risks

MDPI

Article

Investigating Causes of Model Instability: Properties of the Prediction Accuracy Index

Ross Taplin

Curtin Business School, Curtin University, Kent St, Bentley, WA 6102, Australia; r.taplin@curtin.edu.au

Abstract: The Prediction Accuracy Index (PAI) monitors stability, defined as whether the predictive power of a model has deteriorated due to a change in the distribution of the explanatory variables since its development. This paper shows how the PAI is related to the Mahalanobis distance, an established statistic for examining high leverage observations in data. This relationship is used to explore properties of the PAI, including tools for how the PAI can be decomposed into effects due to (a) individual observations/cases, (b) individual variables, and (c) shifts in the mean of variables. Not only are these tools useful for practitioners to help determine why models fail stability, but they also have implications for auditors and regulators. In particular, reasons why models containing econometric variables may fail stability are explored and suggestions to improve model development are discussed.

Keywords: Prediction Accuracy Index (PAI); Basel Accord; IFRS 9; model monitoring; model validation

1. Introduction

Model stability is the extent to which the predictive accuracy of a model deteriorates since its development by examining changes in the distribution of the explanatory variables. This is important because models used in credit risk assessment, such as probability of default (PD) models, are generally developed using one set of data (development data) but implemented using different data collected subsequently. For example, under the Basel Accord (Basel Committee on Banking Supervision 2006) for capital and the International Financial Reporting Standards (IFRS 9) for provisioning (International Accounting Standards Board 2014), models are developed to predict future potential losses and are implemented until they are considered no longer "fit-for-purpose". Model stability is an important feature of fit-for-purpose model reviews because stability does not require the observation of the response variable and therefore is available immediately. Other statistics, such as the calibration of the PD to actual default rates, require the observation of loan defaults. For example, a PD model predicting defaults in the next 12 months will use model inputs that are at least 12 months old. Hence, evaluations of stability are timely and important, not only for banks, insurance, and other companies performing credit or other risk modelling, but also for auditors and regulators. The important question of whether a model remains "fit-for-purpose" when the response variable (e.g., default outcome) is also available is typically not referred to as stability and is beyond the scope of this paper.

Traditionally, model stability has been evaluated using the Population Stability Index (PSI), especially in the context of credit models. However, Taplin and Hunt (2019) showed this had inferior properties for model stability. In particular, by detecting any shift in the distribution of the explanatory variables, it detects changes that have no impact on the predictive accuracy of a model. Indeed, Taplin and Hunt (2019) provided examples of where the PSI indicated low model stability when the accuracy of the model actually improved in the review data compared to the development data. This situation does not indicate a lack of stability of the model, making the PSI of questionable value when assessing whether a model remains fit-for-purpose. Taplin and Hunt (2019) introduced

check for updates

Citation: Taplin, Ross. 2023. Investigating Causes of Model Instability: Properties of the Prediction Accuracy Index. *Risks* 11: 110. https://doi.org/10.3390/risks11060110

Academic Editor: Mogens Steffensen

Received: 24 April 2023
Revised: 29 May 2023
Accepted: 4 June 2023
Published: 7 June 2023

the Predictive Accuracy Index (PAI) to assess the degree of deterioration in the model's predictive accuracy more effectively than the PSI (see Section 1.1 for its definition).

Kruger et al. (2021) recommended the PAI over the PSI due to its superior properties. They particularly recommended the multivariate PAI (MPAI) over the univariate version that considers only one explanatory variable at a time. Becker and Becker (2021) provided further examples where the PSI and PAI produce different results; however, they only considered the univariate version.

Considering the high interest in the PAI, experience gained from using the PAI since its introduction in 2019, and a lack of publications exploring the properties of the PAI, this paper aims to:

(a) explore the properties of the (multivariate) PAI;
(b) reflect on its use in practice;
(c) make further recommendations on how to investigate the cause of instability when a high value of the PAI indicates a lack of model stability.

These results are important in practice in two ways. First, they provide a set of additional statistics and analyses suitable when a model lacks stability. Second, they suggest how techniques used to develop models may lead to model instability. Thus, this paper also contributes by providing advice on model development that is relevant to model developers, auditors, and regulators, providing advice or guidelines. However, this paper does not consider techniques using the response variable to assess whether a model remains fit-for-purpose and it does not consider the broader question of how important variables in a model are to its predictions. While relatively straightforward for techniques such as regression, the machine learning literature contains important research on this topic for other modelling techniques (Lundberg and Lee 2017; Giudici and Raffinetti 2021).

1.1. The Prediction Accuracy Index (PAI)

Taplin and Hunt (2019, p. 5) defined the Prediction Accuracy Index (PAI) as "the average variance of the estimated mean response at review divided by the average variance of the estimated mean response at development". This definition is broad and can be applied to any model. Taplin and Hunt (2019) show that, in the important case of multiple regression (or any model such as logistic regression for probability of default modelling that uses a linear predictor)

$$PAI = \frac{\sum_{j=1}^{N} r_j^T V r_j / N}{\sum_{i=1}^{n} x_i^T V x_i / n} \tag{1}$$

where

r_j is the vector of explanatory variables for the jth observation of the review data (j = 1 to N);
x_i is the vector of explanatory variables for the ith observation of the development data (i = 1 to n);
$V = \text{MSE} \times \left(\text{X}^{\text{T}}\text{X}\right)^{-1}$ is the variance–covariance matrix of the estimated regression coefficients;
X is the design matrix with columns defined by the x_i (the explanatory variables at development).

The vectors r_j and x_i are column vectors that typically have a dimension of $(k+1)$ as they contain not only the k explanatory variables but also a 1 for the intercept.

When interpreting values of the PAI, Taplin and Hunt (2019, pp. 5–6) recommended "values less than 1.1 indicate no significant deterioration; values from 1.1 to 1.5 indicate a deterioration requiring further investigation; and values exceeding 1.5 indicate the predictive accuracy of the model has deteriorated significantly". Note that, in Taplin and Hunt (2019), the PAI is referred to as the Population Stability Index in the title; however, we refer to it as the Prediction Accuracy Index as this is how it is referred throughout Taplin and Hunt (2019) and because it more accurately reflects its purpose to summarise the predictive accuracy of a model.

This paper concentrates on the use of the multivariate PAI (referred to as MPAI in Taplin and Hunt (2019) but referred to here simply as the PAI) in preference to the univariate PAI (Equation (1) when there is only one explanatory variable). This is because most models contain many explanatory variables and because an examination of the impacts of a change in the distribution of one explanatory variable with the PAI (or PSI) in isolation is inconsistent with the properties of the model (unless the model only contains one explanatory variable). We also assume that the model contains a constant term (therefore, the first column of the design matrix X contains a value of 1). In practice, it would be most unusual for a model not to contain such an intercept. Note that this implies that r_j and x_i are vectors with a first entry equal to 1 and that all entries in the first column of the design matrix X equal 1.

1.2. Notation and Illustrative Data Examples

The notation used by Taplin and Hunt (2019) for the explanatory variables (r_j for the review data and x_i for the development data) is problematic since a variable that always contains the value of 1 is not normally considered an explanatory variable. We therefore introduce more intuitive notation, replacing r_j, x_I, and X with ρ_j, ϵ_i, and E, respectively, when the explanatory variables exclude the 1 for the intercept. For example, ϵ_I contains the values of the k explanatory variables (excluding the 1 for the intercept) while I is a vector of dimension $(k+1)$ with an additional value of 1 in front: $x_i^T = \left(1, \epsilon_i^T\right)$. Here we use ϵ and E (e for explanatory variables, in the traditional sense, without a variable with ones) instead of χ (the Greek letter for x) as the capital letter for χ is indistinguishable from the capital for x (which might introduce confusion between the new notation and the notation in Taplin and Hunt (2019).

Throughout this paper, we illustrate techniques using the simple (fictitious) development data in Table 1, which contain three explanatory variables: a, b, and c. For example, there are two observations in this data with $a = 1$, $b = 1$, and $c = 1$ but only one observation with $a = 2$, $b = 1$, and $c = 1$. This latter observation would be denoted as $x_i = (1, 2, 1, 1)^T$ in Taplin and Hunt (2019) and, in our notation, as $\epsilon_i = (2, 1, 1)^T$. The three explanatory variables a, b, and c have means of 3.12, 2.08, and 1.44, respectively, in this development data. The variables a and b are highly correlated ($r = 0.91$) with moderate, negative correlations between a and c ($r = -0.34$) and between b and c ($r = -0.35$). The model uses a linear combination of these three explanatory variables to predict the response variable (e.g., a logistic regression model to predict the default). This development data will typically be the raw observations used to develop the model. For example, for a home loan portfolio a, b, and c could represent the loan-to-value ratio (security), the income-to-repayments ratio (serviceability), and the current interest rate (economic conditions).

Table 1. Number of observations in the illustrative development data for combinations of the three explanatory variables $\epsilon_i = (a, b, c)^T$.

c	***b***	***a***					
		1	**2**	**3**	**4**	**5**	**6**
1	1	2	1	0	0	0	0
1	2	0	2	7	4	0	0
1	3	0	0	0	6	5	1
2	1	9	1	0	0	0	0
2	2	0	2	4	1	0	0
2	3	0	0	0	1	3	1

Together with the development data in Table 1, we considered two different sets of review data. Review data R1 have the same 50 observations as the development data in Table 1 together with an additional 51st observation with $a = 6$, $b = 1$, and $c = 1$. That is, $\rho_{51} = (6, 1, 1)^T$. For the review data R1, the *PAI* equals 1.31, which Taplin and Hunt

(2019) interpret as a deterioration requiring further investigation. The review data R2 are described in Table 2. For this review data, $PAI = 1.58$, which Taplin and Hunt (2019) interpret as a significant deterioration in the predictive accuracy of the model.

Table 2. Number of observations in the illustrative review data R2 for combinations of the three explanatory variables $\rho_j = (a, b, c)^T$.

c	*b*	*a*					
		1	**2**	**3**	**4**	**5**	**6**
1	1	1	2	1	0	0	0
1	2	0	1	3	6	4	0
1	3	0	0	0	1	5	4
2	1	1	8	1	0	0	0
2	2	0	0	2	3	2	0
2	3	0	0	0	1	1	3

These review data are typically not only out-of-sample (i.e., not used to develop the model) but also out-of-time (i.e., relates to a point in time different to the development data). Typically, the review data are later in time (more recent) than the development data. Stress testing models are, by design, intended to be used in situations where some degree of extrapolation is involved, therefore the PAI and the techniques presented in this paper may be of less value. Rather than designating the models using the benchmarks of 1.1 and 1.5 suggested by Taplin and Hunt (2019), for stress testing situations, these techniques may be useful to quantify the extent of the extrapolation, or to determine which econometric variables are responsible for most of the extrapolation.

2. The PAI as a Function of the Squared Mahalanobis Distances

The square of the Mahalanobis distance of a vector of explanatory variables v (relative to the distribution with mean μ and variance–covariance matrix S) equals $M_v = (v - \mu)^T S^{-1} (v - \mu)$. First introduced by Mahalanobis (1936), this distance is useful for detecting multivariate outliers. A multivariate observation v has $M_v = 0$ when $v = \mu$ and M_v is large for an outlier if it deviates a large distance from μ in a direction in which the standard deviation of the multivariate distribution is relatively small.

While the Mahalanobis distance might be well known to modellers, some might be more familiar with the leverage (typically denoted h) used to determine whether outliers exist in the explanatory variables. Due to the monotonic relationship between the squared Mahalanobis distance and leverage, either is suitable as a scale to examine observations.

Model developers may examine the Mahalanobis distance (or, equivalently, the leverage) of observations in the development data, relative to the distribution (mean and variance–covariance matrix) of the development data. This provides useful information concerning the properties of the development data and which (if any) observations have a high leverage on the estimated parameters. Similarly, model reviewers may examine the Mahalanobis distance of observations in the review data (relative to the distribution of the review data).

Table 3 provides the squared Mahalanobis distance M_v based on the development data in Table 1. For example, the squared Mahalanobis distance is 1.0 for the seven observations $\epsilon_i = (3, 2, 1)^T$. This distance is close to 0 because these observations are close to the mean $\bar{\varepsilon} = (3.12,\ 2.08,\ 1.44)^T$ of the development data. M_v increases when the value of any of these variables deviates from the central mean value $\bar{\varepsilon}$, especially when a is high and b is low (or vice versa) due to the high positive correlation between variables a and b.

Table 3. Squared Mahalanobis distance M_v for values of $\epsilon_i = (a, b, c)^T$ using the illustrative development data in Table 1.

c	*b*	*a*					
		1	**2**	**3**	**4**	**5**	**6**
1	1	4.3	5.7	12.5	24.8	42.5	65.7
1	2	12.0	3.8	1.0	3.7	11.8	25.5
1	3	40.3	22.5	10.2	3.3	1.8	5.9
2	1	2.5	4.2	11.4	24.0	42.1	65.7
2	2	11.7	3.8	1.4	4.4	12.9	26.9
2	3	41.6	24.1	12.1	5.5	4.4	8.8

Note that the squared Mahalanobis distance M_v is not strictly defined when the explanatory variables are defined as in Taplin and Hunt (2019) because the data contain a variable always equal to 1; moreover, since this "variable" has a variance of 0, the variance–covariance matrix would not be invertible. Finally, note that, in defining the variance–covariance matrix S for the explanatory variables, we used the definition of variance and covariances for a population (not a sample). For example, the variance of an explanatory variable x (with n values) is defined as $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ (not using $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$). Using this simple mean of squared deviations simplifies mathematical expressions and makes a negligible difference to numerical values for a realistically large sample (n) of development data.

An Expression for the PAI as a Function of the Squared Mahalanobis Distance M_v

The PAI in Equation (1) can be expressed in terms of the squared Mahalanobis distances as follows (see Appendix A for the derivation):

$$PAI = \frac{1 + \overline{M}_r}{1 + \overline{M}_d}. \tag{2}$$

where

$\overline{M}_r$ equals the average of the squared Mahalanobis distance of the review data ρ_j,

$\overline{M}_d$ equals the average of the squared Mahalanobis distance of the development data ϵ_i,

and all Mahalanobis distances are relative to the mean and variance–covariance matrix of the development data. That is, the squared Mahalanobis distance of observation v (either ρ_j for review data or for ε_j for development data) is

$$M_v = (v - \overline{\epsilon})^T V_d^{-1} (v - \overline{\epsilon}) \tag{3}$$

where $\overline{\epsilon}$ contains the mean of the explanatory variables at development and V_d is the variance–covariance matrix of the explanatory variables at development. If the formula for the sample variance–covariance matrix ($n - 1$ instead of n) was used for V_d, the value of 1 on both the numerator and denominator of Equation (2) would need to be replaced with the value of $n/(n-1)$.

For example, using the development data in Table 1, the denominator of Equation (2) equals 3.00 because the average of the squared Mahalanobis distances in the development data equals 2.00. This average of 2.00 is the weighted average of the values in Table 3 (using the values in Table 1 as weights). The numerator of Equation (2) equals the corresponding average of squared Mahalanobis distances for the review data. For example, the 51st observation $\rho_{51} = (6,\ 1,\ 1)^T$ in review data R1 has a squared Mahalanobis distance of 65.7 (see Table 3). This increases the average squared Mahalanobis distance from 2.00 in the

development data to 3.23 in the review data. Thus, from Equation (2), the value of the PAI for the review data R1 is

$$PAI = \frac{1 + 3.23}{1 + 2.00} = \frac{4.23}{3.00} = 1.31$$

In this simple example, the reason for this deterioration is the additional observation $\rho_{51} = (6, 1, 1)^T$, which is an outlier relative to the development data.

For review data R2, the average squared Mahalanobis distance is 5.32. Therefore, from Equation (2), the PAI is

$$PAI = \frac{1 + 5.32}{1 + 2.00} = \frac{5.32}{3.00} = 1.58$$

Equation (2) is a novel use of the Mahalanobis distance because both the review data (numerator) and development data (denominator) use the distribution of the development data (therefore the numerator uses the mean and variance–covariance matrix of the development data even though the distances are calculated for the review data). This differs from the more routine use of the Mahalanobis distance (such as those produced in standard software), where observations are compared to the dataset they come from (observations in the development data are compared to the distribution of the development data and observations in the review data are compared to the distribution of the review data).

3. The Contribution of Individual Observations to the PAI

Equation (2) suggests that a large PAI is due to observations with a large squared Mahalanobis distance (relative to the average squared Mahalanobis distance at development) during the review. Therefore, when further investigation is required (PAI > 1.1), it is prudent to examine the Mahalanobis distance of all the review observations in case the model's deterioration is due to one or several observations. This situation was illustrated using review data R1 in the previous section, where all the excess PAI above 1 was due to one outlier (by construction). The next highest squared Mahalanobis distance is 8.8 for the review observation $\rho_j = (6, 3, 2)^T$.

Rather than examining the squared Mahalanobis distance of an observation, we recommend, in the context of the PAI, an alternative scale that is defined by the effect on the PAI if an observation is removed. Thus, we defined I_k, the influence of review observation ρ_j on the PAI, as

$$I_k = PAI_0 - PAI_{(-k)}$$

where PAI_0 is the PAI using all N observations of the review data (ρ_j; j = 1 to N) and $PAI_{(-k)}$ is the value of the PAI after removing the observation ρ_j. The following properties of I_k also assist with its interpretation (see Appendix B for derivations):

(a) I_k can be either positive or negative;
(b) The average of the I_k is 0;
(c) A useful reference point for a small I_k is 0, since, while I_k can be negative, the PAI does not change if an observation with $I_k = 0$ is removed;
(d) A useful reference point for a large I_k is $PAI_0 - 1$, since the PAI is reduced to 1 when an observation with $I_k = PAI_0 - 1$ is removed;
(e) I_k can be written in terms of the squared Mahalanobis distance m_k of observation ρ_j:

$$I_k = PAI_0 - PAI_{(-k)} = \frac{m_k - \overline{M}_r}{N(1 + \overline{M}_d)}$$

For example, removing the outlier $\rho_{51} = (6, 1, 1)^T$ from the review data R1 results in a value of the PAI equal to 1.0 and $I_k = 0.31$. Thus, this one observation accounts for 100% of the amount the PAI of 1.31 exceeds 1.0. For review data R2, each of the three observations $\rho_i = (6, 3, 2)^T$ (Table 2; bottom right) have a value of I_k equal to 0.02 (the PAI decreases from 1.58 to 1.56 if one of these observations is removed, therefore each of these observations account for only 0.02/0.58 = 3% of the amount the original PAI exceeds 1).

In contrast, the observations $\rho_i = (3,2,2)^T$ have values of $I_k = -0.02$, which are negative because, for these observations, the squared Mahalanobis distance of 1.4 (Table 3) is less than the average of 2.0 for the development data. The distribution of I_k values for all the observations in the review data R2 suggest that the large PAI value for R2 is not due to one or a small number of observations.

Furthermore, the large PAI may be due to a subset of observations with a large Mahalanobis distance rather than just one or a few outliers. Moreover, in this case, it may be useful to examine the squared Mahalanobis distance for the subsets of the review data. For example, even if a variable such as the gender of the applicant is not used in the model, the model may predict some of these subsets better than other subsets. These distances can be examined using the review data and development data, or by assessing whether they have changed from development to review. For example, Table 4 shows the average squared Mahalanobis distance for males and females at development and at review, as well as the PAI for these subsets. While these averages suggest that the model has slightly inferior accuracy when predicting females than males, the difference is considerably higher at review. The resulting large PAI value of 1.31 for females suggests that the model's predictive ability has deterorated for female applicants.

Table 4. Average squared Mahalanobis distances and PAI for males and females, illustrating a model that predicts females slightly less accurately than males at development but considerably less accurately at review.

Subset	Development Data	Review Data	PAI
Male applicants	4.3	4.4	1.02
Female applicants	4.5	6.2	1.31

The subset variable gender in Table 4 can be any variable (irrespective of whether it is an input of the model). Furthermore, if the value of 4.5 at development for females had been 5.9, then the model would be considered stable (for females, PAI = 6.2/5.9 = 1.05). In this situation, the model is considerably less accurate for females than males at development; however, this model's predictive ability has not deteriorated from development to review.

4. The Proportion of the PAI Due to a Shift in Distributions

A simple way for the distribution of the explanatory variables to change from development data to review data is for the distribution to shift using a constant (changing the mean). For example, the distribution of loan to value (LVR) in a property portfolio might increase if property values all fall in a recession. This section therefore investigates how much of a high PAI value is due to a shift in the mean of one or more explanatory variables. From Equation (2), the PAI only depends on the review data through the mean squared Mahalanobis distance $\overline{\mathrm{M}}_\mathrm{r}$. Furthermore, $\overline{\mathrm{M}}_\mathrm{r}$ can be decomposed as follows (to prove this result, write $(\rho_j - \overline{\epsilon})$ as $(\rho_j - \overline{\rho}) + (\overline{\rho} - \overline{\epsilon})$ and then multiply out the quadratic form in the defintion of $\overline{\mathrm{M}}_\mathrm{r}$ and simplify):

$$\overline{\mathrm{M}}_\mathrm{r} = \frac{1}{N}\sum_{j=1}^{N}(\rho_j - \overline{\epsilon})^T V_d^{-1}(\rho_j - \overline{\epsilon}) = \frac{1}{N}\sum_{j=1}^{N}(\rho_j - \overline{\rho})^T V_d^{-1}(\rho_j - \overline{\rho}) + (\overline{\rho} - \overline{\epsilon})^T V_d^{-1}(\overline{\rho} - \overline{\epsilon}) \quad (4)$$

where the right-hand side can be interpreted as the sum of the two components:

(1) $\frac{1}{N}\sum_{j=1}^{N}(\rho_j - \overline{\rho})^T V_d^{-1}(\rho_j - \overline{\rho})$, the mean squared Mahalanobis distance of the review data from the mean in the review data;
(2) $(\overline{\rho} - \overline{\epsilon})^T V_d^{-1}(\overline{\rho} - \overline{\epsilon})$, the squared Mahalanobis distance between the mean of the review data and the mean of the development data.

Interpretability is enhanced if the second component is compared to $\overline{\mathrm{M}}_\mathrm{r} - \overline{\mathrm{M}}_\mathrm{d}$ instead of comparing the second component to the first component (or their sum, $\overline{\mathrm{M}}_\mathrm{r}$) because, from Equation (2), it is the amount that $\overline{\mathrm{M}}_\mathrm{r}$ exceeds $\overline{\mathrm{M}}_\mathrm{d}$ that produces a large PAI. We therefore

define the contribution to the PAI due to the difference in the means of the explanatory variables at review ($\overline{\rho}$) compared to at development ($\overline{\epsilon}$) as:

$$S = \frac{(\overline{\rho} - \overline{\epsilon})^T V_d^{-1} (\overline{\rho} - \overline{\epsilon})}{\overline{\mathrm{M}}_\mathrm{r} - \overline{\mathrm{M}}_\mathrm{d}} \tag{5}$$

Note that the two components in Equation (4) are both quadratic forms and thus cannot be negative; therefore, S varies from 0 (when the second component equals 0; $\overline{\rho} = \overline{\epsilon}$) to 1 (when the difference in the means of the explanatory variables from development and review explains all of the amount the PAI exceeds 1). Furthermore, S only applies when the PAI exceeds 1 (i.e., $\overline{\mathrm{M}}_\mathrm{r} > \overline{\mathrm{M}}_\mathrm{d}$): if $PAI \leq 1$, then the model does not perform less accurately on the review data than on the development data, therefore there is no need to explore how much of the PAI is due to a shift in the means of the explanatory variables.

An alternative equation for S is obtained using (see Appendix C for a derivation):

$$S = \frac{PAI_0 - PAI_t}{PAI_0 - 1} \tag{6}$$

where PAI_0 is the Prediction Accuracy Index for the original review data ρ_j and PAI_t is the Prediction Accuracy Index using the transformed review data $\rho'_j = \rho_j - (\overline{\rho} - \overline{\epsilon})$, therefore the distribution of the review data remains intact other than shifting the mean to coincide with the mean of the development data. Equation (6) has a minor advantage over Equation (5) in that any software that can calculate the PAI can be used to calculate S without the need to calculate any Mahalanobis distances.

For example, for review data R1 with the additional observation $\rho_{51} = (6,\ 1,\ 1)^\mathrm{T}$, $PAI = 1.31$, and $S = 0.02$. This extra observation changes the mean of the explanatory variables from $\overline{\epsilon} = (3.12,\ 2.08,\ 1.44)^T$ to $\overline{\rho} = (3.18,\ 2.06,\ 1.43)^T$; however, transforming the review data to have the same mean as the development data only reduces the PAI by 0.006. Hence, the change in the mean of the variables contributes to only $S = 0.006/0.31 = 2\%$ of the 0.31 the PAI exceeds the baseline value of 1. As expected from the discussion in Section 3, the PAI in this example is not due to a shift in the means of the explanatory variables but due to the single outlier in the review data.

For the review data R2 where $PAI = 1.58$, the means of the three explanatory variables in the review data are $\overline{\rho} = (3.82,\ 2.02,\ 1.44)^T$, therefore the difference relative to development is $\overline{\rho} - \overline{\epsilon} = (0.70, -0.06,\ 0.00)^T$. When subtracting this vector from each of the review observations ρ_j (for example, $\rho_1 = (1,\ 1,\ 1)^\mathrm{T}$, it is transformed to $\rho'_1 = (0.3, 1.06,\ 1)^T$), and produces a value of $PAI = 1.13$ for the transformed review data. Hence, from Equation (6), the proportion of the excess PAI due exclusively to a change in the means of the explanatory variables is (1.58–1.13)/(1.58–1) = 0.77. That is, 77% of the amount the PAI of 1.58 exceeds the value of 1 (when the model performs equally accurately on review and development data) is due to the shift in the mean of the review data relative to the development data.

The difference in means $\overline{\rho} - \overline{\epsilon} = (0.70, -0.06,\ 0.00)^T$ suggests that this is primarily due to an increase in the mean of the first explanatory variable from 3.12 to 3.82; however, since the explanatory variables can be measured using different units, it is recommended that the vector $\overline{\rho} - \overline{\epsilon}$ is standardised by dividing by the corresponding standard deviations. For the development data in Table 1, the standard deviations of the explanatory variables are 1.51, 0.78, and 0.50; therefore, the standardised difference in means are, respectively, $0.70/1.51 = 0.46$, $-0.06/0.78 = -0.08$, and $0.00/0.50 = 0$.

The conclusion that the high value of the PAI for review data R2 is primarily due to a shift in the mean of the first explanatory variable is also evident by applying Equation (6) after transforming, so that the mean of only one explanatory variable is changed. For example, if the review data ρ_j is transformed to $\rho'_j = \rho_j - (0.70,\ 0,\ 0)$, so that the mean for the first explanatory variable a equals the mean in the development data but the mean of the two other explanatory variables are not changed, then the PAI becomes

1.14. Hence, the contribution from a shift in the mean of the first explanatory variable is $S_1 = (1.58 - 1.14)/(1.58 - 1) = 0.75$. That is, the shift in the mean of the first explanatory variable alone explains 75% of the excess PAI. Alternatively, if the mean of the second explanatory variable is shifted by transforming the review data ρ_j to $\rho'_j = \rho_j - (0, -0.06, 0)^T$, then the PAI equals 1.47 and $S_2 = 0.19$. For the third explanatory variable, the means in development and review are equal, therefore $S_3 = 0$.

5. The Contributions of Explanatory Variables to the PAI

Taplin and Hunt (2019) recommended using the univariate PAI to examine which variables contribute to the high values of the multivariate PAI; however, this does not always produce useful insights. For example, with review data R2 (PAI = 1.58), the univariate PAI for the three variables a, b, and c are, respectively, 1.09, 0.99, and 1.00. In all cases these are less than 1.1 and thus are interpreted by Taplin and Hunt (2019) as indicating no significant deterioration. Thus, the univariate PAI fails to diagnose the large PAI value because of a change in the distribution of the first variable a (as discussed in the previous section).

One reason for the univariate PAI not being informative is that models typically include many variables, but the univariate PAI summarises the stability of a model as if the model only contains one variable. This single variable model is most likely very different to the model of interest, and hence less relevant to the multivariate PAI. We therefore recommend calculating the PAI if one variable is removed from the model variables rather than the PAI that includes one variable because removing one variable is a smaller change than removing all but one variable. There is no need to re-estimate the model: the PAI is intended to be calculated using the data without a specific variable.

For the review data R2 (Table 2), the PAI is 1.01 if the first variable is removed and calculated on the remaining variables b and c. Compared to the original value of $PAI = 1.58$, this represents a decrease of $(1.58 - 1.01)/(1.58 - 1) = 98\%$ of the amount the original PAI exceeded 1. Thus, unlike the univariate PAI of 1.09, comparing the multivariate PAI with and without the first variable a suggests that the high PAI is due to the variable a. When the multivariate PAI is calculated without the second variable b, the PAI equals 1.09, and it equals 1.75 if the third variable c is removed. When the second variable is removed, the PAI is small because each of the possible combinations for the other two variables appears in the development data (therefore there is no large extrapolation to predict any of the review data). When the third variable is removed, the PAI is very high because the high correlation between a and b in the development data suggests that many observations (e.g., the six observations with $a = 5$ and $b = 2$) are outliers relative to the development data.

We can draw two conclusions from these examples. First, rather than following the advice in Taplin and Hunt (2019) to use the univariate PAI to investigate which explanatory variables contribute to a large PAI, we recommend calculating the PAI after removing a variable (and calculating the PAI based on all the other variables). In some cases, it may be appropriate to remove more than one variable from the model. Examples include removing all the dummy variables for a categorical variable (see Section 6.1), removing both the variable and its square and then calculating the PAI using the remaining variables for a quadratic relationship, and removing all cross-terms for interaction effects.

Second, we note that the PAI can change considerably depending on which explanatory variables are included; therefore, it may be worth including variables not in the model when calculating the PAI. This advice is consistent with Taplin and Hunt (2019, p. 10): "... we recommend calculating the MPAI using only the variables in the model or using these and a few other variables considered important." For example, suppose the development data (Table 1) produced a model with only b and c as explanatory variables (the first variable a was not considered to have predictive power). Then, for the variables in this model, the PAI for review data R1 is 1.01, therefore the model is stable (the predictive accuracy of the model at review is similar to the predictive accuracy at development). However, if the variable a was considered during the development of the model, its exclusion from the model could

equivalent be considered as a model in which this variable is included but with a coefficient equal to 0 (or insignificantly different to 0). From this perspective, the PAI of 1.31 (including all three variables) is relevant in the sense that the coefficient of (approximately) 0 for variable *a* might not provide accurate predictions for the outlier observation of $\rho_{51} = (6, 1, 1)^{\mathrm{T}}$ in the review data. That is, when the entire model development is considered (rather than just the variables included in the final model), there is a compelling argument that the PAI should be calculated using more variables than just those in the final model. Thus, we suggest that the PAI should also be calculated using all variables considered for inclusion in a model. This is important because the accuracy of predictions from the final model depends not only on the coefficients for variables in the model but also on the choice of which variables are included in the final model. Model stability is arguably relevant for the entire modelling process (not just the final model).

6. Discussion

Both the use of categorical variables and the use of econometric variables in credit models deserve further discussion. Categorical variables are commonly used, highlight some of the conclusions above, and are sometimes constructed from numerical variables. For example, a numeric variable such as an applicant's age (in years) might be converted into a categorical variable with several categories (e.g., young 18–30; middle-aged 31–50; and old 51+). Econometric variables present challenges to model stability.

6.1. Categorical Variables

First, when considering the impact of a categorical variable on the PAI, it may be more logical to exclude all the dummy variables for that categorical variable rather than exclude them one at a time. Second, not only is a shift in the mean of a categorical variable difficult to define, but a reasonable definition involving the proportion of observations in each category amounts to describing all the possible changes in the distribution of that variable. Thus, examining shifts in the mean of a categorical variable is equivalent to examining the effect of excluding the variable.

Third, many modelling techniques and practices involve converting numeric variables into categorical variables. Examples of this include transforming variables prior to a logistic regression and by using the modelling approach itself, such as a classification tree. For example, the three age groups (young 18–30; middle-aged 31–50; and old 51+) might be the variable in the final model while the age (in years: 18, 19, 20, ...) is available in the development data. In these situations, it may be prudent to calculate the PAI using the original numeric variable instead of the categories. This will highlight whether there has been a shift in the distribution of the original age distribution that is not evident from the categories (for example, if the old applicants in the development data were all younger than 55 years old but in the review data they are all over 55 years old).

In the case where a categorical variable (e.g., industry codes) is formed by combining categories, it may be prudent to calculate the PAI using the original (larger) number of categories. This is because changes in the distribution of a categorical variable might be hidden by the combination of a larger number of categories into a smaller number of categories (just as combining many numeric variables into the same category can). It is of interest to investigate the stability of the whole modelling process, not just the stability of the variables and their form in the final model.

6.2. Econometric Variables

Credit models, such as probability of default (PD) models, may contain not only explanatory variables defined by the characteristics of a borrower but also characteristics of economic conditions. For example, under IFRS, nine models use econometric variables such as interest rates or unemployment rates (or recent changes in these rates) to predict future default rates. This enables models to capture changes in default rates through the economic cycle, with the expectation that these will provide more accurate predictions

using the current (or forecast future) economic condition. This is justified by the expectation that response variables such as default rates will be influenced by economic conditions.

However, these econometric variables can be problematic for many reasons. In the context of model stability, econometric variables are likely to take the same values in the review data despite varying in development data. For example, when examining if a PD model is still fit-for-purpose today, model reviewers may examine the current portfolio, but all these observations share the same value as the econometric variables being observed at the same point in time. This distribution with no or little variation will look markedly different to the distribution in the development data (which presumably shows considerable variation, or the econometric variables are unlikely to be significant predictors). Indeed, it would be surprising if the distribution of an econometric variable at review would be similar to the distribution at development. While the PAI might be less problematic than statistics such as the PSI because the PAI measures the extent of extrapolation rather than any differences in the distributions, the PAI is still likely to be high unless the value of the econometric variables at review is near the middle of the distribution at development.

For example, suppose that the three variables in the development data (Table 1) are all econometric variables and we wish to compare the distribution of these variables at review to this distribution at development. If the review data are observed at the most recent time period, then it is likely all these observations have the same value for the econometric variables. We can therefore ask the question: for which values of these econometric variables will the PAI be greater than 1.1? Since the average squared Mahalanobis distance at development is 2.0, from Equation (2), the squared Mahalanobis distance will have to be less than 2.3 if the PAI is to be less than 1.1. This only occurs for three values of the explanatory variables: $\rho = (3, 2, 1)^{\mathrm{T}}$, $(3, 2, 2)^{\mathrm{T}}$, and $(3, 2, 2)^{\mathrm{T}}$, where the squared Mahalanobis distances are 1.0, 1.4, and 1.8, respectively (Table 3). These occur a total of 16 out of 50 times in the development data; therefore, there is only a 32% chance that the PAI will be green (<1.1) even if the review data are selected from the distribution of the development data. All other values for the explanatory variables at review result in a PAI greater than 1.1. For the PAI to be less than 1.5, the squared Mahalanobis distance must be less than 3.5, which adds the observations $\rho = (5, 3, 1)^{\mathrm{T}}$ and $(1, 1, 2)^{\mathrm{T}}$, which occur six times in the development data. Hence, only 44% of the observations in the review data result in a PAI less than 1.5. Thus, even if the value of the explanatory variables at review were randomly selected from the distribution in the development data, there is a 56% chance the PAI is red (>1.5) and only a 32% chance it is green (<1.2).

One solution to this characteristic of the PAI (that is, to be large for models that include econometric variables) is to calculate the PAI without econometric variables (only using the non-econometric variables). Another is to change the cut-offs of >1.1 (amber) and >1.3 (red), recommended by Taplin and Hunt (2019) when econometric variables are included. We do not support these modifications. Instead, we discuss why the large PAI might accurately reflect the instability of the model and how this may be a characteristic of inappropriate modelling.

One reason why the PAI can diagnose models with econometric variables as being unstable is due to an over-confidence in the accuracy of these models. This can be due to a phenomenon referred to as pseudo-replication, which occurs when observations are treated as being statistically independent when they are not. For example, default rates for accounts: if the same customer owns several accounts, then it is likely that they default together (not independently). This is because, if a customer is in default on one account, it is likely they are (or will be determined by a bank to be) in default for all their loans. When modelling with econometric variables, it is equally important to recognise that all observations measured at the same point in time will not be independent: when measured at the same point in time, they are likely to have several characteristics in common, such as unobserved economic conditions. Simple logistic regression models will not capture this dependence structure in the data and will consequently over-estimate the precision

of regression coefficients. For example, consider the development data in Table 1 with 50 observations across 14 combinations of the three explanatory variables. If all these variables are econometric, it is likely that these data result from 14 points in time and it is more realistic, in terms of the sampling of the econometric variables, to view the sample size as closer to 14 than to 50. In practical applications, this discrepancy in sample sizes is likely to be much larger (with thousands or tens of thousands of observations across only dozens of points in time). Hurlbert (1984) discussed the problem of pseudoreplication in ecology, while a more recent discussion of the phenomenon in business can be found in Petersen (2009).

This problem of pseudoreplication can be addressed during model development (for example, using random effects models or using robust standard errors advocated by Petersen (2009)), albeit several consequences. First, fewer (if any) econometric variables will be statistically significant and included in a final model. This is problematic for model developers who are required to follow IFRS 9 (which implies that these variables must be included). Second, the higher standard deviation for the estimated parameters of the econometric variables will change the matrix V in Equation (1) for the PAI, and this will effectively downweight the impact of the econometric variables on the PAI relative to other explanatory variables. Essentially, with regard to Equation (2) for the PAI, distances involving econometric variables will become lower. Thus, correctly modelling the development data to avoid pseudoreplication will not only produce models that more accurately capture econometric variables but at least partially correct the interpretation of the PAI through the use of a more realistic variance–covariance matrix V for the estimated model parameters. That is, a model that has a high PAI due to econometric variables might identify inadequacies in the modelling of the development data. Rather than excluding these econometric variables from the calculation of the PAI, it might be prudent to examine whether the model development was inappropriate due to pseudoreplication.

A quick, simple way to avoid pseudoreplication with econometric variables is to develop a PD model with a two-stage process. First, build a model with non-econometric variables (possibly with dummy variables for different points in time to capture econometric effects). Then, aggregate data to overall default rates at each point in time and model the overall default rate at each point in time using econometric variables and the average predicted PD at that time as a covariate. Note that, if the development data consists of 1000 loans at each of 20 points in time, then the first model will use 20,000 observations while the second will use 20 observations. This approach may be inappropriately simplistic when building a model but might be a quick approach for model reviewers who do not wish to develop a model but rather explore whether pseudoreplication exists in the development data. This is not an unreasonable investigation for a model review, especially if it is suspected that the model development process ignored the presence of pseudoreplication.

However, the possibility that models containing econometric variables are intrinsically unstable cannot be ignored. For example, at the time of writing (early 2023), interest rates in most countries have increased dramatically from economic stimulus conditions following COVID-19; historical data showing such a change in economic conditions are so rare that any model developed on historical data is likely to demonstrate low stability.

7. Conclusions

The Prediction Accuracy Index (PAI) is a useful statistic to detect when the stability of a model has deteriorated; however, the literature provides little guidance on how to investigate reasons for a lack of model stability when the PAI is high. In particular, the suggestion by Taplin and Hunt (2019) to use the univariate PAI to examine which variable(s) contribute to a high multivariate PAI appears simplistic and problematic because it essentially ignores all the other explanatory variables.

This paper has explored the properties of the multivariate PAI, which has led to recommended approaches to examine if a large PAI is due to individual observations, individual explanatory variables, or a shift in the mean of explanatory variables. This

includes the case when several explanatory variables are closely related (such as multiple dummy variables created from one categorical variable). This has several implications for how models should be reviewed, especially when the value of the PAI is high. An important instance of this is when, following IFRS 9 for provisioning, econometric variables are explicitly included in model development. This pracitce is very ambitious (especially compared to the Basel Accord for capital) and may lead to models that fail stability at review due to inadequate modelling practices. This has important implications not only for model developers but also for model reviewers, auditors, regulators, and standard setters.

Funding: This research received no external funding.

Data Availability Statement: The data used are provided in tables within the paper.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Equation (2) follows from the definition of the PAI in Equation (1), the definition of the squared Mahalanobis distance in Equation (3), and the identity for the inverse of a partitioned matrix:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BECA^{-1} & -A^{-1}BE \\ -ECA^{-1} & E \end{bmatrix}$$

where $E = \left(D - CA^{-1}B\right)^{-1}$.

The above identity can be verified by multiplying the matrix by its inverse and simplifying to obtain the identity matrix. Equation (2) then follows directly (after tedious algebra) from Equation (1) after partitioning V (the variance–covariance matrix of the estimated regression coefficients), separating out the intercept from the other explanatory variables (therefore A has dimensions 1 by 1).

The algebra is simplified if we use the fact that the PAI is invariant to the parameterisation of the model; therefore, we choose the parameterisation where each explanatory variable (except the intercept) is mean centred (so that the mean of each explanatory variable equals 0). That is, the sum of the entries in each column of X equals 0 except for the first column (for the intercept), where the sum equals n (the sample size at development). This changes the intercept from the prediction of an observation when all the explanatory variables equal 0 to the prediction when each explanatory variable takes the value equal to the mean of the explanatory variable in the development data. It then follows that the matrix $X^T X$ has the form of a partitioned matrix with entries given by the following:

- Top left entry (A) is n, the sample size at development;
- The off-diagonal entries (B and C) are row and column vectors that contain zeroes;
- Bottom-right entries (D) equal n times the variance–covariance matrix of the explanatory variables (ijth element equal to $\sum_{k=1}^{n}(X_{ik}X_{jk})$.

That is, the matrix $X^T X/n$ is block diagonal with the top-left entry (A) equal to 1, the off-diagonal entries (B and C) are equal to vectors of zeroes, and the bottom right entries (D) are equal to the k by k matrix containing the variance–covariance matrix of the explanatory variables at development.

It then follows that, for any vector v (with the first entry equal to 1),

$$v^T\left(X^T X\right)^{-1} v = v^T\left(X^T X/n\right)^{-1} v/n = (1 + M_v)/n \tag{A1}$$

where M_v is the squared Mahalanobis distance of the vector of the explanatory variables (v without the first value of 1) relative to the variance–covariance matrix of the explanatory variables at development.

Equation (2) follows by applying Equation (A1) to the numerator and denominator of Equation (1).

Appendix B

Each of the properties of I_k are proved as follows:

(a) I_k can be both positive or negative because the squared Mahalanobis distance of observations can be either higher or lower than the average in the development data;

(b) The average of the I_k is 0 because the average of the $PAI_{(-k)}$ equals PAI_0 (to see this, note that each of the $PAI_{(-k)}$ is itself a mean of a set of numbers after leaving out one value, therefore each value in this set is omitted exactly once);

(c) By definition of I_k, when $I_k = 0$, we have $PAI_0 = PAI_{(-k)}$;

(d) By definition of I_k, when $I_k = PAI_0 - 1$, we have $PAI_0 - PAI_{(-k)} = PAI_0 - 1$, therefore $PAI_{(-k)} = 1$;

(e) The identity $I_k = PAI_0 - PAI_{(-k)} = \frac{m_k - \overline{M}_r}{N\left(1+\overline{M}_d\right)}$ follows from Equation (2) to define the PAI in terms of the average of the squared Mahalanobis distances of the review data and the identity for the mean of N observations $\overline{y}$ in terms of the kth observation y_k and the average of the other $(k-1)$ observations $\overline{y}_{(-k)}$:

$$\overline{y} = \frac{\overline{y} + (N-1)\overline{y}_{(-k)}}{N}$$

Appendix C

The equivalence of Equations (5) and (6) is proved by using Equation (2) to define PAI_0 and PAI_t in terms of average squared Mahalanobis distances. Then, by using Equation (4) and cancelling common terms, it follows that

$$PAI_0 - PAI_t = \frac{(\overline{\rho} - \overline{\epsilon})^T V_d^{-1} (\overline{\rho} - \overline{\epsilon})}{1 + \overline{M}_d} \tag{A2}$$

Furthermore, from Equation (4)

$$PAI_0 - 1 = \frac{1 + \overline{M}_r}{1 + \overline{M}_d} = \frac{\overline{M}_r - \overline{M}_d}{1 + \overline{M}_d} \tag{A3}$$

and taking the ratio (Equation (A2) divided by Equation (A3)) yields

$$\frac{PAI_0 - PAI_t}{PAI_0 - 1} = \frac{(\overline{\rho} - \overline{\epsilon})^T V_d^{-1} (\overline{\rho} - \overline{\epsilon})}{\overline{M}_r - \overline{M}_d}$$

proving the equality of the two expressions for S given by Equations (5) and (6).

References

Basel Committee on Banking Supervision. 2006. Basel II: International Convergence of Capital Measurement and Capital Standards, A Revised Framework—Comprehensive Version. Bank for International Settlements. Available online: https://www.bis.org/publ/bcbs128.htm (accessed on 4 February 2018).

Becker, Aneta, and Jarosław Becker. 2021. Dataset shift assessment measures in monitoring predictive models. *Procedia Computer Science* 192: 3391–402. [CrossRef]

Giudici, Paolo, and Emanuela Raffinetti. 2021. Shapley-Lorenz eXplainable Artificial Intelligence. *Expert Systems with Applications* 167: 114104. [CrossRef]

Hurlbert, Stuart H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54: 187–211. [CrossRef]

International Accounting Standards Board. 2014. IFRS 9—Financial Instruments. Available online: http://www.aasb.gov.au/admin/file/content105/c9/AASB9_12-14.pdf (accessed on 4 February 2018).

Kruger, Chamay, Willem Daniel Schutte, and Tanja Verster. 2021. Using Model Performance to Assess the Representativeness of Data for Model Development and Calibration in Financial Institutions. *Risks* 9: 204. [CrossRef]

Lundberg, Scott M., and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. Paper Presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, December 4–9. [CrossRef]
Mahalanobis, Prasanta Chandra. 1936. On the Generalized Distance in Statistics. *Proceedings of the National Institute of Science of India* 2: 49–55.
Petersen, Mitchell A. 2009. Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches. *Review of Financial Studies* 22: 435–80. [CrossRef]
Taplin, Ross, and Clive Hunt. 2019. The Population Accuracy Index: A New Measure of Population Stability for Model Monitoring. *Risks* 7: 53. [CrossRef]