DEUTSCHE
BUNDESBANK
EUROSYSTEM

Discussion Paper

Deutsche Bundesbank
No 04/2022

Calibration alternatives to logistic regression and their potential for transferring the dispersion of discriminatory power into uncertainties of probabilities of default

Jan Henrik Wosnitza

Deutsche Bundesbank, Wilhelm-Epstein-Straße 14, 60431 Frankfurt am Main,
Postfach 10 06 02, 60006 Frankfurt am Main

Tel +49 69 9566-0

Please address all orders in writing to: Deutsche Bundesbank,
Press and Public Relations Division, at the above address or via fax +49 69 9566-3077

Internet http://www.bundesbank.de

ISBN 978–3–95729–870–6
ISSN 2749–2958

Non-technical summary

Research Question

The transformation of credit scores into probabilities of default (PDs) plays an important role in credit risk estimation. The logistic regression, whose logit is a linear function of the credit score, has developed into a standard calibration approach. With the advent of machine learning techniques in the discriminatory phase of credit risk models, however, this standard calibration approach is currently under scrutiny again.

Previous literature has converted the calibration problem into the task of modelling the cumulative accuracy profile (CAP) without any loss of generality. The main objective of this paper is twofold. First, we compare the performance of four calibration approaches on a real-world data set. Second, we explore whether the approach, based on modelling the CAP, provides the opportunity to derive uncertainties of PD estimates stemming from the statistical dispersion of the discriminatory power.

Contribution

The main contribution of this paper is threefold. First, we suggest two new one-parametric families of differentiable functions as candidates for modelling the CAP based on the maximum entropy principle. In so doing, we extend the underlying calibration approach. Second, we benchmark the calibration approach, based on modelling the CAP, against the linear logistic regression on a real-world data set. Third, we develop an approach in order to transfer the statistical dispersion of the discriminatory power into a margin of conservatism for the general estimation error of the PD. In this context, we also provide an alternative representation for the variance of the area under the receiver operating characteristic to the one recently proposed in the literature.

Results

We find that one of the new one-parametric families outperforms the linear logistic regression on the data set in question. In view of the fact that this outperformance is only statistically significant for medium training sample sizes, it is worth noting that even small improvements in calibration performance may translate into significant competitive advantages and into relevant refinements of regulatory capital quantification within sizeable portfolios. Furthermore, the uncertainties of PD estimates, stemming from the statistical dispersion of the discriminatory power, as a function of the sample size fluctuate within a reasonable range.

Nichttechnische Zusammenfassung

Fragestellung

Der Überführung von *Credit Scores* in Ausfallwahrscheinlichkeiten kommt bei der Kreditrisikomessung eine zentrale Bedeutung zu. Die logistische Regression, deren *Logit* eine lineare Funktion des *Credit Scores* ist, hat sich zu einem Standardkalibrierungsverfahren entwickelt. Der vermehrte Einsatz von maschinellen Lernverfahren zur Ausfallprognose von Kreditnehmern stellt dieses Standardkalibrierungsverfahren aber aktuell wieder auf den Prüfstand.

Vorhergehenden Beiträgen in der Literatur ist es gelungen, das Problem der Kalibrierung ganz allgemein in ein Problem der Modellierung des *Cumulative Accuracy Profiles* (CAPs) zu überführen. Der vorliegende Aufsatz verfolgt ein zweifaches Ziel: Zum einen vergleichen wir die Leistungsfähigkeit von vier Kalibrierungsverfahren auf Basis realer Daten. Zum anderen gehen wir der Frage nach, ob der Kalibrierungsansatz auf Basis der Modellierung des CAPs die Möglichkeit bietet, Unsicherheiten der Ausfallwahrscheinlichkeitsschätzung aus der Schwankung der Trennschärfe abzuleiten.

Beitrag

Der Beitrag des vorliegenden Aufsatzes lässt sich in drei Punkte gliedern. Zunächst schlagen wir auf Basis des Maximum Entropie Prinzips zwei neue Familien von differenzierbaren Funktionen zur Modellierung des CAPs vor. Auf diese Weise erweitern wir den zugrundeliegenden Kalibrierungsansatz. Des Weiteren vergleichen wir den auf der Modellierung des CAPs beruhenden Kalibrierungsansatz auf Basis echter Daten mit der logistischen Regression. Außerdem entwickeln wir ein Verfahren, um die statistische Unsicherheit der Trennschärfe in eine Sicherheitsspanne für den allgemeinen Schätzfehler der Ausfallwahrscheinlichkeit zu überführen. In diesem Zusammenhang geben wir auch eine alternative Darstellung für die kürzlich vorgeschlagene Gleichung zur Berechnung der Varianz der *Area under Receiver Operating Characteristic* an.

Ergebnisse

Eine der neuen Familien von differenzierbaren Funktionen führt zu einer besseren Kalibrierungsgüte als die logistische Regression. Obwohl dieser Leistungsunterschied nur für mittelgroße Trainingsstichproben signifikant ist, können bei entsprechenden Portfoliogrößen selbst kleinere Verbesserungen in der Kalibrierungsgüte bedeutende Wettbewerbsvorteile darstellen. Schließlich nimmt die aus der statistischen Schwankung der Trennschärfe resultierenden Unsicherheit der Ausfallwahrscheinlichkeit als Funktion der Stichprobengröße plausible Werte an.

Deutsche Bundesbank Discussion Paper No 04/2022

Calibration alternatives to logistic regression and their potential for transferring the dispersion of discriminatory power into uncertainties of probabilities of default[1]

Jan Henrik Wosnitza
Deutsche Bundesbank

Abstract

The transformation of credit scores into probabilities of default plays an important role in credit risk estimation. The linear logistic regression has developed into a standard calibration approach in the banking sector. With the advent of machine learning techniques in the discriminatory phase of credit risk models, however, the standard calibration approach is currently under scrutiny again. In particular, the assumptions behind the linear logistic regression provide critics with a target. Previous literature has converted the calibration problem into a regression task without any loss of generality. In this paper, we draw on recent academic results in order to suggest two new one-parametric families of differentiable functions as candidates for this regression. The derivation of these two families of differentiable functions is based on the maximum entropy principle and, thus, they rely on a minimum number of assumptions. We compare the performance of four calibration approaches on a real-world data set and find that one of the new one-parametric families outperforms the linear logistic regression. Furthermore, we develop an approach in order to quantify the part of the general estimation error of probabilities of default that stems from the statistical dispersion of the discriminatory power.

Key words: Calibration, credit score, cumulative accuracy profile, logistic regression, margin of conservatism, probability of default.

JEL-Classification: G17, G21, G33

Abbreviations: Area under the receiver operating characteristic (AUROC); Cumulative accuracy profile (CAP); Cumulative distribution function (CDF); Internal rating based approach (IRBA); Linear logistic regression (LLR); Observed default frequency (ODF); Probability density function (PDF); Probability of default (PD).

[1] **Contact address:** Jan.Henrik.Wosnitza@bundesbank.de.
Acknowledgment: We express many thanks to Thomas Kick, an anonymous reviewer as well as to our colleagues Marco van der Burgt, Johannes Diermeier, Christine Fremdt, Raphael Koch, Carsten Merten, Alexander Paduch, Jürgen Prahl, Dirk Tasche, and Johannes Wächtler for providing valuable input at various levels. All remaining errors are solely our own responsibility. Furthermore, we could not have written this paper without being seconded to the Research Centre of Deutsche Bundesbank. Therefore, we also thank Jochen Mankart, Emanuel Mönch, and Christian Schumacher as well as Caroline Knaak, Walter Schauf, and Andreas Schneider. We apologize for any omissions.

1 Introduction

Lending activities are one of the major sources of risks for banks and, thus, the ability to estimate accurately credit risk is essential for them (Aussenegg, Resch, and Winkler 2011). The key variable in credit risk estimation is the probability of default (PD) (Lawrenz 2008). Banks estimate PDs, among other things, in order to support their loan decisions. On the one hand, lending to obligors destined to default can cause substantial losses to banks (Van Gestel, Baesens, Suykens, Espinoza, Baestaens, Vanthienen, and De Moor 2003). On the other hand, denying lending to financially sound obligors most likely means forgoing profitable investment opportunities.

The large number of loan applicants and obligors, in particular in the retail lending business, makes it necessary to rely on statistical methods rather than on human discretion in order to estimate PDs (Huang, Chen, and Wang 2007; Khandani, Kim, and Lo 2010). The model-based estimation of PDs allows banks to assess the credit quality of (potential) obligors at much lower cost (Huang et al. 2007) and it puts them in the position to allocate their rare and expensive human resources to questionable cases of highest importance.

In addition to banks' internal desire for accurate PD estimates, the Basel capital accord also stimulates their usage. Under the internal rating based approach (IRBA), authorised banks quantify their regulatory capital based on own PD estimates (Aussenegg et al. 2011; Van der Burgt 2019; Van Gestel, Baesens, Van Dijcke, Suykens, Garcia, and Alderweireld 2005).

The model-based estimation of PDs usually involves two steps. First, banks condense all critical information about the creditworthiness of obligors into univariate credit scores (Khandani et al. 2010). These credit scores discriminate between obligors of low and high credit quality (i.e. discrimination). Throughout this paper, we follow the convention that low values of credit scores tend to indicate high risk of default and vice versa (cf. e.g. Van der Burgt (2019); Tasche (2010)). Second, banks transform the credit scores into PDs (i.e. calibration).

Banks have a long history of applying data analysis methods to the estimation of credit risk. However, recent technological developments have facilitated the application of machine learning techniques to the calculation of credit scores (Bonini and Caivano 2018). Machine learning techniques use sophisticated mathematical algorithms in order to identify risk drivers providing information about the credit quality of obligors. Then, the machine learning techniques link these risk drivers to the binary default variable through a functional relationship in the training phase and, in so doing, discriminate between defaults and non-defaults (Barboza, Kimura, and Altman 2017; Bequé, Coussement, Gayler, and Lessmann 2017). The trained machine learning model can finally be applied in order to predict the binary default variables of new data over the specified risk horizon.

In their role as principal banks, many financial institutions capture huge volumes of information (e.g. on the payment behaviour of their customers) every day. The digitalisation of banks' business processes in conjunction with the proliferation of computers and mobile phones involves that data volumes increase at an ever-faster pace. Significantly improved storing capacities, vastly increased computational power, and ongoing developments of machine learning algorithms facilitate the implementation of machine learning techniques in order to process these huge amounts of data and, ultimately, to calculate more accurate credit scores (Bonini et al. 2018; European Banking Authority 2020; Hong Kong Monetary Authority and PricewaterhouseCoopers 2019).

As a consequence of its immense practical relevance, a great deal of research has been devoted to the application of different machine learning techniques for default prediction (e.g. Barboza et al. (2017); Butaru, Chen, Clark, Das, Lo, and Siddique (2016); Moscatelli, Parlapiano, Narizzano, and Viggiano (2020); Petropoulos, Siakoulis, Stavroulakis, and Vlachogiannakis (2020)). The general conclusion of these papers is that machine learning techniques outperform traditional models such as logistic regression which are widespread in the banking sector. Alonso and Carbó (2020) summarize that the gains in discriminatory power of machine learning techniques are very heterogeneous, reaching up to 20% measured by the area under the receiver operating characteristic (AUROC). The main reason for this superiority is that machine learning techniques capture the underlying nonlinearities in the relationship between risk drivers and binary default variables better than traditional models (Bazarbash 2019). Furthermore, machine learning techniques are often not subject to series of assumptions and, thus, they are less restrictive than traditional models (Barboza et al. 2017).

However, credit risk estimation does not end with the ordinal ranking of obligors from the most to the least prone to default based on credit scores. Credit risk estimation also involves linking credit scores with accurate PDs, which is referred to as calibration (Bequé et al. 2017; Fonseca and Lopes 2017). In fact, some machine learning techniques (e.g. support vector machines) produce credit scores on arbitrary scales, which disqualifies them from probabilistic interpretations (Bequé et al. 2017; Böken 2021; Caruana and Niculescu-Mizil 2006; Moro, Härdle, and Schäfer 2017). Other machine learning techniques provide predictions in the interval from zero to one, which in principle allow for an interpretation as PD. For example, Kruppa, Schwarz, Arminger, and Ziegler (2013) apply random forests, k-nearest neighbours, and bagged k-nearest neighbours in order to directly estimate PDs of a large data set of short-termed instalment credits. However, Bequé et al. (2017) find that the calibration of direct PD estimates consistently improves the performance in terms of Brier Score (without hurting discriminatory power). Similarly, Leathart, Frank, Holmes, and Pfahringer (2017) point out that direct PD estimates often show poor calibration performance.

In this paper, we follow Tasche (2010) in assuming a strictly monotonically increasing relationship between credit scores and credit quality. An important feature of the discriminatory power measured by the AUROC is its invariance to any strictly monotonic transformation of credit scores (Moro et al. 2017; Van der Burgt 2020). In order to maintain the AUROC, implied by the pairs of credit scores and binary default labels, the mapping of credit scores to PDs should also be strictly monotonic (i.e. rank-order preserving) (Bequé et al. 2017).

There are two fundamental calibration approaches (Lawrenz 2008). The first is to discretize the range of credit scores into a certain number of buckets and, in so doing, to pool obligors with similar credit scores. All the obligors falling into the same bucket receive the same PD, which is usually derived from the observed default frequency (ODF)[2] of the bucket (Nehrebecka 2016). A widely used methodology for grouping obligors and estimating PDs for each group is isotonic regression (Zadrozny and Elkan 2002). For example, Moro et al. (2017) successfully apply the isotonic regression in order to map scores produced by support vector machines into PDs. Similarly, Fonseca et al. (2017) provide empirical evidence in favour of the isotonic regression. However, bucket-based PD estimates by definition discard the individual PDs inside the buckets. Therefore, no risk differentiation within buckets is possible (Böken 2021).

Due to this drawback, we follow the second calibration approach and assign individual PDs to obligors throughout this paper (direct PD estimates). A widely accepted methodology for assigning individual probabilities across several disciplines is the logistic regression (Hosmer, Lemeshow, and Sturdivant 2013). In the banking sector, the linear logistic regression (LLR) (i.e. the logistic regression whose logit is a linear function of the credit score) is the most prevalent calibration methodology (Nehrebecka 2016; Bequé et al. 2017). This approach assumes a sigmoidal relationship between credit scores and PDs.

The assumption of a linear logit as well as the sigmoid shape impose restrictions, which can be unrealistic. For example, Bequé et al. (2017) demonstrate that relaxing the linearity constraint in the logit through penalized regression splines is especially suitable for calibrating classifier predictions. Similarly, Moro et al. (2017) call the linear relationship of the logit as a function of the score a major disadvantage of this calibration approach. Furthermore, they criticize the sigmoid form of the link function between PDs and credit scores as too restrictive. Hence, there is a certain demand for a more flexible calibration regime than the one that the (linear) logistic regression establishes.

[2] The ODF is defined as the number of obligors, which were in the bucket at the beginning of the period and ended up in default at the end of the period, divided by the total number of obligors that were in the bucket at the beginning of the period (Lawrenz 2008).

Falkenstein, Boral, and Carty (2000) introduce a very general calibration approach, which is based on modelling the cumulative accuracy profile (CAP)[3] by means of a differentiable concave function on the interval from zero to one. Tasche (2010) demonstrates that this approach converts the problem of calibration into a regression task without any loss of generality (cf. equation (5.2b) in Tasche (2010)). In this framework, the conditional PD is a function of the obligor's rank. Based on the results of Tasche (2010), we reason that the calibration approach based on the first derivative of the CAP is equivalent to the logistic regression if the theoretical assumptions of the logistic regression are met (i.e. if the credit scores of defaulted and non-defaulted obligors are drawn from two distributions of the exponential family, respectively). In this sense, the approach proposed by Falkenstein et al. (2000) generalises the logistic regression. Particularly against the background of machine learning algorithms entering increasingly the discriminatory phase of credit risk models, a more flexible calibration approach seems to be desirable.

Van der Burgt (2008) models the CAP by means of a one-parametric family of differentiable functions and, then, applies this approach to both synthetic portfolios and a real-life low default portfolio consisting of exposures to 86 sovereigns. The approach of Van der Burgt (2008) is also used and discussed by Agarwal and Taffler (2008); Nehrebecka (2016); Roengpitya and Nilla-Or (2011); Van der Burgt (2019). Tasche (2010) very carefully investigates the approach proposed by Van der Burgt (2008). On the one hand, he acknowledges that "the suggestion by Van der Burgt (2008) leads to a potentially quite useful modification of logit regression in the univariate case". On the other hand, he criticises: "In his paper, van der Burgt (2008) does not spend much time with explaining the why and how of his approach. It is tempting to guess that the approach was more driven by the results than by theoretical considerations." Paragraph 96 of the Guidelines on PD estimation, LGD estimation and treatment of defaulted assets (European Banking Authority 2017) amplifies this point of criticism by requiring financial institutions to "demonstrate that the theoretical assumptions of the probability model underlying the estimation methodology are met to a sufficient extent in practice […]".

The contributions of this paper with regard to calibration are threefold. First, we leverage the results of Tasche (2010) in order to demonstrate that the calibration approach based on modelling the CAP is equivalent to the logistic regression if the theoretical assumptions of the logistic regression are met. In this sense, the approach proposed by Falkenstein et al. (2000) generalises the logistic regression. Second, we justify the one-parametric family of differentiable functions proposed by Van der Burgt (2008). To this end, we use a special case of the result of Brunel (2019), according to which the maximum entropy principle conditional on a given AUROC and assuming independent defaults implies that the conditional PDs can be calculated via the LLR. Instead of the credit score, however, the rank of the obligors serves as explanatory variable

[3] Please note that the CAP is also known as Lorenz curve and power curve.

(Brunel 2019). Then, we empirically show that the approach of Van der Burgt (2008) is very similar to the entropy maximizing logistic regression on the ranks of the obligors for lower unconditional PDs and/or lower AUROCs. Similarity to the entropy maximizing LLR on the ranks of the obligors is theoretically appealing, because maximizing the entropy leads to the distribution of the binary default variable Y that carries the highest uncertainty and, thus, the fewest assumptions about the true distribution. Third, we benchmark the regression proposed by Van der Burgt (2008) and the regression approach leading to maximum entropy against the LLR on a large real-world data set. However, both regression approaches disregard the specific form of the empirical CAP. Therefore, we propose a modification of the regression approach leading to maximum entropy as a third one-parametric family and find that this modified approach outperforms the LLR. This finding supports the result of Bequé et al. (2017) according to which nonlinear but strictly monotonic transformations of credit scores can improve the calibration performance compared to the LLR.

Banking supervisors also take a great interest in the properties of banks' PD estimates. In particular, banking supervisors seek to avoid any underestimation of credit risk. In order to account for the natural uncertainty of PD estimates, Article 179 (1) (f) of the Corrigendum to regulation (EU) No 575/2013 on prudential requirements for credit institutions and investment firms (European Parliament and the Council of the European Union 2013) stipulates that "an institution shall add to its estimates a margin of conservatism that is related to the expected range of estimation errors".[4] Paragraph 43 (b) of the Guidelines on PD estimation, LGD estimation and treatment of defaulted assets (European Banking Authority 2017) further specifies that "the [margin of conservatism] for the general estimation error should reflect the dispersion of the distribution of the statistical estimator."

While the obligation of quantifying an adequate margin of conservatism for the general estimation error is clear, the regulation explicitly does not impose a specific methodology for this (cf. p. 18 in European Banking Authority (2017)). Furthermore, the academic literature on quantifying the general estimation error of PDs is still in its infancy. Lawrenz (2008) assesses the uncertainty of PD estimates by calculating four confidence intervals of the ODF. Blümke (2020) proposes a Bayesian approach, which generates a distribution of the PD estimate and, in so doing, provides an easy way to define a margin of conservatism. More precisely, he points out that "a more conservative estimate than the mean can be obtained by choosing a more conservative percentile of the posterior distribution". In the framework of logistic regression, Hosmer et

[4] The negative consequence of excessively conservative PDs is the loss of competitiveness at international level (Pfeuffer, Nagl, Fischer, and Rösch 2020). In order to solve this dilemma, paragraph 208 (c) of the Guidelines on PD estimation, LGD estimation and treatment of defaulted assets (European Banking Authority 2017) allows financial institutions to exclude the margin of conservatism from their PD estimates when used for internal purposes such as risk management and decision-making processes.

al. (2013) calculate confidence intervals for the logit (in Section 1.4), which translate into confidence intervals of the estimated PDs.

This paper contributes to this emerging strand of literature by exploring whether the calibration methodology proposed by Van der Burgt (2008) allows us to derive uncertainties of PDs resulting from the statistical dispersion of the discriminatory power as required by paragraph 140 (a) of the ECB guide to internal models (European Central Bank 2019). We employ this measure of PD uncertainty to synthetic data sets of pairs of credit scores and default labels.

The remainder of this paper proceeds as follows. Section 2 presents the theoretical background of four calibration approaches. In Section 3, we compare the performance of the four calibration approaches on a real-world data set. Section 4 explores whether the calibration approach of Van der Burgt (2008) provides the opportunity to quantify uncertainties of PD estimates stemming from the statistical dispersion of discriminatory power. Section 5 concludes by briefly summarizing the main results, discussing their implications, and offering suggestions for future research.

2 Calibration theory

This Section, first, introduces definitions and notation with regard to calibration. Second, it revisits a general framework that allows determining PDs conditional on credit scores via regression. Subsection 2.1 then demonstrates that the logistic regression is a special case of this general framework. In Subsection 2.2, we present three one-parametric families of differentiable functions that provide alternatives to logistic regression for calibration.

First, we revisit the one-parametric family introduced by Van der Burgt (2008). Second, we propose the one-parametric family that corresponds to the maximum entropy approach. The maximum entropy principle is well-established in finance (cf. Segoviano Basurto (2006)). It states that the probability distribution with the maximum entropy is the one that makes the fewest assumptions about the true distribution of data. Paragraph 96 of the Guidelines on PD estimation, LGD estimation and treatment of defaulted assets (European Banking Authority 2017) seems to draw attention to calibration approaches, relying on a minimum number of assumptions, by requiring banks to "demonstrate that the theoretical assumptions of the probability model underlying the estimation methodology are met to a sufficient extent." Furthermore, we justify the one-parametric family of differentiable functions proposed by Van der Burgt (2008) by empirically demonstrating its similarity to the entropy maximizing logistic regression on the ranks of the obligors (Brunel 2019) for lower unconditional PDs and/or lower AUROCs. Third, we modify the second one-parametric family in order to fit better the empirical CAP in question. Hence, the theoretical underpinnings of the first and third one-parametric families are less than that of the second one.

As a starting point, we assume the existence of a credit scoring model that maps a multidimensional input vector to a one-dimensional credit score (Van der Burgt 2020).[5] The credit score takes on real values on a continuous scale. Throughout this paper, we follow the convention that low values of credit scores tend to indicate high default risk and vice versa (e.g. Van der Burgt (2019); Tasche (2010)). Hence, banks can use credit scores in order to discriminate between obligors of low and high credit quality.

One main objective of this paper is to compare different approaches for transforming credit scores into PDs (i.e. different calibration approaches). The parameters of the calibration approaches are determined based on historical training data sets and, then, we apply the calibration approaches to new test data sets. The training and test data sets are of the form $\{s_k, y_k\}_{k=1}^{n_D+n_{ND}}$ and consist of pairs of credit scores $s_k \in \mathbb{R}$ and binary default labels $y_k \in \{0; 1\}$ of n_D defaulted and n_{ND} non-defaulted obligors. The binary default label y_k represents the state of default one observation period (usually one year) after the calculation of the credit score s_k. A credit default

[5] For example, the credit score could be a weighted sum of several financial ratios.

of obligor k is labelled as $y_k = 1$, whereas $y_k = 0$ indicates that obligor k has not (yet) fallen into default.

Let $F_D(s) = P(S \leq s|Y = 1)$ and $f_D(s) = P(S = s|Y = 1)$ [resp. $F_{ND}(s) = P(S \leq s|Y = 0)$ and $f_{ND}(s) = P(S = s|Y = 0)$] denote the cumulative distribution function (CDF) and the probability density function (PDF) of the defaults [resp. non-defaults]. Based on these notations and employing the law of total probability, we can express the CDF (i.e. $F_{All}(s)$) and the PDF (i.e. $f_{All}(s)$) of all obligors as follows:

$$\begin{aligned} F_{All}(s) &= P(Y = 1) \cdot F_D(s) + [1 - P(Y = 1)] \cdot F_{ND}(s), \\ f_{All}(s) &= P(Y = 1) \cdot f_D(s) + [1 - P(Y = 1)] \cdot f_{ND}(s). \end{aligned} \tag{1}$$

The objective of calibration is to estimate the PD given a credit score (i.e. $P(Y = 1|S = s)$). This conditional PD can be expressed in terms of the unconditional PD (i.e. $P(Y = 1)$) as well as the PDFs $f_D(s)$ and $f_{All}(s)$ by using the definition of conditional probabilities and the law of total probability:

$$\begin{aligned} P(Y = 1|S = s) &= \frac{P(Y = 1, S = s)}{P(S = s)} \\ &= \frac{P(Y = 1) \cdot P(S = s|Y = 1)}{P(Y = 1) \cdot P(S = s|Y = 1) + [1 - P(Y = 1)] \cdot P(S = s|Y = 0)} \\ &= \frac{P(Y = 1) \cdot f_D(s)}{P(Y = 1) \cdot f_D(s) + [1 - P(Y = 1)] \cdot f_{ND}(s)} \\ &= P(Y = 1) \cdot \frac{f_D(s)}{f_{All}(s)}. \end{aligned} \tag{2}$$

According to equation (2), we can directly calculate the conditional PD by estimating $P(Y = 1)$ as well as the PDFs $f_D(s)$ and either $f_{ND}(s)$ or $f_{All}(s)$. For example, we could estimate these PDFs via maximum likelihood estimation or kernel density estimation. However, both approaches can result in non-monotonic relationships between the credit score and PD which would violate the requirement of rank-order preservation enshrined in paragraph 99 of the Guidelines on PD estimation, LGD estimation and treatment of defaulted assets (European Banking Authority 2017). Furthermore, the former approach requires a-priori specification of potential probability distributions whose parameters are to be determined (Böken 2021). The latter approach can result in a downward biased estimation of the AUROC without manipulating the credit scores (Tasche 2010). Falkenstein et al. (2000) propose an alternative approach, which allows the calculation of the conditional PD through regression. The derivation of their approach starts with the last line of equation (2):

$$
\begin{aligned}
P(Y=1|S=s) &= P(Y=1)\cdot\frac{f_D(s)}{f_{All}(s)} \\
&\overset{(1)}{=} P(Y=1)\cdot\frac{dF_D(s)}{ds}\cdot\frac{1}{f_{All}(s)} \\
&\overset{(2)}{=} P(Y=1)\cdot\frac{dF_D(s)}{ds}\cdot\frac{dF_{All}^{-1}\left(F_{All}(s)\right)}{dF_{All}(s)} \\
&\overset{(3)}{=} P(Y=1)\cdot\frac{dF_D(s)}{ds}\cdot\frac{ds}{dF_{All}(s)} \\
&\overset{(4)}{=} P(Y=1)\cdot\frac{dF_D(s)}{dF_{All}(s)},
\end{aligned}
\tag{3}
$$

where we use

- (1): the definition of $f_D(s)$,
- (2): the derivative of the inverse function[6],
- (3): the definition of inverse functions, and
- (4): the chain rule

in order to arrive at Proposition 5.1 in Tasche (2010). This equation allows us to determine the conditional PD as the product of the unconditional PD and the first derivative of the CAP. Thus, no direct estimation of the PDFs $f_D(s)$ and either $f_{ND}(s)$ or $f_{All}(s)$ is necessary.

So far, we have not made any assumptions on the CDFs $F_D(s)$ and $F_{ND}(s)$ other than differentiability of $F_D(s)$ with respect to $F_{All}(s)$. As we will see in the following subsections, the only difference between the logistic regression and CAP regressions lies in the assumption about the quotient $\frac{f_D(s)}{f_{All}(s)}$ or, alternatively, about the derivative $\frac{dF_D(s)}{dF_{All}(s)}$.

[6] For the sake of completeness, we provide the derivative of the inverse function $\frac{dF_{All}^{-1}(F_{All}(s))}{dF_{All}(s)}$ in the following:

$$
\begin{aligned}
F_{All}^{-1}\left(F_{All}(s)\right) &= s \\
\Rightarrow \frac{d}{ds}F_{All}^{-1}\left(F_{All}(s)\right) &= \frac{d}{ds}s \\
\overset{(4)}{\Leftrightarrow} \frac{dF_{All}^{-1}\left(F_{All}(s)\right)}{dF_{All}(s)}\cdot\frac{dF_{All}(s)}{ds} &= 1 \\
\overset{(1)}{\Leftrightarrow} \frac{dF_{All}^{-1}\left(F_{All}(s)\right)}{dF_{All}(s)} &= \frac{1}{\frac{dF_{All}(s)}{ds}} = \frac{1}{f_{All}(s)}
\end{aligned}
$$

2.1 Calibration based on logistic regression

The LLR is a standard calibration approach. It passes real-valued credit scores through a sigmoid function in order to produce PDs in the range between zero and one. The distributional assumptions of the logistic regression are well known, for example, from the textbook of Bishop (2006) and have recently been revisited by Böken (2021). The purpose of this subsection is to demonstrate that equation (3) leads to the logistic regression if the corresponding assumptions hold.

The logistic regression starts from the premise that the PDFs of the credit scores of defaults ($Y = 1$) and non-defaults ($Y = 0$) are members of the exponential family, i.e. that the PDFs have the following form, respectively (Bishop 2006):

$$f_Y(s) = h(s) \cdot g\left(\vec{\beta}_Y\right) \cdot \exp\left\{\vec{\beta}_Y{}^T \cdot \vec{u}(s)\right\}, \tag{4}$$

where

- the vector $\vec{\beta}_Y$ includes the natural parameters,
- the function $g\left(\vec{\beta}_Y\right)$ can be interpreted as the coefficient ensuring normalisation of the distribution, and
- $\vec{u}(s)$ is some vector function of the credit score s.

Under this assumption, we can first determine $f_{All}(s)$ based on equation (1) and, second, calculate $P(Y = 1|S = s)$ by plugging $f_D(s)$ and $f_{All}(s)$ into the first line of equation (3):

$$P(Y = 1|S = s) = \frac{1}{1 + \exp\left\{-\left[\left(\vec{\beta}_D - \vec{\beta}_{ND}\right)^T \cdot \vec{u}(s) + \ln\left\{\frac{P(Y=1)}{1-P(Y=1)}\right\} + \ln\left\{\frac{g\left(\vec{\beta}_{ND}\right)}{g\left(\vec{\beta}_D\right)}\right\}\right]\right\}}. \tag{5}$$

The right-hand side of equation (5) obviously has the form of the logistic regression function. The argument of the exponential function multiplied by minus one is referred to as logit and is denoted by $l(\cdot)$. When banks use the logistic regression in order to estimate PDs, they assume in the vast majority of cases that the logit is a linear function of s.

2.2 Calibration based on CAP regressions

This subsection presents three calibration approaches that are based on modelling the empirical CAP. The CAP plays an important role in the measurement of discriminatory power of PD models (Engelmann, Hayden, and Tasche 2003; Falkenstein et al. 2000; Van der Burgt 2008, 2019). The construction of the CAP includes two steps. First, we arrange the obligors in ascending order of their credit scores (i.e. from obligors of low credit quality to obligors of high credit quality). Second, we plot the cumulative share of the defaults (i.e. the CDF of the defaults) against the cumulative share of all (i.e. CDF of all) (Engelmann et al. 2003; Falkenstein et al. 2000; Van der Burgt 2008, 2019, 2020). When we go from the lowest to the highest credit score, each default generates a vertical jump whereas any occurrence generates a horizontal move (Brunel 2019). Thus, the specific shape of the CAP depends on the respective portfolio of obligors and on the applied credit scoring methodology. Recent technological developments have facilitated the application of machine learning techniques in order to estimate credit scores (Bonini et al. 2018; European Banking Authority 2020; Hong Kong Monetary Authority et al. 2019). Consequently, the shapes of the CAPs will become more diverse in the future. Therefore, we do not only revisit the calibration approach introduced by Van der Burgt (2008), but we propose on top of this two alternative families of differentiable functions in order to extend the possibilities of modelling the CAP.

According to equation (3), the conditional PD is equal to the product of the unconditional PD and the derivative of the CAP. While we simply approximate the first factor by the ODF, the determination of the second factor involves two steps. First, we construct the empirical CAP from the observations. The set $\{(F_{All}(s); F_D(s))\}$ consists of a finite number of isolated points and, thus, the empirical CAP is obviously not differentiable (Tasche 2010). Therefore, we model the CAP by fitting one-parametric families of differentiable functions to the empirical CAP in the second step. The one-parametric families of differentiable functions have to satisfy the following conditions:

- They have to pass through the points $(F_{All}(s_{\min}) = 0; F_D(s_{\min}) = 0)$ and $(F_{All}(s_{\max}) = 1; F_D(s_{\max}) = 1)$.
- They have to be concave in the range from zero to one. Concavity ensures that the PD is a strictly monotonically decreasing function of the credit score and, thus, it preserves the rank ordering implied by the credit scores.
- According to equation (3), the conditional PD is proportional to the first derivative of the CAP. In order to ensure that the conditional PD is bounded between zero and one, the derivative of the one-parametric families of differentiable functions at $F_{All}(s_{\max}) = 1$ have to be larger than or equal to zero and at $F_{All}(s_{\min}) = 0$ it has to be smaller than or equal to $\frac{1}{P(Y=1)}$.

The parameters determine the specific functional shapes of the three modelled CAPs, respectively. Here, we follow Van der Burgt (2008) and determine the parameters of the three families of differentiable functions in such a way that the AUROC, implied by the pairs of credit scores and default labels, is preserved.[7] To this end, we first use the following relationship in order to convert the AUROC, implied by the pairs of credit scores and default labels, into the corresponding area under the CAP (in the range from zero to one) (Engelmann et al. 2003):

$$\begin{aligned} 2 \cdot AUROC - 1 &= \frac{\int_0^1 CAP(x) \cdot dx - \frac{1}{2}}{1 - \frac{P(Y=1)}{2} - \frac{1}{2}} \\ &= \frac{2 \cdot \int_0^1 CAP(x) \cdot dx - 1}{1 - P(Y=1)} \\ \Leftrightarrow \int_0^1 CAP(x) \cdot dx &= \frac{(2 \cdot AUROC - 1) \cdot \big(1 - P(Y=1)\big) + 1}{2}. \end{aligned} \tag{6}$$

Equations (8), (11), and (14) express the areas under the modelled CAP as invertible functions of the parameters, respectively. Finally, we plug the area under the CAP, implied by the pairs of credit scores and default labels, into the inverse of these functions in order to determine the corresponding parameters (AUROC-matching). In the following, we present three different one-parametric families of differentiable functions for modelling the CAP.

First one-parametric family of differentiable functions

Van der Burgt (2008) suggests the following one-parametric family of differentiable functions in order to fit the CAP:

$$CAP_1(x) = \frac{1 - e^{-a_1 \cdot x}}{1 - e^{-a_1}}, \tag{7}$$

where $x \in [0; 1]$ is the cumulative share of all obligors and $a_1 > 0$ is a parameter. The area under $CAP_1(x)$ in the range from $x = 0$ to $x = 1$ is equal to (Van der Burgt 2008):

[7] Please note that the pairs of default labels and PDs, derived from any of the three CAP regressions, result in exactly the same CAP as the one generated by the pairs of default labels and credit scores (i.e. the solid black line in Figure 3). The reason for this is that the CAP regressions produce PDs, which are strictly monotonically decreasing functions of the credit score (cf. Figure 4). Fitting the empirical CAP is only an intermediate step in order to transform credit scores into PDs.

$$\int_0^1 \frac{1-e^{-a_1 \cdot x}}{1-e^{-a_1}} \cdot dx = \left[\frac{x + \frac{1}{a_1} \cdot e^{-a_1 \cdot x}}{1-e^{-a_1}} \right]_0^1$$
$$= \frac{1 + \frac{1}{a_1} \cdot e^{-a_1}}{1-e^{-a_1}} - \frac{\frac{1}{a_1}}{1-e^{-a_1}} \tag{8}$$
$$= \frac{1 + \frac{1}{a_1} \cdot (e^{-a_1} - 1)}{1-e^{-a_1}}$$
$$= \frac{1}{1-e^{-a_1}} - \frac{1}{a_1}.$$

The derivative of $CAP_1(x)$ with respect to x is (Van der Burgt 2008):

$$\frac{d}{dx} CAP_1(x) = \frac{a_1 \cdot e^{-a_1 \cdot x}}{1-e^{-a_1}}. \tag{9}$$

Second one-parametric family of differentiable functions

The entropy-maximizing one-parametric family of differentiable functions is theoretically appealing, because maximizing the entropy leads to the distribution of the binary default variable Y that carries the highest uncertainty and, thus, the fewest assumptions about the true distribution of data. More specifically, a special case of the result of Brunel (2019) shows that the logistic regression on the ranks of obligors (rather than on the credit scores) maximizes the entropy of the distribution of the binary default variable Y under the assumption of independent defaults and conditional on a specified AUROC. This brings us to our second one-parametric family of differentiable functions for fitting the empirical CAP:

$$CAP_2(x) = \frac{\ln\left\{\frac{1+e^{a_2 \cdot x}}{2}\right\}}{\ln\left\{\frac{1+e^{a_2}}{2}\right\}}, \tag{10}$$

where $a_2 < 0$ is a parameter. The area under $CAP_2(x)$ in the range from $x = 0$ to $x = 1$ is equal to:

$$\begin{aligned}\int_0^1 \frac{\ln\left\{\frac{1+e^{a_2\cdot x}}{2}\right\}}{\ln\left\{\frac{1+e^{a_2}}{2}\right\}}\cdot dx &= \frac{1}{\ln\left\{\frac{1+e^{a_2}}{2}\right\}}\cdot\int_0^1 \ln\left\{\frac{1+e^{a_2\cdot x}}{2}\right\}\cdot dx\\ &= \frac{1}{\ln\left\{\frac{1+e^{a_2}}{2}\right\}}\\ &\cdot\left[x\cdot\ln\left\{\frac{1+e^{a_2\cdot x}}{2}\right\}-\frac{1}{a_2}\cdot Li_2(-e^{a_2\cdot x})-x\cdot\ln\{1+e^{a_2\cdot x}\}\right]_0^1\\ &= \frac{1}{\ln\left\{\frac{1+e^{a_2}}{2}\right\}}\\ &\cdot\left[\ln\left\{\frac{1+e^{a_2}}{2}\right\}-\frac{1}{a_2}\cdot Li_2(-e^{a_2})-\ln\{1+e^{a_2}\}+\frac{1}{a_2}\cdot Li_2(-1)\right],\end{aligned} \tag{11}$$

where $Li_2(\cdot)$ denotes the polylogarithm function of second order, i.e. $Li_2(z)=\sum_{k=1}^{\infty}\frac{z^k}{k^2}$.

The derivative of $CAP_2(x)$ with respect to x is indeed equal to the logistic regression on the ranks multiplied by a factor:

$$\begin{aligned}\frac{d}{dx}CAP_2(x) &= \frac{1}{\ln\left\{\frac{1+e^{a_2}}{2}\right\}}\cdot\frac{d}{dx}\ln\left\{\frac{1+e^{a_2\cdot x}}{2}\right\}\\ &= \frac{a_2}{\ln\left\{\frac{1+e^{a_2}}{2}\right\}}\cdot\frac{1}{1+e^{-a_2\cdot x}}.\end{aligned} \tag{12}$$

In equation (12), x is equal to the cumulative share of all obligors, i.e. equal to the rank divided by the total number of obligors. If we set $a_2=\widetilde{a_2}\cdot n$ with n being the total number of obligors, then we explicitly obtain the logistic regression on the ranks.

Figure 1 and Figure 2 reveal that the one-parametric family of differentiable functions, proposed by Van der Burgt (2008), produces similar PDs for lower unconditional PDs and/or lower AUROCs as the one-parametric family of differentiable functions that maximizes the entropy of the distribution of the binary default variable Y for a given AUROC.

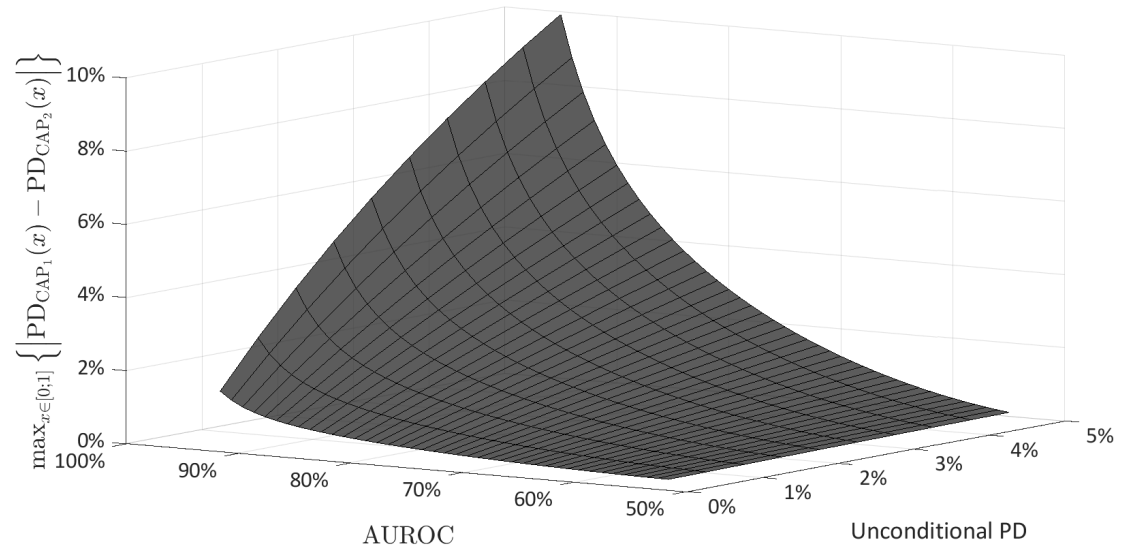

Figure 1: Maximum absolute difference between PDs produced by CAP_1 and CAP_2 as a function of the AUROC and unconditional PD.

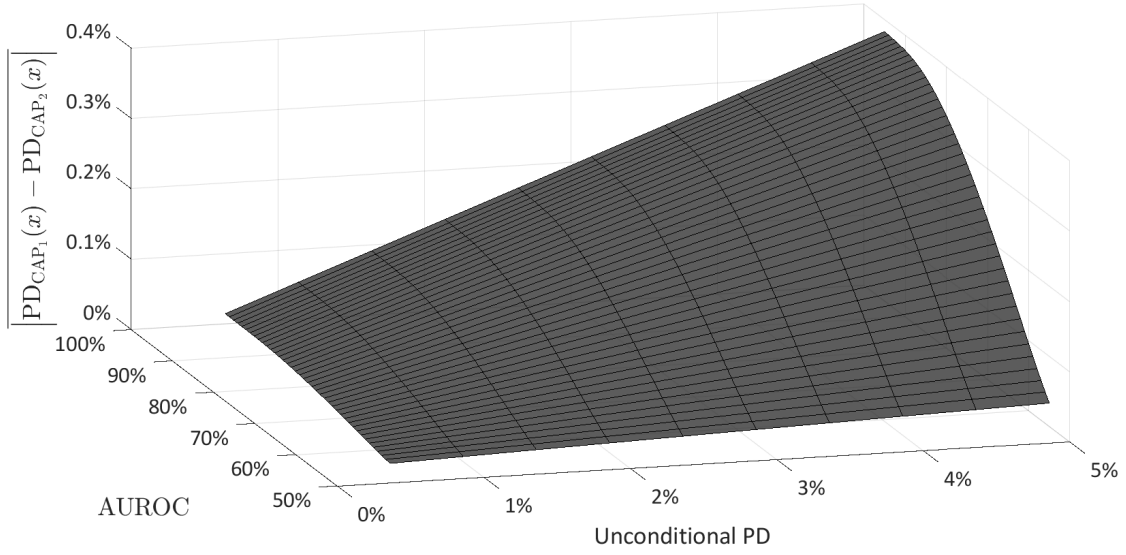

Figure 2: Average absolute difference between PDs produced by CAP_1 and CAP_2 as a function of the AUROC and unconditional PD.

Third one-parametric family of differentiable functions

The second one-parametric family of differentiable functions maximizes the entropy of the distribution of the binary default variable Y conditional only on a specified AUROC. In so doing, it disregards the specific form of the empirical CAP. Hence, we slightly modify the second one-parametric family of differentiable functions in order to represent more accurately the empirical CAP. More precisely, the third one-parametric family of differentiable functions used in this paper is:

$$CAP_3(x) = \frac{\ln(1 + a_3 \cdot x)}{\ln(1 + a_3)}, \tag{13}$$

where $a_3 > 0$ is a parameter. For the area under this function in the range from $x = 0$ to $x = 1$, we get:

$$\begin{aligned} \int_0^1 CAP_3(x) \cdot dx &= \frac{1}{\ln(1 + a_3)} \cdot \int_0^1 \ln(1 + a_3 \cdot x) \cdot dx \\ &= \frac{1}{a_3 \cdot \ln(1 + a_3)} \cdot [(1 + a_3 \cdot x) \cdot \ln(1 + a_3 \cdot x) - a_3 \cdot x]_0^1 \\ &= \frac{(1 + a_3)}{a_3 \cdot \ln(1 + a_3)} \cdot [\ln(1 + a_3) - 1]. \end{aligned} \tag{14}$$

Finally, we derive $CAP_3(x)$ with respect to x:

$$\begin{aligned} \frac{d}{dx} CAP_3(x) &= \frac{a_3}{\ln(1 + a_3)} \cdot \frac{1}{1 + a_3 \cdot x} \\ &= \begin{cases} \dfrac{a_3}{\ln(1 + a_3)}, x = 0 \\ \dfrac{a_3}{\ln(1 + a_3)} \cdot \dfrac{1}{1 + e^{\ln(a_3 \cdot x)}}, x > 0. \end{cases} \end{aligned} \tag{15}$$

The main difference between the second and third one-parametric family is that the logarithm is applied to the product of the parameter and x in the latter case.

3 Comparison of calibration approaches

This section benchmarks the three one-parametric families of differentiable functions, presented in equations (7), (10), and (13), against the LLR using real-world data. The LLR serves as a benchmark, because it is a standard and widely accepted calibration approach in both the banking sector and the academic literature on credit risk estimation. We evaluate the calibration performance of the four approaches out-of-sample.

The analysis is based on a real-world data set, which originates from a European bank. The bank is authorised to follow the IRBA in order to assess the creditworthiness of its obligors in this portfolio. The data set consists of real-valued credit scores and the default statuses one year after the calculation of the credit scores of about 2,450 defaulted and 277,000 non-defaulted obligors. Hence, the unconditional PD in this data set is approximately 0.88%. The AUROC amounts to 0.83.

Figure 3 compares the empirical CAP with three CAP regressions. Obviously, the third one-parametric family of differentiable functions represents the empirical CAP best. Furthermore, the first and second family result in very similar CAPs.

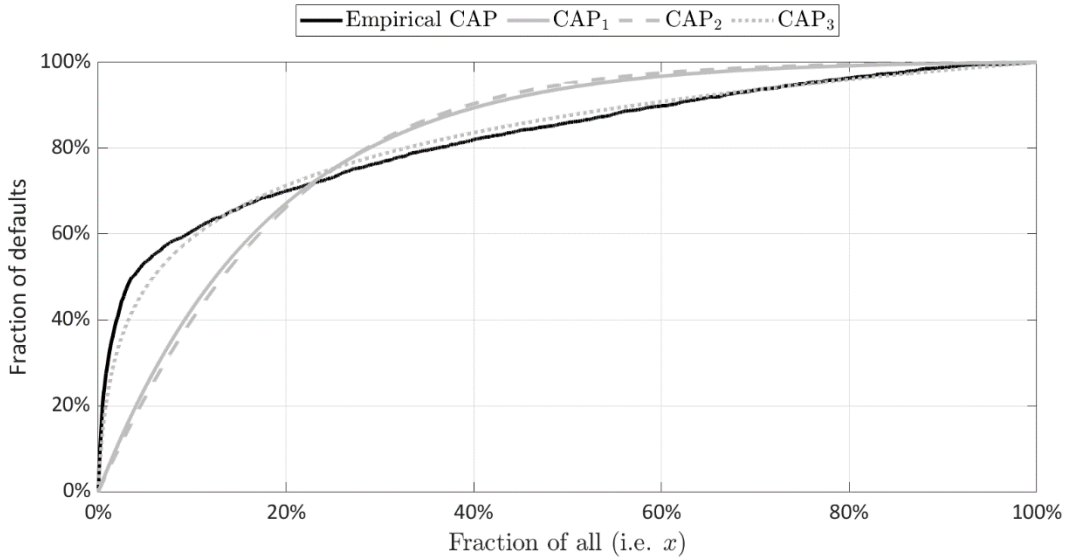

Figure 3: Comparison of the empirical CAP, implied by the real-world data set, and three modelled CAPs.

In the following, we measure relative performance differences between different calibration approaches. An optimal model forecasts the true PDs of individual obligors (Aussenegg et al. 2011; Bequé et al. 2017). However, the true (conditional) PD is a latent variable and, as such, unobservable (Aussenegg et al. 2011; Böken 2021). Therefore, we have to fall back on measures of calibration performance that are based on ex-ante estimated conditional PDs and ex-post observed default labels (Aussenegg et al. 2011; Böken 2021). Specifically, our judg-

ments rest on the Brier Score and the Log Loss which are widely accepted measures of calibration performance (Bequé et al. 2017; Böken 2021; Kruppa et al. 2013). The Brier Score is the average over all squared differences between ex-ante estimated conditional PDs and ex-post observed default labels. The Log Loss is equal to the negative average log-likelihood. Lower values of the Brier Score and Log Loss indicate better calibration quality. However, the absolute value of these two measures is hard to interpret (Böken 2021). Therefore, we calculate the pairwise relative differences between the Brier Score and Log Loss of the LLR and the three CAP regressions, respectively.

We compare the three CAP regressions to the LLR for different training sample sizes expressed as percentage of the total data set. Specifically, the size of the training sample ranges from 5% to 95% in steps of 10% of the total data set. At the beginning, we split the total data set into two subsets that contain all defaults and all non-defaults, respectively. For each of the defined training sample sizes, we then randomly create 1,000 training and test data sets. To this end, we randomly draw a sample from all defaults and a sample from all non-defaults of predefined size (ranging from 5% to 95% in steps of 10%), respectively. The union of these two subsets is the training data set, while the remaining data serve as test data. In so doing, we ensure that the unconditional PDs in the training data set and test data set are equal except for rounding differences. For each draw, we benchmark the three CAP regressions against the LLR, respectively. The two parameters of the linear logit (i.e. the coefficient of the credit score and the constant) are determined by maximizing the likelihood function on the training data.[8]

Due to the random sampling, the AUROC of the training data set fluctuates around the AUROC of the total data set with every draw. For each draw, we determine the three parameters of the three CAP regressions in such a way that the AUROCs induced by the modelled CAPs are equal to the AUROC implied by the pairs of credit scores and binary default labels of the training data set, respectively. After determining the three parameters, we calculate the PDs for the training data set based on equations (9), (12), and (15), respectively. We multiply the estimated PDs by a constant in order to ensure that the average estimated PD is equal to the ODF in the training sample.[9]

We benchmark the three CAP regressions against the LLR on the hold-out test data set (i.e. all the data not considered for training). To this end, we first assign PDs to the test data. If a credit score of the test data set falls between two credit scores of the training data set, we derive its PD via linear interpolation. If a credit score of the test data exceeds the maximum or falls below

[8] More precisely, we use the Matlab-function *glmfit* in order to estimate the parameters of the LLR.

[9] Please note that the ODF in the training sample is approximately equal to the ODF of 0.88% in the total data set due to the construction of the training and test samples.

the minimum credit score of the training data set, we determine its PD via an almost flat extrapolation. This means that we perform a linear regression between the maximum [resp. minimum] credit score of the training data set and a synthetic data point with a credit score of 50 [resp. -50] and a PD that differs by 0.001% from the minimum [resp. maximum] PD of the training data set. In doing so, we preserve the rank ordering of the test data. As an example, Figure 4 shows the PDs of a random test data set (containing 25% of all pairs of credit scores and default labels) as functions of the credit score for the four calibration approaches. The CAP_1- and CAP_2-regression produce PDs in a narrower range than the other two calibration approaches.

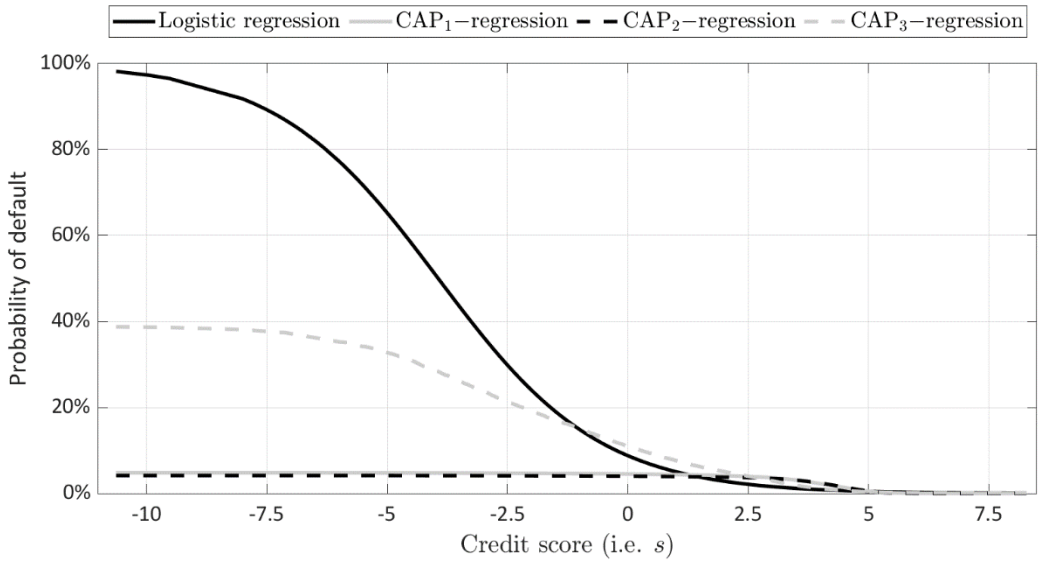

Figure 4: PDs of a random test data set (containing 25% of the total sample) as functions of the credit score for the four calibration approaches.

As a disadvantage, we use the credit scores rather than the ranks in order to assign PDs to the test data set. Indeed, we could alternatively assign PDs to the test data set based on equations (9), (12), and (15), respectively. However, this approach implies that we knew the rank ordering of the entire test data set at once. As obligors in practice constantly enter and exit the application portfolio, we deem this approach unrealistic.

After assigning PDs to the test data set, we calculate the Brier Score and the Log Loss on the test data set for the four calibration approaches. Figure 5 shows the medians of the relative differences in the Brier Score between the LLR and the three CAP regressions over the 1,000 random samplings as a function of the training sample size, respectively. Figure 6 provides the analogous information for the Log Loss as an alternative measure of calibration performance. The main results hold for both measures. On the one hand, Figure 5 and Figure 6 reveal the superiority of the LLR over the first and second CAP regressions. On the other hand, these figures indicate that the third CAP regression outperforms the LLR.

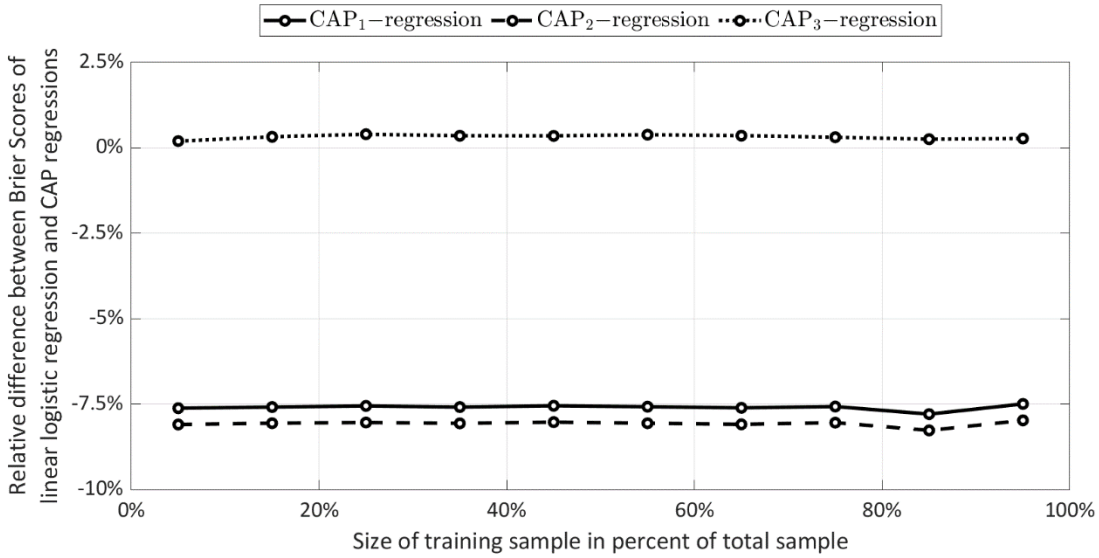

Figure 5: Medians of relative differences in Brier Score between linear logistic regression and CAP regressions as functions of the training sample size.

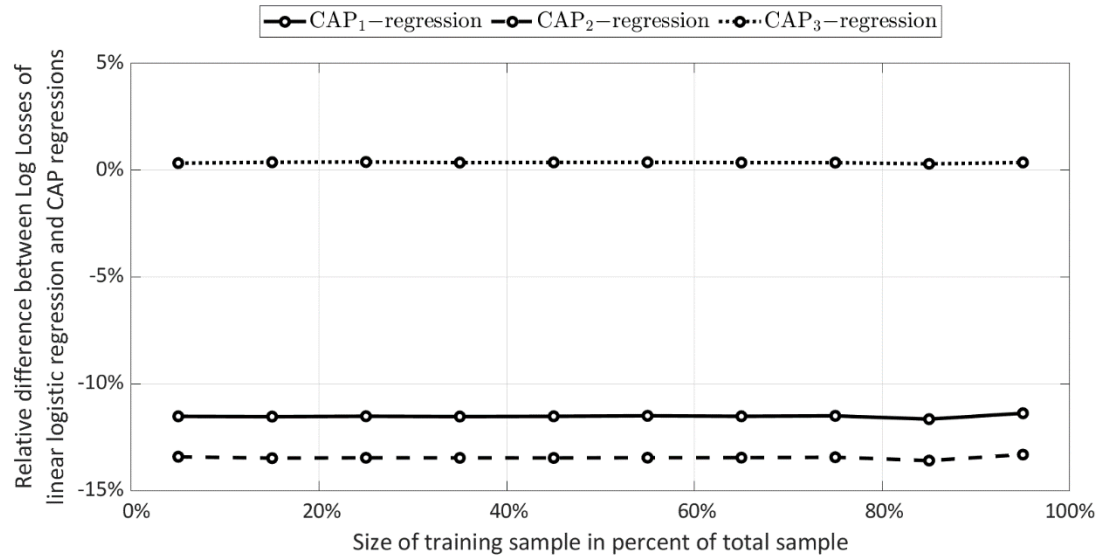

Figure 6: Medians of relative differences in Log Loss between linear logistic regression and CAP regressions as functions of the training sample size.

In order to further assess the stability of this outperformance, Figure 7 and Figure 8 plot the 10%-, 50%- (i.e. the median), and 90%-quantile over the 1,000 relative differences versus the training sample size. Again, the main results hold for both measures of calibration performance. The relative differences of the Brier Scores and Log Losses fluctuate in a narrow range around the medians. In general, smaller sample sizes involve higher uncertainty. Therefore, it seems plausible that the range between the 10%- and 90%-quantile is largest for small training data sets (i.e. on the left of Figure 7 and Figure 8) and for small test data sets (i.e. on the right of Figure 7 and Figure 8). Although the 10%-quantiles are below zero for most of the analysed

training sample sizes, Figure 7 and Figure 8 further substantiate the outperformance of the third CAP regression over the LLR on the data set in question. In particular, the relative difference between Log Losses of the LLR and the third CAP regression is significantly larger than zero with a confidence level of 90% for medium training sample sizes.

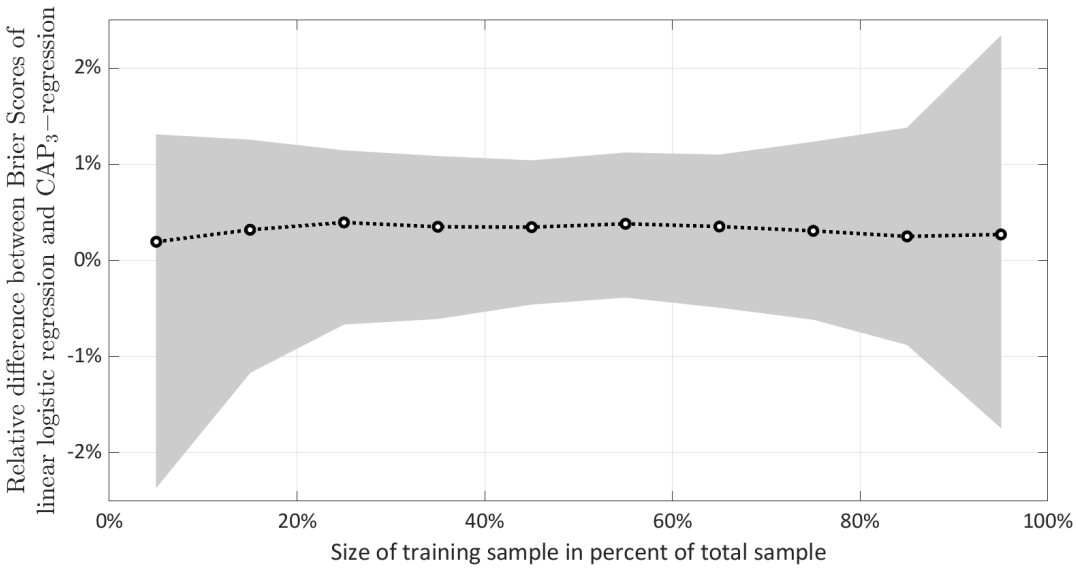

Figure 7: Median and range between 10%- and 90%-quantile of relative differences in Brier Score between linear logistic regression and CAP_3-regression as a function of the training sample size.

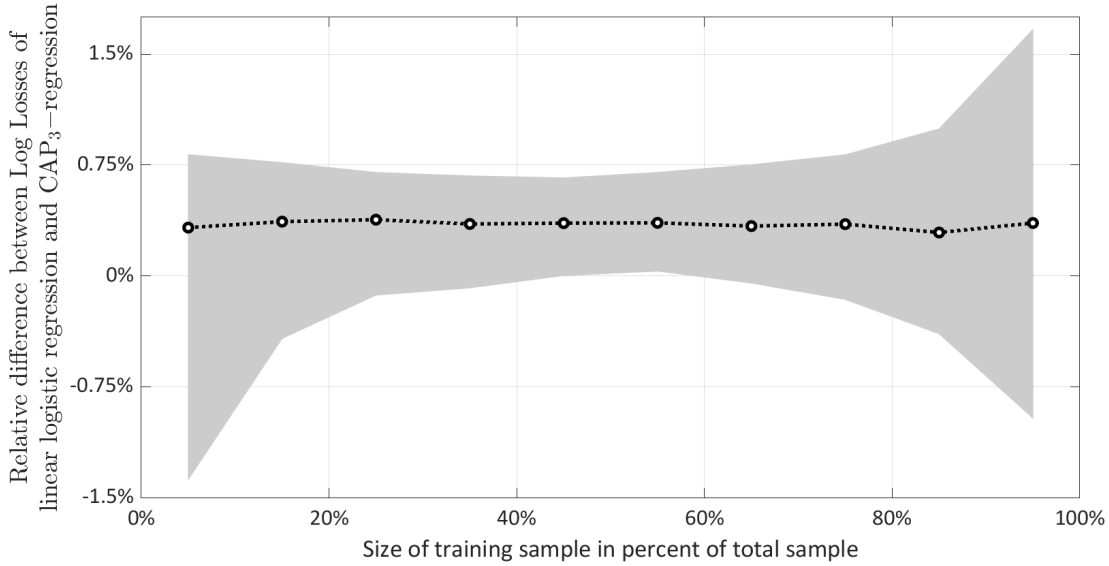

Figure 8: Median and range between 10%- and 90%-quantile of relative differences in Log Loss between linear logistic regression and CAP_3-regression as a function of the training sample size.

4 Propagation of uncertainty from discriminatory power to probabilities of default

In the run-up to the entry into force of the Guidelines on PD estimation, LGD estimation and treatment of defaulted assets (European Banking Authority 2017) on the 1st of January 2022 (European Banking Authority 2019), the quantification of uncertainties inherent in PD estimates has gained momentum. Among other things, paragraphs 42 and 43 (b) of these guidelines require that PD estimates should take a general estimation error into account, which reflects the dispersion of the distribution of the statistical estimator. Paragraph 140 (a) of the ECB guide to internal models further specifies that "when using direct PD estimates, the [margin of conservatism] is based on the distribution of this direct PD estimator (which includes the risk differentiation function), implicitly reflecting the uncertainty of the [long run average]" (European Central Bank 2019).

According to equation (3), uncertainties of PD estimates can indeed stem from two sources. In addition to the statistical dispersion of the unconditional PD, uncertainties of the discriminatory power can transpire into PD estimates via the derivative-term in equation (3). The purpose of this subsection is to explore whether the calibration methodology proposed by Van der Burgt (2008) provides the opportunity to derive uncertainties of PD estimates from the statistical dispersion of discriminatory power. In principle, we could select any of the three CAP regressions presented in Section 2.2 for this analysis. However, we choose the CAP_1-regression as we consider this one-parametric family of differentiable functions as standard. The CAP regression proposed by Van der Burgt (2008) does not adequately represent the CAP of our real-world data set as Figure 3 illustrates. Therefore, we base our analysis on synthetic data sets similar to Böken (2021); Brunel (2019); Tasche (2010). The construction of synthetic data sets comprises the following three steps.

- First, we specify that the credit scores of the defaults and non-defaults follow normal distributions, respectively. In order to comply with the convention that low credit scores tend to represent low creditworthiness and vice versa, the standard deviations of the two normal distributions have to be equal (cf. equations (21) and (24) in the Appendix) and the mean of the non-defaults has to exceed the mean of the defaults. To be exact, the positive real number μ denotes the mean of the non-defaults and $-\mu$ stands for the mean of the defaults.
- Second, we specify the discriminatory power measured by the AUROC. Given the AUROC and the standard deviation of the two normal distributions, the inverse of equation (3.14) in Tasche (2010) allows for the calculation of μ.
- After setting the numerical values of three for the standard deviations, of 0.85 for the AUROC, and of 5% for the unconditional PD, we can draw data sets of arbitrary size from the two normal distributions in the third step.

The synthetic data sets generated in this way provide optimal conditions by excluding any unwanted noisy influence. As an example, Figure 9 illustrates that the one-parametric family of

differentiable functions proposed by Van der Burgt (2008) adequately represents the empirical CAP of a synthetic data set. The differences between the theoretical and empirical CAPs are due to sampling. The sample contains less information than the underlying data generating distribution (Böken 2021).

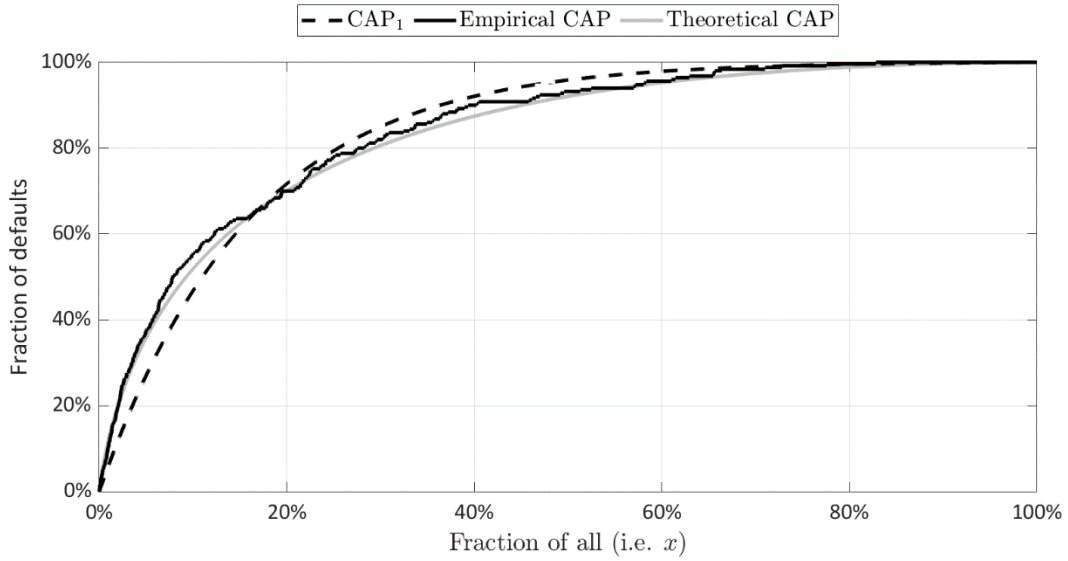

Figure 9: Comparison of theoretical CAP, empirical CAP, and fitted CAP_1 for an exemplary synthetic data set defined by 5,000 obligors, a theoretical AUROC of 0.85, standard deviations equal to three (i.e. $\alpha = 1$), and an unconditional PD of 5%.

Due to the random sampling and the finite size of the data sets, the AUROC of the training data set is subject to statistical uncertainty and, thus, it fluctuates around its theoretical value of 0.85 (Van der Burgt 2020). More precisely, the AUROC is asymptotically normally distributed (Engelmann et al. 2003). We approximate the expected value of this normal distribution by the empirical AUROC of the drawn training data set. Starting from the equation for $Var(U)$ between equations (3) and (4) in the supplementary materials of Fong and Huang (2019), we furthermore derive the variance of the AUROC. Under the more general assumption that the credit score of the non-defaults [resp. defaults] follows a normal distribution with expected value μ [resp. $-\mu$] and standard deviation σ [resp. $\alpha \cdot \sigma$], we arrive at the following equation for the variance of the AUROC:

$$
\begin{aligned}
&Var(\text{AUROC}|n_D; n_{ND}; \alpha) \\
&\quad = \frac{1}{n_D \cdot n_{ND}} \\
&\quad \cdot \Bigg\{ (n_{ND} - 1) \cdot \int_{-\infty}^{\infty} \Phi\left(\frac{y - \sqrt{\frac{1+\alpha^2}{\alpha^2}} \cdot \Phi^{-1}(\text{AUROC})}{\frac{1}{\alpha}} \right)^2 \cdot \varphi(y) \cdot dy \\
&\quad + (n_D - 1) \cdot \int_{-\infty}^{\infty} \Phi\left(\frac{y + \sqrt{1+\alpha^2} \cdot \Phi^{-1}(\text{AUROC})}{\alpha} \right)^2 \cdot \varphi(y) \cdot dy \\
&\quad - (n_{ND} - 1) \cdot (1 - \text{AUROC})^2 - n_D \cdot \text{AUROC}^2 + \text{AUROC} \Bigg\},
\end{aligned}
\tag{16}
$$

where

- n_D denotes the number of defaults in the sample,
- n_{ND} represents the number of non-defaults in the sample,
- $\varphi(\cdot)$ stands for the PDF of the standard normal distribution,
- $\Phi(\cdot)$ denotes the CDF of the standard normal distribution, and
- $\Phi^{-1}(\cdot)$ represents the inverse of the CDF of the standard normal distribution.

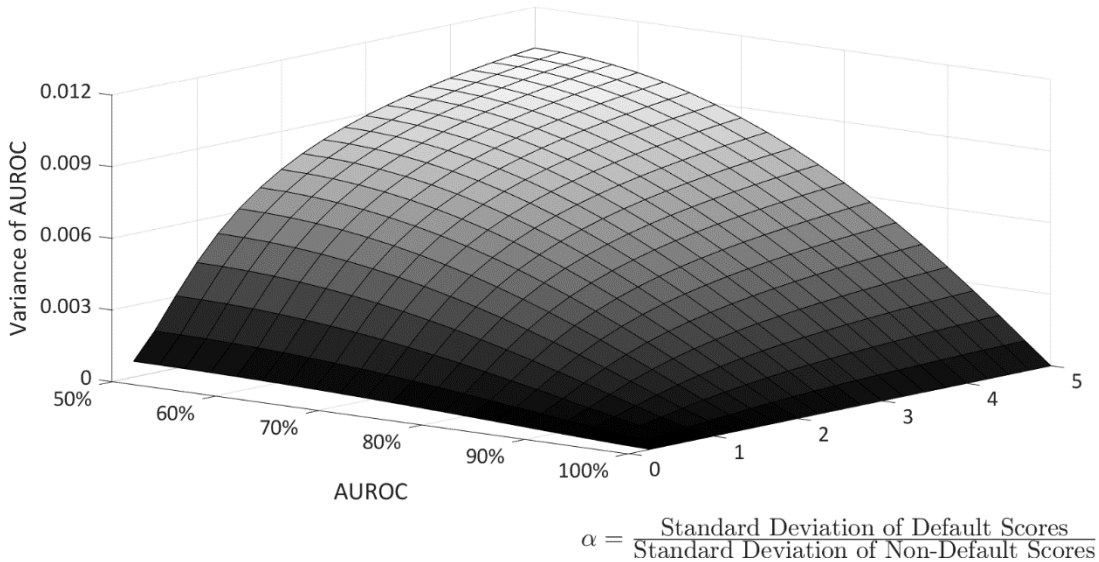

Figure 10: Variance of the AUROC for normally distributed credit scores (with $\sigma = 3$) as a function of the AUROC and the parameter α for a synthetic data set of 1,000 obligors.

Figure 10 shows the variance of the AUROC as a function of the AUROC and of the parameter α. This figure reveals that the variance of the AUROC is a strictly monotonically decreasing function of the AUROC for fixed α.

Under the same assumption that the credit scores of the defaults and non-defaults follow normal distributions, respectively, Van der Burgt (2020) derives the variance of the AUROC based on the relationship between the AUROC and the Mann-Whitney U-statistics. Van der Burgt (2020) eventually expresses the variance of the AUROC by means of Owen's T function. We independently derive the alternative representation of the variance of the AUROC from a different starting point. However, our test calculations suggest that equation (16) and equation (4.10) in Van der Burgt (2020) give exactly the same results and, thus, the two closed-form equations for the sample variance in the observed AUROC are equivalent. In order to comply with the convention that the PD is a strictly monotonic function of the credit score, we set the parameter α in equation (16) equal to one in what follows (cf. Appendix)

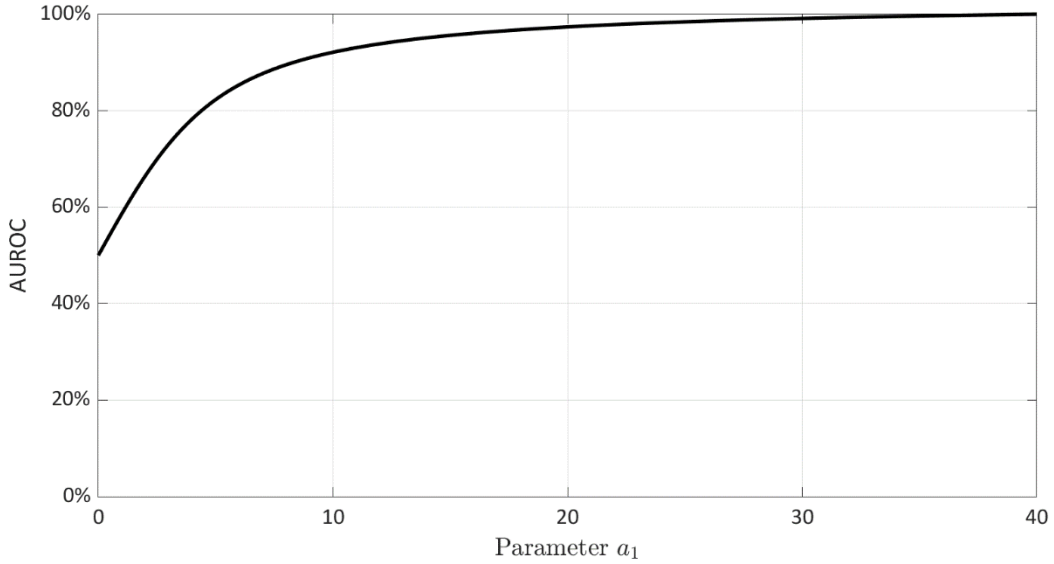

Figure 11: The AUROC as a function of the parameter a_1 in the framework of Van der Burgt (2008).

The uncertainty of the AUROC transpires, first, into uncertainties of the parameter a_1 and, second, into uncertainties of the PD estimates. Figure 11 demonstrates that the AUROC is a strictly monotonically increasing function of the parameter a_1 in the framework of Van der Burgt (2008). Consequently, we can invert this function in order to find the parameter a_1 leading to the empirical AUROC, implied by the pairs of credit scores and default labels of the training data set. Furthermore, we calculate the 5%- and 95%-quantile of the AUROC assuming a normal distribution. As the AUROC is a strictly monotonically increasing function of the parameter a_1, we can also calculate the 5%- and 95%-quantiles of a_1. In the next step, we seek to convert the confidence interval of a_1 into a confidence interval of the PD. According to Blümke (2020), we can obtain a more conservative PD estimate than the expected PD by choosing a more conservative percentile of the PD. In order to derive conservative PDs from the uncertainty of the AUROC, we distinguish the following three cases.

- For small cumulative shares of all obligors (i.e. x), the PD is a strictly monotonically increasing function of the parameter a_1 in the framework of Van der Burgt (2008) (cf. Figure 12). Hence, we can calculate conservative PDs by using the 95%-quantile of a_1.
- For high cumulative shares of all obligors, the PD is a strictly monotonically decreasing function of a_1 (not illustrated). As a consequence, we can calculate conservative PDs by using the 5%-quantile of a_1.
- For intermediate cumulative shares of all obligors, the functional relationship between the PD and a_1 is not strictly monotonic (cf. Figure 12). In this range, we approximate conservative PDs by means of linear interpolation. To this end, we determine the linear function starting at the last cumulative share for which the PD is a strictly monotonically increasing function of a_1 in the relevant range and ending at the first cumulative share for which the PD is a strictly monotonically decreasing function of a_1 in the relevant range (cf. Figure 13).

  Figure 14 shows the range of cumulative shares, for which we cannot calculate conservative PDs based on quantiles of the parameter a_1, as a function of the portfolio size and AUROC for a confidence level of 95%. The lower [resp. upper] surface presents the lower [resp. upper] bounds of these ranges. The figure reveals that the range of non-processable cumulative shares decreases with increasing portfolio size and increasing AUROC. It is worth noting that the approach can even be applied to almost 82% of the obligors in the worst case characterised by an AUROC of 0.6 and a portfolio size of 500.

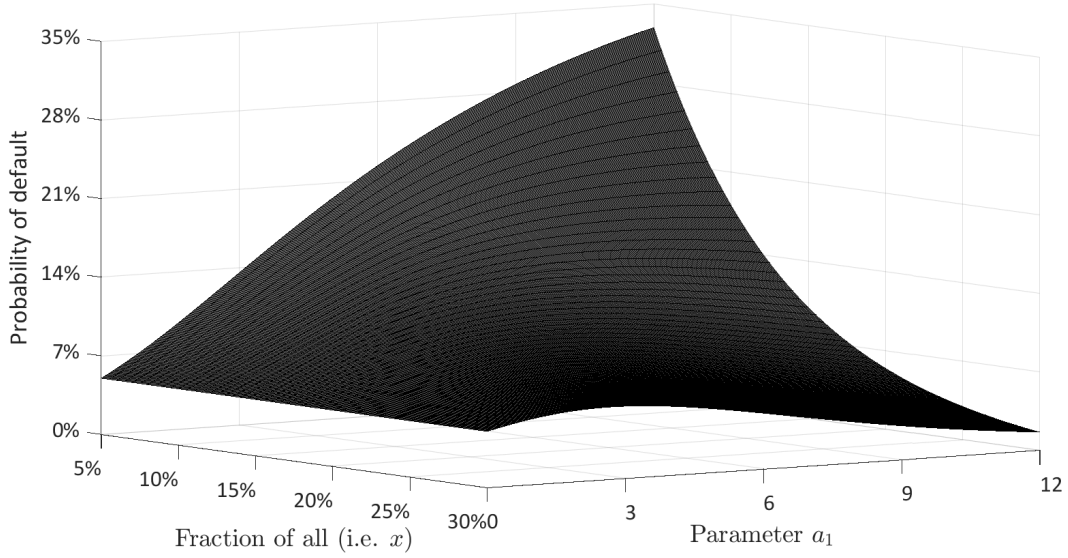

Figure 12: Probability of default as a function of the cumulative share of all obligors (i.e. x) and parameter a_1 in the framework of Van der Burgt (2008).

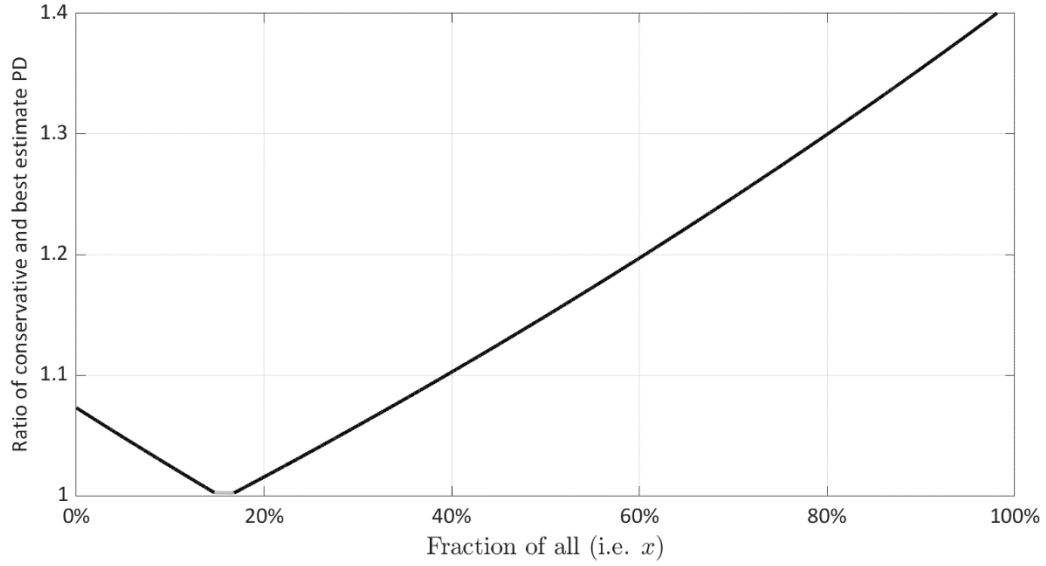

Figure 13: Ratio of conservative and best estimate PD for an exemplary synthetic data set defined by 16,000 obligors, a theoretical AUROC of 0.85, standard deviations equal to three (i.e. $\alpha = 1$), and an unconditional PD of 5% for a confidence level of 95%.

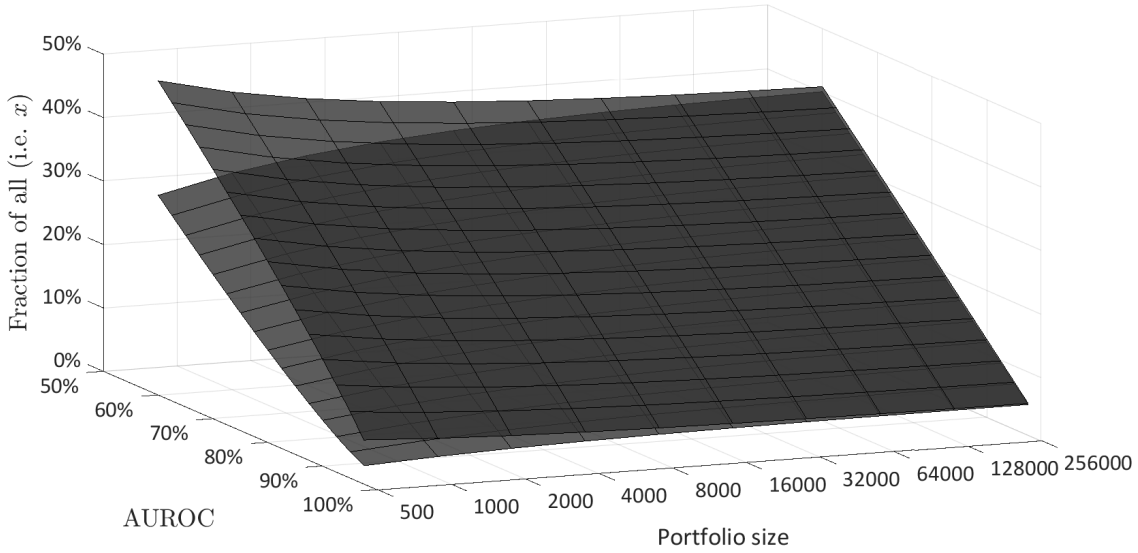

Figure 14: Range of cumulative share, for which we cannot directly estimate conservative PDs, as a function of the AUROC and portfolio size for standard deviations equal to three (i.e. $\alpha = 1$), an unconditional PD of 5%, and a confidence level of 95%. (Please note the logarithmic scale on the axis showing the portfolio size.)

We define different portfolio sizes starting with 500 and doubling until 256,000. For each of the defined portfolio sizes, we then randomly draw 1,000 data sets of credit scores from the two normal distributions. For each synthetic obligor, we calculate the best estimate PD and the conservative PD by plugging the best estimate and conservative value of a_1 into equation (9), respectively (cf. Figure 13). Following this, we average the ratios of conservative PD divided by best estimate PD over all synthetic obligors and plot the 10%-, 50%- (i.e. the median), and 90%-

quantile of these averages over the 1,000 random draws versus the sample size in Figure 15. In so doing, this figure illustrates the degree of reliability of the estimates as a function of the sample size. As intuition suggests, both the median of the average ratios and the range between the 10%- and 90%-quantile decrease for larger sample sizes.

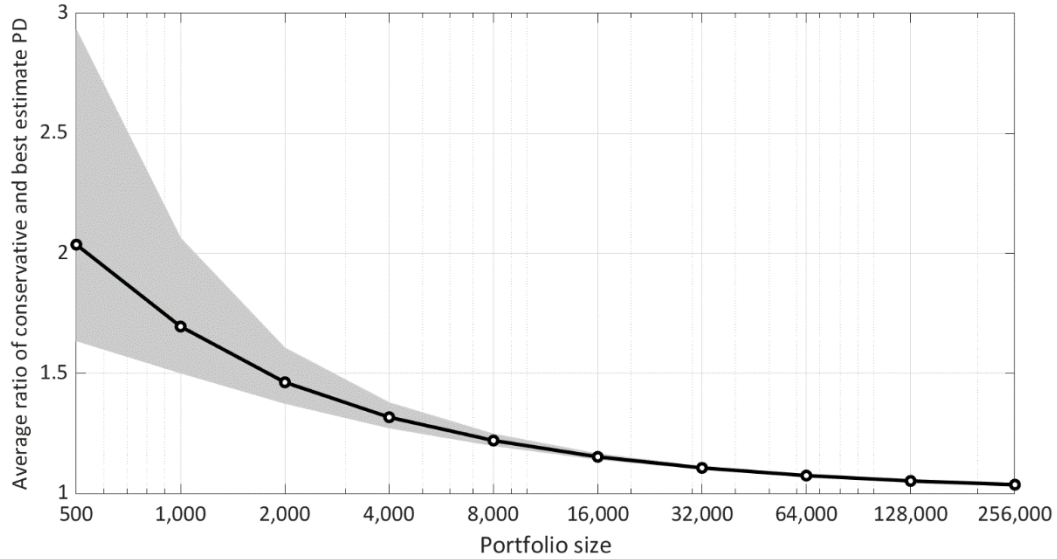

Figure 15: Median and range between 10%- and 90%-quantile of the average ratio of conservative and best estimate PD as a function of the sample size for a confidence level of 95% and a theoretical AUROC of 0.85. (Please note the logarithmic scale on the axis showing the portfolio size.)

So far, we have assumed a deterministic unconditional PD. In fact, however, the unconditional PD is subject to uncertainty, which may amplify the dispersion of the conditional PD. More precisely, the unconditional PD affects the upper endpoint of the confidence interval of the conditional PD in two opposite ways. On the one hand, higher values of the unconditional PD result in higher values of the conditional PD through the first factor on the right hand side of equation (3). In order to account for this source of uncertainty, we can determine a confidence interval of the unconditional PD and use its upper endpoint in equation (3). On the other hand, higher values of the unconditional PD lead to lower values of the variance of the AUROC. In order to see this, we set α equal to one, we substitute the number of defaults by the product of the total number of obligors in the portfolio and the unconditional PD (i.e. $n_D = n \cdot ODF$), and we replace the number of non-defaults n_{ND} by $n - n_D = n \cdot (1 - ODF)$ in equation (16):

$$
\begin{aligned}
&Var(\mathrm{AUROC}, \mathrm{ODF} | n, \alpha = 1) \\
&= \frac{1}{n^2 \cdot ODF \cdot (1 - ODF)} \\
&\cdot \Bigg\{ (n \cdot (1 - ODF) - 1) \cdot \int_{-\infty}^{\infty} \Phi\left(y - \sqrt{2} \cdot \Phi^{-1}(\mathrm{AUROC})\right)^2 \cdot \varphi(y) \cdot dy \\
&+ (n \cdot ODF - 1) \cdot \int_{-\infty}^{\infty} \Phi\left(y + \sqrt{2} \cdot \Phi^{-1}(\mathrm{AUROC})\right)^2 \cdot \varphi(y) \cdot dy \\
&- (n \cdot (1 - OD\,F) - 1) \cdot (1 - \mathrm{AUROC})^2 - n \cdot ODF \cdot \mathrm{AUROC}^2 \\
&+ \mathrm{AUROC} \Bigg\}.
\end{aligned}
\tag{17}
$$

Figure 16 reveals that the higher the unconditional PD, the lower the variance of the AUROC. The lower variance of the AUROC in turn leads to a reduced upper endpoint of the confidence interval of the conditional PD via the second factor on the right hand side of equation (3). Therefore, neglecting the randomness of the unconditional PD in the second factor on the right-hand side of equation (3) simplifies and adds conservatism to the calculation of the margin of conservatism for the general estimation error of the conditional PD.

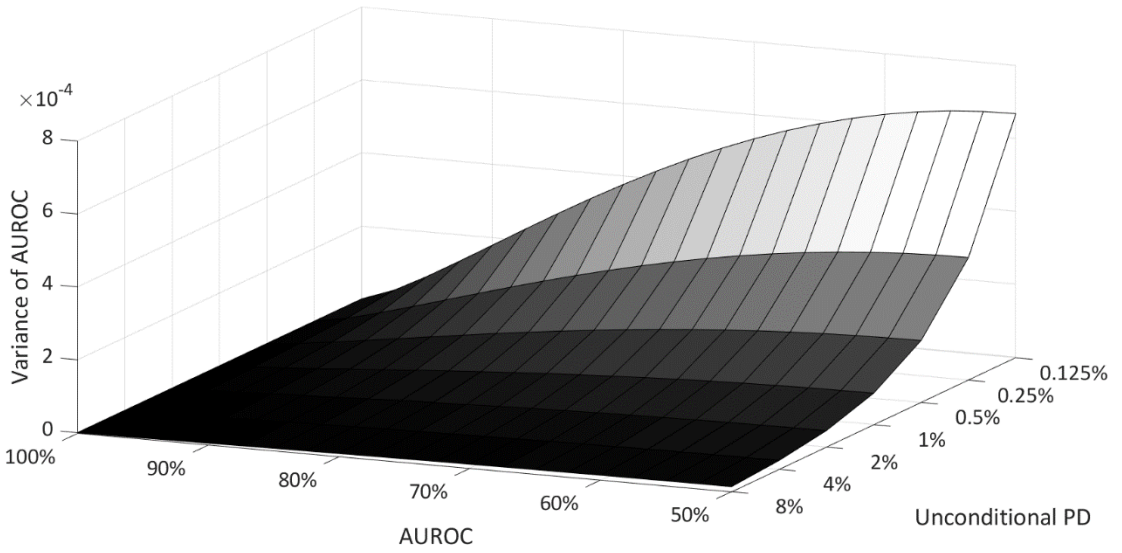

Figure 16: Variance of the AUROC for normally distributed credit scores (with standard deviation equal to three, i.e. $\alpha = 1$) as a function of the AUROC and the unconditional PD for a synthetic data set of 100,000 obligors. (Please note the logarithmic scale on the axis showing the unconditional PD.)

5 Conclusion

The LLR has developed into a standard approach in order to transform credit scores into PD estimates over the last decades. As machine learning techniques increasingly find their way into the discriminatory phase of credit risk models, however, the standard calibration approach is under scrutiny again. For example, Bequé et al. (2017) find that processing the output of different machine learning techniques improves the calibration performance without hurting the discriminatory power. In particular, the authors reveal that a nonlinear but monotonic logit is especially suitable for calibrating the output of machine learning techniques.

Falkenstein et al. (2000) propose a very general calibration methodology, which is based on modelling the empirical CAP through a differentiable function. So far, this calibration methodology has only attracted little attention in the banking sector and in the academic literature. The key question when implementing the general calibration methodology of Falkenstein et al. (2000) is how to model the empirical CAP. Van der Burgt (2008) proposes a one-parametric family of differentiable functions in order to fit the empirical CAP. Based on the result of Brunel (2019), we substantiate this proposal by demonstrating its similarity to the maximum entropy approach (for lower unconditional PDs and/or lower AUROCs). Similarity to the one-parametric family of differentiable functions that maximizes the entropy is desirable as it makes the fewest assumptions about the true distribution of the binary default variable.

However, both regression approaches disregard the specific form of the empirical CAP. Therefore, we propose a third one-parametric family of differentiable functions inspired by the result of Brunel (2019). In order to analyse the practical relevance of these three regression based calibration approaches, we benchmark them against the LLR on a real-world data set. Our results reveal that only the third one-parametric family of differentiable functions outperforms the LLR. Given the fact that the median of the relative difference between the LLR and the third one-parametric family of differentiable functions ranges below 0.5%, it is important to highlight that even small improvements in calibration performance may significantly improve the pricing accuracy of credit products (Alonso et al. 2020). These gains in pricing accuracy, in turn, may translate into competitive advantages and into relevant refinements of regulatory capital quantification. The bottom line of this analysis is that extending the LLR can improve the calibration performance as already demonstrated by Bequé et al. (2017).

Furthermore, we develop an approach, based on the ansatz of Van der Burgt (2008), in order to transfer the statistical dispersion of the discriminatory power into a margin of conservatism for the general estimation error of the PD. In this context, we also provide an alternative representation for the variance of the AUROC to the one proposed by Van der Burgt (2020). In order to demonstrate the effectiveness of our approach, we run a simulation study based on artificially

generated data sets. These synthetic data sets provide optimal conditions by excluding any unwanted noisy influence. More precisely, we generate credit scores of defaults and non-defaults by sampling from two normal distributions. The mean of the defaults is equal to the negative mean of the non-defaults and the standard deviations of the two distributions are equal. This setting has two important properties. First, the PD is a strictly monotonically decreasing function of the credit score as required by convention. Second, we can calculate the true posterior probabilities through equations (5) and (21).

Although the proposed approach is certainly not perfect, it provides an opportunity to relate the uncertainty of discriminatory power to uncertainties of individual PDs as required, for example, by paragraph 140 (a) of the ECB guide to internal models (European Central Bank 2019). In accordance with Article 179 (1) (f) of the Corrigendum to regulation (EU) No 575/2013 on prudential requirements for credit institutions and investment firms (European Parliament and the Council of the European Union 2013), the approach provides larger margins of conservatism where less data induce larger likely ranges of error. Furthermore, the higher the discriminatory power of the credit risk model, the lower is the variance of the discriminatory power in our framework (cf. Figure 10). Therefore, our approach punishes credit risk models with low discriminatory power through higher margins of conservatism and, in so doing, incentivizes banks to improve the discriminatory power of their credit risk models.

In principle, bank internal and supervisory audits of credit risk models can involve empirical investigations, for example, based on challenger models and benchmarking data sets in order to complement the in-depth analyses of the underlying mathematical theory. As the development of challenger models usually requires significant resources, however, this approach is hardly compatible with time-limited supervisory audits (Dupont, Fliche, and Yang 2020). Against this backdrop, benchmarking data sets take on greater significance. Based on the framework described in the second-to-last paragraph, we can generate synthetic data sets of credit scores and default labels for which we know the AUROC, the variance of the AUROC, and the true posterior probabilities. After banks have processed these synthetic data sets, internal auditors and supervisors can check whether the banks' models appropriately reproduce the known output. If this is not the case, the underlying algorithms or their implementations might deserve closer attention. In so doing, the application of synthetic data sets could enhance the efficiency and effectiveness of bank internal and supervisory audits. However, the reproduction of metrics of synthetic data sets only is a necessary (rather than a sufficient) condition for the appropriateness of banks' internal models.

Our paper contains a number of limitations of which some offer avenues for future research. As an example, the paper neglects differences between the empirical and modelled CAPs, which might give rise to further uncertainties of the estimated conditional PDs.

References

Agarwal, Vineet, and Richard Taffler. 2008. "Comparing the performance of market-based and accounting-based bankruptcy prediction models." *Journal of Banking & Finance* 32 (8): 1541-1551. https://doi.org/10.1016/j.jbankfin.2007.07.014.

Alonso, Andrés, and José Manuel Carbó. 2020. "Machine learning in credit risk: Measuring the dilemma between prediction and supervisory cost." *Documentos de Trabajo/Banco de España, 2032*. https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/20/Files/dt2032e.pdf.

Aussenegg, Wolfgang, Florian Resch, and Gerhard Winkler. 2011. "Pitfalls and remedies in testing the calibration quality of rating systems." *Journal of Banking & Finance* 35 (3): 698-708. https://doi.org/10.1016/j.jbankfin.2010.11.016.

Barboza, Flavio, Herbert Kimura, and Edward Altman. 2017. "Machine learning models and bankruptcy prediction." *Expert Systems with Applications* 83: 405-417. https://doi.org/10.1016/j.eswa.2017.04.006.

Bazarbash, Majid. 2019. "FinTech in financial inclusion: Machine learning applications in assessing credit risk." *International Monetary Fund Working Papers*. https://www.imf.org/en/Publications/WP/Issues/2019/05/17/FinTech-in-Financial-Inclusion-Machine-Learning-Applications-in-Assessing-Credit-Risk-46883#:~:text=FinTech%20credit%20has%20the%20potential,4)%20predicting%20changes%20in%20general.

Bequé, Artem, Kristof Coussement, Ross Gayler, and Stefan Lessmann. 2017. "Approaches for credit scorecard calibration: An empirical analysis." *Knowledge-Based Systems* 134: 213-227. https://doi.org/10.1016/j.knosys.2017.07.034.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*. Springer New York.

Blümke, Oliver. 2020. "Estimating the probability of default for no-default and low-default portfolios." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 69 (1): 89-107. https://doi.org/10.1111/rssc.12381.

Böken, Björn. 2021. "On the appropriateness of Platt scaling in classifier calibration." *Information Systems* 95: 101641. https://doi.org/10.1016/j.is.2020.101641.

Bonini, Stefano, and Giuliana Caivano. 2018. "Probability of default modeling: A machine learning approach." In *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, edited by Marco Corazza, María Durbán, Aurea Grané, Cira Perna and Marilena Sibillo, 173-177. Springer, Cham.

Brunel, Vivien. 2019. "From the Fermi–Dirac distribution to PD curves." *Journal of Risk Finance* 20 (2): 138-154. https://doi.org/10.1108/jrf-01-2018-0009.

Butaru, Florentin, Qingqing Chen, Brian Clark, Sanmay Das, Andrew W. Lo, and Akhtar Siddique. 2016. "Risk and risk management in the credit card industry." *Journal of Banking & Finance* 72: 218-239. https://doi.org/10.1016/j.jbankfin.2016.07.015.

Caruana, Rich, and Alexandru Niculescu-Mizil. 2006. "An empirical comparison of supervised learning algorithms." 23rd International Conference on Machine Learning.

Dupont, Laurent, Olivier Fliche, and Su Yang. 2020. *Governance of artificial intelligence in finance.* (Autorité de contrôle prudentiel et de résolution - Banque de France). https://acpr.banque-france.fr/sites/default/files/medias/documents/20200612_ai_governance_finance.pdf.

Engelmann, Bernd, Evelyn Hayden, and Dirk Tasche. 2003. "Testing rating accuracy." *Risk* 16 (1): 82-86.

European Banking Authority. 2017. Guidelines on PD estimation, LGD estimation and treatment of defaulted assets.

---. 2019. *Progress report on the IRB roadmap.* https://www.eba.europa.eu/sites/default/documents/files/documents/10180/2551996/c1eb68d4-a084-486a-9434-70cd9ae43723/Progress%20report%20on%20IRB%20roadmap.pdf.
---. 2020. *EBA report on big data and advanced analytics.* https://www.eba.europa.eu/sites/default/files/document_library//Final%20Report%20on%20Big%20Data%20and%20Advanced%20Analytics.pdf.
European Central Bank. 2019. *ECB guide to internal models.* https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.guidetointernalmodels_consolidated_201910~97fd49fb08.en.pdf.
European Parliament and the Council of the European Union. 2013. Corrigendum to regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending regulation (EU) No 648/2012. In *575/2013*.
Falkenstein, Eric, Andrew Boral, and Lea V. Carty. 2000. RiskCalc for private companies: Moody's default model. Rating Methodology. Moody's Investors Service - Global Credit Research.
Fong, Youyi, and Ying Huang. 2019. "Modified Wilcoxon–Mann–Whitney test and power against strong null." *The American Statistician* 73 (1): 43-49. https://doi.org/10.1080/00031305.2017.1328375.
Fonseca, Pedro G., and Hugo D. Lopes. 2017. "Calibration of machine learning classifiers for probability of default modelling." https://arxiv.org/ftp/arxiv/papers/1710/1710.08901.pdf.
Hong Kong Monetary Authority, and PricewaterhouseCoopers. 2019. *Reshaping banking with artificial intelligence.* https://www.hkma.gov.hk/media/eng/doc/key-functions/finanical-infrastructure/Whitepaper_on_AI.pdf.
Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied logistic regression*. Vol. 3. John Wiley & Sons, Inc.
Huang, Cheng-Lung, Mu-Chen Chen, and Chieh-Jen Wang. 2007. "Credit scoring with a data mining approach based on support vector machines." *Expert Systems with Applications* 33 (4): 847-856. https://doi.org/10.1016/j.eswa.2006.07.007.
Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. 2010. "Consumer credit-risk models via machine-learning algorithms." *Journal of Banking & Finance* 34 (11): 2767-2787. https://doi.org/10.1016/j.jbankfin.2010.06.001.
Kruppa, Jochen, Alexandra Schwarz, Gerhard Arminger, and Andreas Ziegler. 2013. "Consumer credit risk: Individual probability estimates using machine learning." *Expert Systems with Applications* 40 (13): 5125-5131. https://doi.org/10.1016/j.eswa.2013.03.019.
Lawrenz, Jochen. 2008. "Assessing the estimation uncertainty of default probabilities." *Kredit und Kapital* 41 (2): 217.
Leathart, Tim, Eibe Frank, Geoffrey Holmes, and Bernhard Pfahringer. 2017. "Probability calibration trees." Ninth Asian Conference on Machine Learning.
Moro, Russ A., Wolfgang K. Härdle, and Dorothea Schäfer. 2017. "Company rating with support vector machines." *Statistics & Risk Modeling* 34 (1-2). https://doi.org/10.1515/strm-2012-1141.
Moscatelli, Mirko, Fabio Parlapiano, Simone Narizzano, and Gianluca Viggiano. 2020. "Corporate default forecasting with machine learning." *Expert Systems with Applications* 161: 113567. https://doi.org/10.1016/j.eswa.2020.113567.
Nehrebecka, Natalia. 2016. "Probability-of-default curve calibration and validation of internal rating systems." Eighth Irving Fisher Committee Conference on “Statistical implications of the new financial landscape”, Basel. https://www.bis.org/ifc/publ/ifcb43_zd.pdf.
Petropoulos, Anastasios, Vasilis Siakoulis, Evangelos Stavroulakis, and Nikolaos E. Vlachogiannakis. 2020. "Predicting bank insolvencies using machine learning techniques."

International Journal of Forecasting 36 (3): 1092-1113. https://doi.org/10.1016/j.ijforecast.2019.11.005.
Pfeuffer, Marius, Maximilian Nagl, Matthias Fischer, and Daniel Rösch. 2020. "Parameter estimation, bias correction and uncertainty quantification in the Vasicek credit portfolio model." *Journal of Risk* 22 (4): 1-30. https://doi.org/10.21314/jor.2020.429.
Roengpitya, Rungporn, and Pratabjai Nilla-Or. 2011. "Proposal of new hybrid models for PD estimates on low default portfolios (LDPs), empirical comparisons and regulatory policy implications." First International Conference on Credit Analysis and Risk Management.
Segoviano Basurto, Miguel Angel. 2006. "Consistent information multivariate density optimizing methodology." *International Monetary Fund Working Paper*.
Tasche, Dirk. 2010. "Estimating discriminatory power and PD curves when the number of defaults is small." https://arxiv.org/pdf/0905.3928.pdf.
Van der Burgt, Marco. 2008. "Calibrating low-default portfolios, using the cumulative accuracy profile." *Journal of Risk Model Validation* 1 (4): 17-33. https://doi.org/10.21314/jrmv.2008.016.
---. 2019. "Calibration and mapping of credit scores by riding the cumulative accuracy profile." *Journal of Credit Risk* 15 (1): 1-25. https://doi.org/10.21314/jcr.2018.240.
---. 2020. "How accurate is the accuracy ratio in credit risk model validation?" *Journal of Risk Model Validation* 14 (4): 41-63. https://doi.org/10.21314/JRMV.2020.229.
Van Gestel, Tony, Bart Baesens, Johan Suykens, Marcelo Espinoza, Dirk-Emma Baestaens, Jan Vanthienen, and Bart De Moor. 2003. "Bankruptcy prediction with least squares support vector machine classifiers." International Conference on Computational Intelligence for Financial Engineering.
Van Gestel, Tony, Bart Baesens, Peter Van Dijcke, Johan A. K. Suykens, Joao Garcia, and Thomas Alderweireld. 2005. "Linear and non-linear credit scoring by combining logistic regression and support vector machines." *Journal of Credit Risk* 1 (4): 31-60. https://doi.org/10.21314/JCR.2005.025.
Zadrozny, Bianca, and Charles Elkan. 2002. "Transforming classifier scores into accurate multiclass probability estimates." Eighth ACM SIGKDD international conference on Knowledge discovery and data mining.

Appendix

Throughout this paper, we follow the convention that low values of credit scores tend to indicate high default risk and vice versa. This appendix demonstrates that the conditional PD (i.e. $P(y = 1|s)$) does not monotonically decrease with increasing credit score if the credit scores of the defaults and non-defaults follow normal distributions with different variances. In order to verify this, we first specify the parameters and functions in equation (4) as suggested by Bishop (2006), i.e.

- $\vec{\beta}_Y = \begin{pmatrix} \frac{\mu_Y}{{\sigma_Y}^2} \\ \frac{-1}{2\cdot{\sigma_Y}^2} \end{pmatrix}$,
- $\vec{u}(s) = \begin{pmatrix} s \\ s^2 \end{pmatrix}$,
- $h(s) = \frac{1}{\sqrt{2\cdot\pi}}$, and
- $g(\vec{\beta}_Y) = \sqrt{-2\cdot\beta_2}\cdot exp\left[\frac{{\beta_1}^2}{4\cdot\beta_2}\right] = \frac{1}{\sqrt{{\sigma_y}^2}}\cdot exp\left[\frac{{\mu_Y}^2}{{\sigma_Y}^4}\cdot\frac{{\sigma_Y}^2}{-2}\right] = \frac{1}{\sigma_Y}\cdot exp\left[\frac{-{\mu_Y}^2}{2\cdot{\sigma_Y}^2}\right]$.

Under these specifications, the conditional PDFs, defined in equation (4), become univariate normal distributions as examples of exponential family distributions:

$$\begin{aligned} f_Y(s) &= h(s)\cdot g(\vec{\beta}_Y)\cdot \exp\left[{\vec{\beta}_Y}^T\cdot\vec{u}(s)\right] \\ &= \frac{1}{\sqrt{2\cdot\pi\cdot{\sigma_Y}^2}}\cdot\exp\left[\frac{-1}{2\cdot{\sigma_Y}^2}\cdot(s-\mu_Y)^2\right], \end{aligned} \tag{18}$$

where

- μ_Y stands for the expectation and
- ${\sigma_Y}^2$ denotes the variance.

Furthermore, the logit (cf. equation (5)) becomes a quadratic function of the credit score (cf. Section 1.5 in Hosmer et al. (2013)):

$$\begin{aligned} l(s) &= \left(\vec{\beta}_D - \vec{\beta}_{ND}\right)^T\cdot\vec{u}(s) + \ln\left\{\frac{P(Y=1)}{1-P(Y=1)}\right\} + \ln\left\{\frac{g(\vec{\beta}_D)}{g(\vec{\beta}_{ND})}\right\} \\ &= \left(\frac{1}{2\cdot{\sigma_{ND}}^2} - \frac{1}{2\cdot{\sigma_D}^2}\right)\cdot s^2 + \left(\frac{\mu_D}{{\sigma_D}^2} - \frac{\mu_{ND}}{{\sigma_{ND}}^2}\right)\cdot s + \ln\left\{\frac{P(Y=1)}{1-P(Y=1)}\right\} + \ln\left\{\frac{\sigma_{ND}}{\sigma_D}\right\} \\ &\quad + \frac{{\mu_{ND}}^2}{2\cdot{\sigma_{ND}}^2} - \frac{{\mu_D}^2}{2\cdot{\sigma_D}^2}. \end{aligned} \tag{19}$$

If we further specify the variances and expected values of the defaults and non-defaults as follows

$$\begin{aligned} \sigma_D &= \alpha\cdot\sigma_{ND} = \alpha\cdot\sigma, \\ \mu_D &= -\mu_{ND} = -\mu, \end{aligned} \tag{20}$$

then we can write the logit in a more compact form:

$$l(s) = \frac{\alpha^2 - 1}{2 \cdot \alpha^2 \cdot \sigma^2} \cdot s^2 - \frac{\mu \cdot (\alpha^2 + 1)}{\alpha^2 \cdot \sigma^2} \cdot s + \ln\left\{\frac{P(Y = 1)}{1 - P(Y = 1)}\right\} - \ln\{\alpha\} + \frac{\mu^2 \cdot (\alpha^2 - 1)}{2 \cdot \alpha^2 \cdot \sigma^2}. \quad (21)$$

The derivative of the conditional PD with respect to the credit score is:

$$\begin{aligned} \frac{d}{ds} P(Y = 1|s) &= \frac{d}{ds} \frac{1}{1 + \exp\{-l(s)\}} \\ &= l'(s) \cdot \frac{\exp\{-l(s)\}}{(1 + \exp\{-l(s)\})^2}, \end{aligned} \quad (22)$$

with

$$l'(s) = \frac{\alpha^2 - 1}{\alpha^2 \cdot \sigma^2} \cdot s - \frac{\mu \cdot (\alpha^2 + 1)}{\alpha^2 \cdot \sigma^2}. \quad (23)$$

The conditional PD is obviously not a strictly monotonically decreasing function of the credit score which conflicts with our convention:

$$\begin{aligned} &\frac{d}{ds} P(Y = 1|s) \geq 0 \\ \Leftrightarrow\ &\frac{\alpha^2 - 1}{\alpha^2 \cdot \sigma^2} \cdot s \geq \frac{\mu \cdot (\alpha^2 + 1)}{\alpha^2 \cdot \sigma^2} \\ \overset{\alpha > 1}{\Rightarrow}\ &s \geq \mu \cdot \frac{\alpha^2 + 1}{\alpha^2 - 1}. \end{aligned} \quad (24)$$

Figure 17 illustrates the impact of α larger than one on the PD as a function of the credit score. According to equation (21), the quadratic term in s vanishes if α is equal to one. In this case, the conditional PD is a strictly monotonically decreasing function of the credit spread as Figure 18 shows. Section 4 restricts itself to this case.

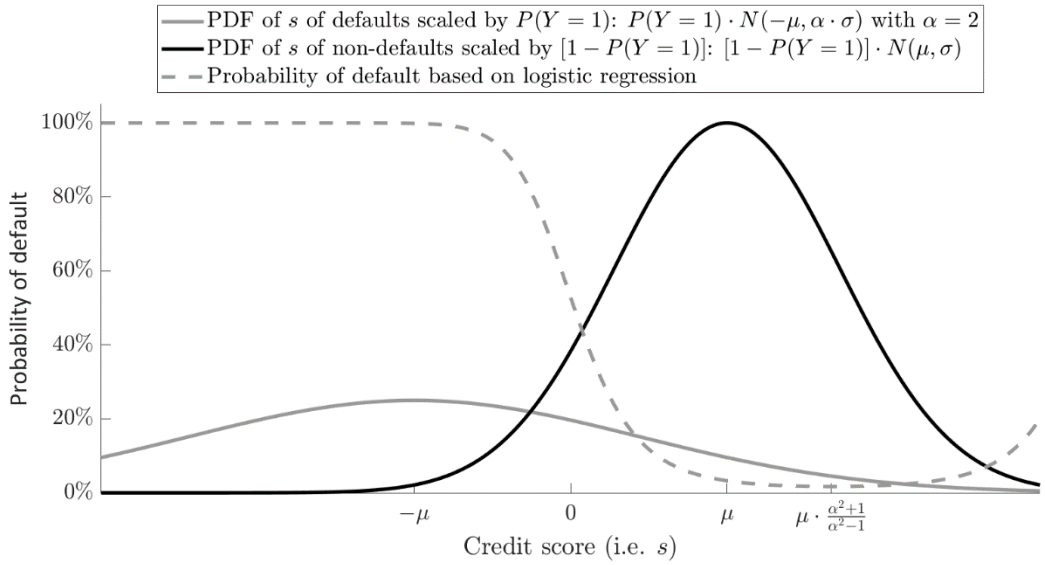

Figure 17: The difference between the standard deviations of the normally distributed credit scores of defaults and non-defaults by the factor $\alpha \neq 1$ induces a non-monotonic relationship between the credit score and PD.

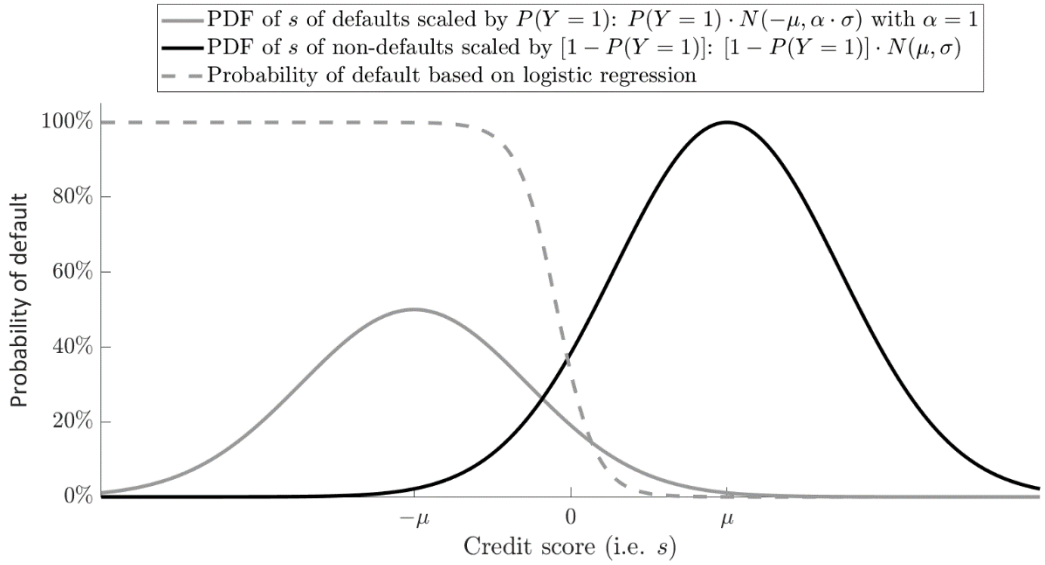

Figure 18: Equal standard deviations of the normally distributed credit scores of defaults and non-defaults ensure that the PD is a strictly monotonically decreasing function of the credit score.