on prediction results.

#### 4. Results

The results of the experiments are made up of performance metrics in tabular and graph formats as well as variables of importance in graph format. The results compare the performances based on the two distinct data sets with three permutations of analytics pipelines. Due to the stochastic nature of classifiers, the results differ slightly for each run. The red dotted line for ROC-AUC graphs represents a random guess for random guess.

The analytics pipelines produce three permutations in search of the best performing classifiers with their respective hyper-parameters. Tables 7–9 and Figures 4–6 list the performance metrics for various permutations. All permutations include data cleansing and standardization:

- Permutation-1—with or without SMOTE using the default hyper-parameters and scoring using full features (all predictors).
- Permutation-2—with or without SMOTE using the best hyper-parameters and scoring using full features (all predictors).
- Permutation-3—with or without SMOTE using the best hyper-parameters and scoring using reduced features (best performing predictors).

**Table 7.** German credit card dataset—Permutation-1.

	Default Hyper-Parameters									
	Full Features (without SMOTE)					Full Features (with SMOTE)				
	M1	M2	M3	M4	M5	M1	M2	M3	M4	M5
knn	0.7000	0.5000	0.3333	0.4000	0.6746	0.6550	0.4536	0.7333	0.5605	0.7073
nb	0.6950	0.4912	0.4667	0.4786	0.7305	0.7100	0.5111	0.7667	0.6133	0.7408
dt	0.6550	0.4211	0.4000	0.4103	0.5821	0.6450	0.4068	0.4000	0.4034	0.5750
lr	0.7500	0.6250	0.4167	0.5000	0.7679	0.7300	0.5366	0.7333	0.6197	0.7630
rf	0.7800	0.7222	0.4333	0.5417	0.7748	0.7500	0.6087	0.4667	0.5283	0.7618

Note: M1—accuracy score, M2—precision score, M3—recall score, M4—F1 score, M5—AUC.

**Table 8.** German credit card dataset—Permutation-2.

	Best Hyper-Parameters									
	Full Features (without SMOTE)					Full Features (with SMOTE)				
	M1	M2	M3	M4	M5	M1	M2	M3	M4	M5
knn	0.7700	0.6667	0.4667	0.5490	0.8227	0.8550	0.6824	0.9667	0.8000	0.9670
nb	0.7200	0.5303	0.5833	0.5556	0.7420	0.7300	0.5333	0.8000	0.6400	0.7614
dt	0.7650	0.6857	0.4000	0.5053	0.7726	0.8250	0.6667	0.8333	0.7407	0.9115
lr	0.7750	0.6596	0.5167	0.5794	0.7979	0.7350	0.5412	0.7667	0.6345	0.7944
rf	0.8900	0.9318	0.6833	0.7885	0.9868	0.8900	0.8519	0.7667	0.8070	0.9457

Note: M1—accuracy score, M2—precision score, M3—recall score, M4—F1 score, M5—AUC.

Table 9. German credit card dataset—Permutation-3.

Best Hyper-Parameters										
	Reduced Features (without SMOTE)					Reduced Features (with SMOTE)				
	M1	M2	M3	M4	M5	M1	M2	M3	M4	M5
knn	0.765	0.6857	0.4000	0.5053	0.8221	0.770	0.5761	0.8833	0.6974	0.8801
nb	0.700	0.5000	0.1833	0.2683	0.6912	0.635	0.4253	0.6167	0.5034	0.6792
dt	0.720	0.5909	0.2167	0.3171	0.7274	0.770	0.6094	0.6500	0.6290	0.8206
lr	0.695	0.4762	0.1667	0.2469	0.6855	0.635	0.4316	0.6833	0.5290	0.6879
rf	0.875	0.9730	0.6000	0.7423	0.9677	0.835	0.6957	0.8000	0.7442	0.9141

Note: M1—accuracy score, M2—precision score, M3—recall score, M4—F1 score, M5—AUC.

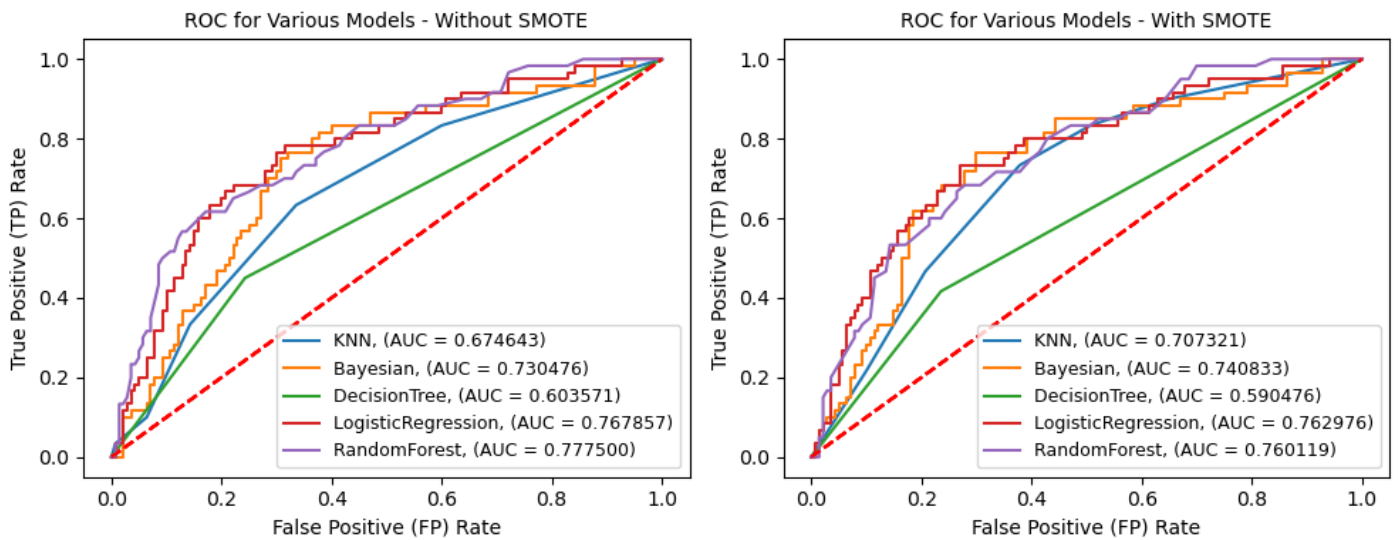


Figure 4. ROC for default hyper-parameters (full features)—German credit data set (Permutation-1).

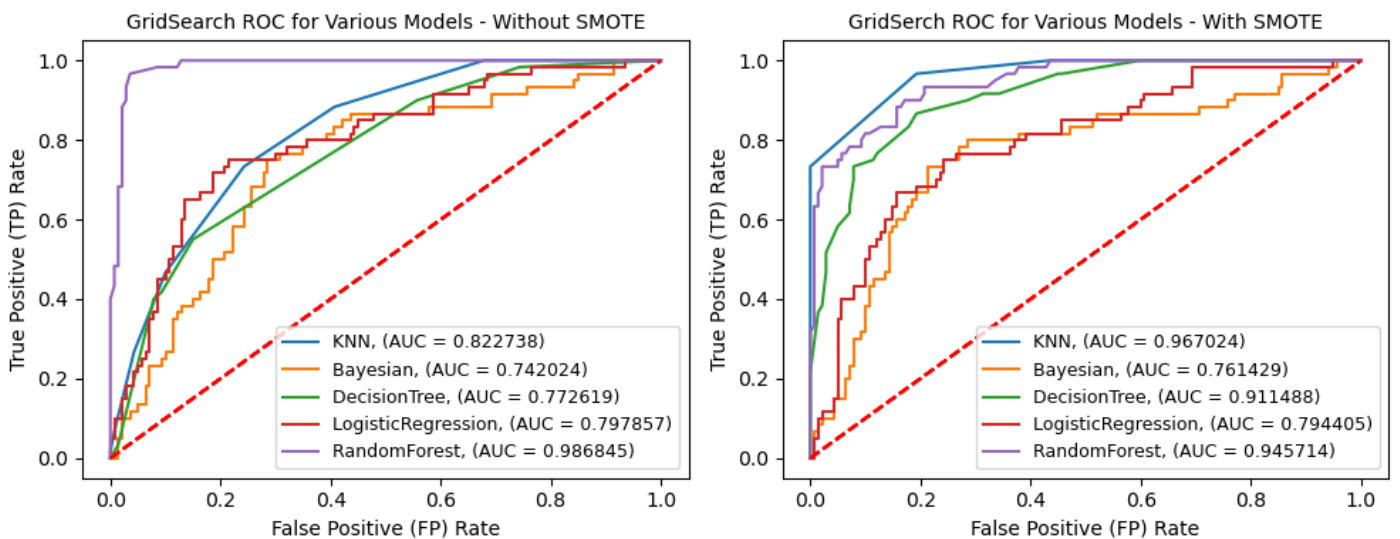


Figure 5. ROC for best hyper-parameters (full features)—German credit data set (Permutation-2).

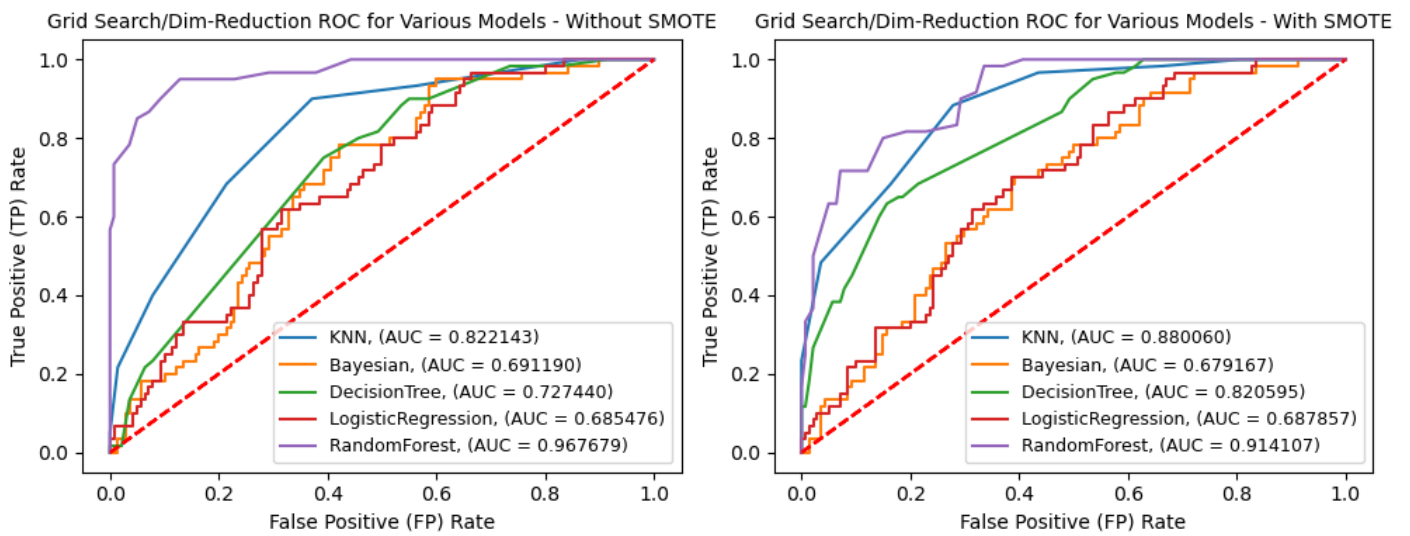


Figure 6. ROC for best hyper-parameters (reduced features)—German credit data set (Permutation-3).

The tables and figures summarize the results:

- Table 7 and Figure 4—Permutation-1 by using German credit data set with the default hyper-parameters and full features.
- Table 8, Figures 5 and 7—Permutation-2 by using German credit data set with the best hyper-parameters and full features.
- Table 9, Figures 6 and 8—Permutation-3 by using German credit data set with the best hyper-parameters and reduced features.
- Table 10, Figures 9 and 10—Permutation-2 using Taiwan credit data set with the best hyper-parameters and full features.

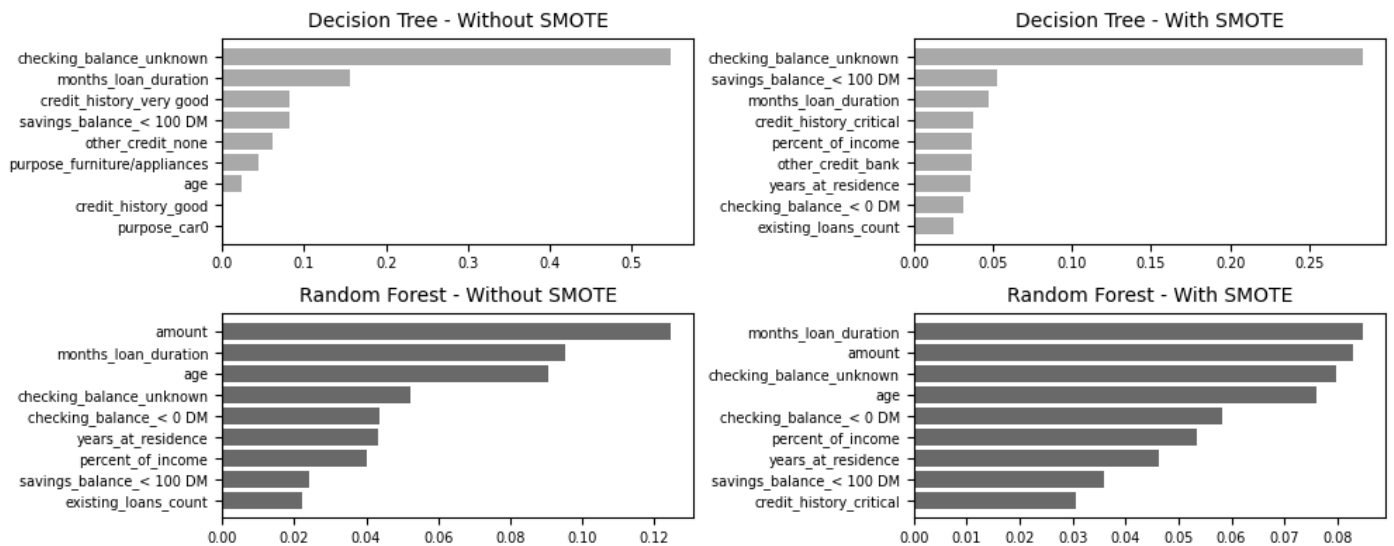
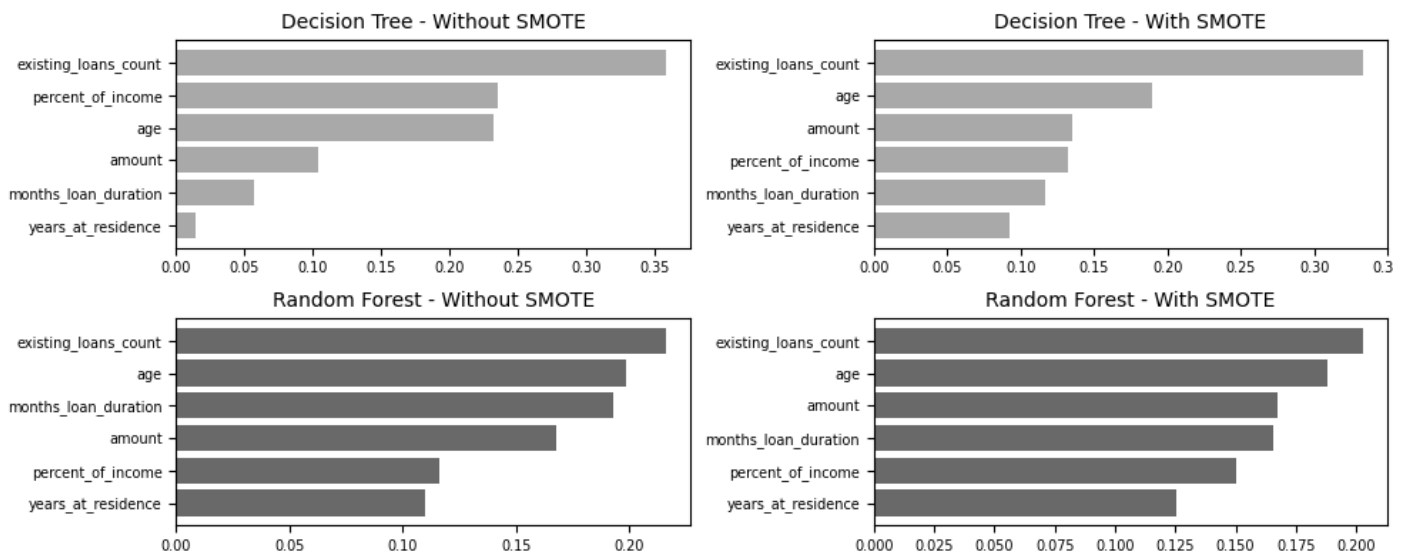
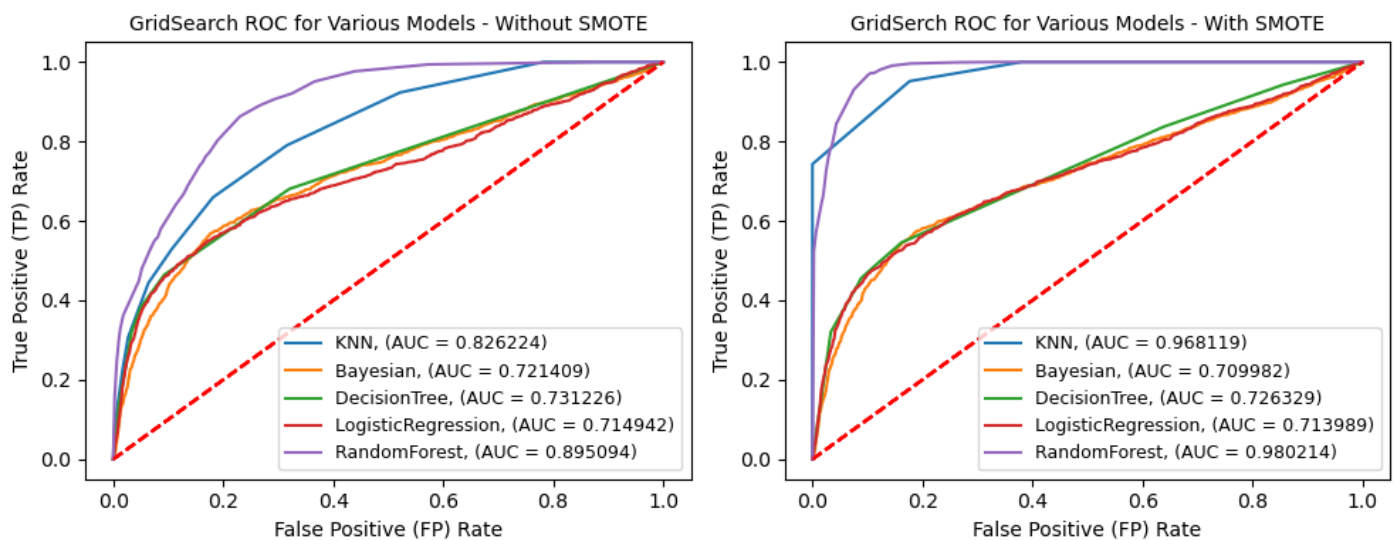


Figure 7. Variable of importance for best hyper-parameters (full features)—German credit data set (Permutation-2).



**Figure 8.** Variable of importance for best hyper-parameters (reduced features)—German credit data set (Permutation-3).

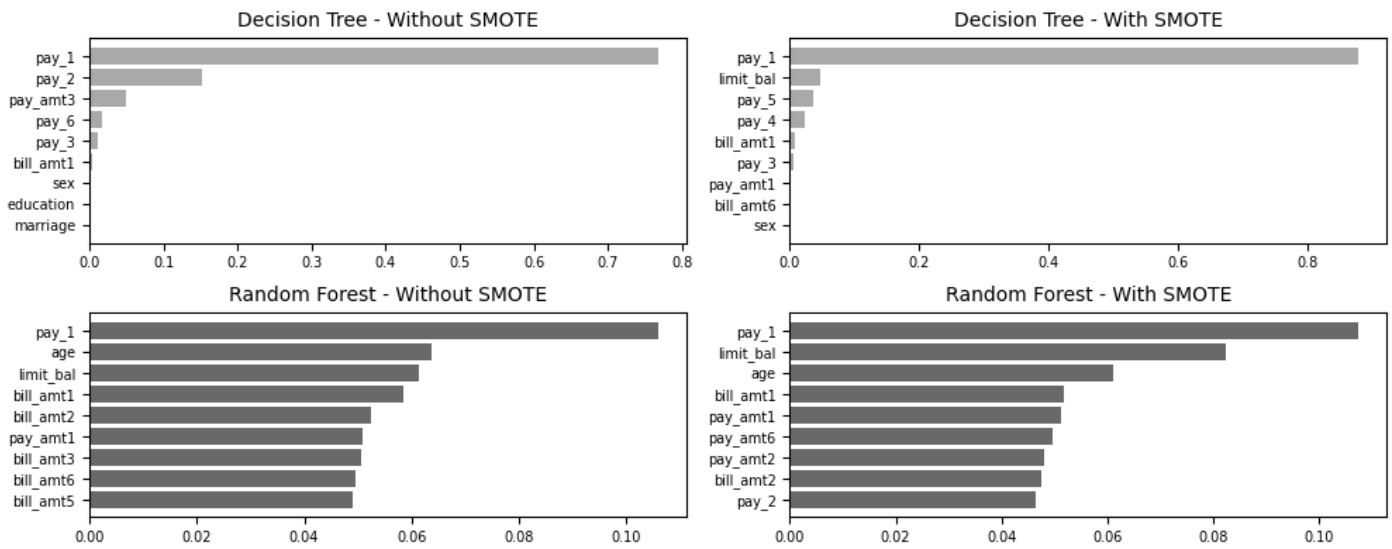


**Figure 9.** ROC for best hyper-parameters (full features)—Taiwan credit data set.

**Table 10.** Taiwan credit card dataset—performance measurement.

Best Hyper-Parameters										
	Full Features (without SMOTE)					Full Features (with SMOTE)				
	M1	M2	M3	M4	M5	M1	M2	M3	M4	M5
knn	0.8265	0.7061	0.3693	0.4849	0.8262	0.8517	0.6046	0.9518	0.7395	0.9681
nb	0.7485	0.4479	0.5893	0.5089	0.7214	0.3915	0.2484	0.8644	0.3859	0.7100
dt	0.8243	0.6842	0.3821	0.4903	0.7312	0.7733	0.4889	0.5456	0.5157	0.7263
lr	0.8135	0.7441	0.2389	0.3617	0.7149	0.6860	0.3759	0.6360	0.4726	0.7140
rf	0.8433	0.7706	0.4152	0.5397	0.8951	0.9292	0.8292	0.8561	0.8424	0.9802

Note: M1—accuracy score, M2—precision score, M3—recall score, M4—F1 score, M5—AUC.



**Figure 10.** Variable of importance for best hyper-parameters (full features)—Taiwan credit data set.

The best hyper-parameters (with or without SMOTE) obtained for various classifiers can be seen below:

- K-nearest neighbour—{‘kn\_n\_neighbours’: 7}
- Naïve-bayes—{‘nb\_priors’: None, ‘nb\_var\_smoothing’:  $1 \times 10^{-9}$ }
- Decision tree—{‘dt\_max\_depth’: 3, ‘dt\_splitter’: ‘best’}
- Logistic regression—{‘lr\_C’: 100, ‘lr\_max\_iter’: 1000}
- Random forest—{‘rf\_max\_features’: ‘sqrt’, ‘rf\_max\_samples’: 0.3, ‘rf\_n\_estimators’: 100}

The three permutations are graphical depictions of the performance metrics for various classifiers.

The last two permutations identify most significant predictors (features of importance) for decision tree and random forest classifiers. The other classifiers produce comparable results.

This study also involves a control data set (Taiwan credit card client) with larger data (30,000 observations). The same analytics pipelines containing data transformations are applied to the data with an identical split ratio, namely 80:20. The results of searching for the best hyper-parameters with full features, ROC graph and most influential predictors can be seen in Table 7, Figures 9 and 10 respectively.

First, it is observed that the default hyper-parameters perform poorly in both smaller (Table 7) and bigger data (Table 10) sets. For example, the five metrics (accuracy, precision, recall, F1 and AUC) hover below 0.8000 for the German credit data set. This shows that the default hyper-parameters are not sufficiently tuned to uncover the hidden pattern in both data sets. Using AUC as a more robust measurement, it is shown that k-nearest neighbour and decision tree are the two worst performing classifiers with 0.6746 and 0.5821, respectively, with untreated and imbalanced data. Interestingly, none of the five classifiers perform better when imbalanced data is treated with SMOTE. Only k-nearest neighbour improves marginally.

Much can be said regarding the full features from the German credit data set being subjected to the best hyper-parameters search. Table 8 shows vast improvements for all five classifiers. Without SMOTE, naïve-bayes is the worst performing classifier with modest improvement alongside logistic regression. However, k-nearest neighbour, decision tree and random forest improve greatly. With imbalanced treatment, k-nearest neighbour improves further. However, the greatest improvement is decision tree which jumps from 0.7726 to 0.9115 followed by k-nearest neighbour which leaps from 0.8227 to 0.9670. What is worth noting, however, is that random forest degrades slightly from 0.9868 to 0.9457. In Permutation-2, both naïve-bayes and logistic regression are indifferent regardless of the inclusion imbalanced data treatment in data pre-processing. The relevant features check

(Figure 7) using the built-in features for decision tree and random forest show the few key features are primarily between “checking balance”, “amount”, “months loan duration”, “age”, “percent of income” and “years of residence.”

Permutation-3 (Table 9, Figures 6 and 8) differs with Permutation-2 in that it further reduces most relevant features from nine to five, where Permutation-2’s nine features are selected by system whilst Permutation-3 is configured to take the best five features. The results are consistent as naïve-bayes and logistic regression are indifferent to imbalanced data treatment whilst k-nearest neighbour and decision tree show big improvements from 0.8221 to 0.8801 and 0.7274 to 0.8206, respectively. The performance for random forest, however, degrades slightly from 0.9677 to 0.9141. In general, the performance of models using the five most relevant features are less optimal than the nine selected by the system.

Finally, comparing with a larger Taiwan credit data set and Permutation-2 (winner), Table 10 and Figures 9 and 10 show that k-nearest neighbour and random forest are the two best performing classifiers across the two data sets. Decision tree performs well in the smaller German credit data set but worse when data is on a larger scale, as in the Taiwan credit data set. Naïve-bayes and logistic regression are indifferent to either smaller or larger data sets, with or without imbalanced data treatment.

## 5. Discussion

The three analytics pipeline permutations used to construct the five models based on five classifiers contain data cleansing and standardization. It is worth noting that the German credit data set contains categorical features and class labels that require encoding, whilst the Taiwan credit data set contains only numeric features.

The main observations and possible explanations of model performance can be summarized as follows:

- Using the default hyper-parameters for the five classifiers does not necessarily produce the best performing metrics. As data sets have distinctive characteristics such as total observations, complexity and patterns, it is rarely the best practice to use the default hyper-parameters settings until certain tuning is implemented based on each data set.
- K-nearest neighbour and random forest perform consistently well across both data sets either with or without imbalanced data treatment. However, k-nearest neighbour’s execution time is a great magnitude faster than random forest. It is likely that random forest requires more processing power and time due to the fact that it is a form of ensemble.
- Naïve-bayes and logistic regression are indifferent to the volume of data sets and imbalanced data treatment, and their performances are mediocre. It can also be attributed to the sub-optimal hyper-parameters selected.
- Decision tree performs really well, which is at par with decision tree and random forest in a smaller data set. However, with imbalanced data, it performs as the worst classifier when a bigger Taiwan credit data set is used. It is highly likely that decision tree overfits with smaller training and testing sets. Moreover, it can be seen that with a smaller data set, decision tree achieves a high AUC score after imbalanced data treatment. This is consistent with its characteristic of being sensitive to imbalanced data.
- The data processing step detects multicollinearity in Taiwan credit data set with all features, “bill\_amtX”’s VIF above 20 (between 20.8453 and 38.2155). The presence of multicollinearity in this data set affects only k-nearest neighbour and logistics regression.
- All analytics pipelines with various permutations include standardization which includes scaling and normalization. Whilst k-nearest neighbour requires and benefits from standardization, naïve-bayes, logistic regression, decision tree and random forest are robust to standardization.
- Forcing the classifiers to pick a smaller set of relevant features will degrade the model performance. This results in insufficient data, which will often affect model accuracy.
- In various scenarios, hyper-parameter selection will determine model performance degradation or remain similar after imbalanced data treatment. It is important to

note that apart from naïve-bayes classifier, which is based on probability, all other classifiers are susceptible to imbalanced data. Due to the limitations of computing resources, obtaining the best hyper-parameters for the larger Taiwan credit data set is not achievable.

## **6. Further Studies**

It is desirable to further study the effect to the classifiers using more refined analytics pipelines such as the inclusion of non-standardized (non-scaling and non-normalizing) approach that includes outliers and missing values in the data sets. The effect of multicollinearity to various classifiers can be explored further. The concept of volatility introduced by Zelaya (2019), which involves including/excluding specific steps in the analytics pipelines, requires further study.

Apart from SMOTE, the use of other treatments suggested by Alam et al. (2020) such as random oversampling, ADASYN, k-means SMOTE, borderline-SMOTE and SMOTE-Tomek is worth exploring since most classifiers, except naïve-bayes, perform sub-optimally with the presence of imbalanced data. Despite the use of truncated singular-value decomposition (truncatedSVD) in the study, the analytics pipelines will also benefit from exploring other dimension reduction techniques including principal component analysis (PCA), t-distributed stochastic neighbour embedding (t-SNE) and linear discriminant analysis (LDA). Further study will benefit by delving into the decision-making criteria used to determine the most relevant predictors for each classifier. Further explorations should be conducted for classifiers using neural network, support vector machine and other modern ensemble techniques such as gradient boosting, extreme gradient boosting and light gradient boosting.

Finally, since achieving the best hyper-parameters for classifiers is key part of the study, it will be worthwhile to include more computing resources to search for the most optimal hyper-parameters for various classifiers. A failure to obtain sufficient system resources will produce sub-optimal hyper-parameters.

## **7. Conclusions**

This study highlights the distinctive characteristics of the five classifiers and how they perform under different data pre-processing steps. The data pre-processing in this study includes data cleansing, features encoding and selection, reduction of dimensions, treatment of imbalanced data and cross validation of training/testing data sets. The final comparisons of the five classifiers demonstrate that data pre-processing steps in conjunction with the data size, complexity and patterns will determine the accuracy of certain classifiers. For example, decision tree performs superbly (overfits) when data size is minor compared to its poor performance when data volume is large. In contrast, the study also shows that random forest does not tend to overfit even with the presence of imbalanced data. In short, the study demonstrates that data distribution and size, multicollinearity, features relevance and imbalanced class contribute to the final scores of models and each classifier reacts to these factors differently (Table 5).

Equally important is the tuning of the hyper-parameters for respective classifiers, with the study concluding that the default hyper-parameters perform sub-optimally. That being said, investing in computing resources to derive the best hyper-parameters is crucial for striving towards the best performing models and achieving cost savings for lending institutes.

Finally, this study concludes that it is mandatory to apply data domain knowledge prior to selecting a classifier of choice. This is primarily due to fact that a data set may have a pattern that suits one classifier but not the other. Hence, it is imperative to understand by unravelling the complexity and patterns of data sets prior to selecting, training and fitting a model.



**Author Contributions:** Conceptualization, R.L. and H.D.; methodology, R.L. and H.D.; software, R.L. and H.D.; validation, R.L. and H.D.; formal analysis, R.L., H.D. and A.W.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Data are contained within the article.

**Acknowledgments:** Sincere thanks to Ong Seng Huat, UCSI, Malaysia and Dat Tran, University of Canberra, for their constructive comments.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

- Alam, Talha Mahboob, Kamran Shaukat, Ibrahim A. Hameed, Suhuai Luo, Muhammad Umer Sarwar, Shakir Shabbir, Jiaming Li, and Matloob Khushi. 2020. An investigation of credit card default prediction in the imbalanced datasets. *IEEE Access* 8: 201173–98. [CrossRef]
- Altman, Edward I. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance* 23: 589–609. [CrossRef]
- Atiya, Amir F. 2001. Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks* 12: 929–35. [CrossRef]
- Beaver, William H. 1966. Financial ratios as predictors of failure. *Journal of Accounting Research* 4: 71–111. [CrossRef]
- Black, Fischer, and Myron Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–54. [CrossRef]
- Breiman, Leo. 2001. Random forests. *Machine Learning* 45: 5–32. [CrossRef]
- Breiman, Leo, and Adele Cutler. 1993. A deterministic algorithm for global optimization. *Mathematical Programming* 58: 179–99. [CrossRef]
- Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16: 321–57. [CrossRef]
- Chen, Tianqi, and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. Paper presented at the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17; pp. 785–94.
- Christoffersen, Peter. 2011. *Elements of Financial Risk Management*. Cambridge, MA: Academic Press.
- Deakin, Edward B. 1972. A discriminant analysis of predictors of business failure. *Journal of Accounting Research* 10: 167–79. [CrossRef]
- Desai, Vijay S., Jonathan N. Crook, and George A. Overstreet, Jr. 1996. A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research* 95: 24–37. [CrossRef]
- Feldman, David, and Shulamith Gross. 2005. Mortgage default: Classification trees analysis. *The Journal of Real Estate Finance and Economics* 30: 369–96. [CrossRef]
- Fernández, Alberto, Salvador Garcia, Francisco Herrera, and Nitesh V. Chawla. 2018. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research* 61: 863–905. [CrossRef]
- Freund, Yoav, and Robert E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55: 119–39. [CrossRef]
- Friedman, Jerome H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29: 1189–232. [CrossRef]
- Han, Jiawei, Jian Pei, and Hanghang Tong. 2022. *Data Mining: Concepts and Techniques*. Burlington: Morgan Kaufmann.
- Hofmann, Hans. 1994. Statlog (German Credit Data). *UCI Machine Learning Repository*. [CrossRef]
- Huang, Zan, Hsinchun Chen, Chia-Jung Hsu, Wun-Hwa Chen, and Soushan Wu. 2004. Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems* 37: 543–58. [CrossRef]
- Jain, Anil K., Robert P. W. Duin, and Jianchang Mao. 2000. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22: 4–37. [CrossRef]
- Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* 30. Montreal: Curran Associates.
- Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* 34: 2767–87.
- Lobo, Jorge M., Alberto Jiménez-Valverde, and Raimundo Real. 2008. AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* 17: 145–51. [CrossRef]
- Martin, Daniel. 1977. Early warning of bank failure: A logit regression approach. *Journal of Banking & Finance* 1: 249–76.
- Nelson, Benjamin J., George C. Runger, and Jennie Si. 2003. An error rate comparison of classification methods with continuous explanatory variables. *IIE Transactions* 35: 557–66. [CrossRef]
- Noor, Jamal A. Mohamed, and Ali I. Abdalla. 2014. The Impact of financial risks on the firms' performance. *European Journal of Business and Management* 6: 97–101.
- Ohlson, James A. 1980. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* 18: 109–31. [CrossRef]

- Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12: 2825–30.
- Peng, Yi, Guoxun Wang, Gang Kou, and Yong Shi. 2011. An empirical study of classification algorithm evaluation for financial risk prediction. *Applied Soft Computing* 11: 2906–15. [CrossRef]
- Vapnik, Vladimir. 1999. *The Nature of Statistical Learning Theory*. Berlin and Heidelberg: Springer Science & Business Media.
- Varoquaux, Gaël, Lars Buitinck, Gilles Louppe, Olivier Grisel, Fabian Pedregosa, and Andreas Mueller. 2015. Scikit-learn: Machine learning without learning the machinery. *GetMobile: Mobile Computing and Communications* 19: 29–33. [CrossRef]
- West, David. 2000. Neural network credit scoring models. *Computers & Operations Research* 27: 1131–52.
- West, Robert Craig. 1985. A factor-analytic approach to bank condition. *Journal of Banking & Finance* 9: 253–66.
- Yeh, I-Cheng. 2016. Default of credit card clients. *UCI Machine Learning Repository* 10: C55S3H. [CrossRef]
- Yeh, I-Cheng, and Che-hui Lien. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications* 36: 2473–80. [CrossRef]
- Yobas, Mumine B., Jonathan N. Crook, and Peter Ross. 2000. Credit scoring using neural and evolutionary techniques. *IMA Journal of Management Mathematics* 11: 111–25. [CrossRef]
- Zelaya, Carlos Vladimiro González. 2019. Towards explaining the effects of data preprocessing on machine learning. Paper presented at the 2019 IEEE 35th International Conference on Data Engineering (ICDE), Macao, China, April 8–11; pp. 2086–90.
- Zhang, Guoqiang, Michael Y. Hu, B. Eddy Patuwo, and Daniel C. Indro. 1999. Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European Journal of Operational Research* 116: 16–32. [CrossRef]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

# The Impact of Non-Financial and Financial Variables on Credit Decisions for Service Companies in Turkey

Ali İhsan Çetin <sup>1,\*</sup> , Arzu Ece Çetin <sup>2</sup> and Syed Ejaz Ahmed <sup>3</sup>

<sup>1</sup> Department of Finance and Banking, Faculty of Business, Ankara Yıldırım Beyazıt University, 06760 Ankara, Turkey

<sup>2</sup> Department of Business, Faculty of Business, Gebze Technical University, 41400 İstanbul, Turkey; aecetin@gtu.edu.tr

<sup>3</sup> Faculty of Mathematics & Statistics, Brock University, Toronto, ON L2S 3A1, Canada; sahmed5@brocku.ca

\* Correspondence: alihsancetin22@gmail.com; Tel.: +90-554-992-2847

**Abstract:** This study aims to analyze and generalize the factors influencing credit decision-making in Turkey's service sector, which has seen substantial growth and increased dynamism post-2000, coinciding with accelerated economic development. The evolving competitive landscape and shifting consumer purchasing perceptions have led companies within this sector to seek differentiation strategies to attain a competitive edge. In this context, access to credit emerges as a crucial enabler for companies to expand and capture market share. The research focuses on the financial and non-financial characteristics of medium-sized service sector firms seeking credit, recognizing that both sets of variables play a pivotal role in the credit allocation process conducted by banks. The core of this study involves applying established assumption tests from extant literature, followed by an extensive regression analysis. The primary objective of this analysis is to identify and underscore the key financial and non-financial factors that significantly impact credit decisions in the service sector. By examining these variables, the study seeks to contribute valuable insights into the credit decision-making process, addressing the nuanced and varied nature of the service sector. This approach not only provides a deeper understanding of the sector's credit dynamics but also assists in formulating more informed strategies for businesses seeking financial support within this evolving economic landscape. The primary conclusion reached by the study is that non-financial variables exert a greater influence on credit decision-making in the service sector compared to financial variables.

**Keywords:** credit decision; determinants of credit; qualitative variables; financials service sector



**Citation:** Çetin, Ali İhsan, Arzu Ece Çetin, and Syed Ejaz Ahmed. 2023. The Impact of Non-Financial and Financial Variables on Credit Decisions for Service Companies in Turkey. *Journal of Risk and Financial Management* 16: 487. <https://doi.org/10.3390/jrfm16110487>

Academic Editor: Svetlozar (Zari) Rachev

Received: 11 September 2023  
Revised: 13 November 2023  
Accepted: 14 November 2023  
Published: 17 November 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Financial analysis plays a crucial role in understanding a company's financial position and performance. By evaluating financial statement accounts and comparing them to established standards and industry averages, financial analysis allows for a comprehensive assessment of a company's liquidity, financial structure, profitability, and activities. These kinds of analyses are also critical for understanding and assessing a company's financial health. This involves evaluating the relationships among accounts in financial statements and interpreting them by comparing them with industry benchmarks and established standards. Financial analysis is an essential tool for understanding and interpreting a company's financial statements (Konstantinidis et al. 2021).

Non-financial data collection and analysis are integral to credit assessment to determine the creditworthiness of borrowers and minimize credit risk (İş Bankası 2012; Vakıfbank 2011; Geçer 2014). Bolkvadze (2019) emphasizes the importance of analytical financial tools in financial analysis, particularly in the study of business entities. Ceran (2019) used financial ratios to predict non-performing loans (NPLs) in advance using artificial neural networks. Mbona, Masimba, and Kong (Mbona and Yusheng 2019) highlight the significance of financial statement analysis in understanding financial performance. Lam et al.

(2021) propose an integrated entropy–fuzzy VIKOR model to evaluate the financial performance of construction companies, identifying ECONBHD as the best-performing firm.

Credit-granting decisions have been extensively explored in academic literature, particularly with a focus on small and medium-sized Enterprises (SMEs). Traditionally, these studies emphasize the importance of financial variables in credit decisions. Financial analyses and ratios are key indicators of a firm’s creditworthiness. However, recent research has also begun to stress the significance of non-financial variables (Jasevičienė et al. 2013).

Yan and Li (2023) introduced a credit risk prediction model for SMEs utilizing a decision tree trained on data sets combined with a linear programming approach. By integrating bank-specific constraints and objectives, the model aims to enhance the precision of credit risk quantification for banks.

Other studies explained that modern firms aim to enhance shareholder value by making decisions on various aspects, such as liquidity status, profitability, financial structure, investment projects, and technological adaptability. Financing decisions can influence firm value, with financial analysis providing information on ratio analysis, liquidity status, financial structure, asset utilization efficiency, and profitability (Altuğ 2010).

Another study investigates the importance of non-financial information in credit decisions, focusing on microentrepreneurs in China. It was discovered that non-financial data such as business characteristics, personal traits, and social relationships play a substantial role in the credit-granting process (Xu et al. 2019).

Similarly, Edem (2017) examined the role of non-financial data in making credit decisions in Macedonian commercial banks. The study revealed that, in addition to financial ratios, non-financial variables such as the company’s reputation, its relationship with the bank, market conditions, and the legal framework had a significant impact on credit decisions (Edem 2017).

Hossain (2023) reviewed the literature from 2016 to 2022 on Big Data analytics in banking and found that IEEE (The Institute of Electrical and Electronics Engineers) is the predominant publisher, with China as the major contributor. Hossain’s study highlights Random Forest techniques as dominant in credit risk management in the financial services sector while noting the need for further research on integrated algorithms.

Erdoğan’s (2020) study investigated firm-specific and macroeconomic factors influencing the profitability of manufacturing firms listed on Borsa Istanbul between 2009 and 2019. A total of 129 firms were grouped by asset size, and 80 firms (20 large, 29 medium, and 31 small) were included in the analysis. Quarterly data from these firms were used, and regression analysis was conducted. The dependent variables are the active profitability rate, equity profitability rate, and pre-tax and interest profit (operational profitability) rates. The independent variables are company size, liquidity, asset structure, total debt ratio (leverage), GDP growth rate, and interest rates. The results of the research show a negative impact of the leverage ratio and fixed asset ratio on all profitability ratios, while the GDP growth rate had a positive effect (Erdoğan 2020).

A systematic review of the literature provides a robust foundation for examining the impact of financial and non-financial variables on the credit decisions of middle-market companies in the service sector.

While the literature highlights the importance of both financial and non-financial variables in credit decisions, there appears to be a gap in studies that specifically focus on the service sector in Turkey. This study aims to fill this gap and provide a comprehensive understanding of the factors influencing mid-segment companies’ credit decisions in the service sector.

This study evaluates the impact of financial and non-financial features on the credit decisions of SMEs in the service sector. The motivation behind this study lies in the changing landscape of competition and consumer behavior, in which companies in the service sector strive to differentiate themselves to gain a competitive edge.

Although academic studies exist on the attributes of companies affecting credit decisions in Turkey, no study has been found to specifically reveal the attributes that are

effective in the credit decisions of companies operating in the service sector. This finding reveals a deficiency in existing literature.

Therefore, the findings of this study contribute to the existing research by shedding light on certain financial and non-financial features that significantly affect credit decisions for mid-segment service sector companies in Turkey. To accomplish this, we conducted a comprehensive analysis of a diverse range of financial and non-financial variables. Statistical techniques and models are used to assess the relative importance and impact of these variables on credit decisions.

This study emphasizes the pivotal importance of a multifaceted array of variables, encompassing financial components, financial ratios, and non-financial data, in molding the credit decisions of enterprises operating within the service sector. While a vast body of literature has delved deeply into the importance of financial and non-financial variables in credit decisions, what remains conspicuously under-explored is the specificity and nuanced understanding of the service sector in Turkey, especially with respect to mid-sized firms. Given the burgeoning role of the service sector in the Turkish economy and its intricate interplay with global market dynamics, this research emerges not just as a filler of an academic void but as an imperative. Our study is novel in its targeted focus on this particular segment, offering insights that transcend conventional binary distinctions of financial and non-financial variables. The contributions of this research are multifaceted. First, it bridges a gap in understanding the unique dynamics of the Turkish service sector. Second, it provides a granular analysis, juxtaposing a myriad of variables to ascertain their influence on credit decisions. Last, our findings will serve as a pragmatic guide for financial institutions, equipping them with a refined lens to evaluate creditworthiness. We believe that this nuanced understanding can foster more resilient, sustainable, and inclusive financial ecosystems in the region.

### *1.1. Financial Analysis*

Financial analysis plays a vital role in financial decision-making by collecting and interpreting data to evaluate the financial performance of businesses. This is important to companies, banks, and governments because it provides a foundation for effective financial planning. Planning activities cannot be conducted effectively without comprehensively analyzing a company's financial situation. Additionally, financial analysis is important for governments and organizations that consider lending, partnering, taxing, and investing in businesses.

Basic financial statements such as balance sheets and income statements were first examined during the credit process. Other auxiliary tables were used to determine whether the companies were suitable for credit.

### *1.2. Financial Statement Items and Ratios*

Banks request financial and non-financial data from companies that apply for credit during their credit processes. Non-financial data consists of information about the company's standing with other banks and in the market, whereas financial data comprises balance sheets and income statement items, financial ratios, and sectoral ratios. Financial ratios are optional because they consist of balance sheet items. In other words, they vary on a sectoral, regional, and firm basis. Therefore, depending on the company and sector, different ratios can be produced and used for each credit evaluation.

Ratio analysis examines the partial relationships between items in financial statements and provides information on the financial condition of a business. This ratio is a mathematical expression of the relationship between two items in financial statements. The calculated ratios were typically expressed as percentages. It is possible to calculate a large number of ratios to indicate the relationships among financial statement items. However, rather than calculating a large number of ratios in a ratio analysis, it is more meaningful to focus on the ratios that have meaningful relationships with each other.

### 1.3. Non-Financial Analysis

Non-financial credit data, also known as intelligence data, are beneficial for getting to know customers well and making accurate, quick, and safe decisions. If the demands of a customer who is not well known cannot be met accurately and safely, the margin of error in the decision to be made is high. When evaluating customer credit requests, it is necessary to consider non-financial data to reach correct and safe decisions. The purposes of using non-financial data in banking can be summarized as follows: to obtain information and opinions about the general conditions of businesses, to discipline credit preparation based on certain procedures and principles, and to ensure that credit risk is eliminated or reduced by determining the business's ability to pay.

There are a number of reasons why non-financial variables can be important predictors of credit risk. First, non-financial variables can provide insights into borrowers' motivations and intentions. For example, a borrower who is unemployed and has a history of loan default is more likely to default on a new loan than a borrower who is employed and has a good credit history (World Bank 2014).

### 1.4. Credit

Economically, credit refers to the purchasing power of legal and real persons. The main reason why various transactions that differ from each other are gathered under the name of credit is because all of these transactions include providing "purchasing power" to the other party (Yürük 2006, p. 63).

We examine the relationship between total credit volume and economic growth in two ways. The first of these empirical studies is the evolution of credit volume as an indicator of financial development or credit to the private sector as a ratio of gross domestic product, and the second focuses on the relationship between direct credit volume and economic growth. With the inferences of these studies and the increase in financial instruments and institutions, the results show that financial development increases, and consequently, economic growth is supported (Merçan 2013, p. 57).

### 1.5. Service Industry in Turkey

In the information age, and owing to the dynamic components of other sectors, services are rapidly growing in importance in Turkey and the global economy. For these reasons, Turkey has adopted a change in its sectoral structure in the planned development model that has been implemented since the 1960s as an important objective, and plans have been prepared within this framework. Since the 1980s, Turkey has increased industrial activities while entering a period of globalization, and with the economic transformation of the 2000s, the importance of the service sector and its role in the economy has increased (İnamoğlu 2013, p. 2).

In the first section of this study, credit, financial items, financial ratios, and non-financial data are introduced, and the service sector is delineated. In the second section, the study's objectives, research design, data structure, hypotheses, detailed information about the data set, sampling methods, and analytical techniques are elaborated. The third section encompasses normality tests, correlation analysis, regression analysis, and hypothesis tests, all serving the main objectives of the study. In the concluding section, the study's key takeaways are discussed, results are analyzed and interpreted, and recommendations for future research are presented.

## 2. Methodology and Data

### 2.1. Aim of the Research

Financial and non-financial analyses play a significant role in credit decisions in the service sector. Given the substantial impact of both types of analysis on credit decisions, companies in today's service sector not only seek credit from banks but also consider them stakeholders for growth. Over the years, the symbiotic relationship between a company

and a bank has had a considerable positive impact on the survival and growth of the company as well as on the bank's earnings.

This study evaluates the significance and impact of financial and non-financial analyses on credit decisions in Turkey's service sector. This study focuses on identifying the most influential financial and non-financial variables among various options. This study aims to ascertain which types of data—financial or non-financial—are more critical to banks' credit decisions. This study seeks to contribute to firms' healthy growth strategies and efforts to reduce non-performing loan (NPL) rates by understanding the effectiveness of the various metrics and ratios used in banking practices.

The first section provides a brief introduction to the research topic and outlines its objectives. The second part includes a literature review covering financial and non-financial analyses, credit, and the service industry. This section establishes the theoretical framework by elaborating on financial ratios and key non-financial variables and detailing their importance in credit decision-making within the service sector.

The section on research design and methodology outlines the study's purpose, scope, methods, and model. Credit decisions serve as the dependent variable, whereas financial and non-financial characteristics are selected as independent variables based on expert opinions. The experts consulted four credit process analysts with at least five years of experience in the bank from which the data were sourced. This section also describes the data collection methods, metrics employed in the study, statistical analysis techniques, and characteristics of the variables under investigation. The conclusion offers answers to the research hypotheses and presents the major findings. We also discuss the contributions and limitations of this study to the existing literature.

## 2.2. Research Methodology and Design

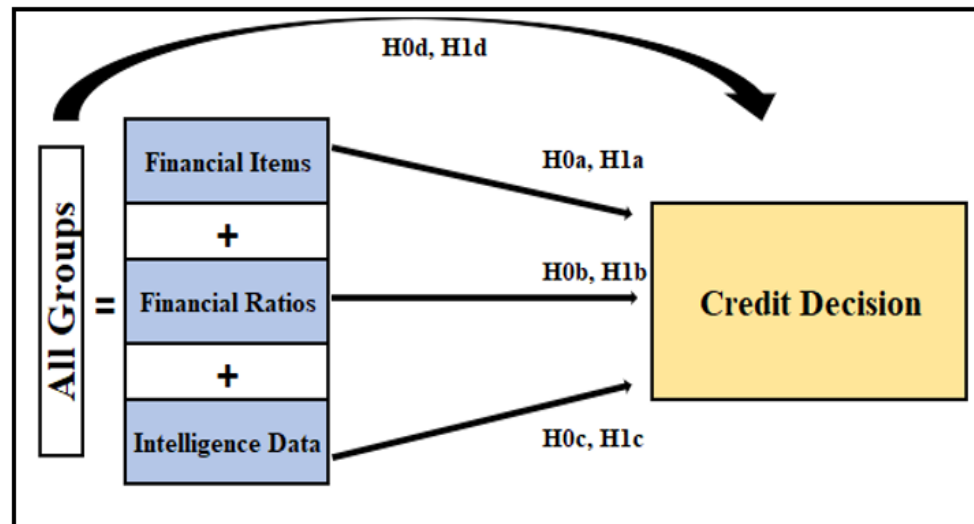
Qualitative and quantitative methods were used in this study. These methods are forms of scientific research aimed at understanding research subjects in a relevant population. Although the overall objectives of the quantitative and qualitative methods are similar, their approaches and focus differ substantially. To elucidate the impact of financial ratios, financial items, and non-financial data on credit decision-making, data from 530 companies were collected from private banks. Subsequently, 34 factors were identified in the analysis. Data were obtained from the bank system in compliance with all necessary ethical protocols, and it was determined that the confidential nature of the data should be maintained and not disclosed. After establishing these factors and completing the calculations, analyses were performed to test the proposed hypotheses. The results clarify the influence of the selected variables on credit decisions.

## 2.3. Hypotheses

1. **H0a:** Financial items do not have a significant effect on credit decisions in the Turkish service sector.
2. **H1a:** Financial items have a significant effect on credit decisions in the Turkish service sector.
3. **H0b:** Financial ratios do not have a significant effect on credit decisions in the Turkish service sector.
4. **H1b:** Financial ratios have a significant effect on credit decisions in the Turkish service sector.
5. **H0c:** Non-financial data do not have a significant effect on credit decisions in the Turkish service sector.
6. **H1c:** Non-financial data have a significant effect on credit decisions in the Turkish service sector.
7. **H0d:** Financial items, financial ratios, and non-financial data do not have a significant effect on credit decisions in the Turkish service sector.

**8. H1d:** Financial items, financial ratios, and non-financial data have a significant effect on credit decisions in the Turkish service sector.

Figure 1 presents the framework for the hypotheses. This figure summarizes whether Financial items, Financial ratios, and Non-financial items affect Credit Decisions. The collective analysis of these groups of variables is also included under the label “All Variables”.



**Figure 1.** Hypothesis and Research Organization of All Variable Groups and Credit Decisions.

#### 2.4. Sampling, Data and Measures

This study explored the relationship between financial and non-financial variables and credit decisions within Turkey’s service sector, focusing specifically on SMEs and commercial firms. Systematic sampling was employed to select companies based on criteria such as business segments, assets, and turnovers. The study incorporated 13 financial variables, 12 financial ratio variables, and nine non-financial variables as independent variables. These were selected through a review of existing literature and consultations with experts from bank allocation departments (Erdoğan 2020; İnamoğlu 2013; Ceran 2019). Missing values were removed from the initial dataset of 1356 firms, resulting in a final dataset comprising 530 data points. The scope of the study encompassed various sub-sectors within the service industry, treating it as a holistic entity.

Figure 2 presents the research organization and offers a visual representation of the theoretical model illustrating the relationship between credit decisions and financial/non-financial variables. In this figure, 34 initial independent variables are categorized into three distinct groups, and their full names are provided. The dependent variable in this study was the credit decisions enacted by the banks. While some studies treat the credit decision as a categorical variable, it is considered a numerical variable, particularly in hybrid studies where both the credit decision and rate are predicted. For the purpose of this study, the target variable was the amount of credit allocated to the firms; therefore, it was treated as a numerical variable to enhance the sensitivity of the analysis.



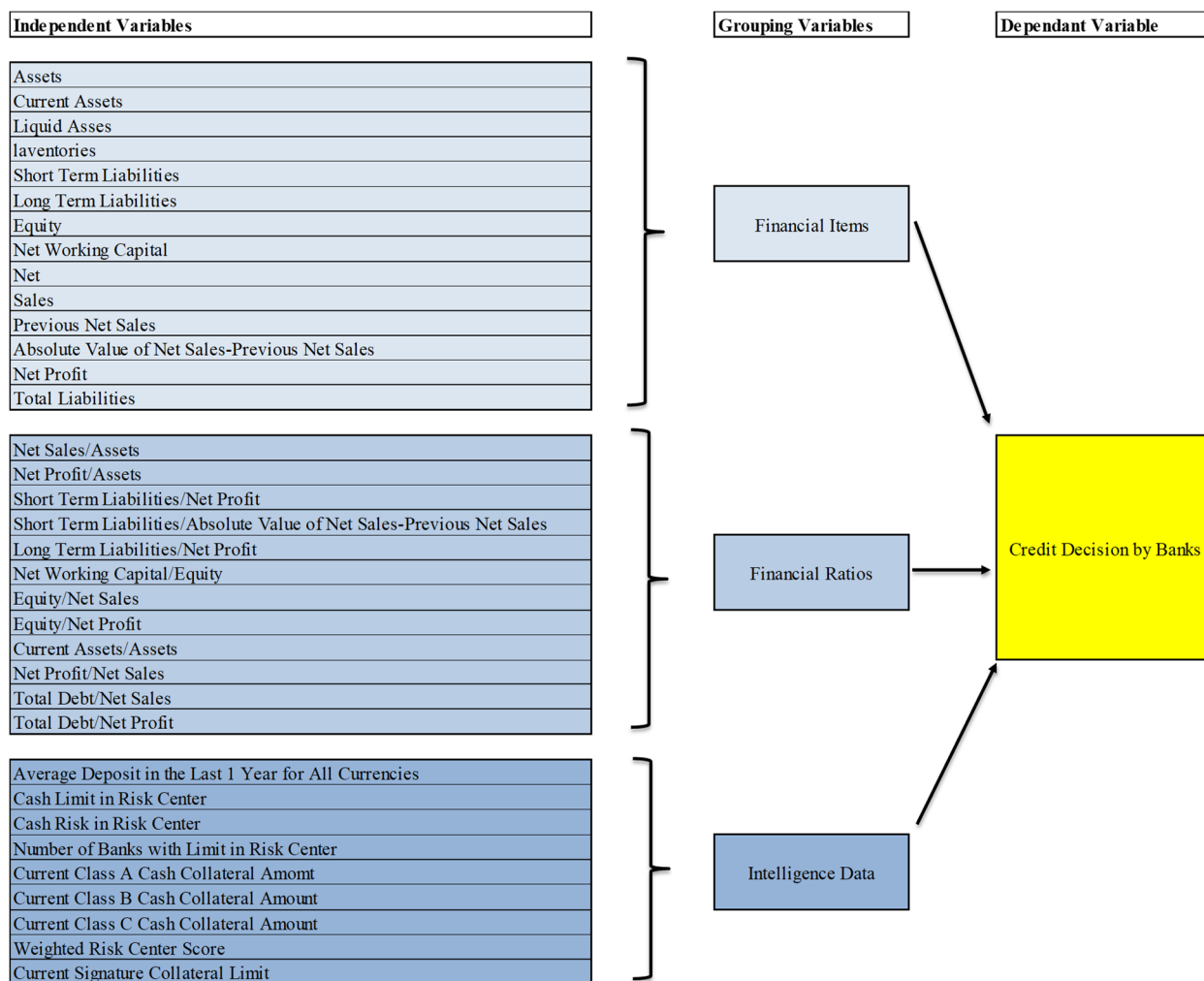


Figure 2. Conceptual Framework of Three Variable Groups and Credit Decisions.

### 2.5. Statistical Methods Used in Data Analysis

The data analysis and hypothesis testing were performed using Python and SPSS 20.0 software packages. The analytical process began with preliminary assumption tests for normality, followed by correlation and regression analyses, and ultimately, hypothesis testing. Finally, the conclusions are presented in the final section.

The data distribution was examined using normality tests to determine the appropriate analytical method. Given that the sample size exceeded 50, the Kolmogorov–Smirnov test was used to assess data normality. However, recognizing the sensitivity of the test to large sample sizes, we decided that exclusive reliance on these results could introduce bias. Consequently, P–P/Q–Q plots were consulted as supplementary tools to corroborate the findings of the normality tests.

A correlation analysis was performed to ascertain the relationships between the variables. The high correlations among the independent variables raised concerns about multicollinearity, which were mitigated by removing highly correlated variables from the analytical model. When deciding which variable to exclude, its relationship with the dependent variable was considered. Also, the fixed effects model was used in this study.

In the employed model, the Bidirectional Elimination (or Stepwise Selection) method was utilized for variable selection within the Stepwise approach. This was implemented with the objective of establishing a refined set of variables for modeling, specifically by eliminating the binary relationships among the variables. This process aims to ensure a more robust and accurate model by focusing on the most relevant and impactful variables, thereby enhancing the overall effectiveness of the modeling exercise.

### 3. Data Analysis and Research Findings

As the normality test, P–P/Q–Q plots were taken into consideration when deciding whether the data was normally distributed or not. The credit decision variable is found to have an approximately normal distribution.

#### 3.1. Normality Analysis

Several methods are available to measure the normality of variables. The most commonly used methods are as follows.

- Shapiro–Wilk Test:

Null Hypothesis: The data follows a normal distribution.

Test Statistic ( $W$ ):

$$W = \frac{\left(\sum_{i=1}^n a_i x_{(i)}\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

The test statistic  $W$  is calculated based on the ordered sample values  $x_{(i)}$  and their corresponding expected values  $a_i$  under normality (Shapiro et al. 1968). Based on the Shapiro–Wilk test, the data appeared to be normally distributed (Shapiro–Wilks Test Statistic: 0.9944,  $p$ -value: 0.5009).

- Kolmogorov–Smirnov Test:

Null Hypothesis: The data follow a specific distribution (e.g., normal distribution).

Test Statistic ( $D$ ):

$$D = \max|F_n(x) - F(x)|$$

Test statistic  $D$  was calculated based on the maximum absolute difference between the cumulative distribution function (CDF)  $F_n(x)$  of the observed data and the CDF  $F(x)$  under the hypothesized distribution. The Kolmogorov–Smirnov test indicated that the data were likely to be normally distributed (Kolmogorov–Smirnov Test Statistic: 0.0259,  $p = 0.5046$ ).

- P–P/Q–Q (Probability–Probability/Quantile–Quantile) plot:

P–P/Q–Q plots are utilized to assess the fit of a dataset to a normal distribution. The PP plot compares observed cumulative probabilities with expected probabilities under a normal distribution. Ideally, the plotted points should be aligned along a straight line, indicating a normal distribution. Deviations from the straight line indicated a departure from normality. These plots helped identify significant deviations from the normality of the data. Figure 3 presents the corresponding plots for the dependent variable.

Figure 3 shows that credit decisions and independent variables exhibit an approximate normal distribution.

The following steps were followed in the analyses.

- Correlation Analysis: The formula for Pearson’s correlation coefficient ( $r$ ) between two variables  $X$  and  $Y$  is given by

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

where  $\bar{X}$  and  $\bar{Y}$  represent the means of variables  $X$  and  $Y$ , respectively.

- Multiple Regression Analysis: the formula for multiple linear regression is represented as follows:

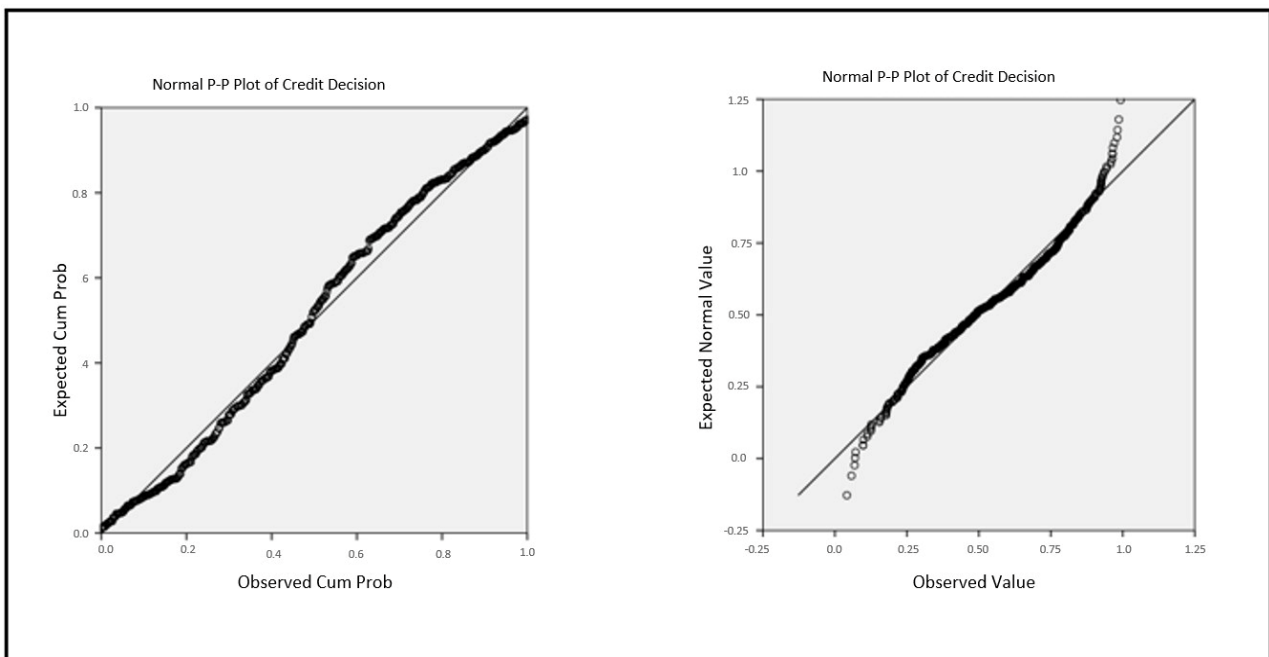
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

where  $Y$  is the dependent variable,  $X_1, X_2, \dots, X_p$  is the independent variable,  $\beta_0, \beta_1, \beta_2, \dots, \beta_p$  is the regression coefficient, and  $\varepsilon$  is the error term.

- Adjusted R-squared (Coefficient of Determination): the formula for adjusted R-squared ( $Adj R^2$ ) in the multiple regression analysis was calculated as

$$Adj\_R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1},$$

where  $R^2 = 1 - \frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2}$  and  $N$  are the total sample sizes;  $p$  is the number of independent variables;  $Y$  represents the observed values of the dependent variable;  $\hat{Y}$  represents the values predicted by the regression model; and  $\bar{Y}$  represents the mean of the dependent variable.



**Figure 3.** Normal P–P/Q–Q Plots of Credit Decisions.

### 3.2. Financial Item Analysis

Correlation analysis was conducted to assess the relationships among these financial variables, thereby minimizing the issue of multicollinearity. Thirteen independent variables were initially considered in the Financial item category. Subsequent to this evaluation, variables with a Pearson correlation coefficient of  $\pm 0.80$  or higher were scrutinized, and it was decided to retain only one of the highly correlated pairs, removing the other from the dataset. Notably, significant and strong correlations were observed between Assets and Current Assets, Assets and Long-Term Liabilities, and Current Assets and Current Liabilities (Pearson’s correlation coefficients were 0.842, 0.811, and 0.863, respectively; all  $p = 0.000$ ). Consequently, two variables—Current Assets and Long-Term Liabilities—were excluded, and the analysis proceeded with 11 financial variables.

Multiple linear regression analysis using a Stepwise Method revealed that the eighth model was statistically significant. The goodness of fit of the model was examined using the coefficient of determination adjusted to the  $R^2$  values. Based on this evaluation, the capability of the selected financial variables to explain the variations in credit decisions is 54%. The selected financial variables are assets, liquid assets, inventories, current liabilities, equity, net working capital, net sales, previous net sales, the absolute value of the change in net sales, net profits, and total liabilities. Additionally, a statistically significant relationship is identified between credit decisions and several financial variables, including assets, current liabilities, equity, net sales, previous net sales, the absolute value of the change

in net sales, net sales, net profit, and total liabilities ( $p < 0.05$ ). Conversely, no significant relationship is observed between liquid assets, inventories, net working capital, and credit decisions ( $p > 0.05$ ).

Based on the regression analysis results presented in Table 1, a statistically significant relationship is observed between credit decisions and financial item variables. The derived model was deemed statistically significant ( $F = 75.143, p < 0.001$ ). Additionally, the model displayed no evidence of autocorrelation, as indicated by a Durbin–Watson statistic of 0.996. Consequently, the model was deemed statistically robust and valid.

**Table 1.** Coefficients and Adj-R<sup>2</sup> of Regression Model for Financial Items.

8. Model F	Dependent Variable	Independent Variables	B	t	p	Adj-R <sup>2</sup>
F = 75.143 p = 0.000	Credit Decision	Constant	−197,186	−2.488	0.013	0.536
		Total Liabilities	0.146	7.720	0.000	
		Net Profit	0.417	8.871	0.000	
		Equity	0.206	4.952	0.000	
		Previous Net Sales	−0.120	−7.026	0.000	
		Net Sales	0.084	5.596	0.000	
		Current Liabilities	0.182	5.084	0.000	
		Assets	0.061	−2.554	0.011	
		Absolute Value of Net Sales–Previous Net Sales	−0.076	−2.198	0.028	

### 3.3. Financial Ratio Analysis

Data pertaining to the financial ratios of 530 companies in the service sector were analyzed. Financial ratio variables, presumed to influence the dependent variable, were employed in the analysis.

Correlation analysis was conducted to ascertain the relationship between the financial ratio variables. This step mitigated the risk of multicollinearity among the independent variables. The dataset initially contains 12 independent variables in the financial ratio category. Upon analysis, variables with a Pearson correlation coefficient of  $\pm 0.80$  or higher were scrutinized, and one variable from each correlated pair was removed from the dataset. Specifically, a significant and strong correlation was observed between “Net Profit/Assets” and “Current Liabilities/Net Profit” as well as between “Net Profit/Assets” and “Equity/Net Profit” (Pearson Correlation = 0.856,  $p < 0.001$ ; Pearson Correlation = 0.872,  $p < 0.001$ , respectively). Consequently, one variable—net profit/assets—was excluded because of its weaker association with the dependent variable.

A multiple linear regression analysis using the stepwise method was conducted to elucidate the relationship between credit decisions and the financial ratio variables. The fifth model was considered statistically significant. Key metrics, such as the relationship coefficient, percentage of the dependent variable explained by the independent variables, and adjusted R<sup>2</sup> values, were examined. According to the results presented in Table 2, the ability of the variables to explain the variations in credit decisions was 21%. The variables are net sales/assets, current liabilities/net profit, long-term liabilities/absolute value of net sales/previous net sales, long-term liabilities/net profit, net working capital/equity, equity/net sales, equity/net profit, current assets/assets, net fit/net sales, total debt/net sales, and total debt/net profit. Furthermore, a significant relationship is identified between Current Liabilities/Net Profit, Long-Term Liabilities/Net Profit, Current Assets/Assets, Total Debt/Net Sales, Total Debt/Net Profit, and Credit Decisions ( $p < 0.05$ ). Conversely, no significant relationship is observed between the variables net sales/assets, long-term liabili-

ties/absolute value of net sales–previous net sales, net working capital/equity, equity/net sales, equity/net profit, net profit/net sales, and credit decisions ( $p > 0.05$ ).

**Table 2.** Coefficients and Adj-R<sup>2</sup> of Regression Model for Financial Ratios.

5. Model F	Dependent Variable	Independent Variables	$\beta$	T	$p$	Adj-R <sup>2</sup>
F = 28.510 $p = 0.000$	Credit Decision	Constant	925,721	4.245	0.000	0.214
		Current Liabilities/Net Profit	−197,123	10.023	0.000	
		Total Debt/Net Profit	−80,574	−5.919	0.000	
		Long-Term Liabilities/Net Profit	−62,880	2.951	0.000	
		Current Assets/Assets	810,149	−2.464	0.014	
		Net Sales/Assets	127,067	2.118	0.035	

Based on the results of the regression analysis presented in Table 2, a significant relationship between credit decisions and financial ratio variables was observed. The derived model was statistically significant ( $F = 28.510, p < 0.001$ ). Additionally, a Durbin–Watson statistic of 0.447 indicated no autocorrelation within the model. Thus, the model was considered statistically valid.

### 3.4. Non-Financial Analysis

Data pertaining to non-financial variables from 530 companies in the service sector were analyzed. Non-financial variables believed to influence the dependent variable were included in the analysis. Correlation analysis was conducted to mitigate multicollinearity among the independent variables. Out of nine initial non-financial variables, one was removed due to a high correlation ( $\pm 0.80$  or above) with another variable. Specifically, a significant and high correlation was found between “Cash Limit in Risk Center” and “Cash Risk in Risk Center” (Pearson Correlation = 0.873,  $p = 0.01$ ). Consequently, “Cash Risk in Risk Center” was excluded from the dataset, resulting in eight variables for subsequent analyses.

A multiple linear regression analysis was conducted using a stepwise method to ascertain the relationship between credit decisions and the remaining non-financial variables. The sixth model is statistically significant. Key statistics, such as the correlation coefficient, explanatory power of the independent variables over the dependent variable, and adjusted R<sup>2</sup> values, were examined. According to these metrics, variables including “Deposit Average in Banks Over the Last Year”, “Cash Limit in Risk Center”, “Number of Banks with Limits in Risk Center”, “Current Class A Cash Collateral Amount”, “Current Class B Cash Collateral Amount”, “Current Class C Cash Collateral Amount”, “Weighted KKB Score” and “Current Signature Collateral Limit” accounted for 71% of the variance in credit decisions. A significant relationship was found between “Cash Limit in Risk Center”, “Number of Banks with Limits in Risk Center”, “Current Class A Cash Collateral Amount”, “Current Class B Cash Collateral Amount”, “Current Class C Cash Collateral Amount”, “Current Signature Collateral Limit” and the dependent variable, Credit Decision ( $p < 0.05$ ). No significant relationship was observed between the “average deposit in banks over the last year”, “weighted KKB score,” and the dependent variable, credit decisions ( $p > 0.05$ ).

According to the results of the regression analysis presented in Table 3, a significant relationship is observed between credit decisions and non-financial variables. The derived model was found to be statistically significant, as evidenced by an F-value of 217.73 and a  $p$ -value of less than 0.05 ( $p = 0.000$ ). Additionally, the Durbin–Watson statistic of 1.442 indicates the absence of autocorrelation within the model. Based on these metrics, the model was deemed statistically valid.

**Table 3.** Coefficients and Adj-R<sup>2</sup> of Regression Model for Non-Financial Data.

5. Model F	Dependent Variable	Independent Variables	$\beta$	T	$p$	Adj-R <sup>2</sup>
F = 217.73 $p = 0.000$	Credit Decision	Constant	570,182	5.937	0.000	0.714
		Current Class A Cash Collateral Amount	0.974	28.23	0.000	
		Cash Limit in Risk Center	0.061	7.076	0.000	
		Weighted KKB Score	1.650	4.370	0.000	
		Current Class C Cash Collateral Amount	-1.313	-3.696	0.000	
		Number of Banks with Limits in Risk Center	-36,496	-2.880	0.004	
		Current Class B Cash Collateral Amount	0.278	2.296	0.022	

### 3.5. All Variable Groups Analysis

A comprehensive regression analysis, including all group variables, was performed. Following prior regression analyses, data related to financial items, financial ratios, and non-financial variables for the 530 companies operating in the service sector were collectively analyzed. Multiple linear regression analysis is subsequently conducted to evaluate the collective influence of financial and non-financial variables on credit decisions. The variables are listed in Table 4.

**Table 4.** Classification of Significant Variables for Financial Items, Ratios, and Non-Financial Data.

All Significant Variables		
Financial Items	Financial Ratios	Non-Financial Variables
Assets	Net Sales/Assets	Deposit Average in Banks Last 1 Year
Liquid Assets	Current Liabilities/Net Profit	Cash Limit in Risk Center
Inventories	Long-Term Liabilities/Absolute Value of Net Sales-Previous Net Sales	Number of Banks with Limits in Risk Center
Current Liabilities	Long-Term Liabilities/Net Profit	Current Class A Cash Collateral Amount
Equity	Net Working Capital/Equity	Current Class B Cash Collateral Amount
Net Working Capital	Equity/Net Sales	Current Class C Cash Collateral Amount
Net Sales	Equity/Net Profit	Weighted KKB Score
Previous Net Sales	Current Assets/Assets	Current Signature Collateral Limit
Absolute Value of Net Sales-Previous Net Sales	Net Profit/Net Sales	
Net Profit	Total Debt/Net Sales	
Total Liabilities	Total Debt/Net Profit	

From the preceding analyses, the variables deemed redundant and subsequently removed included Current Assets and Long-Term Liabilities among the financial item variables, net fit/assets among the financial ratio variables, and cash risk in risk centers among Non-Financial Variables.

Correlation analysis was performed to ascertain intervariable relationships. Based on these results, the Weighted KKB Score was excluded from the Non-Financial Variables, and the Equity variable was removed from the financial variables.

A comprehensive regression analysis is performed to assess the relationship between credit decisions and the remaining variables. The eighth model was significant when using the stepwise method. Key metrics, such as the relationship coefficient, proportion of the dependent variable explained by the independent variables, and adjusted R-squared values, were analyzed. According to the findings, the collective explanatory power of financial items, financial ratios, and non-financial variables for the credit decision variable was 81%. The analysis revealed significant associations between the Credit Decision variable and several variables within Financial Items at a significance level of  $p < 0.05$  (e.g., such as Assets, Equity, Net Sales, Previous Net Sales, Net Profit, Total Liabilities); Financial ratios (e.g., such as Net Sales/Assets, Current Liabilities/Net Profit, Long-Term Liabilities/Net Profit, Net Profit/Net Sales, Total Debt/Net Sales, Total Debt/Net Profit); and Non-Financial variables (e.g., such as Cash Limit in Risk Center, Number of Banks with Limits in Risk Center, Current Class A Cash Collateral Amount, Current Class C Cash Collateral Amount, Weighted KKB Score, Current Signature Collateral Limit).

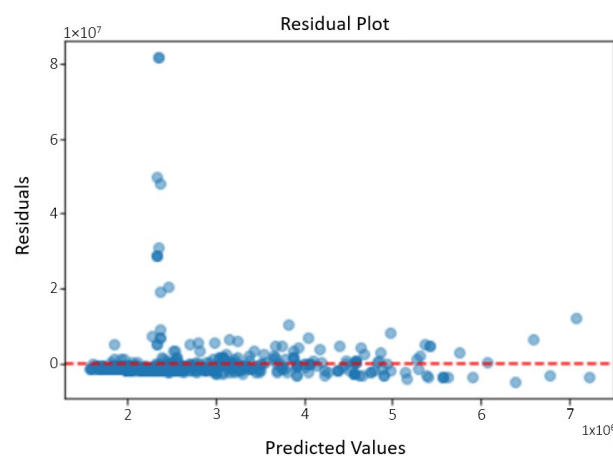
According to the regression analysis results presented in Table 5, a statistically significant relationship is identified between credit decisions and financial items, financial ratios, and non-financial factors. The model was confirmed to be statistically significant, as evidenced by an F-value of 132.641 and a  $p$ -value less than 0.05 ( $p = 0.000$ ). Additionally, the absence of autocorrelation in the model was verified using a Durbin–Watson statistic of 1.710, confirming the statistical validity of the model.

**Table 5.** Coefficients and Adj-R<sup>2</sup> of Regression Model for All Variables.

8. Model F	Dependent Variable	Independent Variables	B	t	p	Adj-R <sup>2</sup>
F = 132.641 p = 0.000	Credit Decision	Constant	245,785	2.322	0.021	0.805
		Current Class A Cash Collateral Amount	0.682	18.025	0.000	
		Total Liabilities	0.152	8.596	0.000	
		Current Liabilities/Net Profit	85,375	7.533	0.000	
		Total Debt/Net Profit	−50,288	−5.732	0.000	
		Net Profit	0.280	4.963	0.000	
		Net Profit/Net Sales	−165,810	−3.097	0.002	
		Current Signature Collateral Limit	1.289	4.081	0.000	
		Previous Net Sales	−0.019	−1.676	0.004	
		Net Sales	0.026	2.275	0.023	
		Number of Banks with Limits in Risk Center	−49,598	−4.540	0.000	
		Cash Limit in Risk Center	0.044	4.396	0.000	
		Assets	0.060	−4.867	0.000	
		Long-Term Liabilities/Net Profit	38,906	3.274	0.001	
		Net Sales/Assets	106,938	3.718	0.000	
		Current Class C Cash Collateral Amount	−0.886	−2.992	0.003	
Total Debt/Net Sales	−99,727	−2.314	0.021			

With a single unit increase in variables such as Current Class A Cash Collateral Amount, Total Liabilities, Current Liabilities/Net Profit, Net Profit, Net Profit/Net Sales, Current Signature Collateral Limit, Net Sales, Cash Limit in Risk Center, Assets, Long-Term Liabilities/Net Profit, and Net Sales/Assets, credit decisions increased by coefficients of 0.682, 0.152, 85,375, 0.280, 165,810, 1.289, 0.026, 0.044, 0.060, 38,906, and 106,938, respectively. Conversely, an increase of one unit in Total Debt/Net Profit, Previous Net Sales, Number of Banks with Limits in Risk Centers, Current Class C Cash Collateral Amount, and Total Debt/Net Sales resulted in a decrease in credit decisions, with coefficients of 50,288, 0.019, 49,598, 0.886, and 99,727, respectively.

The regression model was further corroborated through residual plot analysis, presented in Figure 4, which confirmed the absence of heteroskedasticity. The Jarque–Bera test yielded a *p*-value of 0.0000, further bolstering the reliability of the regression analysis employed in the study.



**Figure 4.** Residual Plot for Credit Decisions.

The red dashed line in a residual plot represents the expected position of residuals if a predictive model’s estimates are perfect. It serves as a benchmark for assessing the model’s prediction accuracy, where deviations indicate prediction errors. Within the framework of econometric modeling, the treatment of outliers is a topic that has generated significant discourse. Outliers can profoundly influence the accuracy of regression estimates. The literature typically divides the discussion on outliers into two distinct perspectives:

**Econometric Perspective:** From a purely statistical standpoint, outliers are observations that notably deviate from the expected pattern of the data. These can unduly influence the model’s performance and potentially lead to misleading interpretations. Quantitative metrics, such as Cook’s distance, are employed to diagnose and assess the influence of these outliers. When such outliers are identified, standard procedure in econometrics often recommends their removal or adjustment, especially if their presence adversely affects the model’s diagnostic tests and predictive accuracy.

**Financial Realism Perspective:** However, outliers often represent genuine economic phenomena in financial econometrics. These outliers could be symptomatic of events or processes that have genuine economic significance, such as unofficial balance sheet adjustments, anomalous sales activities, or taxation anomalies. Removing these outliers might enhance the statistical properties of the model but at the expense of omitting crucial information about the underlying economic process. From this viewpoint, discarding such outliers would strip the model of its ability to capture the full complexity and nuances of the financial reality it seeks to represent.

In summary, while the conventional econometric approach prioritizes the statistical integrity of the model, the financial realism perspective underscores the importance of retaining economically meaningful outliers. Hence, the decision of whether or not to remove outliers should not be based solely on statistical considerations. It is essential



also to weigh the substantive economic context and the specific objectives of the analysis. Furthermore, the impact of outliers needs to be quantified before deciding upon their removal. Given these considerations, it was determined that retaining the outliers would be more conducive to the objectives of the study.

In our analytical process, we also considered the impact of outliers on our regression model. A version of the model was employed after removing these outliers. The results were remarkably consistent with those obtained using the least squares estimation on the full dataset. Given the similarity in outcomes and to maintain conciseness in our presentation, we opted not to report the results of the outlier-removed model in detail within this paper. However, it is worth noting that the presence or removal of outliers did not significantly distort our main findings.

In the presence of outliers, one can employ robust regression techniques, for example, robust M-estimation, among others. However, in this study, we confine ourselves to classical least square estimation.

To further assess the integrity of the model, multicollinearity issues were examined independently for the Financial Items, Financial Ratios, and Intelligence Data variable groups. As evidenced by the Variance Inflation Factors (VIF) presented in Table 6, no VIF values indicative of multicollinearity concerns were identified.

**Table 6.** Variance Inflation Factor (VIF) for each Variable.

Variable Name	VIF	Variable Name	VIF	Variable Name	VIF
Total Liabilities	1.018321	Current Liabilities/Net Profit	1.007908	Current Class A Cash Collateral Amount	1.010088
Net Profit	1.016648	Total Debt/Net Profit	1.010269	Cash Limit in Risk Center	1.013432
Equity	1.057124	Long-Term Liabilities/Net Profit	1.019002	Weighted KKB Score	1.028962
Previous Net Sales	1.077362	Current Assets/Assets	1.217395	Current Class C Cash Collateral Amount	1.055214
Net Sales	1.025715	Net Sales/Assets	1.1246	Number of Banks with Limits in Risk Center	1.06382
Current Liabilities	1.01011			Current Class B Cash Collateral Amount	1.27036
Assets	1.013703				
Absolute Value of Net Sales–Previous Net Sales	1.010068				

## 4. Conclusions and Discussion

### 4.1. Findings and Results

Aligned with most commercial enterprises’ overarching objectives, banks primarily aim to maximize profits. Historically, they have realized this goal through avenues like funding businesses in the marketplace or treasury tool investments. Primarily, it is postulated that banks generate substantial revenue through market funding, coupled with meticulous oversight of credit returns to mitigate the emergence of non-performing assets. During this critical phase, a comprehensive evaluation of both financial and non-financial data furnished by companies becomes instrumental in guiding credit allocation decisions (Villalpando 2014).

The results indicate that both financial and non-financial data have a significant impact on credit decisions. The analysis demonstrates that these data positively influence credit decisions, with non-financial variables having the strongest effect, followed by financial and financial ratio variables. Considering all variable groups, the regression analysis confirms that evaluating these data together yields more effective credit decisions, with the highest

model success rate compared with separate analyses. Thus, it can be concluded that the financial and non-financial data provided by enterprises in Turkey have a positive effect on their credit limits and the banks' credit allocation decisions.

In juxtaposition with the extant literature that scrutinizes the influences of both financial and non-financial variables on credit determinations in specific sectors, the analytical results of this study elucidate certain variables previously unexplored in such contexts yet demonstrably impactful on credit decisions. Beyond confirming the findings of prior research, these results introduce novel variables into the discourse. Given the heterogeneous sub-sector distribution within the service industry, yet the financial congruities among them, the introduction of these unprecedented financial and non-financial variables can potentially enrich the credit decision-making paradigm within the sector (Melnyk et al. 2020; Ceran 2019).

For firms operating in the service sector, the establishment of robust, enduring, and efficacious credit relationships with financial institutions necessitates the comprehensive management of both financial and non-financial reputational factors. This involves the meticulous maintenance of financial records, robustness of key financial metrics, and transparency of non-financial data. Such strategic efforts contribute substantively to constructing a favorable organizational image and reputation from a banking perspective. By showcasing their management of both financial and non-financial variables, firms can position themselves as credible and reliable partners for financial institutions, thereby facilitating long-term, mutually beneficial relationships.

In conclusion, this study underscores the criticality of a multifaceted set of variables, including financial items, financial ratios, and non-financial data, in shaping credit decisions of enterprises in the service sector. These findings emphasize that firms must proactively manage their financial and market reputations to forge durable and advantageous credit affiliations with banks. By providing comprehensive and verifiable financial and non-financial data, these enterprises can increase their creditworthiness, thereby extending their access to higher credit limits and sustaining a continuum of support from financial institutions.

A salient contribution of this study lies in its nuanced approach to disentangling the intertwined influences of financial and non-financial data on credit determinations. Unlike previous studies, which often conducted isolated examinations of these variables or focused within narrow industry boundaries, the current research adopts a distinctive approach by integrating these variables comprehensively, thereby yielding a more holistic comprehension. Specifically, our findings illuminate the differential weightings banks according to these data types, with non-financial metrics emerging as surprisingly dominant determinants. This underscores a shifting paradigm in credit decision-making processes, where subjective and qualitative indicators are increasingly pivotal. Moreover, by introducing previously uncharted financial and non-financial variables into the credit evaluation matrix, this study advances the academic discourse and provides pragmatic insights for the banking sector. The integration of these innovative variables serves not only as an augmentation to the existing scholarly landscape but also equips financial institutions with refined tools and metrics, optimizing their credit allocation endeavors.

As a result, this study's insights offer significant implications for banks and policymakers, particularly in enhancing credit assessment models and fostering service sector growth. Banks can utilize these findings to incorporate a wider range of non-financial metrics in their credit evaluations, potentially improving risk assessment accuracy and supporting viable enterprises. Policymakers could leverage these insights to develop policies promoting transparency in non-financial reporting, aiding in more informed credit decisions, especially beneficial to SMEs.

#### *4.2. Limitations and Future Study*

Although insightful, this study had several limitations that warrant further discussion. The analysis was confined to the service sector and based on limited sample size, thus

constraining the generalizability of the findings beyond this specific industry. Furthermore, this study focuses solely on the impact of variables on credit decisions, and there may be limitations in the applicability of statistical methods to real-world scenarios. The missing values in the dataset further complicate the interpretation of the results.

Notwithstanding these limitations, this study significantly augments the existing literature by comprehensively examining the interplay between financial items, financial ratios, and non-financial variables affecting credit decisions, a domain not extensively explored in previous studies. This study offers both theoretical and empirical contributions by illuminating how various financial and non-financial metrics influence credit decisions within the SME segment of the Turkish service sector. These insights can serve as valuable guides for financial institutions to design credit evaluation models based on the unique characteristics of these enterprises.

A key contribution of this study is its nuanced exploration of the relationships between multiple variables and credit decisions. This comprehensive approach not only fills a research gap but also advances our understanding of the nuanced mechanisms driving credit allocation in Turkey's service sector. Additionally, a comparative analysis of the three variable categories enriches the literature by delineating their relative impacts on credit decisions and emphasizing the need for a multifaceted approach to credit evaluations.

Despite its narrow focus on SMEs in a specific sector, this study offers actionable insights for financial institutions seeking to refine credit allocation mechanisms. This underscores the importance of crafting credit evaluation models tailored to the idiosyncratic needs and characteristics of SMEs in the service sector.

In conclusion, this study posits that non-financial variables have a more pronounced influence on credit decisions than financial variables. This counterintuitive finding paves the way for future research to further explore the relevance of non-financial metrics in credit decision-making, a relatively underexplored area in the existing literature. Overall, this study extends our understanding of the multifaceted influences on credit decisions in the service sector and lays the groundwork for subsequent investigations in other industries.

Theoretically, this research challenges and extends existing credit allocation theories by emphasizing non-financial variables, thereby enriching the literature on financial decision-making. However, the study's focus on Turkey's service sector and a limited sample size calls for further research in diverse settings to validate these findings universally. Such explorations could broaden the theoretical and practical understanding of credit risk assessments globally.

**Author Contributions:** Conceptualization: A.İ.Ç.; Methodology: A.E.Ç.; Software: A.İ.Ç.; Formal analysis: A.E.Ç.; Validation: A.İ.Ç.; Resources: A.İ.Ç.; Original draft Preparation: A.İ.Ç. and S.E.A.; Writing—review and editing: S.E.A.; Supervision: S.E.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** S. Ejaz Ahmed was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

**Data Availability Statement:** Company data were obtained from a private bank's data system to ensure compliance with necessary regulations. The confidentiality of private information was strictly maintained, and data sharing was prohibited.

**Acknowledgments:** The authors thank S. Liu for his time and effort in improving the presentation and quality of this paper.

**Conflicts of Interest:** The authors declare no financial or personal relationships that could inappropriately influence or bias the content of this study.

## References


- Altuğ, Fatih. 2010. Finansal analiz sürecinde sistematik bir yaklaşım ve öneriler (A Systematic Approach and Recommendations in the Financial Analysis Process). Ph.D. thesis, Marmara Üniversitesi, Istanbul, Turkey.
- Bolkvadze, Besik. 2019. The Importance of Financial Ratios in Financial Analysis of Business Entities. *Globalization and Business* 3: 154–157. [CrossRef]

- Ceran, M. 2019. Bankacılıkta Dijitalleşme Kapsamında Öğrenen Yapay Zekâ Desteğiyle Sorunlu Kredilerin Belirlenmesi (Identifying Non-Performing Loans with the Support of Learning Artificial Intelligence within the Scope of Digitalization in Banking). Ph.D. thesis, Marmara Üniversitesi, Social Science Institute, İstanbul, Turkey; pp. 122–28.
- Edem, Daniel Basse. 2017. Liquidity management and performance of deposit money banks in Nigeria (1986–2011): An investigation. *International Journal of Economics, Finance and Management Sciences* 5: 146–61. [CrossRef]
- Erdinç, Nilüfer Yücedağ. 2020. Borsa İstanbul'da İşlem Gören İmalat İşletmelerinin Kârlılığını Etkileyen İşletmeye Özgü ve Makroekonomik Değişkenlerin Analizi (Analysis of Business-Specific and Macroeconomic Variables Affecting the Profitability of Manufacturing Businesses Traded in Borsa İstanbul). *Üçüncü Sektör Sosyal Ekonomi Dergisi* 55: 2109.
- Geçer, Turgay. 2014. Kredi İstihbaratı. *Sosyal Bilimleri Dergisi* 25: 21–35.
- Hossain, Md Junayed. 2023. Implementation of Big Data Analytics in Credit Risk Management in the Banking and Financial Services Sector: A Contemporary Literature Review. Available online: <https://ssrn.com/abstract=4441658> (accessed on 1 September 2023). [CrossRef]
- İnamoğlu, Yusuf. 2013. Türkiye'de Hizmet Sektörünün Gelişimi Ve Ekonomik Büyümeye Etkisi (The Development of the Service Sector in Turkey and Its Impact on Economic Growth). Master's thesis, Bülent Ecevit Üniversitesi, Social Science Institute, Zonguldak, Turkey; pp. 1–2.
- İş Bankası. 2012. *İstihbarat*. İstanbul: Eğitim Müdürlüğü Yayınları.
- Jasevičienė, Filomena, Bronius Povilaitis, and Simona Vidzbelytė. 2013. Commercial banks performance 2008–2012. *Business Management and Economics Engineering* 11: 189–208. [CrossRef]
- Konstantinidis, Christos V., Anastasia Tsolaki, and Nikolaos Giovanis. 2021. Estimating Competitiveness Relations Between Firms of a Multinational Group of Clothing And Footwear Manufacturing Industry In Greece. *TEL* 4: 789–802. [CrossRef]
- Lam, Weng Siew, Weng Hoe Jaaman, and Kah Fai Liew. 2021. Performance Evaluation of Construction Companies Using Integrated Entropy–fuzzy Vikor Model. *Entropy* 3: 320. [CrossRef] [PubMed]
- Mbona, Reginald Masimba, and Kong Yusheng. 2019. Financial statement analysis: Principal component analysis (PCA) approach case study on China telecoms industry. *Asian Journal of Accounting Research* 4: 233–45.
- Melnyk, Mariana, Iryna Leshchukh, Tetyana Medynska, and Nadiya Rushchshyn. 2020. Potential of the sector of financial services in view of the socio-economic growth of Ukrainian regions. *Economic Annals-XXI* 185: 144–54.
- Mercan, Mehmet. 2013. Kredi Hacmindeki Değişimlerin Ekonomik Büyümeye Etkisi: Türkiye Ekonomisi İçin Sınır Testi Yaklaşımı (The Effect of Changes in Credit Volume on Economic Growth: Bounds Test Approach for Turkish Economy). *TBB Bankacılar Dergisi* 84: 54–71.
- Shapiro, Samuel S., Martin B. Wilk, and Hwei J. Chen. 1968. A comparative study of various tests for normality. *Journal of the American Statistical Association* 63: 1343–72. [CrossRef]
- Vakıfbank. 2011. *İstihbarat, Mali Tahlil ve Skorlama Yönetmeliği (Financial Analysis and Scoring Regulation)*. Ankara: Law of Vakıfbank.
- Villalpando, Mario. 2014. Bank credit and productivity: Evidence from Mexican firms. *Revista Mexicana de Economía y Finanzas. Nueva Época/Mexican Journal of Economics and Finance* 9: 195–214.
- World Bank. 2014. *The Use of Non-Financial Information in Credit Scoring: A Review of the Literature*. Washington, DC: World Bank.
- Xu, Yao-Zhi, Jian-Lin Zhang, Ying Hua, and Lin-Yue Wang. 2019. Dynamic credit risk evaluation method for e-commerce sellers based on a hybrid artificial intelligence model. *Sustainability* 11: 5521. [CrossRef]
- Yan, Ying, and Bo Li. 2023. The Research in Credit Risk of Micro and Small Companies with Linear Regression Model. In *Advances in Swarm Intelligence. ICSI 2023*. Edited by Ying Tan, Yuhui Shi and Wenjian Luo. Lecture Notes in Computer Science. Cham: Springer, vol. 13969. [CrossRef]
- Yürük, A. T. 2006. *Banka ve Sigorta Hukuku (Banking and Insurance Law)*. Eskişehir: Anadolu Üniversitesi Yayın, p. 63.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

# Board Gender Diversity and Firm Performance: Recent Evidence from Japan

Kangyi Wang <sup>1</sup>, Jing Ma <sup>1,2</sup>, Chunxiao Xue <sup>1,2,3</sup> and Jianing Zhang <sup>1,2,3,\*</sup> 

<sup>1</sup> College of Business and Public Management, Wenzhou-Kean University, Wenzhou 325060, China; kangyiw0718@163.com (K.W.); majing@wku.edu.cn (J.M.); chunxiaox@wku.edu.cn (C.X.)

<sup>2</sup> Center for Big Data and Decision-Making Technologies, Wenzhou-Kean University, Wenzhou 325060, China

<sup>3</sup> Quantitative Finance Research Institute, Wenzhou-Kean University, Wenzhou 325060, China

\* Correspondence: jianingz@wku.edu.cn

**Abstract:** Gender diversity is increasingly recognized as a critical element in corporate management. However, existing research on its impact on firm performance demonstrates inconsistency in a global context. This study employs 1990 publicly listed Japanese companies from 2006 to 2023 and examines the effect of board gender diversity on firm performance in Japan. Findings from the fixed-effects regression model revealed a significant negative impact of board gender diversity on firm performance. This adverse correlation is more pronounced in smaller firms, those with greater leverage and reduced institutional ownership, and regulated and consumer-focused industries, particularly pre-COVID-19. The detrimental impact of board gender diversity on firm performance is transmitted via corporate social responsibility and firm innovation instead of board independence or CEO duality. Notably, the two-stage least squares estimation addresses potential endogeneity, employing an equal opportunity policy as an instrumental variable. Moreover, the robustness of our results is affirmed via the substitution of return on equity for return on assets as an indicator of firm performance. Lastly, our analysis does not reveal a U-shaped nonlinear relationship between board gender diversity and corporate performance. As Japan progressively promotes women's participation in corporate governance, this research bears significant implications for corporate leaders, investors, and policymakers in Japan.

**Keywords:** gender diversity; firm performance; corporate governance; fixed-effects regression; two-stage least squares; instrumental variable; Chow's test; Japan

**JEL Classification:** G30; J16; M14



**Citation:** Wang, Kangyi, Jing Ma, Chunxiao Xue, and Jianing Zhang. 2024. Board Gender Diversity and Firm Performance: Recent Evidence from Japan. *Journal of Risk and Financial Management* 17: 20. <https://doi.org/10.3390/jrfm17010020>

Academic Editor: Thanasis Stengos

Received: 20 November 2023

Revised: 25 December 2023

Accepted: 2 January 2024

Published: 5 January 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Board attributes have consistently garnered extensive research interest as a pivotal intrinsic element of corporate governance. Recently, the gender diversity of corporate boards has elicited heightened academic focus. Credit Suisse's (2021) Gender 3000 report reveals that the global proportion of female directors on corporate boards escalated from 15.1% in 2015 to 24.0% in 2021, denoting a 59% augmentation. Specifically, this metric surged from 3.6% to 11.5% in Japan, a 219% increase. In comparison, it rose from 23.5% to 34.4% in Europe, while in North America, it advanced from 17.5% to 28.6% (Credit Suisse 2021). Consequently, Japan emerges as the leader in accelerating gender diversity. Its influence may diverge from that in Europe and North America. Thus, comprehending the implications of board gender diversity in Japan is imperative.

In the present study, we delineate four motivations underpinning the investigation. Firstly, Japan has witnessed an extraordinary surge in the ratio of female directors on corporate boards over a notably condensed timeframe. Such accelerated evolution affords a singular lens to examine the repercussions of this rapid alteration in board structure.

EgonZehnder (2022) delineates that the presence of at least one female director on Japanese boards has augmented by 21.4% relative to 2020, starkly contrasting to the global median escalation of merely 4.7%, highlighting Japan's significant progress in the recent two years. Secondly, Japan's corporate culture has traditionally been male-dominated, and patriarchal norms often characterize its society. Therefore, the increasing presence of female directors in such an environment could have different implications than in other countries with different cultural and corporate dynamics. Thirdly, the Japanese administration has overtly endeavored to bolster female participation in leadership roles. As per a draft plan by the Gender Equality Bureau (Reynolds 2023), Japan aspires for women to constitute at least 30% of corporate directorships by 2030. This governmental pledge and its ensuing influence on corporate stewardship and efficacy present a fertile terrain for exploration. Fourthly, Japan is one of the world's largest economies. Understanding how gender diversity impacts corporate performance in such a significant economy can offer valuable insights for economic policies and corporate strategies within and outside Japan.

Numerous investigations have examined the influence of female directors on firm performance, yet their conclusions vary, which are attributable to disparate societal norms, attitudes toward women, and supportive policies across nations. A wealth of empirical research indicates that board gender diversity positively correlates with firm performance. Notably, Carter et al.'s (2003) analysis of 1000 Fortune-selected firms in 1997, Terjesen et al.'s (2016) examination of 3876 publicly traded entities across 47 nations, and Liu et al.'s (2014) study of over 2000 listed companies in China from 1999 to 2011 all affirm this beneficial effect of female directorship on firm performance. Conversely, other studies present differing views, suggesting female board members' detrimental or negligible impact on firm performance. For instance, Carter et al. (2010) identified no significant link between female directors and the financial performance of major U.S. corporations. Adams and Ferreira (2009) found a negative association and posited that increased gender diversity on boards might lead to excessive governance in these firms.

While there is a considerable corpus of literature on board gender diversity, its specific examination in Japan, especially concerning the influence of female directors on corporate performance, remains underexplored. Japan has recently shown a commitment to enhancing female representation in boardrooms. Prime Minister Shinzo Abe, in his address at the Global Leaders Meeting on Gender Equality and Women's Empowerment on 27 September 2015, stated an objective to have women fill about 30% of leadership roles in Japanese society by 2020. Nevertheless, this target was not attained by the designated year. Morgan Stanley Capital International's (2020) report suggests that maintaining the current trajectory, achieving a 30% female representation on corporate boards might be realized by 2029, with a potential to reach 50% by 2045. Nonetheless, Japan's patriarchal societal structure continues to present significant barriers to women's professional ascension.

This study utilizes a sample of 1990 listed Japanese companies from 2006 to 2023 to explore the nexus between board gender diversity and corporate performance, offering novel insights into Japanese corporate governance. Our empirical findings, derived from the fixed-effects regression model, indicate that board gender diversity adversely affects corporate performance in Japan. This diversity is quantified by the proportion of female directors and a binary variable denoting their presence on the board. To address potential endogeneity, we apply a two-stage least squares (2SLS) regression model, which corroborates the negative impact of board gender diversity on firm performance. This outcome persists even when altering corporate performance metrics from return on assets (ROA) to return on equity (ROE). In addition, our analysis does not reveal a nonlinear quadratic relationship between board gender diversity and corporate performance. The detrimental impact of board gender diversity on firm performance is more marked in smaller companies compared to larger ones, in firms with higher leverage as opposed to those with lower leverage, in firms with diminished institutional ownership relative to those with augmented ownership, in regulated and consumer-oriented industries in contrast to innovation-driven industries and was notably more pronounced pre-COVID-19

than during the COVID-19 period. The mediating effects are more pronounced via environmental, social, and governance (ESG) factors and weakly via research and development (R&D) rather than board independence and CEO duality.

Our study contributes to the literature on corporate governance in Japan. Despite numerous global studies exploring the correlation between board gender diversity and corporate performance, the results remain inconsistent. Moreover, given Japan's leading pace in augmenting gender diversity, comprehending its gender-specific effects is vital. However, this aspect has scarcely been the focus of academic scrutiny in Japan. Therefore, our study addresses this gap in the literature concerning the influence of board gender diversity on Japanese corporate performance. Additionally, how firm size, leverage, institutional ownership, and sector classification moderate the impact of gender diversity on firm performance in Japan remains unexplored. Furthermore, the COVID-19 pandemic, an unprecedented global health emergency, necessitates additional exploration of its implications for gender diversity and firm performance in Japan.

The subsequent sections of this paper are structured as follows: Section 2 presents a background of corporate governance and board gender diversity in Japan. Section 3 examines the theoretical framework regarding the impact of board gender diversity on firm performance. Section 4 delves into a review of the extant literature and formulates our hypotheses. Section 5 delineates the data and regression models employed in our analysis. Section 6 presents and discusses the regression results. Finally, Section 7 concludes our research, synthesizing our findings and implications.

## **2. Background**

Japanese corporate governance has traditionally emphasized long-term relationships and consensual decision-making, epitomized by the "keiretsu" system (Aman et al. 2021). Historically, Japan has adhered to a stakeholder-centric governance model, privileging the needs of a broad array of stakeholders over shareholder primacy. Japanese firms have been intricately connected with their primary banks, suppliers, and clients, fostering robust, long-standing alliances. This interdependence between companies and stakeholders has distinctly influenced Japan's corporate governance, setting it apart from Western countries. Nevertheless, Japan's corporate governance landscape has witnessed substantial transformations in response to evolving global contexts. Efforts have been made to align Japanese corporate governance norms with global standards. The 2021 revision of Japan's Corporate Governance Code marked a significant step in this direction, enhancing board independence, fostering diversity, and emphasizing sustainability and ESG considerations (Sawaji 2021).

Augmenting board diversity can be attained by enhancing gender diversity. The Japanese government has set a target of achieving 30% female representation on the boards of companies listed on the prime market by 2030 (Reynolds 2023). The Global Gender Gap Report 2023 indicates that Japan's progress in gender equality lags behind its G7 counterparts (World Economic Forum 2023). In 2023, Japan scored 0.65 in the gender gap, positioning it 125th among 146 countries assessed in the report. This figure is markedly below the G7 average of 0.76. The gender inequality in Japan predominantly arises from women's limited participation in the workforce and scarce representation in political spheres. Per World Bank's (2023) data, the proportion of female to male labor force participation escalated from 64% in 2000 to 76% in 2022. The Gender Equality Bureau's (2022) analysis of gender diversity reveals substantial progress: the proportion of Tokyo Exchange-listed companies lacking female board members has markedly reduced, dropping from 84% in 2013 to 18.7% in 2022. However, women occupied 21.3% of managerial roles and a mere 6.2% of board positions in 2021 (Sawaji 2021). Furthermore, while the gender gap in school enrollment is minimal, a significant gap persists in higher education, particularly at the postgraduate level, where Japan reports the lowest proportion of female master's graduates among OECD countries (OECD 2023).

While Japan has exerted efforts to enhance corporate governance and advance gender diversity, considerable progress is yet to be realized. Despite these initiatives, entrenched gender norms and societal expectations remain impediments to enhancing gender diversity (Binder et al. 2019). Cultural and infrastructural transformations are gradual processes, necessitating persistent endeavors to guarantee enduring advancements in these domains. The Japanese government might encounter obstacles in fulfilling the 30% female directorship objective by 2030, which is attributable to a limited pool of qualified female candidates.

### **3. Theoretical Framework**

#### *3.1. Resource Dependence Theory*

Numerous theories address the influence of board gender diversity on corporate performance. The resource dependence theory, a sociological and organizational concept, argues that organizations require external resources for success and sustainability (Pfeffer and Salancik 1978). Consequently, companies strive to appoint directors capable of providing these essential resources. Prior research indicates that boards with diverse membership amalgamate individuals with varied backgrounds, skills, experiences, expertise, and viewpoints, creating a more extensive resource base. This diversity facilitates more effective decision-making and improves corporate outcomes (Chan and Li 2008; Berger et al. 2014; Delis et al. 2017; Kim and Starks 2016). As women increasingly contribute to societal roles, female directors offer new resources, enabling firms to adapt to contemporary challenges. For instance, Brahma et al. (2021) analyzed FTSE 100 companies in the UK and observed a positive correlation between board gender diversity and firm performance. From a legitimacy standpoint, a gender-diverse board potentially enhances a firm's interactions with stakeholders, including customers, employees, and communities. With the growing prominence of female consumers, gender diversity on boards can help maintain relationships with female clientele or comprehend female consumer purchasing patterns (Süssmuth-Dyckerhoff et al. 2012).

#### *3.2. Agency Theory*

Agency theory, as proposed by Jensen and Meckling (1976), delves into potential conflicts of interest arising from the division of ownership and control between principals (shareholders) and agents (management). A critical function of directors is to alleviate these agency issues via managerial monitoring. Within gender diversity, agency theory facilitates an exploration into whether female directors enhance managerial monitoring efficiency. One segment of literature posits that boardroom gender diversity positively influences corporate performance due to increased vigilance from women directors, the introduction of novel viewpoints, and the avoidance of entrenched "old boys' networks" (Adams and Ferreira 2009; Lara et al. 2017; Gul et al. 2011). Conversely, another body of literature contends that board gender diversity adversely affects companies, attributed either to a scarcity of suitably qualified female directors (Ahern and Dittmar 2012; Bøhren and Staubo 2014) or to potential over-monitoring by women directors (Adams and Ferreira 2009).

#### *3.3. Behavioral Theory*

The behavioral theory of the firm, as articulated by Cyert and March (1963), proposes that firm decision-makers are constrained by their capabilities. Research on group diversity suggests that member heterogeneity can stimulate information processing and enhance problem-solving (Hoffman and Maier 1961; Van Knippenberg and Schippers 2007), leading to heightened innovation efficiency and improved performance (Chen et al. 2018; Alesina and La Ferrara 2005). However, counterarguments exist, contending that diversity may escalate communication expenses and even foster conflicts, thereby deteriorating performance (Wagner et al. 1984; Zenger and Lawrence 1989; Alesina and La Ferrara 2005).



### 3.4. Critical Mass Theory

Critical mass theory emphasizes the need to attain minimum female representation in the boardroom. This threshold, commonly termed critical mass, is deemed crucial for an organization to reap the benefits of gender diversity (Kanter 1977). Absent from this critical mass, including one or two women on a board might be perceived as tokenistic or symbolic merely to satisfy regulatory requirements. As a result, the effectiveness and influence of female directors can be diminished and marginalized in a predominantly male boardroom (Schwartz-Ziv 2017; Konrad et al. 2008). Conversely, appointing three or more women to a board yields more significant contributions and notable positive impacts (Owen and Temesvary 2018). Recent research has pivoted toward identifying this threshold and examining the veracity of the critical mass theory. The threshold is frequently defined as at least three or 30% female directors, equating to roughly one-third of most boards (Torchia et al. 2011; Joecks et al. 2013). In light of these findings, several countries are adopting affirmative measures by implementing gender quotas of 30–40% in boardrooms (Terjesen and Sealy 2016). Nonetheless, critics of these policies argue that companies remain dubious about the efficacy of such regulations, their alignment with corporate structures, and the variability in social, cultural, and legal nuances across different nations (Carter et al. 2010).

In summary, theoretical models forecast both advantageous and detrimental effects of board gender diversity on corporate performance, with empirical studies yielding mixed outcomes.

## 4. Literature Review and Hypotheses Development

The correlation between board gender diversity and corporate performance constitutes a significant and debated topic. The subsequent sections comprehensively review pertinent empirical research in this domain.

### 4.1. Positive Impact of Board Gender Diversity on Firm Performance

Numerous country-specific analyses substantiate the beneficial impact of board gender diversity on corporate performance, with evidence from Mauritius (Mahadeo et al. 2012), China (Liu et al. 2014), France (Sabatier 2015), the UK (Brahma et al. 2021), Russia (Garanina and Muravyev 2021), and India (Sanan 2016; Sarkar and Selarka 2021). These empirical investigations employ accounting performance measures such as ROA and ROE, market performance metric Tobin's Q (Tobin 1969), or a blend of these indicators to assess corporate performance. They consistently illustrate a positive correlation between enhanced corporate performance and an increased proportion of female directors on boards.

Findings from several multi-country investigations also indicate that gender diversity on boards enhances corporate performance. Low et al. (2015) conducted an extensive analysis of board diversification and corporate performance in East Asia, assessing firms in Hong Kong, South Korea, Malaysia, and Singapore. Their study reveals a positive influence of female directors on ROE, particularly in nations where cultural norms limit women's economic involvement. Belaounia et al. (2020), examining listed companies across 24 countries, ascertain that firms with a higher fraction of female directors exhibit superior overall performance, with the addition of a female board member boosting ROA and Tobin's Q. Terjesen et al.'s (2016) research on companies from 47 countries demonstrates that gender-diverse boards significantly enhance corporate performance, with increases in the percentage of female directors correlating with improvements in Tobin's Q and ROA. Pucheta-Martínez and Gallego-Álvarez (2020) analyzed firms from 34 countries and confirmed that the presence of women on boards is associated with better firm performance. In light of the literature reviewed, we propose our first hypothesis.

**Hypothesis 1.** *Board gender diversity has a positive impact on firm performance.*

#### 4.2. Negative Impact of Board Gender Diversity on Firm Performance

Various empirical studies across different national contexts support the notion that board gender diversity negatively impacts corporate performance. Shehata et al. (2017) examine UK-listed companies using four gender diversity measures, all indicating a significant negative correlation with corporate performance. Mirza et al. (2012) analyze a sample of Pakistani companies, discovering negative correlations between female directorship and performance indicators such as ROE and ROA, attributing this to potential information deficits, risk aversion, and societal barriers women face. Similarly, Akram et al. (2020) observe that female directors in Pakistani firms lead to reduced corporate value. In Malaysia, Ahmad et al. (2020) report that an increased proportion of female directors correlates with a decline in ROA. Likewise, Lim et al. (2019) find a negative impact of female directors on Tobin's Q, and Abdullah (2014) identifies a significant negative relationship between board gender diversity and ROA and Tobin's Q. Based on the literature discussed above, we propose our second hypothesis.

**Hypothesis 2.** *Board gender diversity has a negative impact on firm performance.*

#### 4.3. Neutral Impact of Board Gender Diversity on Firm Performance

Country-specific investigations suggest a neutral link between board gender diversity and corporate performance. Kagzi and Guha (2018), assessing listed Indian companies, observe no significant influence of board gender diversity on company performance before and after implementing the 2013 Companies Act, which mandated certain levels of board gender diversity. This finding aligns with earlier studies. Marinova et al. (2016), examining firms in the Netherlands and Denmark, indicate that board gender diversity bears no correlation with corporate performance. Yasser (2012), in an analysis of Pakistani listed companies, detects no association between board gender diversity and corporate performance. Likewise, research in other nations corroborates this absence of correlation. In the United States, Carter et al. (2010) find no empirical evidence supporting a positive or negative causal link between board gender diversity and corporate performance. Ararat and Yurtoglu (2021) investigated Turkish-listed companies. They ascertained no effect of female board presence on Tobin's Q. Similarly, Unite et al. (2019), studying Philippine companies, conclude that board gender diversity does not significantly affect ROA, ROE, or Tobin's Q. These observations underpin our third hypothesis.

**Hypothesis 3.** *Board gender diversity has a neutral impact on firm performance.*

While the scholarly community remains engaged with the effects of board gender diversity on corporate performance, the vast array of empirical studies yields divergent outcomes without a clear consensus. The existing research on the interplay between corporate gender diversity and firm performance in Japan is notably scarce. For instance, Nakagawa and Schreiber (2014), utilizing data from Toyo Keizai and Nikkei NEEDS on 745 Japanese-listed companies, identify a significant positive correlation between firm performance and the ratio of female managers and gender diversity. However, their dataset is dated and no longer reflective of Japan's current gender diversity landscape. Another investigation by Tanaka (2019) suggests that outside female directors enhance firm performance, yet this study focuses primarily on the factors leading to female directorship rather than their impact. Additionally, Tanaka's research, covering the period from 2006 to 2015, does not represent more recent trends. Our study, examining Japanese firms in the recent timeframe of 2006–2023, presents contrasting findings to the two studies above by demonstrating a negative effect of board gender diversity on corporate performance in Japan.

## 5. Research Design

### 5.1. Data Sample

The board composition and financial metrics of Japanese publicly traded firms were extracted from Bloomberg Terminals. Table 1 summarizes the definitions of these variables. The dataset obtained from Bloomberg includes ROA, ROE, market capitalization, total assets, total debts, fixed assets, the total number of directors, the number of female directors, the number of independent directors, the director age, the dual role of the CEO as board chairman, the firm’s explicit commitment to non-discrimination practices, the cash holding, the institutional ownership, R&D, and ESG scores. The final sample includes 25,363 firm-year observations, spanning 2006 to 2023, representing 1990 Japanese entities listed on the Tokyo Stock Exchange.

**Table 1.** Variable definitions.

<b>Variable</b>	<b>Definition</b>
<i>ROA</i>	The net income divided by the total assets.
<i>ROE</i>	The net income divided by the shareholder’s equity.
<i>MktCapChg</i>	The annual percentage change in the market capitalization.
<i>FemaleFrac</i>	The number of female directors divided by the total number of directors
<i>FemaleDum</i>	The dummy variable equals one in the presence of at least one female director and zero in its absence.
<i>FirmSize</i>	The natural logarithm of the total assets.
<i>FirmLev</i>	The total debts divided by the total assets.
<i>Tangibility</i>	The fixed assets divided by the total assets.
<i>BoardSize</i>	The total number of directors.
<i>BoardInd</i>	The number of independent directors divided by the total number of directors.
<i>DirAge</i>	The average director’s age.
<i>Duality</i>	The dummy variable is set to one if the company’s CEO also serves as the board chair; alternatively, it takes a value of zero.
<i>EqOpp</i>	The dummy variable is assigned a value of one if the firm explicitly commits to non-discrimination against any group of people; in other cases, it is set to zero.
<i>CashHold</i>	The cash and cash equivalents divided by the total assets.
<i>InstiOwn</i>	Institutional ownership measures the percentage of a company’s outstanding shares that institutional investors hold.
<i>RD</i>	The research and development expenditure divided by the net sales.
<i>ESG</i>	A metric that evaluates a company’s performance in three key areas: environmental, social, and governance.

The table summarizes the definitions of the variables, where the variable names are italicized.

### 5.2. Fixed-Effects Model

In our analysis, we employed the fixed-effects model, Chow’s (1960) test, and the 2SLS model to examine the effect of board gender diversity on corporate performance. Each firm possesses distinct attributes, such as management style, corporate culture, or brand reputation, which may not be directly quantifiable or observable. As shown below, the firm fixed-effects model accommodates these unseen characteristics, presuming their constancy over time.

$$Perf_{i,t} = \beta_0 + \beta_1 Female_{i,t} + \beta_2 FirmSize_{i,t} + \beta_3 FirmLev_{i,t} + \beta_4 Tangibility_{i,t} + \beta_5 BoardSize_{i,t} + \beta_6 BoardInd_{i,t} + \beta_7 DirAge_{i,t} + \beta_8 Duality_{i,t} + FirmFE + \varepsilon_{i,t} \quad (1)$$

where the subscript  $i$  represents firm  $i$ , and the subscript  $t$  represents year  $t$ .  $Perf$  denotes firm performance proxied by ROA or ROE.  $Female$  denotes the board gender diversity, proxied by  $FemaleFrac$  or  $FemaleDum$ .  $FirmFE$  denotes the firm-fixed-effects. The definitions for all other control variables in Equation (1) are provided in Table 1.

### 5.2.1. Dependent Variable

Per empirical research, ROA is widely utilized as a metric for corporate performance (Adams and Ferreira 2009; Sanan 2016; Terjesen et al. 2016; Brahma et al. 2021; Sarkar and Selarka 2021). Aligning with these studies, ROA is employed in our analysis to gauge firm performance. As an accounting-based metric, ROA represents a company's net income proportion to its total assets. Barber and Lyon (1996) highlighted ROA's merits in evaluating corporate performance. They reveal that ROA facilitates comparative analysis of one company's performance against others. Furthermore, García-Meca et al. (2015) contended that the application of ROA enables the examination of potential market irregularities that might impede the complete, accurate reflection of information in stock prices.

Additionally, we utilize ROE for robustness assessments in measuring corporate performance. ROE, another accounting-based metric, is the ratio of net income to shareholder's equity. This application of ROE aligns with preceding studies on firm performance (Low et al. 2015; Sabatier 2015; Garanina and Muravyev 2021).

### 5.2.2. Explanatory Variables

Board gender diversity is measured via two approaches: (1) the proportion of female directors on the board, calculated by dividing the number of female directors by the total number of directors, and (2) the dummy variable, set to one in the presence of at least one female director, and zero in its absence.

### 5.2.3. Control Variables

In our analysis, control variables are bifurcated into two classifications: firm characteristics and board characteristics. The control variables about firm characteristics encompass firm size, financial leverage, and asset tangibility. Those relating to board characteristics include board size, board independence, average director age, and the dual role of the CEO as board chairman.

The initial category of control variables pertains to firm characteristics. This research quantifies firm size using the natural logarithm of total assets. Doğan (2013) demonstrated a positive correlation between firm size and performance. Financial leverage, the debt ratio, is calculated as total debts over total assets. Das et al. (2022) identified a negative influence of firm leverage on performance. Asset tangibility is derived by dividing fixed assets by total assets. Lee (2010) presented findings indicating a negative effect of fixed asset capital intensity on firm performance.

The second set of control variables relates to board characteristics. Existing empirical research demonstrates that board size adversely affects corporate performance. Conyon and Peck (1998) showed that the correlation between board size and company performance is typically negative. Guest (2009) similarly reported a significant negative effect of board size on firm performance. The influence of independent directors on company performance has been thoroughly investigated in corporate governance literature, yielding mixed outcomes (Aluchna et al. 2020; Reguera-Alvarado and Bravo 2017; Zeng 2018).

### 5.3. Chow's Test

Chow's (1960) test is a statistical test used to determine whether there are significant differences in the intercepts and slopes of two linear regressions across different subgroups. For example, in contrasting regression coefficients between small and large firm subgroups, we designate the *FSD* as one for firms surpassing the median size in a given year and zero for those below. Subsequently, we undertake the prescribed Chow's test by integrating a sequence of interactions with *FSD*.

$$\begin{aligned}
 Perf_{i,t} = & \beta_0 + \beta_1 Female_{i,t} + \beta_2 FirmSize_{i,t} + \beta_3 FirmLev_{i,t} + \beta_4 Tangibility_{i,t} \\
 & + \beta_5 BoardSize_{i,t} + \beta_6 BoardInd_{i,t} + \beta_7 DirAge_{i,t} + \beta_8 Duality_{i,t} \\
 & + \theta_0 FSD_{i,t} + \theta_1 (FSD_{i,t} \times Female_{i,t}) + \theta_2 (FSD_{i,t} \times FirmSize_{i,t}) + \theta_3 (FSD_{i,t} \times FirmLev_{i,t}) \\
 & + \theta_4 (FSD_{i,t} \times Tangibility_{i,t}) + \theta_5 (FSD_{i,t} \times BoardSize_{i,t}) + \theta_6 (FSD_{i,t} \times BoardInd_{i,t}) \\
 & + \theta_7 (FSD_{i,t} \times DirAge_{i,t}) + \theta_8 (FSD_{i,t} \times Duality_{i,t}) + FirmFE + \varepsilon_{i,t}
 \end{aligned} \tag{2}$$

Rather than evaluating the joint hypothesis that all  $\theta$  values are null, we focus on discerning the differential influence of gender diversity. Hence, we examine the null hypothesis asserting  $\theta_1$  equals zero and subsequently disclose corresponding *F*-values and *p*-values. A comparable methodology is employed for other subgroup comparisons.

### 5.4. Instrumental Variables and 2SLS Model

An endogeneity issue may exist between board gender diversity and corporate performance, suggesting a bidirectional causality: board gender diversity might influence corporate performance, and conversely, corporate performance could impact board gender diversity (Hermalin and Weisbach 2003; Adams and Ferreira 2009). In line with Carter et al. (2003), we employed a 2SLS regression to tackle this endogeneity concern. The regression equations are delineated as follows:

$$\begin{aligned}
 Female_{i,t} = & \beta_0 + \beta_1 EqOpp_{i,t} + \beta_2 FirmSize_{i,t} + \beta_3 FirmLev_{i,t} + \beta_4 Tangibility_{i,t} \\
 & + \beta_5 BoardSize_{i,t} + \beta_6 BoardInd_{i,t} + \beta_7 DirAge_{i,t} + \beta_8 Duality_{i,t} + FirmFE + \varepsilon_{i,t}
 \end{aligned} \tag{3}$$

$$\begin{aligned}
 Perf_{i,t} = & \beta_0 + \beta_1 \hat{Female}_{i,t} + \beta_2 FirmSize_{i,t} + \beta_3 FirmLev_{i,t} + \beta_4 Tangibility_{i,t} \\
 & + \beta_5 BoardSize_{i,t} + \beta_6 BoardInd_{i,t} + \beta_7 DirAge_{i,t} + \beta_8 Duality_{i,t} + FirmFE + \varepsilon_{i,t}
 \end{aligned} \tag{4}$$

where all variables remain identical to those in Equation (1), except *EqOpp* represents a dummy variable assigned one if the firm explicitly pledges non-discrimination toward any group and zero otherwise. In the first stage, Equation (3) employs regression to estimate board gender diversity, utilizing the equal opportunity policy as the instrumental variable. The second stage employs the predicted gender diversity from the first stage,  $\hat{Female}$ , to forecast firm performance in Equation (4). As Adams and Ferreira (2009) noted, identifying an instrumental variable is challenging, given that other governance features pertinent to endogenous issues are already incorporated in the performance regression. Our research selects the equal opportunity policy as an instrumental variable. We posit that firms actively pursuing non-discrimination policies are more inclined to appoint female directors, reflecting a corporate culture less prone to gender bias and discrimination. Additionally, the equal opportunity policy does not directly influence corporate performance.

### 5.5. Nonlinear Quadratic Model

Joecks et al.'s (2013) empirical investigation into the critical mass theory posited that the link between board gender diversity and corporate performance is not linear, potentially following a U-shaped pattern. This theory contends that the unique abilities and skills women contribute to a group become significantly impactful only once their representation reaches a certain critical threshold. Consequently, we explore the potential for a U-shaped correlation between board gender diversity and corporate performance, as delineated below.

$$\begin{aligned}
 Perf_{i,t} = & \beta_0 + \beta_1 FemaleFrac_{i,t} + \beta_2 FemaleFrac_{i,t}^2 + \beta_3 FirmSize_{i,t} + \beta_4 FirmLev_{i,t} \\
 & + \beta_5 Tangibility_{i,t} + \beta_6 BoardSize_{i,t} + \beta_7 BoardInd_{i,t} + \beta_8 DirAge_{i,t} \\
 & + \beta_9 Duality_{i,t} + FirmFE + \varepsilon_{i,t}
 \end{aligned}
 \tag{5}$$

where all variables remain identical to those in Equation (1), except *FemaleFrac*<sup>2</sup> denotes the squared term of *FemaleFrac*.

## 6. Empirical Results and Discussion

### 6.1. Descriptive Statistics

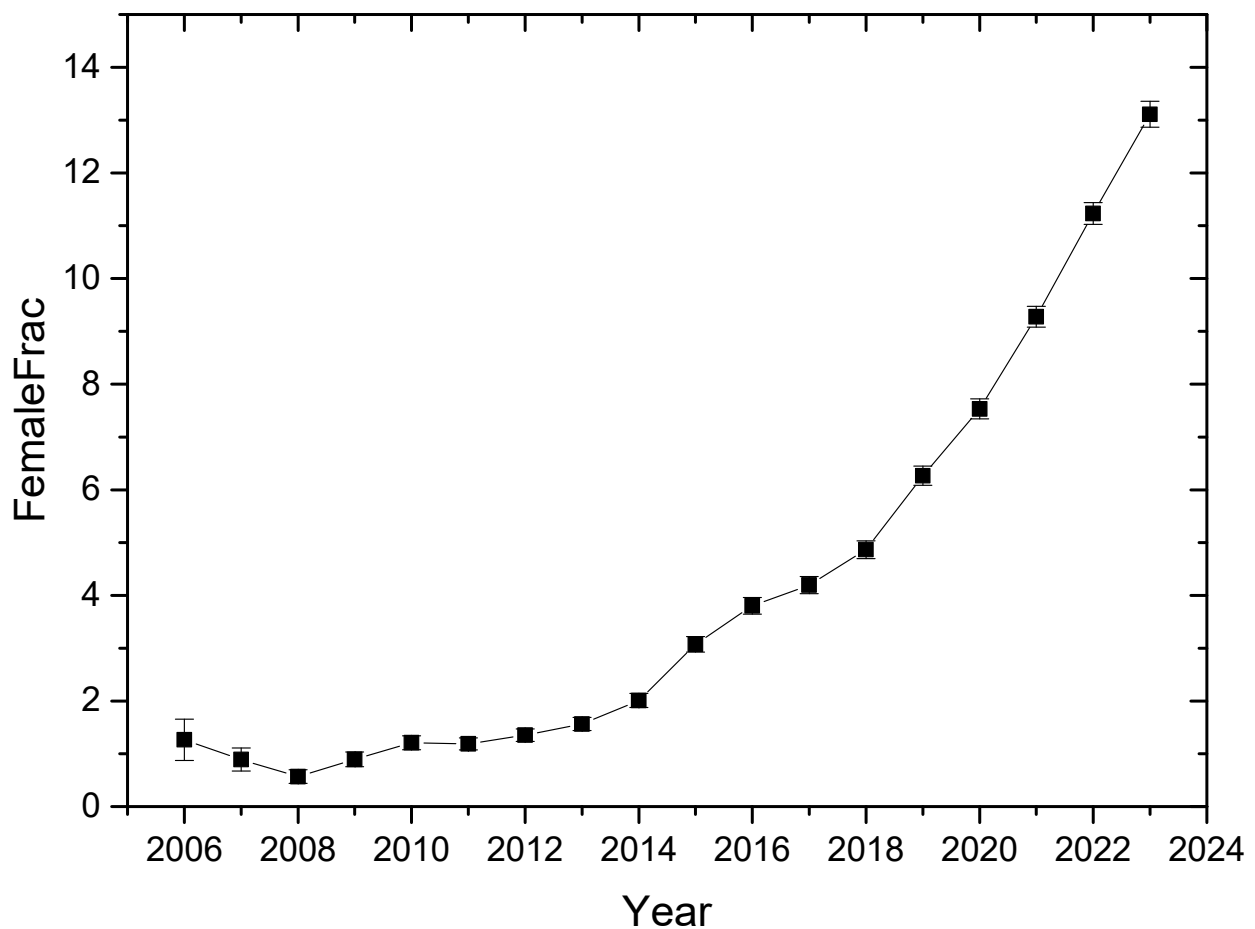
Table 2 presents the descriptive statistics for the variables under study. The mean ROA is recorded at 3.81%, lower than its standard deviation of 4.69%. A meager 5.00% of board directors are female, yet 35.0% of firms have at least one woman on their board. The average financial leverage ratio is calculated to be 18.44%. Asset tangibility is noted at 25.86%. Boards typically comprise about nine directors, with 23.12% classified as independent. The average age of directors is approximately 59.57 years. About 52% of corporations have adopted an equal opportunity policy. Lastly, the variance inflation factor test confirms no multicollinearity concerns in this research, as indicated by all variance inflation factors remaining under five.

**Table 2.** Descriptive statistics of variables.

Variable	Obs.	Mean	Std. Dev.	Min	P25	Median	P75	Max
<i>ROA</i>	25,363	3.813	4.693	−12.572	1.379	3.324	5.820	20.213
<i>ROE</i>	25,363	7.446	9.822	−37.329	3.659	7.114	11.439	37.418
<i>MktCapChg</i>	21,901	0.112	0.391	−0.548	−0.123	0.038	0.252	1.793
<i>FemaleFrac</i>	25,363	5.002	7.832	0.000	0.000	0.000	10.000	96.000
<i>FemaleDum</i>	25,363	0.350	0.477	0.000	0.000	0.000	1.000	1.000
<i>FirmSize</i>	25,363	11.588	1.832	7.925	10.346	11.346	12.615	16.635
<i>FirmLev</i>	25,363	18.443	17.044	0.000	3.615	14.309	28.722	68.934
<i>Tangibility</i>	25,363	25.859	18.254	0.324	11.463	24.256	37.059	76.667
<i>BoardSize</i>	25,363	8.972	2.879	4.000	7.000	9.000	11.000	18.000
<i>BoardInd</i>	25,363	23.122	16.838	0.000	10.000	22.222	33.333	66.667
<i>DirAge</i>	25,363	59.574	4.716	43.690	57.380	60.333	62.667	69.000
<i>Duality</i>	25,363	0.779	0.415	0.000	1.000	1.000	1.000	1.000
<i>EqOpp</i>	22,274	0.517	0.500	0.000	0.000	1.000	1.000	1.000
<i>CashHold</i>	25,363	18.706	14.211	1.402	8.423	14.919	24.751	70.160
<i>InstiOwn</i>	22,263	35.797	18.661	2.341	21.491	33.938	48.938	83.894
<i>RD</i>	23,450	1.707	2.749	0.000	0.000	0.509	2.391	15.640
<i>ESG</i>	5696	2.237	1.044	0.750	1.365	1.995	2.910	5.150

The table reports descriptive statistics of the variables, where the variable names are italicized.

Figure 1 illustrates the temporal progression of *FemaleFrac* from 2006 to 2023 among Tokyo Exchange-listed firms. *FemaleFrac* is the ratio of female directors to the overall director count. Accompanying standard error bars are also depicted. *FemaleFrac* remained subdued until 2012 and escalated exponentially in the recent decade. Despite this rapid growth, the overall level remains below 14% by 2023.



**Figure 1.** Time evolution of the fraction of female board directors in Japan.

### 6.2. Fixed-Effects Regressions

Table 3 presents the regression outcomes from the firm fixed-effects model, following Equation (1). Irrespective of being quantified by the proportion of female directors or via a dummy variable, the findings indicate a negative association between board gender diversity and corporate performance, with the gender diversity coefficient being statistically significant at the 1% or 5% level. In terms of economic importance, a one standard deviation shift in *FemaleFrac*, amounting to 7.832%, correlates with a 0.10% decrease ( $=7.832\% \times 0.013$ ) in ROA, representing approximately 2.7% of the average ROA (3.813%). Similarly, a transition of *FemaleDum* from zero to one corresponds to a 0.156% reduction in ROA, equating to roughly 4.1% of the mean ROA. In summary, the negative impact of board gender diversity on corporate performance is statistically substantial and bears mediocre economic implications.

The reasons for the negative relationship between board gender diversity and firm performance are manifold. Firstly, the presence of female directors may introduce enhanced supervision and excessive oversight, potentially undermining organizational efficacy. Adams and Ferreira (2007) posited that increased oversight could disrupt the flow of communication between directors and management during decision-making processes, adversely impacting firm performance. Moreover, over-monitoring could erode shareholder value (Almazan and Suarez 2003). Secondly, the social identity theory elucidates the dynamics and implications of social identity, including categorizing personal and others' characteristics, such as gender, skin tone, or ethnicity (Abrams and Hogg 2010). Within the context of Japanese culture, women typically occupy comparatively lower status tiers than men, potentially complicating communication and management of this demographic. Female professionals often confront entrenched stereotypes and biases, prompting public

skepticism regarding their leadership capabilities (Thomas 2018). Thirdly, Smith et al. (2006) contended that a gender-diverse board is prone to conflicts, resulting in delayed decision-making processes, whereas the market necessitates prompt reactions. Similarly, Williams Phillips and O'Reilly (1998) argued that gender-diverse groups are more likely to encounter affective conflicts, yielding detrimental effects on team dynamics.

**Table 3.** Fixed-effects regression results.

	(1) ROA	(2) ROA
<i>FemaleFrac</i>	−0.013 *** (0.004)	
<i>FemaleDum</i>		−0.156 ** (0.065)
<i>FirmSize</i>	1.494 *** (0.088)	1.478 *** (0.088)
<i>FirmLev</i>	−0.181 *** (0.003)	−0.181 *** (0.003)
<i>Tangibility</i>	−0.063 *** (0.005)	−0.063 *** (0.005)
<i>BoardSize</i>	0.034 *** (0.012)	0.037 *** (0.012)
<i>BoardInd</i>	0.003 (0.002)	0.002 (0.002)
<i>DirAge</i>	−0.078 *** (0.010)	−0.076 *** (0.010)
<i>Duality</i>	0.018 (0.075)	0.022 (0.075)
<i>Constant</i>	−4.191 *** (1.044)	−4.141 *** (1.045)
Firm FE	Yes	Yes
Observations	25,363	25,363
<i>R-squared</i>	0.146	0.145

The table shows the fixed-effects regression results. The variable names are italicized. The standard errors are reported below the estimated coefficients in parentheses. \*\*\*, \*\*, and \* denotes statistical significance level of 1%, 5%, and 10%, respectively.

### 6.3. Small vs. Large Firms

Table 4 bifurcates our dataset into two subsets based on firm sizes. The smaller firm subsample includes companies whose size falls below the yearly median, while the larger firm subsample comprises those exceeding the median. Subsequently, we apply the fixed-effects regression in line with Equation (1) to these subsamples. The findings indicate that board gender diversity, quantified by the fraction of female directors or as a dummy variable, negatively influences corporate performance, but this effect is predominantly observed in smaller firms. Within this context, the gender diversity coefficient is statistically significant at 1%. Economically, a one standard deviation shift in *FemaleFrac* for smaller firms correlates with a 0.22% reduction in ROA, which is 5.8% of the mean ROA. Altering *FemaleDum* from zero to one in these firms associates with a 0.38% decrease in ROA, amounting to 10.0% of the mean ROA. Chow's test for the divergence between these two coefficients is also significant at the 1% and 10% thresholds. Collectively, the results in Table 4 suggest that the negative relationship between board gender diversity and corporate performance is more pronounced in smaller-sized firms.



**Table 4.** Small vs. large firms.

	(1) Small Firms ROA	(2) Large Firms ROA	(3) Small Firms ROA	(4) Large Firms ROA
<i>FemaleFrac</i>	−0.028 *** (0.006)	0.000 (0.005)		
<i>FemaleDum</i>			−0.383 *** (0.111)	−0.103 (0.075)
<i>FirmSize</i>	1.815 *** (0.141)	1.405 *** (0.125)	1.791 *** (0.141)	1.438 *** (0.124)
<i>FirmLev</i>	−0.203 *** (0.005)	−0.163 *** (0.004)	−0.204 *** (0.005)	−0.163 *** (0.004)
<i>Tangibility</i>	−0.058 *** (0.007)	−0.067 *** (0.006)	−0.058 *** (0.007)	−0.067 *** (0.006)
<i>BoardSize</i>	0.087 *** (0.023)	0.022 * (0.013)	0.098 *** (0.024)	0.023 * (0.013)
<i>BoardInd</i>	−0.003 (0.003)	0.001 (0.003)	−0.004 (0.003)	0.003 (0.003)
<i>DirAge</i>	−0.114 *** (0.015)	−0.022 * (0.013)	−0.112 *** (0.015)	−0.024 * (0.013)
<i>Duality</i>	0.109 (0.151)	0.055 (0.077)	0.118 (0.151)	0.051 (0.077)
<i>Constant</i>	−3.381 ** (1.493)	−8.724 *** (1.643)	−3.345 ** (1.495)	−9.012 *** (1.645)
Firm FE	Yes	Yes	Yes	Yes
Observations	12,676	12,687	12,676	12,687
<i>R-squared</i>	0.149	0.158	0.148	0.158
<i>Chow F-value</i>		10.835 ***		3.739 *

The table shows the fixed-effects regression results for two subsamples based on the median firm size. The variable names are italicized. The standard errors are reported below the estimated coefficients in parentheses. *Chow F-value*, extracted from Chow’s test, assesses the null hypothesis of equal regression coefficients for the key explanatory variable across two subgroups. \*\*\*, \*\*, and \* denotes statistical significance level of 1%, 5%, and 10%, respectively.

#### 6.4. Low vs. High Leverages

Table 5 divides our sample into two groups based on firm leverage, with firms annually categorized by leverage levels. The lower leverage subset includes companies with leverage below the median, and the higher leverage subset comprises those above the median. We conducted the fixed-effects regression following Equation (1) for both subsets. The coefficient for *FemaleDum* is insignificant for low-leverage firms but markedly negative at the 1% threshold for high-leverage firms. Correspondingly, the coefficient for *FemaleFrac* is more significant and more prominent in magnitude for high-leverage firms than those with lower leverage. Chow’s test for the disparity between these two coefficients is significant at 1%. Overall, the findings in Table 5 indicate that the negative impact of board gender diversity on corporate performance is more pronounced in high-leverage firms.

**Table 5.** Low- vs. high-leverage firms.

	(1) Low Leverage ROA	(2) High Leverage ROA	(3) Low Leverage ROA	(4) High Leverage ROA
<i>FemaleFrac</i>	−0.012 ** (0.006)	−0.021 *** (0.006)		
<i>FemaleDum</i>			−0.091 (0.093)	−0.315 *** (0.092)
<i>FirmSize</i>	2.050 *** (0.135)	1.304 *** (0.126)	2.020 *** (0.134)	1.300 *** (0.126)
<i>FirmLev</i>	−0.156 *** (0.011)	−0.198 *** (0.005)	−0.156 *** (0.011)	−0.199 *** (0.005)
<i>Tangibility</i>	−0.087 *** (0.008)	−0.063 *** (0.006)	−0.087 *** (0.008)	−0.063 *** (0.006)
<i>BoardSize</i>	0.030 (0.018)	0.021 (0.017)	0.033 * (0.018)	0.027 (0.017)

**Table 5.** Cont.

	(1) Low Leverage ROA	(2) High Leverage ROA	(3) Low Leverage ROA	(4) High Leverage ROA
<i>BoardInd</i>	0.007 ** (0.003)	−0.006 ** (0.003)	0.005 * (0.003)	−0.007 ** (0.003)
<i>DirAge</i>	−0.140 *** (0.014)	−0.036 ** (0.014)	−0.137 *** (0.014)	−0.034 ** (0.014)
<i>Duality</i>	0.043 (0.108)	0.052 (0.106)	0.051 (0.108)	0.052 (0.106)
<i>Constant</i>	−7.843 *** (1.581)	−2.282 (1.511)	−7.677 *** (1.579)	−2.366 (1.514)
Firm FE	Yes	Yes	Yes	Yes
Observations	12,676	12,687	12,676	12,687
<i>R-squared</i>	0.067	0.168	0.067	0.168
<i>Chow F-value</i>		6.790 ***		7.416 ***

The table shows the fixed-effects regression results for two subsamples based on the median firm leverage. The variable names are italicized. The standard errors are reported below the estimated coefficients in parentheses. *Chow F-value*, extracted from Chow’s test, assesses the null hypothesis of equal regression coefficients for the key explanatory variable across two subgroups. \*\*\*, \*\*, and \* denotes statistical significance level of 1%, 5%, and 10%, respectively.

### 6.5. Low vs. High Cash Holding

Table 6 divides our sample into two groups based on the median cash holding in a given year. We executed the fixed-effects regression per Equation (1) for each subgroup. The results disclose a consistently negative coefficient for *FemaleFrac*, significant at the 1% level for both low and high cash holdings. Nevertheless, Chow’s test for the disparity between the two coefficients is insignificant. In contrast, the coefficient for *FemaleDum* is significantly negative at the 1% level for low cash holding but proves insignificant for high cash holding. Nevertheless, Chow’s test fails to exhibit a substantial divergence between the two coefficients. Overall, cash holding does not appear to influence the negative effect of gender diversity on corporate performance.

**Table 6.** Low vs. high cash holding.

	(1) Low CashHold ROA	(2) High CashHold ROA	(3) Low CashHold ROA	(4) High CashHold ROA
<i>FemaleFrac</i>	−0.019 *** (0.005)	−0.018 ** (0.007)		
<i>FemaleDum</i>			−0.312 *** (0.073)	−0.144 (0.112)
<i>FirmSize</i>	1.449 *** (0.120)	1.682 *** (0.135)	1.455 *** (0.120)	1.644 *** (0.135)
<i>FirmLev</i>	−0.155 *** (0.004)	−0.208 *** (0.006)	−0.155 *** (0.004)	−0.208 *** (0.006)
<i>Tangibility</i>	−0.057 *** (0.006)	−0.067 *** (0.008)	−0.057 *** (0.006)	−0.067 *** (0.008)
<i>BoardSize</i>	0.042 *** (0.013)	0.051 ** (0.023)	0.047 *** (0.013)	0.054 ** (0.023)
<i>BoardInd</i>	0.003 (0.003)	0.002 (0.003)	0.004 (0.003)	0.000 (0.003)
<i>DirAge</i>	−0.009 (0.012)	−0.137 *** (0.015)	−0.008 (0.012)	−0.133 *** (0.015)
<i>Duality</i>	−0.023 (0.079)	0.116 (0.141)	−0.027 (0.079)	0.128 (0.141)
<i>Constant</i>	−9.494 *** (1.523)	−1.745 (1.511)	−9.644 *** (1.525)	−1.602 (1.512)
Firm FE	Yes	Yes	Yes	Yes
Observations	12676	12687	12676	12687
<i>R-squared</i>	0.165	0.138	0.165	0.138
<i>Chow F-value</i>		0.493		1.467

The table shows the fixed-effects regression results for two subsamples based on the median cash holding. The variable names are italicized. The standard errors are reported below the estimated coefficients in parentheses. *Chow F-value*, extracted from Chow’s test, assesses the null hypothesis of equal regression coefficients for the key explanatory variable across two subgroups. \*\*\*, \*\*, and \* denotes statistical significance level of 1%, 5%, and 10%, respectively.

### 6.6. Low vs. High Institutional Ownership

Table 7 bifurcates our dataset into two cohorts based on the median institutional ownership in a specific year. We conducted a fixed-effects regression analysis in alignment with Equation (1) for each subgroup. The findings indicate that the coefficients of *FemaleFrac* and *FemaleDum* are markedly negative at the 1% significance level within the low institutional ownership subset, yet they are not statistically significant for the high institutional ownership group. Chow’s test for the divergence between these two coefficients is also significant at the 1% threshold. It suggests that elevated institutional ownership might engender intensified scrutiny by institutions, thereby eclipsing the governance impact attributable to female directors.

**Table 7.** Low vs. high institutional ownership.

	(1) Low <i>InstiOwn</i> ROA	(2) High <i>InstiOwn</i> ROA	(3) Low <i>InstiOwn</i> ROA	(4) High <i>InstiOwn</i> ROA
<i>FemaleFrac</i>	−0.035 *** (0.007)	−0.000 (0.006)		
<i>FemaleDum</i>			−0.347 *** (0.113)	0.011 (0.086)
<i>FirmSize</i>	1.820 *** (0.154)	1.586 *** (0.133)	1.782 *** (0.154)	1.583 *** (0.132)
<i>FirmLev</i>	−0.188 *** (0.005)	−0.170 *** (0.005)	−0.189 *** (0.005)	−0.170 *** (0.005)
<i>Tangibility</i>	−0.045 *** (0.008)	−0.081 *** (0.007)	−0.045 *** (0.008)	−0.081 *** (0.007)
<i>BoardSize</i>	0.054 ** (0.022)	−0.011 (0.017)	0.064 *** (0.022)	−0.011 (0.017)
<i>BoardInd</i>	0.002 (0.003)	−0.008 *** (0.003)	−0.001 (0.003)	−0.008 *** (0.003)
<i>DirAge</i>	−0.107 *** (0.016)	−0.060 *** (0.015)	−0.102 *** (0.016)	−0.060 *** (0.015)
<i>Duality</i>	0.048 (0.144)	0.162 * (0.095)	0.057 (0.144)	0.162 * (0.095)
<i>Constant</i>	−5.522 *** (1.701)	−6.383 *** (1.648)	−5.522 *** (1.704)	−6.355 *** (1.647)
Firm FE	Yes	Yes	Yes	Yes
Observations	11,129	11,134	11,129	11,134
<i>R-squared</i>	0.132	0.138	0.131	0.138
<i>Chow F-value</i>		14.007 ***		7.879 ***

The table shows the fixed-effects regression results for two subsamples based on the median institutional ownership. The variable names are italicized. The standard errors are reported below the estimated coefficients in parentheses. *Chow F-value*, extracted from Chow’s test, assesses the null hypothesis of equal regression coefficients for the key explanatory variable across two subgroups. \*\*\*, \*\*, and \* denotes statistical significance level of 1%, 5%, and 10%, respectively.

### 6.7. The Impact of COVID-19

Table 8 delineates the influence of the COVID-19 pandemic by dividing the sample into pre-COVID-19 (2006–2019) and during-COVID-19 (2020–2023) subsets. We executed the fixed-effects regression for both subsets according to Equation (1). The findings reveal that before COVID-19, *FemaleFrac* had a significantly negative effect on firm performance at the 1% level. During COVID-19, this negative relationship lost its significance. Chow’s test for the divergence between these two coefficients is also significant at the 5% threshold. *FemaleDum* is negatively significant at the 10% level pre-COVID-19 and becomes insignificant during the COVID-19 period. This modest difference is further evidenced by the insignificant outcome in Chow’s test. In summary, the detrimental impact of board gender diversity on corporate performance was weakly more pronounced pre-COVID-19 than during the pandemic.

**Table 8.** Before vs. during COVID-19.

	(1) Before COVID-19 ROA	(2) During COVID-19 ROA	(3) Before COVID-19 ROA	(4) During COVID-19 ROA
<i>FemaleFrac</i>	−0.023 *** (0.006)	−0.005 (0.008)		
<i>FemaleDum</i>			−0.157 * (0.083)	−0.152 (0.147)
<i>FirmSize</i>	1.580 *** (0.109)	5.511 *** (0.334)	1.555 *** (0.109)	5.531 *** (0.333)
<i>FirmLev</i>	−0.175 *** (0.004)	−0.270 *** (0.010)	−0.175 *** (0.004)	−0.271 *** (0.010)
<i>Tangibility</i>	−0.079 *** (0.006)	−0.156 *** (0.016)	−0.079 *** (0.006)	−0.156 *** (0.015)
<i>BoardSize</i>	0.047 *** (0.015)	0.103 *** (0.036)	0.050 *** (0.015)	0.108 *** (0.036)
<i>BoardInd</i>	0.010 *** (0.003)	−0.007 (0.007)	0.009 *** (0.003)	−0.006 (0.007)
<i>DirAge</i>	−0.060 *** (0.012)	−0.046 * (0.025)	−0.056 *** (0.012)	−0.048 * (0.025)
<i>Duality</i>	0.063 (0.089)	−0.117 (0.185)	0.066 (0.089)	−0.117 (0.185)
<i>Constant</i>	−6.233 *** (1.315)	−49.453 *** (4.047)	−6.168 *** (1.316)	−49.624 *** (4.043)
Firm FE	Yes	Yes	Yes	Yes
Observations	18,153	7210	18,153	7210
<i>R-squared</i>	0.153	0.205	0.152	0.205
<i>Chow F-value</i>		3.879 **		1.993

The table shows the fixed-effects regression results before and during the COVID-19 pandemic: 2006–2019 and 2020–2023. The variable names are italicized. The standard errors are reported below the estimated coefficients in parentheses. *Chow F-value*, extracted from Chow’s test, assesses the null hypothesis of equal regression coefficients for the key explanatory variable across two subgroups. \*\*\*, \*\*, and \* denotes statistical significance level of 1%, 5%, and 10%, respectively.

### 6.8. Different Industries

Table 9 exhibits the fixed-effects regression outcomes according to Equation (1) across eleven disparate industries. The analysis reveals a substantial adverse effect of board gender diversity on firm performance in the energy, materials, consumer discretionary, consumer staples, and utilities sectors. Conversely, this impact is insignificant in the industrials, health care, financials, information technology, communication services, and real estate sectors. The former cluster of industries constitutes regulated and consumer-centric sectors. These fields operate in regulated environments and are closely tied to consumer behaviors and preferences. Diverse perspectives and governance practices can significantly influence their performance, making them sensitive to board composition. In contrast, the latter group of industries is characterized by their innovation-driven nature. Their performance might be more influenced by technological innovation, market adaptability, and industry-specific challenges rather than solely by board composition.

Table 9. Different industries.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Energy ROA	Materials ROA	Industrials ROA	Consumer Discretionary ROA	Consumer Staples ROA	Health Care ROA	Financials ROA	Information Technology ROA	Communication Services ROA	Utilities ROA	Real Estate ROA
<i>FemaleFrac</i>	-0.075 * (0.042)	-0.032 *** (0.012)	-0.006 (0.008)	-0.020 ** (0.009)	-0.022 ** (0.009)	-0.019 (0.020)	-0.006 (0.009)	-0.015 (0.012)	0.041 (0.031)	-0.064 ** (0.026)	-0.020 (0.031)
<i>FirmSize</i>	3.721 *** (1.240)	2.421 *** (0.295)	1.545 *** (0.171)	1.606 *** (0.218)	0.802 *** (0.230)	2.562 *** (0.404)	-0.149 (0.170)	2.497 *** (0.260)	1.004 * (0.523)	-0.301 (0.657)	0.457 (0.468)
<i>FirmLev</i>	-0.096 *** (0.036)	-0.180 *** (0.009)	-0.161 *** (0.006)	-0.217 *** (0.007)	-0.162 *** (0.009)	-0.168 *** (0.014)	-0.080 *** (0.009)	-0.190 *** (0.010)	-0.256 *** (0.024)	-0.106 *** (0.021)	-0.107 *** (0.021)
<i>Tangibility</i>	-0.064 * (0.034)	-0.118 *** (0.013)	-0.084 *** (0.008)	-0.025 ** (0.011)	-0.056 *** (0.011)	0.012 (0.026)	-0.030 ** (0.015)	-0.132 *** (0.015)	0.052 (0.045)	0.002 (0.016)	-0.083 *** (0.022)
<i>BoardSize</i>	-0.147 (0.109)	0.021 (0.031)	0.048 ** (0.020)	-0.058 ** (0.029)	0.027 (0.032)	-0.027 (0.061)	0.011 (0.029)	0.123 *** (0.038)	0.426 *** (0.126)	0.016 (0.073)	0.200 ** (0.098)
<i>BoardInd</i>	0.041 ** (0.019)	-0.005 (0.006)	0.006 * (0.004)	-0.024 *** (0.005)	0.004 (0.005)	-0.010 (0.010)	0.028 *** (0.005)	0.029 *** (0.006)	-0.054 *** (0.018)	0.030 ** (0.015)	0.004 (0.017)
<i>DirAge</i>	-0.165 (0.107)	-0.019 (0.028)	-0.035 ** (0.017)	-0.107 *** (0.023)	-0.013 (0.023)	-0.157 *** (0.051)	-0.079 *** (0.028)	-0.052 * (0.028)	-0.306 *** (0.071)	0.013 (0.062)	-0.067 (0.073)
<i>Duality</i>	-0.250 (0.646)	-0.006 (0.173)	-0.044 (0.121)	-0.295 (0.192)	0.269 (0.190)	-0.082 (0.386)	-0.473 *** (0.165)	0.227 (0.230)	2.142 ** (0.888)	0.401 (0.575)	0.989 (0.621)
<i>Constant</i>	-28.022 * (16.100)	-16.348 *** (3.579)	-7.173 *** (1.941)	-1.696 (2.491)	-0.125 (2.797)	-11.606 *** (4.458)	9.011 *** (2.493)	-16.421 *** (2.984)	10.241 * (6.132)	9.267 (10.274)	7.055 (5.416)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	239	2586	7175	4734	2287	1181	1686	3444	1157	323	551
<i>R-squared</i>	0.150	0.244	0.164	0.186	0.175	0.134	0.070	0.211	0.161	0.134	0.119

The table shows the fixed-effects regression results for different industries. The variable names are italicized. The standard errors are reported below the estimated coefficients in parentheses. \*\*\*, \*\*, and \* denotes statistical significance level of 1%, 5%, and 10%, respectively.

### 6.9. Mediating Effects

Table 10 investigates the mediating effects employing a two-step regression methodology. Concerning *BoardInd* in Columns 1 and 2, the manifested indirect effect stands at 0.004 ( $=1.268 \times 0.003$ ), juxtaposed with a direct effect of  $-0.013$ . The absence of a mediating effect by board independence is inferred from their contrasting signs. Regarding CEO duality in Columns 3 and 4, the indirect effect registers at  $-0.000054$  ( $= -0.003 \times 0.018$ ), while the direct effect maintains at  $-0.013$ . It suggests a negligible mediating impact of CEO duality. Columns 5 and 7 indicate that board gender diversity exerts a notably positive influence on firm innovation (*RD*) and corporate social responsibility (*ESG*). About *RD* in Columns 5 and 6, the indirect effect is  $-0.0022$  ( $=0.003 \times -0.742$ ), against a direct effect of  $-0.012$ , implying a mediating effect of *RD* at 15% ( $=0.0022/(0.0022 + 0.012)$ ). For *ESG* in Columns 7 and 8, the indirect effect is calculated at  $-0.011$  ( $=0.051 \times -0.215$ ), while the direct effect is  $-0.005$ , indicating a mediating effect of *ESG* at 69% ( $=0.011/(0.011 + 0.005)$ ). The findings indicate a modest mediating role via *RD* and a more pronounced one through *ESG*, yet no significant mediation is observed for *BoardInd* and *Duality*.

**Table 10.** Mediating effects.

	(1) <i>BoardInd</i>	(2) <i>ROA</i>	(3) <i>Duality</i>	(4) <i>ROA</i>	(5) <i>RD</i>	(6) <i>ROA</i>	(7) <i>ESG</i>	(8) <i>ROA</i>
<i>FemaleFrac</i>	1.268 *** (0.013)	-0.013 *** (0.004)	-0.003 *** (0.000)	-0.013 *** (0.004)	0.003 *** (0.001)	-0.012 *** (0.004)	0.051 *** (0.001)	-0.005 (0.010)
<i>BoardInd</i>		0.003 (0.002)		0.003 (0.002)		0.002 (0.002)		-0.010 (0.007)
<i>Duality</i>		0.018 (0.075)		0.018 (0.075)		0.031 (0.080)		-0.017 (0.173)
<i>RD</i>						-0.742 *** (0.033)		
<i>ESG</i>								-0.215 ** (0.105)
<i>FirmSize</i>		1.494 *** (0.088)		1.494 *** (0.088)		1.486 *** (0.092)		2.822 *** (0.312)
<i>FirmLev</i>		-0.181 *** (0.003)		-0.181 *** (0.003)		-0.184 *** (0.003)		-0.236 *** (0.010)
<i>Tangibility</i>		-0.063 *** (0.005)		-0.063 *** (0.005)		-0.061 *** (0.005)		-0.120 *** (0.016)
<i>BoardSize</i>		0.034 *** (0.012)		0.034 *** (0.012)		0.033 *** (0.013)		-0.040 (0.033)
<i>DirAge</i>		-0.078 *** (0.010)		-0.078 *** (0.010)		-0.078 *** (0.010)		-0.024 (0.027)
<i>Constant</i>	16.780 *** (0.098)	-4.191 *** (1.044)	0.794 *** (0.002)	-4.191 *** (1.044)	1.694 *** (0.006)	-2.304 ** (1.082)	1.800 *** (0.013)	-20.635 *** (3.980)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	25,363	25,363	25,363	25,363	23,450	23,450	5696	5696
<i>R-squared</i>	0.300	0.146	0.004	0.146	0.001	0.170	0.261	0.159

The table shows the fixed-effects regression results for the mediating effects of *BoardInd*, *Duality*, *RD*, and *ESG*. The variable names are italicized. The standard errors are reported below the estimated coefficients in parentheses. \*\*\*, \*\*, and \* denotes statistical significance level of 1%, 5%, and 10%, respectively.

### 6.10. 2SLS Regressions

Table 11 explores the endogeneity issue concerning board gender diversity within the fixed-effects regression framework. There exists the potential for higher firm performance to affect board gender diversity inversely. Conversely, a third variable might simultaneously elevate board gender diversity and diminish firm performance, creating a perceived negative correlation where none inherently exists. To reassess this dynamic, we introduce an instrumental variable. Initially, we conducted a test for the endogeneity of regression variables. The Hausman *F* test refutes the hypothesis of exogeneity at the 10% level, underscoring the need to address endogeneity and suggesting the 2SLS model's superiority over

ordinary least squares regression. Subsequently, we assess the strength of the instrumental variable, the equal opportunity policy. According to Equation (3), Columns 1 and 3 of Table 11 exhibit the instrumental variable’s significant effect on *FemaleFrac* and *FemaleDum*, achieving statistical significance at the 1% level. The second-stage results, following Equation (4) and displayed in Columns 2 and 4 of Table 11, indicate that predicted board gender diversity adversely affects corporate performance, reaching a 1% level of statistical significance, irrespective of whether it is measured by the percentage of female directors or as a dummy variable.

**Table 11.** Two-stage least squares (2SLS) regression results.

	(1) 1st Stage <i>FemaleFrac</i>	(2) 2nd Stage ROA	(3) 1st Stage <i>FemaleDum</i>	(4) 2nd Stage ROA
<i>EqOpp</i>	1.427 *** (0.110)		0.100 *** (0.006)	
<i>Predicted FemaleFrac</i>		−0.246 *** (0.051)		
<i>Predicted FemaleDum</i>				−3.187 *** (0.649)
<i>FirmSize</i>	4.129 *** (0.163)	2.718 *** (0.255)	0.024 *** (0.002)	2.458 *** (0.203)
<i>FirmLev</i>	−0.002 (0.006)	−0.188 *** (0.004)	−0.001 *** (0.000)	−0.191 *** (0.004)
<i>Tangibility</i>	−0.015 * (0.008)	−0.066 *** (0.006)	0.001 *** (0.000)	−0.067 *** (0.006)
<i>BoardSize</i>	−0.081 *** (0.021)	0.024 * (0.015)	0.023 *** (0.001)	0.096 *** (0.018)
<i>BoardInd</i>	0.222 *** (0.004)	0.059 *** (0.012)	0.014 *** (0.000)	0.049 *** (0.010)
<i>DirAge</i>	−0.390 *** (0.017)	−0.176 *** (0.023)	−0.015 *** (0.001)	−0.147 *** (0.017)
<i>Duality</i>	−0.809 *** (0.127)	−0.192 ** (0.094)	−0.014 * (0.007)	−0.156 * (0.089)
<i>Constant</i>	−24.024 *** (1.980)	−12.424 *** (1.964)	0.373 *** (0.041)	−11.562 *** (1.795)
Firm FE	Yes	Yes	Yes	Yes
Observations	22,274	22,274	22,274	22,274
<i>R-squared</i>	0.348	0.146	0.269	0.146

The table shows the 2SLS regression results using the equal opportunity policy as the instrumental variable. The variable names are italicized. The standard errors are reported below the estimated coefficients in parentheses. \*\*\*, \*\*, and \* denotes statistical significance level of 1%, 5%, and 10%, respectively.

The 2SLS findings mitigate endogeneity concerns. Various factors could underlie the negative impact of board gender diversity on firm performance. One plausible rationale is that female directors might enact stricter supervision, potentially hampering company performance, as posited by Adams and Ferreira (2009). Another hypothesis suggests that gender-mixed groups may encounter more conflicts during decision-making processes, consuming additional time and energy, thereby diminishing the competitive edge of firms with gender-diverse boards (Lim et al. 2019). Additionally, the influence of gender stereotypes, particularly in patriarchal societies like Japan, cannot be overlooked. Culturally, women have historically been consigned to subordinate roles, facing barriers to accessing educational resources. Moreover, prevalent stereotypes often paint women as uninformed, aggressive, and overly emotional. Consequently, the presence of female directors on a board might lead to negative investor perceptions and a loss of confidence in the firm, ultimately adversely affecting corporate performance.

### 6.11. Alternative Performance Measures

Columns 1 and 2 of Table 12 replace ROA with ROE in our analysis to examine robustness, in line with Equation (1). These findings are in harmony with the previous application of ROA for assessing corporate performance, as illustrated in Table 3. The results demonstrate a negative association between board gender diversity and corporate performance, statistically significant at the 1% level. The regression coefficients in Table 12 exhibit magnitudes surpassing those in Table 3, indicating enhanced economic significance. Hence, we affirm that the negative relationship between board gender diversity and corporate performance is robust with an alternative performance measure.

**Table 12.** Alternative performance measures.

	(1) <i>ROE</i>	(2) <i>ROE</i>	(3) <i>MktCapChg</i>	(4) <i>MktCapChg</i>
<i>FemaleFrac</i>	−0.038 *** (0.010)		−0.000 (0.001)	
<i>FemaleDum</i>		−0.585 *** (0.162)		−0.001 (0.009)
<i>FirmSize</i>	2.769 *** (0.219)	2.750 *** (0.218)	−0.336 *** (0.012)	−0.336 *** (0.012)
<i>FirmLev</i>	−0.323 *** (0.008)	−0.323 *** (0.008)	0.006 *** (0.000)	0.006 *** (0.000)
<i>Tangibility</i>	−0.133 *** (0.012)	−0.133 *** (0.012)	0.001 * (0.001)	0.001 * (0.001)
<i>BoardSize</i>	0.086 *** (0.031)	0.099 *** (0.031)	−0.006 *** (0.002)	−0.006 *** (0.002)
<i>BoardInd</i>	−0.010 * (0.005)	−0.010 * (0.005)	−0.000 (0.000)	−0.000 (0.000)
<i>DirAge</i>	−0.158 *** (0.025)	−0.156 *** (0.025)	0.002 * (0.001)	0.002 * (0.001)
<i>Duality</i>	−0.057 (0.188)	−0.053 (0.188)	0.018 * (0.010)	0.018 * (0.010)
<i>Constant</i>	−6.117 ** (2.604)	−6.154 ** (2.606)	3.757 *** (0.147)	3.758 *** (0.147)
Firm FE	Yes	Yes	Yes	Yes
Observations	25,363	25,363	21,901	21,901
<i>R-squared</i>	0.081	0.081	0.061	0.061

The table shows the fixed-effects regression results with alternative performance measures: ROE and *MktCapChg*. The variable names are italicized. The standard errors are reported below the estimated coefficients in parentheses. \*\*\*, \*\*, and \* denotes statistical significance level of 1%, 5%, and 10%, respectively.

Table 12 also incorporates the percentage change in market capitalization as the dependent variable. Nonetheless, the coefficients associated with *FemaleFrac* and *FemaleDum* are insignificant. It likely reflects the distinction between accounting-based performance measures (ROA and ROE) and market-based performance measures (market capitalization variation). While ROA focuses on internal operational performance per accounting records, market capitalization change is swayed by external market forces and expectations. Gender diversity might have a more direct or observable impact on internal management practices and policies (affecting ROA), but its influence on external market valuation (market capitalization change) could be less direct or be overshadowed by other factors. Alternatively, accounting measures like ROA reflect current or short-term operational performance, while market valuations often incorporate long-term expectations and growth potential.

### 6.12. Nonlinear Quadratic Regression

Zhang et al. (2023) employed nonlinear quadratic regression to demonstrate a convex correlation between a CEO’s educational background and corporate risk-taking. Alfar et al. (2023) uncover a nonlinear effect of gender diversity on firm performance in the Palestine Exchange. Consequently, the association between board gender diversity and



corporate performance in Japan may similarly be nonlinear. Table 13 presents the results of nonlinear quadratic regression analyses following Equation (5). These findings indicate that regardless of whether ROA or ROE is utilized to assess corporate performance, the purported quadratic relationship between board gender diversity and corporate performance is not statistically significant. Consequently, we deduce that within our sample, there is no evidence of a nonlinear quadratic relationship between board gender diversity and corporate performance. Thus, the critical mass theory does not appear to be substantiated by our study.

**Table 13.** Nonlinear quadratic regression results.

	(1) ROA	(2) ROE
<i>FemaleFrac</i>	−0.014 * (0.007)	−0.058 *** (0.017)
<i>FemaleFrac</i> <sup>2</sup>	0.000 (0.000)	0.001 (0.001)
<i>FirmSize</i>	1.494 *** (0.088)	2.784 *** (0.219)
<i>FirmLev</i>	−0.181 *** (0.003)	−0.323 *** (0.008)
<i>Tangibility</i>	−0.063 *** (0.005)	−0.133 *** (0.012)
<i>BoardSize</i>	0.034 *** (0.012)	0.088 *** (0.031)
<i>BoardInd</i>	0.003 (0.002)	−0.009 (0.005)
<i>DirAge</i>	−0.078 *** (0.010)	−0.160 *** (0.025)
<i>Duality</i>	0.018 (0.075)	−0.060 (0.188)
<i>Constant</i>	−4.193 *** (1.045)	−6.205 ** (2.604)
Firm FE	Yes	Yes
Observations	25,363	25,363
<i>R-squared</i>	0.146	0.081

The table shows the nonlinear quadratic regression results by adding the squared term of *FemaleFrac*. The variable names are italicized. The standard errors are reported below the estimated coefficients in parentheses. \*\*\*, \*\*, and \* denotes statistical significance level of 1%, 5%, and 10%, respectively.

According to the critical mass theory, the commonly accepted threshold is a minimum of three or 30% female directors. Nevertheless, as illustrated in Figure 1, the average proportion of female directors in Japanese firms significantly lags behind this 30% benchmark. In data not presented, the frequency count of female directors reveals a mere 396 firm-year observations out of 25,363 (1.6%) that meet or exceed the threshold of three female directors in Japan. We conducted a robustness analysis for further validation by regressing ROA against the number of female directors and its squared value. This analysis did not reveal a nonlinear quadratic association. Therefore, the critical mass theory may not apply to Japanese companies due to their low female directorship ratio.

### 6.13. Comparison with Other Countries

Examining the interplay between board gender diversity and corporate performance in Japan versus other countries is imperative. A substantial portion of research reveals a positive impact in Japan (Nakagawa and Schreiber 2014; Tanaka 2019), Mauritius (Mahadeo et al. 2012), China (Liu et al. 2014), France (Sabatier 2015), the UK (Brahma et al. 2021), Russia (Garanina and Muravyev 2021), India (Sanan 2016; Sarkar and Selarka 2021), East Asian territories including Hong Kong, South Korea, Malaysia, and Singapore (Low et al. 2015), across 24 countries (Belaounia et al. 2020), 47 countries (Terjesen et al. 2016), and

34 countries (Pucheta-Martínez and Gallego-Álvarez 2020). Conversely, a minority of studies indicate a detrimental impact in the UK (Shehata et al. 2017), Pakistan (Mirza et al. 2012; Akram et al. 2020), and Malaysia (Ahmad et al. 2020; Abdullah 2014; Lim et al. 2019). Additionally, limited investigations report a neutral influence in India (Kagzi and Guha 2018), the Netherlands and Denmark (Marinova et al. 2016), Pakistan (Yasser 2012), the United States (Carter et al. 2010), Turkey (Ararat and Yurtoglu 2021), and the Philippines (Unite et al. 2019). Notably, disparities exist even within the same nation, as evidenced in the UK, the United States, and Malaysia. Our findings also diverge from established outcomes for Japan based on earlier data (Nakagawa and Schreiber 2014; Tanaka 2019). It is significant to acknowledge Japan's distinctive context, characterized by a historically low ratio of female directors and a remarkable increase in this ratio over the past decade within a predominantly male-centric culture. Hence, a focused study on Japan can yield insights beneficial for other nations with low female director representation and male-dominated environments.

This research also offers pertinent implications for nations exhibiting similar limited female labor force participation patterns and lower gender gap indices. Firstly, the outcomes afford valuable perspectives for such countries. Secondly, despite the distinctive nature of Japanese corporate governance compared to Western standards, its robustness is acknowledged. In this context of stringent corporate governance, enhanced gender diversity may inadvertently foster excessive oversight, potentially detracting from organizational performance. Our observations regarding the adverse effects of board gender diversity on Japanese corporate performance align with prior analyses in jurisdictions characterized by vigorous corporate governance regimes (Ahern and Dittmar 2012; Adams and Ferreira 2009). In contrast, inquiries in locales with lax corporate governance structures have documented beneficial impacts (Liu et al. 2014; Herdhayinta et al. 2021). Consequently, adopting board gender diversity mandates a tailored approach by governments and corporations, reflecting their unique circumstances. It is crucial to recognize the absence of a universally applicable strategy.

Amid global institutional shifts, a reevaluation of corporate governance dynamics is underway. Future research should focus on a dual approach: a macro-level multi-country analysis and a micro-level study of Japanese corporate governance. Variations in the impact of gender diversity on firm performance across nations are influenced by unique national contexts (Terjesen and Singh 2008), with studies highlighting the varying effects of gender diversity quotas on market and accounting performance (Atinc et al. 2021). Further exploration is needed to understand the implementation of these global standards within different social, cultural, and political frameworks (Ansari et al. 2010). Japan's distinctive labor market characteristics and the potential influence of women's educational level and board independence on firm performance warrant deeper investigation (Gull et al. 2018). This nuanced approach will enhance understanding of gender diversity's complex role in corporate governance.

## **7. Conclusions**

Board gender diversity and corporate governance structure have increasingly garnered scholarly interest. While a substantial body of existing literature has investigated the connection between female directors and corporate performance, findings indicate that the influence of female directors on corporate performance varies across diverse national contexts and environments. This paper aimed to contribute novel insights into the relationship between board gender diversity and corporate performance within the Japanese context, a realm hitherto unexplored in prior research.

We employed a sample of 1990 publicly traded Japanese firms from 2006 to 2023 and revealed that female directors significantly and negatively influence corporate performance in Japan. This implies that companies with a higher proportion of female directors underperform relative to those with fewer or no female directors or that firms with at least one female board member fare worse than those with exclusively male boards. This relationship

is more pronounced in smaller firms with higher leverage or lower institutional ownership, within regulated and consumer-oriented industries, and in the pre-COVID-19 period. To address potential endogeneity between board gender diversity and firm performance, we employed the 2SLS methodology. Our findings confirm the robustness of this result, suggesting a causal direction from board gender diversity to firm performance rather than vice versa. We also used ROE as an alternative performance metric. Our fundamental conclusion remains robust. Our study did not identify a U-shaped relationship between board gender diversity and firm performance.

The results of this study are relevant to corporate leaders, investors, and policymakers in Japan. For Japanese policymakers, enacting the 2023 policies that require a 30% female board membership by 2030 poses a significant challenge. There may be a necessity for these policymakers to reassess or modify current regulations to alleviate potential adverse effects on organizational performance. Consequently, it is recommended that policymakers promote cooperative endeavors involving government, private sector entities, and non-profit organizations to formulate an all-encompassing strategy that capitalizes on varied viewpoints and resources. Corporate leaders are faced with the challenge of effectively addressing the international standard of gender quotas. Merely meeting these quota requirements does not automatically lead to the benefits associated with gender diversity. In fact, it could potentially harm corporate performance (Adams and Ferreira 2009). The push to comply with these policy mandates has increased the demand for experienced female directors, surpassing the available pool (Carter et al. 2010). Consequently, this has led to the appointment of less experienced second and third female directors, who may not fully capitalize on the positive impacts on corporate performance (Claessens et al. 2000). Thus, corporate leaders must focus on aligning women's resources, expertise, and viewpoints within the corporate governance framework, accentuating the substantial inclusion of women's contributions beyond their mere presence on the board. For investors, our results indicate the necessity of meticulously considering the changing dynamics in gender diversity regulations and policies. The 30% female board member target by 2030 may affect investment choices, as companies adhering to these requirements could be perceived as more socially responsible and aligned with global expectations. Nonetheless, financial performance may not exhibit uniform progress; thus, investors should engage in more informed investment strategies that align with their ethical standards and risk appetite.

Three potential reasons might explain the observed negative correlation between board gender diversity and corporate performance: female directors could contribute to excessive monitoring, boards with gender diversity might experience more conflicts during decision-making and prevailing social stereotypes about women. Consequently, firms should not anticipate an enhancement in performance merely by appointing female directors. Nevertheless, our research has certain limitations. Firstly, the instrumental variable employed, the equal opportunity policy, may not be the best choice. Future research should consider more potent instrumental variables. Secondly, due to constraints in data availability, our analysis included only a limited set of control variables. Future investigations could benefit from using panel data encompassing a more extensive array of control variables. Finally, this study focused on the nexus between board gender diversity and performance within the Japanese cultural environment. The effects of gender diversity on corporate performance may vary across policy and cultural environments. Hence, cross-country comparisons are warranted in subsequent research endeavors.

**Author Contributions:** Conceptualization, K.W.; methodology, K.W. and J.Z.; software, K.W. and J.Z.; validation, K.W., J.M., C.X. and J.Z.; formal analysis, K.W., J.M., C.X. and J.Z.; investigation, K.W., J.M., C.X. and J.Z.; resources, K.W. and J.Z.; data curation, K.W. and J.Z.; writing—original draft preparation, K.W.; writing—review and editing, K.W., J.M., C.X. and J.Z.; visualization, K.W., J.M., C.X. and J.Z.; supervision, J.M., C.X. and J.Z.; project administration, J.Z.; funding acquisition, J.M., C.X. and J.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Department of Education of Zhejiang Province General Program [Y202249981, Y202353438], the Wenzhou-Kean University Internal Research Support Program [IRSPG202205, IRSPG202206], the Wenzhou-Kean University Student Partnering with Faculty Research Program [WKUSPF2023002, WKUSPF2023004] and the Wenzhou-Kean University International Collaborative Research Program [ICRP2023002, ICRP2023004].

**Data Availability Statement:** The authors used Bloomberg data, which are publicly available through the Bloomberg Terminals.

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

## References

- Abdullah, Shamsul Nahar. 2014. The causes of gender diversity in Malaysian large firms. *Journal of Management & Governance* 18: 1137–59.
- Abrams, Dominic, and Michael A. Hogg. 2010. Social Identity and Self-Categorization. In *The SAGE Handbook of Prejudice, Stereotyping and Discrimination*. Edited by John F. Dovidio, Miles Hewstone, Peter Glick and Victoria M. Esses. Thousand Oaks: SAGE Publications Ltd., pp. 179–93.
- Adams, Renée B., and Daniel Ferreira. 2007. A theory of friendly boards. *Journal of Finance* 62: 217–50. [CrossRef]
- Adams, Renée B., and Daniel Ferreira. 2009. Women in the boardroom and their impact on governance and performance. *Journal of Financial Economics* 94: 291–309. [CrossRef]
- Ahern, Kenneth R., and Amy K. Dittmar. 2012. The changing of the boards: The impact on firm valuation of mandated female board representation. *Quarterly Journal of Economics* 127: 137–97. [CrossRef]
- Ahmad, Maslina, Raja Nur Syazwani Raja Kamaruzaman, Hamdino Hamdan, and Hairul Azlan Annuar. 2020. Women directors and firm performance: Malaysian evidence post policy announcement. *Journal of Economic and Administrative Sciences* 36: 97–110. [CrossRef]
- Akram, Farheen, Muhammad Abrar ul Haq, Vinodh K. Natarajan, and R. Stephen Chellakan. 2020. Board heterogeneity and corporate performance: An insight beyond agency issues. *Cogent Business & Management* 7: 1809299.
- Alesina, Alberto, and Eliana La Ferrara. 2005. Ethnic diversity and economic performance. *Journal of Economic Literature* 43: 762–800. [CrossRef]
- Alfar, Abdelrahman J. K., Nariman Abuatwan, Mohamed Elheddad, and Mohammad Qaki. 2023. The internal determinants of gender diversity and its non-linear impact on firms' performance: Evidence from the listed companies in Palestine Exchange. *Journal of Risk and Financial Management* 16: 28. [CrossRef]
- Almazan, Andres, and Javier Suarez. 2003. Entrenchment and severance pay in optimal governance structures. *Journal of Finance* 58: 519–47. [CrossRef]
- Aluchna, Maria, Jyoti Devi Mahadeo, and Bogumił Kamiński. 2020. The association between independent directors and company value. Confronting evidence from two emerging markets. *Corporate Governance: The International Journal of Business in Society* 20: 987–99. [CrossRef]
- Aman, Hiroyuki, Wendy A. Beekes, and Philip Brown. 2021. Corporate governance and transparency in Japan. *International Journal of Accounting* 56: 2150003. [CrossRef]
- Ansari, Shahzad M., Peer C. Fiss, and Edward J. Zajac. 2010. Made to fit: How practices vary as they diffuse. *Academy of Management Review* 35: 67–92.
- Ararat, Melsa, and B. Burcin Yurtoglu. 2021. Female directors, board committees, and firm performance: Time-series evidence from Turkey. *Emerging Markets Review* 48: 100768. [CrossRef]
- Atinc, Guclu, Saurabh Srivastava, and Sonia Taneja. 2021. The impact of gender quotas on corporate boards: A cross-country comparative study. *Journal of Management and Governance* 26: 685–706. [CrossRef]
- Barber, Brad M., and John D. Lyon. 1996. Detecting abnormal operating performance: The empirical power and specification of test statistics. *Journal of Financial Economics* 41: 359–99. [CrossRef]
- Belaounia, Samia, Ran Tao, and Hong Zhao. 2020. Gender equality's impact on female directors' efficacy: A multi-country study. *International Business Review* 29: 101737. [CrossRef]
- Berger, Allen N., Thomas Kick, and Klaus Schaeck. 2014. Executive board composition and bank risk taking. *Journal of Corporate Finance* 28: 48–65. [CrossRef]
- Binder, Bettina C. K., Terry Morehead Dworkin, Niculina Nae, Cindy A. Schipani, and Irina Averianova. 2019. The plight of women in positions of corporate leadership in the United States, the European Union, and Japan: Differing laws and cultures, similar issues. *Michigan Journal of Gender & Law* 26: 279.
- Brahma, Sanjukta, Chioma Nwafor, and Agyenim Boateng. 2021. Board gender diversity and firm performance: The UK evidence. *International Journal of Finance & Economics* 26: 5704–19.
- Bøhren, Øyvind, and Siv Staubo. 2014. Does mandatory gender balance work? Changing organizational form to avoid board upheaval. *Journal of Corporate Finance* 28: 152–68. [CrossRef]
- Carter, David A., Betty J. Simkins, and W. Gary Simpson. 2003. Corporate governance, board diversity, and firm value. *Financial Review* 38: 33–53. [CrossRef]

- Carter, David A., Frank D'Souza, Betty J. Simkins, and W. Gary Simpson. 2010. The gender and ethnic diversity of US boards and board committees and firm financial performance. *Corporate Governance: An International Review* 18: 396–414. [CrossRef]
- Chan, Kam C., and Joanne Li. 2008. Audit committee and firm value: Evidence on outside top executives as expert-independent directors. *Corporate Governance: An International Review* 16: 16–31. [CrossRef]
- Chen, Jie, Woon Sau Leung, and Kevin P. Evans. 2018. Female board representation, corporate innovation and firm performance. *Journal of Empirical Finance* 48: 236–54. [CrossRef]
- Chow, Gregory C. 1960. Tests of equality between sets of coefficients in two linear regressions. *Econometrica* 28: 591–605. [CrossRef]
- Claessens, Stijn, Simeon Djankov, and Larry H. P. Lang. 2000. The separation of ownership and control in East Asian corporations. *Journal of Financial Economics* 58: 81–112. [CrossRef]
- Canyon, Martin J., and Simon I. Peck. 1998. Board size and corporate performance: Evidence from European countries. *European Journal of Finance* 4: 291–304. [CrossRef]
- Credit Suisse. 2021. The CS Gender 3000 in 2021: Broadening the Diversity Discussion. Available online: <https://www.credit-suisse.com/media/assets/corporate/docs/about-us/research/publications/csri-2021-gender-3000.pdf> (accessed on 24 December 2023).
- Cyert, Richard, and James March. 1963. A behavioral theory of the firm. In *Organizational Behavior* 2. Englewood Cliffs: Prentice Hall, pp. 60–77.
- Das, Nirmol Chandra, Mohammad Ashraf Ferdous Chowdhury, and Md Nazrul Islam. 2022. The heterogeneous impact of leverage on firm performance: Empirical evidence from Bangladesh. *South Asian Journal of Business Studies* 11: 235–52. [CrossRef]
- Delis, Manthos D., Chrysovalantis Gaganis, Iftekhar Hasan, and Fotios Pasiouras. 2017. The effect of board directors from countries with different genetic diversity levels on corporate performance. *Management Science* 63: 231–49. [CrossRef]
- Doğan, Mesut. 2013. Does firm size affect the firm profitability? Evidence from Turkey. *Research Journal of Finance and Accounting* 4: 53–59.
- EgonZehnder. 2022. Who's Really on Board? Spotlight on Japan. Available online: <https://www.egonzehnder.com/global-board-diversity-tracker/regional-spotlight/japan> (accessed on 24 December 2023).
- Garanina, Tatiana, and Alexander Muravyev. 2021. The gender composition of corporate boards and firm performance: Evidence from Russia. *Emerging Markets Review* 48: 100772. [CrossRef]
- García-Meca, Emma, Isabel-María García-Sánchez, and Jennifer Martínez-Ferrero. 2015. Board diversity and its effects on bank performance: An international analysis. *Journal of Banking & Finance* 53: 202–14.
- Gender Equality Bureau. 2022. The White Paper on Gender Equality 2022. Available online: [https://www.gender.go.jp/about\\_danjo/whitepaper/r04/gaiyou/pdf/r04\\_gaiyou\\_en.pdf](https://www.gender.go.jp/about_danjo/whitepaper/r04/gaiyou/pdf/r04_gaiyou_en.pdf) (accessed on 24 December 2023).
- Guest, Paul M. 2009. The impact of board size on firm performance: Evidence from the UK. *European Journal of Finance* 15: 385–404. [CrossRef]
- Gul, Ferdinand A., Bin Srinidhi, and Anthony C. Ng. 2011. Does board gender diversity improve the informativeness of stock prices? *Journal of Accounting and Economics* 51: 314–38. [CrossRef]
- Gull, Ammar Ali, Mehdi Nekhili, Haithem Nagati, and Tawhid Chtioui. 2018. Beyond gender diversity: How specific attributes of female directors affect earnings management. *British Accounting Review* 50: 255–74. [CrossRef]
- Herdhayinta, Heyvon, James Lau, and Carl Hsin-han Shen. 2021. Family female directors versus non-family female directors: Effects on firm value and dividend payouts in an extreme institutional environment. *British Journal of Management* 32: 969–87. [CrossRef]
- Hermalin, Benjamin, and Michael S. Weisbach. 2003. Boards of directors as an endogenously determined institution: A survey of the economic literature. *Economic Policy Review* 9: 7–26.
- Hoffman, L. Richard, and Norman R. F. Maier. 1961. Quality and acceptance of problem solutions by members of homogeneous and heterogeneous groups. *Journal of Abnormal and Social Psychology* 62: 401. [CrossRef]
- Jensen, Michael C., and William H. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs, and ownership structure. *Journal of Financial Economics* 3: 305–60. [CrossRef]
- Joecks, Jasmin, Kerstin Pull, and Karin Vetter. 2013. Gender diversity in the boardroom and firm performance: What exactly constitutes a “critical mass”? *Journal of Business Ethics* 118: 61–72. [CrossRef]
- Kagzi, Muneza, and Mahua Guha. 2018. Does board demographic diversity influence firm performance? Evidence from Indian-knowledge intensive firms. *Benchmarking: An International Journal* 25: 1028–58. [CrossRef]
- Kanter, Rosabeth Moss. 1977. Some effects of proportions on group life: Skewed sex ratios and responses to token women. *American Journal of Sociology* 82: 965–90. [CrossRef]
- Kim, Daehyun, and Laura T. Starks. 2016. Gender diversity on corporate boards: Do women contribute unique skills? *American Economic Review* 106: 267–71. [CrossRef]
- Konrad, Alison M., Vicki Kramer, and Sumru Erkut. 2008. Critical mass: The impact of three or more women on corporate boards. *Organizational Dynamics* 37: 145–64. [CrossRef]
- Lara, Juan Manuel García, Beatriz García Osma, Araceli Mora, and Mariano Scapin. 2017. The monitoring role of female directors over accounting quality. *Journal of Corporate Finance* 45: 651–68. [CrossRef]
- Lee, Seoki. 2010. Effects of Capital intensity on firm performance: The US Restaurant industry. *Journal of Hospitality Financial Management* 18: 2. [CrossRef]

- Lim, Kwee Pheng, Chun-Teck Lye, Yee Yen Yuen, and Wendy Ming Yen Teoh. 2019. Women directors and performance: Evidence from Malaysia. *Equality, Diversity and Inclusion: An International Journal* 38: 841–56. [CrossRef]
- Liu, Yu, Zuobao Wei, and Feixue Xie. 2014. Do women directors improve firm performance in China? *Journal of Corporate Finance* 28: 169–84. [CrossRef]
- Low, Daniel C. M., Helen Roberts, and Rosalind H. Whiting. 2015. Board gender diversity and firm performance: Empirical evidence from Hong Kong, South Korea, Malaysia and Singapore. *Pacific-Basin Finance Journal* 35: 381–401. [CrossRef]
- Mahadeo, Jyoti D., Teerooven Soobaroyen, and Vanisha Oogarah Hanuman. 2012. Board composition and financial performance: Uncovering the effects of diversity in an emerging economy. *Journal of Business Ethics* 105: 375–88. [CrossRef]
- Marinova, Joana, Janneke Plantenga, and Chantal Remery. 2016. Gender diversity and firm performance: Evidence from Dutch and Danish boardrooms. *International Journal of Human Resource Management* 27: 1777–90. [CrossRef]
- Mirza, Hammad Hassan, Sumaira Andleeb, and Farzana Ramzan. 2012. Gender diversity and firm performance: Evidence from Pakistan. *Journal of Social and Development Sciences* 3: 161–66. [CrossRef]
- Morgan Stanley Capital International. 2020. Women on Boards: 2020 Progress Report. Available online: <https://www.msci.com/www/women-on-boards-2020/women-on-boards-2020-progress/02212172407> (accessed on 24 December 2023).
- Nakagawa, Yukiko, and G. M. Schreiber. 2014. Women as drivers of Japanese firms success: The effect of women managers and gender diversity on firm performance. *Journal of Diversity Management* 9: 19–40. [CrossRef]
- OECD. 2023. Joining Forces for Gender Equality. What is Holding Us Back? Available online: <https://www.oecd.org/japan/Gender2023-JPN-En.pdf> (accessed on 24 December 2023).
- Owen, Ann L., and Judit Temesvary. 2018. The performance effects of gender diversity on bank boards. *Journal of Banking and Finance* 90: 50–63. [CrossRef]
- Pfeffer, Jeffrey, and Gerald Salancik. 1978. *The External Control of Organizations: A Resource Dependence Perspective*. Manhattan: Harper & Row.
- Pucheta-Martínez, María Consuelo, and Isabel Gallego-Álvarez. 2020. Do board characteristics drive firm performance? An international perspective. *Review of Managerial Science* 14: 1251–97. [CrossRef]
- Reguera-Alvarado, Nuria, and Francisco Bravo. 2017. The effect of independent directors' characteristics on firm performance: Tenure and multiple directorships. *Research in International Business and Finance* 41: 590–99. [CrossRef]
- Reynolds, Isabel. 2023. Japan Aims for Women to Make Up 30% of Directors at Top Firms. Available online: <https://www.bloomberg.com/news/articles/2023-06-06/japan-aims-for-women-to-make-up-30-of-directors-at-top-firms> (accessed on 24 December 2023).
- Sabatier, Mareva. 2015. A women's boom in the boardroom: Effects on performance? *Applied Economics* 47: 2717–27. [CrossRef]
- Sanan, Neeti Khetarpal. 2016. Board gender diversity and firm performance: Evidence from India. *Asian Journal of Business Ethics* 5: 1–18. [CrossRef]
- Sarkar, Jayati, and Ekta Selarka. 2021. Women on board and performance of family firms: Evidence from India. *Emerging Markets Review* 46: 100770. [CrossRef]
- Sawaji, Osamu. 2021. Revision of Japan's Corporate Governance Code and Guidelines for Investor and Company Engagement. Available online: [https://www.gov-online.go.jp/eng/publicity/book/hlj/html/202111/202111\\_09\\_en.html](https://www.gov-online.go.jp/eng/publicity/book/hlj/html/202111/202111_09_en.html) (accessed on 24 December 2023).
- Schwartz-Ziv, Miriam. 2017. Gender and board activeness: The role of a critical mass. *Journal of Financial and Quantitative Analysis* 52: 751–80. [CrossRef]
- Shehata, Nermeen, Ahmed Salhin, and Moataz El-Helaly. 2017. Board diversity and firm performance: Evidence from the UK SMEs. *Applied Economics* 49: 4817–32. [CrossRef]
- Smith, Nina, Valdemar Smith, and Mette Verner. 2006. Do women in top management affect firm performance? A panel study of 2500 Danish firms. *International Journal of Productivity & Performance Management* 55: 569–93.
- Süssmuth-Dyckerhoff, C., Jin Wang, and Josephine Chen. 2012. Women Matter: An Asian Perspective. *McKinsey & Company*. Available online: <https://www.empowerwomen.org/en/resources/documents/2015/12/women-matter-an-asian-perspective> (accessed on 24 December 2023).
- Tanaka, Takanori. 2019. Gender diversity on Japanese corporate boards. *Journal of the Japanese and International Economies* 51: 19–31. [CrossRef]
- Terjesen, Siri, and Ruth Sealy. 2016. Board gender quotas: Exploring ethical tensions from a multi-theoretical perspective. *Business Ethics Quarterly* 26: 23–65. [CrossRef]
- Terjesen, Siri, and Val Singh. 2008. Female Presence on Corporate Boards: A Multi-Country Study of Environmental Context. *Journal of Business Ethics* 83: 55–63. [CrossRef]
- Terjesen, Siri, Eduardo Barbosa Couto, and Paulo Morais Francisco. 2016. Does the presence of independent and female directors impact firm performance? A multi-country study of board diversity. *Journal of Management & Governance* 20: 447–83.
- Thomas, Debbie A. 2018. Bias in the boardroom: Implicit bias in the selection and treatment of women directors. *Marquette Law Review* 102: 539–74.
- Tobin, James. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit and Banking* 1: 15–29. [CrossRef]
- Torchia, Mariateresa, Andrea Calabrò, and Morten Huse. 2011. Women directors on corporate boards: From tokenism to critical mass. *Journal of Business Ethics* 102: 299–317. [CrossRef]

- Unite, Angelo A., Michael J. Sullivan, and Ailyn A. Shi. 2019. Board diversity and performance of Philippine firms: Do women matter? *International Advances in Economic Research* 25: 65–78. [CrossRef]
- Van Knippenberg, Daan, and Michaela C. Schippers. 2007. Work group diversity. *Annual Review of Psychology* 58: 515–41. [CrossRef] [PubMed]
- Wagner, W. Gary, Jeffrey Pfeffer, and Charles A. O'Reilly III. 1984. Organizational demography and turnover in top-management group. *Administrative Science Quarterly* 29: 74–92. [CrossRef]
- Williams Phillips, Katherine, and Charles A. O'Reilly. 1998. Williams Phillips, Katherine, and Charles A. O'Reilly. 1998. Demography and diversity in organizations: A review of 40 years of research. *Research in Organizational Behavior* 20: 77–140.
- World Bank. 2023. World Development Indicators Database. Available online: <https://data.worldbank.org/indicator/SL.TLF.CACT.FE.ZS?locations=JP> (accessed on 24 December 2023).
- World Economic Forum. 2023. Global Gender Gap Report 2023. Available online: [https://www3.weforum.org/docs/WEF\\_GGGR\\_2023.pdf](https://www3.weforum.org/docs/WEF_GGGR_2023.pdf) (accessed on 24 December 2023).
- Yasser, Qaiser Rafique. 2012. Affects of female directors on firms performance in Pakistan. *Modern Economy* 3: 817–25. [CrossRef]
- Zeng, Chuanyi. 2018. Independent directors, female directors and performance of financial listed companies in China. *Modern Economy* 9: 652–63. [CrossRef]
- Zenger, Todd R., and Barbara S. Lawrence. 1989. Organizational demography: The differential effects of age and tenure distributions on technical communication. *Academy of Management Journal* 32: 353–76. [CrossRef]
- Zhang, Jinyi, Chunxiao Xue, and Jianing Zhang. 2023. The impact of CEO educational background on corporate risk-taking in China. *Journal of Risk and Financial Management* 16: 9. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

# On Financial Distributions Modelling Methods: Application on Regression Models for Time Series

Paul R. Dewick 

Faculty of Science and Technology, University of Canberra, Canberra 2617, Australia;  
paul.dewick@canberra.edu.au

**Abstract:** The financial market is a complex system with chaotic behavior that can lead to wild swings within the financial system. This can drive the system into a variety of interesting phenomenon such as phase transitions, bubbles, and crashes, and so on. Of interest in financial modelling is identifying the distribution and the stylized facts of a particular time series, as the distribution and stylized facts can determine if volatility is present, resulting in financial risk and contagion. Regression modelling has been used within this study as a methodology to identify the goodness-of-fit between the original and generated time series model, which serves as a criterion for model selection. Different time series modelling methods that include the common Box–Jenkins ARIMA, ARMA-GARCH type methods, the Geometric Brownian Motion type models and Tsallis entropy based models when data size permits, can use this methodology in model selection. Determining the time series distribution and stylized facts has utility, as the distribution allows for further modelling opportunities such as bivariate regression and copula modelling, apart from the usual forecasting. Determining the distribution and stylized facts also allows for the identification of the parameters that are used within a Geometric Brownian Motion forecasting model. This study has used the Carbon Emissions Futures price between the dates of 1 May 2012 and 1 May 2022, to highlight this application of regression modelling.

**Keywords:** time series; regression; distribution; volatility; goodness-of-fit



**Citation:** Dewick, Paul R. 2022. On Financial Distributions Modelling Methods: Application on Regression Models for Time Series. *Journal of Risk and Financial Management* 15: 461. <https://doi.org/10.3390/jrfm15100461>

Academic Editor: Thanasis Stengos

Received: 5 September 2022

Accepted: 9 October 2022

Published: 13 October 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The financial market is a complex system that is the result of decisions of interacting agents and traders who speculate and can act impulsively. This collective of chaotic behavior can lead to wild swings within the financial system, Devi (2021). This can drive the system into a variety of interesting phenomena such as phase transitions, bubbles, and crashes and so on. Due to the 2008 financial crisis, there is a renewed interest in the choice of an adequate error distribution, Hambuckers and Heuchenne (2017). More recently, the 2019–2022 crisis due to the COVID-19 pandemic, there is a need in the ability to identify these effects and model this phenomenon.

As a result of these wild swings within the financial system, financial data should be examined using the model specified by their probability distribution, with skewness and excess kurtosis, Fukuda (2021). The appropriate modelling of the time series distribution, being symmetric or asymmetric and in addition, the tail thickness of the distribution as financial time series data is typically heavy tailed and contain time varying volatility, Liu and Heyde (2008). Correct distribution specification of the stylized facts is important as model misspecification can cause an overestimation of the kurtosis in the estimated residuals, Hambuckers and Heuchenne (2017). These stylized facts are often used to support investment decisions, Charpentire (2014).

Time series distributions are generally assumed to be approximately Normal, but the distribution is likely to be of a Student- $t$  or a skewed type distribution that describes a heavy tail or tails. These discrepancies to a Normal distribution must be identified to allow



for correct time series modelling, as deviations from a Normal distribution may indicate volatility, leverage and drift.

Time series modelling methods are typically based on the Box–Jenkins Auto Regressive (AR), the Auto Regressive Moving Average (ARMA) and the Auto Regressive Integrated Moving Average (ARIMA) type models which are used for mean modelling. The Auto Regressive Conditionally Heteroscedastic (ARCH) and Generalized Auto Regressive Conditionally Heteroscedasticity (GARCH) type models are used for variance modelling. When time series contains both mean and variance changes, these models can be combined, as these typical models can be ARMA-ARCH or ARMA-GARCH type models. These methods are mostly forecasting focused, as these models create a mathematical model to allow for forecasting modelling.

Other methods available when modelling financial data can be the Geometric Brownian Motion (GBM) type models and Entropy type models. The GBM (or Exponential Brownian Motion) type models are based on a random walk which follows the Brownian motion model. In undertaking a GBM, the identification of the initial distribution allows for the identification of the parameters  $(\mu, \sigma)$ , that are used within the modelling methodology. The entropy approach in modelling time series uses Tsallis entropy and can be used to determine the underlying distribution using the  $q$ -Gaussian distribution, Tsallis (2017).

The contribution of this study is to explore the utility of using simple linear regression modelling, Equation (1), as a goodness-of-fit criterion to identify a time series model that represents the original dataset by modelling their distributions. Box-Jenkins and Geometric Brownian Motion and Tsallis modelling methods were used as examples in model selection by applying simple linear regression modelling. Identification of the time series distribution also has the utility of allowing further modelling methods to be applied. These include bivariate regression modelling (between two time series datasets), Liu et al. (2020) and/or bivariate copula modelling, Dewick and Liu (2022) apart from the usual forecasting applications.

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{1}$$

In Section 2, I provide common distributions used in financial modelling. In Section 3, I provide an outline on time series modelling; In Section 4, I supply a time series modelling application; In Section 5, I supply the modelling results; In Section 6, I give my conclusions.

## 2. Financial Distributions

Modelling the correct financial distribution when undertaking time series modelling is a significant modelling component as financial time series distributions may contain heavy (fat) tails, volatility clustering nonlinear dependence, Ghani and Rahim (2019). Symmetric distributions available are the Normal, Student- $t$ , see Figure 1 and the  $q$ -Gaussian distributions, see Figure 2 that uses Tsallis entropy.

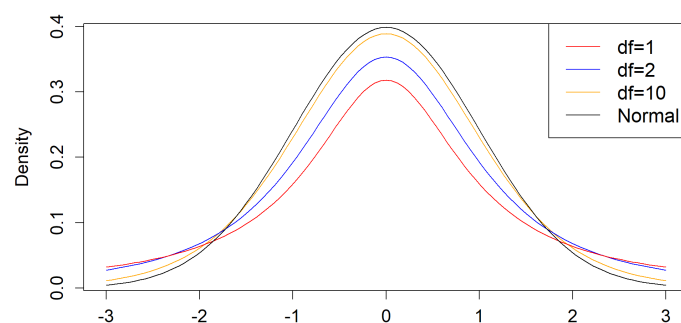
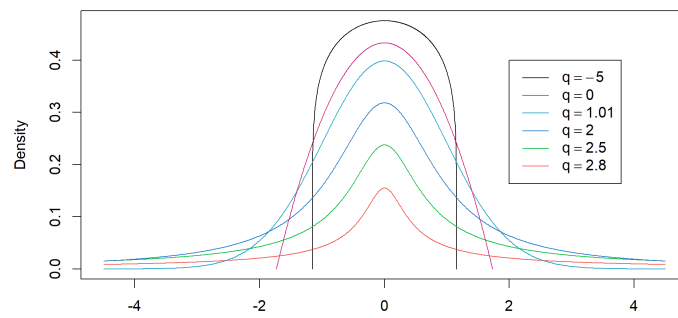


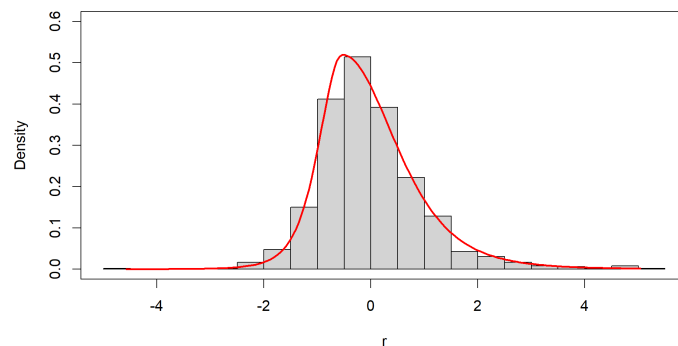
Figure 1. Normal and Student- $t$  Distributions.



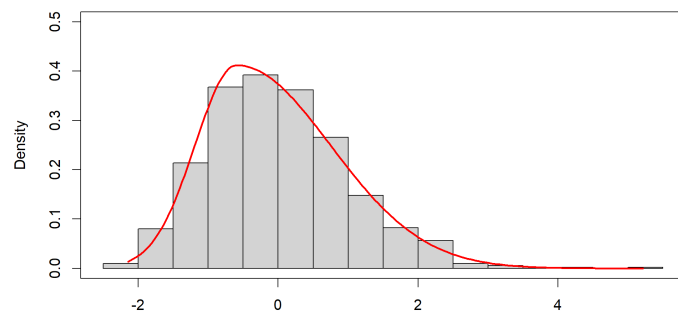
**Figure 2.**  $q$ -Gaussian Distributions.

The extreme losses which occurred in the financial crisis of 2008 highlighted the need to determine the correct distribution. Risk management can be based on any statistical time series model that captures the stylized facts, such as volatility clustering, skewness and tail thickness of their distribution, Stoyanov et al. (2011). Modelling volatility is considered a measure of risk, modelling and forecasting volatility is therefore important, Teräsvirta (2009).

Within the literature it can be noted that certain distributions are used for different financial modelling applications, Fukuda (2021). The Student- $t$  distribution for modelling exchange rates Figure 1, the Skewed Student- $t$  Figure 3 for foreign exchange rates, the Generalized Error distribution, Figure 4 for stock returns. The symmetrical Student- $t$  distribution Figure 1, is regarded as the most common and parsimonious model to use for economic and financial data. The student- $t$  distribution, Afuecheta et al. (2020) offers the ability to fit the leptokurtic properties of financial data, and can describe subtle features such as volatility clustering.



**Figure 3.** Skewed Student- $t$  Distribution.



**Figure 4.** Generalized Error Distribution.

Apart from the symmetric distributions there are asymmetrical distributions such as the skewed normal, skewed student- $t$  see Figure 3 and the skewed Generalized Error distribution, see Figure 4. As financial distributions are generally leptokurtosis distributions which have heavy tails, Heyde and Liu (2001), they can be hyperbolic distributions. Extreme

observations can extend to  $6 \geq$  standard deviations and can be of both interest and concern and have tails which are asymptotically of a Pareto distribution, see Figure 5.

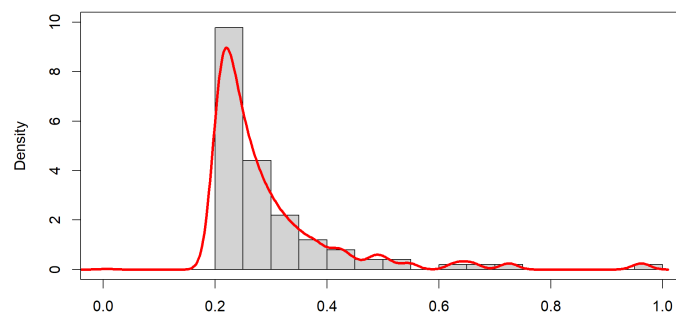


Figure 5. Pareto Distribution.

Initial modelling can identify the time series distribution which allows for applications within other modelling methods, such as bivariate regression, Liu et al. (2020) and/or bivariate copula modelling, Dewick and Liu (2022) as examples. As time series modelling is usually focused on forecasting, the time series data must be modelled to obtain the mathematical model that represents the dataset that will enable forecasting to be undertaken.

#### Financial Time Series Volatility, Leverage and Drift

The commonly used measure for risk within finance, Sheraz and Nasir (2021) is the standard deviation of the return, known as volatility. Volatility means that there are periods of time fluctuations followed by periods of calm, Abdulla and Dhaher Alwan (2022). Volatility interprets market risk, and its prediction is vital for empirical pricing, risk management, and portfolio selection, Sheraz and Nasir (2021). Furthermore, volatility can be broadly defined as the changeableness of the variable under consideration, Bentes et al. (2008). Volatility is not constant over time, volatility is volatile. The volatility can be measured in terms of the standard deviation  $\sigma$ , or variance  $\sigma^2$ , with the larger  $\sigma^2$ , implying higher volatility and risk, Lim and Sek (2013) and is given by;

$$\sigma^2 = \frac{1}{T-1} \sum_{t=1}^T (R_t - \mu)^2 \quad (2)$$

where:  $T$  is the time period,  $t$  denotes the time measures,  $\mu$  and  $R$  are the mean return and return, respectively, Sheraz and Nasir (2021).

Using the standard deviation  $\sigma$ , is the most popular measure of volatility. It has been noted, Bentes et al. (2008) that Equation (2), has the advantage of being easy to estimate but it has some drawbacks. These drawbacks include that large observations can overestimate the volatility and it ignores the nonlinear dynamics. The main body of research recognizes that the standard deviation is still the most popular method used measure.

A leverage effect is a negative correlation between shocks on returns and subsequent shocks on volatility, Caporin and Costola (2019). A negative return shock can produce an increase in volatility and a positive return shock produces a decrease in volatility. A leverage effect can be a special case of asymmetry as under leverage, positive and negative shocks have a different impact on the conditional variance. Often leverage is synonymous for asymmetry and is a common viewpoint.

The leverage effect is often matched with the asymmetry of the GARCH models. This however may not be totally reliable, as several GARCH models are not capable of showing leverage affects. A leverage effect is a special case of asymmetry and has a different impact on the conditional variance, Caporin and Costola (2019). Volatility and leverage effects are two different stylized phenomena. There are different regimes proposed in determining leverage effects within the literature and are outside the scope of this paper.

Time series drift, also referred to as “concept drift” in which the underlying generating process of the time series observations may change, making forecasting models obsolete, Oliveira et al. (2017). The drift parameter in a differenced model is an estimate of the period-to-period growth or stochastic “trend” which may or may not be significantly different.

### 3. Financial Time Series Models

Time series can be defined as a sequence of observations on one or more variables over time. Time is an important dimension because past events can influence future events, Liu et al. (2020). The challenges of time series modelling lie in constructing and applying the appropriate model and data transformations, Charpentire (2014). Financial time series data is non-stationary by nature which needs to be *modelled* out when modelling using the Box-Jenkins methods. The Box-Jenkins methods consist of ARCH, ARMA or the ARIMA type models for mean modelling, and AR or GARCH type models to model the variance, known as the conditional volatility. When financial time series contains both non-stationarity and volatility these models can be combined, such as the ARMA-GARCH type models as an example. Within the literature, the GARCH type models are considered the best models for forecasting stock market volatility, Lim and Sek (2013).

The AR and the ARCH models can be considered as “bursty”, short bursts of variance, then back to the mean, with the GARCH model contains larger “bursts”, longer periods of variance, then back to the mean. These models are based on the standard deviation or variance of the time series data, Bentes et al. (2008). The GARCH models are frequently used for modelling stock price volatility, with the GARCH(1,1) being the most widely used. The GARCH(1,1) model is used under the assumption of *t*-Student distribution.

Another financial time series modelling approach is to use the GBM type models. GBM is a stochastic differential equation with time dependent drift and diffusion parameters. The GBM is often described as a stochastic model with continuous time, where the random variable follows the Geometric Brownian motion, Agustini et al. (2018). Financial modelling using a GBM model may require many simulations to obtain a GBM model that matches the time series dataset.

Additionally, undertaken within this study is the use of Tsallis entropy to generate a *q*-Gaussian distribution that can give an indication of the fat or thin tails within the datasets distribution. A limiting factor in using entropy-based methods is that the entropy method requires a large amount of data. A reliable fitting of a *q*-Gaussian distribution to the empirical data, a large amount of data is needed as fitting return stock volumes, a Tsallis *q*-Gaussian distribution requires 10<sup>6</sup> data points, see de Santa Helena et al. (2018).

#### 3.1. Box-Jenkins Time Series Model Notation

Box-Jenkins time series models typically consist of AR models and MA models type models and may contain combinations of these. These ARMA type models specifies the conditional mean of the process and the GARCH type models specifies the conditional variance of the process, with the models being defined by their notation. These time series models typically consist of AR and MA type models.

The basic Box-Jenkins ARIMA model is a non-seasonal model with the notation as ARIMA(*p, d, q*) model, with *p*; the auto regressive part, *d* being the degree of first differencing and *q*, the order of the moving average. The ARIMA seasonal model is given as:

$$ARIMA = \underbrace{(p, d, q)}_{\text{Nonseasonal Part}} + \underbrace{(P, D, Q)_m}_{\text{Seasonal Part}} \quad (3)$$

where: *m* = length of seasonality, seasonal period time points.

The ARMA model is given as ARMA(*p, q*). If an ARIMA model contained no nonseasonal differences *d* < 0, an ARMA(*p, q*) model can be used. Therefore an ARIMA(*p, 0, q*) = ARMA(*p, q*), Wheelwright et al. (1998). The GARCH model is given as GARCH(*p, q*), where *p* is the number of lag variances to include and *q*, is the number of lag residual errors

to include in the GARCH model. For a GARCH where  $p = 0$ , this reduces the model to an ARCH( $q$ ) model, Bollerslev (1986).

### 3.2. GARCH Type Models

GARCH type models are used to analyze and forecast volatility, Charpentire (2014). The ARCH model describes a volatile variance over time and has all past error terms. The ARCH model is effective for any time series that has increased or decreased variance, Sheraz and Nasir (2021).

The GARCH model is an extension of the ARCH model, Charles and Darné (2019) allowing the conditional variance to be dependent on the previous lags. The GARCH model is widely used to estimate the non-constant volatility, depending on time and provides a good approximation for smooth and persistent changes in volatility, Hongweingjan and Thongtha (2021). If the decay rate is too rapid compared to what is typically observed in financial time series a GARCH model is required, Teräsvirta (2009). The conditional variance that describes an ARCH model of order  $q$ , can be defined as:

$$h_t = \alpha_0 + \sum_{j=1}^q \alpha_j \epsilon_{t-1}^2 \tag{4}$$

where:  $\alpha_0 > 0, \alpha_j \geq 0, j = 1, \dots, q - 1$  and  $\alpha_q > 0$ . The observed random variable  $y_t$  and  $u_t(y_t) = E\{y_t | F_{t-1}\}$  and  $\epsilon_t$  is a random variable that has a mean and variance on the information set  $F_{t-1} = 0$ , with the conditional variance being  $h_t = E\{\epsilon_t^2 | F_{t-1}\}$ , Teräsvirta (2009).

There are variations and a rich abundance of families of GARCH type models which are popular, Sheraz and Nasir (2021) as they are flexible to capture the volatility clustering, also the GARCH type models can capture asymmetries within the data, Abdulla and Dhaher Alwan (2022). The family of GARCH models include, but are not limited to the EGARCH, which is the Exponential GARCH, GJR-GARCH, which is the Glosten-Jagannathan-Runkle GARCH and TGARCH, which is the Threshold GARCH type models. The most popular GARCH model is the GARCH(1,1) model where  $p = q = 1$ , Teräsvirta (2009). The GARCH( $p, q$ ) models with  $p, q \geq 2$ , are rare in practice.

SGARCH–The SGARCH or standard (ordinary) GARCH assumes symmetric effects on volatility, it assumes normality condition for errors, Sheraz and Nasir (2021). As a result, the standard GARCH fails to account for excessive skewness or kurtosis within the modelled distribution. The conditional variance that describes GARCH models can be defined as, Teräsvirta (2009):

$$h_t = \alpha_0 + \sum_{i=1}^p \alpha_i \epsilon_{t-1}^2 + \sum_{j=1}^q \beta_j h_{t-1} \quad \text{for } t \in \mathbb{Z} \tag{5}$$

The standard first-order model GARCH model, GARCH(1, 1) is the most common in practice and the conditional variance ( $h_t = \sigma_t^2$ ) can be given as, Sheraz and Nasir (2021):

$$\sigma_t^2 = w + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{6}$$

where:  $\mathbb{Z}$  are iid random variables,  $w = \alpha_0$  and  $w > 0, \alpha \geq 0, \beta_1 \geq 0$ , are real parameters and ensures that  $\sigma^2 > 0$ .

EGARCH–The exponential EGARCH is another popular GARCH model, Teräsvirta (2009) and does not allow for negative volatility. The EGARCH was proposed to model the financial models leverage effects, Sheraz and Nasir (2021), with the family of EGARCH( $p, q$ ) models can be defined as, Teräsvirta (2009):

$$\ln h_t = \alpha_0 + \sum_{i=1}^p g_i(z_{t-i}) + \sum_{j=1}^q \beta_j \ln h_{t-j} \tag{7}$$

The standard first-order model EGARCH model, EGARCH(1,1) can be given as, Sheraz and Nasir (2021):

$$\ln(\sigma_t^2) = w + \alpha_1(|Z_{t-1}| - E(|Z_{t-1}|)) + \beta_1 \ln(\sigma_{t-1}^2) + \gamma_1 Z_{t-1} \tag{8}$$

where:  $\beta_j$  is a persistence parameter,  $\alpha_1 \geq 0$ ,  $\beta_1 \geq 0$ ,  $|\gamma_1| < 1$ , and  $w > 0$ , and  $\alpha_1$  and  $\gamma_1$  represents the sign and leverage effects. The EGARCH can capture serial dependence and leverage effects in the returns, with the returns being stationary if  $0 < \beta_1 < 1$ .

GJR-GARCH–The GJR-GARCH models are used to model positive and negative shocks on the conditional variance asymmetrically. Applications of the GJR-GARCH is to capture the negative correlation between returns and volatility, Sheraz and Nasir (2021). The conditional variance that describes a GJR-GARCH model can be defined as, Teräsvirta (2009):

$$h_t = \alpha_0 + \sum_{i=1}^p \{\alpha_j + \delta_j I(\epsilon_{t-j} > 0)\} \epsilon_{ij}^2 + \sum_{j=1}^q \beta_j \ln h_{t-j} \tag{9}$$

The standard first-order model GJR-GARCH model, GJR-GARCH(1,1) can be given as, Sheraz and Nasir (2021):

$$\sigma_t^2 = w + (\alpha_1 + \gamma_1 I_{t-1}) \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{10}$$

where:  $\alpha_1 > 0$ ,  $\beta_1 > 0$ ,  $\gamma_1 > 0$ ,  $w > 0$  and  $\gamma$  indicates the asymmetry of returns. The  $I_{t-1}$  assumes value equals to 1 for  $\eta_{t-1}^2 < 0$  (negative-shock), and zero otherwise. For positive and significant  $\gamma_1$ , a leverage effect exists.

TGARCH–The threshold GARCH is similar to the GJR model, different only because of the standard deviation, instead of the variance. The TGARCH allows for the analysis of negative and positive return shocks on the volatility, Lim and Sek (2013) with the family of TGARCH( $p, q$ ) models can be defined as, Teräsvirta (2009):

$$h_t^{1/2} = \alpha_0 + \sum_{i=1}^p (\alpha_j^+ \epsilon_{t-j}^+ - \alpha_j^- \epsilon_{t-j}^-) + \sum_{j=1}^q \beta_j \ln h_{t-j}^{1/2} \tag{11}$$

The standard first-order model TGARCH model, TGARCH(1,1) can be given as, Sheraz and Nasir (2021):

$$\sigma_t^2 = w + (\alpha_1 + \gamma_1 I_{t-1}) \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{12}$$

where:  $\alpha_{1+} \geq 0$ ,  $\alpha_{1-} \geq 0$ ,  $\beta_1 \geq 0$ , and  $w > 0$  are real numbers. The volatility depends on both the modulus and the sign of the past returns through  $\alpha_{1,+}$  and  $\alpha_{1,-}$ .

### 3.3. Geometric Brownian Motion Type Models

The Option pricing industry was largely fueled by the success of Black and Scholes (1973), in obtaining an analytical pricing formula for European Options under the Geometric Brownian Motion (GBM) model, Heyde and Liu (2001). The GBM process can be defined as a stochastic process where  $X_t \geq 0$ , Khamis et al. (2017). Black and Scholes postulated a log normal model for stock prices, Heyde et al. (2001), and that the stock returns process  $X_t$  is given by:

$$X_t = \log \frac{S_t}{S_{t-1}} \tag{13}$$

A stochastic process  $S_t$ , is said to follow a GBM if it satisfies the following stochastic differential equation, where  $\mu$  is the percentage drift and  $\sigma$  is the percentage volatility, for arbitrary initial values of  $S_0$ , Ermogenous (2006):

$$dS_t = S_t(\mu dt + \sigma dB_t) \tag{14}$$

With the analytical solution given as:

$$dS_t = S_0 e^{(\mu - \frac{\sigma^2}{2})t + \sigma dB_t} \tag{15}$$

If the stochastic process, Islam and Nguye (2021) is defined as  $X_t = \log S_t$  and  $\{W(t) : 0 \leq t \leq T\}$  it is a standard Brownian motion on  $[0, T]$  then:

$$dS_t = \mu S_t dt + \sigma S_t dW_t \tag{16}$$

For any time  $t > 0$ , the differential can be written as:

$$\log S_t = \log S_0 + \left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma W_t, \quad \text{or} \tag{17}$$

$$S_t = S_0 e^{(\mu - \frac{1}{2}\sigma^2)t + \sigma W_t} \tag{18}$$

For a time set,  $t_0 = 0 < t_1 < t_2 \dots < t_n$ , a stock price  $S(t)$  at time  $t_0, T_1, \dots, t_n$  can be generated by, Islam and Nguye (2021);

$$S(t_{i+1}) = S(t_i) e^{(\mu - \frac{1}{2}\sigma^2)(t_{i+1} - t_i) + \sigma \sqrt{(t_{i+1} - t_i)} Z_{i+1}} \tag{19}$$

where:  $Z_1, Z_2, \dots, Z_n$  are iid standard normals and the time interval  $t_{i+1} - t_i = 1$  for all  $i = 0, (n - 1)$ , since predicting next day price is given as;

$$S(t_{i+1}) = S(t_i) e^{(\mu - \frac{1}{2}\sigma^2) + \sigma Z_{i+1}} \tag{20}$$

where:  $\mu$  is the amount of change over time (called the drift), and  $\sigma$  is the volatility.

### 3.4. Tsallis Entropy Type Models

Using an entropy approach to time series modelling is through the use of the concept of Tsallis entropy which captures the nature of volatility, Bentes et al. (2008). The entropy process consists of using Tsallis entropy models which is based on Shannon’s entropy. The term entropy can be viewed as the measure of disorder, uncertainty, or ignorance, Sheraz and Nasir (2021) of a system which also resembles the features associated with the stock market with entropy being used to study stock market volatility, Bentes et al. (2008). The Shannon entropy corresponding to a discrete random variable  $X$ , of probability measure  $P = \{p_1, p_2, \dots, p_n\}$ , can be defined as;

$$S(X) = - \sum_{i=1}^n p_i \ln p_i \tag{21}$$

Tsallis derived a generalized form of entropy, known as Tsallis entropy, Bentes et al. (2008). When the entropy takes a non-additive form that involves a parameter  $q$ , this reduces the entropy in the limit of  $q = 1$ , which is referred to as Tsallis statistics, Kapusta (2021). Tsallis entropy is a non-extensive entropy, Sheraz and Nasir

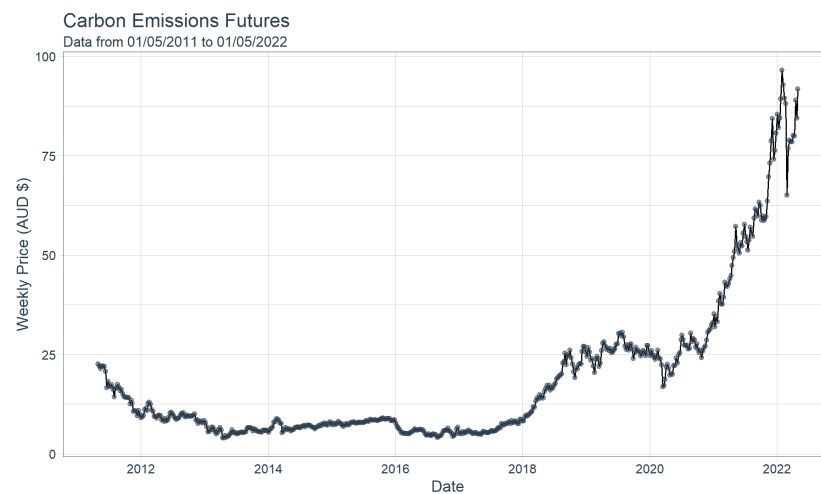
(2021). The Shannon entropy recovers as  $q \rightarrow 1$ , where  $q$  is the parameter of the Tsallis entropy. The Tsallis entropy is defined as;

$$S_q(X) = \frac{1 - \sum_{i=1}^m p_i^q}{q - 1} \quad (22)$$

Tsallis entropy under the constraint of normalization and variance, Sato (2010) leads to a  $q$ -Gaussian distribution and the  $q$ -Gaussian distribution has power-law tails when  $q > 1$ , as shown at Figure 4. The underlying statistical dynamics is Gaussian if  $q = 1$ , Pavlos et al. (2014). As the system moves away from equilibrium, the underlying statistical dynamics become non-Gaussian,  $q \neq 1$ . A normal diffusion is when  $q = 1$ , anomalous sub-diffusion (resulting from thin tails) for  $q < 1$  and super-diffusion (resulting from heavy tails) for  $1 < q < 3$ , Tsallis (2017). As the value of the parameter  $q$  decreases to 1, the frequency of the data decreases and values where  $1 \leq q \leq 2$  emphasize highly volatile signals, Sheraz and Nasir (2021).

#### 4. A Financial Time Series Modelling Application

The financial Carbon Emissions Futures price between the dates of 1 May 2012 and 1 May 2022 which is shown at Figure 6, has been modelled using the Box–Jenkins, GBM and the Tsallis time series modelling methods. Modelling to determine the datasets distribution and undertaking simple linear regression modelling allows for a goodness-of-fit assessment between the original and generated time series models, which can serve as a criterion for model selection.



**Figure 6.** Time Series Plot.

##### 4.1. Determining the Initial Distribution

Determining the time series distribution was undertaken using the Box-Jenkins methodology to stabilize the time series dataset, see Figure 7. The resulting distribution of the log-differenced residual distribution at Figure 8, shows that the residuals are normal slightly skewed with a long thin tail. The log-differenced results of the Carbon Emissions price distribution at Table 1, produced a mean  $\hat{\mu} = 0.0024$  and the standard deviation  $\hat{\sigma} = 0.0681$  indicates a low level of dispersion. The kurtosis of the distribution suggests that low values of the Carbon price resulted in volatility, see Figures 8 and 9.



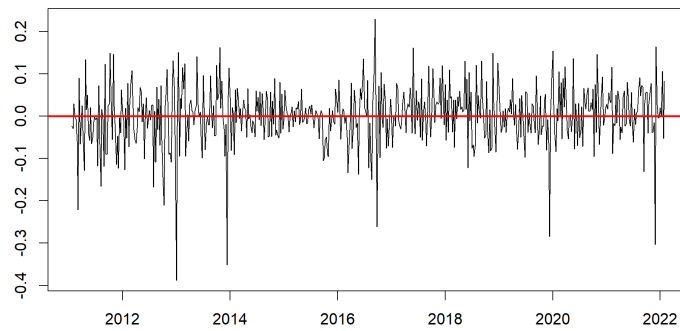


Figure 7. Log–Differenced Residuals.

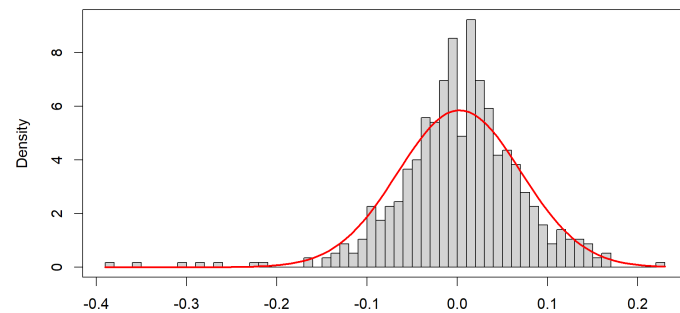


Figure 8. Log–Differenced Residual Distribution.

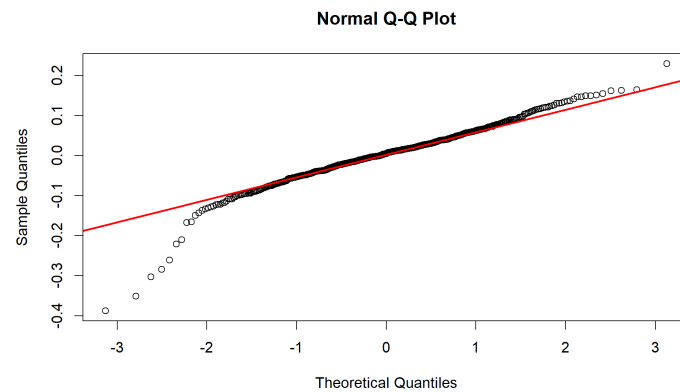


Figure 9. Log–Differenced Residual Q-Q Plot.

Table 1. Modelling Results.

Model	$\hat{\mu}$	$\hat{\sigma}$	Intercept	Slope	Std Error	$R^2$	$p$ -Value	AIC	BIC
Log-diff.	0.002	0.068	−0.011	0.000	0.068	0.011	0.007	−1457.4	−1444.4
ARIMA	0.000	0.084	0.000	0.000	0.084	−0.002	0.986	−1206.6	−1193.6
SGARCH	0.061	0.999	−0.156	0.001	0.993	0.014	0.003	1627.4	1640.4
TGARCH	0.026	1.000	−0.237	0.001	0.990	0.021	0.000	1624.3	1637.4
GJR-GARCH	0.057	1.036	−0.323	0.001	1.013	0.043	0.000	1651.1	1664.1
GBM	0.002	0.068	−0.006	0.000	0.063	0.011	0.011	−1532.1	−1519.0
Tsallis	0.033	0.806	−0.452	0.000	0.890	0.001	0.644	725.8	736.6

The distribution at Figure 8, is the time series distribution for the Carbon price. Further modelling is required in constructing a mathematical model that will allow for forecasting. The distribution at Figure 8, can be used within a bivariate regression model, Liu et al. (2020) and/or a bivariate copula model, Dewick and Liu (2022) if required.

Regression modelling has been used as a methodology in determining a suitable distribution that represents the initial distribution, see Figure 8 which can be used in model selection. The regression model for the initial distribution is shown at Figure 10, with the

results shown at Table 1. This allows the initial distribution to be used as a baseline in comparing other distributions produced from using the Box-Jenkins, Geometric Brownian Motion and Tsallis methods.

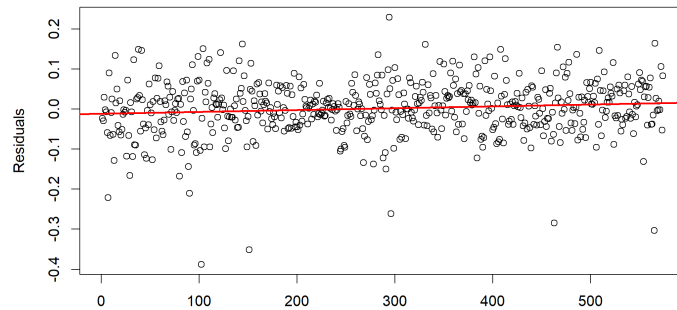


Figure 10. Regression Plot for the Log-Differenced Residual Distribution.

#### 4.2. Box-Jenkins Time Series Modelling Methodology

The Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) plots identified at Figure 11, shows that there is a persistence of volatility, therefore a GARCH type model is required.

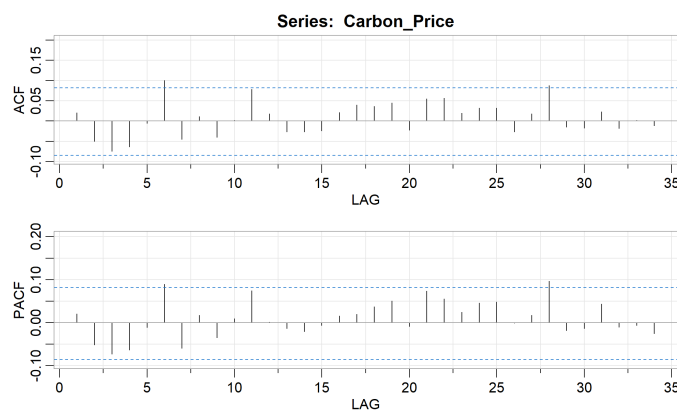


Figure 11. ACF and PACF of the Log-Differenced Carbon Emissions Dataset.

Initial ARIMA modelling of the dataset at Figure 6, was undertaken using the **R** package function *auto.arima* and gave the best fitting model as:  $ARIMA(1, 2, 0)$ . Further modelling was undertaken using the **R** package function *rugarch*. This package produced the best fitting models as:  $SGARCH(1, 0, 2)(2, 1)$ ,  $TGARCH(1, 0, 2)(1, 1)$  and the  $GJR-GARCH(1, 0, 2)(2, 1)$ . These models were modeled for their distributions and regression modelling was undertaken. The TGARCH model residuals and Q-Q plot is shown at Figures 12 and 13. Further GARCH type models could also have been used but this paper has used the ones shown as examples.

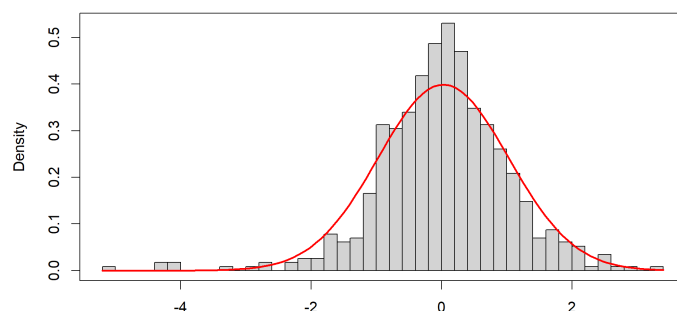


Figure 12. TGARCH Model Residual Distribution.

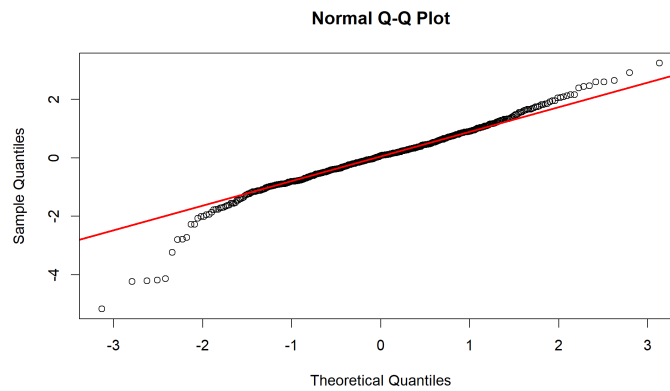


Figure 13. TGARCH Residual Distribution Q–Q Plot.

Simple linear regression modelling using the original and TGARCH distributions, see Equation (1), are shown at Figures 10 and 14. The results for all the modelled distributions using simple linear regression undertaken are given at Table 1.

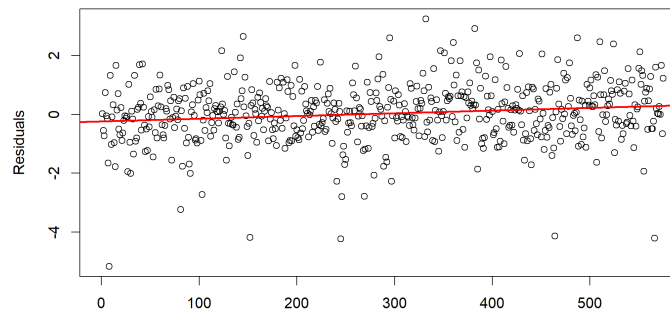


Figure 14. Regression Plot for the TGARCH Residual Distribution.

Overall, the results show that the time series distribution is symmetric. The lower tail being thin, suggests that the skew represents rare events. This is highlighted in comparing the Q–Q plots similarities at Figures 9 and 13.

#### 4.3. Time Series Brownian Motion Results

A Time Series Geometric Brownian Motion simulation was undertaken with input parameters for the GBM being identified using the initial Box-Jenkins model results ( $\hat{\mu} = 0.0024$ ,  $\hat{\sigma} = 0.0681$ ), shown at Figure 8. To allow for the simulation to represent the dataset, hundreds of simulations Khamis et al. (2017), may be required. GBM simulations tend to fluctuate wildly, this is highlighted at Figure 15, as 518 simulations were required to produce the better fitting model which is shown at Figure 16.

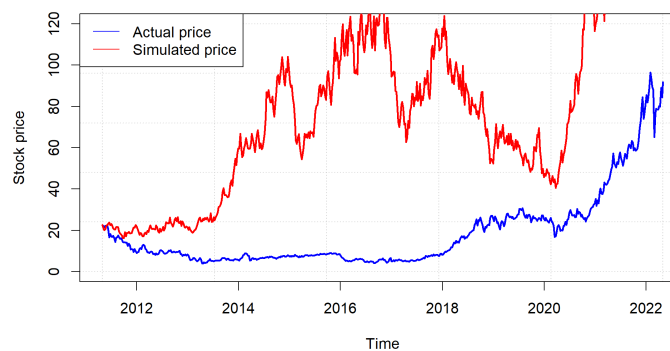
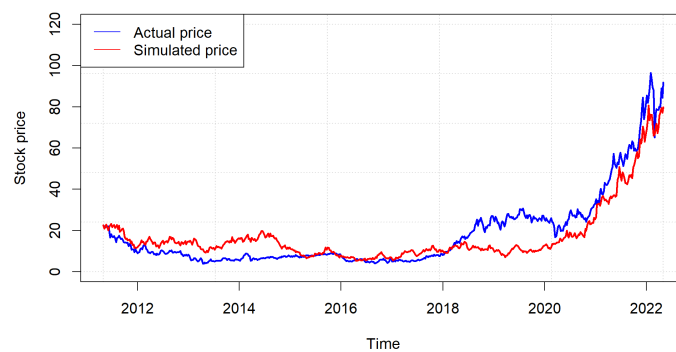


Figure 15. GBM Simulation No.105.



**Figure 16.** GBM Simulation No.518.

The Geometric Brownian Motion model required the distribution required to be log-differenced to stabilize the data using the same methodology shown at Figures 7 and 8. The simple linear regression modelling results for the Geometric Brownian Motion are given at Table 1.

#### 4.4. Tsallis Entropy Results

A Tsallis entropy estimation was undertaken using the **R** package *q-Gaussian*, de Santa Helena and de Lim (2018). This **R** package allows for the calculation of the Tsallis  $q$  value in determining the  $q$ -Gaussian distribution see Figure 2, that would indicate if the time series distribution contained fat or thin tails. The simple linear regression model for the Tsallis modelling results is at Table 1.

The **R** package *tsallisexp* is a package that can be used to determine the quality of fit to an exponential distribution, as  $q \rightarrow 1$ , an exponential distribution is obtained, Shalizi and Dutang (2021). Due to the small dataset size 575 data points, far less than what is required,  $10^6$  data points. Entropy Modelling is a methodology that can be easily undertaken should a large enough dataset be available.

### 5. Modelling Results Summary

Simple linear regression was undertaken on the resulting distributions shown at Table 1. The results show the GBM model, Figure 16 gave the best modelling results that matched the original log-differenced dataset. The results also shows that modelling using the GARCH family, the TGARCH model bests goodness-of-fit with the original log-differenced dataset as there is slight volatility. This slight volatility (skew) was highlighted within the original distribution at Figures 8 and 9.

### 6. Conclusions

This study has modelled time series data using three different modelling methods in identifying the underlying distribution that can result due to the different phenomenon that affect the financial market. The purpose and utility of determining the underlying distribution is three-fold, firstly, these distributions can be used with regression modelling as a goodness-of-fit criterion when undertaking forecasting modelling. Secondly, the distribution can be applied in other modelling methods, such as regression and copula modelling. Thirdly, by understanding the initial distribution of the dataset will give insight to the possible presents of volatility (and leverage affects) and drift that can result from different phenomenon's acting within the financial market.

Using Tsallis entropy for time series modelling could be a good option in determining the distribution, however it does require a large dataset ( $10^6$ ), which may not be practically available. This study used Tsallis Entropy modelling in determining the  $q$ -Gaussian distribution, but it failed to reproduce the valid results. It is unsure how small a dataset could be to produce a  $q$ -Gaussian distribution with reasonable accuracy in determining the underlying distribution using this application, this could be a topic for further research. In an

environment using Big-Data, Tsallis entropy could be a quick methodology in identifying the underlying distribution.

Future applications may include that time series modelling, not just be focused on determining a predictive forecasting model, but rather being more focused on determining and identifying the underlying distribution and parameters which aids in model selection by using regression methods, then proceed to either other modelling methods, such as regression, copula modelling or to forecasting methods and methodologies.

Limitations of this study is that only a few GARCH type models were used, being the SGARCH, TGARCH and GJR-GARCH type models. This study has highlighted that there are many GARCH type models that can be used. Familiarity for all the GARCH type models should be obtained that will allow for the modelling the stylized facts from the time series distribution.

**Funding:** This research received no external funding.

**Data Availability Statement:** Publicly available dataset was analysed in this study. This data can be found here: <https://au.investing.com/commodities/carbon-emissions-historical-data>, accessed on 11 May 2022.

**Acknowledgments:** I would like to acknowledge the reviewers for their constructive comments.

**Conflicts of Interest:** The author declare no conflict of interest.

## References

- Abdulla, Suhail Najm, and Heba Dhaher Alwan. 2022. Using apgarch/avgarch models Gaussian and non-Gaussian for modeling volatility exchange rate. *International Journal of Nonlinear Analysis and Applications* 13: 3029–38.
- Afuecheta, Emmanuel, Artur Semeyutin, Stephen Chan, Saralees Nadarajah, and Diego Andrés Pérez Ruiz. 2020. Compound distributions for financial returns. *PLoS ONE* 15: e0239652. [CrossRef] [PubMed]
- Agustini, W. Farida, Ika Restu Affianti, and Endah R. M. Putri. 2018. Stock price prediction using geometric Brownian motion. *Journal of Physics: Conference Series* 974: 012047.
- Bentes, Sónia R., Rui Menezes, and Diana A. Mendes. 2008. Stock market volatility: An approach based on Tsallis entropy. *arXiv*, arXiv:0809.4570.
- Bollerslev, Tim. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31: 307–27. [CrossRef]
- Caporin, Massimiliano, and Michele Costola. 2019. Asymmetry and leverage in GARCH models: A news impact curve perspective. *Applied Economics* 51: 3345–64. [CrossRef]
- Charles, Amélie, and Olivier Darné. 2019. The accuracy of asymmetric GARCH model estimation. *International Economics* 157: 179–202. [CrossRef]
- Charpentier, Arthur. 2014. *Computational Actuarial Science with R*. New York: CRC Press.
- de Santa Helena, Emerson Luis, C. M. Nascimento, and G. Gerhardt. 2015. Alternative way to characterize a q-Gaussian distribution by a robust heavy tail measurement. *Physica A: Statistical Mechanics and Its Applications* 435: 44–50. [CrossRef]
- de Santa Helena, Emerson Luis, and Wagner Santos de Lima. 2018. Package ‘qGaussian’. R Package Version 0.1.8. Available online: <https://CRAN.R-project.org/package=qGaussian> (accessed on 24 April 2022).
- Devi, Sandhya. 2021. Asymmetric Tsallis distributions for modeling financial market dynamics. *Physica A: Statistical Mechanics and Its Applications* 578: 126109. [CrossRef]
- Dewick, Paul R., and Liu Shuangzhe. 2022. Copula modelling to analyse financial data. *Journal of Risk and Financial Management* 15: 104. [CrossRef]
- Ermogenous, Angeliki. 2006. Brownian Motion and Its Applications in the Stock Market. Available online: [https://ecommons.udayton.edu/cgi/viewcontent.cgi?article=1010&context=mth\\_epumd](https://ecommons.udayton.edu/cgi/viewcontent.cgi?article=1010&context=mth_epumd) (accessed on 12 May 2022).
- Fukuda, Kosei. 2021. Selecting from among 12 alternative distributions of financial data. *Communications in Statistics-Simulation and Computation* 51: 3943–3954. [CrossRef]
- Ghani, I. M. Md and H. A. Rahim. 2019. Modeling and forecasting of volatility using arma-garch: Case study on malaysia natural rubber prices. *IOP Conference Series: Materials Science and Engineering* 548: 012023. [CrossRef]
- Hambuckers, Julien, and Cedric Heuchenne. 2017. A robust statistical approach to select adequate error distributions for financial returns. *Journal of Applied Statistics* 44: 137–61. [CrossRef]
- Heyde, C. Heyde, and Shuangzhe Liu. 2001. Empirical Realities for Minimal Description Risky Asset Model. The Need for Fractal Features. *Journal of the Korean Mathematical Society* 38: 1047–59.
- Heyde, C. Heyde, Roger Gay, and Shuangzhe Liu. 2001. Fractal scaling and Black-Scholes [A new view of long-range dependence in stock prices]. *JASSA* 1: 29–32.

- Hongwiengjan, Warunya, and Dawud Thongtha. 2021. An analytical approximation of option prices via TGARCH model. *Economic Research-Ekonomiska Istraživanja* 34: 948–69. [CrossRef]
- Islam, Mohammad Rafiqul, and Nguyet Nguyen. 2021. Comparison of Financial Models for Stock Price Prediction. *Joint Mathematics Meetings (JMM)* 13: 181. [CrossRef]
- Kapusta, Joseph I. 2021. Perspective on Tsallis statistics for nuclear and particle physics. *International Journal of Modern Physics E* 30: 2130006. [CrossRef]
- Khamis, Azme, M. A. A. Sukor, M. E. Nor, S. N. A. M. Razali, and R. M. Salleh. 2017. Modeling and Forecasting Volatility of Financial Data using Geometric Brownian Motion. *International Journal of Advanced Research in Science, Engineering and Technology* 4: 4599–605.
- Lim, Ching Mun, and Siok Kun Sek. 2013. Comparing the performances of GARCH-type models in capturing the stock market volatility in Malaysia. *Procedia Economics and Finance* 5: 478–87. [CrossRef]
- Liu, Shuangzhe, and Chris C. Heyde. 2008. On estimation in conditional heteroskedastic time series models under non-normal distributions. *Statistical Papers* 49: 455–69. [CrossRef]
- Liu, Timina, Shuangzhe Liu, and Lei Shi. 2020. *Time Series Analysis Using SAS Enterprise Guide*. Singapore: Springer Nature.
- Oliveira, Gustavo H. F. M., Rodolfo C. Cavalcante, George G. Cabral, Leandro L. Minku, and Adriano L. I. Oliveira. 2017. Time series forecasting in the presence of concept drift: A pso-based approach. Paper presented at the 2017 IEEE 29th International Conference on Tools with Artificial Intelligence ICTAI), Boston, MA, USA, November 6–8. pp. 239–46.
- Pavlos, G. P., L. P. Karakatsanis, M. N. Xenakis, E. G. Pavlos, A. C. Iliopoulos, and D. V. Sarafopoulos. 2014. Universality of non-extensive Tsallis statistics and time series analysis: Theory and applications. *Physica A: Statistical Mechanics and Its Applications* 395: 58–95. [CrossRef]
- Sato, Aki-Hiro. 2010. q-Gaussian distributions and multiplicative stochastic processes for analysis of multiple financial time series. *Journal of Physics: Conference Series* 201: 012008. [CrossRef]
- Shalizi, Cosma, and Christophe Dutang. 2021. tsallisqexp: Tsallis Distribution. R Package Version 0.9-4. Available online: <https://CRAN.R-project.org/package=tsallisqexp> (accessed on 15 May 2020).
- Sheraz, Muhammad, and Imran Nasir. 2021. Information-Theoretic Measures and Modeling Stock Market Volatility: A Comparative Approach. *Risks* 9: 89. [CrossRef]
- Stoyanov, Stoyan V., Svetlozar T. Rachev, Boryana Racheva-Yotova, and Frank J. Fabozzi. 2011. Fat-tailed models for risk estimation. *The Journal of Portfolio Management* 37: 107–17. [CrossRef]
- Teräsvirta, Timo. 2009. An introduction to univariate GARCH models. In *Handbook of Financial Time Series*. Berlin and Heidelberg: Springer, pp. 17–42.
- Tsallis, Constantino. 2017. Economics and Finance: q-Statistical stylized features galore. *Entropy* 19: 457. [CrossRef]
- Wheelwright, Steven, Spyros Makridakis, and Rob J. Hyndman. 1998. *Forecasting: Methods and Applications*. Hoboken: John Wiley & Sons.

Article

# Modeling Bivariate Dependency in Insurance Data via Copula: A Brief Study

Indranil Ghosh <sup>1</sup>, Dalton Watts <sup>1</sup> and Subrata Chakraborty <sup>2,\*</sup>

<sup>1</sup> Department of Mathematics and Statistics, University of North Carolina, Wilmington, NC 28403, USA; ghoshi@uncw.edu (I.G.); dow3346@uncw.edu (D.W.)

<sup>2</sup> Department of Statistics, Dibrugarh University, Assam 786004, India

\* Correspondence: subrata\_stats@dibru.ac.in

**Abstract:** Copulas are a quite flexible and useful tool for modeling the dependence structure between two or more variables or components of bivariate and multivariate vectors, in particular, to predict losses in insurance and finance. In this article, we use the VineCopula package in R to study the dependence structure of some well-known real-life insurance data and identify the best bivariate copula in each case. Associated structural properties of these bivariate copulas are also discussed with a major focus on their tail dependence structure. This study shows that certain types of Archimedean copula with the heavy tail dependence property are a reasonable framework to start in terms modeling insurance claim data both in the bivariate as well as in the case of multivariate domains as appropriate.

**Keywords:** bivariate copula; measures of association; dependence modeling; Kendall's  $\tau$ ; Blomqvist's  $\beta$



**Citation:** Ghosh, Indranil, Dalton Watts, and Subrata Chakraborty. 2022. Modeling Bivariate Dependency in Insurance Data via Copula: A Brief Study. *Journal of Risk and Financial Management* 15: 329. <https://doi.org/10.3390/jrfm15080329>

Academic Editors: Shuangzhe Liu, Tiefeng Ma and Seng Huat Ong

Received: 27 May 2022

Accepted: 20 July 2022

Published: 25 July 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Modeling insurance data via copula is not new in the literature. For example, Alexeev et al. (2021) studied dependence among insurance claims arising from different lines of business via copula. Shi et al. (2016) discussed a multilevel modeling of insurance claims using copula. Pfeifer and Neslehova (2003) discussed at length in a survey paper the role of copula in modeling dependence in finance and insurance. In exploring dependence structures related to insurance (from any business domain, such as healthcare sector, travel industry, etc.) one pertinent aspect is the assessment of various types of risk (for example, portfolio risk) arising out of each of these domains. There is no denying of the fact that without the proper assessment of risk, insurance coverage to public and private property/organization (as the case may be) as well as for individuals associated cannot be evaluated effectively. Consequently, in the literature, there are several instances of using bivariate and/or multivariate copula and studying their tail dependence behavior. For a detailed study on copula and associated bivariate (as well as multivariate) dependence based on copula theory, see the books by Joe (1997) and Nelsen (2006). A non-exhaustive list of such references may be cited as follows. Mensi et al. (2017) has discussed via a wavelet-based copula approach the dependence structure across oil, wheat, and corn. The authors have established time varying asymmetric tail dependence (at different time zones) between the pair of cereals as well as between oil and the two cereals. Naeem et al. (2021) studied the asymmetric and extreme tail dependence between five energy markets and green bonds using a time-varying optimal copula. This serves as a motivation for the current work. In this article, we focus on studying the dependence structure between two components resulting from insurance claim datasets. Specifically, we consider Australian automobile insurance data and the Swedish motor insurance data. There is little or no evidence of studying automobile insurance data that are asymmetric in nature via copula. This is another motivation to carry out this work. These datasets are selected from a wide collection of CAS datasets available in R. Here, we consider a copula-based modeling of

insurance claims data, especially the tail dependence and through a specific selection criteria in R, popularly known as the VineCopula package, to select the best fitted copula in each of these datasets. This paper investigates dependence among insurance claims arising from the auto industry with datasets selected from two different countries. Interestingly, for the first dataset, the Australian automobile insurance data, we examine the dependence among (pairwise) four different variables; each such comparison is useful in the context of claims assessed from the insured as well as the insurer. The details are provided in each model description in Section 3. For a detailed study on the use of copula, see Shi et al. (2016) and the references cited therein. The second dataset is taken from two different countries on motor insurance claims. In this case, we study the tail dependence between claims submitted and the number of insured motorists. When modeling dependency between components of insurance claims using copula, we aim to select copulas that are capable of generating upper- and or lower-tail dependence, that is, when several components of the insurance claims have a strong tendency to exhibit extreme losses simultaneously. We expect that the outcomes of this study provide valuable insights with regards to the nature of dependence and satisfy one of the primary objectives of the general insurance providers aiming at assessing total risk of an aggregate portfolio of losses when components of insurance are correlated. General insurance (for example, property–casualty) protects individuals and organizations from financial losses due to property damage or legal liabilities, in our case, due to auto accident. Consequently, it allows policyholders to exchange the risk of a large loss for the certainty of smaller periodic payments of premiums. Next, insurers allocates the bulk of premium dollars into investment and claims payments. As it is for an insurer to manage its investment portfolio, it is equally important for the insurer to manage its claim portfolio. It is the counterpart of asset management for the claims on the insurer’s book. Claim management is the analytics of insurance costs. It requires applying statistical techniques in the analysis and interpretation of the claims data. In the data-driven industry of general insurance, claim management provides useful insights for insurers to make better business decisions. From the above, it is quite evident as to why a study of insurance claims via copula is important.

In this article, we aim to model the dependence structure (in the bivariate domain) of data arising out of financial domains, precisely, from the insurance domain via copula. Insurance data from the automobile sector are selected for these purposes that are asymmetric in nature. We consider the application of vine copulas (in two dimensions) for several types of insurance data which are asymmetric in nature by utilizing the Vine Copula package in R. It appears that the resultant most appropriate bivariate copulas are members of the *C* and *D*-vine copulas, and among them, some are Archimedean as well. A vine copula is a copula constructed from a set of  $\frac{d(d-1)}{2}$  bivariate copulas by using successive mixing according to a tree structure on finite indexes  $1, \dots, d$ . Depending on the types of trees, various vine copulas can be constructed. The remainder of the paper is organized as follows. In Section 2, we discuss some basic definitions and useful preliminaries on copula theory. In Section 3, we discuss in details two different datasets, subsequently fitting an appropriate bivariate copula to each of them. In Section 4, we discuss some useful structural properties of these copulas, in particular tail dependence structures that are pertinent in the study of insurance claims dependence structure. Finally, some concluding remarks are made in Section 5.

## 2. Bivariate Copula and Its Properties

We begin this section by reviewing some basic definitions and concepts related to copula. The utility of Sklar’s theorem is that the modeling of the marginal distributions can be conveniently and efficiently separated from the dependence modeling in terms of the copula. Interestingly, the major task that lies in practical applications is how to identify this copula. For the bivariate case, a rich collection of copula families is available and well-investigated (see, for details, Joe 1997; Nelsen 2006). Sklar’s theorem establishes the link between multivariate distribution functions and their univariate margins. We state



this theorem at first. Let  $F$  be the  $p$ -dimensional distribution function of the random vector  $\underline{X} = (X_1, \dots, X_p)^T$  with marginals  $F_1, \dots, F_p$ . Then, there exists a copula  $C$  such that for all  $\underline{x} = (x_1, \dots, x_p)^T \in [-\infty, \infty]^p$ ,

$$F(\underline{x}) = C(F_1(x_1), \dots, F_p(x_p)). \tag{1}$$

Note that  $C$  is unique if  $F_1, \dots, F_p$  are continuous. Conversely, if  $C$  is a copula and  $F_1, \dots, F_p$  are distribution functions, then the function  $F$  defined by (1) is a joint distribution function with marginals  $F_1, \dots, F_p$ . Precisely,  $C$  can be interpreted as the distribution function of a  $p$ -dimensional random variable on  $[0, 1]^p$  with uniform marginals. Associated densities are denoted by a lower case  $c$ . In addition, the random variables  $X_1, \dots, X_p$  are assumed to be continuous in the following. By setting  $p = 2$ , one may easily obtain a bivariate version of the Sklar’s theorem as a special case.

We now provide some basic properties of a copula. For details on this, see Nelsen (1999, 2006).

**Definition 1.** A copula is a function  $C$  whose domain is the entire unit square with the following properties:

1.  $C(u, 0) = C(0, v) = 0$ , for all  $(u, v) \in [0, 1]$ .
2.  $C(u, 1) = C(1, u) = u$ , for all  $(u, v) \in [0, 1]$ .
3.  $C(u_1, v_1) - C(u_1, v_2) - C(u_2, v_1) + C(u_2, v_2) \geq 0$ , for all  $(u_1, v_1, u_2, v_2) \in [0, 1]$ . for every  $u_1 \leq u_2, v_1 \leq v_2$ .

Sklar (1973) established that any bivariate distribution function, say,  $F_{XY}(x, y)$ , can be represented as a function of its marginals, say,  $F_X(x)$  and  $F_Y(y)$ , by using a two-dimensional copula  $C(.,.)$  in the following way:

$$F_{XY}(x, y) = C(F_X(x), F_Y(y)).$$

If  $F_X(x)$  and  $F_Y(y)$  are absolutely continuous, then the associated copula  $C$  is unique. Moreover,  $C(u, v)$  is ordinarily invariant, which implies that if  $\delta(x)$  and  $\Phi(y)$  are strictly increasing functions, the copula of  $(\delta(X), \Phi(Y))$  is also that of  $(X, Y)$ . Therefore, both the marginals of  $F_{XY}(x, y)$  are absolutely continuous. Then, by selection of  $\delta(x) = F_X(x)$  and  $\Phi(y) = F_Y(y)$ , we can say that every copula is a distribution function whose marginals are uniform on the interval  $[0, 1]$ . Consequently, it represents the dependence structure between two variables by eliminating the influence of the marginals, and hence of any monotone transformation on the marginals.

### Dependence Structures

Copulas are instrumental in understanding the dependence between random variables. With them, we can separate the underlying dependence from the marginal distributions. It is well-known that a copula which characterizes dependence is invariant under strictly monotone transformations; subsequently, a better global measure of dependence would also be invariant under such transformations. Among other dependence measures, Kendall’s  $\tau$  and Spearman’s  $\rho$  are invariant under strictly increasing transformations, and, as we see in the next, they can be expressed in terms of the associated copula.

- **Kendall’s  $\tau$ :** Kendall’s  $\tau$  measures the amount of concordance present in a bivariate distribution. Suppose that  $(X, Y)$  and  $(\tilde{X}, \tilde{Y})$  are two pairs of random variables from a joint distribution function. We say that these pairs are concordant if large values of one tend to be associated with large values of the other and small values of one tend to be associated with small values of the other. The pairs are called discordant if large goes with small or vice versa. Algebraically, we have concordant pairs if  $(X - \tilde{X})(Y - \tilde{Y}) > 0$  and discordant pairs if we reverse the inequality. The formal definition is:

$$\tau(X, Y) = P\{((X - \tilde{X})(Y - \tilde{Y}) > 0)\} - P\{((X - \tilde{X})(Y - \tilde{Y}) < 0)\},$$

where  $(\tilde{X}, \tilde{Y})$  is an independent copy of  $(X, Y)$ . Let  $X$  and  $Y$  be continuous random variables with copula  $C$ . Then, Kendall's  $\tau$  is given by

$$\tau(X, Y) = 4 \iint_{[0,1]^2} C(u, v) dC(u, v) - 1. \tag{2}$$

- **Spearman's  $\rho$ :** Let  $X$  and  $Y$  be continuous random variables with copula  $C$ . Then, Spearman's  $\rho_s$  is given by

$$\rho_s = 12 \iint_{[0,1]^2} C(u, v) dudv - 3. \tag{3}$$

Alternatively,  $\rho_s$  can be written as  $\rho_s = 12 \int_0^1 \int_0^1 [C(u, v) - uv] dudv$ . Moreover, as mentioned earlier, one can equivalently show that  $\rho_s(U, V) = \rho(F_1(X), F_2(Y))$ .

- **Tail dependence property:** Let  $X$  and  $Y$  be two continuous r.v.'s with  $X \sim F$ , and  $Y \sim G$ . The upper-tail dependence coefficient (parameter)  $\lambda_U$  is the limit (if it exists) of the conditional probability that  $Y$  is greater than the  $100\alpha$  th percentile of  $G$  given that  $X$  is greater than the  $100\alpha$  th percentile of  $F$  as  $\alpha$  approaches 1.

$$\lambda_U = \lim_{\alpha \uparrow 1} P\left(Y > G^{-1}(\alpha) | X > F^{-1}(\alpha)\right). \tag{4}$$

If  $\lambda_U > 0$ , then  $X$  and  $Y$  are upper-tail dependent and asymptotically independent otherwise. Similarly, the lower-tail dependence coefficient is defined as

$$\lambda_L = \lim_{\alpha \downarrow 0} P\left(Y \leq G^{-1}(\alpha) | X \leq F^{-1}(\alpha)\right). \tag{5}$$

Let,  $C$  be the copula of  $X$  and  $Y$ . Then, equivalently, we can write

$\lambda_U = \lim_{u \downarrow 0} \frac{\tilde{C}(u, u)}{u}$ , and  $\lambda_L = \lim_{u \downarrow 0} \frac{C(u, u)}{u}$  where  $\tilde{C}(u, u)$  is the corresponding joint survival function given by

$$\tilde{C}(u, u) = 1 - 2u + C(u, u).$$

- **Blomqvist's  $\beta$ :** Suppose that  $\tilde{X}_n$  and  $\tilde{Y}_n$  are the medians of the samples  $X_1, \dots, X_n$  and  $Y_1, \dots, Y_n$ , respectively. In order to summarize information about the dependence between  $X$  and  $Y$ , Blomqvist (1950) suggested dividing the  $x - y$  plane into four regions by drawing the lines  $x = \tilde{X}_n$  and  $y = \tilde{Y}_n$  and comparing the following quantities:
  - $n_1$  : the number of points lying in either the lower left quadrant or the upper right quadrant;
  - $n_2$  : the number of points in either the upper left quadrant or the lower right quadrant.

Consequently, the definition of  $\beta_n$ , which is equivalently called Blomqvist's beta, is given by

$$\beta_n = \frac{n_1 - n_2}{n_1 + n_2} = -1 + 2 \frac{n_1}{n_1 + n_2}.$$

If  $n$  is even, then no sample point falls on either of the lines  $x = \tilde{X}_n$  and  $y = \tilde{Y}_n$ , and it follows that both  $n_1$  and  $n_2$  are even. If  $n$  is odd, however, then either one or two sample points lie on the lines defined by the sample medians. In the case of a single point lying on a median, Blomqvist (1950) proposed not to count the point altogether. In the latter case, one point has to fall on each line: one of them is assigned to the quadrant touched by the two points, and the other is not counted. This allows both  $n_1$  and  $n_2$  to remain even. The population analogue of  $\beta_n$  is

$$\beta = P[(X - \tilde{x})(Y - \tilde{y}) > 0] - P[(X - \tilde{x})(Y - \tilde{y}) < 0],$$

where  $\tilde{x}$  and  $\tilde{y}$  denote the population medians of  $X$  and  $Y$ , respectively. Next, on using the facts that

–

$$P[(X - \tilde{x})(Y - \tilde{y}) > 0] = P[(X - \tilde{x}) > 0, (Y - \tilde{y}) > 0] + P[(X - \tilde{x}) < 0, (Y - \tilde{y}) < 0];$$

and  $P[X > \tilde{x}, Y > \tilde{y}] = Pr[X < \tilde{x}, Y < \tilde{y}]$ ;

– From the fundamental Sklar’s (1959) theorem  $H(x, y) = C(F(x), G(y))$ ; one can write

$$\beta = 4C\left(\frac{1}{2}, \frac{1}{2}\right) - 1. \tag{6}$$

As  $\beta$  is only a function of  $C$ , it is possible to write it in terms of  $\underline{\alpha}$  whenever  $C \in C_{\underline{\alpha}}$ , where  $\underline{\alpha}$  is the set of parameters associated with the copula  $C$ .

- **Left-Tail decreasing property and Right-Tail increasing property:** Nelsen (1999) showed that  $X(Y)$  is left-tail decreasing i.e.,  $LTD(Y|X)$  and  $LTD(X|Y)$  if and only if for all  $u, u', v, v'$  such that  $0 < u \leq u' \leq 1$  and  $0 < v \leq v' \leq 1$ , if  $\frac{C(u,v)}{uv} \geq \frac{C(u',v')}{u'v'}$ . Again, from Nelsen (2006), Theorem 5.2.5,  $X(Y)$  is right-tail increasing if

- $RTI(Y|X)$  if and only if for any  $v \in (0, 1)$   $\frac{1-u-v+C(u,v)}{1-u}$  is nondecreasing in  $u$ .
- $RTI(Y|X)$  if and only if for any  $u \in (0, 1)$   $\frac{1-u-v+C(u,v)}{1-v}$  is nondecreasing in  $v$ .

For an alternative criteria see (Nelsen 2006, p. 197, Theorem 5.2.12 and Corollary 5.2.11). Moreover, regarding stochastically increasing, left-tail decreasing and right-tail increasing properties, we provide the following equivalent conditions (see, Nelsen 2006, p. 197, Corollary 5.2.11 and Theorem 5.2.12), which are utilized later on in determining the dependence structure for the best fitted bivariate copula:

In the next, we discuss the stochastic increasing (SI) property for a copula beginning with the definition given in the following result.

**Result 1.** Let  $X$  and  $Y$  be continuous random variables with a copula  $C$ . Then

- $SI(Y|X)$  if and only if for any  $v \in [0, 1]$ ,  $C(u, v)$  is a concave function of  $u$ ;
- $SI(X|Y)$  if and only if for any  $u \in [0, 1]$ ,  $C(u, v)$  is a concave function of  $v$ .

**Result 2.** Let  $X$  and  $Y$  be continuous random variables with a copula  $C$ . Then:

- $SI(Y|X)$ , then  $LTD(Y|X)$  and  $RTI(Y|X)$ ,
- $SI(X|Y)$ , then  $LTD(X|Y)$  and  $RTI(X|Y)$ .

Regarding the LTD (RTI) property, they are also discussed in Section 4 In the next section, we briefly discuss the types of insurance data selected for our study and associated goodness of fit based on a best-fitted bivariate copula for each of the scenarios considered in this paper.

### 3. Application to Insurance Data

#### 3.1. Data and Variable Selection

We particularly focus on insurance claim data that are related to auto/motor accidents. The reason for selecting this specific domain is already established in the introduction. All of the datasets referred to in this paper can be found in the Computational Actuarial Science collection and are accessible through the “CASdatasets” package in R. Additionally, we used the “VineCopula” package to find the best-fitted copula model for each pair of variables in each dataset used. The “VineCopula” package takes the selected variables and finds the best copula model from the families available in the package. This choice of copula is based on test diagnostics such as AIC, BIC, and the log-likelihood value. A generic R-code

based on the Vine Copula package which is used for selecting the best possible bivariate copula on the different insurance datasets is provided in the Appendix A. The next section details how the variables from each dataset were selected and the associated best-fitted bivariate copula models.

### 3.1.1. Dataset 1 (Australian Automobile Claim Data)

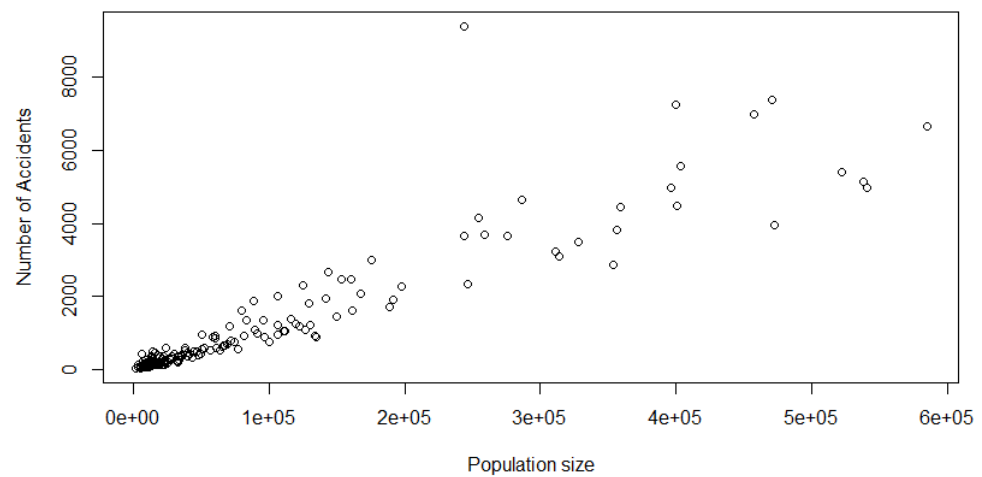
This dataset records the number of third-party claims in a 12 month period between 1984 and 1986 in each of the 176 local government areas in New South Wales, Australia. Additionally, the dataset includes the name of the local government, the number of third-party claims filed, the number of people killed or injured in automobile accidents, the population size, and the population density. Australia is historically known for its low population density. This is due to extreme climate of the continent. With this in mind, we decided to include the population size of each city in New South Wales as opposed to the population density because the density is skewed by the lack of inhabitants in Australia. For this dataset, we plan to study dependence measure among 4 different variables in pairwise comparison structure. We argue that the selection of these pairwise comparisons are legitimate in nature. The Table 1 below provides a key for the abbreviations we use for each variable throughout our study.

**Table 1.** Variable description.

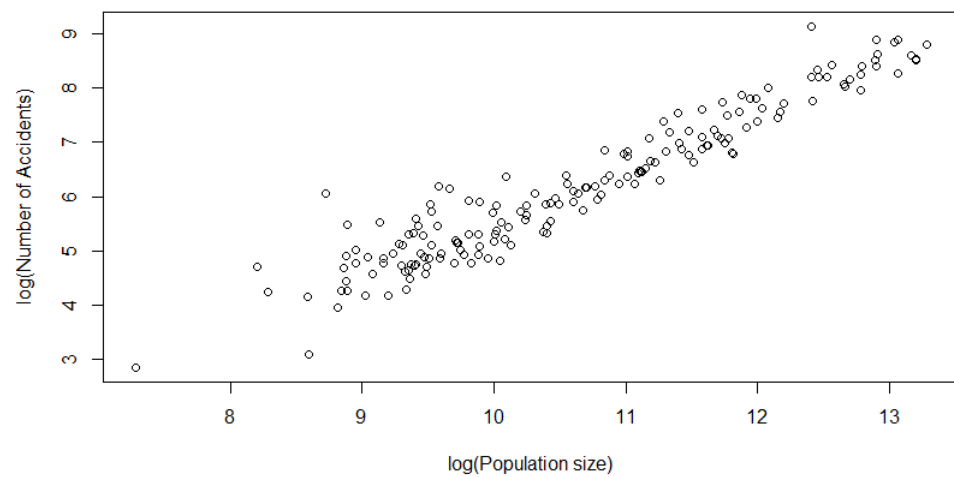
Abbreviation	Variable
ACC	Number of accidents
TPC	Number of third-party claims filed
K/I	Number of people killed or injured in an accident
Pop	Population of the area

Furthermore, in each of these bivariate modeling setups, we provide scatterplots (on actual values as well as on a logarithmic scale) to have an initial glimpse of their dependence structure. The scatterplot based on a log transformation of the original variable is due to the fact that in visualizing numerical data which ranges over several magnitudes, conventional wisdom says that a log transformation of the data can often result in a better visualization. As such, the scatterplots in logarithmic scale are also provided to see the skewness of the original data values. Next, we provide each pairwise model description to be considered in our analysis.

**Model 1 (AUS 1):** The first pair of variables that were selected were the number of accidents (ACC) and the population size (POP). These were chosen because we expected more accidents to occur in regions with a higher population relatively speaking. From the scatterplot in Figures 1 and 2, it appears that there exists a strong linear relationship between these two variables, which is also supported by the associated Kendall’s  $\tau$  and Spearman’s  $\rho$  values in Table 2 (Column 4, 5, Row 1).



**Figure 1.** Scatterplot of the Model 1 data.



**Figure 2.** Scatterplot of the Model 1 data (on a natural log scale).

**Model 2 (AUS 2):** In this model, we consider the two concomitant variables under study, namely the number of Third-Party Claims filed (TPC) and POP. TPC happens if a driver’s negligence results in the injury or death of another driver; the affected party or their family have the ability to file a claim against the guilty driver’s insurance company. We consider studying the dependence between TPC and the population of a given region in Australia because one would expect a larger volume of third-party claims to be filed in regions with higher populations. Needless to say, this is a good source of information for car insurance providers. From the scatterplot in Figures 3 and 4, it appears that there exists a strong linear relationship between these two variables, which is also supported by the associated Kendall’s  $\tau$  and Spearman’s  $\rho$  values in Table 2 (Column 4, 5, Row 2).

**Model 3 (AUS 3):** Next, we consider studying the dependence of a region’s population (POP) and the number of people killed or injured in an accident (K/I). Once we discovered that there was a strong dependence relationship between the number of third-party claims and the population size of a region, we realized that since third-party claims are a result of accidents with injuries involved, the number of people killed or injured could be greater in higher-populated areas where more third-party claims are filed. From the scatterplot in Figures 5 and 6, it appears that there exists a strong linear relationship between these two variables, which is also supported by the associated Kendall’s  $\tau$  and Spearman’s  $\rho$  values in Table 2 (Column 4, 5, Row 3).

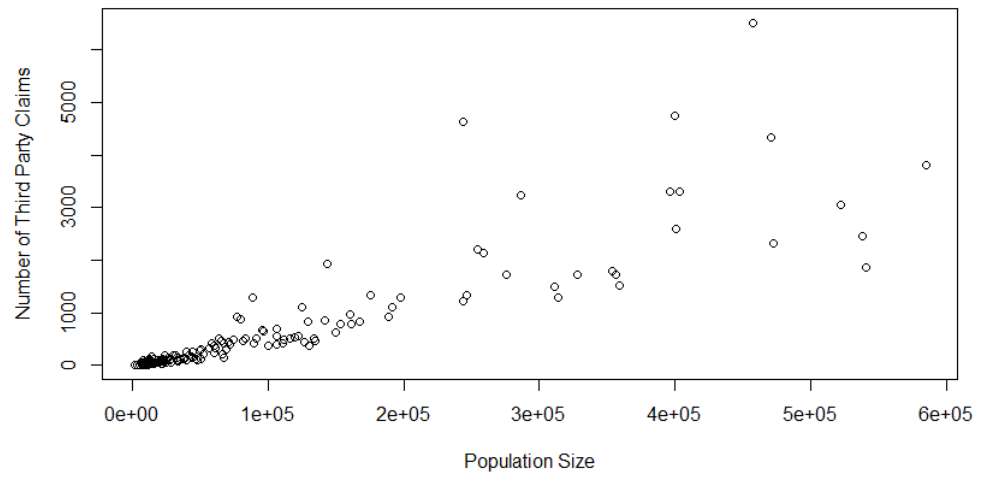


Figure 3. Scatterplot of the Model 2 data.

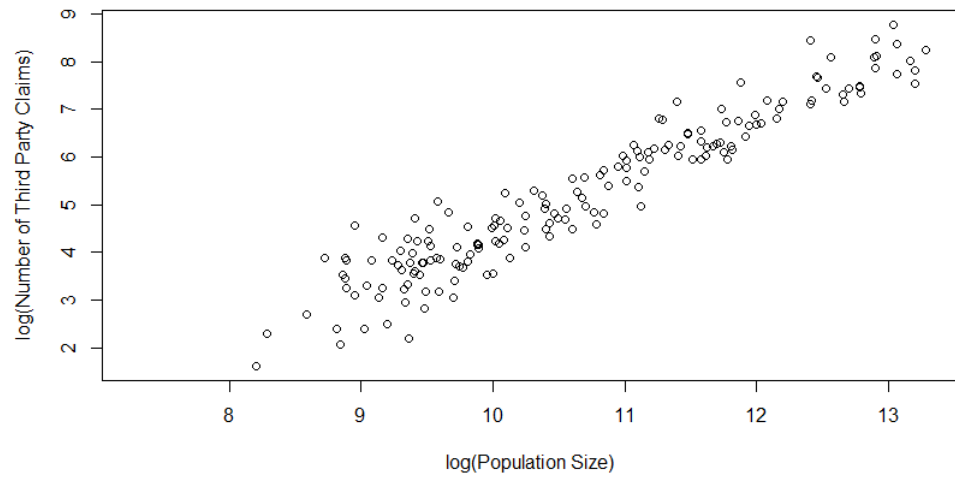


Figure 4. Scatterplot of the Model 2 data (on a natural log scale).

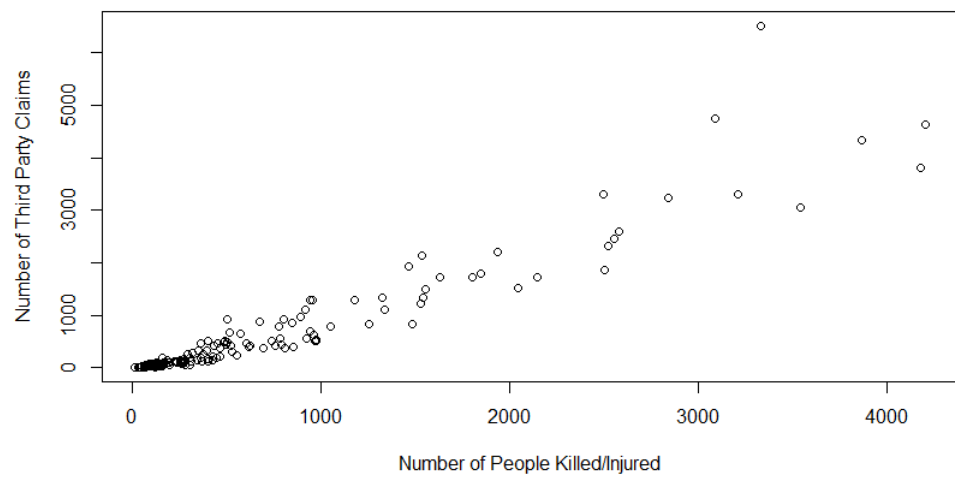
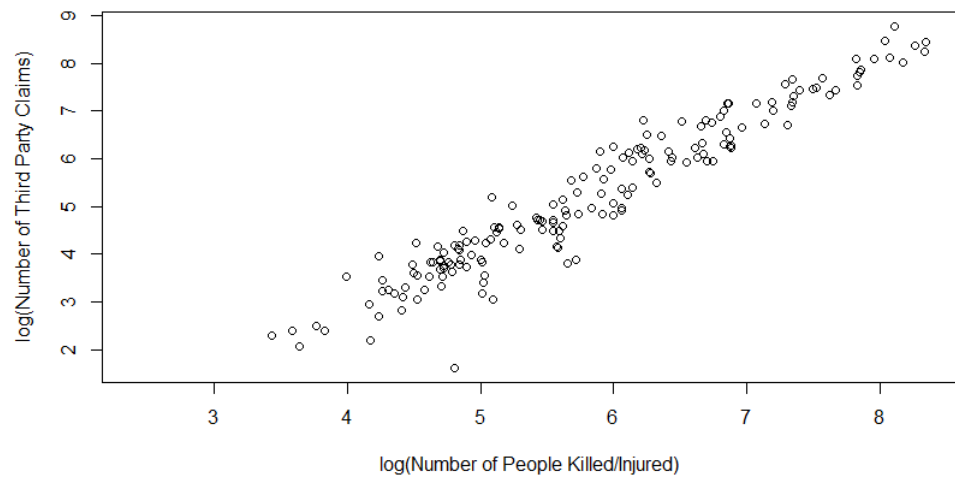


Figure 5. Scatterplot of the Model 4 data.

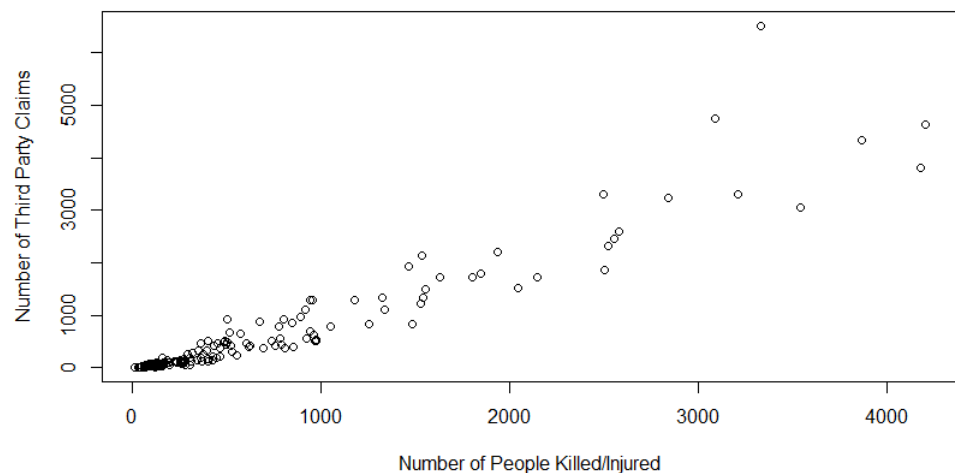


**Figure 6.** Scatterplot of the Model 4 data (on a natural log scale).

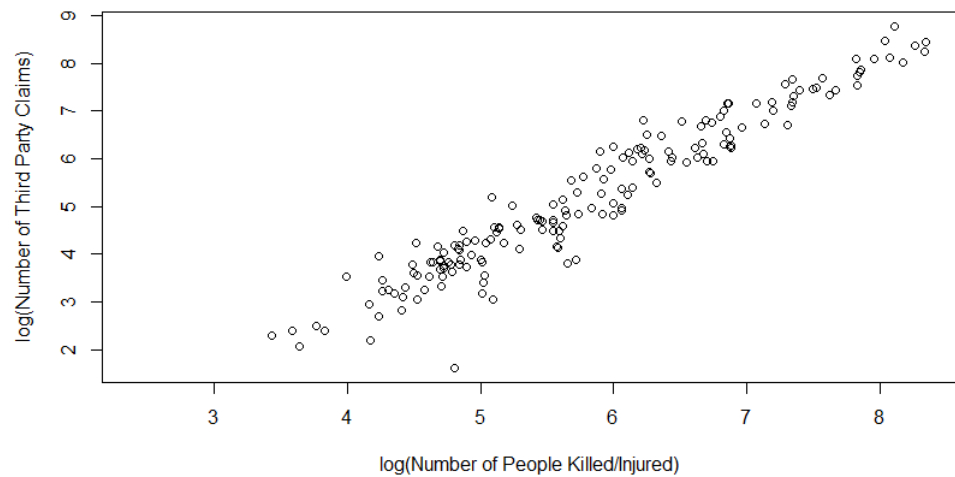
**Model 4 (AUS 4):** In this model, we consider studying the level of dependence between the number of people injured or killed in an automobile accident (K/I) and the corresponding number of third-party claims filed (TPC). As defined above in Model 2, third-party claims are filed in the event of an accident in which other drivers suffer injury from the negligence of another. While injury and death are not exclusive to the third party, we found a positive trend in the scatterplot of these two variables (Figures 7 and 8). Hence, we chose to fit a copula to these two concomitant variables.

The non-exhaustive reasons for selecting four different models are as follows:

- For multicomponent insurance claim data, instead of a single representative value for the tail dependence measure, which would not reveal the partial dependence structure(s), it is better to observe the tail dependence structure pairwise. This way, one can eliminate to some extent the effect of lurking variable(s).
- Pairwise dependence measures help to identify (possibly one way or the other) which of the two components would be the most important to influence the associated portfolio risk.
- A class of bivariate copulas can be listed adequately for dealing with such types asymmetric insurance data, for example, where a specific class of extreme-value copulas or Archimedean copulas could be useful.



**Figure 7.** Scatterplot of the Model 4 data.



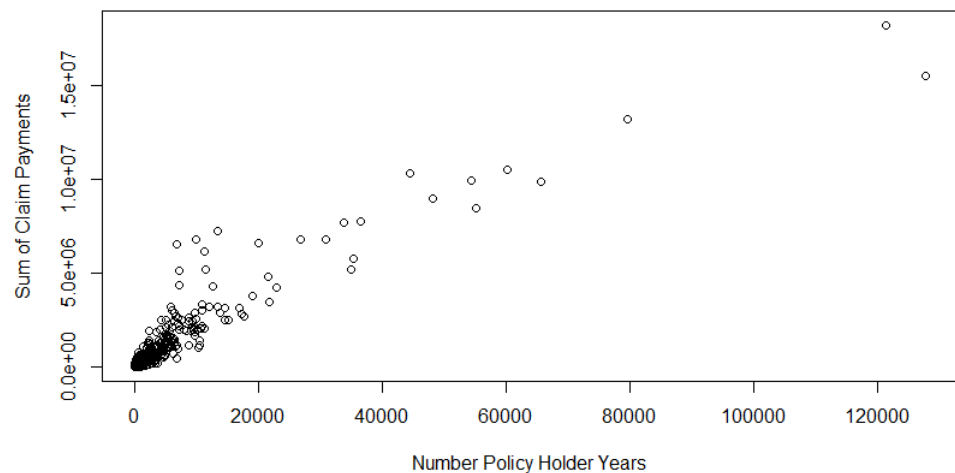
**Figure 8.** Scatterplot of the Model 4 data (on a natural log scale).

### 3.1.2. Dataset 2 (Swedish Motor Insurance Data)

This dataset represents the insurance information of 2182 motorists collected by the Swedish Committee on the Analysis of Risk Premium in 1977. It consists of the number of kilometers driven by a motorist (grouped into 5 categories), the geographical zone of a vehicle (grouped into 7 categories), the bonus variable (grouped into 7 categories), the make of the vehicle, the number of years that a motorist has been insured, the number of claims a motorist has filed, and the sum of the payments made by a motorist. We excluded the geographic zone and make of the vehicle variables from our consideration because while they are quantitatively defined, they describe qualitative variables and do not have a defined ordering. Due to the way the kilometers' variable was defined, we were unable to come up with a model that showed a large amount of dependence, so the results of that model are excluded from this paper. Instead, we chose to study the dependence and subsequently search for the best possible bivariate copula model with the following variables of interest:

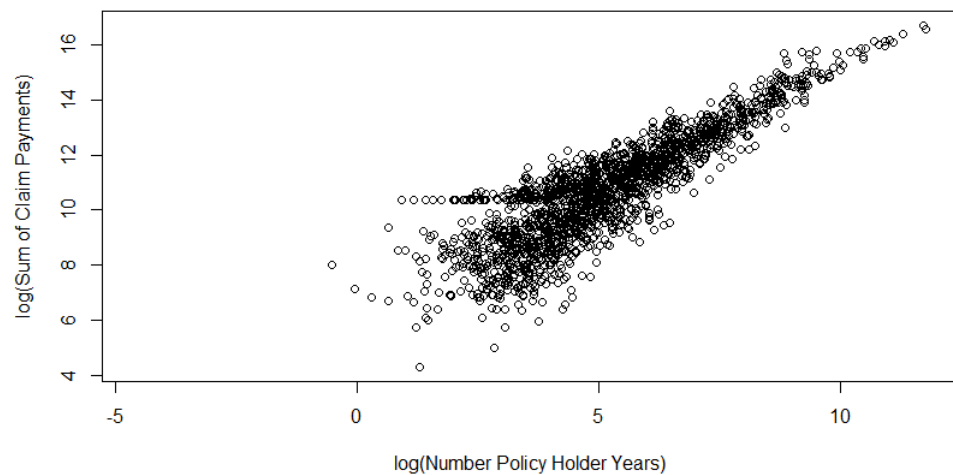
1.  $X_1$ : Insured (number of years a motorist has been insured).
2.  $X_2$ : Claims (sum of claim payments).

The scatterplots in Figures 9 and 10 show a linear relationship between the variables.



**Figure 9.** Scatterplot of the Swedish motor data.





**Figure 10.** Scatterplot of the Swedish motor data (on a natural log scale).

The Tables 2 and 3 below detail the results for each model. Note that all of these computations were performed in R.

**Table 2.** Level of dependence between model variables.

Dataset/Model	$X_1$	$X_2$	Kendall's $\tau$	Spearman's $\rho$
AUS 1	ACC	Population	0.8123	0.9452
AUS 2	TPC	Population	0.8078	0.9479
AUS 3	K/I	Population	0.7981	0.9373
AUS 4	K/I	TPC	0.8372	0.9611
Swedish Motor	Policy Holder Years	Sum of Payments	0.7411	0.9030

**Table 3.** Model diagnostics and goodness of fit statistics.

Dataset/Model	Best-Fitted Copula	Parameter Estimates	AIC	BIC	Log Likelihood
AUS 1/Model 1	Frank	(18.42)	−377.38	−374.21	189.69
AUS 2/Model 2	Frank	(18.33)	−376.58	−373.41	189.29
AUS 3/Model 3	Tawn 1	(5.01, 0.95)	−373.11	−366.76	188.55
AUS 4/Model 4	Student t	(0.96, 4.61)	−442.33	−435.99	223.16
Swedish Motor	BB6	(1.59, 2.81)	−4095.96	−4084.58	2049.98

Table 2 outlines the level of concordance between each pair of variables in each model. When two variables are concordant, this means that higher values of one variable are associated with higher values of the other and vice versa for lower values. If these coefficients are closer to 0, this indicates low dependence or even independence. Conversely, if these coefficients are closer to 1, it tells us that the variables are dependent upon one another. From Table 2, we see that each pair of variables exhibits a strong level of dependence, since the concordance coefficients are close to 1. Table 3 represents various model diagnostics along with parameter estimates corresponding to the best-fitted bivariate copula. We expect

the AIC and BIC to be minimal and the log likelihood to be maximal. Each copula shown in Table 3 represents the best fit for the pair of variables that were being tested according to the AIC, BIC, and log-likelihood criteria. The c.d.f. and p.d.f. plots corresponding to the best-fitted bivariate copulas listed in Table 3 are also provided in the Figures 11–14.

**Gaussian CDF (0.8)**

**Gaussian PDF (0.8)**

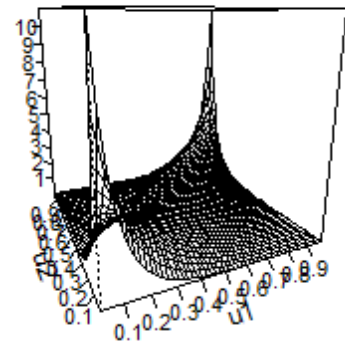
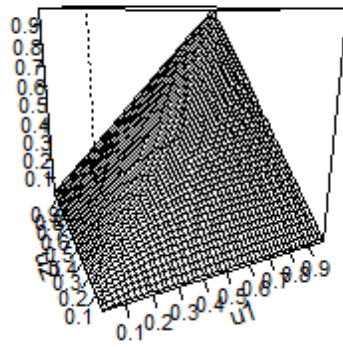


Figure 11. Gaussian(0.8) c.d.f. and p.d.f.

**Frank (18.42) CDF**

**Frank (18.42) PDF**

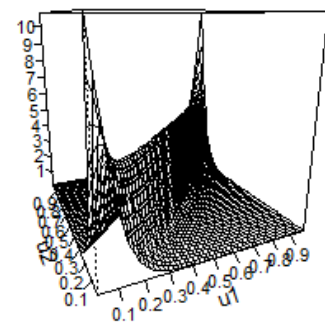
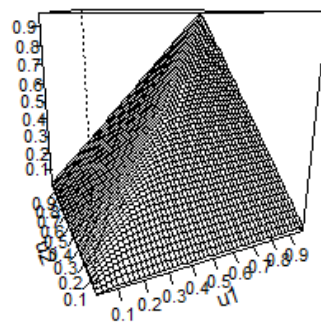


Figure 12. Frank c.d.f. and p.d.f. with  $\alpha = 18.42$ .

**Frank (18.33) CDF**

**Frank (18.33) PDF**

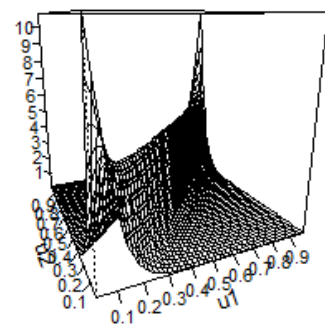
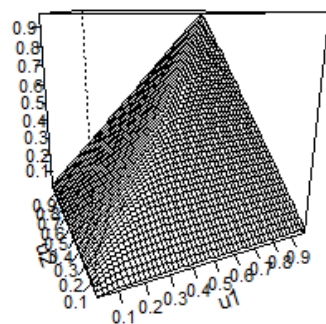


Figure 13. Frank c.d.f. and p.d.f. with  $\alpha = 18.33$ .

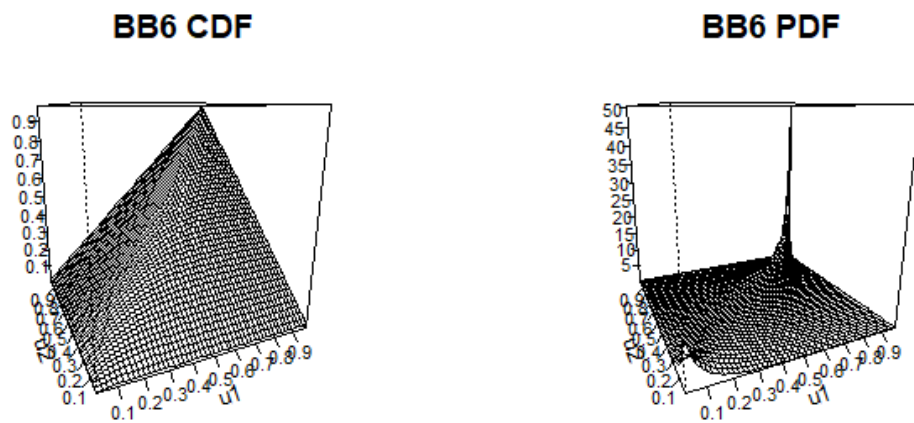


Figure 14. BB6 c.d.f. and p.d.f. with  $\theta = 1.59$  and  $\delta = 2.81$ .

#### 4. Structural Properties of the Fitted Bivariate Copula

This section presents the analysis of certain structural properties of the copulas. We begin our discussion with the Tawn type-1 copula.

##### 4.1. Tawn Type-1 Copula

We can say the following regarding Tawn type-1 copula that was found to be the best-fitted bivariate copula for Model 3 from the dataset 1:

- The Tawn copula is a nonexchangeable extension of the Gumbel copula with three parameters (also known as the asymmetric logistic copula).
- Tawn copula’s definition is based around so-called Pickands dependence functions, see Franc et al. (2011) for pertinent details. Equation (4) in Franc et al. (2011) presents the way one can compute the density in the probability space using a Pickands function  $M$ :

$$C(u, v) = (u, v)^{A(w)},$$

with  $w = \frac{\log(u)}{\log(uv)}$ .

- The Tawn copula’s Pickand function can be written as

$$M(t) = (1 - \psi_2)(1 - t) + (1 - \psi_1)t + \left[ (\psi_1(1 - t))^\theta + (\psi_2)^\theta \right]^{1/\theta},$$

with  $t \in [0, 1]$ ,  $0 \leq \psi_1, \psi_2 \leq 1$ , and  $\theta \in [1, \infty)$ . The Tawn copula is in actuality a Gumbel copula with two additional asymmetry parameters:  $\psi_1$  and  $\psi_2$ . If we set  $\psi_1$  and  $\psi_2$  equal to unity, the Gumbel copula is obtained. In the *VineCopula* package in R, the Tawn type-1 copula refers to  $\psi_1 = 1$ .

- For this copula, the lower-tail dependent  $\lambda_L = 0$ . The corresponding upper-tail dependent  $\lambda_U = 0.8288$ , for the AUS3/Model 3 data for which Tawn type-1 copula appeared to be the best fit.

##### 4.2. Frank Copula

The Frank copula (see, Nadarajah et al. 2017) has the following form:

$$C(u, v) = \log_\alpha \left[ 1 + \frac{(\alpha^u - 1)(\alpha^v - 1)}{\alpha - 1} \right],$$

for  $\alpha > 0$ . In this case, the positive dependence corresponds to  $0 < \alpha < 1$ , independence corresponds to  $\alpha \rightarrow 1$ , and negative dependence corresponds to  $\alpha > 1$ . Next, for this bivariate copula, we can write the following:

- The associated dual of the copula is denoted by  $\tilde{C}(u, v)$  and is given by

$$\tilde{C}(u, v) = u + v - C(u, v) = u + v - \log_{\alpha} \left[ 1 + \frac{(\alpha^u - 1)(\alpha^v - 1)}{\alpha - 1} \right],$$

- Again, the associated co-copula denoted by  $C^*(u, v)$  and is given by

$$C^*(u, v) = 1 - C(1 - u, 1 - v) = 1 - \log_{\alpha} \left[ 1 + \frac{(\alpha^{1-u} - 1)(\alpha^{1-v} - 1)}{\alpha - 1} \right].$$

For the bivariate Frank copula, we have the following:

- The Kendall's  $\tau$  will be  $\tau = 4 \left[ \frac{3 \log(\alpha)(-2 \log(1-\alpha) + \log(\alpha) + 2) - 6 \text{Li}_2(\alpha) + \pi^2}{6 \log^2(\alpha)} \right] - 1$ , where  $\text{Li}_2(\alpha)$  is the PolyLog function (on using Mathematica).
- The Spearman's  $\rho$  will be

$$\rho = \int_0^1 \frac{v((1-\alpha) \log(\alpha-1)\alpha^v + (\alpha-1)\alpha^v \log((\alpha-1)\alpha^v) - \alpha \log(\alpha)(\alpha^v-1))}{\log(\alpha)(\alpha^v-1)(\alpha^v-\alpha)} dv - 3.$$

- The Blomqvist's  $\beta$  expression will be

$$\beta = 4 \log_{\alpha} \left[ 1 + \frac{(\alpha^{\frac{1}{2}} - 1)^2}{\alpha - 1} \right].$$

**Tail dependence property of the bivariate Frank copula:**

- For the upper-tail dependence (using L'Hôpital's rule)

$$\begin{aligned} \lambda_U &= \lim_{u \uparrow 1} \frac{1 - 2u + \log_{\alpha} \left[ 1 + \frac{(\alpha^u - 1)^2}{\alpha - 1} \right]}{1 - u} \\ &\stackrel{H}{=} - \lim_{u \uparrow 1} \left( -2 + \left( \frac{1}{1 + \frac{(\alpha^u - 1)^2}{\alpha - 1}} \right) \left( \frac{2(\alpha^u - 1)(\alpha^u \log \alpha)}{(\alpha - 1) \log \alpha} \right) \right) \\ &= 0. \end{aligned}$$

Therefore, Frank's copula is upper-tail dependent. Next, we determine if it is lower-tail dependent. Consider the limit

$$\begin{aligned} \lambda_L &= \lim_{u \downarrow 0} \frac{\log_{\alpha} \left[ 1 + \frac{(\alpha^u - 1)^2}{\alpha - 1} \right]}{u} \\ &\stackrel{H}{=} \lim_{u \downarrow 0} \frac{2(\alpha^u - 1)(\alpha^u)(\log \alpha)^2}{1 + \frac{(\alpha^u - 1)^2}{\alpha - 1}} \\ &= 0, \end{aligned}$$

again, on using L'Hôpital's rule. Consequently, the Frank copula is also lower-tail independent. Therefore, the bivariate Frank copula is asymptotically independent.

**LTD and RTI property of the bivariate Frank copula:**

Consider the following:

$$\frac{\partial^2}{\partial u^2} C_{\alpha}(u, v) = - \frac{\log(\alpha) \alpha^u (\alpha^v - 1) (\alpha^v - \alpha)}{(\alpha - \alpha^u + \alpha^{u+v} - \alpha^v)^2}.$$

Therefore,  $\frac{\partial^2}{\partial u^2} C_{\alpha}(u, v) \leq 0$  for  $0 < \alpha < 1$ ; thus,  $C_{\alpha}(u, v)$  is a concave function of  $u$  for  $0 < \alpha < 1$ . It follows that if  $X$  and  $Y$  are continuous with the Frank family copula, then  $SI(Y|X)$  (and by symmetry  $SI(X|Y)$  as well). Again, from Theorem 5.2.12 (Nelsen (2006)) this implies the associated BB8 family copula also holds the LTD and

RTI property, i.e.,  $LTD(Y|X)$  and  $RTI(Y|X)$ , and because of symmetry,  $LTD(X|Y)$  and  $RTI(X|Y)$ .

Furthermore, we see that both models are highly correlated in the center of their respective distributions. The Table 4 below summarizes the dependence structures discussed above and displays the generator function of this particular copula:

**Table 4.** Dependence of the Frank copula.

Generator Function	$\phi(t) = -\log\left(\frac{\exp(-\alpha t)-1}{\exp(-\alpha)-1}\right)$
Blomqvist $\beta$ (AUS 1)	0.9348
Blomqvist $\beta$ (AUS 2)	0.9329
Upper-Tail Dependence	0
Lower-Tail Dependence	0
Kendall's $\tau$	given earlier

### 4.3. Bivariate $t$ Copula

The  $t$  copula (see Embrechts et al. (2001) or Fang and Fang (2002) and the references cited therein) can be thought of as representing the dependence structure implicit in a multivariate  $t$  distribution. The two-dimensional unique  $t$  copula associated with a bivariate random vector  $Y = (Y_1, Y_2)^T$ , is given by

$$C_{\delta}^t(u, v) = \int_{-\infty}^{t_{\delta}^{-1}(u)} \int_{-\infty}^{t_{\delta}^{-1}(v)} \frac{\Gamma((\delta + 2)/2)}{\Gamma(\delta/2)\sqrt{\{(\pi\delta)^2|\Sigma|\}}} \left[1 + \frac{y^T \Sigma^{-1} y}{\delta}\right]^{-\frac{\delta+2}{2}} dy_1 dy_2,$$

where  $t_{\delta}^{-1}(\cdot)$  denotes the quantile function of a standard univariate  $t_{\delta}(\cdot)$  distribution. Furthermore,  $\Sigma$  is the correlation matrix given by

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

where  $\rho$  is the correlation coefficient between  $Y_1$ , and  $Y_2$ . The determinant of this matrix, denoted by  $|\Sigma|$ , is given by  $|\Sigma| = 1 - \rho^2$ . Next, one may verify the following regarding the dependence structure for a bivariate  $t$  copula

- Kendall's  $\tau$  will be

$$\tau = \frac{2}{\pi} \arcsin \rho,$$

for the proof, see Fang and Fang (2002).

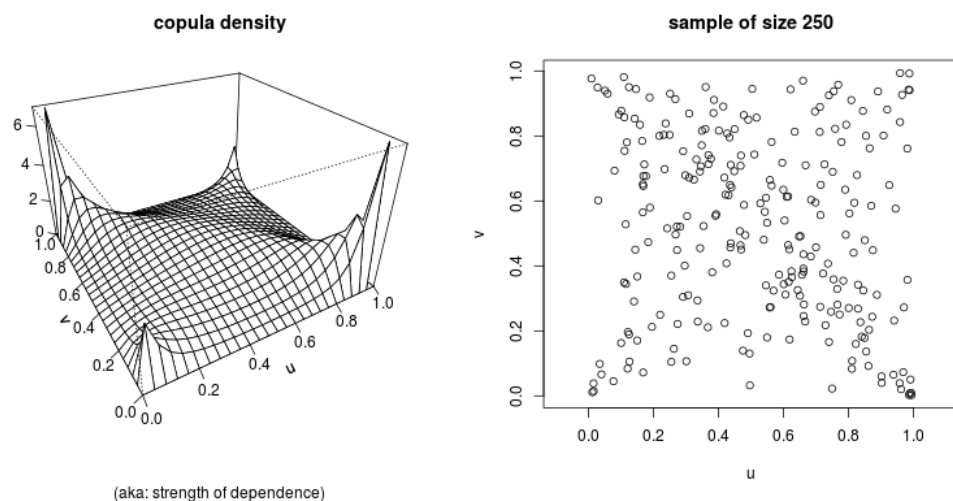
- The tail dependence coefficient (associated with a bivariate  $t$  copula as given earlier)  $\lambda$  is given by

$$\lambda = 2t_{\delta+1}\left(-\frac{\sqrt{\{\delta + 1\}}\sqrt{\{1 - \rho\}}}{\sqrt{\{1 + \rho\}}}\right),$$

where  $t_{\delta+1}$  is the univariate central student  $t$  distribution with  $(\delta + 1)$  degrees of freedom and  $\rho$  is the correlation coefficient. It is important to note that a student  $t$ -copula may exhibit both the positive and negative tail dependence, although the "overall" association is negative  $\rho < 0$ . Furthermore, a student  $t$ -copula with a large value of  $\delta$  tends to have a 0 tail dependence even though the correlation is 0. The  $t$ -copula can capture the asymptotic dependence even when the variables are negatively (inversely) associated (see, Embrechts et al. 2001). In the  $t$ -copula formula, as  $\delta$  increases, the tail dependence weakens, and thus, the probability of occurrence of extreme values reduces.

For illustrative purposes, we provide the following picture in Figure 15 (generated through the <https://copulatheque.shinyapps.io/copulas/> (accessed on 15 June 2022) created by BenGraeler) showing a student  $t$ -copula with  $\rho = -0.3$  and  $\delta = 1$ , which gives the value of Kendall's  $\tau = -0.19$  and upper- and lower-tail dependence of  $\lambda = 0.19$ .

### t-copula



#### Dependence properties

Kendall's tau: -0.19  
 lower and upper tail dependence: 0.19, 0.19

**Figure 15.** Student  $t$  – copula density with  $\delta = 2$  and with Kendall's  $\tau = -0.19$ .

The estimation of a student  $t$ -copula is quite difficult. Noticeably, the marginal tails (for bivariate and/or multivariate data distributions) of financial data are usually heavy tailed, and hence this should be fitted by a  $t$ -distribution and not by a Gaussian distribution. In addition, the dependence in joint extremes of bivariate and/or multivariate financial data suggests a dependence structure allowing for tail dependence. Consequently, the use of  $t$ -copulas has become popular for modeling dependencies in financial data. Some recent applications have been: analysis of nonlinear and asymmetric dependence in the German equity market Sun et al. (2008); estimation of large portfolio loss probabilities Chan and Kroese (2010); and risk modeling for future cash flow Pettere and Kollo (2011). See also Dakovic and Czado (2011).

One may subsequently obtain the expressions for the upper-tail as well as lower-tail dependence from the above. For details, see (Demarta and McNeil 2005, p. 4, Proposition 1).

#### 4.4. BB6 (Joe–Gumbel) Copula

The BB6 copula (see Joe 1997) has the following form:

$$C(u, v) = 1 - (1 - \exp(-[(-\log(1 - \bar{u}^\theta))^\delta + (-\log(1 - \bar{v}^\theta))^\delta]^\frac{1}{\delta}))^\frac{1}{\delta}, \quad u \geq 0, v \leq 1, \theta \geq 1, \delta \geq 1.$$

where  $\bar{u} = 1 - u$  and  $\bar{v} = 1 - v$ .

**Tail dependence property of the bivariate BB6 copula**

The lower-tail and upper-tail dependence coefficients can be calculated using the same methodology that we used for the Frank copula. For the upper-tail dependence coefficient, we obtain the following:

$$\lambda_U = \lim_{u \uparrow 1} \frac{1 - 2u + 1 - (1 - \exp(-[2(-\log(1 - \bar{u}^\theta))^\delta]))^{\frac{1}{\delta}}}{1 - u}$$

$$\stackrel{H}{=} \lim_{u \uparrow 1} 2 - 2^{\frac{1}{\delta}}(1 - u)^{\theta-1} \exp [2^{\frac{1}{\delta}} \log(1 - (1 - u)^\theta)](1 - \exp [2^{\frac{1}{\delta}} \log(1 - (1 - u)^\delta)])^{\frac{1}{\delta}-1} = 2 - 2^{\frac{1}{\delta}}.$$

Similarly, for the lower-tail dependence coefficient:

$$\lambda_L = \lim_{u \downarrow 0} \frac{1 - (1 - \exp(-[2(-\log(1 - \bar{u}^\theta))^\delta]))^{\frac{1}{\delta}}}{u}$$

$$\stackrel{H}{=} \lim_{u \downarrow 0} 2^{\frac{1}{\delta}}(1 - u)^{\theta-1} \exp [2^{\frac{1}{\delta}} \log(1 - (1 - u)^\theta)](1 - \exp [2^{\frac{1}{\delta}} \log(1 - (1 - u)^\delta)])^{\frac{1}{\delta}-1} = 0.$$

Therefore, the BB6 copula is not asymptotically independent. However, from the given expression for  $\lambda_U$ , it is quite clear that as both  $\theta$  and  $\delta$  are close to 1,  $\lambda_U$  is close to zero. This would imply that the BB6 copula is asymptotically independent in such a case. Furthermore, from the expression for  $\lambda_U$ , it appears that as the values of  $\delta$  and  $\theta$  increases, the value of  $\lambda_U$  increases. This implies the fact that this copula might not be that useful to model the dependence structure for financial data in general, as such tend to exhibit tail dependence, especially lower-tail dependence. However, if the data suggests that the estimated values of the parameters  $\delta$  and  $\theta$  are larger than one, then it may be utilized to model financial data, such as insurance data that exhibit some amount of tail dependence.

**LTD and RTI property of the bivariate Frank copula:**

Consider the following:

$$\frac{\partial^2}{\partial u^2} C_{\theta,\delta}(u, v)$$

$$= \left[ \left( (1 - u)^\theta - 1 \right)^2 \left( \exp \left( \left( \left( -\log(1 - (1 - u)^\theta) \right)^\delta + \left( -\log(1 - (1 - v)^\theta) \right)^\delta \right)^{\frac{1}{\delta}} \right) - 1 \right)^2 \right]^{-1}$$

$$\times \left[ A_1 \times A_2 \left\{ A_3 + (1 - u)^\theta \times (B_1 + B_2 + B_3) \right\} \right],$$

where

$$A_1 = (1 - u)^{\theta-2} \left( -\log(1 - (1 - u)^\theta) \right)^{\delta-2}$$

$$\times \left[ 1 - \exp \left( - \left( \left( -\log(1 - (1 - u)^\theta) \right)^\delta + \left( -\log(1 - (1 - v)^\theta) \right)^\delta \right)^{\frac{1}{\delta}} \right) \right]^{\frac{1}{\delta}},$$

$$A_2 = \left[ \left( -\log(1 - (1 - u)^\theta) \right)^\delta + \left( -\log(1 - (1 - v)^\theta) \right)^\delta \right]^{\frac{1}{\delta}-2},$$

$$A_3 = \left( \theta + (1 - u)^\theta - 1 \right) \log(1 - (1 - u)^\theta) \left( \left( -\log(1 - (1 - u)^\theta) \right)^\delta + \left( -\log(1 - (1 - v)^\theta) \right)^\delta \right)$$

$$\times \left[ \exp \left( \left( \left( -\log(1 - (1 - u)^\theta) \right)^\delta + \left( -\log(1 - (1 - v)^\theta) \right)^\delta \right)^{\frac{1}{\delta}} \right) - 1 \right],$$

$$B_1 = - \left( -\log(1 - (1 - u)^\theta) \right)^\delta \left( \left( -\log(1 - (1 - u)^\theta) \right)^\delta + \left( -\log(1 - (1 - v)^\theta) \right)^\delta \right)^{\frac{1}{\delta}},$$

$$\begin{aligned}
 B_2 &= \theta \left[ \left( -\log(1 - (1 - u)^\theta) \right)^\delta \left( \left( -\log(1 - (1 - u)^\theta) \right)^\delta + \left( -\log(1 - (1 - v)^\theta) \right)^\delta \right)^{\frac{1}{\delta}} \right. \\
 &\quad \left. - \delta \left( -\log(1 - (1 - v)^\theta) \right)^\delta + \left( -\log(1 - (1 - v)^\theta) \right)^\delta \right] \\
 &\quad \times \exp \left( \left( \left( -\log(1 - (1 - u)^\theta) \right)^\delta + \left( -\log(1 - (1 - v)^\theta) \right)^\delta \right)^{\frac{1}{\delta}} \right), \\
 B_3 &= (\delta - 1)\theta \left( -\log(1 - (1 - v)^\theta) \right)^\delta.
 \end{aligned}$$

Therefore,  $\frac{\partial^2}{\partial u^2} C_{\theta,\delta}(u, v) \leq 0$  for  $\theta > 1$  and for any  $\delta \geq 1$ ; thus,  $C_{\theta,\delta}(u, v)$  is a concave function of  $u$  for  $\theta > 1$  and for any  $\delta \geq 1$ . It follows that if  $X$  and  $Y$  are continuous with the BB6 family copula, then  $SI(Y|X)$  (and by symmetry  $SI(X|Y)$  as well). Again, from Theorem 5.2.12 (Nelsen 2006, p. 197) this implies the associated BB8 family copula also holds the LTD and RTI property, i.e.,  $LTD(Y|X)$  and  $RTI(Y|X)$  and because of symmetry  $LTD(X|Y)$  and  $RTI(X|Y)$ . Note that for  $0 < \theta \leq 1$ , it is inconclusive for this copula family. In Table 5, below, we provide the the summary of the dependence measures for the BB6 copula for the Swedish motor insurance data.

**Table 5.** Dependence structures of the BB6 copula for the Swedish motor insurance data.

Generator Function	$\phi(t) = (-\log[1 - (1 - t)^\theta])^\delta$
Blomqvist Beta (General)	$\beta = 4 \left( 1 - \left( 1 - e^{-[(-\log(1 - \frac{1}{2}^\theta))^\delta + (-\log(1 - \frac{1}{2}^\theta))^\delta]^{\frac{1}{\delta}}} \right)^{\frac{1}{\theta}} \right)$
Blomqvist $\beta$ (Swedish Auto)	0.7397
Upper-Tail (Swedish Auto)	$2 - 2^{\frac{1}{\theta}} = 0.8321$
Lower-Tail Dependence	0
Kendall's $\tau$	$1 + \frac{4}{\delta\theta} \int_0^1 (-\log(1 - (1 - t)^\theta)(1 - t)(1 - (1 - t)^{-\theta})dt$

### 5. Concluding Discussion and Remarks

In this article, we considered several well-known bivariate copulas, including the Tawn type-1, Frank, and BB6 families of copula based on the R package VineCopula for fitting two well-known insurance datasets arising out of automobile insurance. In addition, we also provided several useful structural properties of the selected bivariate copulas such as the LTD and RTI property, primarily focusing on the tail dependence properties, which are very important for studying dependence for insurance claims. This study shows that certain types of Archimedean copula with heavy tail dependence property are a reasonable framework to start with in terms of modeling insurance claim data, both in the bivariate as well as in the case of multivariate domains as appropriate. The goodness-of-fit statistics are provided in terms of AIC and BIC values as well as the log-likelihood values. As future research, we will be focusing on datasets from other domains (such as health care data), and we will consider the fitting to a trivariate and in higher dimensions as well based on the vine copula methodology. We will report our findings in a separate article. The tail-dependence coefficient has several applications, including: validation and verification of weather and climate models in reproducing extreme events; analysis of simultaneous extremes; probabilistic assessment of occurrences of extremes; and understanding climate variability. For example, by deriving tail-dependence coefficients for simulations of a numerical weather prediction model or a climate model, one can evaluate whether these models produces dependencies as seen in the observations. These approaches are not limited to precipitation, they also include a wide variety of earth science variables. This study of extreme tail dependence on local, regional, and global scales can assist in planning and policy making as well as validating numerical models, thus providing a valuable tool



for understanding how extreme events impact society. For future policy implementation out of this study, one may mention the following:

- Classes of extreme values of copulas (such as Tawn type-1, Frank, and BB6) are useful in modeling dependence for insurance claim data from the automobile industry that are asymmetric in nature.
- All the fitted bivariate copulas have one property in common, which is the nonzero value of the upper-tail dependence measure. This implies the fact that one observes an extremely large value for one component together with an extremely large value for the other component, a feature which is expected for insurance claim dataset-generated dependence structure. As a consequence, one can start with bivariate and multivariate copulas (as the case may be) that have a nonzero value of the upper-tail dependence measure  $\lambda_U$  when examining the dependence structure related to insurance claim data from the automobile industry.
- As a future study, it will be interesting to see whether such a class of extreme value copulas can be useful for insurance claims from other industries. Furthermore, for portfolio risk assessment, the effectiveness of such classes of copulas would be the subject matter of future research.

**Author Contributions:** Conceptualization: I.G.; Methodology: I.G. and D.W.; Software: I.G. and D.W.; Formal analysis: I.G.; Validation: I.G. and S.C.; Original draft Preparation: I.G. and D.W.; Writing—review and editing: I.G. and S.C.; Supervision: I.G. and S.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** No funding was available for all the authors.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** All the data are widely available and the data sources are also properly mentioned in the manuscript.

**Acknowledgments:** The authors would like to thank all the anonymous reviewers for their constructive comments and suggestions.

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

#### Appendix A. R Package: Vine Copula

Here, we provide a generic R-code based on the Vine Copula package which is used in the main body of the text for selecting the best possible bivariate copula of the four different insurance datasets:

```
install.packages("copula")
library("copula")
m<-pobs(a)
n<-pobs(b)
install.packages("VineCopula")
library("VineCopula")
selectedCopula <- BiCopSelect(m, n, familyset = NA)
summary(selectedCopula)
```

**Remark A1.** In the above code,  $a$  and  $b$  are the transformed (on a log (to the base  $e$ ) scale) variable values corresponding to two components of the associated bivariate data.

The best-fitted bivariate copulas mentioned here do not possess a closed form of expression in terms of their density function (i.e., the p.d.f.). However, in order to obtain the p.d.f. of each of these copulas, one may use *R*. Next, we provide an example as to how one can simulate from the p.d.f. of a Survival BB1 copula with specific parameter choices in *R*.

```
Simulate from a bivariate rotated BB1 copula
(180 degrees; ''survival BB1'')
install.packages("VineCopula")
library("VineCopula")
SBB1<- BiCop(family = 17, par =0.63, par2 =1.09)
sim<- BiCopSim(1000, SBB1)
```

## References

- Alexeev, Vitali, Katja Ignatieva, and Thusitha Liyanage. 2021. Dependence Modelling in Insurance via Copulas with Skewed Generalised Hyperbolic Marginals. *Studies in Nonlinear Dynamics and Econometrics* 25: 2. [CrossRef]
- Blomqvist, Nils. 1950. On a measure of dependence between two random variables. *Annals of Mathematical Statistics* 21: 593–600. [CrossRef]
- Chan, Joshua C. C., and Dirk P. Kroese. 2010. Efficient estimation of large portfolio loss probabilities in t-copula models. *European Journal of Operational Research* 205: 361–67. [CrossRef]
- Dakovic, Rada, and Claudia Czado. 2011. Comparing point and interval estimates in the bivariate t-copula model with application to financial data. *Statistical Papers* 52: 709–31. [CrossRef]
- Demarta, Stefano, and Alexander J. McNeil. 2005. The t Copula and Related Copulas. *International Statistical Review* 73: 111–29. [CrossRef]
- Embrechts, Paul, Alexander McNeil, and Daniel Straumann. 2001. Correlation and dependency in risk management: Properties and pitfalls. In *Risk Management: Value at Risk and Beyond*. Cambridge, UK: Cambridge University Press, pp. 176–223. Available online: <http://www.math.ethz.ch/~mcneil> (accessed on 15 June 2022).
- Fang, Hong-Bin, Kai-Tai Fang, and Samuel Kotz. 2002. The meta elliptical distributions with given marginals. *Journal of Multivariate Analysis* 82: 1–16. [CrossRef]
- Franc, Jean-Pierre, Michel Riondet, Ayat Karimi, and Georges L. Chahine. 2011. Impact load measurements in an erosive cavitating flow. *Journal of Fluids Engineering* 133: 121301. [CrossRef]
- Joe, Harry. 1997. *Multivariate Models and Dependence Concepts*. New York: Chapman & Hall.
- Mensi, Walid, Aviral Tiwari, Elie Bouri, David Roubaud, and Khamis H. Al-Yahyaee. 2017. The dependence structure across oil, wheat, and corn: A wavelet-based copula approach using implied volatility indexes. *Energy Economics* 66: 122–39. [CrossRef]
- Nadarajah, Saralees, Afuecheta Emanuel, and Chan Stephen. 2017. A compendium of copulas. *Statistica* 77: 279–328.
- Naeem, Muhammad Abubakr, Elie Bouri, Mabel D. Costa, Nader Naifar, and Syed Jawad Hussain Shahzad. 2021. Energy markets and green bonds: A tail dependence analysis with time-varying optimal copulas and portfolio implications. *Resources Policy* 74: 102418. [CrossRef]
- Nelsen, Roger B. 1999. *An Introduction to Copulas*, 1st ed. New York: Springer.
- Nelsen, Roger B. 2006. *An Introduction to Copulas*, 2nd ed. New York: Springer.
- Pettere, Gaida, and Tõnu Kollo. 2011. Risk modeling for future cash flow using skew t-copula. *Communications in Statistics-Theory and Methods* 40: 2919–25. [CrossRef]
- Pfeifer, Dietmar, and Johana Nešlehová. 2003. Modeling dependence in finance and insurance: The copula approach. *Blätter der DGVFM* 26: 177–91. [CrossRef]
- Shi, Peng, Xiaoping Feng, and Jean-Philippe Boucher. 2016. Multilevel modeling of insurance claims using copulas. *The Annals of Applied Statistics* 10: 834–63. [CrossRef]
- Sklar, Abe. 1959. Fonctions de rpartition À n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris* 8: 229–231.
- Sklar, Abe. 1973. Random variables, joint distribution functions, and copulas. *Kybernetika (Prague)* 9: 449–60.
- Sun, Wei, Svetlozar Rachev, Stoyan V. Stoyanov, and Frank J. Fabozzi. 2008. Multivariate skewed Student's t copula in the analysis of nonlinear and asymmetric dependence in the German equity market. *Studies in Nonlinear Dynamics & Econometrics* 12: 1–37.

Review

# On Asymmetric Correlations and Their Applications in Financial Markets

Linyu Cao <sup>1</sup>, Ruili Sun <sup>1,\*</sup>, Tiefeng Ma <sup>2</sup> and Conan Liu <sup>3</sup>

<sup>1</sup> College of Mathematics and Information Science, Zhengzhou University of Light Industry, Zhengzhou 450001, China

<sup>2</sup> School of Statistics, Southwestern University of Finance and Economics, Chengdu 611130, China

<sup>3</sup> UNSW Business School, University of New South Wales, Sydney, NSW 2052, Australia

\* Correspondence: sunruili2009@163.com

**Abstract:** Progress on asymmetric correlations of asset returns has recently advanced considerably. Asymmetric correlations can cause problems in hedging effectiveness and overstate the value of diversification. Furthermore, considering the asymmetric correlations in portfolio construction significantly enhances performance. The purpose of this paper is to trace developments and identify areas that require further research. We examine three aspects of asymmetric correlations: first, the existence of asymmetric correlations between asset returns and their significance tests; second, the test on the existence of asymmetric correlations between different markets and financial assets; and third, the root cause analysis of asymmetric correlations. In the first part, the contents of extreme value theory, the H statistic and a model-free test are covered. In the second part, commonly used models such as copula and GARCH are included. In addition to the GARCH and copula formulations, many other methods are included, such as regime switching, the Markov switching model, and the multifractal asymmetric detrend cross-correlation analysis method. In addition, we compare the advantages and differences between the models. In the third part, the causes of asymmetry are discussed, for example, higher common fundamental risk, correlation of individual fundamental risk, and so on.

**Keywords:** asymmetric correlation; statistical test; copula; GARCH



**Citation:** Cao, Linyu, Ruili Sun, Tiefeng Ma, and Conan Liu. 2023. On Asymmetric Correlations and Their Applications in Financial Markets. *Journal of Risk and Financial Management* 16: 187. <https://doi.org/10.3390/jrfm16030187>

Academic Editors: Ștefan Cristian Gherghina and Robert Brooks

Received: 9 January 2023  
Revised: 18 February 2023  
Accepted: 7 March 2023  
Published: 9 March 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Several recent studies corroborating asset returns have three asymmetric characteristics: the asymmetries in volatility, correlations, and betas. Notably, Black (1976) was the first researcher to consider asymmetry in volatility. Since then, asymmetric GARCH-type models have become popular when investigating the characteristics of financial time series, and a significant number of asymmetric GARCH models have been proposed (Choy et al. 2012). In addition, there is notable relevance between beta coefficients and asset pricing theories, and beta coefficients help to understand the riskiness of the associated asset stocks (Hong et al. 2007); see Ball and Kothari (1989), Conrad et al. (1991), and Bekaert and Wu (2000) for literature covering asymmetries in the betas.

This paper focuses on asymmetric correlations, the study of which is important for three reasons. Firstly, hedging mainly depends on the correlations between assets and financial instruments, and the existence of asymmetric correlations may lead to problems in hedging effectiveness (Hong et al. 2007). Second, in an optimal portfolio selection problem, if all stocks tend to fall with the decline of the market, the value of diversification may be exaggerated without considering the increase of downside correlations (Ang and Chen 2002). Third, taking the asymmetric correlations into account enhances the portfolio performance significantly (Taamouti and Tsafack 2009).

Let  $\{r_{1t}, r_{2t}\}$  denote two assets returns during time period  $t$ . For convenience of computation and statistical analysis, the returns are normalized to zero mean and unit

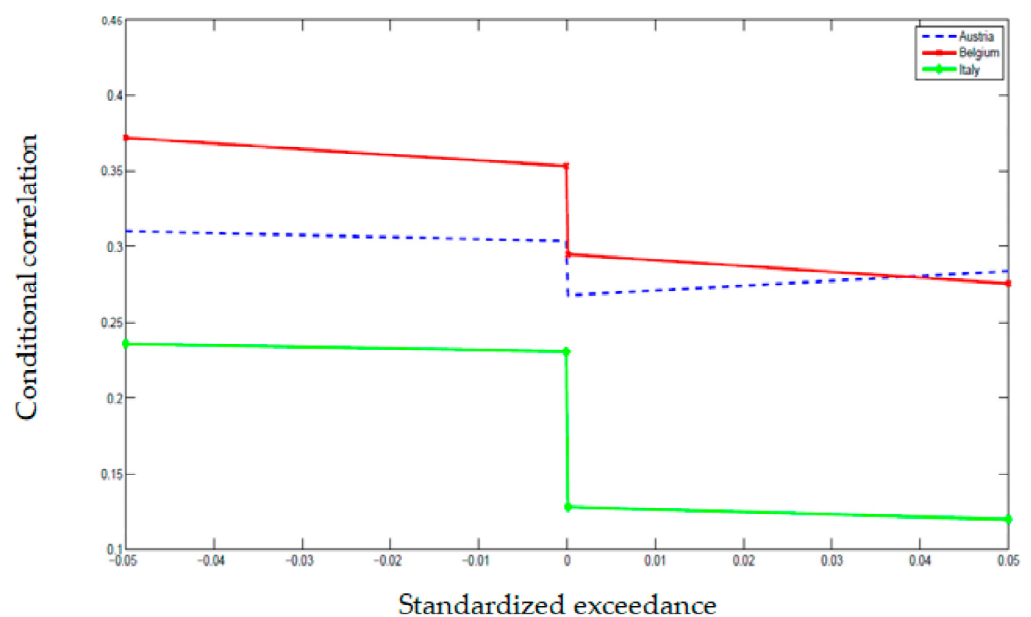
variance. Using the same notation as in Longin and Solnik (2001), Ang and Chen (2002), and Hong et al. (2007), we define the exceeding correlation at a given level  $c$  as follows:

$$\begin{aligned} \rho^+(c) &= \text{corr}(r_{1t}, r_{2t} | r_{1t} > c, r_{2t} > c), \\ \rho^-(c) &= \text{corr}(r_{1t}, r_{2t} | r_{1t} < -c, r_{2t} < -c), \end{aligned}$$

where  $\rho^+(c)$  measures the correlation between two returns above a certain exceedance level  $c$ , and  $\rho^-(c)$  measures the correlation below a certain exceedance level  $c$ . Additionally,  $\rho^+(c)$  represents the correlation during market upturns, and  $\rho^-(c)$  denotes the correlation during downturns.

If  $\rho^+(c) = \rho^-(c)$  for all  $c \geq 0$ , the correlation between the positive returns are the same as those with negative returns. This is called symmetric correlation. However, if  $\rho^+(c) \neq \rho^-(c)$  for all  $c \geq 0$ , then there are asymmetric correlations. Specifically, certain literature on co-movement also indicates asymmetric correlations.

The asymmetric correlations of Austrian, Belgian, and Italian government bonds with US government bonds from January 1976 to March 2010 are shown in Figure 1, taken from Ozsoy (2013). On the one hand, the three curves share some similar patterns, indicating these countries' exhibit larger conditional correlations on the negative standardized exceedances than those on the positive standardized exceedances. On the other hand, they differ from each other with Belgium's conditional correlations intersecting Austria's at standardized exceedances of 0 and about 0.038, and with Italy's conditional correlations at the bottom.



**Figure 1.** Asymmetric correlations.

This paper is organized as follows. In Section 2, we outline the existence of asymmetric correlations of asset returns and its significance test. In Section 3, we review the test on the existence of asymmetric correlations between different markets and financial assets. In Section 4, we present the root cause analysis of asymmetric correlations. In Section 5, we provide our conclusions of this study and directions for future research.

## 2. Existence of Asymmetric Correlations

From the introduction, it is clear that asymmetric correlations are a crucial topic in the research of portfolio selection-related issues. Therefore, in this section, we review the discovery of asymmetric correlations and their existence tests. We then summarize and provide some problems worthy of comprehensive study.

2.1. Literatures Review

In this subsection, we consider the existing research on asymmetric correlations. Some important literature is shown in Table 1.

**Table 1.** Selected work on existence of asymmetric correlations.

Author (Year)	Paper Title
Longin and Solnik (2001)	Extreme Correlation of International Equity Markets
Ang and Chen (2002)	Asymmetric Correlations of Equity Portfolios
Campbell et al. (2002)	Increased Correlation in Bear Markets
Hong et al. (2007)	Asymmetries in Stock Returns: Statistical Tests and Economic Evaluation
Pan et al. (2014)	Testing Asymmetric Correlations in Stock Returns via Empirical Likelihood Method
Jondeau (2016)	Asymmetry in Tail Dependence in Equity Portfolios

Erb et al. (1994) considered the behavior of correlation over time and predicting correlation to be of importance. Therefore, the changing international correlations in the G7 countries were investigated, and the results showed that correlations during recessions were higher than those during periods of growth, and that correlations were not symmetrical in up and down markets.

In order to verify the hypothesis that the correlation between international equity markets increases during fluctuation periods, Longin and Solnik (2001) used extreme value theory to model the tail of multivariate distribution, derived the extreme correlation distribution of the broad category distributions, and found that correlation was related to the market trend and that correlation increased in bear markets. Since Longin and Solnik (2001), asymmetric correlations have garnered more and more research attention.

However, Forbes and Rigobon (2002) found that a correlation calculated conditional on some variables was a biased estimator for the corresponding unconditional correlation.

Ang and Chen (2002) found that correlations between U.S. stocks and the aggregate U.S. market were much greater during declines than during market rallies. A new H statistic was developed to test conditional correlation asymmetries, which could correct for conditioning biases. Moreover, they established several empirical models about asymmetric correlation in the U.S. equity market. The results showed that mall stocks, value stocks, and past loser stocks had more asymmetric movements, and that stocks with lower betas exhibited greater asymmetric correlations by controlling for size.

To overcome estimator bias for implied correlation, Campbell et al. (2002) derived the quantile correlation estimator, which, based on the quantiles of the multivariate distribution, used the unbiased quantile correlation estimates to explore the correlations in international equity markets, and found that correlation in international equity returns increased significantly in bear markets.

Hong et al. (2007) emphasized that the H statistic proposed by Ang and Chen (2002) only answered the question of whether the asymmetry could be explicated by a given mode. Therefore, Hong et al. (2007) provided a model-free test for asymmetric correlations of stock returns in which stocks fluctuated with the market more often when the market fell than when it rose; the test also had a simple asymptotic chi-square distribution and could easily be applied to test the symmetries of beta and covariance. There existed significant asymmetries in size and momentum portfolios. To account for parameter and model uncertainties, a Bayesian framework was proposed to model them and evaluate their economic value. The results showed that taking the asymmetric characteristics of assets into consideration could significantly improve the performance of portfolio selection.

To investigate the robustness of recent empirical results that indicated a structural breakdown of correlation, Campbell et al. (2008) derived theoretical truncated and ex-

ceedance correlations, evaluated the performance of the truncated and exceedance correlation estimators, and found important asymmetry evidence of the conditional correlation functions.

Based on detrended fluctuation analysis (DFA), Alvarez-Ramirez et al. (2009) developed a DFA extension to study asymmetric correlations in nonstationary time-series, and the DFA version separated positive trends and negative trends to analyze the individual contributions to the overall scaling behavior. The results showed that the asymmetries of three different time-series were scale-dependent, and that there were different correlation properties depending on whether the signal trending was positive or negative.

Based on a conditional version of Kendall's tau and copula method, Manner (2010) proposed two tests for symmetric dependence; these tests outperformed the one proposed by Hong et al. (2007) in a Monte Carlo study. When the tests were applied to stock market returns and quarterly US GNP and unemployment data, the results showed that there was evidence of asymmetries and nonlinearities.

Livan and Rebecchi (2012) investigated the spectral properties of correlation matrices between distinct statistical systems, in which the correlation matrices were intrinsically nonsymmetrical, and extended the spectral analyses to the realm of complex eigenvalues. Random matrix theory was used to differentiate the noise and nontrivial correlation structures. The above results were applied to study the asymmetric correlation matrix of daily prices of the US and UK stock exchanges.

In order to analyze the asymmetric correlation of sovereign bond yield dynamics between eight Eurozone countries pair-wise, Dajčman (2013a) provided a dynamic version of the test proposed by Hong et al. (2007) and identified time periods when the correlation of Eurozone sovereign bond yield dynamics became asymmetric. They found that correlation between the positive and the negative yield dynamics between sovereign bonds became asymmetric after the start of the Eurozone debt crisis.

Pan et al. (2014) stressed that the model-free test proposed by Hong et al. (2007) seemed to be under-rejected in the size value and had low power in a finite sample. Therefore, they used an empirical likelihood method to conduct a model-free statistic that could test asymmetric correlations of stock returns, corrected the size performance using a bootstrap method, which improved the performance of Hong et al.'s (2007) test, and analyzed the asymmetric correlations of the China stock market and international stock markets, respectively. The results showed that asymmetric correlations occurred in the China stock market and international stock markets.

Jondeau (2016) considered that standard nonparametric measures of tail dependence had poor finite-sample properties in view of the limited number of observations in the tails of a joint distribution. Therefore, Jondeau (2016) developed a parametric model to measure and test asymmetry in tail dependence based on a multivariate noncentral *t* distribution. The proposed model accommodated situations in which the volatilities or the correlations between different asset returns changed over time. Applying the above model to real data, they found that the correlation between the international markets and Fama–French portfolios in bear markets was greater than that in bull markets.

Based on the statistic originally proposed by Hong et al. (2007), Alcock and Hatherley (2016) used a linear ( $\beta$ ) dependence invariant metric to investigate the price of asymmetric dependence on the cross section of Wall Street stocks, and found that the existence of asymmetric dependence between the firm's returns and those of the market would lead to corresponding price discounts or premiums, and that failing to recognize the impact of asymmetric dependence of the cost of capital may cause low pricing or insufficient subscription of public capital offerings.

Miyazaki and Hamori (2016) implemented the model-free test proposed by Hong et al. (2007) to study the asymmetric cross-asset correlations of the gold market. The results showed that gold exhibited asymmetric correlation with stocks and the US dollar, and by dividing the sample into three characteristic periods, the exceedance correlation also exhibited significant time variation even under similar market stress of the same asset pairs.

Jiang et al. (2018a) emphasized that the test proposed by Hong et al. (2007) did not solve asymmetry problems beyond the second moment and had low power. Therefore, to measure the asymmetric co-movement between returns on a single asset and the market returns, they proposed a model-free entropy measure, which provided a direct test for asymmetry in the joint distribution, generalizing the correlation-based test proposed by Hong et al. (2007). The results showed that many common portfolios such as size, book value, and momentum portfolios had significant asymmetry in statistics.

Jiang et al. (2018b) considered that the test proposed by Hong et al. (2007) captured only linear dependence. To characterize the general asymmetric dependence between two random variables, they proposed a modified information measure, provided a test of asymmetric dependence and examined its finite sample performance. The results showed that common stock portfolios and market returns in the US and other similarly developed countries existed obvious asymmetric correlations, and when these markets were in a downturn, they exhibited higher correlation with each other.

## *2.2. Conclusions and Further Research*

Erb et al. (1994) and Longin and Solnik (2001) played a pioneering role in the discovery of asymmetric correlations. However, Forbes and Rigobon (2002) found that there existed conditioning biases in the estimation of correlation. To correct the biases of correlation, Ang and Chen (2002), Campbell et al. (2002), and Hong et al. (2007) proposed a new H statistic, quantile correlation estimator, and a model-free test, respectively. Furthermore, Campbell et al. (2008) derived theoretical truncated and exceedance correlations to verify the robustness of recent empirical results. The other studies are mostly based on the research of Hong et al. (2007) and improve some of its shortcomings such as low power in a finite sample, or linear dependence.

However, there are still some problems worth considering and studying in the verification of the existence of asymmetric correlations. First, does the exceedance level  $c$  affect the results of all the test statistics mentioned above? If so, how does the exceedance level affect the results? How do we choose a reasonable and accurate exceedance level? Second, as pointed out by Dajčman (2013a), the model-free test proposed by Hong et al. (2007) depends on the time interval. The interesting question is whether the model-free test is consistent with the time interval and whether there are certain methods and criteria for the selection of time intervals.

## **3. Asymmetric Correlations between Different Markets and Financial Assets**

With the discovery of asymmetric correlations, especially the corresponding asymmetric correlation test statistics, more and more scholars are beginning to pay attention to the asymmetric correlations of asset returns. In the research of asymmetric correlations, the two most used models are GARCH family models and copula. In the first two subsections, we focus on asymmetric GARCH family models and copula. In the third subsection, we introduce some other research methods related to asymmetric correlations. Finally, we make a summary and comparative analysis, and put forward some new and open issues worth studying.

### *3.1. Asymmetric GARCH Formulations*

In this subsection, we first review the development of GARCH family models. Then, we represent the use of GARCH formulation in capturing the asymmetric correlations between different financial markets. Some pioneering research is summarized in Table 2, shown below.

**Table 2.** Selected works on correlation/covariance and GARCH.

Author (Year)	Paper Title
Engle (1982)	Autoregressive Conditional Heteroscedasticity and Estimates of UK Inflation
Bollerslev (1986)	Generalized Autoregressive Conditional Heteroscedasticity
Bollerslev (1990)	Modelling the Coherence in Short-run Nominal Exchange Rates: A Multivariate Generalized ARCH Model
Engle and Kroner (1995)	Multivariate Simultaneous Generalized ARCH
Tse and Tsui (2002)	A Multivariate Generalized Autoregressive Conditional Heteroscedasticity Model with Time-varying Correlations
Engle (2002)	Dynamic Conditional Correlation (DCC): A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models
Cappiello et al. (2006)	Asymmetric Dynamics in the Correlations of Global Equity and Bond Returns
Wang and Nie (2016)	Research of Asymmetric Dynamics in the Correlations of the Chinese Stock Markets
Chen et al. (2021)	On a Bivariate Hysteretic AR-GARCH Model with Conditional Asymmetry in Correlations

Engle (1982) first introduced the autoregressive conditional heteroscedasticity (ARCH) model, and Bollerslev (1986) subsequently extended the ARCH model to the generalized autoregressive conditional heteroskedasticity (GARCH) model. Bollerslev et al. (1988) proposed a multivariate GARCH (MGARCH) model and used it to estimate the earnings of bills, bonds, and stocks. To ensure that the conditional covariance was positive definite, Bollerslev (1990) proposed the constant conditional correlations (CCC) MGARCH model. However, many researchers found that practical financial data violated certain assumptions of the CCC MGARCH model. Engle and Kroner (1995) proposed a BEKK method for multivariate ARCH processes and derived the sufficiency constraints to ensure the conditional covariance matrices were positive definite. Kroner and Ng (1998) compared the restrictions of VECH, BEKK, factor ARCH, and CCC GARCH models; introduced a group of robust conditional moment tests to check whether the model was specified properly; and proposed a generalized adoption model that allowed for asymmetric influences on the variances and covariances. Many researchers have found that the correlation is not invariant, which means the correlation is time-varying. Tse and Tsui (2002) proposed a MGARCH model whose correlation could be changed over time, in which they decomposed the conditional variance–covariance matrix into a product of two parts: one was a conditional variance matrix, and the other was a conditional correlation coefficient matrix. They also stuck each term of the conditional variance matrix to a single variable GARCH model and engineered each element of the conditional correlation coefficient matrix to follow an ARMA model. Meanwhile, Engle (2002) suggested a DCC MGARCH model to estimate time-varying correlations. Since then, GARCH family models and its generations, especially the asymmetric version of the DCC MGARCH model, have been widely used in asymmetric correlations measurement and testing. For additional GARCH family models, see, e.g., Liu and Heyde (2008), Liu and Neudecker (2009), and Dewick (2022).

Next, let us briefly introduce the asymmetry generalized dynamic conditional correlation multivariate GARCH (AG-DCC-MVGARCH) model. Assume  $r_t$  is the  $p$ -dimensional asset returns at time  $t$ . Then,  $r_t$  obeys the multivariate normal distribution

$$r_t | \Omega_{t-1} \sim N(0, H_t),$$



where  $\Omega_{t-1}$  represents the information set at time  $t - 1$ ;  $H_t$  is the conditional variance-covariance matrix; and it can be decomposed as

$$H_t = D_t R_t D_t,$$

where  $D_t$  is a  $p \times p$  diagonal variance matrix of asset returns;  $D_t = \text{diag}\{\sqrt{h_{i,t}}\}$ ;  $h_{i,t}$  is the time-varying variance obtained from the single-variable GARCH model;  $R_t$  is the time-varying conditional correlation coefficient matrix defined as

$$R_t = Q_t^{*-1} Q_t Q_t^{*-1},$$

$$Q_t = (\bar{Q} - A' \bar{Q} A - B' \bar{Q} B - G' \bar{N} G) + A' \varepsilon_{t-1} \varepsilon_{t-1}' A + B' Q_{t-1} B + G' \eta_{t-1} \eta_{t-1}' G,$$

where  $Q_t^*$  is a diagonal matrix;  $Q_t^* = [\sqrt{q_{i,i,t}}]$ ,  $q_{i,i,t}$  is the corresponding diagonal element of  $Q_t$ ;  $A$ ,  $B$ , and  $G$  are  $p \times p$  parameter matrices;  $\varepsilon_{i,t} = \frac{r_{i,t}}{\sqrt{h_{i,t}}}$ ,  $\bar{Q}$  is the unconditional covariance matrix of  $\varepsilon_{i,t}$ ;  $\eta_{i,t} = I[\varepsilon_{i,t}] \circ \varepsilon_{i,t}$ ,  $I[\cdot]$  is the indicator function;  $\circ$  is the Hadamard product;  $\bar{N}$  is the unconditional variance-covariance matrix of  $\eta_{i,t}$  and can capture the asymmetric characteristics of conditional correlation.

Butler and Joaquin (2002) used three popular bivariate distributions (the normal, RiskMetrics' restricted GARCH(1,1) distribution) to investigate the correlations with monthly returns observed in bear, calm, and bull markets. The results showed that the correlation during the market declines was obviously higher than that predicted by the normal distribution and RiskMetrics distributions, and the correlation during the bear market was significantly higher than that during bull market.

Kearney and Potì (2005) focused on country-level market index correlations, applied the symmetric and asymmetric version of the DCC MGARCH model to capture dynamic correlations, and found mixed evidence of asymmetric correlation reactions to news types simulated by the traditional asymmetric DCC MGARCH formulations.

Cappiello et al. (2006) implemented an asymmetric version of the DCC MGARCH model proposed by Engle (2002) to investigate asymmetric correlations in international capital stock and bond returns. The results illustrated that both bonds and international capital stock exhibited asymmetric correlation.

In the presence of asymmetry in the tail dependence, Tsafack (2009) considered that the DCC MGARCH model was a symmetric model, and that symmetrical portfolio models of this kind would cause investors to undervalue the value at risk (VaR) or expected shortfall (ES) of the portfolio, concluding that it was important to adopt an asymmetric portfolio model, e.g., the Gumbel copula, to deal with the asymmetric correlation problem.

To study the correlation between some notable indices and bonds in the United States, Yang et al. (2010) applied an asymmetric generalized DCC MGARCH model to a series of daily data, such as the S&P 500 and corporate bonds, and their real estate counterparts. They found that the correlation between REIT and stock returns exhibited asymmetries.

Horvath and Poldauf (2012) used multivariate GARCH models to investigate the co-movements of certain stock markets among various countries. The results showed that during 2008–2010, the correlation between stock returns increased, and that the correlation between the stock markets in the US and China was basically zero before the crisis, but slightly increased during the crisis.

Choy et al. (2012) used a bivariate GARCH model with DCC and leverage effect to model financial data, and proposed a new modified multivariate t-distribution, which offered independent marginal Student-t distributions, to highlight the relationship between different stock returns. The empirical study showed that the correlations between the oil price shocks and stock returns from 2008 to 2009 increased significantly.

Chen (2013) employed the asymmetry generalized dynamic conditional correlation multivariate GARCH (AG-DCC-MVGARCH) model, quasi-maximum likelihood estimation, and LR test to investigate the asymmetric and dynamic correlation of stock returns in

the US and China, and found that the correlation between different stock returns enhanced during bear markets.

Toyoshima and Hamori (2013) used the asymmetric DCC MGARCH model to describe the correlation of stock markets in Japan and Singapore, and found that financial integration had advanced due to the Japan–Singapore Economic Partnership Agreement, and that the investment portfolio in Asia had increased since the recent global financial crisis.

Gjika and Horváth (2013) used the asymmetric DCC MGARCH model to study stock market co-movements in central Europe. The results showed that the correlations increased over time, and that the stock markets exhibited asymmetric correlations to a certain degree.

Since the mean variance model was the most important model in the portfolio optimization, Kalotychou et al. (2014) explored its economic value in modeling conditional correlations and evaluated its dynamic strategies. They found that, by characterizing the change of correlation properly, fund managers could improve risk-adjusted returns by accurately capturing correlation time variation.

El Abed (2016) adopted a multivariate asymmetric DCC EGRACH framework to investigate the correlations of US dollar exchange rates and three European stock prices, and found that there were asymmetric responses in correlations, and that the correlation between exchange rates and stock prices increased during times of crisis.

Chen (2016) used the AG-DCC-MGARCH model to analyze the correlations among the four main stock markets in China and the impacts of the major economic events on the dynamics of the correlation coefficients of the four main stock markets. The results showed that the conditional correlations between Hong Kong and Shanghai, Hong Kong and Shenzhen, and Shanghai and Shenzhen were asymmetric.

Wang and Nie (2016) built EGARCH and an asymmetric version of the DCC MGARCH model to investigate dynamics and asymmetries in conditional variance and correlations in the Chinese stock markets. They found that A and B shares significantly existed dynamics and asymmetry in conditional correlation.

By generalizing the time-varying conditional correlation model proposed by Tse and Tsui (2002), Chen et al. (2021) suggested a new MHAR-A-GARCH-T model and used it to investigate the correlations with conditionally dynamic asymmetric structure. Moreover, by employing an adaptive Bayesian MCMC method, they found that adopting the asymmetric effects made a difference in estimation of dynamic correlations.

### *3.2. Copula Formulations*

In this subsection, we first review the advancement of copula, and then introduce the application of copula in asymmetric correlations.

Sklar (1959) proposed copula to verify the structure of dependency, especially the latent nonlinear correlation. Many researchers find that copula works well in capturing the correlation of financial data, so it is widely used in correlation measurement of financial data (Embrechts 1999). Since then, different copulas have been developed and are used in financial data exploration (Mashal and Zeevi 2002; Van den Goorbergh et al. 2005; Bartram et al. 2007; Chen and Tu 2013; Pastpipatkul et al. 2018). For more details about copula and its applications, see, e.g., Dewick and Liu (2022). Some important publications are listed in Table 3.

**Table 3.** Selected works on copula.

Author (Year)	Paper
Sklar (1959)	Fonctions Derépartitionà Dimensions et Leurs Marges
Embrechts (1999)	An Introduction to Copulas
Mashal and Zeevi (2002)	Beyond Correlation: Extreme Co-movements between Financial Assets
Patton (2006)	Modelling Asymmetric Exchange Rate Dependence
Christoffersen et al. (2012)	Is the Potential for International Diversification Disappearing? A Dynamic Copula Approach

We assume  $F(x_1, \dots, x_p)$  is an arbitrary  $p$  dimension joint distribution function,  $F_1(x_1), \dots, F_p(x_p)$  are marginal distribution functions of  $F(x_1, \dots, x_p)$ , and  $C(u_1, \dots, u_p)$  is a  $p$  dimension copula of  $F(x_1, \dots, x_p)$  if they satisfy the equation

$$F(x_1, \dots, x_p) = C(F_1(x_1), \dots, F_p(x_p))$$

where the above function  $C$  is called a copula of  $F$ . Moreover, if the marginal distributions are continuous, then there is a unique copula  $C$  corresponding to the joint distribution  $F(x_1, \dots, x_p)$ , which can be obtained from

$$C(u_1, \dots, u_p) = F(F_1^{-1}(u_1), \dots, F_p^{-1}(u_p))$$

On the contrary, the corresponding density function of joint distribution  $F(x_1, \dots, x_p)$  is calculated with

$$f(x_1, \dots, x_p) = c(F_1(x_1), \dots, F_p(x_p)) \prod_{i=1}^p f_i(x_i)$$

given the density functions exist, where  $f_i(x_i)$  represents the marginal density functions and  $c$  is the density function of the copula and can be obtained by the equation

$$c(u_1, \dots, u_p) = \frac{f(F_1^{-1}(u_1), \dots, F_p^{-1}(u_p))}{\prod_{i=1}^p f_i(F_i^{-1}(u_i))}$$

Patton (2004) considered the portfolio selection problem for investors with constant relative risk aversion, used models that could depict fourth order time-varying moments, and constructed time-varying dependence structure models allowing for different dependencies during bear markets and bull markets using copula theory. They found that the understanding of higher moments and asymmetric dependence would, in some cases, bring significant economic and statistical benefits to investors without short-selling restrictions.

Based on the GARCH model and regime-switching (RS) copula function, Wei and Zhang (2005) constructed the RS-copula–GARCH model to investigate the asymmetric tail dependence structure in Chinese stock markets and found that tail dependence structure of Shanghai and Shenzhen stock markets were asymmetric, and that RS-copula–GARCH model was superior to static copula model in describing dependence.

To test asymmetry of dependence between the German mark and the Japanese yen, Patton (2006) generalized the copulas theory to adopt conditioning variables and built conditional dependence models to fit the dependence of these exchange rates. The results showed that the exchange rates were more correlated when depreciating against the dollar than when appreciating.

To capture time-varying and nonlinear relationships among European stock markets, Bartram et al. (2007) used a time-varying copula model in which a GARCH formulation us-

ing a Gloston Jagannatha Runkle-generalized autoregressive conditional heteroskedasticity-moving average-t model was used to model the marginal distributions and the Gaussian copula was adopted to model the joint distribution. The results showed that market dependence increased after the introduction of the common currency only for large equity markets.

In order to investigate the dual dependence of exchange rates against the dollar, Boero et al. (2011) employed nonparametric plots and a robust semiparametric method to obtain the copula function. The results showed that the model captured asymmetric tail dependence well.

Using four parametric copulas to model the dependence structure at different investment horizons, Kang et al. (2010) reexamined the asymmetric correlations within hedge fund returns and market returns at a range of investment, and found that the dependence asymmetry was not limited to a specific time range but emerged clearly at all investment periods, and that the size of asymmetry was not invariable to the investment period, and its degree decreased significantly with the extension of investment period.

Garcia and Tsafack (2011) proposed an alternative RS copula of extreme dependence asymmetry. The copula-based model included one normal regime where dependence was symmetric and a second regime in which it was characterized by asymmetric dependence. Applying the above model to the capital stock and bond markets, they observed significant asymmetric behavior between different markets.

Christoffersen et al. (2012) considered that international equity markets were characterized by nonlinear dependence and asymmetries, proposed a new dynamic asymmetric copula model that allowed for asymmetric and dynamic tail dependence, and found that correlations had increased significantly in all markets.

To investigate the asymmetric dependence of financial data, Uhm et al. (2012) employed the copula approach for directional dependence. They found that the exchange rates correlation between the Republic of Korea and Japan was asymmetric due to the influence of the 2008 financial crisis and concluded that the direction-dependent copula method could be supplemented to interpret the asymmetric dependence.

Cho and Lee (2022) considered that default probabilities of credit portfolios were seriously affected by system risk, so they used the GJR-GARCH model and copula method to fit the volatility and dependence, respectively, proposing a new time-varying credit risk model. The results showed that the suggested model outperformed the existing model, and that there was strong evidence to show the existence of asymmetric correlation of asset returns.

### *3.3. Other Asymmetric Formulations*

Except the GARCH and copula formulations, many other methods are used to explore the asymmetric correlations, such as regime switching, the Markov switching model, and the multifractal asymmetric detrend cross-correlation analysis method (MF-ADCCA).

In order to characterize the risk and return in risk arbitrage, Mitchell and Pulvino (2001) used piecewise linear regressions to analyze 4750 mergers from 1963 to 1998. The results showed that risk arbitrage returns in most environments were uncorrelated with market returns, and that the correlation between market returns and risk arbitrage returns increased dramatically during market downturn.

The existence of asymmetric correlation made investors question the correctness of international diversification. In order to investigate the above result, Ang and Bekaert (2002) introduced RS model to deal with the asset allocation problem within a dynamic international situation and found that international diversification was still useful under regime changes.

Yuan (2005) presented a rational expectations equilibrium model to cope with the determinants of asset market crises and contagion. They found that market return distributions were asymmetric and that correlations between different asset returns tended to increase during crashes.

Michayluk et al. (2006) examined the volatility spillover effects and the inherent correlation among the US- and UK-securitized real estate indices, and found the correlation of different markets exhibited asymmetry.

To verify whether asymmetric correlations existed and determine the explanation of asymmetry, Taamouti and Tsafack (2009) used the generalized impulse response function under an autoregressive model framework to quantify the relationship among return, volatility, and correlation, and tested the asymmetric correlations between return and volatility against correlation. The results showed that considering the asymmetric correlation between return and correlation could obtain improved financial gain.

Abid and Bahloul (2011) used the discrete time Markov switching model to analyze the behavior of equity returns correlations, investigated the effect of this behavior on international portfolio allocation, and found that the correlations in a bear market showed obvious difference with correlations in the bull market.

Lee et al. (2011) examined the performance of asset correlation with the market returns in the asymptotic single risk factor (ASRF) approach of the Basel II accord on regulatory capital requirement and found that asset correlations were asymmetric.

By comparing the equity market in Croatia in good (bull, clam) and bad (bear, turbulent) market conditions, Kunovac (2012) found that correlations between stock prices during bear markets more than doubled those exhibited during bull markets. In addition, they found that the losses might occur if the asymmetry was ignored in practice by the research of taking asymmetric correlation into consideration and assessing the performance of the portfolio selection model.

Cao et al. (2013) used asymmetric multifractal detrended fluctuation analysis to test the asymmetry of China's stock markets in the upward or downward trend. They found that asymmetric correlation was more obvious in large fluctuations.

Dajčman (2013b) examined the asymmetric correlation pair-wise between the Eurozone's stock market returns, and investigated if the results were sensitive to a time span of returns. The results showed the asymmetric correlation test relied on the time span of returns.

By using the Chinese market data, Cao et al. (2014) proposed the MF-ADCCA model to study the asymmetric correlations in stock and exchange market. The empirical results showed that there was asymmetric cross-correlation between Chinese stock market and the Chinese RMB exchange market.

Based on theoretical derivation, Chen et al. (2014) studied the time varying correlation between the Chinese stock market and the broader macroeconomy. The results showed that there was indeed asymmetric correlation between the Chinese stock market and global economies.

To verify whether the strength of the co-movements caused by market declines and market rallies were significantly different, Li (2014) developed a nonparametric test, and the proposed test could be applied to verify whether there were asymmetric co-movements resulting from a linear or nonlinear dependence. The results showed significant evidence of asymmetric co-movements in the stock markets of the U.S. and other developed countries.

To study the correlation of gold prices and oil prices with COVID-19, Mensi et al. (2020) used the asymmetric multifractal detrended fluctuation analysis (A-MF-DFA) method to investigate the impacts between them and found obvious evidence of asymmetric multifractality that increased as the fractality scale increased.

Kristjanpoller et al. (2020) used the MF-ADCCA approach to study asymmetric multifractality and found significant evidence of asymmetric multifractality in the cross-correlation between five main cryptocurrencies and six equity ETFs.

Based on the autoregressive distributed lag model, Thampanya et al. (2020) investigated the asymmetric influences of gold and cryptocurrency returns on the Thai stock market, and studied whether hedging functions of gold or cryptocurrency were still effective in the event of a stock market decline or rally. The results showed that gold and cryptocurrencies were not good tools for stock market hedging.

Given the fact that industry and market portfolios showed the asymmetry in correlations, Kim et al. (2021) developed a novel optimal consumption and portfolio selection framework and found that neglecting asymmetric correlations could cause loss to investors.

Xu et al. (2021) used the multifractal cross-correlation analysis method to investigate the asymmetric cross-correlations between international stock markets such as the China and US markets. The empirical results showed that the cross-correlations between markets were asymmetric, and that the cross-correlations were more stable and stronger in bear markets than those exhibited in bull markets.

Chuang et al. (2022) suggested nonparametric tests to verify asymmetric co-movements, applied them to daily return of SP 500 and 29 individual stocks, and found that most stock returns showed the showed asymmetric co-movements.

### *3.4. Conclusions and Further Research*

Since Bollerslev et al. (1988) proposed the MGARCH model, MGARCH is widely used in the research of multiple asset returns. In particular, DCC model proposed by Engle (2002), a new family of multivariate GARCH models, constructs the model based on using the MGARCH model to study asymmetric correlation. Much of the research on asymmetric correlations based on GARCH model use the asymmetric version of DCC model.

Copula has unique advantages in the study of correlation, especially for nonlinear relationships. In the research of asymmetric correlations, copula is often combined with other models, such as GARCH model and regime switching.

In addition, regime switching, the Markov switching model, and the MF-ADCCA model are also used to investigate asymmetric correlations.

Through the review of research on asymmetric correlations, we compare the difference and advantages of the aforementioned models:

(a) GARCH family models are usually used to interpret covariance asymmetry. The most used GARCH family model with asymmetric correlation is an asymmetric version of DCC model proposed by Engle (2002). The asymmetric DCC MGARCH model could consider the asymmetric effects on conditional second moments, adopt asymmetric dynamics in the correlation as well as the asymmetric response in variances, and accommodate different news impact patterns for correlations between different assets. However, traditional GARCH family models are constructed using the conditionally normal distribution assumption of asset returns, have too many unknown parameters to estimate, and usually impose limited scope or significant parameter restrictions.

(b) Copula is a more effective measurement of dependence between multivariate variables; since the joint distribution is nonelliptical the conventional correlation cannot capture the dependence structure appropriately. In addition, when decomposing multivariate distribution into marginal distributions, copula can construct a better distribution of stock returns than existing multivariate distributions. However, copula needs the assumption of marginal distributions and needs to specify an affirmatory dependence structure about the asset returns.

(c) The multivariate regime switching model is a useful parametric alternative to copula models. In the regime switching model, the Markov switching model is a special case of regime switching model in which the discrete state variable follows a Markov chain process. The regime switching model is versatile and effective in capturing nonlinear relationships. However, the regime switching model assumes that the observations come from a mixture of parametric distributions and constant transition probabilities for the unobserved states. Furthermore, the identification of the number of regimes is also important but difficult. Both copula and regime switching models are usually combined with other models, such as the GARCH model.

(d) The multifractal asymmetric detrend cross-correlation analysis method is model-free and easy to implement. It can be used to analyze the nonlinear and highly volatile nature of and investigate asymmetric multifractality between financial time series data.

#### 4. Root Cause Analysis of Asymmetric Correlations

##### 4.1. Literature Review

As the asymmetric correlations garner the attention of many researchers, the root cause of asymmetric correlations also increases in popularity. To our knowledge, we classify the literature on the root cause of asymmetric correlations into four categories: the first is cashflow related causes, the second is firm-level return dispersions, the third is skewness-related causes, the fourth is other causes. Some important publications on root cause of asymmetric correlations are listed in Table 4.

**Table 4.** Selected works on root cause of asymmetric correlations.

Author (Year)	Paper
Yu and Wu (2001)	Economic Sources of Asymmetric Cross-correlation among Stock Returns
Demirer and Lien (2004)	Firm-level Return Dispersion and Correlation Asymmetry: Challenges for Portfolio Diversification
Ding et al. (2011)	Asymmetric Correlations in Equity Returns: a Fundamental-based Explanation
Albuquerque (2012)	Skewness in Stock Returns: Reconciling the Evidence on Firm Versus Aggregate Returns
Chung et al. (2019)	What Causes the Asymmetric Correlation in Stock Returns

Campbell (1991) and Vuolteenaho (2002) showed that stock returns could be decomposed into the following components: the expected return, shocks to expected cashflows, and shocks to discount rates. However, Vuolteenaho (2002) and Campbell and Vuolteenaho (2004) found that the first two components of stock returns were related, and pointed out that stock returns were caused by cashflow news. Therefore, cash flow related causes are first investigated. Yu and Wu (2001) suggested an alternative framework to explain and verify major causes of asymmetric cross-correlation and found the asymmetric cross-correlation was mostly attributed to differences in sensitivity of stock prices to market information and the differences in quality of cash flow information of differently sized firms. Chung et al. (2019) considered the latent causes of the asymmetric correlation in stock returns. They found that the correlation of firms' cash flow news variable and other accounting measures of firm performance was asymmetric, and that only the asymmetric correlation in firm performance could explain the asymmetric correlation in stock returns.

Unlike the cashflow-related causes, firm-level return dispersions were only studied by Demirer and Lien (2004). Demirer and Lien (2004) studied the question of whether firm-level return dispersions could explain asymmetric correlations in stock returns significantly. The results showed that asymmetric correlations were correlated with asymmetric firm-level return dispersions, and that portfolio managers need to take the asymmetry in return correlation and firm-level return dispersions into account.

Skewness of financial data is another cause of asymmetric correlations. Emphasizing that significant literature explained aggregate stock market returns, displayed negative skewness, and ignored the fact that firm stock returns displayed positive skewness, Albuquerque (2012) provided a unified theory, built a stationary asset pricing model of firm announcement events, and found that cross-sectional heterogeneity could result in asymmetric correlations in stock returns. Chung and Kim (2017) thought that asymmetric correlations led to negative skewness of portfolios, provided asymmetric correlation measurements by using portfolio skewness, and found that asymmetric correlation was generated at the asset level of individual firms.

The other causes of asymmetric correlations include increasing common fundamental risk, investor sentiment, variance and earning price ratio, and so on. However, they can only partially explain the asymmetric correlation. Ding et al. (2011) offered an explanation

to the potential fundamental causes of the asymmetric correlations of stock portfolio returns. They found that several sources caused the asymmetry during market decline, such as increasing common fundamental risk, and they also concluded that these key factors were only part of the causes of asymmetric correlation. Wang et al. (2021) considered that the tests proposed to verify the existence of asymmetric correlations in previous literature could not be used in practical investment due to the natures of time-varying and unpredictable of asymmetric correlations. Therefore, they constructed a unified state–space model to measure in-sample and out-of-sample asymmetric correlations. They found that there were many factors that resulted in asymmetric correlations, such as investor sentiment, variance and price-to-earnings ratio, but they all could not fully explain the asymmetric correlation.

#### *4.2. Conclusions and Further Research*

Through the above review, we can see that researchers have conducted extensive research on the causes of asymmetric correlations and that various factors may cause or partially cause the asymmetric phenomenon of asset return. In our view, the financial market is rapidly changing. Therefore, during different periods of time, especially with the different financial policy guidance of each country, the causes for the asymmetry of asset returns may not be unique and certain; that is, different periods of time and different countries have different causes. There may not be a uniform determining cause for the asymmetric correlations of asset return, but there is a common applicable research framework, which can contain various causes and methods that need to be verified one by one according to the actual situation.

### **5. Conclusions**

Since Markowitz (1952), the portfolio selection problem has been a hot topic. However, when the asset returns show asymmetry in the correlations, the portfolio selection problem should be reconsidered seriously. Therefore, asymmetric correlations of stock returns play an important role in portfolio selection and risk management. In this paper, we review the development and application of asymmetric correlations in financial markets and identify the directions for future research. This review focuses on three aspects: (a) the existence of asymmetric correlations between stock returns and its significance test; (b) the test on the existence of asymmetric correlations between different markets and financial assets; (c) the root cause analysis of asymmetric correlations.

Abundant empirical research verifies that the correlations of stock returns are higher in bear markets than in bull markets. Longin and Solnik (2001) are among the first to show the existence of asymmetric correlations after controlling for bias resulting from conditioning. The relevant methods and tools used on testing the existence of asymmetric correlations are extreme value theory, quantile, Kendall’s tau, the copula method, detrended fluctuation analysis, etc. For the test on the existence of asymmetric correlations, GARCH family models and the copula method are two main methods. In addition, regime switching, the Markov switching model, and multifractal asymmetric detrend cross-correlation analysis method are also important tools. Asymmetric correlations also become a stylized fact of asset returns. In order to deepen the study of asymmetric correlations, the root causes of asymmetric correlations have also attracted the interest of researchers. According to the contents of root causes of asymmetric correlations, we divide them into four categories: the cash flow related causes, the firm-level return dispersions, the skewness related causes and other causes.

However, there are still many open issues worthy of consideration and research. For example, for the hypothesis testing of asymmetric correlations, how does the exceedance level affect the results of all the test statistics mentioned above, and how can we choose a reasonable and accurate exceedance level? In addition, Kang et al. (2010) found that the dependence of asymmetry was not limited to a specific time range but emerged clearly at all investment periods, that the size of asymmetry was not invariable to the investment period, and its degree decreased significantly with the extension of an investment period.



Is there an appropriate way to measure the change degree of the size of asymmetry? Is the change degree of the size of asymmetry linear or nonlinear?

As mentioned at the beginning of this paper, the asset returns do exist asymmetrically in the volatility and correlations, but will the performance of portfolio selection be improved by taking the asymmetry in the volatility and correlations into account simultaneously? In addition, it is well known that the rolling window method proposed by DeMiguel and Nogales (2009) is widely used in testing the performance of portfolio selection. If we combine the above two methods and apply them to the portfolio selection problem, how can we detect the upturns and downturns of asset returns pairwise for a certain time window?

As to the cause of asymmetric correlations, how do we build a common applicable research framework, which can contain various causes and methods that need to be verified one by one according to the actual situation?

The asymmetric correlations only measure the asymmetry in terms of time; however, we consider that the asymmetry in correlation has two levels: the first is the time level, the second is the individual level, which means the asymmetry in different asset returns. For instance, on the stock market, the leader stock in one industry has a significant positive impact on other stocks, while other stocks in the same industry have a rather small positive impact on the leader stock. How do we measure the asymmetry at an individual level and combine the asymmetry in two levels of asset returns? Moreover, Chatterjee (2021) introduced a simple new rank correlation coefficient, which is not symmetric in two random variables. How can we use it in the portfolio selection problem?

**Author Contributions:** Original draft Preparation: L.C. and R.S.; Writing-review and editing: T.M. and C.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** The first two authors received support from the National Natural Science Foundation of [China (12201579)].

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors sincerely thank the editors and reviewers for their insightful comments which aided in improving the paper.

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

## References

- Abid, Fathi, and Slah Bahloul. 2011. Regime switching, asymmetric correlation and international portfolio choices. *International Journal of Monetary Economics and Finance* 4: 172–94. [CrossRef]
- Albuquerque, Rui. 2012. Skewness in stock returns: Reconciling the evidence on firm versus aggregate returns. *The Review of Financial Studies* 25: 1630–73. [CrossRef]
- Alcock, Jamie, and Anthony Hatherley. 2016. Characterizing the asymmetric dependence premium. *Review of Finance* 21: 1701–37. [CrossRef]
- Alvarez-Ramirez, Jose, Eduardo Rodriguez, and Juan Carlos Echeverria. 2009. A DFA approach for assessing asymmetric correlations. *Physica A: Statistical Mechanics and Its Applications* 388: 2263–70. [CrossRef]
- Ang, Andrew, and Geert Bekaert. 2002. International asset allocation with regime shifts. *Review of Financial Studies* 15: 1137–87. [CrossRef]
- Ang, Andrew, and Joseph Chen. 2002. Asymmetric correlations of equity portfolios. *Journal of Financial Economics* 63: 443–94. [CrossRef]
- Ball, Ray, and S. P. Kothari. 1989. Nonstationary expected returns: Implications for tests of market efficiency and serial correlation in returns. *Journal of Financial Economics* 25: 51–74. [CrossRef]
- Bartram, Söhnke, Stephen Taylor, and Yaw-Huei Wang. 2007. The euro and European financial market integration. *Journal of Banking and Finance* 31: 1461–81. [CrossRef]
- Bekaert, Geert, and Guojun Wu. 2000. Asymmetric volatility and risk in equity markets. *Review of Financial Studies* 13: 1–42. [CrossRef]
- Black, Fisher. 1976. Studies of stock market volatility changes. In *Proceedings of the 1976 Meeting of the Business and Economic Statistics Section*. Washington DC: American Statistical Association, pp. 177–81.
- Boero, Gianna, Param Silvapulle, and Ainura Tursunaliyeva. 2011. Modelling the bivariate dependence structure of exchange rates before and after the introduction of the euro: A semi-parametric approach. *International Journal of Finance and Economics* 16: 357–74. [CrossRef]
- Bollerslev, Tim. 1986. Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics* 31: 307–27. [CrossRef]

- Bollerslev, Tim. 1990. Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. *Review of Economics and Statistics* 72: 498–505. [CrossRef]
- Bollerslev, Tim, Robert F. Engle, and Jeffrey M. Wooldridge. 1988. A capital asset pricing model with time-varying covariances. *The Journal of Political Economy* 96: 116–31. [CrossRef]
- Butler, Kirt, and Domingo Joaquin. 2002. Are the gains from international portfolio diversification exaggerated? The influence of downside risk in bear markets. *Journal of International Money and Finance* 21: 981–81. [CrossRef]
- Campbell, John Y. 1991. A variance decomposition for stock returns. *The Economic Journal* 101: 157–79. [CrossRef]
- Campbell, John Y., and Tuomo Vuolteenaho. 2004. Bad beta, good beta. *American Economic Review* 94: 1249–75. [CrossRef]
- Campbell, Rachel A. J., Kees Koedijk, and Paul Kofman. 2002. Increased Correlation in Bear Markets. *Financial Analysts Journal* 58: 87–94. [CrossRef]
- Campbell, Rachel A. J., Catherine S. Forbes, Kees G. Koedijk, and Paul Kofman. 2008. Increasing correlations or just fat tails? *Journal of Empirical Finance* 15: 287–309. [CrossRef]
- Cao, Guangxi, Jie Cao, and Longbing Xu. 2013. Asymmetric multifractal scaling behavior in the Chinese stock market: Based on asymmetric MF-DFA. *Physica A: Statistical Mechanics and Its Applications* 392: 797–807. [CrossRef]
- Cao, Guangxi, Jie Cao, Longbing Xu, and Lingyun He. 2014. Detrended cross-correlation analysis approach for assessing asymmetric multifractal detrended cross-correlations and their application to the Chinese financial market. *Physica A: Statistical Mechanics and Its Applications* 393: 460–69. [CrossRef]
- Cappiello, Lorenzo, Robert F. Engle, and Kevin Sheppard. 2006. Asymmetric dynamics in the correlations of global equity and bond returns. *Journal of Financial Econometrics* 4: 537–72. [CrossRef]
- Chatterjee, Sourav. 2021. A new coefficient of correlation. *Journal of the American Statistical Association* 116: 2009–22. [CrossRef]
- Chen, Yun. 2013. Empirical study on asymmetric dynamic correlations among stock returns in the US, Hong Kong and Mainland China. *Management Science* 26: 79–88.
- Chen, Menglong. 2016. Asymmetric dynamic in China's stock markets correlations. *Journal of Financial Research* 9: 41–48.
- Chen, Yi-Hsuan, and Anthony H. Tu. 2013. Estimating hedged portfolio value-at-risk using the conditional Copula: An illustration of model risk. *International Review of Economics and Finance* 27: 514–28. [CrossRef]
- Chen, Shoudong, Xiaowei Yi, and Yang Liu. 2014. Research on asymmetric effects of correlation between China's stock markets and Macro-Economy under uncertain policies. *Journal of Contemporary Finance and Economics* 1: 45–55.
- Chen, Cathy W. S., Hong Than-Thi, and Manabu Asai. 2021. On a bivariate hysteretic AR-GARCH model with conditional asymmetry in correlations. *Computational Economics* 58: 413–33. [CrossRef]
- Cho, Yongbok, and Yongwoong Lee. 2022. Asymmetric asset correlation in credit portfolios. *Finance Research Letters* 49: 1030–37. [CrossRef]
- Choy, S. T. Boris, Cathy W. S. Chen, and Edward M. H. Lin. 2012. Bivariate asymmetric GARCH models with heavy tails and dynamic conditional correlations. *Quantitative Finance* 14: 1–17. [CrossRef]
- Christoffersen, Peter, Vihang Errunza, Kris Jacobs, and Hugues Langlois. 2012. Is the potential for international diversification disappearing? A dynamic Copula approach. *Review of Financial Studies* 25: 3711–51. [CrossRef]
- Chuang, O-Chia, Xiaojun Song, and Abderrahim Taamouti. 2022. Testing for asymmetric comovements. *Oxford Bulletin of Economics and Statistics* 84: 1153–80. [CrossRef]
- Chung, Y. Peter, and Thomas S. Kim. 2017. Asymmetric correlation as an explanation for the effect of asset skewness on equity returns. *Asia-Pacific Journal of Financial Studies* 46: 686–99. [CrossRef]
- Chung, Y. Peter, Hyun A. Hong, and S. Thomas Kim. 2019. What causes the asymmetric correlation in stock returns? *Journal of Empirical Finance* 54: 190–212. [CrossRef]
- Conrad, Jennifer, Mustafa N. Gultekin, and Gautam Kaul. 1991. Asymmetric predictability of conditional variances. *Review of Financial Studies* 4: 597–622. [CrossRef]
- Dajčman, Silvo. 2013a. A formal test of asymmetric correlation in stock market returns and the relevance of time interval of returns—A case of Eurozone stock markets. *Acta Polytechnica Hungarica* 10: 9–19.
- Dajčman, Silvo. 2013b. Asymmetric correlation of sovereign bond yield dynamics in the Eurozone. *Panoeconomicus* 60: 775–89. [CrossRef]
- DeMiguel, Victor, and Francisco J. Nogales. 2009. Portfolio selection with robust estimation. *Operations Research* 57: 560–77. [CrossRef]
- Demirer, Riza, and Donald Lien. 2004. Firm-level return dispersion and correlation asymmetry: Challenges for portfolio diversification. *Applied Financial Economics* 14: 447–56. [CrossRef]
- Dewick, Paul R. 2022. On financial distributions modelling methods: Application on regression models for time series. *Journal of Risk and Financial Management* 15: 461–76. [CrossRef]
- Dewick, Paul R., and Shuangzhe Liu. 2022. Copula modelling to analyse financial data. *Journal of Risk and Financial Management* 15: 104–15. [CrossRef]
- Ding, Liang, Hiroyoki Miyake, and Hao Zou. 2011. Asymmetric correlations in equity returns: A fundamental-based explanation. *Applied Financial Economics* 21: 389–99. [CrossRef]
- El Abed, Riadh. 2016. On the comovements among European exchange rates and stock prices: A multivariate time-varying asymmetric approach. *Journal of Applied Finance and Banking* 6: 53–79.
- Embrechts, Paul. 1999. Extreme value theory as a risk management tool. *North American Actuarial Journal* 3: 30–41. [CrossRef]

- Engle, Robert F. 1982. Autoregressive conditional heteroscedasticity and estimates of UK inflation. *Econometrica* 50: 987–1008. [CrossRef]
- Engle, Robert F. 2002. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business and Economic Statistics* 20: 339–50. [CrossRef]
- Engle, Robert F., and Kenneth F. Kroner. 1995. Multivariate simultaneous generalized ARCH. *Econometric Theory* 11: 122–50. [CrossRef]
- Erb, Claude B., Campbell R. Harvey, and Tadas E. Viskanta. 1994. Forecasting international equity correlations. *Financial Analysts Journal* 50: 32–45. [CrossRef]
- Forbes, Kristin, and Roberto Rigobon. 2002. Nontagion, only interdependence: Measuring stock market co-movements. *Journal of Finance* 57: 2223–61. [CrossRef]
- Garcia, René, and Georges Tsafack. 2011. Dependence structure and extreme comovements in international equity and bond markets. *Journal of Banking and Finance* 35: 1954–70. [CrossRef]
- Gjika, Dritan, and Roman Horváth. 2013. Stock market comovements in Central Europe: Evidence from the asymmetric DCC model. *Economic Modelling* 33: 55–64. [CrossRef]
- Hong, Yongmiao, Jun Tu, and Guofu Zhou. 2007. Asymmetries in stock returns: Statistical tests and economic evaluation. *Review of Financial Studies* 20: 1547–81. [CrossRef]
- Horvath, Roman, and Petr Poldauf. 2012. International stock market comovements: What happened during the financial crisis? *Global Economy Journal* 12: 1–21. [CrossRef]
- Jiang, Lei, Wu Ke, and Guofu Zhou. 2018a. Asymmetry in stock comovements: An Entropy Approach. *Journal of Financial and Quantitative Analysis* 53: 1479–507. [CrossRef]
- Jiang, Lei, Esfandiar Maasoumi, Jiening Pan, and Ke Wu. 2018b. A test of general asymmetric dependence. *Journal of Applied Econometrics* 33: 1026–43. [CrossRef]
- Jondeau, Eric. 2016. Asymmetry in tail dependence in equity portfolios. *Computational Statistics and Data Analysis* 100: 351–68. [CrossRef]
- Kalotychou, Elena, Sotiris K. Staikouras, and Gang Zhao. 2014. The role of correlation dynamics in sector allocation. *Journal of Banking and Finance* 48: 1–12. [CrossRef]
- Kang, Byoung Uk, Francis In, Gunky Kim, and Tong Suk Kim. 2010. A longer look at the asymmetric dependence between hedge funds and the equity market. *Journal of Financial and Quantitative Analysis* 45: 763–89. [CrossRef]
- Kearney, Colm, and Valerio Poti. 2005. Correlation dynamics in European equity markets. *Research in International Business and Finance* 20: 305–21. [CrossRef]
- Kim, Myeong Hyeon, Seyoung Park, and Jong Mun Yoon. 2021. Industry portfolio allocation with asymmetric correlations. *European Journal of Finance* 27: 178–98. [CrossRef]
- Kristjanpoller, Werner, Elie Bouri, and Tetsuya Takaishi. 2020. Cryptocurrencies and equity funds: Evidence from an asymmetric multifractal analysis. *Physica A: Statistical Mechanics and Its Applications* 545: 123711. [CrossRef]
- Kroner, Kenneth F., and Victor K. Ng. 1998. Modeling asymmetric comovements of asset returns. *Review of Financial Studies* 11: 817–44. [CrossRef]
- Kunovac, Davor. 2012. Asymmetric correlations on the Croatian equity market. *Financial Theory and Practice* 35: 1–24.
- Lee, Shih-Cheng, Chien-Ting Lin, and Chih-Kai Yang. 2011. The asymmetric behavior and procyclical impact of asset correlations. *Journal of Banking and Finance* 35: 2559–68. [CrossRef]
- Li, Fuchun. 2014. Identifying asymmetric comovements of international stock market returns. *Journal of Financial Econometrics* 12: 507–43. [CrossRef]
- Liu, Shuangzhe, and Chris Heyde. 2008. On estimation in conditional heteroskedastic time series models under non-normal distributions. *Statistical Papers* 49: 455–69.
- Liu, Shuangzhe, and Heinz Neudecker. 2009. On pseudo maximum likelihood estimation for multivariate time series models with conditional heteroskedasticity. *Mathematics and Computers in Simulation* 79: 2556–65.
- Livan, Giacomo, and Luca Rebecchi. 2012. Asymmetric correlation matrices: An analysis of financial data. *European Physical Journal B* 85: 213–24. [CrossRef]
- Longin, Francois, and Bruno Solnik. 2001. Extreme correlation of international equity markets. *Journal of Finance* 56: 649–76. [CrossRef]
- Manner, Hans. 2010. Testing for asymmetric dependence. *Studies in Nonlinear Dynamics and Econometrics* 14: 1–32. [CrossRef]
- Markowitz, Harry. 1952. Portfolio selection. *Journal of Finance* 7: 77–91.
- Mashal, Roy, and Assaf Zeevi. 2002. *Beyond Correlation: Extreme Co-Movements between Financial Assets*. Technical Report. New York: Columbia University.
- Mensi, Walid, Ahmet Sensoy, Xuan Vinh Vo, and Sang Hoon Kang. 2020. Impact of COVID-19 outbreak on asymmetric multifractality of gold and oil prices. *Resources Policy* 69: 101829. [CrossRef]
- Michayluk, David, Patrick J. Wilson, and Ralf Zurbrugg. 2006. Asymmetric volatility, correlation and returns dynamics between the U.S. and U.K. securitized real estate markets. *Real Estate Economics* 34: 109–32. [CrossRef]
- Mitchell, Mark, and Todd Pulvino. 2001. Characteristics of risk and return in risk arbitrage. *Journal of Finance* 56: 2135–75. [CrossRef]
- Miyazaki, Tomomi, and Shigeyuki Hamori. 2016. Asymmetric correlations in gold and other financial markets. *Applied Economics* 48: 4419–25. [CrossRef]
- Ozsoy, Sati Mehmet. 2013. Asymmetric Correlations in Financial Markets. Ph.D. thesis, Department of Economics, Duke University, Durham, NC, USA.

- Pan, Zhiyuan, Xu Zheng, and Qiang Chen. 2014. Testing asymmetric correlations in stock returns via empirical likelihood method. *China Finance Review International* 4: 42–57. [CrossRef]
- Pastpipatkul, Pathairat, Woraphon Yamaka, and Songsak Sriboonchitta. 2018. Portfolio selection with stock, gold and bond in Thailand under vine Copulas functions. *Econometrics for Financial Applications* 760: 698–711.
- Patton, Andrew J. 2004. On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. *Journal of Financial Econometrics* 2: 130–68. [CrossRef]
- Patton, Andrew J. 2006. Modelling asymmetric exchange rate dependence. *International Economic Review* 47: 527–56. [CrossRef]
- Sklar, Abe. 1959. Fonctions de répartition à dimensions et leurs marges. *Publication Institute Statistics University Paris* 8: 229–31.
- Taamouti, Abderrahim, and Georges Tsafack. 2009. Asymmetric Effects of Return and Volatility on Correlation between International Equity Markets. Available online: <https://ssrn.com/abstract=1344416> (accessed on 16 February 2009).
- Thampanya, Natthinee, Muhammad Ali Nasir, and Toan Luu Duc Huynh. 2020. Asymmetric correlation and hedging effectiveness of gold & cryptocurrencies: From pre-industrial to the 4th industrial revolution. *Technological Forecasting and Social Change* 159: 120195.
- Toyoshima, Yuki, and Shigeyuki Hamori. 2013. Asymmetric dynamics in stock market correlations: Evidence from Japan and Singapore. *Journal of Asian Economics* 24: 117–23. [CrossRef]
- Tsafack, Georges. 2009. Asymmetric dependence implications for extreme risk management. *Journal of Derivatives* 17: 7–20. [CrossRef]
- Tse, Yiu Kuen, and Albert K. C. Tsui. 2002. A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. *Journal of Business and Economics Statistics* 20: 351–62. [CrossRef]
- Uhm, Daiho, Jong-Min Kim, and Yoon-Sung Jung. 2012. Large asymmetry and directional dependence by using copula modeling to currency exchange rates. *Model Assisted Statistics and Applications* 7: 327–40. [CrossRef]
- Van den Goorbergh, Rob W. J., Christian Genest, and Bas J. M. Werker. 2005. Bivariate option pricing using dynamic copula. *Mathematics and Economics* 37: 101–14. [CrossRef]
- Vuolteenaho, Tuomo. 2002. What drives firm-level stock returns? *Journal of Finance* 57: 233–64. [CrossRef]
- Wang, Xiang, and Fuqiang Nie. 2016. Research of asymmetric dynamics in the correlations of the Chinese stock markets. *Journal of Applied Statistics and Management* 5: 907–15.
- Wang, Nianling, Lijie Zhang, Zhuo Huang, and Yong Li. 2021. Asymmetric correlations in predicting portfolio returns. *International Review of Finance* 21: 97–120. [CrossRef]
- Wei, Yanhua, and Shiyang Zhang. 2005. Research on asymmetric tail dependence structure in financial markets. *Chinese Journal of Management* 5: 601–5.
- Xu, Lin, Xiaoying Lin, and Wan Xiao. 2021. Asymmetric multifractal cross-correlations analysis on global stock market based on modified MF-ADCCA model. *Statistics and Information Forum* 10: 41–54.
- Yang, Jian, Yinggang Zhou, and Wai Kin Leung. 2010. Asymmetric correlation and volatility dynamics among stock, bond, and securitized real estate markets. *Journal of Real Estate Finance and Economics* 45: 491–521. [CrossRef]
- Yu, Chih-Hsien, and Chunchi Wu. 2001. Economic sources of asymmetric cross-correlation among stock returns. *International Review of Economics and Finance* 10: 19–40. [CrossRef]
- Yuan, Kathy. 2005. Asymmetric price movements and borrowing constraints: A rational expectations equilibrium model of crises, contagion, and confusion. *Journal of Finance* 60: 379–411. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



MDPI  
St. Alban-Anlage 66  
4052 Basel  
Switzerland  
[www.mdpi.com](http://www.mdpi.com)

*Journal of Risk and Financial Management* Editorial Office

E-mail: [jrfm@mdpi.com](mailto:jrfm@mdpi.com)  
[www.mdpi.com/journal/jrfm](http://www.mdpi.com/journal/jrfm)



Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.





Academic Open  
Access Publishing

[mdpi.com](http://mdpi.com)

ISBN 978-3-7258-0481-8