JOURNAL OF FINANCIAL AND QUANTITATIVE ANALYSIS Vol. 47, No. 5, Oct. 2012, pp. 1089–1123
COPYRIGHT 2012, MICHAEL G. FOSTER SCHOOL OF BUSINESS, UNIVERSITY OF WASHINGTON, SEATTLE, WA 98195
doi:10.1017/S0022109012000324

Validation of Default Probabilities

Andreas Blöchlinger*

Abstract

Well-performing default predictions show good discrimination and calibration. Discrimination is the ability to separate defaulters from nondefaulters. Calibration is the ability to make unbiased forecasts. I derive novel discrimination and calibration statistics to verify forecasts expressed in terms of probability under dependent observations. The test statistics' asymptotic distributions can be derived in analytic form. Not accounting for cross correlation can result in the rejection of actually well-performing predictions, as shown in an empirical application. I demonstrate that forecasting errors must be serially uncorrelated. As a consequence, my multiperiod tests are statistically consistent.

I. Introduction

For a lender the use of well-performing default forecasting models that can carry consistent predictive information about credit defaults is of crucial importance in terms of profitability and, ultimately, for survival in the marketplace. It is a long-standing theory in economics (empirically corroborated by the numerous bank failures after the financial market crisis in 2008) that agents who do not predict as accurately as others are driven out of the market (see Alchian (1950), Friedman (1953), and Sandroni (2000)).

The default forecast expressed in terms of the probability of default (PD) is a key input parameter in credit risk management (e.g., to compute the regulatory capital). As a consequence, the validation of PDs is a crucial component of the supervisory review process. Banks must demonstrate that they can assess

*Blöchlinger, andreas.bloechlinger@zkb.ch, Zürcher Kantonalbank, Josefstrasse 222, CH-8010 Zurich, Switzerland. I thank the participants of the 1st European Conference of the Society for Financial Econometrics (2009, Geneva) and the 12th Conference of the Swiss Society for Financial Market Research (2009, Geneva), the group Treasury Engineering at Zürcher Kantonalbank, and the team Credit Risk Analytics at Credit Suisse, in particular Daniel Aunon, Peter Bolleter, Manuele de Gennaro, Aleksandar Georgiev, Jérôme Koller, Basile Maire, Thomas Nittner, Markus Rech, Maria Stepanova, Daniel Straumann, and Urs Wolf for valuable discussions and numerous workshops. I especially thank Basile Maire for assistance in the computation of Merton's probabilities of default (PDs). Furthermore, I thank Hendrik Bessembinder (the editor), Markus Betschart, Christian Bluhm, Peng Cheng, Sanjiv Das (the referee), Daniel Egloff, Markus Leippold, and Paolo Vanini for their valuable comments and suggestions. The content of this paper reflects the personal view of the author. In particular, it does not necessarily represent the opinion of Zürcher Kantonalbank.

the performance of their forecasts consistently and meaningfully. More detailed requirements demand, for instance, that realized default rates have to be within an expected interval (see Basel Committee on Banking Supervision (2005)). The committee states that "the area of validation will prove to be a key challenge for banking institutions in the foreseeable future." In this paper I exactly address this key challenge by developing new statistical tools to validate PDs as suggested in a paper by the Department of the Treasury, Federal Reserve System, and Federal Deposit Insurance Corporation (2003) on supervisory guidance for the internal ratings-based (IRB) approach:

> At this time, there is no generally agreed-upon statistical test of the accuracy of IRB systems. Banks must develop statistical tests to back-test their IRB rating systems.

In weather forecasting, validation methods for probability forecasts are well established (see, e.g., Cooke (1906), Brier (1950)). What makes the verification of default probabilities more challenging than precipitation probabilities? In weather forecasting, over the course of 1 year I have 365 daily probability forecasts. Perhaps, I have 2 separate precipitation forecasting sequences for the north and the south of a state. The forecasting errors between the 2 geographical regions may be cross correlated (e.g., when a cold front brings rainfall for both regions faster than anticipated). The forecasting errors over time, however, must be uncorrelated (i.e., the early arrival of the cold front is incorporated into the prediction when making the following day's rain probability forecasts). In this weather example, I have a long time series and a small cross section. With default probabilities it is reversed. I have a large cross section as measured by the number of borrowers but usually only a few periods of data. Accounting for small cross correlation can reverse the validation outcome. That is, the default predictions are considered well performing when accounting for cross correlation but not so when assuming zero correlation. Further complicating matters, I may have an average of 25% rainy days, but default events are scarcer.

According to the existing literature (e.g., Hosmer and Lemeshow (1989), Harrell (2001), Basel Committee on Banking Supervision (2005)), there are 2 major validation aspects that need to be assessed: discrimination and calibration. Models that distinguish well between borrowers who default and those who survive are said to have good discrimination. Calibration refers to the ability of a model to match predicted and observed default rates across the entire spread of the data. A model in which the number of observed defaults aligns well with the number of defaults expected by the model demonstrates good calibration. A commonly used measure of discrimination is the Gini (1921) index. The higher the value of the Gini index, the higher the discriminatory power. Bamber (1975) provides a method for testing the significance of a model against the naive model. DeLong, DeLong, and Clarke-Pearson (1988) extend this method to the comparison of 2 models. Common measures of calibration are the χ^2 statistics of Pearson (1900) and Hosmer and Lemeshow (1989). Both statistics compare observed with predicted outcomes.

Up to now, according to the Basel Committee on Banking Supervision (2005), commonly used tests have had at least one or often several shortfalls:

i) Many tests are derived under stochastic independence (like the Kolmogorov (1933) statistic or the Spiegelhalter (1986) test based on the Brier score), which is a problematic assumption for validating default predictions, as tests can be biased (i.e., true type I error rates can be substantially higher than nominal significance levels).[1]

ii) Asymptotic distributions are often not valid because of sparseness of defaults (e.g., the goodness-of-fit statistics of Pearson (1900) and Hosmer and Lemeshow (1989)).

iii) Some test statistics rely heavily on numerical methods (e.g., Balthazar (2004)).

iv) Some statistics require borrowers to be grouped (e.g., Pearson's χ^2 and its extensions as in Pollard (1979), Andrews (1988)). Grouping can heavily affect test results.[2]

v) Other tests are applicable to single time periods only, or if a summary statistic over several periods is obtainable, I must first determine a time-series process (e.g., Blochwitz, Hohl, Tasche, and Wehn (2004)).

Due to these shortcomings, the Basel Committee on Banking Supervision (2005) considers current quantitative validation approaches insufficient:

> [A]ll tests based on the independence assumption are rather conservative, with even well-behaved rating systems performing poorly in these tests. 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.
>
> Basel Committee on Banking Supervision ((2005), p. 3)

They conclude that "the validation process is mainly qualitative in nature and should rely on the skills and experience of typical banking supervisors" (p. 9). The experiences from the credit crisis, however, call into question this somehow fuzzy validation framework and call for a more powerful validation procedure. Blöchlinger and Leippold (2011) have recently developed a 1-period calibration statistic under dependent observations that proved to be more powerful than existing tests. I pursue a different goal: Given probability forecasts over several time periods, how can I come to a summary statistic for whether the prediction system is calibrated? I develop multiperiod statistics to test the calibration hypothesis and thereby offer a solution to the issue raised by the Basel Committee: I account for cross correlation, but forecasting errors must be serially uncorrelated so that miscalibrated forecasts can be detected over time.

I am concerned not only with calibration but also with discrimination. In fact, I propose a new calibration test that is directly derived from a discrimination

[1] Andrews (1997) extends the Kolmogorov test but keeps the independence assumption. Andrews (2005) also considers test statistics under common shocks but limits the scope to linear models.

[2] By means of regrouping, Bertolini, Damico, Nardi, Tinazzi, and Apolone (2000) report for a single data set p-values ranging from 1% to 95%.

statistic. That is, I present novel test statistics for both discrimination and calibration under a common set of assumptions that do not suffer from the previous shortcomings: First, I assume a conditional IID (independent and identically distributed) as opposed to an unconditional IID setup. Second, my asymptotic distributions are valid for typical sample sizes in credit risk management. Third, the distributions are either available in analytic forms or can be easily derived with numerical methods. Fourth, my tests do not require grouped data. Fifth, I show that forecasting errors must be serially uncorrelated so that a series of test statistics for single periods can be easily aggregated into a multiperiod summary statistic.

The paper is structured as follows: Section II illustrates the concepts of discrimination and calibration. Section III sets out the 3 key assumptions. Sections IV and V derive new statistics for discrimination and calibration, respectively. Section VI addresses the dependence between observations. Section VII evaluates the performance of the test statistics in a simulation exercise. Section VIII comprises an empirical application. Section IX concludes.

II. Discrimination and Calibration

Discrimination is a concept for ordinal measures of risk (e.g., rating classes), whereas calibration is applicable only for risk measures on a ratio scale (e.g., PDs). The Lorenz (1905) curve is the standard way to depict the discriminatory power, and the related Gini (1921) index is arguably the best-known statistic to quantify the discriminatory power in a single figure. Calibration is a concept widely used in probability forecasting (see, e.g., Dawid (1982), (1985)). The standard measure of calibration is Pearson's χ^2 goodness-of-fit test, which examines the sum of the squared differences between the observed and expected number of defaults per group divided by its standard error. Unfortunately, the Gini index and Pearson's χ^2 are only valid discrimination or calibration measures, respectively, when observations are independent. I create simple examples to illustrate the problem of validating predictions with respect to discrimination and calibration under dependent observations.

A. Illustrative Examples

To generate my examples, I introduce a sequence of N independent triples $\{(U_i, V_i, W_i) : i = 1, \ldots, N\}$, where U_i, V_i, and W_i are mutually independent and all 3 variables are uniformly distributed on $[0, 1]$. System A generates the following predictions:

$$A_i \quad = \quad 0.01\left[1 + \mathbf{1}_{\{U_i \leq 0.5, V_i \leq 0.25\}} + \mathbf{1}_{\{U_i > 0.5, V_i > 0.25\}}\right],$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. Thus, A_i is 1% for 1 group of observations (the low-risk group) and 2% for the other group (the high-risk group). I can expect the 2 groups to be equally sized. Systems B and C produce the following forecasts, respectively:

$$B_i \quad = \quad 0.04\left[0.25 + \mathbf{1}_{\{U_i > 0.5\}}\right], \quad \text{and} \quad C_i \quad = \quad \mathbf{1}_{\{V_i \leq 0.03\}}.$$

The forecast B_i is either 1% or 5% for each observation i. In system B the 2 groups are expected to be equally sized. System C generates 1s for 3% of the population and 0s for the rest. Furthermore, I include a common shock, $X \sim U[0,2]$, to induce default dependencies between observations. The shock variable X is stochastically independent from U_i, V_i, and W_i and can be interpreted as the state of the economy. A value of X above 1 (below 1) represents an economic outcome worse (better) than originally expected at the time the forecasts are made. Default indicators are generated by

$$Y_i \quad = \quad \mathbf{1}_{\{W_i \leq B_i X\}}.$$

A positive shock, $X > 1$, increases the default likelihood for each observation, since the state of the economy turned out to be worse than anticipated. A negative shock, $X < 1$, decreases the likelihood of default. Hence, in a good (bad) state of the economy the conditional default probabilities are lower (higher). By straightforward calculations, I have the following conditional expectations: $\mathbb{E}\left[Y_i \mid B_i, X\right] = B_i X$, $\mathbb{E}\left[Y_i \mid A_i, X\right] = 2A_i X$, and $\mathbb{E}\left[Y_i \mid C_i, X\right] = \mathbb{E}\left[Y_i\right] X$. The mean default rate $\mathbb{E}\left[Y_i\right]$ is 3%. By iterated expectations it is straightforward to show that the 1st group of observations in system A has an unconditional default probability of 2% and the other group a probability of 4%. The corresponding unconditional default probabilities in system B are 1% and 5%, since $\mathbb{E}\left[Y_i \mid B_i\right] = B_i$. The naive predictions C_i are completely unable to separate defaults from nondefaults, since Y_i and C_i are independent.

Overall, I create a sequence of quadruples, $\{(Y_i, A_i, B_i, C_i) : i = 1, \ldots, N\}$, which is conditionally IID. Thus, given X, 2 different quadruples are stochastically IID. However, it is worth noting that the predictions A_i and B_i for the same observation i are dependent, even conditional on X, through their common dependence on U_i. The same is true for A_i and C_i through V_i.

B. Discrimination

In system A I expect to observe 66.7% of all defaults in the high-risk group, but system B assigns 83.3% of all defaults to the high-risk group on average. Therefore, I can draw the expected Lorenz (1905) curves as depicted in Figure 1.[3]

[3] The empirical Lorenz (1905) curve for the sample $\{(Y_i, P_i) : i = 1, \ldots, N\}$ is the following 2-dimensional graph:

$$\text{(1)} \qquad \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{P_i \leq p\}}, \frac{1}{N_1} \sum_{i=1}^{N} \mathbf{1}_{\{P_i \leq p\}} Y_i \right),$$

over all $p \in \mathbb{R}$, where Y_i is the default indicator and P_i the probability forecast of observation i, $N_1 = \sum_{i=1}^{N} Y_i$ denotes the number of defaults, and $N_0 = N - N_1$ denotes the number of nondefaults. The Gini (1921) index is given by a ratio of the areas on the Lorenz curve diagram:

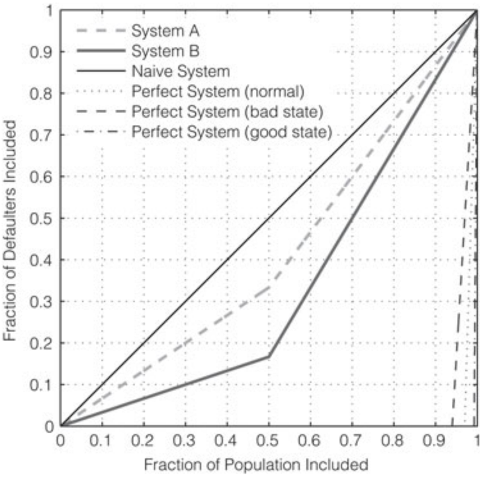
$$2 \frac{1}{N_0 N_1} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[\mathbf{1}_{\{P_i > P_j\}} + \frac{1}{2} \mathbf{1}_{\{P_i = P_j\}} \right] Y_i (1 - Y_j) - 1.$$

The Gini index is 1 in case of perfect forecasts (i.e., $P_i = Y_i$ for all i), and it is 0 for naive forecasts (i.e., $P_i = p$ for all i and for any $p \in \mathbb{R}$).

FIGURE 1

Lorenz Curves to Compare the Discriminatory Power

The Lorenz (1905) curve illustrates the discriminatory power of a prediction system. The Lorenz curve is a graph showing the fraction of defaulters as a percentage of the total population (y-axis) among the x% best rated borrowers (x-axis). The Lorenz curve of a naive system without any informational content with respect to default corresponds to the diagonal line. The perfect system can completely separate defaulters from nondefaulters, and its Lorenz curve goes from the point $(0, 0)$ over $(1 - p, 0)$ to $(1, 1)$, whereas p is the default rate. In a good state of the economy, the default rate p is smaller and the Lorenz curve is closer to the point $(1, 0)$, as compared to a bad state of the economy with a higher default rate p. Systems A and B have 2 equally sized rating classes, 1/3 or 1/6 of all defaulters, respectively, that can be found in the better rated class. The Gini (1921) index is defined as the ratio of 2 areas (i.e., the area between the Lorenz curve and the diagonal and the area between the Lorenz curve of the perfect system and the diagonal). In a normal state of the economy, prediction system A (system B) has a Gini index of 17.2% (34.4%).



In large samples and in a normal state of the economy (i.e., $N \to \infty$, $X = 1$), system B has a Gini (1921) index of 34.4% and system A a Gini index of 17.2%. Unfortunately, the Gini index is not suitable under dependence. If the population's default rate doubles in a bad state of the economy ($X = 2$), then the Gini index in system B increases to 35.5%. In a good state of the economy ($X = 0.5$), the default rates are cut in half, and the Gini index drops to 33.8%. In other words, the Gini index does not only depend on the discriminatory power of the prediction system per se but also on the state of the economy. To filter out the dependence on the state of the economy, I propose to measure the discriminatory power by the area above the Lorenz curve. I have an expected area above the Lorenz curve of 66.7% for system B and 58.3% for system A for all states of the economy. Thus, both systems show discrimination ability, but system B demonstrates higher discriminatory power than system A.

C. Calibration

The common shock X has a simultaneous impact on all observations and distorts Pearson's χ^2. The common shock will almost surely be different from 1, and the mean empirical default frequency will not converge toward the (unconditional) expected default frequency, that is, by the law of large numbers the mean default rate $(1/N)\sum_{i=1}^{N} Y_i$ converges in probability toward $\mathbb{E}[Y_i]X$ and not $\mathbb{E}[Y_i]$. As a consequence, Pearson's χ^2 will diverge with increasing sample sizes, and the calibration hypothesis will be rejected with probability 1. The convergence is only guaranteed when X is almost surely 1 (i.e., when there is no dependence).

I decompose the calibration definition into 2 components that will allow the derivation of test statistics under dependent observations. First, a calibrated system makes unbiased forecasts of the number of defaults. That is, a set of $N = 1{,}000$ observations with a mean prediction of 3% experiences on average 30 defaults. I refer to this component as calibration with respect to the level. Second, a calibrated system differentiates correctly between low and high probability forecasts. That is, a set of observations with a forecast of 5% averages 5 times the default rate compared to a set of observations with a prediction of 1%, as in prediction system B. I refer to this component as calibration with respect to the shape. Level and shape cover 2 different calibration aspects. The level is an absolute and the shape a relative consideration. To clarify the difference, in system A the low-risk class has a true default probability of 2%, the high-risk class 4%. Now, system A assigns 1% to the low-risk class and 2% to the other. The shape is correct (i.e., high-risk observations are indeed twice as likely to default as low-risk observations), but the level is wrong (i.e., the mean prediction is 1.5%, but the true mean is 3%).

I make use of the Lorenz (1905) curve for shape calibration purposes. If the predictions are indeed shape-calibrated, then the expected Lorenz curve can be computed solely from the probability forecasts. In system A the forecasts are either 1% or 2% and the 2 classes are expected to be of equal size. Thus, I expect ⅓ of the defaulters to stem from the 1st class, with the rest from the other class. Conversely, in system B the predictions are 1% and 5%, and I expect only ⅙ of all defaulters to come from the 1st class. Therefore, I compare the realized distributions with the expected distributions by means of the Lorenz curve diagram in order to test the shape calibration hypothesis. With this crucial insight, I propose a new calibration test that I directly derive as a corollary from a discrimination test.

D. Definition of Discrimination and Calibration

To offer formal definitions of discrimination, calibration, and level and shape calibration, I introduce the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and the σ-field $\mathcal{G} \subseteq \mathcal{F}$. Let $Y_i \in \{0, 1\}$ be the indicator of default for observation i and let $P_i \in [0, 1]$ be the probability forecast associated with that observation. I define discriminatory power and calibration as follows:

Definition 1 (Discriminatory Power). I say the probability forecast P_i is $\mathcal{G}$-powerful if there is a strictly increasing function $g : \mathbb{R} \to \mathbb{R}$ so that $P_i = g\left(\mathbb{E}\left[Y_i \middle| \mathcal{G}\right]\right)$ holds true almost surely.

Definition 2 (Calibration). I say the probability forecast P_i is calibrated if $P_i = \mathbb{E}\left[Y_i \middle| P_i\right]$ holds true almost surely.

Now, let me return to the previous examples to illustrate these 2 formal definitions. With $\mathcal{G}$ I denote the σ-algebra generated by prediction system A and system B (i.e., $\mathcal{G} = \sigma\{(A_i, B_i) : i = 1, \ldots, N\}$). Since $\mathbb{E}[Y_i \mid \mathcal{G}] = B_i$, B_i is $\mathcal{G}$-powerful but A_i is not. The marginal information provided by prediction system A beyond the information already provided by system B is irrelevant in forecasting defaults. In other words, system B is more powerful in predicting defaults than system A. However, system B is not $\mathcal{F}$-powerful since $\mathbb{E}\left[Y_i | B_i, Y_i\right] = Y_i$. The perfect system is by definition more powerful than system B. Both predictors Y_i and B_i are calibrated, however, since $\mathbb{E}\left[Y_i | B_i\right] = B_i$ and $\mathbb{E}\left[Y_i | Y_i\right] = Y_i$. On the other hand, the predictors A_i and C_i are not calibrated, since $\mathbb{E}\left[Y_i | A_i\right] = 2A_i$ and $\mathbb{E}\left[Y_i | C_i\right] = \mathbb{E}\left[Y_i\right]$. Definition 2 does not yet imply statistically testable properties. I obtain 2 anchors for a test design by decomposing calibration into level and shape calibration as in Blöchlinger and Leippold (2011):

Definition 3 (Level and Shape Calibration). I say the probability forecast P_i is level calibrated if

$$\mathbb{E}\left[Y_i\right] \quad = \quad \mathbb{E}\left[P_i\right]. \tag{2}$$

I say the probability forecast P_i is shape calibrated if

$$\mathbb{P}\left\{P_i \leq p | Y_i = 1\right\} \quad = \quad \frac{\mathbb{E}\left[\mathbf{1}_{\{P_i \leq p\}} P_i\right]}{\mathbb{E}\left[P_i\right]}, \tag{3}$$

for any $p \in \mathbb{R}$.

Level and shape can be combined to obtain a global test on calibration:

Proposition 1. The probability forecast P_i is calibrated if and only if P_i is both level and shape calibrated.

All proofs can be found in the Appendix. From the previous examples, C_i is level calibrated but not shape calibrated, A_i is shape calibrated but not level calibrated, and B_i and Y_i are both level calibrated and shape calibrated. With respect to an information set, a predictor that is level calibrated, shape calibrated, and powerful is also efficient with respect to the mean squared error (MSE) criterion:

Proposition 2. The conditional expectation $\mathbb{E}\left[Y_i | \mathcal{G}\right]$ is the only predictor that is both $\mathcal{G}$-powerful and calibrated among all $\mathcal{G}$-measurable functions.

Since the conditional expectation is the only MSE efficient predictor, Proposition 2 implies that if I test for discrimination and calibration then I do not need a separate test for MSE efficiency.

III. Assumptions

In this section I state my key assumptions that are needed for the derivation of my validation tests on discrimination and calibration under dependence. A sample of observations, denoted $\{(Y_i, P_i) : i = 1, \ldots, N\}$, is often regarded as a sequence of IID random variables. The variable P_i is the probability forecast, and Y_i the default indicator of observation i. For default risk purposes, these observations are better thought of as conditionally IID. There is a K-dimensional vector $\mathbf{X} = [X_1, \ldots, X_K]^\top$ of latent factors representing macroeconomic or other shocks common to all observations. Given $\mathbf{X}$, observations are assumed to be IID. I am interested in testing whether the set of default predictions $\{P_i : i = 1, \ldots, N\}$ has more discriminatory power than a set of benchmark predictions $\{B_i : i = 1, \ldots, N\}$

and whether the predictions $\{P_i : i = 1, \ldots, N\}$ are calibrated. The testing is supposed to be out-of-sample, that is, the set of default indicators $\{Y_i : i = 1, \ldots, N\}$ was not used for estimating/calibrating the default prediction system.

The K latent factors are assumed to be mutually independent. Each factor is a continuous and positive random variable with a mean of 1. The vector $\mathbf{W}_i$ is the K-dimensional vector of latent factor loadings such that $\mathbf{W}_i^\top \mathbf{1} \leq 1$, and $\min\{\mathbf{W}_i\} \geq 0$, for all $i \in \{1, \ldots, N\}$. The variable $U_i = \mathbf{W}_i^\top \mathbf{X} + 1 - \mathbf{W}_i^\top \mathbf{1}$ is the random effect of observation i inducing dependence between observations. The introduction of the random effects specification is the main difference with respect to the assumptions between my approach and traditional approaches. Three key assumptions are made: Bernoulli mixture, exchangeability, and orthogonality.

Assumption 1 (Bernoulli Mixture). Conditional on the predictor variable, P_i, the benchmark predictor variable, B_i, the vector of factor loadings, $\mathbf{W}_i$, and the vector of factors, $\mathbf{X}$, the default variable, Y_i, is Bernoulli distributed with

$$\mathbb{P}\{Y_i = 1 | P_i, B_i, \mathbf{W}_i, \mathbf{X}\} = \mathbb{P}\{Y_i = 1 | P_i, B_i\} U_i, \tag{4}$$

for all $i \in \{1, \ldots, N\}$.

The multiplicative setup and the factor structure in equation (4) are borrowed from the well-known portfolio model CreditRisk+ (as in Gordy (2000), eq. (1), p. 122). The intuition behind the random effects' specification is that U_i serves to "scale up" or "scale down" the average PD. A high draw of U_i (over 1) increases the PD. A low draw of U_i (under 1) scales down the default probability.

Assumption 2 (Exchangeability). Conditional on $\mathbf{X}$, the observations, $\{(Y_i, P_i, B_i, \mathbf{W}_i) : i = 1, \ldots, N\}$, are IID.

Assumption 2 is adopted from Andrews ((2005), Assump. 1, p. 1555). Under Assumption 2, the random variables $\{(Y_i, P_i, B_i, \mathbf{W}_i) : i = 1, \ldots, N\}$ are exchangeable. That is, $\left\{\left(Y_{\pi(i)}, P_{\pi(i)}, B_{\pi(i)}, \mathbf{W}_{\pi(i)}\right) : i = 1, \ldots, N\right\}$ has the same joint probability distribution as $\{(Y_i, P_i, B_i, \mathbf{W}_i) : i = 1, \ldots, N\}$ for every permutation π of $\{1, \ldots, N\}$ for all $N \geq 2$. The assumption that the data are conditionally IID given $\mathbf{X}$ implies that, unconditionally, the sequence $\{(Y_i, P_i) : i = 1, \ldots, N\}$ is exchangeable (as opposed to IID). As pointed out by Stein (2003), unconditional IID causes problems if traditional inference procedures are used.

Assumption 2 is surprisingly general, as shown in Andrews ((2005), Sect. 7). When observations are sampled randomly from the population (and that is the standard way of doing validation), Assumption 2 is compatible with arbitrary stochastic dependence in the population (e.g., observations in different regions or industry sectors). It is also compatible with any forms of heterogeneity (i.e., nonidentical distributions). Furthermore, Assumption 2 is compatible with common shocks that have different influences on different observations. My last assumption deals with orthogonality.

Assumption 3 (Orthogonality). The sequence of the triples with predictor variables, benchmark predictor variables, and weight variables, $\{(P_i, B_i, \mathbf{W}_i) : i = 1, \ldots, N\}$, and the factors, $\mathbf{X}$, are independent. The sequence of weight variables,

$\{\mathbf{W}_i : i=1,\ldots,N\}$, and the sequence of the pairs with predictors and benchmark predictors, $\{(P_i, B_i) : i=1,\ldots,N\}$, are also stochastically independent.

The independence assumption between the triple $(P_i, B_i, \mathbf{W}_i)$ and the vector of common shocks $\mathbf{X}$ in Assumption 3 comes very naturally. The common shock vector $\mathbf{X}$ must be unanticipated, otherwise the expression "shock" would be misleading. Anticipated effects must be incorporated into the predictions.

In the same vein, I have orthogonality between predictor variables (P_i, B_i) and the vector of weight variables $\mathbf{W}_i$. That is, the random effect U_i and the variables (P_i, B_i) are stochastically independent. On the other hand, the predictor P_i and the benchmark predictor B_i can be dependent even when conditioned on the common factors, which is not excluded by Assumption 3. For instance, dependence can result when the models underlying the 2 predictors are nested. Of course, the default indicators are affected by $\mathbf{X}$, which is also not excluded by Assumption 3. My examples from the previous section fulfill all 3 assumptions.

IV. Discrimination Testing

I test whether the predictor in question, say P_i, can better discriminate than a benchmark predictor, say B_i. If a bank uses the predictor P_i to make default forecasts but its closest competitor uses the more powerful predictor B_i, then the bank runs the risk of adverse selection. Depending on the price sensitivity of borrowers, the potential losses can be huge even for small differences in discriminatory power as analyzed by Blöchlinger and Leippold (2006). In order to obtain a relevant benchmark, I choose the most powerful predictor among a set of well-known forecasting systems for the population in question. For large corporations, natural benchmarks are the Z-score in Altman (1968), Merton's (1974) distance to default, agency issuer ratings, Moody's KMV expected default frequency, or any other logit or discriminant model (see, e.g., Wiginton (1980)).

Discrimination is traditionally measured by the Lorenz (1905) curve or, alternatively, by the closely related receiver operating characteristic (ROC) (see, e.g., Swets (1988)). Formally, the Lorenz curve is a 2-dimensional graph:

$$(5) \qquad \left(\mathbb{P}\left\{P_i \leq p\right\}, \mathbb{P}\left\{P_i \leq p \middle| Y_i = 1\right\}\right),$$

as a function of $p \in \mathbb{R}$. When P_i and Y_i are independent, the predictor P_i has no discriminatory power at all, I have $\mathbb{P}\left\{P_i \leq p \middle| Y_i = 1\right\} = \mathbb{P}\left\{P_i \leq p\right\}$ for any $p \in \mathbb{R}$ (i.e., the defaulters' distribution coincides with the population's distribution). In other words, the proportion of defaulters to nondefaulters is not higher in high-risk classes than in low-risk categories. The Lorenz curve is then the diagonal from $(0,0)$ to $(1,1)$. On the other hand, P_i is a perfect discriminator if there is a p^* such that $\mathbb{P}\left\{P_i > p^* \middle| Y_i = 0\right\} = 0$ and $\mathbb{P}\left\{P_i > p^* \middle| Y_i = 1\right\} = 1$. All defaulters are in the high-risk class and all nondefaulters in the low-risk class.

My summary statistic of the Lorenz (1905) curve is the area above the Lorenz curve θ_P:

$$(6) \qquad \theta_P = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Psi\left(p_1, p\right) d\mathbb{F}_{P|Y=1}(p_1) d\mathbb{F}_P(p),$$

where $\mathbb{F}_P(\cdot)$ denotes the cumulative distribution function (CDF) of P_i (i.e., the population's distribution) and $\mathbb{F}_{P|Y=1}(\cdot)$ is the conditional CDF of P_i given $Y_i = 1$ (i.e., the defaulters' distribution), and $\Psi\left(p_1, p\right) := \mathbf{1}_{\{p_1 > p\}} + \tfrac{1}{2}\mathbf{1}_{\{p_1 = p\}}$. Due to exchangeability, the graph in expression (5) is the same for all i, so that the sample $\{(Y_i, P_i) : i = 1, \ldots, N\}$ can be used to obtain an empirical estimate. Even by the presence of random factors, the empirical Lorenz curve as given in expression (1) is a consistent estimator of the true Lorenz curve in expression (5):

Proposition 3. The empirical CDF of P_i and the empirical CDF of $(P_i|Y_i = 1)$ are consistent estimators for the theoretical CDFs, so that for $N \to \infty$:

$$(7) \qquad \sup_{p\in\mathbb{R}} \left| \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{P_i \leq p\}} - \mathbb{P}\left\{P_j \leq p\right\} \right| \overset{P}{\to} 0, \qquad \text{and}$$

$$\sup_{p\in\mathbb{R}} \left| \frac{1}{N_1} \sum_{i=1}^{N} \mathbf{1}_{\{P_i \leq p\}} Y_i - \mathbb{P}\left\{P_j \leq p | Y_j = 1\right\} \right| \overset{P}{\to} 0,$$

for any $j \in \{1, \ldots, N\}$ and where $N_1 = \sum_{i=1}^{N} Y_i$.

As a consequence of Proposition 3, the empirical Lorenz (1905) curve converges in probability toward the true Lorenz curve even under dependent observations (i.e., for any realization of the state of the economy $\mathbf{X}$).

The area above the Lorenz (1905) curve θ_P is estimated by replacing the expectation with the sample average (i.e., the area above the empirical Lorenz curve $\hat{\theta}_{P,N}$). Analogously, I construct $\hat{\theta}_{B,N}$ for the benchmark:

$$(8) \qquad \hat{\theta}_{P,N} = \frac{1}{N_1 N} \sum_{i=1}^{N} \sum_{j=1}^{N} \Psi\left(P_i, P_j\right) Y_i, \qquad \text{and}$$

$$\hat{\theta}_{B,N} = \frac{1}{N_1 N} \sum_{i=1}^{N} \sum_{j=1}^{N} \Psi\left(B_i, B_j\right) Y_i.$$

When comparing the discrimination ability of 2 predictors (P_i and B_i), I look at the difference $\hat{\theta}_{P,N} - \hat{\theta}_{B,N}$. If the predictor P_i is indeed more powerful than the benchmark predictor B_i, then the difference in the areas above the Lorenz (1905) curve is expected to be greater than 0. Under the null hypothesis that both predictors have the same discriminatory power, $\theta_P = \theta_B$, the standardized difference is asymptotically Gaussian distributed:

Proposition 4 (Discrimination Test). If the predictor P_i and the predictor B_i have the same discriminatory power, $\theta_P = \theta_B$, then for any $s \in \mathbb{R}$,

$$(9) \qquad \lim_{N\to\infty} \mathbb{P}\left\{ \frac{\hat{\theta}_{P,N} - \hat{\theta}_{B,N}}{\sqrt{\hat{V}\left[\hat{\theta}_{P,N} - \hat{\theta}_{P,N} \middle| N_1\right]}} \leq s \middle| \mathbf{X} \right\} = \Phi\left(s\right),$$

where $\Phi(s) = \int_{-\infty}^{s} (1/\sqrt{2\pi}) \exp\left(-\tfrac{1}{2}\xi^2\right) d\xi$ is the CDF of a standard Gaussian variable, $\hat{\theta}_{P,N}$, $\hat{\theta}_{B,N}$ are given in expression (8), and $\hat{V}\left[\cdot| N_1\right]$ denotes the empirical variance estimator given N_1 defaulters. The estimation of the standard error is found in the Appendix.

It is noteworthy that the convergence in Proposition 4 is guaranteed for any realization of the state of the economy $\mathbf{X}$.

V. Calibration Testing

In this section I derive statistical tests to accept or reject the calibration hypothesis. The level calibration hypothesis is stated in equation (2), the shape calibration hypothesis in equation (3). I provide tests for level and shape as well as a combined statistic to have a summary test for both hypotheses. If a predictor P_i passes the combined test, then the calibration hypothesis in Definition 2 cannot be rejected. Unlike well-known calibration tests that are only valid under IID observations, I work under the more general assumption of conditional IID as stated in Section III.

If the sample $\{(Y_i, P_i) : i = 1, \ldots, N\}$ is unconditionally IID, then Spiegelhalter (1986) provides asymptotic confidence intervals of the well-known Brier (1950) score S_N:

$$S_N \quad := \quad \frac{1}{N}\sum_{i=1}^{N}(Y_i - P_i)^2 .$$

Mean and variance of S_N can easily be computed:

$$\begin{aligned}\mathbb{E}[S_N] &= \frac{1}{N}\sum_{i=1}^{N} P_i(1-P_i),\\ \mathbb{V}[S_N] &= \frac{1}{N^2}\sum_{i=1}^{N} P_i(1-P_i)(1-2P_i)^2 .\end{aligned}$$

By the unconditional IID assumption, the standardized Brier score is asymptotically standard Gaussian distributed. Hence, I have for any $s \in \mathbb{R}$,

$$\lim_{N\to\infty}\mathbb{P}\left\{\frac{S_N - \mathbb{E}[S_N]}{\sqrt{\mathbb{V}[S_N]}} \le s\right\} \quad = \quad \Phi(s) . \tag{10}$$

Alternatively, if the sample $\{(Y_i, P_i) : i = 1, \ldots, N\}$ is unconditionally IID, I simply resort to the law of large numbers, and I have for any $s \in \mathbb{R}$,

$$\lim_{N\to\infty}\mathbb{P}\left\{\frac{\sqrt{N}\left(\overline{Y}_N - \overline{P}_N\right)}{\sqrt{\overline{P}_N\left(1-\overline{P}_N\right)}} \le s\right\} \quad = \quad \Phi(s) , \tag{11}$$

where $\overline{Y}_N := (1/N)\sum_{i=1}^{N} Y_i$ denotes the average default frequency and $\overline{P}_N := (1/N)\sum_{i=1}^{N} P_i$ denotes the average default forecast of sample size N. The statistic in equation (11) is a special case of the Pearson (1900) statistic and of the Hosmer and Lemeshow (1989) statistic when all observations are summarized into a single group.

In general, Pearson's goodness-of-fit test is computed using the covariate patterns in the data as groups. The Hosmer-Lemeshow test identifies groups via the quantiles of the response variable. However, the expected number of defaults in the deciles with low-risk credit ratings is typically too small to warrant the use of the asymptotic distribution. The expected number of defaults per group must be 5 or greater. Hence, with a group's mean default forecast of, say, 0.02%, I needed 25,000 observations for only this particular class. I want to point out that for the following calibration tests, no grouping or bucketing will be required.

A. Level Calibration Testing

Under unconditional IID, the mean default frequency $\overline{Y}_N$ converges in probability toward a degenerate random variable. However, under conditional IID, the asymptotic variable is nondegenerate:

Proposition 5. If P_i is level calibrated, then I have for $N \to \infty$,

$$(12)\quad \overline{Y}_N \quad \overset{P}{\to} \quad \mathbb{P}\{Y_i = 1 | \mathbf{X}\} \quad = \quad \mathbb{E}[P_i]\left(1 - \mathbb{E}[\mathbf{W}_i]^\top \mathbf{1} + \mathbb{E}[\mathbf{W}_i]^\top \mathbf{X}\right),$$

for any $i \in \{1, \ldots, N\}$.

Since the factors are assumed to be continuous and stochastically independent, the CDF $\mathrm{F}(\cdot)$ of the variable $\mathbb{P}\{Y_i = 1 | \mathbf{X}\}$ in expression (12) is also continuous. With the help of the quantile transformation theorem, $\mathrm{F}\left(\overline{Y}_N\right)$ is asymptotically uniformly distributed on $[0, 1]$.[4] A further application of the transformation theorem leads me to

$$(13)\qquad \lim_{N\to\infty} \mathbb{P}\left\{\Phi^{-1}\left(\mathrm{F}\left(\overline{Y}_N\right)\right) \le s\right\} \quad = \quad \Phi(s), \qquad \text{for any} \quad s \in \mathbb{R},$$

where $\Phi^{-1}(\cdot)$ is the quantile function of the standard Gaussian CDF $\Phi(\cdot)$. Therefore, my standardized level statistic is given with expression (13).

If I make some sensible distributional assumptions, the asymptotic distribution of expression (12) is easily derivable even for observations spread over multiple periods. By the law of large numbers, the sample mean prediction $\overline{P}_N$ converges (in probability) toward the population mean $\bar{\pi} := \mathbb{E}[P_i]$. If the scaled factors, $\{\bar{\pi} X_k : k = 1, \ldots, K\}$, are assumed to be IID beta variables with corresponding factor weights, $\mathbb{E}[\mathbf{W}_i] = (1/K)[\omega, \ldots, \omega]^\top$, then I get

$$(14)\qquad \frac{\overline{Y}_N - \overline{P}_N(1-\omega)}{\overline{P}_N \omega} \quad \overset{P}{\to} \quad \frac{1}{K}\sum_{k=1}^{K} X_k,$$

$$\text{with} \qquad \mathbb{V}[X_k] \quad = \quad \sigma^2, \qquad \mathbb{E}[X_k] \quad = \quad 1, \qquad \text{for each } k.$$

The distribution of $(1/K)\sum_{k=1}^{K} X_k$ can be obtained with numerical methods (e.g., Monte Carlo simulations) or analytically (e.g., in case of $K = 1$).[5] If the

[4] The quantile transformation theorem can be found in Karr (1993).

[5] By Slutsky's theorem, if $N \to \infty$ then $\overline{P}_N X_k$ has the same asymptotic distribution as $\bar{\pi} X_k$ for each k, since $\overline{P}_N$ converges in probability toward the constant $\bar{\pi}$. Therefore, I simulate the sum of K IID beta variables with mean $\overline{P}_N$ and standard deviation $\sigma \overline{P}_N$ in order to approximate the asymptotic distribution in expression (14).

sample size N is too small to warrant the use of the asymptotic distribution, I can resort to small sample equivalents, for example, I can approximate the Bernoulli-beta mixture with a Poisson-gamma mixture as in Wilde (1997) in order to get a closed-form solution for the distribution in expression (14).

B. Shape Calibration Testing

I propose to compare the expected Lorenz (1905) curve with the empirical Lorenz curve. With the addition of observations, the empirical Lorenz curve converges to the true Lorenz curve for any state of the economy (see Proposition 3). Thus, the area above the empirical Lorenz curve converges to the true area:

Corollary 1 (Shape Calibration Test). I have for any $s \in \mathbb{R}$,

$$(15) \qquad \lim_{N\to\infty} \mathbb{P}\left\{ \frac{\hat{\theta}_{P,N} - \theta_P}{\sqrt{\mathbb{V}\left[\hat{\theta}_{P,N}\middle| N_1\right]}} \leq s \middle| \mathbf{X} \right\} \quad = \quad \Phi\left(s\right),$$

where $\hat{\theta}_{P,N}$ and θ_P are given in expressions (8) and (6). The computation of the standard error is found in the Appendix.

The defaulters' CDF, $\mathbb{F}_{P|Y=1}(\cdot)$, to compute θ_P in equation (6) is obtained from the distribution of P_i on the right-hand side of equation (3) if shape calibrated. Corollary 1 provides an asymptotic statistic to test the null hypothesis that P_i has a true area above the Lorenz (1905) curve of θ_P. Corollary 1 implies that there is a relation between the distribution of P_i and θ_P when P_i is calibrated. In other words, under the null hypothesis of a shape-calibrated predictor P_i, the distribution of P_i determines the true Lorenz curve. For instance, if P_i is Dirac distributed, then the predictor P_i has no discriminatory power at all and the area above the true Lorenz curve θ_P is 0.5. If P_i is shape calibrated and Bernoulli distributed, then P_i must be the perfect discriminator, $Y_i = P_i$ for all i.

C. Combined Calibration Testing: Level and Shape

I now combine the level and the shape statistic into a summary statistic. From expressions (13) and (15) I have 2 standardized and asymptotically Gaussian distributed random variables:

$$(16) \qquad T_{\pi,N} \quad := \quad \Phi^{-1}\left(\mathrm{F}\left(\overline{Y}_N\right)\right) \quad \text{and} \quad T_{\theta,N} \quad := \quad \frac{\hat{\theta}_{P,N} - \theta_P}{\sqrt{\mathbb{V}\left[\hat{\theta}_{P,N}\middle| N_1\right]}}.$$

The level statistic, $T_{\pi,N}$, and the shape statistic in the form of the standardized area above the Lorenz (1905) curve, $T_{\theta,N}$, are asymptotically independent, which allows for the straightforward combination into a global statistic:

Proposition 6 (Combined Calibration Test). When testing against the alternative hypothesis that level, $T_{\pi,N}$, and standardized area above the Lorenz (1905) curve, $T_{\theta,N}$, are different from 0, I obtain an asymptotically χ^2 distributed combined test statistic with 2 degrees of freedom: $Q_N := T^2_{\pi,N} + T^2_{\theta,N} \xrightarrow{d} \chi^2_{\langle 2 \rangle}$.

Proposition 6 provides my χ^2 distributed test statistic Q_N with 2 degrees of freedom for the calibration hypothesis as stated in Definition 2.

VI. Dependence between Observations

There is a huge amount of literature on dependent credit events (see, e.g., Li (2000), Zhou (2001), Chen and Sopranzetti (2003), Crouhy, Jarrow, and Turnbull (2008), and Blöchlinger (2011), to list just a few). The kind of dependence between observations must be understood to determine the degree of correlation and to aggregate my new test statistics over multiple periods. For instance, the default of the Italian dairy and food corporation Parmalat in Dec. 2003 certainly had less influence on the failure of Washington Mutual on Sept. 26, 2008, than did the default of Lehman Brothers on Sept. 23, 2008. Parmalat operated in a different industry sector. Furthermore, Parmalat defaulted in a different year and a different time sector, respectively.

A. Serial Correlation and Noncyclical Industries

Assume observation i stems from an underlying borrower that operates in a noncyclical industry. Observation i is nonsensitive toward common shocks in economy $\mathbf{X}$. Consequently, the realization of the random effect U_i will be 1 for any realizations of $\mathbf{X}$. Conversely, the realization of $\mathbf{W}_j^\top \mathbf{1}$ for borrower j in a strongly cyclical industry will be higher than average and therefore above $\mathbb{E}\left[\mathbf{W}_j^\top\right]\mathbf{1}$.

If the predictor P_i is calibrated, then I have an unbiased estimate of Y_i, and the forecast error, $Y_i - P_i$, has a mean of 0. Moreover, these forecast errors must be orthogonal over time, since the information with respect to past predictions and defaults must be taken into account when making the following period's probability forecasts. That is, if observation j stems from a later time period than observation i (i.e., at the time the prediction P_j is made the realization of Y_i is known), then the realization of the scalar product $\mathbf{W}_i^\top \mathbf{W}_j$ must be 0. In such an instance, the observations i and j are from different time sectors. Technically speaking, the stochastic time series of forecasting errors form a martingale difference sequence. This derivation leads to the following proposition:

Proposition 7. If observation i or observation j stems from a noncyclical sector or if observation i and observation j stem from different time sectors, then the realization of $\mathbf{W}_i^\top \mathbf{W}_j$ must be 0.

If P_i is the 1-year default forecast at the beginning of the year and Y_i is the default indicator at the end of the year, then Proposition 7 implies that when I have T years of data I must have at least T (independent) factors. If defaults in any given year are driven by S industry factors, then I must have $T \times S$ factors.

B. Cross Correlation

The Basel Committee on Banking Supervision (2001) measures the degree of cross correlation with the so-called asset correlation ϱ in a 1-factor asset value model. If I impose a 1-factor assumption, there is a relation to transform ϱ into the volatility parameter σ of the latent factor X_1 as required in my framework. Assuming 2 borrowers with a default probability of $\bar{\pi}$ each and a constant factor weight of ω, these 2 borrowers have the same cross correlation in both frameworks when the following equality holds true (see Gordy (2000), Prop. 1, p. 133):

$$\omega^2\sigma^2\bar{\pi}^2 \quad = \quad \Phi_2(\Phi^{-1}(\bar{\pi}), \Phi^{-1}(\bar{\pi}); \varrho) - \bar{\pi}^2, \tag{17}$$

where $\Phi_2(x, y; \varrho)$ denotes the bivariate standard Gaussian CDF at point (x, y) with correlation coefficient ϱ. To obtain a historical estimate of ϱ, I insert the mean default rate, $\bar{\pi}$, and the standard deviation of yearly default rates, $\omega\sigma\bar{\pi}$, into equation (17) and solve for ϱ.

According to historical experience, the standard deviation of the number of defaults observed year on year among corporate borrowers is typically of the same order as the average annual number of defaults (see Wilde (1997), p. 44). For the population of Standard & Poor's (S&P) rated corporations, I observe an average yearly default rate of 1.54% and a standard deviation of 1.05%. According to equation (17), this ratio of 0.68 (= 1.54%/1.05%) translates into a historical estimate for the asset correlation ϱ of about 6.5%. A higher standard deviation of 1.54% would mean a ϱ of about 12%. The lower (higher) the correlation value ϱ, the more conservative (progressive) are the validation results; that is, a correct calibration hypothesis is rejected too often (too rarely). Therefore, I consider $\varrho = 0.06$ a sensible, moderate parameterization for a population of large corporations. To induce an asset correlation of $\varrho = 0.06$ with $\omega = 0.8$ and $\bar{\pi} = 0.02$, the volatility σ of the latent factor in equation (17) must be 0.7889.

To investigate the effects of different default dependencies, I work in the following with a set of asset correlation values, $\varrho \in \{0, 0.05, 0.1, 0.2\}$, for observations in the same time sector. A correlation of $\varrho = 0$ is equivalent to an unconditional IID setup. The Basel Committee on Banking Supervision (2005) argues that zero correlation is too conservative an assumption for the purpose of validation. An asset correlation of 5% represents a moderate correlation regime that is slightly lower than the historical point estimate. An asset correlation regime of 10% is a bit stronger than observed in the past. The highest correlation coefficient was in line with the calculation of regulatory capital under the Basel II framework and translates into rather strong default dependence.

C. Multiperiod Summary Statistics

In light of Proposition 7 and the previous elucidations, the following statement is true for a sample from a single period of observations but not so when the sample consists of multiple periods of observations:

> 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.
>
> Basel Committee on Banking Supervision (2005)

It is true that cross-correlated errors must be accounted for, but the autocorrelation of forecasting errors in a calibrated and powerful prediction system must be 0. Given T time sectors, I have T uncorrelated clusters, rendering the detection of miscalibrations efficient over the course of time (i.e., when $T \to \infty$). Consequently, sequences of 1-period statistics can be easily aggregated into multiperiod summary statistics.

VII. Simulation

In a simulation exercise I demonstrate the functioning and performance of my discrimination and calibration tests. Within a simulation environment, I know the true data generating process and therefore bias and statistical power can be measured under various (asset) correlation regimes. The statistical power measures a test's ability to detect an alternative hypothesis. The power of a statistical test is 1 minus its type II error rate. I expect the statistical power to decrease when the correlation between observations increases. A test is said to be unbiased or correctly sized when the probability of rejecting a correct null hypothesis corresponds to its nominal significance level.

It is highly desirable to work with unbiased statistics. In statistical testing I control the type I error rate, that is, if I assume a nominal significance level of 5%, I want the true significance level to be 5% without negative or positive bias. Among all unbiased tests, I want to choose the one with the greatest power in detecting a wrong null hypothesis. Pearson's χ^2 statistic and the Brier score are derived under stochastic independence, so I expect these tests to be biased when the asset correlation is nonzero (i.e., I will reject a correct null hypothesis too often). Tests can also be biased when the number of observations is too small (i.e., when the asymptotic distributions are not yet valid). In this case, I could resort to small sample distributions (e.g., by way of Monte Carlo simulations) to obtain unbiased tests. I first present the simulation setup and then discuss the results.

A. Simulation Setup

I have T time periods and N observations per time period. The sample size is increased by enlarging the number of time periods T. I simulate with a set of time periods, $T \in \{1, 5, 10, 20, 50\}$, N = 1,000, and I investigate different asset correlation values, $\varrho \in \{0, 0.05, 0.1, 0.2\}$, under 100,000 Monte Carlo paths. The dependent defaults are generated through conditionally independent Bernoulli variables:

$$Y_i \quad \sim \quad B\left(1, P_i U_i\right), \tag{18}$$

where $P_i = ⅓ \left(R_{1,i} + R_{2,i} + R_{3,i}\right)$ is the default forecast and $U_i = (1 - \omega) + \mathbf{W}_i^\top \mathbf{X}$ the random effect with mean 1 and $\omega = 0.8$. The variables $R_{\ell,i}$ are IID binary variables, and the realization is either $\pi_u = 3.5\%$ or $\pi_d = 0.5\%$ for all ℓ and i. Both states, π_u and π_d, are equally likely, so that the expected value of P_i is

$\bar{\pi} = ½ (\pi_u + \pi_d) = 2\%$. Furthermore, I have T IID common factors $\mathbf{X} = [X_1, \ldots, X_T]$ for each time period with

$$\text{(19)} \quad \bar{\pi} X_t \quad \sim \quad \beta \left(\bar{\pi} \left[\frac{1-\bar{\pi}}{\bar{\pi}\sigma^2} - 1 \right], (1-\bar{\pi}) \left[\frac{1-\bar{\pi}}{\bar{\pi}\sigma^2} - 1 \right] \right), \quad \text{where } \sigma^2 = \mathbb{V}[X_t].$$

The variance σ^2 is chosen to obtain the desired degree of asset correlation ϱ as derived in equation (17). As shown in Section VI, when 2 observations are in different time periods, the scalar product of factor weights must be 0. To fulfill this restriction, $T-1$ of the T elements of the vector of factor weights $\mathbf{W}_i$ are 0, the nonzero element is chosen at random and is set equal to ω.

By construction, the predictor P_i is shape and level calibrated as defined in equations (2) and (3). That is, the mean of P_i and the mean of Y_i are equal, and both the area above the Lorenz (1905) curve defined on the left-hand side of equation (3) and the area above the curve defined on the right-hand side of equation (3) yield a value of 61.72%. I introduce 4 further predictor variables:

$$\begin{aligned} A_i &= ⅓ (R_{1,i} + R_{2,i} + \bar{\pi}), \\ B_i &= q (R_{1,i} + R_{2,i}) + \bar{\pi}(1-2q), \\ C_i &= q (R_{2,i} + R_{3,i} + \bar{\pi}), \qquad \text{and} \\ D_i &= \frac{\bar{p}}{\bar{\pi}} q (R_{1,i} + R_{3,i}) + \bar{p}(1-2q), \end{aligned}$$

where $\bar{p} = 1.5\%$ and $q = ¼$. The predictors A_i and B_i are level calibrated, and C_i and D_i are not. The mean of C_i and D_i, $\mathbb{E}[C_i] = \mathbb{E}[D_i] = \bar{p} = 1.5\%$, is lower than the true mean default probability of $\bar{\pi} = 2\%$. The true area above the Lorenz curve for all predictors A_i, B_i, C_i, and D_i is 59.38% (from the left-hand side of equation (3)). That is, all 4 predictors, A_i, B_i, C_i, and D_i, have the same discriminatory power, but the predictor P_i is more powerful. The true area above the Lorenz curve is 61.72% for P_i versus 59.38% for A_i, B_i, C_i, and D_i. The forecasts A_i and C_i are shape calibrated, but B_i and D_i are not. For B_i and D_i, the area above the curve as defined on the right-hand side of equation (3) is only 57.03% (but again, the true area above the Lorenz curve is 59.38%). As can be seen from these examples, discrimination and calibration are not offsetting (i.e., when one improves it does not need to be at the expense of the other).

B. Simulation Results

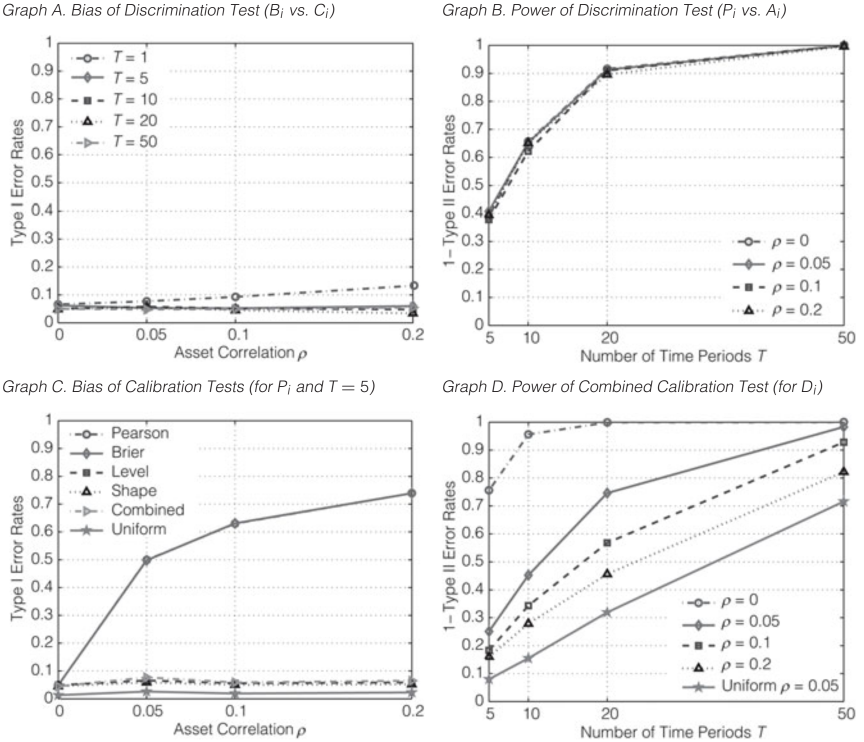
In Figure 2 the simulation results are summarized.[6] In Graph A of Figure 2 I have the simulated type I error rate when testing the discrimination ability between B_i and C_i. Under the null hypothesis, both predictors have the same discrimination ability, and since this is indeed the case, I expect to reject this null hypothesis in 5% of all cases, which corresponds to the nominal significance

[6] The standard error of the Monte Carlo analysis is $\sqrt{p(1-p)/n}$, where n denotes the number of Monte Carlo draws and p the true error rate. The standard error of the type I error analysis is therefore 0.07% when the nominal significance level is $p = 5\%$ and $n = 100{,}000$. For the simulated type II error rate, I have a maximal standard error of 0.16% when $p = 50\%$.

FIGURE 2

Bias and Statistical Power of Discrimination and Calibration Tests in Simulation Exercise

Unless the sample size is small, $T = 1$ period, the discrimination test is correctly sized (Graph A of Figure 2). The discrimination test's power converges toward 1 with increasing sample sizes (Graph B). My multiperiod calibration tests on level and shape as well as the combined test are correctly sized even under dependent observations. On the other hand, Pearson's χ^2 and the Brier score are only unbiased under independence or $\varrho = 0$, respectively (Graph C). The power of my calibration test decreases with stronger default dependence but converges nonetheless toward 1 over time (Graph D). Applying uniform correlation among all observations, even for observations in different time periods, when in fact they are uncorrelated, yields biased and less powerful test statistics.



Graph A. Bias of Discrimination Test (B_i vs. C_i)

Graph B. Power of Discrimination Test (P_i vs. A_i)

Graph C. Bias of Calibration Tests (for P_i and $T = 5$)

Graph D. Power of Combined Calibration Test (for D_i)

level. Unless the sample size is very small ($T = 1$), the discrimination test is unbiased. With only 1,000 observations and only 20 defaults on average ($T = 1$), the discrimination test is particularly biased under an asset correlation of 20%, and the true type I error is above 10% when it should be 5%. The statistical power of the discrimination test converges quite quickly toward 1, as can be seen from Graph B. The type II error increases only marginally when ϱ increases. When the asset correlation ϱ is high, I have scenarios in which the number of defaults is very low, which causes the type II error to increase slightly.

In Graph C of Figure 2, one sees the type I error for the calibration tests. My 2 calibration tests on level and shape, as well as the combined test, are quite correctly sized (i.e., they are virtually unbiased). The marginal bias arises from the level statistic when the sample size is rather small. Pearson's χ^2 and the Brier score, however, are hugely biased when correlation comes into play. When $\varrho = 0.2$, the true type I errors are around 75% when they should be 5%. This distortion makes the Pearson statistic and the Brier score unsuitable for calibration testing under dependent observations. With my combined calibration test, the power to detect wrongly calibrated predictors decreases when defaults are dependent (Graph D). That is, it takes longer to tell a wrongly calibrated level of 1.5% from a true level of 2%. With 20 time periods, the type II error is close to 0 when observations are independent but still above 50% when defaults are strongly correlated ($\varrho = 20\%$). Nevertheless, since the autocorrelation must be 0 for observations in different time periods, as explained in Section VI, the statistical power of the combined calibration test goes to 1 over time for any asset correlation regime. Distinguishing between observations in different time sectors is the key; my multiperiod approach accounts for orthogonality between observations in different periods. If I were to apply a uniform asset correlation ϱ to all observations, in fact implying that all the observations are from the same time sector as in a 1-period setup, this would give rise to biased and less powerful test statistics, as depicted in Figure 2.

My newly developed tests are correctly sized for typical sample sizes in credit portfolio management. The stronger the cross correlation between observations, the weaker the statistical power, but all of my tests are statistically consistent over time. As analyzed by the Basel Committee on Banking Supervision (2005), existing tests were either too conservative (i.e., they reject well-performing systems too often like Pearson's χ^2 statistic and the Brier score) or too progressive (i.e., they accept inferior systems too often when uniform correlation is assumed). The final decision as to whether a prediction system was acceptable from a regulatory viewpoint had to be based on the qualitative assessment of experienced banking supervisors. My multiperiod approach now offers a valuable quantitative alternative: My multiperiod statistics are unbiased and more powerful than existing approaches.

VIII. Empirical Analysis

The goal of the empirical exercise is to demonstrate the previously derived theoretical validation framework with a substantive credit application. I assess the performance of 2 default forecasting methods: S&P's issuer ratings and the distance-to-default model. The distance-to-default model applies the framework of Merton (1974), in which the equity of a firm is a call option on the underlying asset value of the firm with a strike price equal to the face amount of the firm's debt.

Default probabilities are typically estimated with a sample that includes the recent history of observations. For example, the default rate is estimated for each rating class, and this estimate is applied to make today's forecasts. To validate today's forecasts, the whole sample is usually split into an estimation and a validation sample. All observations before a certain threshold date are in the estimation sample, and later observations are in the validation sample. The default rate per rating class is again estimated in-sample (i.e., in the estimation sample) and then applied out-of-sample (i.e., in the validation sample). My new test statistics allow the out-of-sample assessment of whether the forecasts in the form of default probabilities are calibrated and if the predictions are at least as powerful in discriminating between defaulters and nondefaulters as benchmark forecasts. If so, the predictor passes the validation test. If not, the prediction model may be overfitting in the estimation sample and therefore produces inferior forecasts in the validation sample. A more parsimonious model may be needed.

A. The Merton Distance-to-Default Model

The Merton (1974) distance-to-default model produces a default forecast for borrowers with market-traded equity at any given point in time. To compute the distance to default, I basically subtract the logarithmic face value of debt from the firm's expected logarithmic asset value, and then I divide this difference by the asset volatility of the firm:

$$(20) \qquad D_i \quad = \quad \frac{\log V_i + \mu_i - \log F_i - \frac{1}{2}\sigma_i^2}{\sigma_i},$$

where V_i is the market value of the ith firm's total assets, F_i is the face amount of the firm's debt, σ_i is the volatility of the firm's total assets, and μ_i is the corresponding mean asset return. To implement the distance-to-default model, I closely follow the procedure as detailed in Bharath and Shumway (2008). The distance to default is transformed into a default probability by the link function $G(\cdot)$ from $\mathbb{R}$ onto $(0, 1)$:

$$(21) \qquad P_i \quad = \quad G\left(-D_i\right) \quad = \quad G\left(-\frac{\log V_i + \mu_i - \log F_i - \frac{1}{2}\sigma_i^2}{\sigma_i}\right).$$

In fact, if the assumptions of Merton hold true, then the link function $G(\cdot)$ must be the standard Gaussian distribution function $\Phi(\cdot)$, and in this case the probability forecast P_i is calibrated. Bharath and Shumway use the Gaussian distribution to convert distances to default into probability forecasts. I will use an empirical estimate for the link function $G(\cdot)$, which I describe later.

B. Data

My data set includes all S&P rated nonfinancial corporations between Jan. 1981 and Dec. 2010 for which equity and debt data from Bloomberg are available. I focus on 1-year default probabilities. At the end of each month I observe the S&P rating and the distance to default; 1 year later, each firm has either defaulted, not defaulted, or the S&P rating was withdrawn during the course of the year. I count as default a firm that was downgraded to D by S&P at some point during that year. Ratings that were withdrawn are excluded from the sample. I use the data before Dec. 2000 as the sample to estimate the PDs per S&P rating class and to estimate the mapping from distances to default into default probabilities. The estimation sample consists of 70,590 observations and 847 defaults. The 10 years from 2001 to 2010 represent my validation sample for which my newly developed validation tools can be applied. For the validation sample, I retrieve S&P ratings and Merton's (1974) PDs on Dec. 31 in years 2000–2009 and the default indicators on Dec. 31 in years 2001–2010. Summary statistics with respect to S&P ratings and Merton PDs can be found in Table 1.

TABLE 1

Summary Statistics of S&P Ratings and Merton's Distances to Default

In Table 1, from 2001 to 2010, there are 14,654 observations and 228 defaults. According to the estimated PD per S&P rating class, obtained by a sample from 1981 to 2000, I expect an average default frequency of 2.12%, but the realized frequency is 1.56%. Under independence such a difference between expectation and realization would be a 4.7-sigma event; under the more reasonable assumption of dependence, however ($\varrho = 0.06$ for observations in the same time sector), it is a 1.4-sigma event and therefore not significant. The Merton (1974) PD is within the 1-sigma range even under independence.

Panel A. S&P Ratings						*Panel B. Merton's Distances to Default*					
Rating	No. of Obs.	Freq.	Def.	PD	Emp. PD	Distance	No. of Obs.	Freq.	Def.	PD	Emp. PD
AAA	102	0.70%	0	0.001%	0.00%	$[7.25, \infty)$	4,371	29.83%	1	0.020%	0.02%
AA+	42	0.29%	0	0.010%	0.00%	$[6.83, 7.25)$	514	3.51%	0	0.022%	0.00%
AA	201	1.37%	0	0.020%	0.00%	$[6.44, 6.83)$	514	3.51%	0	0.029%	0.00%
AA–	364	2.48%	0	0.030%	0.00%	$[6.07, 6.44)$	514	3.51%	0	0.036%	0.00%
A+	591	4.03%	0	0.040%	0.00%	$[5.71, 6.07)$	515	3.51%	0	0.045%	0.00%
A	1,031	7.04%	0	0.050%	0.00%	$[5.39, 5.71)$	514	3.51%	0	0.056%	0.00%
A–	1,196	8.16%	0	0.060%	0.00%	$[5.09, 5.39)$	514	3.51%	0	0.067%	0.00%
BBB+	1,474	10.06%	2	0.070%	0.14%	$[4.77, 5.09)$	514	3.51%	0	0.081%	0.00%
BBB	1,848	12.61%	2	0.096%	0.11%	$[4.45, 4.77)$	514	3.51%	3	0.097%	0.58%
BBB–	1,364	9.31%	3	0.201%	0.22%	$[4.14, 4.45)$	514	3.51%	3	0.120%	0.58%
BB+	982	6.70%	2	0.356%	0.20%	$[3.82, 4.14)$	515	3.51%	1	0.150%	0.19%
BB	1,236	8.43%	6	0.977%	0.49%	$[3.46, 3.82)$	514	3.51%	2	0.194%	0.39%
BB–	1,503	10.26%	11	2.333%	0.73%	$[3.11, 3.46)$	514	3.51%	1	0.257%	0.19%
B+	1,240	8.46%	29	4.318%	2.34%	$[2.74, 3.11)$	514	3.51%	4	0.351%	0.78%
B	746	5.09%	35	7.889%	4.69%	$[2.34, 2.74)$	514	3.51%	6	0.504%	1.17%
B–	443	3.02%	38	13.789%	8.58%	$[1.89, 2.34)$	514	3.51%	4	0.766%	0.78%
CCC+	146	1.00%	38	21.443%	26.03%	$[1.39, 1.89)$	514	3.51%	10	1.302%	1.95%
CCC	89	0.61%	26	28.560%	29.21%	$[0.80, 1.39)$	515	3.51%	8	2.395%	1.55%
CCC–	22	0.15%	14	38.036%	63.64%	$[0.12, 0.80)$	514	3.51%	26	4.785%	5.06%
CC	34	0.23%	22	42.424%	64.71%	$[-0.84, 0.12)$	514	3.51%	35	10.215%	6.81%
C	0	0.00%	0	—	—	$(-\infty, -0.84)$	514	3.51%	124	23.535%	24.12%
Sum/avg.	14,654	100.00%	228	2.12%	1.56%	Sum/avg.	14,654	100.00%	228	1.58%	1.56%

C. From Ordinal Risk Measures to Default Probabilities

I use the sample from 1981 to 2000 to obtain an empirical estimate of the link function $G(\cdot)$ in expression (21) and to estimate the default rate for each S&P rating class. The mapping from S&P rating classes into the 1-year default probabilities is found in Graph A of Figure 3; the mapping for the distance to default is depicted in Graph B. As one can see in Graph B, the Gaussian distribution fails to fit the empirical default frequencies. However, it must be noted that the transformation from distances to default into PDs is an important input for calibration testing, but it is not required for measuring the discriminatory power. As long as the transformation is strictly monotone, the distance to default D_i in equation (20) and the forecast P_i in expression (21) have the same discriminatory power.

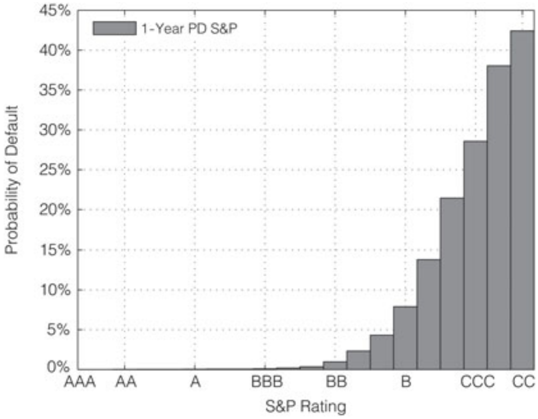
Given the estimation of the credit curves in Figure 3, I now plot the development of S&P PDs and Merton (1974) PDs over time for any firm. In its promotional material, Moody's KMV points to the example of Enron to illustrate how the Merton PD is superior to that of agency ratings, as depicted in Graph C of Figure 4. When it became public knowledge that Enron had serious accounting problems, Enron's equity price began to fall and its distance to default steadily decreased. S&P waited several months to downgrade Enron's creditworthiness. Obviously, using market information to infer default probabilities allows the Merton model to reflect information faster than traditional agency ratings. On the

FIGURE 3

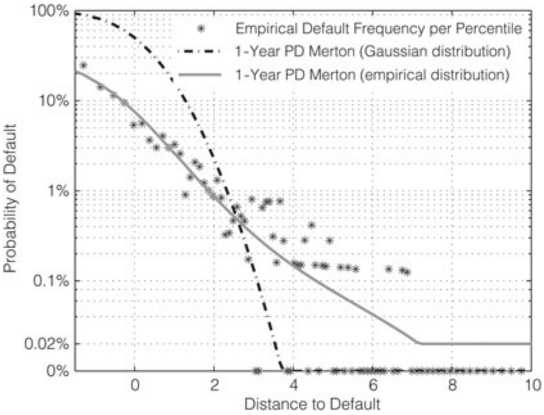
Default Probability Curves

The 1-year default probabilities are estimated with a sample period from Jan. 1981 to Dec. 1999 with 70,590 observations and 847 defaults. The PD per S&P rating class is depicted in Graph A of Figure 3, with PD as a function of the distance to default in Graph B. A Gaussian link function between distance to default and PD would yield an inferior fit. Several borrowers with a distance to default of 4 or higher defaulted on their debt even though the Gaussian link function indicates a PD of less than 0.01%. Following the approach of Moody's KMV, I estimate an empirical link based on exponential functions and allow the PD not to fall below 0.02%.

Graph A. S&P PD Curve



Graph B. Merton PD Curve

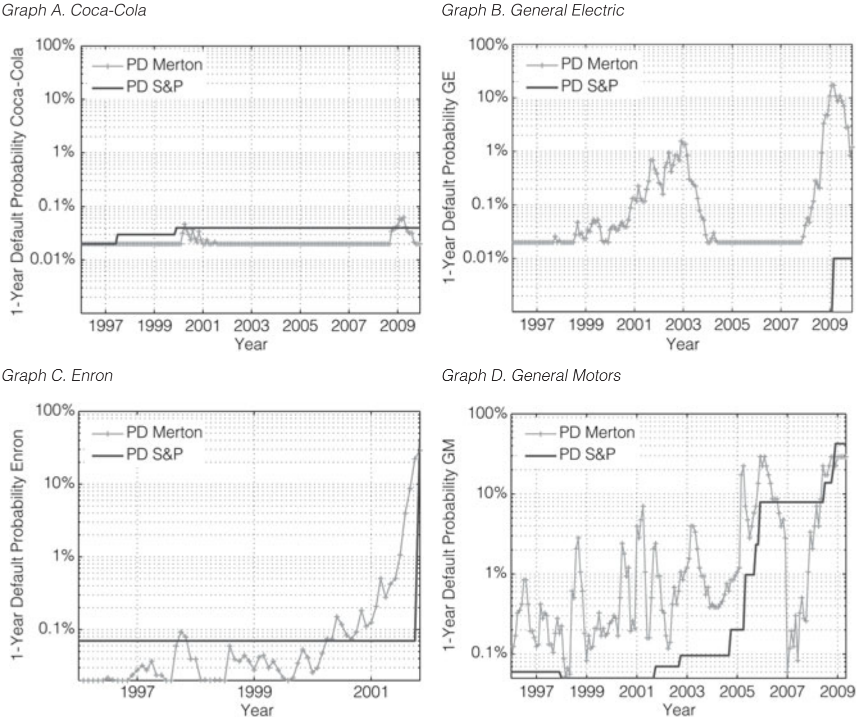


other hand, in March 2009 General Electric (GE) was downgraded from AAA to AA+, but the distance to default indicated a high likelihood of an imminent default triggered by concerns about GE's short-term liquidity, as shown in Graph B. By the end of 2010, however, GE had not defaulted on its debt. GE could issue billions of dollars of debt guaranteed by the Federal Deposit Insurance Corporation during the financial crisis because it was considered too big to fail. Thus, unlike the distance to default, the S&P rating reflects the anticipated government backing of GE due to its systematically important subsidiary GE Capital. The default predictions of Coca-Cola (Graph A) and General Motors (Graph D) are also shown in Figure 4. However, the superiority/inferiority of one system over

FIGURE 4

Time Series of 1-Year Default Probabilities

Figure 4 depicts the development of 1-year default probabilities according to S&P and Merton (1974) for 4 selected companies: Coca-Cola (Graph A), General Electric (Graph B), Enron (Graph C), and General Motors (Graph D). General Motors defaulted in June 2009 and Enron in Dec. 2001. Coca-Cola and General Electric have not defaulted on their debt.



Graph A. Coca-Cola

Graph B. General Electric

Graph C. Enron

Graph D. General Motors

the other is not obvious from those 4 arbitrary examples. Fortunately, with my newly developed validation tools I can rigorously test the performance of both systems.

D. Validating under Independence and Dependence

The calibration testing exercise aims at validating the PD curves as depicted in Figure 3 and tabulated in Table 1. Under the null hypothesis, the PD curves are calibrated. For discriminatory power testing, I compare both S&P ratings and Merton (1974) PDs against the naive model as well as against each other. The 1st discrimination analysis compares the predictions with the random or naive model. Do the probability forecasts have discriminatory power at all? The 2nd discrimination analysis juxtaposes 2 different forecasting methods. Are Merton PDs as good as agency issuer ratings in distinguishing between defaulters and nondefaulters? Under the null hypothesis, the 2 approaches have the same discriminatory power. I work under equivalent assumptions, as for my simulation exercise in the previous section. That is, the factors follow scaled beta distributions, and the factor loadings are assumed to be equal across industry sectors. Forecasting errors in different years must be uncorrelated, according to Proposition 7.

I first analyze the performance of S&P forecasts. The results for S&P predictions are tabulated in Table 2. In 2001 I have 1,174 observations and 48 defaults. The expected default frequency of 2.29% is lower than the realized frequency of 4.09%. Under unconditional IID in equation (11), such an outcome represents a 4.1-sigma event, and the null hypothesis of a level-calibrated predictor is definitely rejected. In 2006 and 2007 I also have 4-sigma events. A 4-sigma event occurs less then once every 10,000 years. Hence, if defaults were indeed independent and the predictions calibrated, I would have experienced three 4-sigma events within 10 years! Over all 10 years, I have 14,654 observations with 228 defaults. The difference between the realized and expected default frequency is 0.66% $(= 2.12\% - 1.56\%)$, and the difference would represent a 4.7-sigma event under independence. The null hypothesis of a level-calibrated predictor is clearly rejected.

TABLE 2

Validation of S&P Ratings and Merton's Distances to Default

In Table 2, I perform an out-of-sample validation exercise for the time period from 2001 to 2010. Here, N denotes the number of observations, and N_1 is the number of defaulters. The mean prediction $\bar{P}_N$ in % is compared with the empirical default frequency N_1/N. The expected area above the Lorenz (1905) curve θ is compared with the empirical area $\hat{\theta}$ (both in percentage points). The standard error σ_θ in % is computed under the null hypothesis of calibrated forecasts, T_Δ is the standard Gaussian distributed test statistic comparing the discriminatory power between the 2 forecasting systems, $T_{\pi,\varrho}$ denotes the level statistic under correlation regime ϱ, T_θ is the shape statistic, and Q_ϱ denotes the combined calibration statistic under correlation regime ϱ. The critical value of the $\chi^2_{\langle 2\rangle}$ distributed combined statistic under a significance level of 5% (1%) is 5.9915 (9.2103). * and ** indicate significance at the 5% and 1% levels, respectively.

Year	N	N_1	$\bar{P}_N$	θ	$\hat{\theta}$	σ_θ	T_Δ	$T_{\pi,0}$	$T_{\pi,0.06}$	T_θ	Q_0	$Q_{0.06}$
Panel A. S&P Ratings												
2001	1,174	48	2.29	89.55	90.14	1.68	0.32	4.13**	1.22	0.35	17.21**	1.62
2002	1,252	38	2.32	90.70	88.72	1.84	−1.27	1.68	0.69	−1.08	3.99	1.64
2003	1,320	25	2.19	90.15	92.72	2.37	0.39	−0.73	0.04	1.09	1.70	1.18
2004	1,508	14	1.94	89.80	93.34	3.28	−0.84	−2.85**	0.85	1.08	9.29**	1.88
2005	1,596	11	1.81	89.24	92.14	3.80	−1.20	−3.37**	−1.21	0.76	11.91**	2.06
2006	1,626	6	1.80	89.13	93.13	5.23	−1.49	−4.34**	−3.28**	0.76	19.41**	11.33**
2007	1,656	5	2.00	89.07	98.88	5.69	0.17	−4.94**	$-\infty$**	1.72	27.35**	∞**
2008	1,559	23	2.12	88.25	90.41	2.69	−1.24	−1.76	−0.28	0.80	3.74	0.73
2009	1,528	44	2.26	88.40	92.78	1.88	0.45	1.64	0.65	2.33*	8.14*	5.86
2010	1,435	14	2.65	88.64	98.01	3.30	1.87	−3.95**	−1.26	2.84**	23.67**	9.68**
2001–2010	14,654	228	2.12	89.33	91.93	0.82	−1.11	−4.75**	−1.43	3.19**	32.69**	12.21**
Panel B. Merton's Distances to Default												
2001	1,174	48	3.42	90.33	89.38	1.50	−0.32	1.27	0.54	−0.64	2.01	0.70
2002	1,252	38	1.66	92.08	92.47	1.83	1.27	3.81**	1.27	0.21	14.60**	1.65
2003	1,320	25	2.81	90.46	91.93	2.23	−0.39	−2.02*	−0.33	0.66	4.51	0.54
2004	1,508	14	0.36	94.10	95.01	3.79	0.84	3.70**	1.93	0.24	13.72**	3.78
2005	1,596	11	0.36	94.55	95.81	4.08	1.20	2.19*	1.35	0.31	4.90	1.92
2006	1,626	6	0.40	95.45	97.99	5.10	1.49	−0.22	0.13	0.50	0.30	0.26
2007	1,656	5	0.28	93.73	98.83	6.77	−0.17	0.14	0.36	0.75	0.59	0.70
2008	1,559	23	0.96	94.97	92.45	2.05	1.24	2.10**	0.97	−1.23	5.91	2.44
2009	1,528	44	5.86	83.04	91.99	2.36	−0.45	−4.97**	−0.78	3.80**	39.08**	15.03**
2010	1,435	14	0.51	90.62	92.84	4.14	−1.87	2.51*	1.37	0.54	6.58*	2.15
2001–2010	14,654	228	1.58	93.91	92.91	0.60	1.11	−0.28	−0.01	−1.68	2.89	2.81

Under conditional IID in expression (14), however, the p-value in 2001 is 22%, computed from the beta distribution $0.0229 X_1 \sim \beta(1.55, 66.15)$ and a realization of X_1 of $1.98 = (4.09\% - 2.29\% \times 0.2)/(2.29\% \times 0.8)$, so that the calibration hypothesis is no longer rejected. Conversely, in 2004 I have 1,508 observations and 14 defaults, an expected default frequency of 1.94%, but now a realized rate of only 0.94%. Under IID such an outcome is highly unlikely (2.9-sigma event). Under conditional IID, however, the mean default rate of 0.94% is well within the 95% confidence band (even in the 1-sigma band). In 2001 and 2004, the level statistics under IID in expression (11) and unconditional IID in expression (14) produce contradictory outcomes with respect to the calibration of S&P ratings. My multiperiod statistic under a moderate correlation of $\varrho = 0.06$ results in less than a 2-sigma event. Consequently, the level calibration hypothesis is no longer rejected. This outcome is based on 10 years of data or 10 uncorrelated time sectors. Ten data points of yearly default rates are regarded as a long time series, and 5-year series are seen as sufficient (see Basel Committee on Banking Supervision (2005), p. 29). But even for a long time series, cross correlation must be considered.

The shape statistics listed in Table 2 are not affected by the degree of dependence and are within the 2-sigma confidence bounds for all years except 2009 and 2010. In 2009 (2010) I have an area above the empirical Lorenz (1905) curve of $\hat{\theta} = 92.78\%$ (98.01%), an expected area of $\theta = 88.40\%$ (88.64%), and a standard error of 1.88% (3.30%). Over the 10-year period, the multiperiod shape statistic indicates the rejection of the shape calibration hypothesis due to the outliers in 2009 and 2010. The area above the expected Lorenz curve is only 89.33%, whereas the corresponding empirical area is 91.93% under a standard error of 0.82%. In other words, defaults are empirically more heavily concentrated in the high-risk classes than predicted. As a consequence, there may be no constant functional dependence between S&P ratings and default probabilities. In a good business cycle, borrowers who are rated say BBB may have the lower default probability as compared to BBB-rated obligors in a bad business cycle.

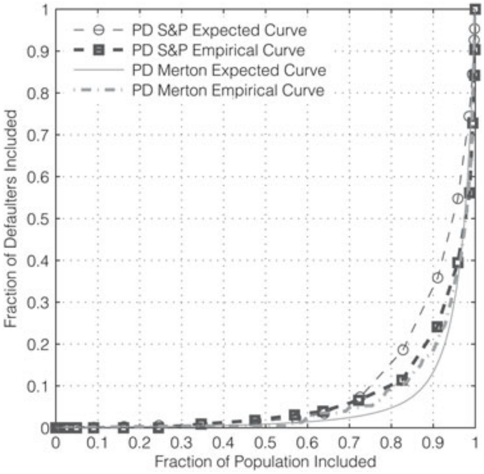
The results look a bit different for Merton (1974) PDs as tabulated in Table 2. Over all 10 years and under the assumption of default dependencies, the only outlier is observed in 2009. At the beginning of 2009, at the height of the financial market crisis, the decline in asset values and the increase in volatilities lead to a predicted default rate of 5.9% even though the realization at the end of 2009 was only 2.9%. Over all 10 years and taking into account a correlation of 6% leads to the conclusion that Merton PDs are calibrated. However, under independence, the level calibration hypothesis would be rejected in 7 out of 10 years!

The shape is rather poorly fitted in 2009, which can be seen by observing the difference $\theta - \hat{\theta}$ and the standard error σ_θ in Table 2. The shape in 2009 indicates that low-risk borrowers have lower default rates and high-risk borrowers have higher default rates than forecast (the sign of $\hat{\theta} - \theta$ is positive). However, over all 10 years I cannot reject the calibration hypothesis. Both the multiperiod shape and level statistics are well within the 2-sigma confidence bands. In Figure 5 I plot the empirical and the expected Lorenz (1905) curves for S&P ratings and Merton (1974) PDs. The area above the empirical Lorenz curve $\hat{\theta}$ for S&P ratings is 91.93%, and the corresponding area for Merton PDs is 92.91%. Since the standard errors are smaller than 2%, both areas are significantly greater than 50%. Both prediction methods are definitely more powerful than naive forecasts. The difference between the 2 areas of 0.98% $(= 91.93\% - 92.91\%)$

FIGURE 5

Lorenz Curves to Measure Discrimination and Shape Calibration

In Figure 5 the empirical and expected Lorenz (1905) curves are constructed with 14,654 observations from 2001 to 2010 based on S&P ratings, Merton's (1974) distance to default, and 1,239 credit defaults, from the data in Table 1. The area above the expected Lorenz curve under S&P's approach (Merton's approach) θ is 89.33% (93.91%), and the area above the empirical Lorenz curves $\hat{\theta}$ is 91.93% (92.91%). Both approaches demonstrate discriminatory power; they clearly outperform the naive model. Even though Merton's approach has the higher point estimate for the area above the Lorenz curve, the difference to S&P ratings is insignificant. Furthermore, the difference between empirical and expected Lorenz curve is highly significant for S&P ratings but insignificant for Merton's distance-to-default model. Hence, the shape calibration hypothesis is rejected with respect to S&P ratings but not so for Merton PDs.



is insignificant given the standard error of 0.88%. Thus, the null hypothesis that both prediction systems have the same discriminatory power cannot be rejected.

In general, my findings favor Merton (1974) PDs over S&P ratings for the validated forecast horizon of 1 year. Both prediction methods have about the same discriminatory power, but unlike S&P ratings, Merton PDs produce calibrated forecasts. As assumed, the functional dependence between the distance to default and the PD seems to be constant over the business cycle. The relation between S&P ratings and PDs, however, seems to be inconstant over time. I implicitly assume that borrowers in the same rating category are homogeneous with respect to the default likelihood irrespective of time or sector. Given my validation results, this homogeneity for S&P ratings seems to be an inadequate assumption. Actually, the superiority of Merton PDs could be caused by the fundamental difference between the through-the-cycle approach as followed by S&P and the point-in-time approach of Merton PDs. Unlike the Merton PD, the S&P rating does not reflect transitory effects or short-term fluctuations (see Treacy and Carey (2000) for a discussion).

To conclude this section, if one agrees with the overwhelming literature on dependent defaults as discussed in Section VI, the discrepancies between unconditional IID and conditional IID underscore the need to consider cross correlation between observations when it comes to the verification of default probabilities. Even small cross correlations can reverse the validation outcome.

IX. Conclusion

This paper introduces novel test statistics for the verification of default probabilities. Although emphasis in this paper is exclusively on default risk, the method has the potential to be applied to other fields. Well-performing probability forecasts demonstrate good discrimination and calibration. I show that calibrated and powerful forecasts are also MSE efficient. Calibration can be decomposed into level calibration and shape calibration. I derive statistical tests for calibration and discrimination as well as subtests for shape and level calibration under the assumption of a *conditional IID* setup. My global test on calibration is asymptotically χ^2 distributed with 2 degrees of freedom. The 2 subtests on calibration as well as the discrimination test are asymptotically standard Gaussian distributed.

There is overwhelming evidence that defaults are dependent, but commonly used performance tests to validate probability forecasts were developed outside the field of finance and economics and under the assumption of *unconditional IID*. This assumption may well apply to meteorological, psychological, or medical studies. I show that not accounting for default dependence can result in the rejection of actually well-performing default prediction systems. As demonstrated in this paper, well-known discrimination figures like the Gini (1921) index or the area under the ROC curve can differ in different states of the economy. When comparing the discriminatory power of 2 models, however, the dependence on the state of the economy can be filtered out.

I derive new summary statistics for observations spread over multiple time periods. I show that a calibrated and powerful default prediction system must generate serially uncorrelated forecasting errors. As a consequence, over time my multiperiod statistics are consistent for the calibration and discrimination hypothesis. That is, with the addition of observations from new time periods, an inferior prediction system is rejected with a probability converging toward 1. This finding contrasts with the view of the Basel Committee on Banking Supervision (2005) that suggests that "there even seems to be no way to develop such [powerful] tests" (p. 34). My validation approach is demonstrated with a simulation exercise and applied to S&P's corporate rating data and Merton's (1974) distance-to-default model. I highlight the need to consider default dependence for observations in the same time period to obtain reasonable validation results and show the discrepancy between dependence and independence. Comprehensive statistical reports and comments on discrimination and calibration for S&P ratings and Merton's distance-to-default model are presented.

Appendix. Proofs and Estimation of Standard Errors

1. Proofs

Proof of Proposition 1. Assume the predictor is level and shape calibrated. Then the "if" part is proven by multiplying both sides of equation (3) with $\mathbb{E}\left[P_i\right] = \mathbb{E}\left[Y_i\right]$ from equation (2). This multiplication yields $\mathbb{E}\left[\mathbf{1}_{\{P_i \leq p\}} Y_i\right] = \mathbb{E}\left[\mathbf{1}_{\{P_i \leq p\}} P_i\right]$ for any $p \in \mathbb{R}$, so that $P_i = \mathbb{E}\left[Y_i \middle| P_i\right]$ according to the conditional expectation's definition. The "only if" part is trivially true. A similar proof is found in Blöchlinger and Leippold (2011). □

Proof of Proposition 2. Proposition 2 is a direct consequence of a powerful predictor (Definition 1) and calibration (Definition 2). If I apply a strictly increasing transformation to the conditional expectation, then it is still powerful but no longer calibrated. If a transformation is made so that the transformed predictor is still calibrated, then the transformed predictor must be a conditional expectation on a different σ-field but can no longer be $\mathcal{G}$-powerful. Only the conditional expectation based on the σ-field $\mathcal{G}$ fulfills both definitions. □

Lemma 1. The bivariate, conditional random variables $\{(P_i, B_i|Y_i = 1) : i \in \mathbb{M}\}$ are IID, so that

$$\begin{aligned}
&\mathbb{P}\left\{P_{M_1} \le p_1, \dots, P_{M_n} \le p_n, B_{M_1} \le b_1, \dots, B_{M_n} \le b_n|\ Y_{M_1} = 1, \dots, Y_{M_n} = 1\right\} \\
&\quad = \prod_{k=1}^{n} \mathbb{P}\left\{P_j \le p_k, B_j \le b_k|\, Y_j = 1\right\},
\end{aligned}$$

for any $p_1, \dots, p_n, b_1, \dots, b_n \in \mathbb{R}$, any set $\mathbb{M} \subseteq \{1, \dots, N\}$ with n elements, and any $j \in \mathbb{M}$, M_k the kth element of $\mathbb{M}$.

Proof of Lemma 1. Note:

$$\begin{aligned}
&\mathbb{P}\left\{P_{M_1} \le p_1, \dots, P_{M_n} \le p_n, B_{M_1} \le b_1, \dots, B_{M_n} \le b_n, Y_{M_1} = 1, \dots, Y_{M_n} = 1|\right. \\
&\quad \left. U_{M_1}, \dots, U_{M_n}\right\} = \prod_{k=1}^{n} U_{M_k} \mathbb{E}\left[\mathbb{P}\left\{Y_{M_k} = 1|\, P_{M_k}, B_{M_k}\right\} \mathbf{1}_{\left\{P_{M_k} \le p_k, B_{M_k} \le b_k\right\}}\right] \\
&\quad = \prod_{k=1}^{n} U_{M_k} \mathbb{P}\left\{P_{M_k} \le p_k, B_{M_k} \le b_k, Y_{M_k} = 1\right\},
\end{aligned}$$

where the equalities follow from my assumptions in Section III, that is, Bernoulli mixture, exchangeability, orthogonality, as well as from iterated expectations. I obtain

$$\begin{aligned}
&\mathbb{P}\left\{P_{M_1} \le p_1, \dots, P_{M_n} \le p_n, B_{M_1} \le b_1, \dots, B_{M_n} \le b_n|\right. \\
&\quad \left. Y_{M_1} = 1, \dots, Y_{M_n} = 1, U_{M_1}, \dots, U_{M_n}\right\} \\
&\quad = \frac{\prod_{k=1}^{n} U_{M_k} \prod_{k=1}^{n} \mathbb{P}\left\{P_{M_k} \le p_k, B_{M_k} \le b_k, Y_{M_k} = 1\right\}}{\prod_{k=1}^{n} U_{M_k} \prod_{k=1}^{n} \mathbb{P}\left\{Y_{M_k} = 1\right\}} \\
&\quad = \prod_{k=1}^{n} \mathbb{P}\left\{P_j \le p_k, B_j \le b_k|\, Y_j = 1\right\}.
\end{aligned}$$

Reducing the fraction, Bayes' theorem and exchangeability explain the last equality. □

Proof of Proposition 3. I have 2 IID sequences of random variables, that is, the sequence of probability forecasts, P_i, and the sequence of forecasts given a subsequent default, $(P_i|Y_i = 1)$. The former sequence is IID according to exchangeability in Assumption 2 and orthogonality in Assumption 3, the latter according to Lemma 1. As a consequence, I just need to apply the Glivenko-Cantelli theorem (see, e.g., Theorem 7.28 in Karr (1993) for a proof). □

Proof of Proposition 4. I have N_1 draws from the defaulters' distribution and N_0 draws from the nondefaulters' distribution such that $N_1/(N_0 + N_1) \xrightarrow{P} \mathbb{P}\{Y_i = 1|\, \mathbf{X}\}$. Neither the defaulters' distribution of (P_i, B_i), Lemma 1, nor the population's distribution, Assumption 3, depends on $\mathbf{X}$. Thus, I have for any $p, b \in \mathbb{R}$, and for any $i \in \{1, \dots, N\}$,

$$\begin{aligned}
\text{(A-1)} \qquad \mathbb{F}_{P,B}(p, b) &= \mathbb{P}\{Y_i = 1|\, \mathbf{X}\}\, \mathbb{F}_{P,B|Y=1}(p, b) \\
&\quad + \mathbb{P}\{Y_i = 0|\, \mathbf{X}\}\, \mathbb{F}_{P,B|Y=0,\mathbf{X}}(p, b),
\end{aligned}$$

where $\mathbb{F}_{P,B}(p,b) = \mathbb{P}\{P_i \le p, B_i \le b\}$, $\mathbb{F}_{P,B|Y=1}(p,b) = \mathbb{P}\{P_i \le p, B_i \le b|\, Y_i = 1\}$, and the nondefaulter's CDF $\mathbb{F}_{P,B|Y=0,\mathbf{X}}(p,b) = \mathbb{P}\{P_i \le p, B_i \le b|\, Y_i = 0, \mathbf{X}\}$.

First, I derive the 2nd moment of $\hat{\theta}_{P,N} - \hat{\theta}_{B,N}$ for a known number of defaulters and nondefaulters as well as conditional on random factors. To avoid long-winded notations, I do not explicitly write the conditioning on $\sigma(\mathbf{X}, N_0, N_1)$ in the following. Conditional on $\mathbf{X}$, the observation sequence $\{(Y_i, P_i, B_i) : i = 1, \ldots, N\}$ is IID. Second, I then let the number of observations tend to infinity to obtain the limiting distribution. Now, I define a linear transformation of $\hat{\theta}_{P,N}$, $\hat{\vartheta}_{P,N} = (1/N_1N_0)\sum_{i=1}^{N}\sum_{j}^{N} \Psi(P_i, P_j)\, Y_i\,(1 - Y_j)$. Hence,

$$\begin{aligned}
\hat{\theta}_{P,N} &= \frac{1}{N_1 N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\{\Psi(P_i, P_j)\, Y_i\,(1 - Y_j) + \Psi(P_i, P_j)\, Y_i Y_j\right\} \\
&= \frac{N_0}{N}\hat{\vartheta}_{P,N} + \frac{1}{2}\frac{N_1}{N}.
\end{aligned}$$

The statistic $\hat{\vartheta}_{P,N}$ denotes the area under the ROC curve. The Gini (1921) coefficient is given by $2\hat{\vartheta}_{P,N} - 1$. Expectation and variance between $\hat{\theta}_{P,N}$ and $\hat{\vartheta}_{P,N}$ are also linearly related:

$$\text{(A-2)} \qquad \mathbb{E}\left[\hat{\theta}_{P,N}\right] = \frac{N_0}{N}\mathbb{E}\left[\hat{\vartheta}_{P,N}\right] + \frac{1}{2}\frac{N_1}{N}, \quad \text{and} \quad \mathbb{V}\left[\hat{\theta}_{P,N}\right] = \frac{N_0^2}{N^2}\mathbb{V}\left[\hat{\vartheta}_{P,N}\right].$$

I show that

$$\text{(A-3)} \qquad \frac{\hat{\theta}_P - \hat{\theta}_B - \mathbb{E}\left[\hat{\theta}_P - \hat{\theta}_B\right]}{\sqrt{\hat{\mathbb{V}}\left[\hat{\theta}_P - \hat{\theta}_B\right]}} = \frac{\hat{\vartheta}_{P,N} - \hat{\vartheta}_{B,N} - \mathbb{E}\left[\hat{\vartheta}_{P,N} - \hat{\vartheta}_{B,N}\right]}{\sqrt{\mathbb{V}\left[\hat{\vartheta}_{P,N} - \hat{\vartheta}_{B,N}\right]}}\,\frac{\sqrt{\mathbb{V}\left[\hat{\vartheta}_{P,N} - \hat{\vartheta}_{B,N}\right]}}{\sqrt{\hat{\mathbb{V}}\left[\hat{\vartheta}_{P,N} - \hat{\vartheta}_{B,N}\right]}}$$

is asymptotically Gaussian distributed. The 2nd factor on the right-hand side of equation (A-3) converges in probability toward 1, since by IID, the empirical variance is a consistent estimator for the true variance. Hence, if I can show that the 1st factor is Gaussian distributed, I can resort to Slutsky's theorem and the proposition is proven.

It remains to be shown that

$$\lim_{N_1, N_0 \to \infty} \mathbb{P}\left\{\frac{\hat{\vartheta}_{P,N} - \hat{\vartheta}_{B,N}}{\sqrt{\mathbb{V}\left[\hat{\vartheta}_{P,N} - \hat{\vartheta}_{B,N}\right]}} \le s\right\} = \Phi(s), \qquad \text{for any } s \in \mathbb{R}.$$

Under the null hypothesis I have $\mathbb{E}\left[\hat{\vartheta}_{P,N} - \hat{\vartheta}_{B,N}\right] = 0$. I construct 2 auxiliary quantities:

$$\begin{aligned}
\tilde{\vartheta}_{P,N_1} &= \frac{1}{N_1}\sum_{\{i:Y_i=1\}} \mathbb{E}\left[\Psi(P_i, P_j)|\, Y_i = 1, Y_j = 0, P_i\right], \qquad \text{and} \\
\tilde{\vartheta}_{B,N_1} &= \frac{1}{N_1}\sum_{\{i:Y_i=1\}} \mathbb{E}\left[\Psi(B_i, B_j)|\, Y_i = 1, Y_j = 0, B_i\right].
\end{aligned}$$

I have a sequence of independent and bounded random variables:

$$\nu_i \;:=\; \mathbb{E}\left[\Psi(P_i, P_j)|\, Y_i = 1, Y_j = 0, P_i\right] - \mathbb{E}\left[\Psi(B_i, B_j)|\, Y_i = 1, Y_j = 0, B_i\right].$$

Hence, the sequence satisfies the central limit theorem. With the number of draws N_1 from the defaulters' distribution, $\mathbb{F}_{P,B|Y=1}(\cdot,\cdot)$, going to infinity, I have for any $s \in \mathbb{R}$,

$$\begin{aligned}
\lim_{N_1 \to \infty} \mathbb{P}\left\{\frac{\sqrt{N_1}\sum_{\{i:Y_i=1\}} \nu_i}{\sqrt{\mathbb{V}\left[\sum_{\{i:Y_i=1\}} \nu_i\right]}} \le s\right\} &= \lim_{N_1 \to \infty} \mathbb{P}\left\{\frac{\tilde{\vartheta}_{P,N_1} - \tilde{\vartheta}_{B,N_1}}{\sqrt{\mathbb{V}\left[\tilde{\vartheta}_{P,N_1} - \tilde{\vartheta}_{B,N_1}\right]}} \le s\right\} \\
&= \Phi(s),
\end{aligned}$$

where I make use of $(1/N_1)\sum_{\{i:Y_i=1\}}\nu_i = \tilde{\vartheta}_{P,N_1} - \tilde{\vartheta}_{B,N_1}$. Now, I show that

$$\frac{\tilde{\vartheta}_{P,N_1} - \tilde{\vartheta}_{B,N_1}}{\sqrt{\mathbb{V}\left[\tilde{\vartheta}_{P,N_1} - \tilde{\vartheta}_{B,N_1}\right]}} \quad \text{and} \quad \frac{\hat{\vartheta}_{P,N} - \hat{\vartheta}_{B,N}}{\sqrt{\mathbb{V}\left[\hat{\vartheta}_{P,N} - \hat{\vartheta}_{B,N}\right]}}$$

converge in quadratic mean when the number of draws, N_0, from the nondefaulters' distribution, $\mathbb{F}_{P,B|Y=0,\mathbf{x}}(\cdot,\cdot)$, tends to infinity.

Thus, I show that

$$\text{(A-4)} \qquad 0 = \lim_{N_0\to\infty} \mathbb{E}\left[\left(\frac{\tilde{\vartheta}_{P,N_1} - \tilde{\vartheta}_{B,N_1}}{\sqrt{\mathbb{V}\left[\tilde{\vartheta}_{P,N_1} - \tilde{\vartheta}_{B,N_1}\right]}} - \frac{\hat{\vartheta}_{P,N} - \hat{\vartheta}_{B,N}}{\sqrt{\mathbb{V}\left[\hat{\vartheta}_{P,N} - \hat{\vartheta}_{B,N}\right]}}\right)^2\right].$$

Convergence in quadratic mean implies convergence in distribution. Rewriting equation (A-4),

$$\begin{aligned}\text{(A-5)} \quad 1 &= \lim_{N_0\to\infty} \frac{\mathbb{E}\left[\left(\tilde{\vartheta}_{P,N_1} - \tilde{\vartheta}_{B,N_1}\right)\left(\hat{\vartheta}_{P,N} - \hat{\vartheta}_{B,N}\right)\right]}{\sqrt{\mathbb{E}\left[\left(\tilde{\vartheta}_{P,N_1} - \tilde{\vartheta}_{B,N_1}\right)^2\right]\mathbb{E}\left[\left(\hat{\vartheta}_{P,N} - \hat{\vartheta}_{B,N}\right)^2\right]}} \\ &= \lim_{N_0\to\infty} \frac{N_1\left(\mathbb{E}\left[\tilde{\vartheta}_{P,N_1}\hat{\vartheta}_{P,N}\right] - \mathbb{E}\left[\tilde{\vartheta}_{P,N_1}\hat{\vartheta}_{B,N}\right] - \mathbb{E}\left[\tilde{\vartheta}_{B,N_1}\hat{\vartheta}_{P,N}\right] + \mathbb{E}\left[\tilde{\vartheta}_{B,N_1}\hat{\vartheta}_{B,N}\right]\right)}{\sqrt{N_1^2\left(\mathbb{E}\left[\tilde{\vartheta}^2_{P,N_1}\right] - 2\mathbb{E}\left[\tilde{\vartheta}_{P,N_1}\tilde{\vartheta}_{B,N_1}\right] + \mathbb{E}\left[\tilde{\vartheta}^2_{B,N_1}\right]\right)\left(\mathbb{E}\left[\hat{\vartheta}^2_{P,N}\right] - 2\mathbb{E}\left[\hat{\vartheta}_{P,N}\hat{\vartheta}_{B,N}\right] + \mathbb{E}\left[\hat{\vartheta}^2_{B,N}\right]\right)}}.\end{aligned}$$

Evaluating expectations, I obtain

$$\begin{aligned}\text{(A-6)} \quad N_0N_1\mathbb{E}\left[\hat{\vartheta}^2_{P,N}\right] &= (N_0-1)(N_1-1)\mathbb{E}\left[\varPsi(P_i,P_j)\middle| Y_i=1, Y_j=0\right]^2 \\ &\quad + (N_0-1)\mathbb{E}\left[\varPsi(P_i,P_j)\varPsi(P_i,P_l)\middle| Y_i=1, Y_j=0, Y_l=0\right] \\ &\quad + (N_1-1)\mathbb{E}\left[\varPsi(P_i,P_j)\varPsi(P_k,P_j)\middle| Y_i=1, Y_j=0, Y_k=1\right] \\ &\quad + \mathbb{E}\left[\varPsi(P_i,P_j)^2\middle| Y_i=1, Y_j=0\right],\end{aligned}$$

$$\begin{aligned}\text{(A-7)} \quad N_0N_1\mathbb{E}\left[\hat{\vartheta}^2_{B,N}\right] &= (N_0-1)(N_1-1)\mathbb{E}\left[\varPsi(B_i,B_j)\middle| Y_i=1, Y_j=0\right]^2 \\ &\quad + (N_0-1)\mathbb{E}\left[\varPsi(B_i,B_j)\varPsi(B_i,B_l)\middle| Y_i=1, Y_j=0, Y_l=0\right] \\ &\quad + (N_1-1)\mathbb{E}\left[\varPsi(B_i,B_j)\varPsi(B_k,B_j)\middle| Y_i=1, Y_j=0, Y_k=1\right] \\ &\quad + \mathbb{E}\left[\varPsi(B_i,B_j)^2\middle| Y_i=1, Y_j=0\right],\end{aligned}$$

and

$$\begin{aligned}\text{(A-8)} \quad & N_0N_1\mathbb{E}\left[\hat{\vartheta}_{P,N}\hat{\vartheta}_{B,N}\right] \\ &= (N_0-1)(N_1-1)\mathbb{E}\left[\varPsi(P_i,P_j)\middle| Y_i=1, Y_j=0\right] \\ &\quad \times \mathbb{E}\left[\varPsi(B_i,B_j)\middle| Y_i=1, Y_j=0\right] \\ &\quad + (N_0-1)\mathbb{E}\left[\varPsi(P_i,P_j)\varPsi(B_i,B_l)\middle| Y_i=1, Y_j=0, Y_l=0\right] \\ &\quad + (N_1-1)\mathbb{E}\left[\varPsi(P_i,P_j)\varPsi(B_k,B_j)\middle| Y_i=1, Y_j=0, Y_k=1\right] \\ &\quad + \mathbb{E}\left[\varPsi(P_i,P_j)\varPsi(B_i,B_j)\middle| Y_i=1, Y_j=0\right].\end{aligned}$$

In the limit, when $N_0 \to \infty$,

$$\begin{aligned}\lim_{N_0\to\infty} N_1\mathbb{E}\left[\hat{\vartheta}^2_{P,N}\right] &= (N_1-1)\mathbb{E}\left[\varPsi(P_i,P_j)\middle| Y_i=1, Y_j=0\right]^2 \\ &\quad + \mathbb{E}\left[\varPsi(P_i,P_j)\varPsi(P_i,P_l)\middle| Y_i=1, Y_j=0, Y_l=0\right], \\ \lim_{N_0\to\infty} N_1\mathbb{E}\left[\hat{\vartheta}^2_{B,N}\right] &= (N_1-1)\mathbb{E}\left[\varPsi(B_i,B_j)\middle| Y_i=1, Y_j=0\right]^2 \\ &\quad + \mathbb{E}\left[\varPsi(B_i,B_j)\varPsi(B_i,B_l)\middle| Y_i=1, Y_j=0, Y_l=0\right],\end{aligned}$$

$$\begin{aligned}
\lim_{N_0\to\infty} N_1\mathbb{E}\left[\hat{\vartheta}_{P,N}\hat{\vartheta}_{B,N}\right] &= (N_1-1)\mathbb{E}\left[\Psi\left(P_i,P_j\right)\middle|\, Y_i=1,Y_j=0\right] \\
&\quad \times \mathbb{E}\left[\Psi\left(B_i,B_j\right)\middle|\, Y_i=1,Y_j=0\right] \\
&\quad + \mathbb{E}\left[\Psi\left(P_i,P_j\right)\Psi\left(B_i,B_l\right)\middle|\, Y_i=1,Y_j=0,Y_l=0\right].
\end{aligned}$$

Furthermore,

$$\begin{aligned}
N_1\mathbb{E}\left[\tilde{\vartheta}^2_{P,N_1}\right] &= (N_1-1)\mathbb{E}\left[\Psi\left(P_i,P_j\right)\middle|\, Y_i=1,Y_j=0\right]^2 \\
&\quad + \mathbb{E}\left[\Psi\left(P_i,P_j\right)\Psi\left(P_i,P_l\right)\middle|\, Y_i=1,Y_j=0,Y_l=0\right], \\
N_1\mathbb{E}\left[\tilde{\vartheta}^2_{B,N_1}\right] &= (N_1-1)\mathbb{E}\left[\Psi\left(B_i,B_j\right)\middle|\, Y_i=1,Y_j=0\right]^2 \\
&\quad + \mathbb{E}\left[\Psi\left(B_i,B_j\right)\Psi\left(B_i,B_l\right)\middle|\, Y_i=1,Y_j=0,Y_l=0\right], \\
N_1\mathbb{E}\left[\tilde{\vartheta}_{P,N_1}\tilde{\vartheta}_{B,N_1}\right] &= (N_1-1)\mathbb{E}\left[\Psi\left(P_i,P_j\right)\middle|\, Y_i=1,Y_j=0\right] \\
&\quad \times \mathbb{E}\left[\Psi\left(B_i,B_j\right)\middle|\, Y_i=1,Y_j=0\right] \\
&\quad + \mathbb{E}\left[\Psi\left(P_i,P_j\right)\Psi\left(B_i,B_l\right)\middle|\, Y_i=1,Y_j=0,Y_l=0\right],
\end{aligned}$$

and

$$\begin{aligned}
N_1\mathbb{E}\left[\tilde{\vartheta}_{P,N_1}\hat{\vartheta}_{P,N}\right] &= (N_1-1)\mathbb{E}\left[\Psi\left(P_i,P_j\right)\middle|\, Y_i=1,Y_j=0\right]^2 \\
&\quad + \mathbb{E}\left[\Psi\left(P_i,P_j\right)\Psi\left(P_i,P_l\right)\middle|\, Y_i=1,Y_j=0,Y_l=0\right], \\
N_1\mathbb{E}\left[\tilde{\vartheta}_{B,N_1}\hat{\vartheta}_{B,N}\right] &= (N_1-1)\mathbb{E}\left[\Psi\left(B_i,B_j\right)\middle|\, Y_i=1,Y_j=0\right]^2 \\
&\quad + \mathbb{E}\left[\Psi\left(B_i,B_j\right)\Psi\left(B_i,B_l\right)\middle|\, Y_i=1,Y_j=0,Y_l=0\right], \\
N_1\mathbb{E}\left[\tilde{\vartheta}_{P,N_1}\hat{\vartheta}_{B,N}\right] &= (N_1-1)\mathbb{E}\left[\Psi\left(P_i,P_j\right)\middle|\, Y_i=1,Y_j=0\right] \\
&\quad \times \mathbb{E}\left[\Psi\left(B_i,B_j\right)\middle|\, Y_i=1,Y_j=0\right] \\
&\quad + \mathbb{E}\left[\Psi\left(P_i,P_j\right)\Psi\left(B_i,B_l\right)\middle|\, Y_i=1,Y_j=0,Y_l=0\right], \\
N_1\mathbb{E}\left[\tilde{\vartheta}_{B,N_1}\hat{\vartheta}_{P,N}\right] &= N_1\mathbb{E}\left[\tilde{\vartheta}_{P,N_1}\hat{\vartheta}_{B,N}\right].
\end{aligned}$$

By inserting those terms into the numerator and denominator in equation (A-5), I see that the quotient is indeed 1 in the limit. □

Proof of Proposition 5. The convergence in probability follows from the law of large numbers for exchangeable random variables (see, e.g., Hall and Heyde (1980), eq. (7.1), p. 202). The equality then follows from the multiplicative setup in Assumption 1, from level calibration, $\mathbb{E}\left[Y_i\right]=\mathbb{E}\left[P_i\right]$, and the assumed linearity of U_i. □

Proof of Corollary 1. Corollary 1 is a consequence of Proposition 4 by replacing $\hat{\theta}_{B,N}$ with θ_P. I have $\theta_P=\mathbb{E}\left[\hat{\theta}_{P,N}\middle|\,\mathbf{X}\right]$ for any realizations of $\mathbf{X}$, since neither the CDF of $(P_i|Y_i=1)$ (Lemma 1) nor the CDF of P_i (Assumption 3) depends on $\mathbf{X}$, and the sequences P_i and $(P_i|Y_i=1)$ are IID (Proposition 3). □

Proof of Proposition 6. By Proposition 5, $\overline{Y}_N$ converges in probability toward $\mathbb{P}\left\{Y_i=1\middle|\,\mathbf{X}\right\}$ so that $T_{\pi,N}$ is in the limit only a function of the K factors $\mathbf{X}$, but the standardized area above the Lorenz (1905) curve is independent from $\mathbf{X}$. Conditional on $\mathbf{X}$, I therefore have

$$\begin{aligned}
\lim_{N\to\infty}\mathbb{E}\left[\mathbf{1}_{\left\{T_{\pi,N}\le t_1\right\}}\mathbf{1}_{\left\{T_{\theta,N}\le t_2\right\}}\middle|\,\mathbf{X}\right] &= \mathbb{E}\left[\mathbf{1}_{\left\{\lim_{N\to\infty}T_{\pi,N}\le t_1\right\}}\mathbf{1}_{\left\{\lim_{N\to\infty}T_{\theta,N}\le t_2\right\}}\middle|\,\mathbf{X}\right] \\
&= \mathbf{1}_{\left\{\lim_{N\to\infty}T_{\pi,N}\le t_1\right\}}\mathbb{P}\left\{\lim_{N\to\infty}T_{\theta,N}\le t_2\middle|\,\mathbf{X}\right\} \\
&= \mathbf{1}_{\left\{\lim_{N\to\infty}T_{\pi,N}\le t_1\right\}}\Phi\left(t_2\right),
\end{aligned}$$

so that $\lim_{N\to\infty}\mathbb{P}\left\{T_{\pi,N}\le t_1,T_{\theta,N}\le t_2\right\}=\Phi\left(t_1\right)\Phi\left(t_2\right)$ for any $t_1,t_2\in\mathbb{R}$ by iterated expectations and dominated convergence. The dominated convergence theorem is also applied to the 1st and 3rd equalities. The 2nd equality follows from Proposition 5. The 3rd equality exploits Corollary 1. In the limit, I therefore have 2 independent standard Gaussian variables. □

2. Estimation of Standard Errors

The variance of $\hat{\theta}_{P,N} - \hat{\theta}_{B,N}$ is given by

$$\mathbb{V}\left[\hat{\theta}_{P,N} - \hat{\theta}_{B,N}\right] = \frac{N_0^2}{N^2}\left\{\mathbb{E}\left[\hat{\vartheta}_{P,N}^2\right] - 2\mathbb{E}\left[\hat{\vartheta}_{P,N}\hat{\vartheta}_{B,N}\right] + \mathbb{E}\left[\hat{\vartheta}_{B,N}^2\right]\right\}, \tag{A-9}$$

as shown in the previous section. The 3 expectation terms are derived in equations (A-6), (A-7), and (A-8). To estimate these 3 expectation terms, I simply replace theoretical means with empirical means. In the same manner, I can estimate the variance in Corollary 1. By Slutsky's theorem, both the limit distribution in Proposition 4 and the limit distribution in Corollary 1 are unchanged when working with empirical instead of theoretical means for the computation of the standard error.

There is also an alternative. The standard error in Corollary 1 can be directly calculated via the expectation terms derived in the previous section. Thus, the variance of $\hat{\theta}_{P,N}$,

$$\mathbb{V}\left[\hat{\theta}_{P,N}\right] = \frac{N_0^2}{N^2}\left\{\mathbb{E}\left[\hat{\vartheta}_{P,N}^2\right] - \mathbb{E}\left[\hat{\vartheta}_{P,N}\right]^2\right\}, \tag{A-10}$$

is computed from the following expectation terms:

$$\mathbb{E}\left[\hat{\vartheta}_{P,N}\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Psi\left(p_1, p_0\right) d\mathbb{F}_{P|Y=1}\left(p_1\right) d\mathbb{F}_{P|Y=0,\mathbf{X}}\left(p_0\right),$$

and

$$\begin{aligned}\mathbb{E}\left[\hat{\vartheta}_{P,N}^2\right] &= \mathbb{E}\left[\hat{\vartheta}_{P,N}\right]^2 + \frac{1-N_0-N_1}{N_0N_1}\mathbb{E}\left[\hat{\vartheta}_{P,N}\right]^2 \\ &\quad + \frac{N_0-1}{N_0N_1}\mathbb{E}\left[\Psi\left(P_i,P_j\right)\Psi\left(P_i,P_l\right)\middle|\, Y_i=1, Y_j=0, Y_l=0\right] \\ &\quad + \frac{N_1-1}{N_0N_1}\mathbb{E}\left[\Psi\left(P_i,P_j\right)\Psi\left(P_k,P_i\right)\middle|\, Y_i=1, Y_j=0, Y_k=1\right] \\ &\quad + \frac{1}{N_0N_1}\mathbb{E}\left[\Psi\left(P_i,P_j\right)^2\middle|\, Y_i=1, Y_j=0\right],\end{aligned}$$

with

$$\begin{aligned}&\mathbb{E}\left[\Psi\left(P_i,P_j\right)\Psi\left(P_i,P_l\right)\middle|\, Y_i=1, Y_j=0, Y_l=0\right] \\ &\quad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Psi\left(p_1,p_0\right)\Psi\left(p_1,p\right) d\mathbb{F}_{P|Y=1}\left(p_1\right) \\ &\quad\quad \times d\mathbb{F}_{P|Y=0,\mathbf{X}}\left(p_0\right) d\mathbb{F}_{P|Y=0,\mathbf{X}}\left(p\right), \\ &\mathbb{E}\left[\Psi\left(P_i,P_j\right)\Psi\left(P_k,P_i\right)\middle|\, Y_i=1, Y_j=0, Y_k=1\right] \\ &\quad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Psi\left(p_1,p_0\right)\Psi\left(p,p_0\right) d\mathbb{F}_{P|Y=1}\left(p_1\right) \\ &\quad\quad \times d\mathbb{F}_{P|Y=0,\mathbf{X}}\left(p_0\right) d\mathbb{F}_{P|Y=1}\left(p\right),\end{aligned}$$

and

$$\mathbb{F}_{P|Y=1}(p_1) = \frac{\mathbb{E}\left[P_i\mathbf{1}_{\{P_i \le p_1\}}\right]}{\mathbb{E}\left[P_i\right]}, \tag{A-11}$$

$$\mathbb{F}_{P|Y=0,\mathbf{X}}(p_0) = \frac{1}{\mathbb{P}\left\{Y_i=0\middle|\,\mathbf{X}\right\}}\mathbb{F}_P(p_0) - \frac{\mathbb{P}\left\{Y_i=1\middle|\,\mathbf{X}\right\}}{\mathbb{P}\left\{Y_i=0\middle|\,\mathbf{X}\right\}}\mathbb{F}_{P|Y=1}(p_0). \tag{A-12}$$

The unconditional CDF $\mathbb{F}_P(\cdot)$ allows, if shape calibrated, the derivation of the defaulter's distribution function $\mathbb{F}_{P|Y=1}(\cdot)$ according to equation (A-11). An asymptotically equivalent derivation for $\mathbb{F}_{P|Y=1}(\cdot)$ is obtained when using the empirical CDF instead of the theoretical CDF $\mathbb{F}_P(\cdot)$. The default frequency N_1/N converges in probability toward $\mathbb{P}\{Y_i = 1|\mathbf{X}\}$. Thus, this convergence provides an approximation for the nondefaulter's distribution function in equation (A-12) by replacing $\mathbb{P}\{Y_i = 1|\mathbf{X}\}$ with N_1/N and $\mathbb{P}\{Y_i = 0|\mathbf{X}\}$ with N_0/N. The CDFs in equations (A-11) and (A-12) together with the previously derived expectation terms allow the computation of the variance in equation (A-10).

References

Alchian, A. A. "Uncertainty, Evolution, and Economic Theory." *Journal of Political Economy*, 58 (1950), 211–221.

Altman, E. I. "Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy." *Journal of Finance*, 23 (1968), 589–609.

Andrews, D. W. K. "Chi-Square Diagnostic Tests for Econometric Models: Theory." *Econometrica*, 56 (1988), 1419–1453.

Andrews, D. W. K. "A Conditional Kolmogorov Test." *Econometrica*, 65 (1997), 1097–1128.

Andrews, D. W. K. "Cross-Section Regression with Common Shocks." *Econometrica*, 73 (2005), 1551–1573.

Balthazar, L. "PD Estimates for Basel II." *Risk Magazine*, 17 (2004), 84–85.

Bamber, D. "The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph." *Journal of Mathematical Psychology*, 12 (1975), 387–415.

Basel Committee on Banking Supervision. "The Internal Ratings-Based Approach." Consultative Document, Bank for International Settlements (2001).

Basel Committee on Banking Supervision. "Studies on the Validation of Internal Rating Systems." Working Paper No. 14, Bank for International Settlements (2005).

Bertolini, G.; R. Damico; D. Nardi; A. Tinazzi; and G. Apolone. "One Model, Several Results: The Paradox of the Hosmer-Lemeshow Goodness-of-Fit Test for the Logistic Regression Model." *Journal of Epidemiology and Biostatistics*, 5 (2000), 251–253.

Bharath, S. T., and T. Shumway. "Forecasting Default with the Merton Distance to Default Model." *Review of Financial Studies*, 21 (2008), 1339–1369.

Blöchlinger, A. "Arbitrage-Free Credit Pricing Using Default Probabilities and Risk Sensitivities." *Journal of Banking and Finance*, 35 (2011), 268–281.

Blöchlinger, A., and M. Leippold. "Economic Benefit of Powerful Credit Scoring." *Journal of Banking and Finance*, 30 (2006), 851–873.

Blöchlinger, A., and M. Leippold. "A New Goodness-of-Fit Test for Event Forecasting and Its Application to Credit Defaults." *Management Science*, 57 (2011), 487–505.

Blochwitz, S.; S. Hohl; D. Tasche; and C. S. Wehn. "Validating Default Probabilities on Short Time Series." Working Paper, Deutsche Bundesbank (2004).

Brier, G. W. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review*, 78 (1950), 1–3.

Chen, R.-R., and B. J. Sopranzetti. "The Valuation of Default-Triggered Credit Derivatives." *Journal of Financial and Quantitative Analysis*, 38 (2003), 359–382.

Cooke, W. E. "Forecasts and Verifications in Western Australia." *Monthly Weather Review*, 34 (1906), 23–24.

Crouhy, M.; R. A. Jarrow; and S. M. Turnbull. "The Subprime Credit Crisis of 2007." *Journal of Derivatives,* 16 (2008), 81–110.

Dawid, A. P. "The Well-Calibrated Bayesian." *Journal of the American Statistical Association*, 77 (1982), 605–610.

Dawid, A. P. "Calibration-Based Empirical Probability." *Annals of Statistics*, 13 (1985), 1251–1274.

DeLong, E. R.; D. M. DeLong; and D. L. Clarke-Pearson. "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach." *Biometrics*, 44 (1988), 837–845.

Department of the Treasury, Federal Reserve System, Federal Deposit Insurance Corporation. "Internal Ratings-Based Systems for Corporate Credit and Operational Risk Advanced Measurement Approaches for Regulatory Capital." Draft Supervisory Guidance with Request for Comment, *Federal Register,* 68, No. 149 (Aug. 4, 2003), 45949–45988.

Friedman, M. *Essays in Positive Economics*. Chicago, IL: University of Chicago Press (1953).

Gini, C. "Measurement of Inequality of Incomes." *Economic Journal*, 31 (1921), 124–126.

Gordy, M. B. "A Comparative Anatomy of Credit Risk Models." *Journal of Banking and Finance*, 24 (2000), 119–149.

Hall, P., and C. C. Heyde. *Martingale Limit Theory and Its Application*. New York, NY: Academic Press (1980).

Harrell, F. E. *Regression Modeling Strategies*. New York, NY: Springer-Verlag (2001).

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

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

Kolmogorov, A. N. "Sulla Determinazione Empirica Di Una Legge Di Distribuzione." *Giornale dell'Istituto Italiano degli Attuari*, 4 (1933), 83–91.

Li, D. X. "On Default Correlation: A Copula Function Approach." *Journal of Fixed Income*, 9 (2000), 43–54.

Lorenz, M. O. "Methods of Measuring the Concentration of Wealth." *Publications of the American Statistical Association*, 9 (1905), 209–219.

Merton, R. C. "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates." *Journal of Finance*, 29 (1974), 449–470.

Pearson, K. "On the Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling." *London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, 50 (1900), 157–175.

Pollard, D. "General Chi-Square Goodness-of-Fit Tests with Data-Dependent Cells." *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 50 (1979), 317–331.

Sandroni, A. "Do Markets Favor Agents Able to Make Accurate Predictions?" *Econometrica*, 68 (2000), 1303–1341.

Spiegelhalter, D. J. "Probabilistic Prediction in Patient Management and Clinical Trials." *Statistics in Medicine*, 5 (1986), 421–433.

Stein, R. M. "Are the Probabilities Right?" Working Paper, Moody's KMV (2003).

Swets, J. A. "Measuring the Accuracy of Diagnostic Systems." *Science*, 240 (1988), 1285–1293.

Treacy, W. F., and M. Carey. "Credit Risk Rating Systems at Large U.S. Banks." *Journal of Banking and Finance*, 24 (2000), 167–201.

Wiginton, J. C. "A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior." *Journal of Financial and Quantitative Analysis*, 15 (1980), 757–770.

Wilde, T. "CreditRisk+ A Credit Risk Management Framework." Technical Document, Credit Suisse (1997).

Zhou, C. "An Analysis of Default Correlations and Multiple Defaults." *Review of Financial Studies*, 14 (2001), 555–576.