Testing Probability Calibrations*

Andreas Blöchlinger†

Credit Suisse

First Version: July, 2005

This Version: November 30, 2005

*The content of this paper reflects the personal view of the author, in particular, it does not necessarily represent the opinion of Credit Suisse. The author thanks Markus Leippold, and the "quants" at Credit Suisse for valuable and insightful discussions.

†**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

Testing Probability Calibrations

Abstract

Probability calibration is the act of assigning probabilities to uncertain events. We develop a testing procedure consisting of two components to check whether the ex-ante probabilities are in line with the ex-post frequencies. The first component tests the level of the probability calibration under dependencies. In the long run 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. Out of it we construct a goodness-of-fit statistic which is asymptotically χ^2-distributed and further a traffic light system for a time-series of forecasts.

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.

According to Foster and Vohra [1998] probability calibration is the act of assigning probabilities to an uncertain event. Since 1965, the US National Weather Service has been in the habit of making and announcing "probability of precipitation" forecasts. Such a forecast is interpreted to be the probability that precipitation, defined to be at least 0.01 inches, will occur in a specified time period and area. 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]. The aggregated subjective probability that a horse wins a race is forcasted from the odds of that horse.[1] Today, probability forecasts include various applications: Medicine (e.g. Lemeshow and Le Gall [1994], and Rowland, Ohno-Machad, and Ohrn [1998]), weather prediction tasks (e.g. DeGroot and Fienberg [1983]), game theory (e.g. Fudenberg and Levine [1999]), in the context of pattern classification (e.g. Zadrozny and Elkan [2001]), and Zadrozny and Elkan [2002]), and credit scoring (e.g. Stein [2002]). In this paper we limit ourselves to the consideration of probability calibration of credit scoring models even though the validation procedures we are presenting can be applied to various fields.

A credit scoring system is mainly an ordinal measurement instrument that distinguishes between low and high default risk – the risk that a borrower does not comply with the contractual loan agreement, i.e. by not paying interest. Upfront, credit scoring is meant to deliver a ranking of

[1]This is true since betting on horses does not involve systematic risk, i.e. the amount of money lost equals the amount won among the aggregate of bettors and race track. Therefore, a horse bet wager is not rewarded with a risk premium and probabilities can be derived from the odds (see Harrison and Kreps [1979] on the relationship between pricing, systematic risk, probabilities and equivalent martingale measures).

obligors, i.e. the higher the score the worse the creditworthiness. But when it comes to pricing of loans or to quantitative risk assessments one needs to map the ordinal score into a metric measure or into a probability of default (PD), respectively.

A major obstacle to backtesting of PDs is the scarcity of data, caused by the infrequency of default events and the impact of default clusterings. Due to correlation between defaults in loan portfolios caused by economic up- and downswings, observed default rates can systematically exceed the critical values if these are determined under the assumption of independence. This can happen easily for otherwise well-calibrated rating systems. As a consequence, on one hand, tests based on the independence assumption are rather conservative, with even well-behaved rating systems performing poorly. On the other hand, tests that take into account correlation between defaults will only allow the detection of relatively obvious cases of rating system miscalibration.

An accurate PD calibration of rating models is primarily required by competition among banks. Competition brings prices down. A correctly calibrated and powerful credit scoring systems has the capability to significantly increasing profits by both reducing losses and increasing revenues even in a saturated and competitive market. On the other side a bank operating under a poorly calibrated model experiences adverse selection by attracting bad loans. According to Stein [2005] and Blöchlinger and Leippold [2005] small differences in accuracy between banks result in several millions of profit differences. Hence, the testing procedure on probability calibration must be powerful against alternatives with large economic impact. If two probability calibrations will only result in small profit differences then the test statistics do not need to be very powerful. Secondarily, an accurate PD calibration is also required by regulating authorities like the Basel Committee on Banking Supervision.

For the validation of probabilites of default, the Basel Committee on Banking Supervision [2005] differentiates between two stages: validation of the discriminatory power of a rating system and validation of the accuracy of the PD quantification. The two stages are highly interrelated. For instance, a rating systems with no discriminatory power results in a flat or "horizontal" PD function – all obligors get the same PD irrespective of their credit score.[2] A perfect scoring system necessitates a set of score values with PD one and the complementary set with a probability of zero (what we call "vertical" PD function).

A recent example of a test on the accuracy of the PD quantification is given by Balthazar [2004], relying heavily on simulation methods. Tasche [2003] presents a method avoiding simulations but requiring approximations. The Basel Committee on Banking Supervision [2005] has in detail reviewed the literature with respect to calibration tests (i.e. binomial test, χ^2-test, normal test and the traffic lights approach of Blochwitz, Hohl, and Wehn [2005]), but the committee has to conclude that "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." Other studies come to similar conclusions, e.g. Blochwitz, Hohl, Tasche, and Wehn [2004] note "that further developments in the field of PD validation might not reach much improvement. Nevertheless, this is only a conjecture so that further research for its verification is needed." A further shortcoming of the reviewed methods, not mentioned by the two studies, is that they are only applicable under grouping of obligors into rating classes or other weighting schemes. If the PD calibration is continuous in the sense that two obligors have almost surely different PDs then all tests reviewed by the Basel Committee fail.

[2]The Basel Committee also uses the term "pool PD" for the "horizontal" PD function.

Altogether, approaches to the validation have to be made that should be understandable by a bank's practitioners as well as by examiners who are responsible for auditing the appropriateness and adequacy of the estimation, modeling, and calibration procedures.

In a well-calibrated model, the estimated default frequency is equivalent to the default probability. This observation needs to be transformed into a statistical hypothesis that allows a powerful testing procedure. Note, a well-calibrated model implies two testable properties. First, a well-calibrated system predicts on average the realized number of events. Second, it also forecasts on average the realized number of events for an arbitrary subpopulation (e.g. only observations with low probabilities). We call the former property probability calibration with respect to the level – the second property with respect to the shape, and we deduce test statistics for probability level and probability shape as well as a global test statistic. Further, we derive a traffic light tool in order to backtest the probability calibration over a time series of probability forecasts. This traffic light system generalizes the approach described by Blochwitz, Hohl, and Wehn [2005].

We contribute to the literature by deriving new test statistics that are not subject to the above mentioned shortcomings – e.g. our testing procedure allows continuous PDs, we explicitly take default correlation into account and we do not rely on Monte Carlo simulations. We proceed the following way: In Section 1 we outline basic assumptions and definitions. Section 2 derives test statistics on a one-period basis for level and shape, and the two tests are combined into a global test statistic. 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 3 generalizes the global test statistic so that it can be applied over a time-series of default forecasts. Finally, Section 4 outlines our conclusions.

1 Assumptions and Definitions

We make three basic assumptions regarding homogeneity, orthogonality, and monotonicity.

Assumption 1.1 (Homogeneity). The loan portfolio consists of n obligors. To each obligor i we assign a binary default indicator Y_i and a credit score S_i. Further, we assume $k < n$ systematic risk factors $\mathbf{V}$. $\mathbf{S}$, $\mathbf{Y}$, and $\mathbf{V}$ are random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The portfolio is homogeneous in the sense that the random vector $(\mathbf{S}, \mathbf{Y}, \mathbf{V})$ is exchangeable, so that

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

for any permutation $(\Pi(1), ..., \Pi(n))$ of $(1, ..., n)$.

Assumption 1.2 (Orthogonality). The conditional distributions of credit score S_i and default indicator Y_i are so that

$$\begin{aligned} S_i|\mathbf{S}, \mathbf{V}, \mathbf{Y} &\backsim S_i|Y_i \\ Y_i|\mathbf{S}, \mathbf{V}, \mathbf{Y} &\backsim Y_i|S_i, \mathbf{V}. \end{aligned}$$

On one hand, defaults are correlated through the dependence on common factors. This means that with respect to default prediction, the credit score does not subsume all the information generated by macroeconomic drivers. There are some economic-wide noise factors influencing the true creditworthiness of obligors which are not predictable by the credit score. Since these factors affect all obligors they induce default clusterings over time. A good state of the overall economy leads to a low number of defaults and vice versa. On the other hand, conditional on the default indicator Y_i the scores S_i form an independent sequence of random variables. Therefore, regarding the forecast of the credit score all the information is contained by

the default state. In the following, we write $S_D = (S_i|Y_i = 1)$ for the credit score of defaulters and correspondingly $S_{ND} = (S_i|Y_i = 0)$ for the score of non-defaulters. Note also, unless degenerated cases, our orthogonality assumptions imply that it is generally not true that $S_i|\mathbf{V} \backsim S_i$.

Since we have a homogeneous loan portfolio, according to Assumption 1.1, the probability of default does not depend on i. Hence, we define the PD function,

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

Unfortunately, in practice $\mathrm{PD}(s)$ is not observable and has to be defined or estimated, respectively.

Definition 1.3 (Probability calibration). The act of estimating or approximating $\mathrm{PD}(s)$ by a measurable function

$$\widetilde{\mathrm{PD}}(s) : \mathbb{R} \to [0, 1]$$

is called probability calibration.

The PD function $\widetilde{\mathrm{PD}}(s)$ links the credit score with the estimated default frequency. Technically we need to assume that $\widetilde{\mathrm{PD}}(s)$ is a measurable function. We call this mapping probability calibration since an ordinal measure is mapped into a metric measure. In practice, the score is usually mapped into a one-year PD. Many financial institutions apply a step function, but other well-known parametric links are logistic distribution function (logit model), Gaussian distribution (probit model) or identity link (linear probability model, discriminant function), but nonparametric links are today also very common. Regarding the PD function and probability calibration we make the third and last assumption that concerns monotonicity.

Assumption 1.4 (Monotonicity). The PD function is montonic, so that either

$$\begin{aligned} &\mathrm{PD}(s) \geq \mathrm{PD}(t) \text{ for all } s \geq t \quad \text{or} \\ &\mathrm{PD}(s) \geq \mathrm{PD}(t) \text{ for all } s \leq t. \end{aligned}$$

Therefore, the PD function is assumed to be either entirely non-increasing or non-decreasing. If the probability calibration is performed correctly then we have a functional equivalence to the true PD function.

Definition 1.5 (Functional Equivalence). The PD functions $\widetilde{\mathrm{PD}}(s)$ and $\mathrm{PD}(s)$ are functionally equivalent, if

$$\widetilde{\mathrm{PD}}(s) \quad = \quad \mathrm{PD}(s)$$

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

In hypotheses, it is unnecessary or even impossible to assume that something is true for every outcome, in our case for every s, but rather only that it is true of outcomes belonging to an event of probability one. Correspondingly, that some property holds on an event of probability one is, ordinarily, all that one can establish. That is way we define a weaker property than functional equivalence – almost sure equivalence.

Definition 1.6 (Almost Sure Equivalence). The PD functions $\widetilde{\mathrm{PD}}(s)$ and $\mathrm{PD}(s)$ are almost surely equivalent, if

$$\widetilde{\mathrm{PD}}(s) \quad = \quad \mathrm{PD}(s)$$

for almost all $s \in \mathbb{R}$.

From a practical perspective, it is inherently impossible to distinguish two PD functions that are equivalent almost surely but not functionally.

Even if we have two almost surely inequivalent PD functions the testing for almost all s may become cumbersome due to a lack of defaulters and/or observations per s, even for a finite number of rating classes. Therefore, we focus our attention on two other important properties of the PD function – level and shape – that will allow a statistical validation procedure.

The PD level, an estimate of the long-run aggregate probabilities of default for an economy, is the first anchor for a models validity.

Definition 1.7 (Level Equivalence). The PD functions $\widetilde{\mathrm{PD}}(s)$ and $\mathrm{PD}(s)$ are equivalent with respect to the PD level, if

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

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

We assume that the distribution function $\mathbb{F}_S(t)$ is known/observable or it is replaced by the actual distribution, respectively. Note, the PD function $\widetilde{\mathrm{PD}}(s)$ and the true PD function, the one under which defaults are generated, are equivalent with respect to the PD level, if $\mathbb{P}\{Y_i = 1\} = \int_{-\infty}^{\infty} \widetilde{\mathrm{PD}}(s) d\mathbb{F}_S(s)$.

The second anchor of the PD function is the shape – the inherent property of distinguishing between non-defaulters and defaulters. The distribution function of defaulters' and non-defaulters' $\mathbb{F}_{S_D}(t)$, and $\mathbb{F}_{S_{ND}}(t)$ are a function of $\mathrm{PD}(s)$. This can be derived explicitly by,

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

and,

$$\begin{aligned}\mathbb{F}_{S_{ND}}(t) &= \mathbb{P}\left\{S_i \leq t | Y_i = 0\right\} \\ &= \int_{-\infty}^{t} \frac{1}{\mathbb{P}\left\{Y_i = 0\right\}} \mathbb{P}\left\{Y_i = 0 | S_i = s\right\} d\mathbb{F}_S(s) \qquad (2) \\ &= \frac{\int_{-\infty}^{t} \left[1 - \mathrm{PD}(s)\right] d\mathbb{F}_S(s)}{1 - \int_{-\infty}^{\infty} \mathrm{PD}(s) d\mathbb{F}_S(s)}.\end{aligned}$$

If credit score S_i and default indicator Y_i are two independent random variable then the non-defaulters' and defaulters' distribution function coincide with the unconditional distribution function of the credit score. In this case we say the credit score has no discriminatory power and it is irrelevant with respect to the prediction of a loan failure. The discriminatory power is visualized by the Receiver Operating Characteristic (ROC) curve. The two-dimensional graph generated by the survival functions for non-defaulters and defaulters,

$$\left\{1 - \mathbb{F}_{S_{ND}}(t), 1 - \mathbb{F}_{S_D}(t)\right\} \text{ for all } t \in \mathbb{R}, \qquad (3)$$

is called the ROC curve. From the definition we see immediately that the range of the ROC graph is restricted to the unit square. Accordingly, the area below the curve is limited from above by one and from below by zero. It is easy to see from (1) and (2) that two almost surely equivalent PD functions engender the same ROC graph. Further, we can establish that the ROC curve itself as well as the slope and the area below the graph depend on the PD function. The area under the ROC curve (AUROC) is calculated as (see e.g. Bamber [1975], or Blöchlinger and Leippold [2005])

$$\begin{aligned}\mathrm{AUROC} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[\mathbf{1}_{\{x>y\}} + \frac{1}{2}\mathbf{1}_{\{x=y\}}\right] d\mathbb{F}_{S_D}(x)\, d\mathbb{F}_{S_{ND}}(y) \\ &= \mathbb{P}\left\{S_D > S_{ND}\right\} + \frac{1}{2}\mathbb{P}\left\{S_D = S_{ND}\right\}. \qquad (4)\end{aligned}$$

The last equality follows by orthogonality established in Assumption 1.2. By the fact that $1 - \mathbf{1}_{\{x<y\}} = \mathbf{1}_{\{x>y\}} + \mathbf{1}_{\{x=y\}}$ we can also write (4) the following way,

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

The AUROC figure represents our quantitative measure for shape equivalence.

Definition 1.8 (Shape Equivalence). Two PD functions $\widetilde{\text{PD}}(s)$ and $\text{PD}(s)$ are equivalent with respect to the PD shape, if

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

Figure 1 shows examples of two ROC curves of two PD functions that are equivalent with respect to shape (and level). It is straightforward to show that if the function $\text{PD}(s)$ is constant the resulting AUROC is equal to 0.5. Table 1 tabulates five examples of PD functions that are equivalent the one way or the other – two of the PD functions have AUROC figures equal to 0.5. In general we can state the following relationships among functional equivalence, almost sure equivalence, level equivalence, and shape equivalence

Theorem 1.9. *Let $\widetilde{PD}(s)$ and $PD(s)$ be two PD functions.*

a) If the two PD functions are functionally equivalent then they are also almost surely equivalent.

b) If the two PD functions are almost surely equivalent then they are also equivalent with respect to the PD level.

c) If the two PD functions are almost surely equivalent then they are also equivalent with respect to the PD shape.

Proof. The proof can be found in the Appendix. ■

Hence, two functionally equivalent PD functions have the same level and shape.

2 One-Period-Based Statistic Inference

In this section we derive statistical tests in order to address the problem whether the empirical default frequency corresponds to the expected default frequency. We start by comparing these figures for only one observation in time, typically on a yearly basis.

2.1 Testing of PD Level

One naive approach would be by directly assuming an approximate distribution for the one-period default frequency $\hat{\pi}$, e.g. a β-distribution,

$$\mathbb{P}\left\{\hat{\pi} \leq t\right\} \quad \cong \quad \int_0^t \beta(a,b)^{-1} z^{a-1} \left(1-z\right)^{b-1} dz, \tag{5}$$

where $\beta(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$ and with a corresponding calibration, i.e. choosing values for a and b. The following section is supposed to give some insights regarding the distribution of $\hat{\pi}$ or the number of defaulters N_1, respectively.

We start with restrictive distributional assumptions and over the course of the section we will relax step by step some of these constraints. We proceed by deriving test statistics with the following distributional constraints,

i) $Y_i|\mathbf{S},\mathbf{V},\mathbf{Y} \backsim Y_i$,

ii) $Y_i|\mathbf{S},\mathbf{V},\mathbf{Y} \backsim Y_i|\mathbf{V}$,

iii) $Y_i|\mathbf{S},\mathbf{V},\mathbf{Y} \backsim Y_i|S_i$,

iv) $Y_i|\mathbf{S},\mathbf{V},\mathbf{Y} \backsim Y_i|S_i,\mathbf{V}$.

First i), we assume that the default indicator is orthogonal to credit scores and systematic factors as well as default indicators of other obligors. In this case Y_i form an independent and identically distributed Bernoulli sequence with parameter π. Hence, we are in a position to deduce the limiting distribution in three steps. Firstly, for the number of defaults N_1 in a portfolio of n obligors, by the very definition of a binomial distribution, we derive

$$N_1 \quad \backsim \quad B\left(n,\pi\right). \tag{6}$$

Secondly, according to the De-Moivre-Laplace global limit theorem we arrive at

$$\lim_{n\to\infty} \mathbb{P}\left\{\frac{N_1 - n\pi}{\sqrt{n\pi(1-\pi)}} \leq t\right\} = \Phi\left(t\right). \tag{7}$$

Thridly, according to a basic convergence theorem of Cramér[3], we can replace the theoretical standard deviation with the empirical one and we still have an asymptotic Gaussian distribution,

$$\lim_{n\to\infty} \mathbb{P}\left\{\frac{N_1 - n\pi}{\sqrt{\frac{n}{n-1}n\hat{\pi}(1-\hat{\pi})}} \leq t\right\} = \Phi\left(t\right). \tag{8}$$

Second ii), we still maintain that credit score and default indicator are independent, in particular $Y_i|\mathbf{S},\mathbf{V},\mathbf{Y} \backsim Y_i|\mathbf{V}$, but induce default clustering through the supposition of a Bernoulli mixture model. Economic history shows that the basic assumption of the binomial model is not fulfilled as

[3]If 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)

borrower defaults tend to default together. As such, default correlations exist and have to be taken into account. In a mixture model the default probability of an obligor is assumed to depend on a set of common factors (typically one). Given the common factors default events of different obligors are independent. Dependence between defaults hence stems from the dependence on a set of common factors.

Definition 2.1 (Bernoulli Mixture Model). Given some $k < n$ and a k dimensional random vector $\mathbf{V} = (V_1, ..., V_k)'$, the random vector $\mathbf{Y} = (Y_1, ..., Y_n)'$ follows a Bernoulli mixture model if there are functions $Q_i : \mathbb{R}^k \to [0, 1]$, such that conditional on $\mathbf{V}$ the default indicators $\mathbf{Y}$ are a vector of independent Bernoulli random variables with $\mathbb{P}\{Y_i = 1|\mathbf{V}\} = Q_i(\mathbf{V})$.

Due to our assumption of a homogeneous loan portfolio the functions $Q_i(\mathbf{V})$ are all identical, so that $\mathbb{P}\{Y_i = 1|\mathbf{V}\} = Q(\mathbf{V})$ for all i. It is 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

$$\mathbb{P}\{N_1 = m\} = \binom{n}{m} \int_0^1 z^m (1 - z)^{n-m} \, dG(z). \tag{9}$$

Further simple calculations give the probability of default π and the joint probability of default π_2

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

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

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

which means that in an exchangeable Bernoulli mixture model the so-called default correlation ρ_Y is always nonnegative. Any value of ρ_Y in $[0, 1]$ can be obtained by an appropriate choice of the mixing distribution. The following one-factor exchangeable Bernoulli mixture models are frequently used in practice:

- Probit-normal mixing-distribution with $Z = \Phi(V)$ and $V \backsim N(\mu, \sigma^2)$ (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 $V \backsim N(\mu, \sigma^2)$ (CreditPortfolioView model; see Wilson [1998]),
- Beta mixing-distribution with $Z \backsim \text{Beta}(a, b)$ with density $g(z) = \beta(a, b)^{-1} z^{a-1} (1 - z)^{b-1}$ where $a, b > 0$ (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

$$\begin{aligned} \mathbb{P}\{N_1 = m\} &= \binom{n}{m} \frac{1}{\beta(a,b)} \int_0^1 z^{a+m-1} (1-z)^{b+n-m-1} \, dz \\ &= \binom{n}{m} \frac{\beta(a+m, b+n-m)}{\beta(a,b)}, \end{aligned} \tag{10}$$

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 $\pi_2 = \mathbb{E}\left[Z^2\right]$ 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 seminal paper of Merton [1974]. The following proposition shows how asset correlation and default correlation are related.

Proposition 2.2. *Given a homogeneous portfolio, the unconditional probability of default π as well as the asset correlation ρ in the one-factor CreditMetrics framework, the joint probability of default π_2 and the default correlation ρ_Y can be calculated as*

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

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

Proof. The proof can be found in the Appendix. ■

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 2.3. *Denote by $G^{-1}(\alpha)$ the α-quantile of the mixing distribu-*

tion G of Z, i.e. $G^{-1}(\alpha) = \inf\{z : G(z) \geq \alpha\}$, and assume that the quantile function $\alpha \to G^{-1}(\alpha)$ is continuous in α, so that

$$G(G^{-1}(\alpha) + \delta) > \alpha \text{ for all } \delta > 0, \tag{11}$$

then

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

Proof. The proof can be found in Frey and McNeil [2001]. ■

In particular, if G admits a density g (continuous random variable) which is positive on $[0, 1]$ the condition (11) is satisfied for any $\alpha \in (0, 1)$.

Third iii), we now work under the assumption $Y_i|\mathbf{S}, \mathbf{V}, \mathbf{Y} \backsim Y_i|S_i$, so that the 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

$$\lim_{n\to\infty} \mathbb{P}\left\{ \frac{N_1 - \mathbb{E}\left[N_1|\,\mathbf{S}\right]}{\sqrt{\mathbb{V}\left[N_1|\mathbf{S}\right]}} < t \middle| \mathbf{S} \right\} = \Phi\left(t\right), \tag{12}$$

where

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

Fourth iv), in the most general case, $Y_i|\mathbf{S}, \mathbf{V} \backsim Y_i|S_i, \mathbf{V}$, so that defaults are clustered in the sense that the default indicator depends on the business

cycle then we can deduce

$$\mathbb{P}\{N_1 = m|\mathbf{S}\} = \int_{\mathbb{R}^k} \sum_P \prod_{i=1}^n \mathbb{P}\{Y_i = \Pi(i)|S_i, \mathbf{V} = \mathbf{v}\}\, d\mathbb{F}_{\mathbf{V}}(\mathbf{v}), \quad (13)$$

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 $\{\Pi(1), ...\Pi(m), \Pi(m+1), ..., \Pi(n)\}$ of $\{1, ..., 1, 0, ..., 0\}$. Usually, the derivation of the distribution of (13) requires Monte-Carlo simulations or numerical integration procedures. Therefore, we suggest to approximate the distribution by a beta-binomial distribution derived in (10). In order to calibrate the beta-binomial distribution we fix the asset correlation ρ and we set π equal to the average default probability

$$\begin{aligned} \pi &= \int_{-\infty}^{\infty} \mathrm{PD}(s) d\mathbb{F}_S(s) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{P}\{Y_i = 1|S_i = s\}. \end{aligned} \quad (14)$$

The choice of the parameter ρ is not so obvious. The higher ρ the more are defaults clustered in time. For instance in the German speaking area and middle market corporate loans, $\rho = 0.05$ appears to be appropriate for a one-year-horizon (see also Tasche [2003]). Internationally, the Basel Committee on Banking Supervision [2005] considers default correlations, ρ_Y, between 0.5% and 3% as typical.

A remark regarding the selection of the various level statistics: If the level testing of the PD functions span a long period of time, possibly a whole credit cycle, then the independence assumption for the test statistics in (6), (7), (8), and (12) is warranted. This is true since by assuming mean ergodicity for the process the averaged yearly default rate over a business cycle converges to the unconditionally expected default frequency, and within a cycle 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

$$\frac{\sum_{t=1}^{T}\left(\bar{p}_t - \mathbb{E}\left[\bar{p}_t|\mathcal{F}_{t-1}\right]\right)}{\sqrt{\sum_{t=1}^{T}\mathbb{V}\left[\bar{p}_t|\mathcal{F}_{t-1}\right]}}, \tag{15}$$

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 default correlations have to be taken into account. In this instance the test statistics in (10) and (13) are more appropriate.

2.2 Testing of PD Shape

The shape of the PD function is visualized by the ROC curve. The realized or empirical ROC curve can be plotted against the theoretical ROC graph and PD miscalibrations can be detected 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

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

can be compared to the theoretical one as defined in (3).[4] The empirical and true ROC curve are, under the assumptions outlined in Section 1, asymptotically equivalent what is stated in the following theorem:

[4]Note the empirical distribution functions are unbiased since

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

where the first equality follows by orthogonality (Assumption 1.2). The rest is computational.

Theorem 2.4. *The empirical and theoretical ROC curve converge almost surely, so that*

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

as $n\to\infty$.

Proof. The proof can be found in the Appendix. ■

If 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, 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. This can be stated as a proposition:

Proposition 2.5. *If we have two monotonic PD functions* $\widetilde{PD}(s)$ *and* $PD(s)$*, so that*

$$\widetilde{PD}(s) \leq PD(s) \text{ for all } s\in\mathbb{S} \tag{16}$$

$$\widetilde{PD}(s) \geq PD(s) \text{ for all } s\in\mathbb{S}^c, \tag{17}$$

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* (16) *and* (17) *for some* s *with positive probability measure, so that*

$$0<\int_{\mathbb{S}}\widetilde{PD}(s)d\mathbb{F}_S(s) < \int_{\mathbb{S}}PD(s)d\mathbb{F}_S(s) \tag{18}$$

$$0<\int_{\mathbb{S}^c}PD(s)d\mathbb{F}_S(s) < \int_{\mathbb{S}^c}\widetilde{PD}(s)d\mathbb{F}_S(s), \tag{19}$$

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*

$$\widetilde{AUROC} > AUROC$$

Proof. The proof can be found in the Appendix. ■

We are also in a position to construct confidence bands for the ROC curve (see e.g. Macskassy, Provost, and Littman [2004]). Out of robustness considerations we focus our attention on the area below the curve and not the curve itself. We denote the empirical AUROC figure by $\widehat{\text{AUROC}}_n$. This estimator is given by

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

where the index i (j) indicates summation over defaulters (non-defaulters) and where $N_1 = \sum_{i=1}^{n} Y_i$ and $N_0 = \sum_{i=1}^{n} (1 - Y_i)$ denote the number of defaulters and non-defaulters, respectively. Only for notational convenience we added a subscript D and ND for the defaulter's and non-defaulter's score, respectively. The AUROC estimator is consistent and unbiased as derived in the following proposition:

Proposition 2.6. *The (conditional) expectation and variance of the estimator $\widehat{AUROC}_n$ is equal to*

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

Further,

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

Proof. The proof can be found in the Appendix. ■

Note, the corresponding event probabilities for the calculation of B, B_{001}, and B_{110} are computed 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 $\widehat{\text{AUROC}}_n$ is Gaussian:

Proposition 2.7. *The AUROC statistic has the following limiting distribution*

$$\lim_{n\to\infty} \mathbb{P}\left\{ \frac{\widehat{AUROC}_n - AUROC}{\sqrt{\mathbb{V}\left[\widehat{AUROC}_n | \mathbf{Y}\right]}} \middle| \mathbf{Y} \right\} = \Phi(t). \tag{20}$$

Proof. The proof can be found in Lehmann [1951]. ■

The theoretical standard deviation in the denominator in equation (20) of proposition 2.7 can be replaced by the empirical counterpart and the limiting distributions is still Gaussian according to a basic theorem of Cramér [1946] (Theorem 20.6, see also Bamber [1975]). Proposition 2.6 and Proposition 2.7 are generalizations of the seminal papers of Wilcoxon [1945] as well as Mann

and Whitney [1947]. The following Wilcoxon-Mann-Whitney Corollary is therefore appropriate in case of the "horizontal" PD function.[5]

Corollary 2.8 (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*

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

with the limiting distribution

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

There are a number of standard statistical measures to describe how different defaulters and non-defaulters are in their characteristics. These measure how well the PD function separates the two groups, we looked at one measure – the ROC statistic. In Thomas, Edelman, and Crook [2002] we find other measures – like the Mahalanobis distance and Kolmogorov-Smirnov statistics. Theoretically, these statistics are suited as well to perform shape tests.

2.3 Goodness-of-Fit

In the two previous sections we have derived level and shape statistics. Usually the limiting distributions of the test statistics are standard normal. If the distribution is (asymptotically) different from a standard Gaussian one

[5]Note that the expectation for the AUROC statistic is also 0.5 for the case when 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 2.6. However, a non-diagonal ROC graph with AUROC 0.5 violates the monotonicity assumption of the PD function.

transforms the realized estimate into a standard normal quantile according to the following lemma.

Lemma 2.9. *If the random variable X is distributed according to the continuous distribution function G, then*

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

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

Proof. The proof can be found in the Appendix. ■

The shape statistic is based on scores conditional on the default indicators. According to the orthogonality assumptions (Assumption 1.2) 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 that[6]

$$S_i|\mathbf{S}, \mathbf{V}, \mathbf{Y} \backsim S_i|Y_i, N_1, \mathbf{V} \backsim S_i|Y_i.$$

This means that level and shape statistics are independent. A high figure in the PD level statistic does not on average imply a high (or a low) number for the PD shape statistic. We are now in a position to deduce a summary statistic in order to test globally the null hypothesis of a correctly calibrated PD function with respect to 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 $1-(1-\alpha)^2$. In case of a 5% significance level, there is a chance of 9.75% of at least one of the two tests being declared significant under the null hypothesis. One very simple method, due to Bonferroni [1936], to circumvent this problem is to divide the test-wise significance level by the number of

[6]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 by $\sigma(\mathbf{S}, \mathbf{V}, \mathbf{Y})$, in particular it is true that $\sigma(\mathbf{S}, \mathbf{V}, \mathbf{Y}) \supseteq \sigma(Y_i, N_1, \mathbf{V}) \supseteq \sigma(Y_i)$.

tests. Unfortunately, Bonferroni's method generally does not result in the most powerful test, meaning that there are critical regions with the same size but higher power according to Neyman-Pearson's lemma. That is why we resort to the likelihood ratio Λ,

$$\Lambda = \exp\left[-\frac{1}{2}\left(T_{level}^2 + T_{shape}^2\right)\right], \tag{21}$$

where T_{level} denotes one of the level statistics in (5), (6), (7), (8), (10), (12), (13), and (15), T_{shape} denotes the shape statistic in (20). The statistics are first transformed into a standard Gaussian quantile according to Lemma 2.9. The likelihood-ratio test rejects the null hypothesis if the value of the statistic in (21) is too small, and is justified by the Neyman-Pearson lemma. If the null hypothesis is true, then $-2\log\Lambda$ will be asymptotically distributed with degrees of freedom equal to the difference in dimensionality. Hence, we derive asymptotically

$$T_{level}^2 + T_{shape}^2 \quad \backsim \quad \chi^2\langle 2\rangle\,. \tag{22}$$

Therefore, the critical value for the global test in (22) on a confidence level of 95% (99%) is 5.9915 (9.2103).

2.4 Simulation Study

As the design of our test procedure is based on assumptions as outlined in Section 1, we check its robustness with respect to violations. A simulation study allows us to draw conclusions on the robustness of the validation procedure in case of misspecifications and approximations. 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. The performance of our approach is then compared to the performance of a benchmark statistic, the well-known and well-documented Hosmer-Lemeshow's χ^2-goodness-of-fit test (see e.g.

Hosmer, Hosmer, le Cessie, and Lemeshow [1997]). A common feature of both these tests is the suitability of being applied to several rating categories simultaneously. Hosmer-Lemeshow's χ^2-test is based on the assumption of independence and a normal approximation. Due to the dependence of default events that are observed in practice and the generally low frequency of default events, Hosmer-Lemeshow's χ^2-test is likely to underestimate the true type I error, i.e. the proportion of erroneous rejections of PD forecasts will be higher than expected from the formal confidence level of the test. Hosmer-Lemeshow's χ^2-test statistic is defined as

$$T = \sum_{j=1}^{C} \frac{n_j \left(\hat{\pi}_j - \pi_j\right)^2}{\pi_j(1-\pi_j)}, \tag{23}$$

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.

Both Hosmer-Lemeshow's χ^2-test as well as our global test statistic are derived by asymptotic considerations with regard to the portfolio size. As a consequence, even in the case of complete independence in the loan portfolio it is not clear that the type I errors observed with the tests are dominated by the nominal error levels. When compliance with the nominal error level for the type I error is confirmed, the question has to be examined which test is the more powerful, i.e. for which test the type II errors is lower. Of course, the compliance with the nominal error level is much more an issue in the case of dependencies of the default events in the portfolio. The tests should have small type I and type II error rates for calibrations with economic significance. The higher the profit impact at stake, the more powerful the statistics have to be.

We now turn to the simulation setup in order to address the question of

size and power of the test statistics under various circumstances. To induce default correlation we model the asset value Y_i^* for each obligor 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. The asset correlation between two obligors is denoted by ρ. The higher the asset correlation, the more the systematic risk factor X dominates, thus resulting in a collapse of the default rates in either a high or a low overall default rate in the portfolio. The default event is defined by

$$Y_i = \begin{cases} 0 & : \quad Y_i^* > D_i \\ 1 & : \quad Y_i^* \leq D_i \end{cases}, \tag{24}$$

where D_i denotes the distance to default calculated by the standard Gaussian quantile of the default probability and is the same value for all obligors in a given rating category. For the simulation study we assume that D_i is orthogonal to both X and ϵ_i. The distance to default can therefore be interpreted as a "through the business cycle" credit score. This setup imposes quite strong assumptions because credit scores are usually computed from balance sheet information and since the aggregate of balance sheets make up the economy one might very well argue that such a credit systems is never fully "through the cycle", and it also violates the orthogonality assumption established in Assumption 1.2.

We consider 4 different correlation regimes (0, 0.05, 0.10, and 0.15) and 3 different numbers of rating classes (15, 10, and 5) resulting in 12 scenarios. We run 10'000 Monte Carlo simulations under each scenario where the Hosmer-Lemeshow test and our validation procedure are two independent simulation series. The (unconditional) expected default frequency under the

data generating process is fixed for all scenarios at 3% (the average default probability is 2.5% in case of type II error analyses), and the size of the portfolio is set at 10'000 obligors. The true (alternative) AUROC figures are 0.6112, 0.6279, 0.6509 (0.6354, 0.6551, 0.6816) for 15, 10, and 5 rating classes, respectively. Table 2 outlines 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 (22) we rely on a "beta"-approximation for the level T_{level} as in (10) and the statistic T_{shape} in (20) for testing the shape. We calibrate the beta-binomial distribution according to Proposition 2.2 with an average default probability of 3% (2.5%), as computed by (14), for the type I error analyses (type II) and a fixed asset correlation ρ of 5% for all but one correlation regime. This gives us the parameters $a = 3.4263$ (3.2203) and $b = 110.7850$ (125.5922) for type I error considerations (type II). In case of zero asset correlation we omit the "beta"-approximation and we work with the approximate level statistic as outlined in (12).

Table 2 and Table 4 report the simulation results under a nominal error level of 5% and 1%, respectively. The results indicate that under independence all test methodologies, Hosmer-Lemeshow's χ^2, global, level, and shape statistics, seem to be more or less in compliance with the nominal error levels. However the former test fits the levels worse than the latter ones – the true type I errors are in absolute terms up to 3% higher than the nominal levels. Under low asset correlation regimes of up to 5%, the global test statistic is still essentially compliant with the nominal error levels whereas Hosmer-Lemeshow's χ^2-test is distorted. When compliance with the nominal type I error is established the power of the test statistics are assessed via type II error. The global test procedure is more powerful under independence with true type II error levels around 10% (23%) at 5% (1%)

nominal level, than Hosmer-Lemeshow's χ^2 resulting in type II errors of up to about 37% (55%).

Under asset correlation regimes higher than 5% both overall test procedures, Hosmer-Lemeshow's χ^2 and global, 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 and therefore inducing a conservative distortion. But the distortion is quite high for Hosmer-Lemeshow's χ^2-test.

The power of all test statistics decrease with the size of the asset correlation. A test is said to be unbiased if the power for the alternative exceeds the level of significance. Under asset correlation regimes higher than 5%, Hosmer-Lesmeshow's χ^2 is biased, the sum of true type I and type II error exceeds one or is close to one rendering it virtually useless for practical considerations. This is not the case for the global test statistic even though the applicability of the procedure might also 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 unless the special case of zero asset correlation. According to the orthogonality assumption established in Section 1 the shape statistic is consistent even for short time horizons. Over time, also the level analysis, e.g. (15), provides us with consistent estimators.

Altogether, the global test statistic is more robust and more powerful against misspecifications than Hosmer-Lemeshow's χ^2. Unlike Hosmer-Lemeshow's χ^2 the global test is unbiased for the scenarios considered in the simulation setup. This is mainly driven by the fact that the shape statistic is not very vulnerable to misspecifications. Especially for typical scenarios encountered in practice, ten to fifteen rating classes and asset correlations around 5%, the shape statistic performs reasonably well. The shape-test is more or less in line with the nominal error level and it does not lose power

under small default dependency structures. Credit scores for corporates anticipate, at least to some extent, economic recessions due to the incorporation of financial figures resulting only in small but significant residual default dependencies. Hence, for scenarios with the highest economic and pracitical relevance, the global test statistic performs better than Hosmer-Lemeshow's χ^2.

3 Multi-Period-Based Statistic Inference

In general we can state that the longer the observation time the more reliable the test results. However, the question might arise whether our proposed global test statistic over a time-period of T years (one-period approach) should be split up into T statistics at one year each (multi-period approach). The reasons are many, for some borrowers we do not know the whole T-year credit history because they have entered the loan portfolio later or left it beforehand. This leaves us with the problem of missing observations. It is also the case that banks are validating and aligning their credit scoring systems quite regularly by means of incorporating additional information and/or changing the weighting schemes of input variables. For some scoring systems a complete default term structure might not be available forcing the controller to resort to the one-year probabilities of default. In the medium to long-term a controller, supervisor or developer might rather want to validate the holistic rating systems than a particular rating model. We therefore introduce a traffic light system enabling the flexible validation over time.

In Blochwitz, Hohl, and Wehn [2005], a traffic light approach is presented as a tool to identify poorly calibrated rating grades over a multiple of data points. Their procedure is applied to one single rating category at any one time. We extend their approach to simultaneous monitoring of several rating grades since for rating systems with many grades a purely random rejection of appropriate estimation for one or two grades becomes very likely.

Our proposal is based on the assumption of no correlation in time for the goodness-of-fit statistics in (22). For the traffic-light-statistic, probabilities with $\pi_g+\pi_y+\pi_o+\pi_r = 1$ (corresponding to the colors green, yellow, orange, and red), $\pi_g > \pi_y > \pi_o > \pi_r > 0$ and a color mapping $C(x)$ are defined by

$$C(x) = \begin{cases} g & \text{if} \quad x \leq F^{-1}_{\chi^2\langle 2\rangle}(\pi_g) \\ y & \text{if} \quad F^{-1}_{\chi^2\langle 2\rangle}(\pi_g) < x \leq F^{-1}_{\chi^2\langle 2\rangle}(\pi_y) \\ o & \text{if} \quad F^{-1}_{\chi^2\langle 2\rangle}(\pi_y) < x \leq F^{-1}_{\chi^2\langle 2\rangle}(\pi_o) \\ r & \text{if} \quad F^{-1}_{\chi^2\langle 2\rangle}(\pi_o) < x \end{cases},$$

where $F^{-1}_{\chi^2\langle 2\rangle}(\pi_y)$ denotes the quantile function of the χ^2-distribution function with two degrees of freedom. With this definition, under the assumption of independence of the periodical test statistic in (22), the vector (L_g, L_y, L_o, L_r) with L_c counting the appearances of color $c \in \{g, y, o, r\}$ will be multinomially distributed with

$$\mathbb{P}\{(L_g = l_g, L_y = l_y, L_o = l_o, L_r = l_r)\} \quad = \quad \frac{(l_g + l_y + l_o + l_r)!}{l_g! l_y! l_o! l_r!} \pi_g^{l_g} \pi_y^{l_y} \pi_o^{l_o} \pi_r^{l_r}.$$

Now the only thing that is left is to construct an order function to all quadruples for ranking all of them according to the difference between empirical and theoretical PD function. Blochwitz, Hohl, and Wehn [2005] decided to apply a quite intuitive approach to an order of the quadruples, namely

$$\lambda(L_g, L_y, L_o, L_r) \quad = \quad \pi_g L_g + \pi_y L_y + \pi_o L_o + \pi_r L_r.$$

With the severity measure $\lambda(L_g, L_y, L_o, L_r)$ it is decided if a scoring model is correctly calibrated for a time series of default forecasts. The smaller the value of $\lambda(L_g, L_y, L_o, L_r)$ the more severe we judge the underlying observation of deviations between empirical and theoretical default frequencies. Although this represents a quite intuitive approach to an order of the quadruples and despite the fact that counter examples to the order can be

constructed, various different constellations in Blochwitz, Hohl, and Wehn [2005] showed that more sophisticated techniques did not lead to deeper insights. The concept of ordering the quadruples is illustrated in Table 3 for the example of $L = 4$ as published in Blochwitz, Hohl, and Wehn [2005].

4 Conclusions

The validation of the probability calibration has several components. Our goal is to provide a comprehensible tool for backtesting probability calibrations in a quantitative way. We therefore focus on two important quantitative components – level and shape. The level evaluation is based on a comparison of ex ante expected frequencies and the realized ex post rates. We propose level statistics that are derived under dependencies, e.g. credit default correlations are modeled via Bernoulli-mixture models. The second component, the shape, compares the theoretical area below the receiver operating characteristic curve (AUROC) with the empirical one. This approach has the great advantage of visualizing the results through a graph. That allows us to visually detecting probability miscalibrations and facilitates the selection of samples for a deeper examination.

In statistics, a test procedure is said to be consistent against a certain class of alternatives if for each alternative the power of the test tends to one as the sample size tends to infinity. Consistency, even though it is usually a rather weak property, is not granted in case of credit scoring. Due to cross-correlations between obligors consistency can only be achieved over time. However, our proposed validation procedure is not designated to distinguish between all functionally different calibrations but those that are economically relevant. In credit scoring where default events are scarce compels the model-controller to focus on the most serious miscalibrations from an economic perspective. A financial institution has to avoid adverse selection of loans that is mainly caused by level and shape deviations of the

probability calibration. If the level is too high, the financial institution will systematically lose market shares to competitors. A too low level might force the bank out of the loan market since risk-premiums do not cover losses in the long run. If the probability mapping is wrongly shaped then some groups of obligors subsidize other groups. For instance, investment grade obligors are charged too low in comparison to sub-investment rated borrowers. In the worst case, this might lead to a bank-failure because competitors will exploit the resulting mispricing of loans.

Our test procedure is meant to be applied to the whole population at a time but has the great flexibility to be only employed for a subpopulation (e.g. separation into investment grade and non-investment grade borrowers). 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. The comparison of the global test statistic and the well-known Hosmer-Lemeshow's χ^2 was carried out by means of a simulation study. Reliability with respect to type I error levels as well as power measured by type II error sizes were examined. Overall, the performance of the global test statistic is better than the performance of Hosmer-Lemeshow's χ^2. We show that the global test is more robust against misspecifications especially when it comes to validating of credit scoring systems where event clusterings have to be taken into account.

Our testing procedure is also applicable in the case of continuous probabilities where two events have almost surely different (conditional) probabilities. If we are confronted with continuous probabilities or a lot of categories then existing calibration tests (i.e. binomial test, normal test, Hosmer-Lemeshow's χ^2-test, and the traffic lights approach of Blochwitz, Hohl, and Wehn [2005]) are virtually powerless – the true probability might be anywhere between zero and one. In this case our procedure offers a viable alternative.

We extend our methodology for a single data point with the inclusion of the dimension of time. Missing observations over time, changing the underlying explanatory variables on a regular basis, or the absence of a probability term structure, might force the model-controller to divide a long-period of time (single data point) into sub-periods (multiple data points) in order to validate the system as a whole. Therefore, we combine a time-series of global test statistics into a single multi-period test. The proposed traffic light approach is a rule-based system assessing the differences between theoretical and empirical event frequencies which has the merit that it is easy to implement. This approach is an extension of the methodology suggested by Blochwitz, Hohl, Tasche, and Wehn [2004]. Their method is applicable to a single rating class at any one time. Our extension allows a simultaneous monitoring of several rating grades what represents a major step forward since for systems with many grades/classes a purely random rejection of appropriate estimation for one or two grades becomes very likely.

We leave it to future studies to further check the robustness and reliability of our validation procedure. For instance, the performance of the traffic-light approach needs to be compared to the power of the global test statistic. Such a study should also comment on the subdivision of a single period into sub-periods for which the global test statistic are applied resulting in a time-series of global tests. Such an optimal subdivision might prove to be difficult to derive due to a decreasing default correlation over time. Another string of research might deal with the exact distribution of the proposed test statistics or some bootstrap methods in order to derive statistics under less stringent assumptions.

A Appendix

Proof of Theorem 1.9. 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 $A \subseteq \Omega$. 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 2.2. Let Y_i^* and Y_j^* be the CreditMetrics latent variables for two obligors, $i \neq j$. There is only one systematic risk factor X and, since we have a homogeneous portfolio, the two obligors 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

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

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

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

where the first line follows by iterating expectations (see Proposition 8.13 of

Karr [1993]), so that $\mathbb{E}\left[\mathbb{P}\left\{Y_i^* \leq C, Y_j^* \leq C|X\right\}\right] = \mathbb{P}\left\{Y_i^* \leq C, Y_j^* \leq C\right\}$. ■

Proof of theorem 2.4. Consider the inequality

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

if we apply the Glivenko-Cantelli Theorem for the first term on the right hand side and the theorem of Dvoretzky, Kiefer, and Wolfowitz [1956] and then the Borel-Cantelli Lemma for the second term, the theorem is proved. ■

Proof of Proposition 2.5. From (16) and (17) as well as the basic integration rule of monotonicity[7] we can derive that

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

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

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

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

$$\tilde{\mathbb{F}}_{S_D}(t) \leq \mathbb{F}_{S_D}(t) \text{ for all } t\in\mathbb{R}, \tag{25}$$

[7] If either $0 \leq g \leq h$ or g and h are integrable and $g \leq h$, then $\int g dF \leq \int h dF$.

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

$$\tilde{\mathbb{F}}_{S_{ND}}(t) \quad \geq \quad \mathbb{F}_{S_{ND}}(t) \text{ for all } t \in \mathbb{R}, \tag{26}$$

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

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

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. ■

Proof of Proposition 2.6. The estimate $\widehat{\text{AUROC}}_n$ is unbiased since

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

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

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

Now, we can differentiate between four different instances:

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

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

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

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

   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

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

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

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

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 cancellations lead to the desired result. ■

Proof of Lemma 2.9. From two well-known theorems, see for instance Theorems 2.47 and 2.48 in Karr [1993] for the proofs, we know that a) $G(X)$ is uniformly distributed, and that b) $\Phi^{-1}\left(G(X)\right)$ is standard Gaussian distributed. ■

References

Balthazar, L. (2004): "PD estimates for Basel II," *Risk Magazine*, 17, 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, D. Tasche, and C. S. Wehn (2004): "Validating Default Probabilities on Short Time Series," Working paper, Deutsche Bundesbank.

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: Princeton University Press.

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 Clas-

sical Multinomial Estimator," *The Annals of Mathematical Statistics*, 27, 642–669.

Epstein, E. S. (1969): "A Scoring System for Probability Forecasts of Ranked Categories," *Journal of Applied Meteorology*, 8, 985–987.

Foster, D. P. and R. V. Vohra (1998): "Asymptotic Calibration," *Biometrika*, 85, 379–390.

Frey, R. and A. J. McNeil (2001): "Modelling dependent defaults," Working paper, Unversity of Zurich and ETH Zurich.

Fudenberg, D. and D. Levine (1999): "An easier way to calibrate," *Games and Economic Behavior*, 29, 131–137.

Gupton, G. M., C. C. Finger, and M. Bhatia (1997): "CreditMetrics," Technical document, J.P. Morgon & Co.

Harrison, J. and D. Kreps (1979): "Martingales and Arbitrage in Multiperiod Securities Markets," *Journal of Economic Theory*, 20, 381–408.

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, 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, 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," .

Hosmer, D. W. and S. Lemeshow (1989): *Applied Logistic Regression*, New York: John Wiley & Sons, Inc.

Karr, A. F. (1993): *Probability*, New York: Springer Verlag.

Lehmann, E. L. (1951): "Consistency and unbiasedness of certain nonparametric tests," *Annals of Mathematical Statistics*, 22, 165–179.

Lemeshow, S. and J. R. Le Gall (1994): "Modeling the Severity of Illness of ICU patients," *Journal of the American Medical Association*, 272, 1049–1055.

Macskassy, S. A., F. J. Provost, and M. L. Littman (2004): "Confidence Bands for ROC Curves," Working paper, Stern School of Business, New York University.

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, 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.

Stein, R. M. (2002): "Benchmarking Default Prediction Models: Pitfalls and Remedies in Model Validation," Tech. rep., Moody's KMV company.

Tasche, D. (2003): "A traffic lights approach to PD validation," Working paper, Deutsche Bundesbank, Frankfurt am Main, Germany.

Thomas, L. C., D. B. Edelman, and J. N. Crook (2002): *Credit Scoring and Its Applications*, Philadelphia: Society for Industrial and Applied Mathematics.

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*, 131–137.

Rating	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8
$\mathrm{PD}(s)$	0.0009	0.0018	0.0043	0.0108	0.0339	0.0767	0.1110	0.1500
$\widetilde{\mathrm{PD}}(s)$	0.0002	0.0008	0.0028	0.0135	0.0339	0.0669	0.0922	0.1500
$\widehat{\mathrm{PD}}(s)$	0.0009	0.0018	0.0043	0.0108	0.0339	0.0767	0.1110	0.2000
$\mathrm{PD}_1(s)$	0.0154	0.0154	0.0154	0.0154	0.0154	0.0154	0.0154	0.0154
$\mathrm{PD}_2(s)$	0.0175	0.0175	0.0175	0.0175	0.0175	0.0175	0.0175	0.0175
$d\mathbb{F}_S(s)$	0.0300	0.0700	0.2900	0.3800	0.1900	0.0300	0.0100	0

Table 1: The PD functions $\mathrm{PD}(s)$ and $\widetilde{\mathrm{PD}}(s)$ have the same PD level (=1.54% average PD) and the same PD shape (=75.99% AUROC) even though they are functionally and almost surely not equivalent. The PD functions $\mathrm{PD}(s)$ and $\widehat{\mathrm{PD}}(s)$ are equivalent with respect to level, shape and almost surely, but not functionally. $\mathrm{PD}(s)$ and $\mathrm{PD}_1(s)$ have the same PD level but different PD shapes whereas $\mathrm{PD}_1(s)$ and $\mathrm{PD}_2(s)$ have the same shape but different levels. Figure 1 depicts the ROC graphs of the PD functions $\mathrm{PD}(s)$ and $\widetilde{\mathrm{PD}}(s)$.

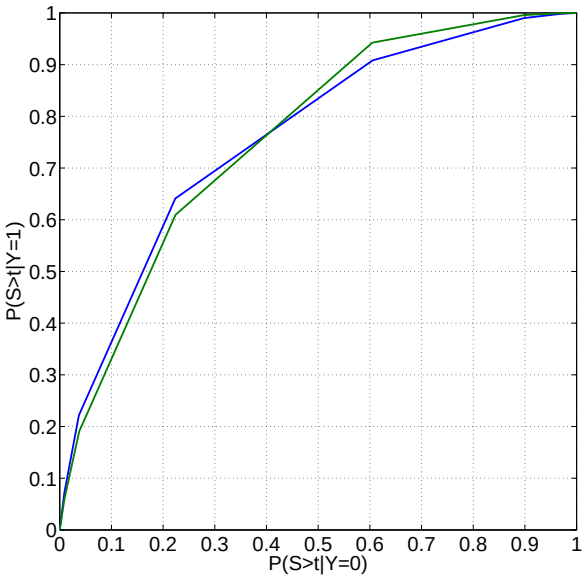

Figure 1: If two PD functions have the same shape (=area under the ROC curve) then this does not imply that they have the same ROC graph. The graph depicts the PD functions $\mathrm{PD}(s)$ and $\widetilde{\mathrm{PD}}(s)$ as tabulated in Table 1.

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 2: 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 obligors. The table outlines the rating distribution along 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 analyses.

Rank	*gyor*	#	λ	Π	Rank	*gyor*	#	λ	Π
1	0004	1	0.2000	0.0000	19	1111	24	1.0000	0.0857
2	0013	4	0.3000	0.0001	20	0310	4	1.0500	0.1019
3	0022	6	0.4000	0.0004	21	2002	6	1.1000	0.1057
4	0103	4	0.4500	0.0006	22	1120	12	1.1000	0.1462
5	0031	4	0.5000	0.0012	23	1201	12	1.1500	0.1732
6	0112	12	0.5500	0.0026	24	2011	12	1.2000	0.1957
7	0040	1	0.6000	0.0031	25	0400	1	1.2000	0.2038
8	1003	4	0.6500	0.0034	26	1210	12	1.2500	0.2848
9	0121	12	0.6500	0.0074	27	2020	6	1.3000	0.3185
10	0202	6	0.7000	0.0088	28	2101	12	1.3500	0.3635
11	1012	12	0.7500	0.0110	29	1300	4	1.4000	0.4175
12	0130	4	0.7500	0.0151	30	2110	12	1.4500	0.5525
13	0211	12	0.8000	0.0232	31	3001	4	1.5500	0.5775
14	1021	12	0.8500	0.0299	32	2200	6	1.6000	0.7125
15	0220	6	0.9000	0.0421	33	3010	4	1.6500	0.7875
16	1102	12	0.9000	0.0466	34	3100	4	1.8000	0.9375
17	0301	4	0.9500	0.0520	35	4000	1	2.0000	1.0000
18	1030	4	0.9500	0.0587					

Table 3: The table is taken from Blochwitz et al. [2005] and displays all realizations of the extended traffic light approach for a time series of $L = 4$. Note, that $(\pi_g, \pi_y, \pi_o, \pi_r) = (0.50, 0.30, 0.15, 0.05)$, # is the number of realizations of the quadruple with severity λ, Π is the cumulative probability of observing events of at least the same severity, i.e. quadruples with the same rank or lower.

		Type I error				Type II error			
ρ	C	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 4: 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.

		Type I error				Type II error			
ρ	C	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 5: 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.