

National Centre of Competence in Research
Financial Valuation and Risk Management

Working Paper No. 323

Testing Probability Calibrations: Application to Credit Scoring Models

Andreas Blöchinger

Markus Leippold

First version: July 2005
Current version: May 2006

This research has been carried out within the NCCR FINRISK project on
“Credit Risk and Non-Standard Sources of Risk in Finance”

Testing Probability Calibrations: Application to Credit Scoring Models*

Andreas Blöchlinger[†]

Credit Suisse

Markus Leippold[‡]

Federal Reserve Bank of New York and University of Zurich, Switzerland

First Version: July, 2005

This Version: May 6, 2006

*Preliminary version. The content of this paper reflects the personal view of the authors. In particular, it does not necessarily represent the opinion of Credit Suisse. Andreas Blöchlinger thanks the “quants” at Credit Suisse for valuable and insightful discussions. Markus Leippold acknowledges the financial support of the Swiss National Science Foundation (NCCR FINRISK) and of the University Research Priority Program “Finance and Financial Markets” of the University Zurich.

[†]**Correspondence Information:** Andreas Blöchlinger, Head of Credit Risk Analytics, Credit Suisse, Bleicherweg 33, CH-8070 Zurich, Switzerland, tel: +41 1 333 45 18, <mailto:andreas.bloechlinger@credit-suisse.com>

[‡]**Correspondence Information:** Markus Leippold, Swiss Banking Institute, University of Zürich, Switzerland, tel: +41 (01) 634 39 62, <mailto:leippold@isb.unizh.ch>

Testing Probability Calibrations: Application to Credit Scoring Models

Abstract

The validation of probability calibration is an inherently difficult task. We develop a testing procedure for credit-scoring models. The models comprise two components to check whether the ex-ante probabilities support the ex-post frequencies. The first component tests the level of the probability calibration under dependencies. In the long term, the number of events should equal the sum of assigned probabilities. The second component validates the shape, measuring the differentiation between high and low probability events. We construct a goodness-of-fit statistic for both level and shape together with a global statistics, which is asymptotically χ^2 -distributed.

JEL Classification Codes: C12, C52, and G21

KEY WORDS: Receiver Operating Characteristic (ROC); Credit scoring; Probability of Default (PD) validation; Basel Committee on Banking Supervision; Bernoulli mixture models.

In this paper we derive new test statistics for probability calibration. We focus on the probability calibration of credit-scoring models, although our validation procedure is much more general and is applicable to various other fields.¹ We define a credit-scoring system as one that has an ordinal measurement instrument that distinguishes between low and high default risk, i.e., the risk that a borrower does not comply with the contractual loan agreement. For risk-management purposes or for loan pricing, we need to map the ordinal score into either a metric measure or a probability of default (PD). This mapping is called PD quantification or probability calibration.

In today's competitive credit markets, the validation of the probability of default is a cornerstone of modern risk management. Inaccurately calibrated probabilities result in substantial losses, even if the inaccuracy is very small (see, Stein (2005) and Blöchlinger and Leippold (2005)). Furthermore, PDs directly enter the pricing of credit-rating dependent instruments such as bonds, loans, and credit derivatives. In addition to the competitive aspect, regulatory authorities such as the Basel Committee on Banking Supervision require banks to report the accuracy of their probability calibration. Badly calibrated models are penalized with higher regulatory capital.

In this paper, we derive test statistics that are not subject to the shortcomings of other currently available tests. Our testing procedure allows continuous PDs, we explicitly take default correlation into account, and we do not rely on Monte Carlo simulations or an approximation scheme. To define our test statistics, we argue along the following lines. In a well-calibrated model, the estimated default frequency is equivalent to the default probability. We transform this observation into a statistical hypothesis that allows a powerful testing procedure.

A well-calibrated model implies two testable properties. First, on average, it predicts the realized number of events. We refer this property as probability calibration with respect to the level. Second, on average, a well-calibrated model also forecasts the realized number of events for an arbitrary subpopulation (e.g., only observations with low probabilities). We refer to this second property as probability calibration with respect to the shape. In what follows, we derive test statistics for both calibrations and we present a global test statistic.

For the paper proceeds as follows. Section 1 presents the background for our study. Section 2 outlines the basic assumptions and definitions. In Section 3, we derive the test statistics for the level and shape, and we combine these two tests into a global test statistic. In Section 4, we provide a simulation study on the robustness of our proposed framework and compare it to the χ^2 -test of Hosmer and Lemeshow (1989). Section 5 concludes.

1. BACKGROUND

The Basel Committee on Banking Supervision (2005) (BIS) reviews in detail the studies on PD calibration tests and concludes:

“At present no really powerful tests of adequate calibration are currently available. Due to the correlation effects that have to be respected there even seems to be no way to develop such tests. Existing tests are rather conservative [...] or will only detect the most obvious cases of miscalibration.”

As the BIS notes, one of the main obstacles to backtesting PDs is the impact of correlation effects, which leads to default clusterings. When correlation effects are present, the default rates systematically exceed the critical values obtained from a model calibrated under the independence assumption. Therefore, these tests are rather conservative. At the same time, currently available tests that take into account correlation between defaults only allow us to detect relatively obvious cases of rating-system miscalibration.

There are additional shortcomings of the methods reviewed by the BIS. They are only applicable under grouping of debtors into rating classes or other weighting schemes. If the PD calibration is continuous in the sense that two debtors almost surely have different PDs, then all tests reviewed by the BIS fail. Furthermore, recent validation tests on PD calibration such as Balthazar (2004) rely heavily on simulation methods. Tasche (2003) circumvents the need for simulation, but his method only holds approximatively.

For the validation of probabilities of default, the BIS differentiates between two stages: validation of the discriminatory power of a rating system and validation of the accuracy of

the PD quantification. For the assessment of the discriminatory power, the usual technique is the Receiver Operating Characteristic (ROC), a technique originally used in medicine, engineering, and psychology to assess the performance of diagnostic systems and signal recovery techniques. A corresponding summary statistic is the area under the ROC curve (AUROC), which condenses the information of the ROC curve into one single number. Contrary to the discriminatory power of the rating system, the validation of a rating system's probability calibration is much more difficult and the methods for doing so are only in early development. However, against a backdrop of increasingly competitive loan markets, these methods have attracted considerable interest.

For probability validation, the BIS reviews the following tests: the binomial, normal, and χ^2 -test, and the traffic-lights approach recently suggested by Blochwitz, Hohl, and Wehn (2005). Efforts to account for dependence in the binomial test² yield tests of moderate power, even for low levels of correlation. Furthermore, the binomial test can be applied to only one single rating category at a time. For example, if we simultaneously test twenty categories at a 5% significance level, we must expect one erroneous rejection of the null hypothesis "correct probability calibration."

To check several credit categories simultaneously, we can use the χ^2 -test (or Hosmer and Lemeshow (1989) test). This test is based on the assumption of independence and a normal approximation. Therefore, the χ^2 -test gives also conservative results.

The normal test is a multiperiod test of the correctness of a default probability forecast for a single rating category. This test is applied under the assumption that the mean default rate does not vary much over time, and that default events in different years are independent. However, the test has only moderate power, particularly for short time series (for example, five years).

In contrast to the normal test, the traffic-lights approach is independent of any assumption of constant PDs over time. It is a multiperiod backtesting tool for a single rating category that is based on the assumption of cross-sectional and intertemporal independence of default

events. Thus, the traffic-lights approach is very conservative and gives more false alerts than it fails to detect bad calibrations.

2. ASSUMPTIONS AND DEFINITIONS

We make three basic assumptions on homogeneity, orthogonality, and monotonicity. We also introduce the definitions necessary to derive the test statistics for probability validation.

Our loan portfolio consists of n debtors. To each debtor i we assign a binary default indicator Y_i and a credit score S_i . We assume k systematic risk factors \mathbf{V} . $\mathbf{S}, \mathbf{Y} \in \mathbb{R}^n$ and $\mathbf{V} \in \mathbb{R}^k$ are random vectors on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

ASSUMPTION 1 (Homogeneity and Orthogonality): *The portfolio is homogeneous in the sense that the random vector $(\mathbf{S}, \mathbf{Y}, \mathbf{V})$ is exchangeable, i.e.,*

$$(S_1, \dots, S_n, Y_1, \dots, Y_n, V_1, \dots, V_k) \sim (S_{\Pi(1)}, \dots, S_{\Pi(n)}, Y_{\Pi(1)}, \dots, Y_{\Pi(n)}, V_1, \dots, V_k)$$

for any permutation $(\Pi(1), \dots, \Pi(n))$ of $(1, \dots, n)$. Furthermore, the conditional distributions of credit score S_i and default indicator Y_i are such that

$$(1) \quad S_i | \mathbf{S}, \mathbf{V}, \mathbf{Y} \sim S_i | Y_i,$$

$$(2) \quad Y_i | \mathbf{S}, \mathbf{V}, \mathbf{Y} \sim Y_i | S_i, \mathbf{V}.$$

The implications of Assumption 1 are the following. First, since we have a homogeneous loan portfolio, the probability of default does not depend on i . Therefore, we can write the PD function as

$$\text{PD}(s) = \mathbb{P}\{Y_i = 1 | S_i = s\}.$$

Second, equation (1) states that, conditional on the default indicator Y_i , the scores S_i form an independent sequence of random variables. To forecast the credit score, all information is contained in the default state.³

Finally, equation (2) states that defaults are correlated through their dependence on common factors, implying that the credit score does not subsume all the information generated by macroeconomic drivers.

ASSUMPTION 1 (Monotonicity): *The PD function is monotonic, so that either*

$$PD(s) \geq PD(t) \text{ for all } s \geq t \quad \text{or}$$

$$PD(s) \geq PD(t) \text{ for all } s \leq t.$$

In practice, the probability function $PD(s)$ is not observable, but we can estimate $PD(s)$ by a measurable function

$$(3) \quad \widetilde{PD}(s) : \mathbb{R} \rightarrow [0, 1].$$

A perfectly calibrated probability density function yields functional equivalence to the true PD function, i.e., $\widetilde{PD}(s)$ and $PD(s)$ is functionally equivalent if $\widetilde{PD}(s) = PD(s)$, for all $s \in \mathbb{R}$. However, in hypotheses testing, it is often unnecessary or even impossible to assume that something is true for every possible outcome. Therefore, we use the weaker property of almost sure equivalence, $\widetilde{PD}(s) = PD(s)$, for almost all $s \in \mathbb{R}$.

From a practical perspective, it is inherently impossible to distinguish between two PD functions that are almost surely equivalent. Therefore, to design our statistical validation procedure, we focus on two other important properties of the PD function, the level and the shape.

The PD level is an estimate of the long-run aggregate probabilities of default and serves as the first anchor for our test design.

DEFINITION 1 (Level Equivalence): *The PD functions $\widetilde{PD}(s)$ and $PD(s)$ are equivalent with respect to the PD level, if*

$$\int_{-\infty}^{\infty} \widetilde{PD}(s) d\mathbb{F}_S(s) = \int_{-\infty}^{\infty} PD(s) d\mathbb{F}_S(s),$$

where $\mathbb{F}_S(t) = \mathbb{P}\{S_i \leq t\}$.

The second anchor of the PD function is its shape, which allows us to distinguish between non-defaulters and defaulters. In the following, we write $S_D = (S_i | Y_i = 1)$ for the credit score of defaulters and $S_{ND} = (S_i | Y_i = 0)$ for the score of non-defaulters, respectively. The

distribution function of defaulters' and non-defaulters' $\mathbb{F}_{S_D}(t)$, and $\mathbb{F}_{S_{ND}}(t)$ are a function of $\text{PD}(s)$, i.e.,

$$(4) \quad \begin{aligned} \mathbb{F}_{S_D}(t) &= \mathbb{P}\{S_i \leq t | Y_i = 1\} = \int_{-\infty}^t \frac{1}{\mathbb{P}\{Y_i = 1\}} \mathbb{P}\{Y_i = 1 | S_i = s\} d\mathbb{F}_S(s) \\ &= \frac{\int_{-\infty}^t \text{PD}(s) d\mathbb{F}_S(s)}{\int_{-\infty}^{\infty} \text{PD}(s) d\mathbb{F}_S(s)}, \end{aligned}$$

$$(5) \quad \begin{aligned} \mathbb{F}_{S_{ND}}(t) &= \mathbb{P}\{S_i \leq t | Y_i = 0\} = \int_{-\infty}^t \frac{1}{\mathbb{P}\{Y_i = 0\}} \mathbb{P}\{Y_i = 0 | S_i = s\} d\mathbb{F}_S(s) \\ &= \frac{\int_{-\infty}^t [1 - \text{PD}(s)] d\mathbb{F}_S(s)}{1 - \int_{-\infty}^{\infty} \text{PD}(s) d\mathbb{F}_S(s)}. \end{aligned}$$

If the credit score S_i and default indicator Y_i are two independent random variables, then the non-defaulters' and defaulters' distribution function coincide with the unconditional distribution function of the credit score. In this case, the credit score has no discriminatory power. We can visualize the discriminatory power using the Receiver Operating Characteristic (ROC) curve. The ROC curve is a two-dimensional graph generated by the survival functions for non-defaulters and defaulters,

$$(6) \quad \{1 - \mathbb{F}_{S_{ND}}(t), 1 - \mathbb{F}_{S_D}(t)\} \text{ for all } t \in \mathbb{R}, .$$

The range of the ROC curve is restricted to the unit square. Accordingly, the area below the curve, the AUROC, is limited from above by one and from below by zero. The AUROC depends on the PD function as follows:

$$(7) \quad \begin{aligned} \text{AUROC} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[\mathbf{1}_{\{x > y\}} + \frac{1}{2} \mathbf{1}_{\{x = y\}} \right] d\mathbb{F}_{S_D}(x) d\mathbb{F}_{S_{ND}}(y) \\ &= \mathbb{P}\{S_D > S_{ND}\} + \frac{1}{2} \mathbb{P}\{S_D = S_{ND}\} \\ &= \frac{1}{2} [1 - \mathbb{P}\{S_D < S_{ND}\} + \mathbb{P}\{S_D > S_{ND}\}]. \end{aligned}$$

The second equality in (7) follows from orthogonality in Assumption 1 and the last equality from $1 - \mathbf{1}_{\{x < y\}} = \mathbf{1}_{\{x > y\}} + \mathbf{1}_{\{x = y\}}$. The AUROC serves as our quantitative measure for the shape equivalence defined below.

DEFINITION 2 (Shape Equivalence): *Two PD functions $\widetilde{\text{PD}}(s)$ and $\text{PD}(s)$ are equivalent with respect to the PD shape, if*

$$\widetilde{\text{AUROC}} = \text{AUROC}.$$

With all the definitions at hand, we can now state the following relationships between functional equivalence, almost-sure equivalence, level equivalence, and shape equivalence.

THEOREM 1: *Let $\widetilde{PD}(s)$ and $PD(s)$ be two PD functions.*

- a) If the two PD functions are functionally equivalent, they are also almost surely equivalent.*
- b) If the two PD functions are almost surely equivalent, they are also equivalent with respect to the PD level.*
- c) If the two PD functions are almost surely equivalent, they are also equivalent with respect to the PD shape.*

The Appendix provides the proof of Theorem 1 and all subsequent results.

From Theorem 1 it follows that two functionally equivalent PD functions have the same level and shape. We note that above, we derive the ROC curve from the score contribution and the PD function. Conversely, we could also derive the PD function and the score distribution from the distribution of defaulters and non-defaulters' distribution, and unconditional default probability (PD level) (see Appendix B).

3. STATISTICAL INFERENCE

In this section, we derive statistical tests to validate one-period default probabilities. For expository purposes, we restrict ourselves to the one-period setting. However, our test statistics can be directly extended to a multiperiod setting, e.g., along the lines of Blochwitz, Hohl, and Wehn (2005)'s traffic-lights approach.

3.1. Testing of PD Level

For illustrative purposes, we assume an approximate distribution for the one-period default frequency $\hat{\pi}$. A common choice is the β -distribution,

$$(8) \quad \mathbb{P} \{ \hat{\pi} \leq t \} \cong \int_0^t \beta(a, b)^{-1} z^{a-1} (1-z)^{b-1} dz,$$

where $\beta(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$. A calibration exercise then gives us the values for a and b .

We begin with restrictive distributional assumptions and relax these step by step. In particular, we proceed by deriving test statistics with four different distributional constraints.

3.1.1. Case I: $Y_i|\mathbf{S}, \mathbf{V}, \mathbf{Y} \sim Y_i$

We assume that the default indicator is orthogonal to the credit scores, the systematic factors, and to all other debtors' default indicator, i.e., Y_i forms an iid Bernoulli sequence with parameter π . To obtain the limiting distribution, we first calculate the number of defaults N_1 as

$$(9) \quad N_1 \sim B(n, \pi).$$

Then, according to the De-Moivre-Laplace global limit theorem we obtain

$$(10) \quad \lim_{n \rightarrow \infty} \mathbb{P} \left\{ \frac{N_1 - n\pi}{\sqrt{n\pi(1-\pi)}} \leq t \right\} = \Phi(t).$$

From the basic convergence theorem of Cramér (1946),⁴ we can replace the theoretical standard deviation with the empirical one, which is still asymptotic Gaussian,

$$(11) \quad \lim_{n \rightarrow \infty} \mathbb{P} \left\{ \frac{N_1 - n\pi}{\sqrt{\frac{n}{n-1}n\hat{\pi}(1-\hat{\pi})}} \leq t \right\} = \Phi(t).$$

3.1.2. Case II: $Y_i|\mathbf{S}, \mathbf{V}, \mathbf{Y} \sim Y_i|\mathbf{V}$

Next, we allow for default clustering through the supposition of a Bernoulli mixture model. Empirical evidence on defaults suggests that the basic assumption of the binomial model is not fulfilled, because borrower defaults tend to default together. In a mixture model, the default probability of an debtor depends on a set of common factors (typically one). This dependence can cause dependence of defaults.

DEFINITION 3 (Bernoulli Mixture Model): *Given a k dimensional random vector $\mathbf{V} = (V_1, \dots, V_k)'$, the random vector $\mathbf{Y} = (Y_1, \dots, Y_n)'$ follows a Bernoulli mixture model, if there are*

functions $Q_i : \mathbb{R}^k \rightarrow [0, 1]$, such that conditional on \mathbf{V} the default indicators \mathbf{Y} form a vector of independent Bernoulli random variables with $\mathbb{P}\{Y_i = 1 | \mathbf{V}\} = Q_i(\mathbf{V})$.

Given our homogeneity assumption, the functions $Q_i(\mathbf{V})$ are all identical and $\mathbb{P}\{Y_i = 1 | \mathbf{V}\} = Q(\mathbf{V})$ for all i . Here, we find it convenient to introduce the random variable $Z = Q(\mathbf{V})$. By G we denote the distribution function of Z . To calculate the unconditional distribution of the number of defaults N_1 , we integrate over the mixing distribution of Z to get

$$(12) \quad \mathbb{P}\{N_1 = m\} = \binom{n}{m} \int_0^1 z^m (1-z)^{n-m} dG(z).$$

Then, we obtain the probability of default π and the joint probability of default π_2 as

$$\begin{aligned} \pi &= \mathbb{P}\{Y_i = 1\} \\ &= \mathbb{E}[Y_i] = \mathbb{E}[\mathbb{E}[Y_i | Z]] = \mathbb{E}[\mathbb{P}\{Y_i = 1 | Z\}] = \mathbb{E}[Z], \\ \pi_2 &= \mathbb{P}\{Y_i = 1, Y_j = 1\} \\ &= \mathbb{E}[Y_i Y_j] = \mathbb{E}[\mathbb{E}[Y_i Y_j | Z]] = \mathbb{E}[\mathbb{P}\{Y_i = 1, Y_j = 1 | Z\}] = \mathbb{E}[Z^2]. \end{aligned}$$

where $i \neq j$. Moreover, for $i \neq j$,

$$\rho_Y = \text{COV}[Y_i, Y_j] = \pi_2 - \pi^2 = \mathbb{V}[Z] \geq 0.$$

Hence, in an exchangeable Bernoulli mixture model the so-called default correlation ρ_Y is always nonnegative. In practice, the following one-factor exchangeable Bernoulli mixture models are frequently used:

- Probit-normal mixing-distribution with $Z = \Phi(V)$ and $V \sim N(\mu, \sigma^2)$ (CreditMetrics and KMV-type models; see Gupton, Finger, and Bhatia (1997) and Crosbie (1997)),
- Logit-normal mixing-distribution with $Z = \frac{1}{1+\exp(-V)}$ and $V \sim N(\mu, \sigma^2)$ (CreditPortfolioView model; see Wilson (1998)),
- Beta mixing-distribution with $Z \sim \text{Beta}(a, b)$ with density $g(z) = \beta(a, b)^{-1} z^{a-1} (1-z)^{b-1}$, where $\beta(a, b)$ is the beta function and $a, b > 0$ (see Frey and McNeil (2001)).

With a beta mixing-distribution the number of defaults N_1 has a so-called beta-binomial distribution with probability function

$$\begin{aligned}
 \mathbb{P}\{N_1 = m\} &= \binom{n}{m} \frac{1}{\beta(a, b)} \int_0^1 z^{a+m-1} (1-z)^{b+n-m-1} dz \\
 (13) \qquad &= \binom{n}{m} \frac{\beta(a+m, b+n-m)}{\beta(a, b)},
 \end{aligned}$$

where the second line follows from the definition of the β -function. If Z follows a beta-distribution then the expectation and variance are given by

$$\begin{aligned}
 \mathbb{E}[Z] &= \frac{a}{a+b} \\
 \mathbb{V}[Z] &= \frac{ab}{(a+b)^2(a+b+1)}.
 \end{aligned}$$

Thus given two of the following three figures, the unconditional probability of default $\pi = \mathbb{E}[Z]$, the joint probability of default $\pi_2 = \mathbb{E}[Z^2]$ and/or the default correlation $\rho_Y = \mathbb{V}[Z]$ we can calibrate the beta-distribution,

$$\begin{aligned}
 a &= \mathbb{E}[Z] \left[\frac{\mathbb{E}[Z]}{\mathbb{V}[Z]} (1 - \mathbb{E}[Z]) - 1 \right] \\
 b &= a \frac{1 - \mathbb{E}[Z]}{\mathbb{E}[Z]}.
 \end{aligned}$$

Bernoulli mixture models are often calibrated via the asset correlation ρ (e.g. CreditMetrics) and are motivated by the paper of Merton (1974). The following proposition shows how asset correlation and default correlation are related.

PROPOSITION 1: *Given a homogeneous portfolio, the unconditional probability of default π as well as the asset correlation ρ in the one-factor CreditMetrics framework, we can calculate the joint probability of default π_2 , and the default correlation ρ_Y as*

$$\begin{aligned}
 \pi_2 &= \Phi_2(\Phi^{-1}(\pi), \Phi^{-1}(\pi), \rho) \\
 \rho_Y &= \Phi_2(\Phi^{-1}(\pi), \Phi^{-1}(\pi), \rho) - \pi^2,
 \end{aligned}$$

where $\Phi_2(\cdot, \cdot, \rho)$ denotes the bivariate standard Gaussian distribution function with correlation ρ , $\Phi(\cdot)$ is the distribution function of a standard Gaussian variable, and $\Phi^{-1}(\cdot)$ denotes the corresponding quantile function.

For an exchangeable Bernoulli mixture model and if the portfolio is large enough, the quantiles of the number of defaulters are essentially determined by the quantiles of the mixing distribution.

PROPOSITION 2: We denote by $G^{-1}(\alpha)$ the α -quantile of the mixing distribution G of Z , i.e. $G^{-1}(\alpha) = \inf \{z : G(z) \geq \alpha\}$, and assume that the quantile function $\alpha \rightarrow G^{-1}(\alpha)$ is continuous in α , so that

$$(14) \quad G(G^{-1}(\alpha) + \delta) > \alpha \text{ for all } \delta > 0,$$

then

$$\lim_{n \rightarrow \infty} \mathbb{P} \{ \hat{\pi} \leq G^{-1}(\alpha) \} = \mathbb{P} \{ Z \leq G^{-1}(\alpha) \} = \alpha.$$

In particular, if G admits a density g (continuous random variable), which is positive on $[0, 1]$, the condition (14) is satisfied for any $\alpha \in (0, 1)$.

3.1.3. Case III: $Y_i | \mathbf{S}, \mathbf{V}, \mathbf{Y} \sim Y_i | S_i$

Next, we work under the assumption that default indicators $Y_i | S_i$ represent an independent and uniformly bounded sequence, since $|Y_i| \leq 1$ for each i . Hence, the Lindeberg condition is satisfied and the number of defaulters N_1 converges to a Gaussian distribution (see i.e. Proposition 7.13. of Karr (1993)), so that

$$(15) \quad \lim_{n \rightarrow \infty} \mathbb{P} \left\{ \frac{N_1 - \mathbb{E}[N_1 | \mathbf{S}]}{\sqrt{\mathbb{V}[N_1 | \mathbf{S}]}} < t \mid \mathbf{S} \right\} = \Phi(t),$$

where

$$\begin{aligned} \mathbb{E}[N_1 | \mathbf{S}] &= \sum_{i=1}^n \mathbb{P} \{ Y_i = 1 | S_i \} \\ \mathbb{V}[N_1 | \mathbf{S}] &= \sum_{i=1}^n \mathbb{P} \{ Y_i = 1 | S_i \} \mathbb{P} \{ Y_i = 0 | S_i \}. \end{aligned}$$

3.1.4. Case IV: $Y_i|\mathbf{S}, \mathbf{V}, \mathbf{Y} \sim Y_i|S_i, \mathbf{V}$

Case IV is the most general setup. Defaults are clustered in the sense that the default indicator depends on the business cycle. Then,

$$(16) \quad \mathbb{P}\{N_1 = m|\mathbf{S}\} = \int_{\mathbb{R}^k} \sum_P \prod_{i=1}^n \mathbb{P}\{Y_i = \Pi(i)|S_i, \mathbf{V} = \mathbf{v}\} d\mathbb{F}_{\mathbf{V}}(\mathbf{v}),$$

where $\mathbb{F}_{\mathbf{V}}(\mathbf{v})$ denotes the distribution function of \mathbf{V} . P denotes the set of the permutations with m ones and $n-m$ zeros $\{\Pi(1), \dots, \Pi(m), \Pi(m+1), \dots, \Pi(n)\}$ of $\{1, \dots, 1, 0, \dots, 0\}$. Usually, the derivation of the distribution of (16) requires Monte-Carlo simulations or numerical integration procedures. Therefore, we approximate the distribution by the beta-binomial distribution derived in (13). To calibrate the beta-binomial distribution, we fix the asset correlation ρ and we set π equal to the average default probability

$$(17) \quad \begin{aligned} \pi &= \int_{-\infty}^{\infty} \text{PD}(s) d\mathbb{F}_S(s) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{P}\{Y_i = 1|S_i = s\}. \end{aligned}$$

The choice of the parameter ρ is not so obvious. The higher ρ , the more do defaults cluster in time.⁵

If the level testing of the PD functions spans a long period of time, possibly a whole credit cycle, then the independence assumption for the test statistics in equations (9), (10), (11), and (15) is warranted. By assuming mean ergodicity for the default process, the average yearly default rate over a business cycle converges to the unconditionally expected default frequency, and within a cycle, the defaults are approximately uncorrelated. Even more subtly, if the yearly default events are stochastically dependent, but if the annual default rates \bar{p}_t are uncorrelated over time, then the quotient

$$(18) \quad \frac{\sum_{t=1}^T (\bar{p}_t - \mathbb{E}[\bar{p}_t|\mathcal{F}_{t-1}])}{\sqrt{\sum_{t=1}^T \mathbb{V}[\bar{p}_t|\mathcal{F}_{t-1}]}}$$

where \mathcal{F}_t is a filtration, converges in distribution to a standard Gaussian random variable. On the other hand, if the aim is to make inference on short time intervals (typically, on a yearly basis), then we must take default correlations into account. In this instance, the test statistics in (13) and (16) are more appropriate.

3.2. Testing of PD Shape

The shape of the PD function is visualized by the ROC curve. We can plot the realized or empirical ROC curve against the theoretical ROC graph and detect PD miscalibrations visually. Therefore, the empirical ROC curve

$$\left\{ 1 - \hat{\mathbb{F}}_{S_{ND}}(t), 1 - \hat{\mathbb{F}}_{S_D}(t) \right\} \text{ for all } t \in \mathbb{R},$$

where

$$\hat{\mathbb{F}}_{S_D}(t) = \frac{\sum_{i:Y_i=1} \mathbf{1}_{\{S_i \leq t\}}}{\sum_{i=1}^n Y_i} \quad \text{and} \quad \hat{\mathbb{F}}_{S_{ND}}(t) = \frac{\sum_{j:Y_j=0} \mathbf{1}_{\{S_j \leq t\}}}{\sum_{j=1}^n (1 - Y_j)},$$

can be compared to the theoretical ROC curve as defined in equation (6). We note that the empirical distribution functions are unbiased since

$$\mathbb{E} [\mathbf{1}_{\{S_i \leq t\}} | \mathbf{V}, \mathbf{Y}] = \mathbb{E} [\mathbf{1}_{\{S_i \leq t\}} | Y_i] = \mathbb{P} \{S_i \leq t | Y_i\},$$

where the first equality follows by orthogonality (Assumption 1).

The empirical and true ROC curves are, under the assumptions outlined in Section 2, asymptotically equivalent:

THEOREM 2: *The empirical and theoretical ROC curves almost surely converge, so that*

$$\sup_{0 \leq \beta \leq 1} \left| \hat{\mathbb{F}}_{S_D} \left(\hat{\mathbb{F}}_{S_{ND}}^{-1}(1 - \beta) \right) - \mathbb{F}_{S_D} \left(\mathbb{F}_{S_{ND}}^{-1}(1 - \beta) \right) \right| \rightarrow 0,$$

as $n \rightarrow \infty$.

For example, consider the situation in which the assigned default probabilities are too low for investment-graded obligors (too high for sub-investment-rated borrowers), but well calibrated with respect to the level. Then, we expect the empirical ROC curve to be below the theoretical ROC curve implied by the PD function. Consequently, the area below the curve is lower than expected.

PROPOSITION 3: *If we have two monotonic PD functions $\widetilde{PD}(s)$ and $PD(s)$, so that*

$$(19) \quad \widetilde{PD}(s) \leq PD(s) \text{ for all } s \in \mathbb{S}$$

$$(20) \quad \widetilde{PD}(s) \geq PD(s) \text{ for all } s \in \mathbb{S}^c,$$

for any $\mathbb{S} \subset \mathbb{R}$, where all elements in \mathbb{S} are smaller than the elements in \mathbb{S}^c , and if the inequalities are strict in (19) and (20) for some s with positive probability measure, so that

$$(21) \quad 0 < \int_{\mathbb{S}} \widetilde{PD}(s) d\mathbb{F}_S(s) < \int_{\mathbb{S}} PD(s) d\mathbb{F}_S(s)$$

$$(22) \quad 0 < \int_{\mathbb{S}^c} PD(s) d\mathbb{F}_S(s) < \int_{\mathbb{S}^c} \widetilde{PD}(s) d\mathbb{F}_S(s),$$

and if the two PD functions have the same PD level, so that $\int_{-\infty}^{\infty} \widetilde{PD}(s) d\mathbb{F}_S(s) = \int_{-\infty}^{\infty} PD(s) d\mathbb{F}_S(s)$, then

$$\widehat{AUROC} > AUROC$$

The estimator for the empirical AUROC figure, \widehat{AUROC}_n , is given by

$$\widehat{AUROC}_n = \frac{1}{N_0 N_1} \sum_{i=1}^{N_1} \sum_{j=1}^{N_0} \left[\mathbf{1}_{\{S_{D_i} > S_{ND_j}\}} + \frac{1}{2} \mathbf{1}_{\{S_{D_i} = S_{ND_j}\}} \right],$$

where the index i (j) indicates summation over defaulters (non-defaulters) and $N_1 = \sum_{i=1}^n Y_i$ and $N_0 = \sum_{i=1}^n (1 - Y_i)$ denote the number of defaulters and non-defaulters, respectively. The AUROC estimator is consistent and unbiased:

PROPOSITION 4: *The (conditional) expectation and variance of the estimator \widehat{AUROC}_n is equal to*

$$\begin{aligned} \mathbb{E} \left[\widehat{AUROC}_n | \mathbf{Y} \right] &= AUROC \\ \mathbb{V} \left[\widehat{AUROC}_n | \mathbf{Y} \right] &= \frac{1}{4N_0 N_1} [B + \{N_1 - 1\} B_{110} + \{N_0 - 1\} B_{001} \\ &\quad - 4 \{N_0 + N_1 - 1\} \{AUROC - 0.5\}^2]. \end{aligned}$$

Furthermore,

$$\begin{aligned} B &= \mathbb{P} \{S_D \neq S_{ND}\} \\ B_{110} &= \mathbb{P} \{S_{D_1}, S_{D_2} < S_{ND}\} + \mathbb{P} \{S_{ND} < S_{D_1}, S_{D_2}\} \\ &\quad - \mathbb{P} \{S_{D_1} < S_{ND} < S_{D_2}\} - \mathbb{P} \{S_{D_2} < S_{ND} < S_{D_1}\} \\ B_{001} &= \mathbb{P} \{S_{ND_1}, S_{ND_2} < S_D\} + \mathbb{P} \{S_D < S_{ND_1}, S_{ND_2}\} \\ &\quad - \mathbb{P} \{S_{ND_1} < S_D < S_{ND_2}\} - \mathbb{P} \{S_{ND_2} < S_D < S_{ND_1}\}. \end{aligned}$$

We note that we compute the corresponding event probabilities for the calculation of B , B_{001} , and B_{110} out of the distribution functions $\mathbb{F}_{S_{ND}}(t)$ and $\mathbb{F}_{S_D}(t)$, respectively, e.g.,

$$\mathbb{P}\{S_D \neq S_{ND}\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbf{1}_{\{x \neq y\}} d\mathbb{F}_{S_D}(x) d\mathbb{F}_{S_{ND}}(y).$$

The limiting distribution of \widehat{AUROC}_n is Gaussian:

PROPOSITION 5: *The AUROC statistic has the following limiting distribution*

$$(23) \quad \lim_{n \rightarrow \infty} \mathbb{P} \left\{ \frac{\widehat{AUROC}_n - AUROC}{\sqrt{\mathbb{V}[\widehat{AUROC}_n | \mathbf{Y}]}} \middle| \mathbf{Y} \right\} = \Phi(t).$$

The theoretical standard deviation in the denominator in equation (23) of Proposition 5 can be replaced by the empirical counterpart. The limiting distribution remains Gaussian, according to a basic theorem of Cramér (1946) (Theorem 20.6, see also Bamber (1975)). Proposition 4 and Proposition 5 generalize Wilcoxon (1945) and Mann and Whitney (1947). Their results are applicable if there is a “horizontal” PD function.⁶

COROLLARY 1 (Wilcoxon-Mann-Whitney): *If S_{D_i} and S_{ND_j} form two independent as well as identically and continuously distributed sequences and if they are independent among one another then*

$$\begin{aligned} \mathbb{E}[\widehat{AUROC}_n | \mathbf{Y}] &= \frac{1}{2} \\ \mathbb{V}[\widehat{AUROC}_n | \mathbf{Y}] &= \frac{N_1 + N_0 + 1}{12N_1N_0}, \end{aligned}$$

with the limiting distribution

$$\lim_{n \rightarrow \infty} \mathbb{P} \left\{ \frac{\widehat{AUROC}_n - \frac{1}{2}}{\sqrt{\frac{N_1 + N_0 + 1}{12N_1N_0}}} \middle| \mathbf{Y} \right\} = \Phi(t).$$

3.3. Goodness-of-Fit

In the previous section, we derived level and shape statistics. But usually, the limiting distributions of the test statistics are standard normal. If the distribution is (asymptotically) different from a standard Gaussian, we can transform the realized estimate into a standard normal quantile according to the following lemma.

LEMMA 1: *If the random variable X is distributed according to the continuous distribution function G , then*

$$\mathbb{P}\{\Phi^{-1}(G(X)) \leq t\} = \Phi(t)$$

for all $t \in \mathbb{R}$.

We base this shape statistic on scores conditional on the default indicators. According to the orthogonality assumptions (Assumption 1), this distribution is unaffected by both the number of defaulters N_1 and the business cycle \mathbf{V} , i.e., it is true for all i that⁷

$$(24) \quad S_i | \mathbf{S}, \mathbf{V}, \mathbf{Y} \sim S_i | Y_i, N_1, \mathbf{V} \sim S_i | Y_i.$$

Therefore, the level and shape statistics are independent. On average, a high figure in the PD level statistic does not imply a high (or a low) number for the PD shape statistic.

We can now deduce a summary statistic to globally test the null hypothesis of a correctly calibrated PD function for both level and shape. When performing two independent significance tests, each with size α , the probability of making at least one type I error (rejecting the null hypothesis inappropriately) is $1 - (1 - \alpha)^2$. If there is a 5% significance level, there is a chance of 9.75% that at least one of the two tests will be declared significant under the null hypothesis. One very simple method, attributable to Bonferroni (1936), used to circumvent this problem is to divide the test-wise significance level by the number of tests. Unfortunately, Bonferroni's method does not generally result in the most powerful test, meaning that there are critical regions with the same size, but at a higher power according to Neyman-Pearson's lemma, which is why we resort to the likelihood ratio Λ ,

$$(25) \quad \Lambda = \exp \left[-\frac{1}{2} (T_{level}^2 + T_{shape}^2) \right],$$

where T_{level} denotes one of the level statistics in Section 3.1 and T_{shape} denotes the shape statistic in (23).

We first transform all the statistics into a standard Gaussian quantile according to Lemma 1. The likelihood-ratio test rejects the null hypothesis if the value of the statistic in (25) is too small, and is justified by the Neyman-Pearson lemma. If the null hypothesis is true, then

$-2 \log \Lambda$ will be asymptotically χ^2 -distributed with degrees of freedom equal to the difference in dimensionality. Hence, we derive asymptotically

$$(26) \quad T_{level}^2 + T_{shape}^2 \sim \chi^2 \langle 2 \rangle.$$

Therefore, the critical value for the global test in equation (26) on a confidence level of 95% (99%) is 5.9915 (9.2103).

4. SIMULATION STUDY

In this simulation study, we make robustness checks for violations of the assumptions in Section 2 that underly our test statistics. For this purpose, we simulate the true type I error (size of the test) and type II error (power of the test) at given nominal levels. We then compare the performance of our approach to the performance of a benchmark statistic, the Hosmer-Lemeshow's χ^2 -goodness-of-fit test (see e.g. Hosmer, Hosmer, le Cessie, and Lemeshow (1997)). Hosmer-Lemeshow's χ^2 -test statistic is defined as

$$(27) \quad T = \sum_{j=1}^C \frac{n_j (\hat{\pi}_j - \pi_j)^2}{\pi_j (1 - \pi_j)},$$

where $\hat{\pi}_j$ are observed default rates, π_j are corresponding expected rates, n_j are the number of observations in class j and C is the number of classes for which frequencies are being analyzed. The test statistic is distributed approximately as a χ^2 random variable with C degrees of freedom.

A common feature of Hosmer-Lemeshow's χ^2 -test and our test statistic is that they are both suitable for application to several rating categories simultaneously. Hosmer-Lemeshow's χ^2 -test is based on the assumption of independence and a normal approximation. Given the empirical evidence on default dependency and the low frequency of default events, Hosmer-Lemeshow's χ^2 -test is likely to underestimate the true type I error. Therefore, the proportion of erroneous rejections of PD forecasts will be higher than expected from the formal confidence level of the test.

Both Hosmer-Lemeshow's χ^2 -test and our global test statistic are derived under asymptotic considerations with regard to the portfolio size. As a consequence, even in the case of default

independence, it is not clear that the type I errors we observe with the tests are dominated by the nominal error levels. When compliance with the nominal error level for the type I error is confirmed, we must ask which test is more powerful, i.e., for which test is the type II errors lower. Of course, complying with the nominal error level is much more of an issue if there are dependencies of the default events in the portfolio.

Next, we examine the simulation setup to address the question of size and power of the test statistics under various settings. To generate default correlation, we model the asset value Y_i^* for each debtor i ,

$$Y_i^* = \sqrt{\rho}X + \sqrt{1 - \rho}\epsilon_i,$$

where ϵ_i form an independent sequence that is also orthogonal to the systematic risk driver X . Both X and ϵ_i follow a standard Gaussian distribution. We denote the asset correlation between two debtors by ρ . The higher the asset correlation, the more the systematic risk factor X dominates. The default event is defined by

$$(28) \quad Y_i = \begin{cases} 0 & : Y_i^* > D_i \\ 1 & : Y_i^* \leq D_i \end{cases},$$

where D_i denotes the distance to default calculated by the standard Gaussian quantile of the default probability. It is the same value for all debtors in a given rating category. For the simulation study, we assume that D_i is orthogonal to both X and ϵ_i . Therefore, we can think of the distance to default as a "through the business cycle" credit score.

We consider four different correlation regimes (0, 0.05, 0.10, and 0.15) and three different sizes of rating classes (15, 10, and 5) resulting in 12 scenarios. We run 10,000 Monte Carlo simulations under each scenario. The (unconditional) expected default frequency under the data generating process is fixed for all scenarios at 3% (the average default probability is 2.5% for the type II error analysis), and the size of the portfolio is set at 10,000 debtors. The true (alternative) AUROC figures are 0.6112, 0.6279, and 0.6509 (0.6354, 0.6551, and 0.6816) for 15, 10, and 5 rating classes, respectively. Table 1 reports the rating distribution with the assigned rating class PDs under the null hypotheses (the data-generating distributions) and the alternative hypotheses.

For the composition of the global test statistic in (26), we rely on a beta-approximation for testing the level T_{level} as in equation (13) and on the statistic T_{shape} in equation (23) for testing the shape. We calibrate the beta-binomial distribution according to Proposition 1 with an average default probability of 3% (2.5%), as computed by equation (17), for the type I error analysis (type II) and a fixed asset correlation ρ of 5% for all but one of the correlation regime. Doing so gives us the parameters $a = 3.4263$ (3.2203) and $b = 110.7850$ (125.5922) for type I error considerations (type II). If there is a of zero asset correlation we omit the "beta"-approximation we work with the approximate level statistic as outlined in (15).

Tables 2 and 3 report the simulation results under nominal error levels of 5% and 1%, respectively. The results indicate that under independence all test methods, Hosmer-Lemeshow's χ^2 , global, level, and shape statistics, seem to mostly comply with the nominal error levels. However Hosmer-Lemeshow's χ^2 test fits the levels less well than do our test statistics: the true type I errors are, in absolute terms, up to 3% higher than the nominal levels. Under asset correlation regimes below or equal to 5%, the global test statistic still essentially complies with the nominal error levels, but Hosmer-Lemeshow's χ^2 -test is distorted. Once we establish compliance with the nominal type I error, we can assess the power of the test statistics via the type II error. Again, the global test procedure is more powerful under independence with true type II error levels around 10% (23%) at 5% (1%) nominal level. In comparison, Hosmer-Lemeshow's χ^2 -test results in type II errors of up to about 37% (55%).

Under asset correlation regimes above 5%, both the Hosmer-Lemeshow's χ^2 and our global test tend to underestimate the true type I error. As a consequence, the true type I errors are higher than the nominal levels of the test, inducing a conservative distortion. We observe that the power of all test statistics decrease with the size of the asset correlation. However, the distortion of Hosmer-Lemeshow's χ^2 -test greatly exceeds the distortion of our global test.

Next, we address the problem of biasedness and how consistent the test statistics are. A test is said to be unbiased if the power for the alternative exceeds the level of significance. Under asset correlation regimes above 5%, Hosmer-Lemeshow's χ^2 is biased. The sum of true type I and type II error exceeds one or is close to one, which renders the test virtually useless

for practical considerations. Such a bias does not occur for our global test statistic, even though the applicability of our procedure might be limited under very high asset correlations.

A test is considered consistent against a certain class of alternatives, if the power of the test tends to one as the sample size tends to infinity. By our stringent simulation setup, none of the test statistics are consistent except for the special case of zero asset correlation. Under the orthogonality assumption established in Section 2, the shape statistic is consistent even for short time horizons. Over time, the level analysis, e.g. equation (18), also provides us with consistent estimators.

In summary, our results show that our global test statistic is more robust and more powerful against misspecifications than Hosmer-Lemeshow's χ^2 . Unlike Hosmer-Lemeshow's χ^2 , our global test is unbiased for the scenarios considered in the simulation setup. The reason for this observation is that the shape statistic is not vulnerable to misspecifications. Especially for typical scenarios encountered in practice, i.e., ten to 15 rating classes and asset correlations around 5%, the shape statistic performs well. The shape statistic agrees with the nominal error level and it does not lose power under small default dependency structures. Empirical evidence suggests that defaults exhibit a small, but statistically significant, correlation. Hence, for scenarios with the highest economic and practical relevance, we conclude that our global test statistic performs better than Hosmer-Lemeshow's χ^2 .

5. CONCLUSIONS

The validation of the probability calibration has several components. Our goal is to provide a comprehensive tool for backtesting probability calibrations in a quantitative way. Therefore, we focus on two important quantitative components, level and shape. We base our level evaluation on a comparison of ex-ante expected frequencies and the realized ex-post rates. We propose level statistics that are derived under dependencies. The second component, the shape, compares the theoretical area below the receiver operating characteristic curve (AUROC) with the empirical area. We then combine the two components into a global test statistic and show that it is asymptotically χ^2 -distributed with two degrees of freedom.

In a simulation study, we compare our global test statistic with the well-known Hosmer-Lemeshow's χ^2 . We examine both tests' reliability with respect to type I error levels, and both tests' power measured by type II error sizes. Overall, we find that the performance of our global test statistic is better than the performance of Hosmer-Lemeshow's χ^2 . We show that our global test is more robust against misspecifications, especially when defaults tend to cluster.

From a more practical viewpoint, and in addition to its applicability to situations in which defaults cluster, one of the main advantages of our test statistics is that they can handle ratings systems with multiple rating categories. With a large number of categories, previous calibration tests such as, e.g., the binomial test, the normal test, and the Hosmer-Lemeshow's χ^2 -test are virtually powerless.

APPENDIX A

Proof of Theorem 1. To prove a): Functional equivalence denotes an equivalence for all $\omega \in \Omega$ whereas almost sure equivalence denotes an equivalence on $\omega \in A$ where $\mathbb{P}\{A\} = 1$ and $A \subseteq \Omega$. To prove b) and c): Level and shape of a PD function denote two expectation measures of a random variable. Two almost surely equal random variables have the same expectation. ■

Proof of Proposition 1. Let Y_i^* and Y_j^* be the CreditMetrics latent variables for two debtors, $i \neq j$. There is only one systematic risk factor X and, since we have a homogeneous portfolio, the two debtors have the same weight $\sqrt{\rho}$ on that risk factor. Thus,

$$\begin{aligned} Y_i^* &= \sqrt{\rho}X - \sqrt{1-\rho}\epsilon_i \\ Y_j^* &= \sqrt{\rho}X - \sqrt{1-\rho}\epsilon_j, \end{aligned}$$

where X , ϵ_i , and ϵ_j are independent standard Gaussian variables. Hence, $(Y_i^*, Y_j^*)'$ follows a bivariate Gaussian distribution function with correlation ρ , also called asset correlation. A default event occurs if Y_i^* is lower than a predetermined threshold value $C := \Phi^{-1}(\pi)$, the so-called distance-to-default. Thus,

$$\mathbb{P}\{Y_i = 1|X\} = \mathbb{P}\{Y_i^* \leq C|X\} = \Phi\left(\frac{C - \sqrt{\rho}X}{\sqrt{1-\rho}}\right) = \Phi(V),$$

where $V := \frac{C - \sqrt{\rho}X}{\sqrt{1-\rho}}$. Note, conditional on X default events are independent, so that

$$\mathbb{P}\{Y_i = 1, Y_j = 1|X\} = \mathbb{P}\{Y_i^* \leq C, Y_j^* \leq C|X\} = \Phi(V)^2.$$

Hence, we deduce the variance of $Z = \Phi(V)$,

$$\begin{aligned} \mathbb{V}[\Phi(V)] &= \mathbb{P}\{Y_i^* \leq C, Y_j^* \leq C\} - \mathbb{P}\{Y_i^* \leq C\}^2 \\ &= \Phi_2(C, C, \rho) - \pi^2 = \pi_2 - \pi^2 = \rho_Y, \end{aligned}$$

where the first line follows by iterating expectations (see Proposition 8.13 of Karr (1993)), so that $\mathbb{E}\left[\mathbb{P}\{Y_i^* \leq C, Y_j^* \leq C|X\}\right] = \mathbb{P}\{Y_i^* \leq C, Y_j^* \leq C\}$. ■

Proof of Theorem 2. Consider the inequality

$$\begin{aligned} & \sup_{0 \leq \beta \leq 1} \left| \hat{\mathbb{F}}_{S_D} \left(\hat{\mathbb{F}}_{S_{ND}}^{-1}(1 - \beta) \right) - \mathbb{F}_{S_D} \left(\mathbb{F}_{S_{ND}}^{-1}(1 - \beta) \right) \right| \\ \leq & \sup_{0 \leq \beta \leq 1} \left| \hat{\mathbb{F}}_{S_D} \left(\hat{\mathbb{F}}_{S_{ND}}^{-1}(1 - \beta) \right) - \mathbb{F}_{S_D} \left(\hat{\mathbb{F}}_{S_{ND}}^{-1}(1 - \beta) \right) \right| \\ & + \sup_{0 \leq \beta \leq 1} \left| \mathbb{F}_{S_D} \left(\hat{\mathbb{F}}_{S_{ND}}^{-1}(1 - \beta) \right) - \mathbb{F}_{S_D} \left(\mathbb{F}_{S_{ND}}^{-1}(1 - \beta) \right) \right|. \end{aligned}$$

If we apply the Glivenko-Cantelli Theorem for the first term on the right hand side of the above inequality, then the theorem of Dvoretzky, Kiefer, and Wolfowitz (1956) and the Borel-Cantelli Lemma prove our claim. \blacksquare

Proof of Proposition 3. From (19) and (20) as well as the basic integration rule of monotonicity⁸ we can derive that

$$\begin{aligned} \int_{-\infty}^t \widetilde{\text{PD}}(s) d\mathbb{F}_S(s) & \leq \int_{-\infty}^t \text{PD}(s) d\mathbb{F}_S(s) \text{ for all } t \in \mathbb{S} \\ \int_t^{\infty} \widetilde{\text{PD}}(s) d\mathbb{F}_S(s) & \geq \int_t^{\infty} \text{PD}(s) d\mathbb{F}_S(s) \text{ for all } t \in \mathbb{S}^c. \end{aligned}$$

Thus, it follows for all $t \in \mathbb{R}$,

$$\int_{-\infty}^t \widetilde{\text{PD}}(s) d\mathbb{F}_S(s) \leq \int_{-\infty}^t \text{PD}(s) d\mathbb{F}_S(s).$$

Since the PD functions are equivalent with respect to the PD level, so that $\int_{-\infty}^{\infty} \widetilde{\text{PD}}(s) d\mathbb{F}_S(s) = \int_{-\infty}^{\infty} \text{PD}(s) d\mathbb{F}_S(s)$, we can normalize the above inequality to arrive at

$$(29) \quad \tilde{\mathbb{F}}_{S_D}(t) \leq \mathbb{F}_{S_D}(t) \text{ for all } t \in \mathbb{R},$$

for some t^* the inequality is strict, so that $\tilde{\mathbb{F}}_{S_D}(t^*) < \mathbb{F}_{S_D}(t^*)$. With the similar reasoning we can deduce that

$$(30) \quad \tilde{\mathbb{F}}_{S_{ND}}(t) \geq \mathbb{F}_{S_{ND}}(t) \text{ for all } t \in \mathbb{R},$$

where the inequality is strict for some t^* . Hence, it follows that the difference in AUROC is

$$\begin{aligned}
\widehat{\text{AUROC}} - \text{AUROC} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[\mathbf{1}_{\{x>y\}} + \frac{1}{2} \mathbf{1}_{\{x=y\}} \right] \\
&\quad d \left[\tilde{\mathbb{F}}_{S_D}(x) - \mathbb{F}_{S_D}(x) \right] d \left[\tilde{\mathbb{F}}_{S_{ND}}(y) - \mathbb{F}_{S_{ND}}(y) \right] \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[\mathbf{1}_{\{-z>y\}} + \frac{1}{2} \mathbf{1}_{\{-z=y\}} \right] \\
&\quad \underbrace{d \left[\mathbb{F}_{S_D}(-z) - \tilde{\mathbb{F}}_{S_D}(-z) \right]}_{\geq 0} \underbrace{d \left[\tilde{\mathbb{F}}_{S_{ND}}(y) - \mathbb{F}_{S_{ND}}(y) \right]}_{\geq 0}.
\end{aligned}$$

The first equality comes from the definition of the AUROC figure. The second equality follows by the substitution rule. The last term is positive since the integrand is nonnegative and positive for some values and therefore proving the proposition. \blacksquare

Proof of Proposition 4. The estimate $\widehat{\text{AUROC}}_n$ is unbiased since

$$\begin{aligned}
\mathbb{E} \left[\widehat{\text{AUROC}}_n | \mathbf{Y} \right] &= \mathbb{P} \{ S_D > S_{ND} \} + \frac{1}{2} \mathbb{P} \{ S_D = S_{ND} \} \\
&= \frac{1}{2} [1 - \mathbb{P} \{ S_D < S_{ND} \} + \mathbb{P} \{ S_D > S_{ND} \}] \\
&= \text{AUROC}.
\end{aligned}$$

For the computation of the variance we start with the squared $\widehat{\text{AUROC}}_n$ figure

$$\begin{aligned}
\widehat{\text{AUROC}}_n^2 &= \frac{1}{N_0^2 N_1^2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_0} \sum_{k=1}^{N_1} \sum_{l=1}^{N_0} \frac{1}{4} \left[1 - \mathbf{1}_{\{S_{D_i} < S_{ND_j}\}} \right. \\
&\quad + \mathbf{1}_{\{S_{ND_j} < S_{D_i}\}} - \mathbf{1}_{\{S_{D_k} < S_{ND_l}\}} + \mathbf{1}_{\{S_{ND_l} < S_{D_k}\}} \\
&\quad + \mathbf{1}_{\{S_{D_i} < S_{ND_j}, S_{D_k} < S_{ND_l}\}} + \mathbf{1}_{\{S_{ND_j} < S_{D_i}, S_{D_k} < S_{ND_l}\}} \\
&\quad \left. + \mathbf{1}_{\{S_{D_i} < S_{ND_j}, S_{ND_l} < S_{D_k}\}} + \mathbf{1}_{\{S_{ND_j} < S_{D_i}, S_{ND_l} < S_{D_k}\}} \right].
\end{aligned}$$

Now, we can differentiate between four different cases:

1. In $N_0(N_0 - 1)N_1(N_1 - 1)$ cases the defaulters' indices and non-defaulters' ones are different, so that $i \neq k$ and $j \neq l$. In this instance the expectation of the summand in squared brackets is AUROC^2 or

$$\frac{1}{4} [1 - \mathbb{P} \{ S_D < S_{ND} \} + \mathbb{P} \{ S_D > S_{ND} \}]^2.$$

2. In $N_1N_0(N_0 - 1)$ cases the defaulters' indices are equal but the non-defaulters' ones are different, so that $i = k$ and $j \neq l$. In this instance the expectation of the summand is

$$\begin{aligned} & \frac{1}{2} [1 - \mathbb{P}\{S_D < S_{ND}\} + \mathbb{P}\{S_D > S_{ND}\}] - \frac{1}{4} \\ & + \frac{1}{4} \mathbb{P}\{S_{D_1}, S_{D_2} < S_{ND}\} - \frac{1}{4} \mathbb{P}\{S_{D_1} < S_{ND} < S_{D_2}\} \\ & + \frac{1}{4} \mathbb{P}\{S_{ND} < S_{D_1}, S_{D_2}\} - \frac{1}{4} \mathbb{P}\{S_{D_2} < S_{ND} < S_{D_1}\}, \end{aligned}$$

what can be rewritten as $\text{AUROC} - \frac{1}{4} + \frac{1}{4}B_{110}$.

3. In $N_0N_1(N_1 - 1)$ cases the defaulters' indices are different but the non-defaulters' ones are equal, so that $i \neq k$ and $j = l$. In this instance the expectation of the summand is

$$\begin{aligned} & \frac{1}{2} [1 - \mathbb{P}\{S_D < S_{ND}\} + \mathbb{P}\{S_D > S_{ND}\}] - \frac{1}{4} \\ & + \frac{1}{4} \mathbb{P}\{S_{ND_1}, S_{ND_2} < S_D\} - \frac{1}{4} \mathbb{P}\{S_{ND_1} < S_D < S_{ND_2}\} \\ & + \frac{1}{4} \mathbb{P}\{S_D < S_{ND_1}, S_{ND_2}\} - \frac{1}{4} \mathbb{P}\{S_{ND_2} < S_D < S_{ND_1}\}, \end{aligned}$$

what can be rewritten as $\text{AUROC} - \frac{1}{4} + \frac{1}{4}B_{001}$.

4. In N_1N_0 cases the the defaulters' indices and the non-defaulters' ones are equal, so that $i = k$ and $j = l$. In this instance the expectation of the summand is

$$\mathbb{P}\{S_{ND} < S_D\} + \frac{1}{4} \mathbb{P}\{S_{ND} = S_D\} = \text{AUROC} - \frac{1}{4} + \frac{1}{4} \mathbb{P}\{S_{ND} \neq S_D\}.$$

Now, the fact that

$$\mathbb{V} \left[\widehat{\text{AUROC}}_n | \mathbf{Y} \right] = \mathbb{E} \left[\widehat{\text{AUROC}}_n^2 | \mathbf{Y} \right] - \text{AUROC}^2,$$

as well as simple arithmetic summations and cancelations lead to the final result. ■

Proof of Lemma 1. From two well-known theorems, see for instance Theorem 2.47 and 2.48 in Karr (1993) for the proofs, we know that a) $G(X)$ is uniformly distributed, and that b) $\Phi^{-1}(G(X))$ is standard Gaussian distributed. ■

APPENDIX B

We can derive the PD function and the score distribution from the distribution of defaulters and non-defaulters' distribution, and unconditional default probability (PD level) as follows:

$$\mathbb{F}_S(t) = \mathbb{P}\{Y_i = 1\} \mathbb{F}_{S_D}(t) + \mathbb{P}\{Y_i = 0\} \mathbb{F}_{S_{ND}}(t).$$

In addition, with the slope of the ROC curve $m(t)$,

$$\begin{aligned} m(t) &:= \lim_{\Delta \rightarrow 0} \frac{\mathbb{F}_{S_D}(t + \Delta) - \mathbb{F}_{S_D}(t)}{\mathbb{F}_{S_{ND}}(t + \Delta) - \mathbb{F}_{S_{ND}}(t)} \\ &= \frac{\mathbb{P}\{Y_i = 0\}}{\mathbb{P}\{Y_i = 1\}} \lim_{\Delta \rightarrow 0} \frac{\int_t^{t+\Delta} \text{PD}(s) d\mathbb{F}_S(s)}{\int_t^{t+\Delta} [1 - \text{PD}(s)] d\mathbb{F}_S(s)} \\ &= \frac{\mathbb{P}\{Y_i = 0\}}{\mathbb{P}\{Y_i = 1\}} \frac{\text{PD}(t)}{1 - \text{PD}(t)}, \end{aligned}$$

we can infer the PD function as

$$\text{PD}(t) = \frac{\mathbb{P}\{Y = 1\} m(t)}{\mathbb{P}\{Y = 0\} [m(t) - 1] + 1}.$$

REFERENCES

- BALTHAZAR, L. (2004): “PD estimates for Basel II,” *Risk Magazine*, 17(4), 84–85.
- BAMBER, D. (1975): “The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Graph,” *Journal of Mathematical Psychology*, 12, 387–415.
- BASEL COMMITTEE ON BANKING SUPERVISION (2005): “Studies on the Validation of Internal Rating Systems,” Working paper No. 14, Bank for International Settlements.
- BLÖCHLINGER, A., AND M. LEIPPOLD (2005): “Economic Benefit of Powerful Credit Scoring,” *Journal of Banking and Finance*, forthcoming.
- BLOCHWITZ, S., S. HOHL, AND C. S. WEHN (2005): “Reconsidering Ratings,” Working paper, Deutsche Bundesbank.
- BONFERRONI, C. E. (1936): “Teoria statistica delle classi e calcolo delle probabilità,” *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8, 3–62.
- BRIER, G. W. (1950): “Verification of Forecasts Expressed in Terms of Probability,” *Monthly Weather Review*, 78, 1–3.
- CRAMÉR, H. (1946): *Mathematical methods of statistics*. Princeton University Press, Princeton.
- CROSBIE, P. (1997): “Modeling Default Risk,” Technical document, KMV Corporation.
- DEGROOT, M., AND S. FIENBERG (1983): “The comparison and evaluation of forecasters,” *The Statistician*, 32, 12–22.
- DVORETZKY, A., J. KIEFER, AND J. WOLFOWITZ (1956): “Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator,” *The Annals of Mathematical Statistics*, 27(3), 642–669.
- EGAN, J. (1975): *Signal Detection Theory and ROC Analysis*, Series in Cognition and Perception. Academic Press, New York.

- EPSTEIN, E. S. (1969): “A Scoring System for Probability Forecasts of Ranked Categories,” *Journal of Applied Meteorology*, 8, 985–987.
- FREY, R., AND A. J. MCNEIL (2001): “Modelling dependent defaults,” Working paper, University of Zurich and ETH Zurich.
- FUDENBERG, D., AND D. LEVINE (1999): “An easier way to calibrate,” *Games and Economic Behavior*, 29, 131–137.
- GORDY, M. B. (2003): “A Risk-Factor Model Foundation for Ratings-Based Bank Capital Rules,” *Journal of Financial Intermediation*, 12(3), 199–232.
- GUPTON, G. M., C. C. FINGER, AND M. BHATIA (1997): “CreditMetrics,” Technical document, J.P. Morgan & Co.
- HENERY, R. J. (1985): “On the Average Probability of Losing Bets on Horses with Given Starting Price Odds,” *Journal of the Royal Statistical Society*, 148(4), 342–349.
- HOERL, A. E., AND H. K. FALLIN (1974): “Reliability of Subjective Evaluations in a High Incentive Situation,” *Journal of the Royal Statistical Society*, 137(2), 227–231.
- HOSMER, D. W., T. HOSMER, S. LE CESSIE, AND S. LEMESHOW (1997): “A comparison of goodness-of-fit tests for the logistic regression model,” *Statistics in Medicine*, 16, 965–980.
- HOSMER, D. W., AND S. LEMESHOW (1989): *Applied Logistic Regression*. John Wiley & Sons, Inc., New York.
- KARR, A. F. (1993): *Probability*. Springer Verlag, New York.
- LEMESHOW, S., AND J. R. LE GALL (1994): “Modeling the Severity of Illness of ICU patients,” *Journal of the American Medical Association*, 272(13), 1049–1055.
- MANN, H., AND D. WHITNEY (1947): “On a Test Whether One of Two Random Variables is Stochastically Larger Than the Other,” *Annals of Mathematical Statistics*, 18, 50–60.
- MERTON, R. (1974): “On the Pricing of Corporate Debt: The Risk Structure of Interest Rate,” *Journal of Finance*, 2, 449–470.

- MURPHY, A. H. (1970): “The Ranked Probability Score and the Probability Score: A Comparison,” *Monthly Weather Review*, 98, 917–924.
- MURPHY, A. H., AND E. S. EPSTEIN (1967): “Verification of Probabilistic Predictions: A Brief Review,” *Journal of Applied Meteorology*, 6, 748–755.
- ROWLAND, T., L. OHNO-MACHAD, AND A. OHRN (1998): “Comparison of Multiple Prediction Models for Ambulation Following Spinal Cord Injury,” Proceedings of the american medical informatics association, American Medical Informatics Association, Orlando.
- SNYDER, W. W. (1978): “Horse Racing: Testing the Efficient Markets Model,” *Journal of Finance*, 33(4), 1109–1118.
- STEIN, M. R. (2005): “The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing,” *Journal of Banking and Finance*, 29, 1213–1236.
- TASCHE, D. (2003): “A traffic lights approach to PD validation,” Working paper, Deutsche Bundesbank, Frankfurt am Main, Germany.
- THOMAS, L. (2000): “A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Consumers,” *International Journal of Forecasting*, 16, 149–172.
- THOMAS, L. C., D. B. EDELMAN, AND J. N. CROOK (2002): *Credit Scoring and Its Applications*. Society for Industrial and Applied Mathematics, Philadelphia.
- WILCOXON, F. (1945): “Individual Comparisons by Ranking Methods,” *Biometrics*, 1, 80 – 83.
- WILSON, T. C. (1998): “Portfolio Credit Risk,” *FRBNY Economic Policy Review*, 10, 1–12.
- WINKLER, R. L., AND A. H. MURPHY (1968): “Evaluation of Subjective Precipitation Probability Forecasts,” Proceedings of the first national conference on statistical meteorology, American Meteorological Society, Boston.
- ZADROZNY, B., AND C. ELKAN (2001): “Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers,” *International Conference on Machine Learning*, 29, 131–137.

——— (2002): “Transforming classifier scores into accurate multiclass probability estimates,” *Knowledge Discovery and Data Mining*, pp. 131–137.

NOTES

¹For a recent survey on the use of credit-scoring models, see Thomas (2000) and Thomas, Edelman, and Crook (2002). The earliest known reference to proper probability forecasting dates back to the meteorological statistician Brier (1950) and much of the early literature on proper probability forecasting is inspired by meteorology as in Murphy and Epstein (1967), Winkler and Murphy (1968), Epstein (1969), Murphy (1970) and works cited in them. Later, game theory and in particular horse racing attracted the interest of probability forecasters as in Hoerl and Fallin (1974), Snyder (1978), and Henery (1985). In addition, probability forecasts also include applications in medicine (Lemeshow and Le Gall (1994), Rowland, Ohno-Machad, and Ohrn (1998)), weather prediction (DeGroot and Fienberg (1983)), game theory (Fudenberg and Levine (1999)), and pattern classification (Zadrozny and Elkan (2001), Zadrozny and Elkan (2002)).

²Such extension of the basic model include, e.g., imposing a one-factor dependence structure and a granularity adjustment to account for the finiteness of the sample (see Gordy (2003)).

³Except for degenerated cases, the orthogonality assumptions imply that $S_i|\mathbf{V} \sim S_i$ is generally not true.

⁴If X_n converges in distribution to X and if Y_n converges in distribution to a constant $c > 0$ then X_n/Y_n converges in distribution to X/c (see Cramér (1946) for a proof)

⁵For instance, in some situations, $\rho = 0.05$ appears to be appropriate for a one-year-horizon (see also Tasche (2003)). However, the Basel Committee on Banking Supervision (2005) considers default correlations, ρ_Y , between 0.5% and 3% as typical.

⁶We note that the expectation for the AUROC statistic is also 0.5 for the case in which the two continuous distributions are not identical but have only the medians in common, resulting in a non-diagonal ROC curve. But in this case, the variance has to be derived as shown in Proposition 4. However, a non-diagonal ROC graph with AUROC 0.5 violates the monotonicity assumption of the PD function.

⁷We note that the σ -algebra generated by Y_i, N_1 and \mathbf{V} , $\sigma(Y_i, N_1, \mathbf{V})$, and the σ -algebra generated by Y_i are both contained in $\sigma(\mathbf{S}, \mathbf{V}, \mathbf{Y})$, in particular it is true that $\sigma(\mathbf{S}, \mathbf{V}, \mathbf{Y}) \supseteq \sigma(Y_i, N_1, \mathbf{V}) \supseteq \sigma(Y_i)$.

⁸If either $0 \leq g \leq h$ or g and h are integrable and $g \leq h$, then $\int gdF \leq \int hdF$.

15 classes			10 classes			5 classes		
PD	PD_β	#	PD	PD_β	#	PD	PD_β	#
0.0053	0.0027	1	0.0058	0.0030	20	0.0075	0.0042	625
0.0068	0.0038	9	0.0084	0.0049	176	0.0144	0.0096	2500
0.0088	0.0053	56	0.0120	0.0077	703	0.0263	0.0205	3750
0.0113	0.0072	222	0.0169	0.0119	1641	0.0455	0.0403	2500
0.0144	0.0097	611	0.0235	0.0180	2460	0.0746	0.0735	625
0.0181	0.0130	1222	0.0320	0.0264	2460			
0.0227	0.0173	1831	0.0430	0.0380	1641			
0.0281	0.0226	2096	0.0569	0.0535	703			
0.0347	0.0293	1831	0.0740	0.0735	176			
0.0424	0.0376	1222	0.0948	0.0989	20			
0.0515	0.0477	611						
0.0620	0.0598	222						
0.0742	0.0742	56						
0.0882	0.0911	9						
0.1039	0.1107	1						

Table 1: For the simulation study we consider 3 different numbers of rating classes (15, 10, and 5). The expected default frequency is fixed for all scenarios at 3%, and the size of the portfolio is set at 10'000 debtors. Entries report the rating distribution together with the assigned rating class PDs. PD denotes the default probability under the data generating process whereas PD_β is the assumed PD for type II error analysis.

ρ	C	Type I error				Type II error			
		χ^2	Global	Level	Shape	χ^2	Global	Level	Shape
0	15	0.083	0.047	0.049	0.047	0.374	0.118	0.125	0.665
0	10	0.065	0.052	0.046	0.050	0.244	0.099	0.120	0.577
0	5	0.052	0.050	0.045	0.051	0.126	0.072	0.123	0.436
0.05	15	0.721	0.064	0.037	0.077	0.275	0.753	0.935	0.693
0.05	10	0.741	0.065	0.038	0.083	0.231	0.711	0.939	0.640
0.05	5	0.766	0.081	0.035	0.097	0.185	0.635	0.942	0.552
0.10	15	0.801	0.155	0.147	0.098	0.208	0.739	0.844	0.740
0.10	10	0.821	0.161	0.142	0.115	0.183	0.714	0.849	0.692
0.10	5	0.844	0.175	0.140	0.142	0.151	0.663	0.858	0.629
0.15	15	0.845	0.254	0.251	0.117	0.168	0.710	0.758	0.777
0.15	10	0.862	0.267	0.255	0.142	0.145	0.679	0.757	0.734
0.15	5	0.884	0.286	0.242	0.182	0.127	0.655	0.766	0.692

Table 2: Nominal level $\alpha = 0.05$: For the simulation study we consider 4 different asset correlation regimes (0, 0.05, 0.1, and 0.15) as well as 3 different numbers of rating classes (15, 10, 5) resulting in 12 scenarios. The estimated type I and type II error rates based on 10,000 Monte Carlo simulations at given nominal error level of 0.05 are tabulated.

ρ	C	Type I error			Type II error				
		χ^2	Global	Level	Shape	χ^2	Global	Level	Shape
0	15	0.032	0.010	0.011	0.009	0.553	0.265	0.285	0.845
0	10	0.019	0.011	0.012	0.010	0.422	0.230	0.284	0.782
0	5	0.010	0.009	0.010	0.010	0.259	0.187	0.272	0.660
0.05	15	0.652	0.018	0.006	0.022	0.340	0.859	0.984	0.835
0.05	10	0.682	0.018	0.007	0.020	0.302	0.825	0.983	0.785
0.05	5	0.706	0.027	0.006	0.030	0.258	0.761	0.986	0.705
0.10	15	0.755	0.060	0.055	0.029	0.256	0.845	0.933	0.850
0.10	10	0.776	0.062	0.050	0.033	0.233	0.814	0.939	0.807
0.10	5	0.803	0.073	0.050	0.048	0.198	0.773	0.936	0.748
0.15	15	0.805	0.122	0.131	0.034	0.208	0.821	0.876	0.869
0.15	10	0.826	0.134	0.125	0.045	0.185	0.798	0.877	0.830
0.15	5	0.850	0.147	0.118	0.069	0.163	0.772	0.883	0.790

Table 3: Nominal level $\alpha = 0.01$: For the simulation study we consider 4 different asset correlation regimes (0, 0.05, 0.10, and 0.15) as well as 3 different numbers of rating classes (15, 10, 5) resulting in 12 scenarios. The estimated type I and type II error rates based on 10,000 Monte Carlo simulations at given nominal error level of 0.01 are tabulated.