

Evaluating credit risk models: A critique and a proposal

Hergen Frerichs^{a,*}

Gunter Löffler^a

University of Frankfurt (Main)

First Version: February, 2001

This version: May 2002

^a Chair of Banking and Finance, University of Frankfurt (Main), P.O. Box 11 19 32, 60054 Frankfurt (Main), Germany.

* Corresponding author: Tel.: ++49-69-79828959, facsimile: ++49-69-79822143. *E-mail addresses:* frerichs@wiwi.uni-frankfurt.de, gloeffler@wiwi.uni-frankfurt.de

We wish to thank Ron Anderson, Wolfgang Bühler, Jan Pieter Krahn, Thilo Liebig, Ludger Overbeck, Peter Raupach, Mark Wahrenburg and participants at the 2001 meeting of the European Financial Management Association, the 2001 research conference of the International Association of Financial Engineers, and seminars at the universities of Frankfurt (Main), Mannheim, and Amsterdam (Free University), for helpful comments.

Evaluating credit risk models: A critique and a proposal

Abstract

Evaluating the quality of credit portfolio risk models is an important issue for both banks and regulators. Lopez and Saidenberg (2000) suggest cross-sectional resampling techniques in order to make efficient use of available data. We show that their proposal disregards cross-sectional dependence in resampled portfolios, which can invalidate standard statistical inference. We proceed by suggesting the Berkowitz (2001) procedure, which relies on standard likelihood ratio tests performed on transformed loss data. We simulate the power of this approach in various settings including one in which the test is extended to incorporate cross-sectional information. Monte Carlo simulations show that a loss history of ten years can be sufficient to resolve uncertainties currently present in credit risk modeling.

Key words: credit risk, density forecasts; model validation, bank regulation

JEL classification: G2; G28; C52

1. Introduction

In the literature on portfolio credit risk models, it is customary to refer to the difficulties of evaluating the quality of these models. Several years after the first models have been proposed, there is only one paper that empirically examines their predictive ability (Nickell, Perraudin and Varotto, 2001). One explanation for the scarcity of research are concerns that evaluation procedures developed for market risk models have little power when applied to credit data sets. The available time series on credit portfolio losses are believed to be too short to produce reliable results.

In order to overcome the lack of credit data in the time dimension Lopez and Saidenberg (2000) propose to evaluate credit portfolio risk models based on cross-sectional simulation. Given a data set covering T years, the idea is to resample, for each of the T years, a large number of portfolios. After predicting the loss distribution for each resampled portfolio, the predictions can be evaluated using a variety of statistical tests. In the construction of the tests, Lopez and Saidenberg assume that prediction errors for portfolios resampled from the loss experience of one year are independent.

We demonstrate that the independence assumption made by Lopez and Saidenberg will not be fulfilled in a typical setting. If the economy moves into recession, for instance, losses will be above average both in the entire sample and in randomly drawn subsamples, which can invalidate the tests proposed by Lopez and Saidenberg. Subsequently, we show that ten years of data can be sufficient for model evaluation if we use the information of the complete distribution. For this purpose, we recommend using Berkowitz' (2001) test procedure. Observed credit

losses are transformed such that they are independent and identically distributed standard normal random variables under the null hypothesis that the model is correct. Standard likelihood ratio tests can then be used to test this hypothesis. For a market risk setting, Berkowitz shows that powerful tests can be constructed with a sample size as small as 100.

Our simulations indicate that as few as ten observations are sufficient to detect misspecifications in credit risk models, a finding that can be illustrated through the following examples. Many credit risk models capture credit event correlations through correlated latent variables. These latent variables are usually interpreted as the borrower's asset values. According to the Basel Committee on Banking Supervision (2001), an average asset correlation of 20% is consistent with industry practice. In a calibration exercise for US loan portfolios, however, Gordy (2000) obtains correlation estimates that vary between 1.5% and 12.5%. With ten years of data on annual losses, a true correlation of 5% and a significance level of 10%, the probability of rejecting a correlation assumption of 20% can be above 90% in our examples. Another currently debated issue is whether the latent variables are normally distributed. If they follow a t-distribution with 10 degrees of freedom, the probability of rejecting the normal assumption is again above 90%.

We follow Lopez and Saidenberg in trying to exploit information contained in the cross-section of losses. Specifically, we consider cases where evaluators have a-priori information on the nature of possible misspecifications. This information can be used to split the portfolio into subportfolios. If the number of subportfolios is not too large, the Berkowitz procedure can be adapted to jointly test the validity of predictions for subportfolio losses. To gain an intuition for the approach, consider a portfolio whose obligors are evenly split across two sectors. The true default

probability is 1% in the first sector, and 3% in the second. Now assume that an analyst uses the default experience of this portfolio to evaluate a model that posits a uniform default probability of 2%. If the test is based only on the average default rate of the entire portfolio, or random subsets thereof, the inadequacy of the model will not be revealed because the expected default rate will be 2% in either case. If the analyst conjectures that the default probability differs across sectors, she could examine the default experience of single sector subportfolios. She would then be in a much better position to identify the inadequacy of the model.

The related literature includes Nickel, Perraudin and Varotto (2001), who use two different credit risk models to predict the credit risk of a large portfolio of dollar-denominated eurobonds. The authors compare the predictions to the observed losses, but do not conduct a formal test of the models' validity. Carey (1998) and Carey (2001) discuss various resampling strategies for constructing expected loss distributions from a default history. Carey (2001) uses the Moody's database (1970-98) to simulate credit portfolios in order to evaluate the relevance of several dimensions of credit risk. Carey (1998) performs a similar task on the database of the Society of Actuaries (1986-92). Gordy (2000) and Kiesel, Perraudin and Taylor (2001) use stylized portfolios to study how risk measures vary across different portfolio types. Crouhy, Galai and Mark (2000) and Gordy (2000) compare risk measures calculated for the same portfolio but using different models. Sobehart, Keenan and Stein (2000) propose techniques for assessing the quality of individual default rate estimates, an important input to credit risk models. A useful summary of available credit risk models is given in Crouhy, Galai and Mark (2000).

Besides being related to the credit risk literature, our paper also builds on the literature on the evaluation of density forecasts: Clements and Smith (2000) compare

the performance of models to forecast macroeconomic variables. They compare three different validation techniques: the approach of Diebold, Gunter and Tay (1998)¹, Berkowitz (2001) and a normality test recommended by Doornik and Hansen (1994). The authors suspect that the Berkowitz (2001) test and the normality test might be sensitive to outlier observations. De Gooijer and Zerom (2000), however, cannot confirm this conjecture.

The paper is organized as follows. Section 2 describes the framework for the evaluation of test procedures. Section 3 discusses the tests proposed by Lopez and Saidenberg (2000). Section 4 presents our proposals and assesses their power using Monte Carlo simulations. Section 5 concludes.

2. Framework for the evaluation of test procedures

A natural way for evaluating the power of test procedures is to employ a Monte Carlo study. We simulate a large number of artificial credit histories that are all generated by one specific credit portfolio risk model. We then state the null hypothesis that the history is governed by some model specification, choose a significance level, and apply a statistical test separately for each simulated history. The performance of the test is judged by two criteria: if the H_0 -model is the one that has generated the history, the rejection frequency should equal the chosen significance level, i.e. the size of the test. If the H_0 -model is incorrect, the rejection frequency, i.e. the power of the test, should be as large as possible.

¹ Diebold, Gunter and Tay (1998) propose to use the probability integral transform to transform observed data into a series of iid $U(0,1)$ distributed variables under the true model. The independence assumption and the uniformity assumption can be tested together or separately. The authors argue for a separate test and graphical methods in order to identify the source of a possible deviation.

We examine models that capture correlations in credit events through latent variables. Following Merton (1974), these latent variables are usually thought of as the firms' asset values. In the option-theoretic approach of Merton, a firm defaults if its asset value falls below a critical threshold defined by the value of liabilities. Asset value correlations thus translate into correlations of credit quality changes. Such models have been investigated by, among others, Gordy (2000), Lucas et al. (2001) and Frey and McNeil (2001). The asset value approach to modeling portfolio credit risk underlies the risk weights proposed by the Basel Committee on Banking Supervision (2001) as well as industry models such as CreditMetrics and KMV PortfolioManager.²

We examine two variants, which differ in their complexity:

- (i) We neglect both migration risk and recovery rate uncertainty. Recovery rates are assumed to be zero for all loans. In consequence, the loss distribution is fully described by the distribution of the number of defaults within a portfolio. The rationale for choosing a two-state model is that it poses little data requirements, and so lends itself more easily to empirical tests. Many banks do not mark to market their loan positions, or did not do so until recently. Also, consistent data on recovery rates may not be available. By contrast, most banks should be able to collect the number of defaults that occurred in the recent past. Note, too, that the risk weights proposed by the Basel Committee are based on such a two-state model.³

² See Gupton, Finger and Bhatia (1997) for a description of CreditMetrics, and Crouhy, Galai and Mark (2000) for a comparison of the KMV and CreditMetrics models.

³ Basel Committee on Banking Supervision (2001), S. 36.

- (ii) We derive the full distribution of portfolio losses by accounting for the risk of default, the risk of migration, and both systematic and unsystematic recovery risk. As in the previous literature, we neglect general interest rate risk and specific spread risk in order to focus on the risk from credit events.

In a two-state world, available credit portfolio risk models like CreditRisk⁺, CreditMetrics, KMV PortfolioManager or CreditPortfolioView are similar in structure and produce almost identical outputs when parameterized consistently.⁴ For this reason, we are confident that our results are applicable to a broad range of credit risk models. Even though we restrict the analysis to one particular class of portfolio credit risk models, we will nevertheless speak of various ‘models’ we are going to evaluate. In the following, the term ‘models’ will thus refer to different parameterizations of the basic latent variable approach.

In our framework asset value changes $\Delta\tilde{A}_i$, depend on only one systematic factor \tilde{Z} (e.g. the growth rate of the economy) and idiosyncratic factors $\tilde{\varepsilon}_i$.⁵

$$\Delta\tilde{A}_i = w_i\tilde{Z} + \sqrt{1-w_i^2}\tilde{\varepsilon}_i, \quad (1)$$

where \tilde{Z} and $\tilde{\varepsilon}_i$ are iid $N(0,1)$, as is the asset value change $\Delta\tilde{A}_i$. A borrower defaults whenever $\Delta\tilde{A}_i < \Phi^{-1}(p_i)$, where p_i is the unconditional default probability and $\Phi(\cdot)$ denotes the cumulative standard normal distribution function. For a given realization of the systematic factor Z the conditional default probability $p_i | Z$ equals

⁴ Cf. Finger (1998), Koyluoglu and Hickman (1998), and Gordy (2000).

⁵ The extension to a multi-factor model is straightforward.

$$p_i | Z = \text{Prob} \left(\varepsilon_i \leq \frac{\Phi^{-1}(p_i) - w_i Z}{\sqrt{1 - w_i^2}} \right) = \Phi \left[\frac{\Phi^{-1}(p_i) - w_i Z}{\sqrt{1 - w_i^2}} \right]. \quad (2)$$

The factor loadings w_i determine asset correlations. In the case of a uniform loading, $w_i = w$ for all i , the asset correlation is equal to w^2 for all pairs of borrowers. Default correlations can be calculated via the bivariate normal distribution.⁶ We also examine a case where the factor \tilde{Z} follows an autoregressive process, rather than being iid. Even though general credit risk is likely to be cyclical in practice, assuming the factor to be uncorrelated seems to be more appropriate when it comes to evaluating actual credit risk models used by banks. In accordance with bank practice in assigning internal ratings (Carey and Hrycay, 2001), the Basle Committee on Banking Supervision (2001) proposes that default probability estimates reflect the current status of a borrower. In terms of the model, this means that default probability estimates for period t are conditioned on information about the realizations of \tilde{Z} up to t . Any predictability in general credit conditions would thus be accounted for by default probability estimates. The case is different when default probability estimates are based on agency ratings. Rating agencies typically employ a through-the-cycle approach, that is, intentionally neglect cyclical variation in credit quality (see Carey and Hrycay, 2001).

Since Gordy (2000), Lucas et al. (2001) and Frey and McNeil (2001) show that the multivariate normal assumption for asset returns is critical for the results, we will also investigate a case in which asset returns follow a t-distribution. The t-distribution converges to the normal as the degrees of freedom approach infinity, which means

⁶ Cf. Finger (1999), Koyluoglu and Hickman (1998), and Belkin, Suchower and Forest (1998b) for applications of this model.

that choosing the shape of the distribution is one step in parameterizing the asset value model (1).

For both two-state models (i) and the general multi-state models (ii) we need to specify the factor sensitivity w_i and the distribution of the common factor. For a two-factor model, all we need in addition is the individual default probabilities p_i . To model the full loss distributions, we assume R rating categories; category R corresponds to default. The probability of moving from category k to category l is given by p_{kl} . The portfolios we analyze contain only simple, fixed-rate loans. The initial maturity of each loan is set to five years. The value effects of rating transition are derived from assumptions on rating-specific zero yields. We set the annual coupon rate such that loans are valued at par at the beginning of horizon. At the end of horizon, the position is revalued using implied forward rates, taking into account that interest has accrued, and that maturity has decreased to four years. In the case of default, we take R_i , the recovery rate of loan i , to be a fraction of the principal. Frye (2000) shows that recovery risk has both idiosyncratic and systematic components. We therefore follow Frye (2000) and model recovery rates as

$$R_i = m_R + s \left(q_i \tilde{Z} + \sqrt{1 - q_i^2} \tilde{\omega}_i \right) \quad (3)$$

where m_R is the mean recovery rate assumed for the loans. \tilde{Z} is the common factor from (1); it introduces systematic recovery risk. Idiosyncratic recovery risk is modeled through the component $\tilde{\omega}_i$, which is iid $N(0,1)$. The factor sensitivities q_i determine the relative importance of systematic and unsystematic recovery risk, while the parameter s determines the overall magnitude of recovery risk. With this formulation, recoveries are normally distributed, meaning that they can fall below zero. This seems unproblematic for large portfolios such as the ones analyzed in this paper,

where the realized mean recovery rate is unlikely to become negative.

Having parameterized the models, the loss distributions are derived using Monte Carlo simulations. A trial involves drawing an asset value for each obligor, according to (1). In a two-state model, default occurs when an asset value falls below a threshold given by $\Phi^{-1}(p_i)$.⁷ For the R -state models, there are $R - 1$ thresholds. They split the asset value distribution into R bins, each of which corresponds to a rating category. The thresholds are chosen such that the probability of falling into a bin is equal to the corresponding probability of rating transition, p_{kl} . (A detailed description of the procedure is given in Gupton, Finger and Bhatia (1997).) Depending on the bins in which simulated asset values fall, loans are mapped into ratings, and revalued. In case of default, a random recovery rate is drawn according to (3). One trial produces one scenario for portfolio losses; the distribution of portfolio losses is constructed from a multitude of scenarios. We conduct Monte Carlo simulations with 1,000,000 trials to ensure that the simulation error is negligible.

In the following, we describe the parameters used for most of the analyses; we refer to this set of assumptions as the base case. In the base case, we consider two-state models with zero recovery in the case of default. The portfolios are homogeneous in terms of default probability, asset correlations, and loan size. We assume that the available data sets comprise ten years of annual data on the number of defaults within homogeneous portfolios of 10,000 borrowers. We set the unconditional annual

⁷ If the portfolio is homogeneous, a quicker way to perform the simulations is i) draw $N(0,1)$ -distributed random numbers for the factor realizations, ii) calculate the conditional default probability, and iii) draw the number of defaults from a binomial distribution given the number of loans and the conditional default probability. The closed-form solution of Vasicek (1997) holds quite well for the portfolio sizes we use in this paper, but there are some discrepancies when asset correlations are small (e.g. 0.5%).

default probability equal to 1% for each obligor, and assume that the common factor is serially uncorrelated. Asset values follow a standard normal distribution. The asset correlation parameter is the only one that is varied in the base case. For the model that generates the simulated default histories, we use a uniform asset correlation of $w^2 = 5\%$ for all pairs of borrowers. In the alternative models, we vary the asset correlation in the range $w^2 \in [0\%, 20\%]$.

A test's power is assessed based on 10,000 10-year default histories, generated independently from the true credit risk model. In most cases, the size of the test is chosen to be 10%. A size of 5% or 1% may be more common in other settings, but we believe that the data problems associated with the evaluation of credit risk models will make evaluators choose a larger size to increase the power. The base case assumptions are summarized in Table 1. They will be varied to check the robustness of the results.

3. The proposal of Lopez and Saidenberg

The main problem when evaluating credit risk models is the scarcity of data in the time dimension. Lopez and Saidenberg (2000) suggest cross-sectional resampling techniques to increase the power of evaluation procedures. Given a credit data set covering T years of data for N_t loans, a large number R of subportfolios is simulated within each year t . Lopez and Saidenberg suggest to draw the loans of a subportfolio without replacement from all loans contained in the portfolio in year t .⁸ They also

⁸ Lopez and Saidenberg (2000) characterize their procedure as drawing with replacement, referring to the fact that all loans drawn for a subportfolio are put back into the pool before the next subportfolio is resampled. The technical description they provide on p. 158, however, makes it clear that the loans for a single subportfolio are drawn without replacement from the loan pool. That is, each loan can enter a

recommend to draw 'large' subportfolios, but do not discuss this issue in detail. For each subportfolio, the loss distribution is forecasted conditional on information available at the end of the previous year.⁹ Various statistical tests can then be used to examine whether the predictions are accurate. In a sense, the number of observations available for model evaluation is thus multiplied by the factor R .

The following example illustrates the procedure as well as a difficulty associated with it. For a credit portfolio with 10,000 borrowers, we have a default history for the past ten years:

	Year 1	Year 2	...	Year 10
Portfolio defaults	200	100	...	50

Assuming an unconditional annual default probability of 1%, the number of defaults is high in the first year, average in the second and low in the last. For each out of the ten years we randomly draw 1,000 subportfolios S_i , with 1,000 borrowers each. We count the number of defaults in each subportfolio:

	Year 1	Year 2	...	Year 10
Defaults in S_1	18	9	...	6
Defaults in S_2	20	13	...	5
...
Defaults in $S_{1,000}$	22	7	...	5

In their proposal, Lopez and Saidenberg effectively treat all observations from this matrix as if they were independent. However, if overall portfolio defaults are high, as in year one, the number of defaults in resampled subportfolios will be high as well.

specific subportfolio only once.

⁹ We will comment on conditioning the prediction on contemporaneous observations not included in a specific subportfolio.

Similarly, the low number of portfolio defaults in year ten shows in subportfolios drawn from that year. Obviously, defaults in the 1,000 subportfolios resampled from one year's default experience are not independent. Such cross-sectional dependencies pose a problem for all tests that compare loss predictions with observed losses. This intuition is confirmed by the formal analysis in the appendix. There we show that the unbiasedness tests proposed by Lopez and Saidenberg neglect correlation in subportfolio defaults. To examine whether cross-sectional dependence is relevant in practical applications of the Lopez and Saidenberg procedure, we conduct Monte Carlo simulations. Specifically, we implement the quantile test proposed by Lopez and Saidenberg (2000, p. 160) on simulated data sets.

Under the assumption that the predicted quantiles of the loss distribution are accurate and observed violations of the quantiles are independent, these violations are draws from a binomial distribution. Whether or not the percentage of observed violations $\hat{\alpha}$ is equal to the chosen confidence level α can be tested using the likelihood ratio statistic

$$LR(\alpha) = 2 \left[\log(\hat{\alpha}^y (1 - \hat{\alpha})^{T \cdot R - y}) - \log(\alpha^y (1 - \alpha)^{T \cdot R - y}) \right], \quad (4)$$

where y is the number of violations across the $T \cdot R$ subportfolios. The statistic is referred to the chi-squared distribution with one degree of freedom.

In the simulations, we apply this test to validate credit risk models that differ only in their asset correlation w^2 (all parameters as in Table 1). For the test, we use the 90%-quantile, i.e. $\alpha = 10\%$, and proceed as follows:

1. Simulate a 10-year default history using the true model with an asset correlation of $w^2 = 5\%$.

2. Draw 1,000 random subportfolios for each year as proposed by Lopez and Saidenberg: “generate the $(N \times 1)$ vector w_i which is the set of portfolio weights for resampled portfolio i , by generating N independent draws from the uniform distribution over the interval $[0,1]$. For each draw above ρ , the associated credit is assigned a weight of zero and is not included in the resampled portfolio. For each draw below ρ , the associated credit is assigned a weight of one and is included in the resampled portfolio (Lopez and Saidenberg, 2000, p. 158).” We choose ρ to be 0.2, 0.5 or 0.8, corresponding to an average subportfolio size of 2,000, 5,000 and 8,000, respectively.
3. For each resampled portfolio, use Monte Carlo simulations (1,000,000 trials) to determine the 90% quantile of defaults predicted by the model under analysis. (We thus take into account that the quantile depends on the number of loans contained in a subportfolio.)
4. Implement the LR-test (4) for a specific credit risk model by counting the number of violations of the predicted 90% quantile of defaults.
5. Repeat steps 1. - 4. 10,000 times.

In step 2 we depart from Lopez and Saidenberg (2000, footnote 8) in that we do not avoid having the same subportfolio composition more than once. First, it is unclear why a repetition should bias the test. Second, the probability of having no repetition is close to unity. If the subportfolio size is 8,000 (and constant), for example, the probability that, within each of the 10 years, each of the 1,000 subportfolios is

different is equal to 0.999997.¹⁰

The results are summarized in Table 2.¹¹ Since we use a test size of 10%, the credit risk model with the true asset correlation $w^2 = 5\%$ ought to be rejected with a relative frequency of 10%. Yet, depending on the subportfolio size these numbers vary between 74% and 90%. The intuition for the results is that, by assuming independence across simulated subportfolios, the test overestimates the information contained in the data. In consequence, the test is biased towards rejection. Note that the true model is not rejected because the predicted subportfolio quantiles are inaccurate on average; this can be verified by counting the average number of subportfolio exceptions across all simulated data sets, which is 10% for the true model.

A quick check of the numbers can be performed by examining the extreme case in which the average subportfolio size is 9,999. The subportfolios, being drawn without replacement, are then almost identical to the overall portfolio, and the violation frequency across the subportfolios will more or less be equal to the one in the original data. With $T=10$, $R=1,000$ and a size of 10%, the test (4) rejects the null whenever the observed frequency of violations is lower than 9.5% or larger than 10.5%. To pass the test, the frequency of violations in the ten years of the original data thus has to be close to the interval [9.5%, 10.5%]; but the only way of getting close is to have exactly one violation in the ten years. This happens with a probability of

¹⁰ The figure obtains through $\left[\prod_{i=0}^{999} \left(\frac{\binom{10000}{8000} - i}{\binom{10000}{8000}} \right) \right]^{10}$.

¹¹ The test statistic (3) is not defined if there are no violations across all subportfolios, but it is obvious that the model should be rejected. (With independent binomial draws, the probability of observing no violations if the sample size is 10,000 and the probability of a violation is 0.1 is less than 10^{-300} .)

$10 \times 0.1 \times 0.9^9 = 36.7\%$. With a subportfolio size of 9,999, the rejection frequency would therefore be close to $1 - 36.7\% = 63.3\%$. In Table 2, the rejection frequency is generally higher than 63.3%, and it increases when we reduce the subportfolio size. Since we draw without replacement, the randomness in subportfolio defaults that is introduced through resampling increases with decreasing subportfolio size. This increase in variation reduces the probability that the overall frequency of exceptions falls within the narrow interval [9.5%, 10.5%], and thereby leads to higher rejection frequencies.

We conclude that the cross-sectional dependence across subportfolios is not only of theoretical concern. It can severely affect the performance of test statistics proposed by Lopez and Saidenberg (2000). Note that each of the tests proposed by Lopez and Saidenberg presupposes independent draws. We therefore caution against the application of any of these tests.

One might think of modifying the procedure by conditioning the forecasts of subportfolio defaults on the default experience of those borrowers that are not included in this specific subportfolio. While this might be a valid and useful procedure in some cases, it would fail to detect false models in others. For example, it would be impossible to discriminate between one-factor models that differ in the value assumed for the factor sensitivity w . The intuition is as follows: default correlation arises through variations in the conditional default rate. Using conditional default rates instead of unconditional ones amounts to purging the default data of default correlation, making it impossible to discriminate between two simple one-factor models which differ in their assumptions about correlation. Another possible modification of the procedure is to draw subportfolios *with* rather than *without* replacement. This would not eliminate the problem of cross-sectional correlation

across subportfolios.

The data sets in our simulations cover ten years. Increasing the sample length would reduce the documented biases as the dependence brought about by cross-sectional resampling would be mitigated by a larger number of independent observations across time. Even if the tests were asymptotically valid, however, they would gain little appeal. Asymptotically, that is, for an increasing sample length T , the cross-sectional information that the tests are meant to exploit loses importance.

4. Evaluating credit risk models based on the entire distribution

Lopez and Saidenberg aimed at increasing the number of observations, assuming that existing approaches are inadequate for sample sizes typically available. In this section we show that it is possible to design powerful tests if we use the information of the complete loss distribution.¹²

We recommend the Berkowitz (2001) test procedure. In this approach, the loss history is transformed so that one obtains a series of standard normally distributed variables when using the correct credit risk model. Standard tests can be performed to test this characteristic.

Berkowitz (2001) applies a simple twist to the so-called Rosenblatt (1952) transformation of observed data. First, the estimated cumulative distribution function $\hat{F}(\cdot)$ is applied to observed losses

¹² Simple quantile tests as in (4) are of little use if the sample size is small. This is intuitive for the case where the H_0 distribution is riskier than the true one. The number of violations will be smaller than expected; in the extreme, there will be no violation at all. With only ten observations, however, observing no quantile violation is not sufficient evidence (at the 10% significance level) for rejecting the H_0 if one tests for violations of the 90%, 95% or 99% quantiles.

$$x_t = \hat{F}(y_t) = \int_{-\infty}^{y_t} \hat{f}(u) du, \quad (5)$$

where y_t are observed losses and $\hat{f}(u)$ is the forecasted probability of a loss of u . If the estimated loss distribution is equal to the true one, the transformed variable x_t is iid $U(0,1)$, where $U(\cdot)$ denotes the uniform distribution.

In a second step, Berkowitz suggests to apply another transformation using the inverse of the standard normal distribution function $\Phi(\cdot)$:

$$z_t = \Phi^{-1}(x_t) \quad (6)$$

If the predicted distribution function is correct, the transformed observations z_t are iid $N(0,1)$.¹³ Berkowitz recommends using a likelihood ratio test for testing whether the series z_t is serially uncorrelated with mean zero and unit variance. In the following, we apply such tests to simulated credit loss data in order to assess their power.

4.1 Two-state models

4.1.1 Alternative models differ in asset correlation assumption

In the base case, we compare asset value models with one systematic factor and a uniform mutual asset correlation (all parameters as in Table 1). The asset correlation of the true model equals $w^2 = 5\%$. We define different null hypotheses by changing the correlation parameter w^2 on the interval $[0\%, 20\%]$.

The test statistic is calculated based on the log-likelihood function of the univariate normal distribution for the transformed variable z_t :

¹³ See Berkowitz (2001) for a proof.

$$\log L = -\frac{T}{2} \log 2\pi - \frac{T}{2} \log \sigma^2 - \sum_{t=1}^T \frac{(z_t - \mu)^2}{2\sigma^2}, \quad (7)$$

where T is the number of years. Since both the true model and the H_0 do not exhibit serial correlation, we do not need to test for it in this case. The maximum likelihood estimators for the mean and variance of the transformed variable are given by

$$\hat{\mu}_{ML} = \frac{\sum z_t}{T};$$

$$\hat{\sigma}_{ML}^2 = \frac{\sum (z_t - \hat{\mu}_{ML})^2}{T}. \quad (8)$$

The LR-test is then structured to test the joint hypothesis that the z_t have zero mean and unit variance. It is given by

$$\lambda = 2 \left[\log L(\mu = \hat{\mu}_{ML}, \sigma^2 = \hat{\sigma}_{ML}^2) - \log L(\mu = 0, \sigma^2 = 1) \right] \quad (9)$$

The statistic is referred to the chi-squared distribution with two degrees of freedom.

Figure 1 shows the simulated power of the test statistic in the base case. If the null hypothesis posits a zero default correlation, it is rejected in 100% of all cases. For models that are close to the correct 5%, the power is lower. However, it is larger than 50% if the assumed correlation is below 2.5% or above 10.5%.

If the null hypothesis coincides with the true model, the power equals 12%, which is slightly higher than the significance level of 10%. Due to the small sample size, the test statistic is not exactly chi-squared distributed. The inaccuracy seems to be small, and is probably acceptable in many practical applications. It could be eliminated by simulating the critical values for the test statistic.

The results depicted in Figure 1 are also shown in column three of Table 3, along with some additional information that puts them into perspective. The second column

contains the 99% quantiles of the loss distribution under the various null hypotheses to illustrate how different these distributions are from the true model. Columns 4-9 of Table 3 report the simulated power when the size of the test, the available database, or the portfolio structure is changed. We examine the following, non-accumulating variations:

- we use a significance level of 5% instead of 10%
- the portfolio contains loans to 1,000 or 5,000 borrowers, respectively (instead of 10,000)
- the available history comprises only five years instead of ten
- the default rate is 0.5% instead of 1%
- the portfolio is heterogeneous in terms of default probabilities. Rather than assuming a uniform default rate of 1% we split the portfolio into seven rating classes (Table 4). The structure is based on the high quality credit portfolio in Gordy (2000). Compared to the Gordy portfolio, we adjust the number of obligors in rating classes A and B to achieve a mean default rate of 1%.

As should be expected, the power decreases if we lower the size of the test, increase idiosyncratic risk by lowering the number of obligors in the portfolio, shrink the available data history, or lower the default rate. The loss of power is fairly small when the number of borrowers is 5,000 instead of 10,000. With 1,000 borrowers, the power is still above 75% in some cases. The same holds when the chosen size of the test is 5% instead of 10%, or when the number of years in the observed default history is five instead of ten. With heterogeneous default rates, the power decreases modestly.

Is the documented power of the tests satisfactory? One of the most pressing questions in parameterizing credit risk models is to choose an appropriate value for the asset correlation. While the Basel Committee on Banking Supervision (2001)

favors an asset correlation of 20%, calibration exercises (cf. Gordy, 2000) typically lead to much lower correlation estimates.¹⁴ Often, the estimates are smaller than 5%. In Table 3, the probability of rejecting an asset correlation of 20%, if the correct one is 5%, ranges from 74% to 97%. Such rejection rates appear to be satisfactory.

Contrary to the base case, estimates of default probabilities will be noisy in practice, and one might suspect that this reduces the power of detecting misspecifications of the asset correlation. We therefore examine a case in which the risk model not only falsely assumes an asset correlation of 20% but is also misspecified with respect to the default probabilities. The true default probabilities are those of the heterogeneous portfolio from above (see Table 4). Under H_0 , we underestimate the default probability by 50% for one half of the borrowers of each rating class, and overestimate it by the same percentage for the other half.¹⁵ Recall that the test's power equals 93% when the heterogeneous default probabilities are correctly specified (see Table 3). When we introduce noise the power decreases slightly to 90%. This suggests that the results presented above are robust to the introduction of estimation error.

4.1.2 Alternative models differ in parameters other than the asset correlation

So far, we have illustrated the power of rejecting models that diverged from the true model in their assumptions about asset correlations. In the following, we present some results on the test's power if other elements of the parameter space are misestimated. We start by examining a situation in which the models to be tested

¹⁴ The asset correlations are calibrated to match the observed default rate volatility. Recently, the Basel Committee proposed to use asset correlations between 10% and 20%.

¹⁵ For example, the H_0 default probabilities for obligors rated BB are 0.53% or 1.59% instead of 1.06%.

differ from the true model only with respect to the unconditional default probability. As before the true default probability is 1%, while the default rates assumed under the null hypotheses span from 0.2% to 2.4%. The other variables are set as in the base case (uniform correlation of 5%, 10,000 borrowers per year, ten observations). The simulated power is presented in Table 5.

When comparing the power to the previous results, it is illustrative to compare null hypotheses that produce similar errors in predicting extreme losses, e.g. the 99% quantile. The true model is the same in both setups. An asset correlation of 5% and a default probability of 1.6% lead to roughly the same 99% quantile as an asset correlation of 10% and a default probability of 1%. In the latter case, the power is 44% (see Table 3), while it amounts to 74% in the former case. Contrary to a false correlation assumption, missing the default probability also leads to a wrong prediction of the mean default rate. Since the Berkowitz test utilizes the entire distribution rather than focusing on extreme events, this explains the observed differences in power.

Even if default probabilities and asset correlations are correctly specified, a credit risk model can still be a poor predictor of credit losses. Lucas et al. (2001) and Frey and McNeil (2001) document that the distribution of the latent variable heavily influences the probability of extreme events. Until now we followed the standard approach and assumed the latent variable to be normally distributed. A more general specification is to model the latent variables as following a t-distribution. Since the t-distribution is a continuous mixture of normal distributions, where the mixing distribution is the chi-squared, this can be achieved by transforming the asset value changes as follows (see Frey and McNeil, 2001):

$$\Delta \tilde{A}'_i = \sqrt{\frac{\nu}{\tilde{w}}} \Delta \tilde{A}_i, \quad \tilde{w} \sim \chi^2(\nu), \quad (10)$$

where ν denotes the degrees of freedom assumed for the t-distribution. The distribution approaches the normal as ν approaches infinity. A borrower defaults when $\Delta \tilde{A}'_i < t_\nu^{-1}(p)$, where p is the unconditional default probability and t_ν is the cumulative t-distribution with ν degrees of freedom. For the simulation experiments, we choose $\nu = \infty$ to describe the true model, and vary the degrees of freedom assumed under the null hypothesis. In Table 5, it can be seen that the test's power is larger than 50% if the degrees of freedom under H_0 are less than forty. An example shall help to assess the power. The standard approach in credit risk modeling is to assume that latent variables are normally distributed. One piece of evidence against this assumption is the observed leptokurtosis of stock returns. The excess kurtosis of the S&P 500 index, for example, is 0.74 when computed with annual log returns from 1971 to 2000. This could lead a risk manager to favor a t-distribution with twelve degrees of freedom because then the excess kurtosis would be 0.75. If the normal assumption is correct, and there are ten years of credit data to check whether a t-distribution with twelve degrees of freedom is appropriate, the power is close to 100%. Conclusions do not change when we look at the opposite case in which the true asset value distribution is a t-distribution. For example, if the true asset value distribution is a t with ten degrees of freedom and we test the null hypothesis that the asset value distribution is normal, the test's power equals 99.6% (all other parameters as in the base case; the asset correlation equals 5%).

Finally, we modify the base case by introducing autocorrelation into the time series of the systematic factor \tilde{Z} . In simulating the default histories, we use the following

autoregressive process for \tilde{Z}_t :

$$\tilde{Z}_t = 0.5\tilde{Z}_{t-1} + 0.866\tilde{u}_t, \quad \tilde{u}_t \sim N(0,1), \quad \tilde{Z}_1 \sim N(0,1) \quad (11)$$

The choice of parameters is based on the study of Belkin, Suchower and Forest (1998a), who fit such a process on rating transition matrices and obtain an autocorrelation coefficient of 0.46. A credit risk model should incorporate such autocorrelation, that is, take the current position in the credit cycle into account when predicting default rates. Evaluators should thus be interested in testing whether the prediction errors are indeed uncorrelated across time. As in Berkowitz (2001), we augment the density function for the transformed losses z_t by allowing them to follow a first-order autoregressive process:

$$\begin{aligned} \log L = & -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \left[\frac{\sigma^2}{1-\rho^2} \right] - \frac{(z_1 - \mu/(1-\rho))^2}{2\sigma^2/(1-\rho^2)} - \frac{T-1}{2} \log 2\pi \\ & - \frac{T-1}{2} \log \sigma^2 - \sum_{t=2}^T \frac{(z_t - \mu - \rho z_{t-1})^2}{2\sigma^2} \end{aligned} \quad (12)$$

As the estimator for the autocorrelation coefficient ρ is downward biased in small samples (cf. Quenouille, 1949 or Andrews, 1993), we first use Monte Carlo simulations to identify the bias. If the null hypothesis is correct and there are ten observations as in the base case, the median maximum likelihood estimator of ρ equals -0.114. We therefore test the restrictions $\mu=0$, $\sigma^2=1$ and $\rho=-0.114$.¹⁶ The statistic is referred to the chi-squared distribution with three degrees of freedom.

A simulation study, where we set all parameters (except for the autocorrelation) as in

¹⁶ In practical applications, one will have to determine the bias associated with the number of observations at hand. Using the mean bias (-0.108) instead of the median for defining the restriction does not change the results significantly.

the base case, produces the following result: if the factor is governed by the process described in (11), but the null hypothesis assumes that there is no autocorrelation, the probability of rejecting the null is 38%. The figure is rather low, which is not surprising given that there are only ten time periods to estimate the autocorrelation.

Should one nevertheless routinely test for autocorrelation? To answer this question, it is interesting to know whether testing for autocorrelation can actually decrease the power of the test. We use the base case setup, that is, a situation where neither the true model nor the H_0 models contain autocorrelated factors. If the H_0 posits an asset correlation of 10% (true being 5%), the power is 44% if we do not test for autocorrelation. The figure drops to 35% once the test includes the restriction $\rho = -0.114$. If one routinely tests for serial correlation, it might therefore be advisable to conduct parallel tests that exclude serial correlation.

4.1.3 Alternative tests

Under the null hypothesis, the transformed variables should be standard normally distributed. Following Berkowitz (2001), however, we only tested whether they have mean zero and unit variance. One could presume that the power of the test could be increased by testing for normality as well. To check whether this is indeed the case, we perform two additional tests. First, we test the transformed variable z_t for normality using the test described in Doornik and Hansen (1994). The test is based on skewness and kurtosis, but transforms these statistics in order to improve the small-sample performance of the test.

Second, we use an alternative testing procedure that will typically not be feasible, but

provides a useful benchmark in the example considered here.¹⁷ In the base case, the only unknown parameter was the factor sensitivity w . Using the original, untransformed default data we can determine a maximum likelihood estimate for w :

$$\hat{w}_{ML} = \arg \max_w \sum_{t=1}^T \log[f_w(u_t)], \quad (13)$$

where $f_w(u_t)$ is the density function of portfolio defaults u for a specific factor sensitivity w . Maximization is done through a simple search procedure in which we evaluate the likelihood for each correlation assumption $w^2 \in [0\%, 0.5\%, \dots, 100\%]$. This estimate can be used to construct a standard likelihood ratio test against a specific H_0 .

$$\lambda_{Alt} = 2[\log L_{\hat{w}_{ML}} - \log L_{w_{H_0}}] = 2\left[\sum_{t=1}^T \log f_{\hat{w}_{ML}}(u_t) - \sum_{t=1}^T \log f_{w_{H_0}}(u_t)\right] \quad (14)$$

Asymptotically, the statistic will be distributed chi-squared with one degree of freedom. Since we cannot rely on the asymptotic properties to hold, we simulate the distribution under H_0 and obtain critical values from this simulated distribution.

We simulate the power of the Doornik-Hansen normality test as well as that of the standard maximum likelihood test λ_{Alt} . Results for the base case setting are shown in Figure 2. To facilitate comparison, the graph also contains the power curve of the Berkowitz test already shown in Figure 1. The power of the normality test is very low; the power of the standard likelihood test λ_{Alt} is not substantially higher than when

¹⁷ Typically, the number of free parameters will be too large to estimate them based on aggregate portfolio data. We also used a random effects probit model to determine maximum likelihood estimates of w , but found them to suffer from a small sample bias.

testing the transformed variables for a mean of zero and a variance of one. The evidence supports the view that we do not lose significant information by (i) transforming the loss data and (ii) testing only a subset of the restrictions that the transformed data should obey if the null hypothesis is correct.

4.2 Multi-state models

In this section, we analyze the power of the Berkowitz test when applied to credit risk models that incorporate migration and recovery rate uncertainty. We take the heterogeneous portfolio from above (see Table 4). The probabilities of rating transition are shown in Table 6; they are taken from Lando and Skodeberg (2001, Table 3). As described in Lando and Skodeberg, the estimates are based on continuous rating data from Moody's over the period 1988-98. They are preferable to other available estimates, which average transition frequencies observed within discrete time intervals, and thus do not make efficient use of the data. We analyze portfolios of simple, fixed-rate loans with an initial maturity of five years. The yields necessary for loan valuation are taken from the CreditMetrics web site. We use the yield spreads for bonds of US corporates, and the yield on US treasuries. Required yields that are not provided on the web site (4-year yields), are obtained through linear interpolation. Coupon rates are set such that loans are initially valued at par. The conditional one-year ahead loan values are computed using implied forward yields; they are reported in Table 7. The mean recovery rate m_R is set to 0.521, which equals the mean bank loan value in default for senior unsecured loans in Gupton, Gates, and Carty (2000). The parameters describing recovery rate uncertainty are taken from the estimates in Frye (2000): the volatility of individual recovery rates s is set to 0.32; the correlation of recovery rates is set to 2.89%, which corresponds to $q=0.17$ in equation (3).

Figure 3 displays the power of the Berkowitz test for the multi-state case with systematic recovery rate risk. As in the base case, the asset correlation of the true model equals $w^2 = 5\%$. The null hypotheses are defined by changing the correlation parameter w^2 on the interval $[0\%, 20\%]$. The test's power reaches almost 100% if the null hypothesis specifies a zero asset correlation, and 69% if the asset correlation under the null hypothesis is set to 20%. Comparing these results with Table 3, it can be seen that incorporating migration and recovery rate uncertainty reduces the test's power.

Since the test's power depends on the difference between the correct loss distribution and the one under H_0 , a closer look at these distributions helps to explain the results. We compare unexpected losses, which we define as the 1% quantile of portfolio value minus expected portfolio value. In the multi-state analysis from above, an asset correlation of 20% leads to an unexpected loss that is 1.7 times higher than the unexpected loss that obtains with a 5% asset correlation. In the two-state base case (see section 4.1.1) the corresponding ratio is 3.0, which means that misspecifications of the asset correlation have a much stronger impact than in the multi-state case. In consequence, the power of the Berkowitz test is lower in the multi-state case.

4.3 Testing cross-sectional predictions

Consider evaluating a model that assumes a uniform asset correlation across obligors. Using the test procedure described above, the evaluator cannot reject the validity of the model. However, she has some a-priori information indicating that the true correlations differ across obligors. How could she incorporate this information?

As an illustration we get back to two-state models, but change our base case setup

slightly. Instead of assuming a uniform asset correlation of 5% in the true model, we split the portfolio into two equally sized sectors with intra-sector asset correlations of 2% and 9%, respectively:

$$\Delta \tilde{A}_i = w_i \tilde{Z} + \sqrt{1 - w_i^2} \tilde{\mathcal{E}}_i, \quad w_i^2 = 0.02 \text{ for } i \in \text{sector 1}, w_i^2 = 0.09 \text{ for } i \in \text{sector 2} \quad (15)$$

We simulate 10-year default histories using this two-sector model and use the Berkowitz test (9) to check whether we can reject a model that posits a uniform asset correlation of 5%. With a size of 10%, the power is only 16% (Figure 4). This result is due to the fact that the aggregate expected loss distributions of the true model and the null hypothesis are almost identical, even though the sector portfolio distributions differ.

If the evaluator conjectures that factor sensitivities differ across the two sectors, she could form two subportfolios consisting of just one sector and proceed as though she were to test default predictions for two different portfolios. Applying the Berkowitz transformation to the sector defaults yields two series of transformed default data z_t . Since both sectors are subject to the same common factor, actual losses will be contemporaneously correlated. Under the null, they follow a bivariate standard normal distribution, which has the following log-likelihood:

$$\begin{aligned} \log L = & -T \log 2\pi - T \log \sigma_1 - T \log \sigma_2 - \frac{T}{2} \log(1 - \rho_{12}^2) \\ & - \frac{1}{2(1 - \rho_{12}^2)} \sum_{t=1}^T \left[\left(\frac{z_{t1} - \mu_1}{\sigma_1} \right)^2 - 2\rho_{12} \left(\frac{z_{t1} - \mu_1}{\sigma_1} \right) \left(\frac{z_{t2} - \mu_2}{\sigma_2} \right) + \left(\frac{z_{t2} - \mu_2}{\sigma_2} \right)^2 \right] \end{aligned} \quad (16)$$

We obtain maximum likelihood estimators for the parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ and ρ_{12} and construct a likelihood ratio statistic to jointly test the restrictions $\mu_1 = 0, \mu_2 = 0, \sigma_1^2 = 1, \sigma_2^2 = 1$. The statistic is referred to the chi-squared distribution with four degrees of freedom.

Applying this methodology to our example of a one-factor model with two intra-sector correlations of 2% and 9%, ten years of data are sufficient to reject the H_0 of a uniform asset correlation of 5% in 99.6% of all cases. The reason for this substantial improvement is that the correlation parameters are sufficiently different from each other within each sector. We repeat the power calculations for other null hypotheses, which differ in the assumption about the value of the uniform asset correlation. The results are shown in Figure 4. Regardless of the asset correlation assumed under H_0 , the power is close to 100% if the test is based on sector defaults.

The example has shown that the Berkowitz procedure can be extended to test cross-sectional predictions. Since we propose to base the test on judiciously chosen subportfolios, there is no general rule for structuring an evaluation procedure. However, we believe that the choice of subportfolios will often be evident. If one wants to test whether a model is too parsimonious (as in the example) one would split the portfolio into sectors one believes to be different. Similarly, the choice will arise naturally once evaluators have defined a benchmark model for evaluation purposes. In such a case, evaluators would determine the portfolio-split such that the intra-sector differences between the benchmark model and the model under analysis are maximized. If the model default probabilities differ from the benchmark ones, for example, one could form two subportfolios according to whether the difference is positive, or negative.

By extending the bivariate likelihood (16) to the M -variate case, such tests can be based on M subportfolios instead on just two as in the example. Of course, there is a limit to the number of subportfolios one can form because the number of parameters in the likelihood function ($M(M-1)/2 + 2M$) grows faster than the number of usable observations ($M \times T$).

One possible way of exploiting the cross-section without needing a-priori information is to utilize the idea of Lopez and Saidenberg, and apply the Berkowitz procedure to randomly drawn portfolio subsets. There are two problems associated with such an approach. First, we have to account for cross-sectional correlations, which imposes a limit on the number of subportfolios we can draw. Second, drawing random subportfolios means that we hardly ever get extreme subportfolio compositions. If there are two sectors, and we draw a large number of reasonably large subportfolios (say, with 2,000 borrowers each), the probability that we obtain at least one subportfolio that consists only of borrowers of one sector is close to zero.¹⁸ As the above example has shown, such extreme portfolio compositions may have the greatest informational value for the purpose of model evaluation. Even if we obtained some extreme portfolio compositions through resampling, their informational value would be lost by averaging across all subportfolios.

¹⁸ Consider a portfolio of 10,000 obligors, one half of which belongs to one sector, the other half to another. Drawing a subportfolio of 2,000 obligors without replacement, the probability that all obligors belong to single sector is lower than 10^{-300} . By contrast, the probability of obtaining an even mixture of sectors is 2%.

5. Concluding remarks

We have described procedures for evaluating credit risk models. Monte Carlo simulations show that the power of the tests is satisfactory. With ten years of annual data, some of the questions currently debated by credit risk managers can be resolved with a probability larger than 90%. An application of the test procedure could, for example, be to validate the assumptions underlying the new capital adequacy framework (Basel Committee on Banking Supervision, 2001).

A test should meet other criteria than a large power, for instance ease of implementation and general applicability. The tests are computationally simple. They require only the predicted cumulative loss distribution and some elementary transformations. The simplest form of the test, which is based only on aggregate portfolio losses, provides a benchmark that is generally applicable. To exploit additional information contained in the cross-section of defaults, we propose to test the model's prediction for judiciously chosen subportfolios. The subportfolio choice can, for example, be based on a benchmark model favored by the evaluators. Note, too, that the test procedures can be directly applied to models that include any form of risk, including spread risk, interest rate risk and other market risks.

A possible criticism is that the tests are based on the entire range of the distribution, whereas risk managers and regulators are mainly concerned about the probability of extreme events. Why should one thus want to rely on the tests? First, data problems can be so severe that there is no alternative. By using a censored likelihood, the Berkowitz (2001) procedure can be based on the tail of the distribution only. However, if the data set is limited, it may not contain the extreme events necessary to conduct such a test. Second, differences in the tails of two distributions will often go along with predictable differences in the rest of the distribution. If default

correlation is increased, for example, the probability of catastrophe losses rises, but so does the probability of very small losses. A good example in point is the choice of the distribution of the latent variables. Choosing a fat-tailed distribution can have substantial impacts on the probability of extreme credit events. As shown in the paper, ten data points give good guidance on choosing the distribution even though such a small sample will typically not contain the extreme events risk managers are concerned about.

References

- Andrews, D.W.K., 1993, Exactly median-unbiased estimation of first order autoregressive/unit root models. *Econometrica* 61, 139-165.
- Basel Committee on Banking Supervision, 2001, The internal ratings-based approach, Basel.
- Belkin, B., Suchower, S., Forest, L.R. Jr., 1998a, A one-parameter representation of credit risk and transition matrices. *CreditMetrics Monitor*, Third Quarter, 46-56.
- Belkin, B., Suchower, S., Forest, L.R. Jr., 1998b, The effect of systematic credit risk on loan portfolio value-at-risk and loan pricing. *CreditMetrics Monitor*, First Quarter, 17-28.
- Berkowitz, J., 2001, Testing density forecasts with applications to risk management. *Journal of Business & Economic Statistics* 19, 465-474.
- Carey, M., 1998, Credit risk in private debt portfolios. *Journal of Finance* 53, 1363-1387.
- Carey, M., 2001, Dimensions of credit risk and their relationship to economic capital requirements. In: Mishkin, F.S. (ed.), *Prudential supervision: what works and what doesn't*. NBER and UC Press.
- Carey, M., Hrycay, M., 2001, Parameterizing credit risk models with rating data, *Journal of Banking and Finance* 25, 197-270.
- Clements, M.P., Smith, J., 2000, Evaluating the forecast densities of linear and non-linear models: Applications to output growth and unemployment. *Journal of Forecasting* 19, 255-276.

- Crouhy, M., Galai, D., Mark, R., 2000, A comparative analysis of current credit risk models. *Journal of Banking and Finance* 24, 59-117.
- De Gooijer, J.G., Zerom, D., 2000, Kernel-based multistep-ahead predictions of the US short-term interest rate. *Journal of Forecasting* 19, 335-353.
- Diebold, F.X., Gunther, T.A. and Tay, A.S., 1998, Evaluating density forecasts with applications to financial risk management. *International Economic Review* 39, 863-883.
- Doornik, J.A., Hansen, H., 1994, An omnibus test for univariate and multivariate normality. Working paper, University of Oxford, University of Copenhagen.
- Finger, C.C., 1998, Sticks and stones. Working paper, The RiskMetrics Group, New York.
- Finger, C.C., 1999, Conditional approaches for CreditMetrics portfolio distributions. *CreditMetrics Monitor*, First Quarter, 14-33.
- Frey, R., McNeil, A.J., 2001, Modelling dependent defaults. Working paper, University of Zurich, ETH Zentrum Zurich.
- Frye, J., 2000, Depressing recoveries. Working paper, Federal Reserve Bank of Chicago.
- Gordy, M., 2000, A comparative anatomy of credit risk models. *Journal of Banking and Finance* 24, 119-149.
- Gupton, G.M., Finger, C.C., Bhatia, M., 1997, *CreditMetrics – Technical document*, New York.
- Gupton, G.M., Gates, D., Carty, L.V., 2000, Bank loan loss given default, Moody's Investors Service, Global Credit Research, November.

- Kiesel, R., Perraudin, W., Taylor, A. 2001, The structure of credit risk. Working paper, Birkbeck College, Bank of England.
- Koyluoglu, H.U., Hickman, A., 1998, Reconcilable differences. *Risk* 11, No 10, 56-62.
- Lando, D., Skodeberg, T., 2001, Analyzing rating transitions and rating drift with continuous observations. *Journal of Banking and Finance*, forthcoming.
- Lopez, J.A., Saidenberg, M.R., 2000, Evaluating credit risk models. *Journal of Banking and Finance* 24, 151-165.
- Lucas, A., Klassen, P., Spreij, P., Straetmans, S., 2001, An analytic approach to credit risk of large corporate bond and loan portfolios. *Journal of Banking and Finance* 25, 1635-1664.
- Merton, R.C., 1974, On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29, 449-470.
- Nickel, P., Perraudin, W., Varotto, S., 2001, Ratings- versus equity-based credit risk modeling: An empirical analysis. Working paper, Bank of England, Birkbeck College.
- Quenouille, M.H., 1949, Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society B* 11, 68-84.
- Rosenblatt, M., 1952, Remarks on a multivariate transformation. *Annals of Mathematical Statistics* 23, 470-472.
- Sobehart, J.R., Keenan, S.C., Stein, R.M., 2000, Benchmarking quantitative default risk models: a validation methodology. Moody's Investors Service, New York.
- Vasicek, O., 1997, The loan loss distribution. Working paper, KMV Corporation.

Appendix

For one specific test suggested by Lopez and Saidenberg, we formally demonstrate that it neglects cross-sectional dependence in resampled subportfolios. As in the base case, we examine only losses from default; recovery rates are set to zero. Default correlations are modeled via correlated asset values as in (1). Let L_{it+1} denote the losses in subportfolio i drawn from the default experience of year $t+1$. If, conditional on the information at time t , credit risk model m forecasts an expected loss $\hat{\mu}_{mit}$ for subportfolio i in year $t+1$, unbiasedness implies that the prediction errors

$$e_{it+1} = L_{it+1} - \hat{\mu}_{mit} \tag{A1}$$

have mean zero. Lopez and Saidenberg suggest to test this hypothesis through the regression $L_{it+1} = \alpha + \beta \hat{\mu}_{mit} + u_{it+1}$, in which one tests for $\alpha = 0$ and $\beta = 1$. In the following, we show that the errors u_{it+1} are typically not independent across subportfolios.

Assume that the underlying portfolio has N obligors in each year. For all obligors, the unconditional default probability is p ; the sensitivity towards the single common factor Z is w for each borrower. The factor Z is independently distributed across time. Let M denote the number of obligors drawn for subportfolio i in the course of the Lopez-Saidenberg resampling. If the credit risk model is correct, we can rewrite (A1) as:

$$e_{it+1} = L_{it+1} - M p \tag{A2}$$

because, conditional on the information in time t , the prediction of the expected loss is equal to the unconditional default probability times the number of obligors within the portfolio.

Realized portfolio losses in one period are independent draws from a Bernoulli

distribution whose success probability is determined by the factor sensitivity w and the factor realization in that period. For the entire portfolio, losses can be written as

$$L_{t+1} = N \Phi \left[\frac{\Phi^{-1}(p) - wZ_{t+1}}{\sqrt{1-w^2}} \right] + \varepsilon_{t+1} \quad (\text{A3})$$

The error term ε_{t+1} represents idiosyncratic risk. It is independently distributed across time, and has zero mean. For a mean-preserving resampling procedure like drawing with or without replacement, a random subportfolio i has the same expected loss rate as the one suffered in the entire portfolio. The subportfolio loss can thus be written as:

$$L_{it+1} = M \left[\frac{L_{t+1}}{N} \right] + \eta_{it+1} = M \left[\Phi \left[\frac{\Phi^{-1}(p) - wZ_{t+1}}{\sqrt{1-w^2}} \right] + \frac{\varepsilon_{t+1}}{N} \right] + \eta_{it+1} \quad (\text{A4})$$

where η_{it+1} has mean zero and captures the randomness introduced through resampling. It is also independently distributed across subportfolios and time, provided that one puts the borrowers drawn in one replication back into the portfolio, and does not impose conditions on the composition of the subportfolios. Substituting (A4) into (A2) we obtain the following expression for a subportfolio prediction error

$$e_{it+1} = M \left[\Phi \left[\frac{\Phi^{-1}(p) - wZ_{t+1}}{\sqrt{1-w^2}} \right] + \frac{\varepsilon_{t+1}}{N} \right] + \eta_{it+1} - Mp \quad (\text{A5})$$

The covariance between the predictions errors for two subportfolios that belong to the same year and contain the same number of loans M is

$$\text{cov} [e_{it+1}, e_{jt+1}] = \text{cov} \left[M \Phi \left[\frac{\Phi^{-1}(p) - wZ_{t+1}}{\sqrt{1-w^2}} \right] + \frac{M}{N} \varepsilon_{t+1} + \eta_{it+1} - Mp, M \Phi \left[\frac{\Phi^{-1}(p) - wZ_{t+1}}{\sqrt{1-w^2}} \right] + \frac{M}{N} \varepsilon_{t+1} + \eta_{jt+1} - Mp \right]$$

$$\begin{aligned}
&= \text{cov} \left[M\Phi \left[\frac{\Phi^{-1}(p) - wZ_{t+1}}{\sqrt{1-w^2}} \right], M\Phi \left[\frac{\Phi^{-1}(p) - wZ_{t+1}}{\sqrt{1-w^2}} \right] \right] + \text{cov} \left[\frac{M}{N} \varepsilon_{t+1}, \frac{M}{N} \varepsilon_{t+1} \right] \\
&= M^2 \text{var} \left[\Phi \left[\frac{\Phi^{-1}(p) - wZ_{t+1}}{\sqrt{1-w^2}} \right] \right] + \left(\frac{M}{N} \right)^2 \text{var}[\varepsilon_{t+1}] \tag{A6}
\end{aligned}$$

The first variance term is the variance of the conditional default rate, the second is the variance of idiosyncratic losses. For a non-zero subportfolio size $M > 0$, the covariance is greater than zero whenever (i) the factor sensitivity is positive, or (ii) the portfolio is not perfectly diversified, i.e. when the idiosyncratic losses have a positive variance.

In the derivation of (A6) we assumed that the number of loans contained in a subportfolio is a constant M . In the procedure described by Lopez and Saidenberg (2000), by contrast, the subportfolio size is M only on average. This does not question the conclusions drawn from (A6). The variation in subportfolio size is independent across subportfolios and would thus not affect the direction of the covariances. Further, Lopez and Saidenberg suggest not to draw the same subportfolio more than once. This introduces a dependence between the error terms η_{it+1} and η_{jt+1} . There is no reason to suspect, though, that this will make the overall covariance equal to zero, especially when considering the fact that there is only a very small probability of drawing the same subportfolio twice (cf. section 3).

To sum up, we can rewrite equation (A1) as

$$e_{it+1} = L_{it+1} - \hat{\mu}_{mit} = \gamma_{t+1} + u_{it+1} \tag{A7}$$

where the time-specific error γ_{t+1} represents the cross-sectional dependence derived in (A6). Any statistical test which does not take account of time effects, for instance a

standard regression $L_{it+1} = \alpha + \beta \hat{\mu}_{mit} + u_{it+1}$ will lead to biased standard errors. One way of tackling the problem of cross-sectional dependencies is through a panel regression which augments $L_{it+1} = \alpha + \beta \hat{\mu}_{mit} + u_{it+1}$ through either fixed or random time effects. While this might be a valid procedure in some cases, it will fail to detect misspecifications in others. Consider a model that overestimates the factor sensitivity w . The predicted variance of the time effect γ_{t+1} will thus be too high relative to the correct model, but this prediction is not subject of the unbiasedness test $\alpha = 0$ and $\beta = 1$.

Table 1: Base case setup

Parameter	Value
Number of possible states	2
Recovery in case of default	0
Portfolio size / number of borrowers (N)	10,000
Constant unconditional 1-year default probability (p)	1%
Uniform asset correlation in true model (w^2)	5%
Uniform asset correlation under H_0 (w^2)	[0%, 20%]
Asset value distribution	$N(0,1)$
Serial correlation of systematic factor	None
Forecast horizon (years)	1
Length of default history (years)	10
Test size / Type-I error	10%
Number of simulated default histories for power calculations	10,000
Number of scenarios for loss distributions	1,000,000

Table 2: Simulated performance of the Lopez and Saidenberg quantile test

Average subportfolio size	H ₀ (asset correlation)	Rejection frequency of H ₀
2,000	1.0%	95.1%
	2.5%	92.5%
	5.0% = true	89.9%
	7.5%	89.2%
	20.0%	92.6%
5,000	1.0%	91.4%
	2.5%	85.2%
	5.0% = true	81.0%
	7.5%	80.7%
	20.0%	86.7%
8,000	1.0%	85.1%
	2.5%	76.4%
	5.0% = true	73.6%
	7.5%	73.2%
	20.0%	81.6%

The Lopez and Saidenberg (2000) test procedure is implemented for the base case (see Table 1). For each year of a simulated 10-year default history, 1,000 subportfolios are drawn randomly without replacement. A likelihood ratio test is performed to test the null hypothesis that the observed number of 90%-quantile violations is equal to 10%; the significance level of the test is 10%. The reported rejection frequencies are based on 10,000 simulated default histories.

Table 3: Simulated power of Berkowitz test

Correlation	Power in variations of base case							
	H ₀ 99% quantile of default distribution (base case)	Power in base case	Size = 5% (vs. 10%)	1000 borrowers (vs. 10,000)	5,000 borrowers (vs. 10,000)	5-year history (vs. 10)	0.5% default probability (vs 1%)	Heterogeneous default probabilities
0%	123	100%	100%	87.1%	100%	99.7%	100%	100%
1%	181	92.3%	88.9%	53.2%	89.1%	74.5%	90.1%	90.5%
2%	221	60.9%	50.6%	28.4%	53.8%	42.5%	56.5%	56.9%
3%	256	29.5%	20.0%	16.2%	27.1%	24.2%	27.5%	27.0%
4%	289	15.7%	8.8%	12.7%	14.7%	17.4%	14.9%	14.4%
5% = true	321	12.6%	6.9%	13.3%	12.4%	15.8%	12.5%	12.3%
6%	352	15.3%	8.4%	16.9%	15.0%	17.3%	15.2%	14.4%
7%	382	20.6%	12.1%	21.7%	19.9%	19.9%	20.2%	19.4%
8%	413	27.5%	16.9%	27.1%	26.5%	23.2%	27.0%	25.7%
9%	441	35.2%	22.7%	33.6%	34.0%	26.9%	34.8%	32.8%
10%	471	43.8%	29.5%	41.2%	42.4%	31.1%	43.5%	40.3%
11%	497	52.0%	36.7%	48.1%	50.7%	35.2%	51.9%	47.7%
12%	525	60.6%	44.6%	55.6%	59.8%	39.2%	61.0%	55.3%
13%	558	68.8%	52.6%	62.0%	67.0%	43.3%	68.9%	63.0%
14%	585	76.0%	60.6%	67.8%	73.9%	47.8%	76.0%	69.4%
15%	614	81.9%	67.8%	73.3%	79.5%	52.4%	81.7%	75.3%
20%	758	97.1%	92.2%	92.1%	95.4%	74.0%	96.8%	93.3%

Columns 1-3 refer to the base case setting (see Table 1). The other columns refer to separate variations of the base case. In the last column the assumption of homogeneous default probabilities is replaced by a heterogeneous portfolio (see Table 4) that is similar to the high quality credit portfolio in Gordy (2000).

Table 4: Composition of heterogeneous portfolio

Rating	Unconditional default probability	Number of borrowers
AAA	0.01%	382
AA	0.02%	590
A	0.06%	2.256
BBB	0.18%	3.792
BB	1.06%	1.908
B	4.94%	942
CCC	19.14%	130

Table 5: Simulated power of Berkowitz test

<i>Varying default probabilities under H_0</i>			<i>Varying the asset value distribution under H_0</i>		
Default probability under H_0	99% quantile of default distribution	Power	Degrees of freedom of t-distribution under H_0	99% quantile of default distribution	Power
0.2%	79	100%	10	911	100%
0.4%	145	99.5%	20	646	92.3%
0.6%	207	76.4%	30	547	71.8%
0.8%	265	29.1%	40	496	55.2%
1.0% = true	321	12.6%	50	463	44.5%
1.2%	376	22.8%	60	441	37.3%
1.4%	428	48.3%	70	426	32.5%
1.6%	481	73.8%	80	413	28.9%
1.8%	531	89.9%	90	404	26.2%
2.0%	581	96.9%	100	395	24.1%
2.2%	630	99.1%	200	361	16.6%
2.4%	678	99.8%	∞ = true	321	12.6%

The true model is as in the base case (see Table 1). The models that are evaluated are identical to the true model except for the unconditional default probability or the type of the asset value distribution.

Table 6: One-year probabilities of rating transition

Initial rating	One-year ahead rating							
	AAA	AA	A	BBB	BB	B	CCC	D
AAA	0.9223	0.0656	0.0093	0.0005	0.0021	0.0001	0.0000	0.0001
AA	0.0063	0.9080	0.0786	0.0059	0.0006	0.0004	0.0000	0.0001
A	0.0004	0.0135	0.9341	0.0455	0.0049	0.0011	0.0002	0.0002
BBB	0.0003	0.0024	0.0507	0.8967	0.0416	0.0067	0.0008	0.0008
BB	0.0000	0.0013	0.0100	0.0962	0.8083	0.0677	0.0093	0.0072
B	0.0001	0.0026	0.0028	0.0099	0.0758	0.7939	0.0630	0.0519
CCC	0.0029	0.0004	0.0062	0.0020	0.0251	0.0786	0.4462	0.4387

The transition matrix is taken from Lando and Skodeberg (2001, Table 3). It is based on continuous rating data from Moody's over the period 1988-98. The probability mass of the not-rated category has been apportioned to the other rating classes according to their probability mass.

Table 7: One-year ahead loan values

Initial rating	Spread	Coupon	One-year ahead loan value conditional on rating						
			AAA	AA	A	BBB	BB	B	CCC
AAA	0.0043	0.0495	1.0376	1.0319	1.0170	0.9998	0.9372	0.8600	0.7296
AA	0.0058	0.0510	1.0442	1.0386	1.0236	1.0064	0.9436	0.8661	0.7352
A	0.0098	0.0548	1.0619	1.0561	1.0411	1.0237	0.9603	0.8822	0.7500
BBB	0.0148	0.0598	1.0843	1.0785	1.0632	1.0457	0.9816	0.9026	0.7688
BB	0.0333	0.0776	1.1657	1.1596	1.1438	1.1254	1.0589	0.9767	0.8372
B	0.0591	0.1023	1.2779	1.2715	1.2549	1.2355	1.1654	1.0789	0.9314
CCC	0.1100	0.1502	1.4961	1.4890	1.4709	1.4495	1.3725	1.2775	1.1148

All loans mature in five years. The yields necessary for loan valuation were taken from the CreditMetrics web site on 8 April 2002. We use the yield spreads for US corporates, and the yield on US treasuries. Required yields that are not provided on the website are obtained through linear interpolation. Coupon rates are set such that loans are initially valued at par.

Figure 1: Power of Berkowitz test in base case

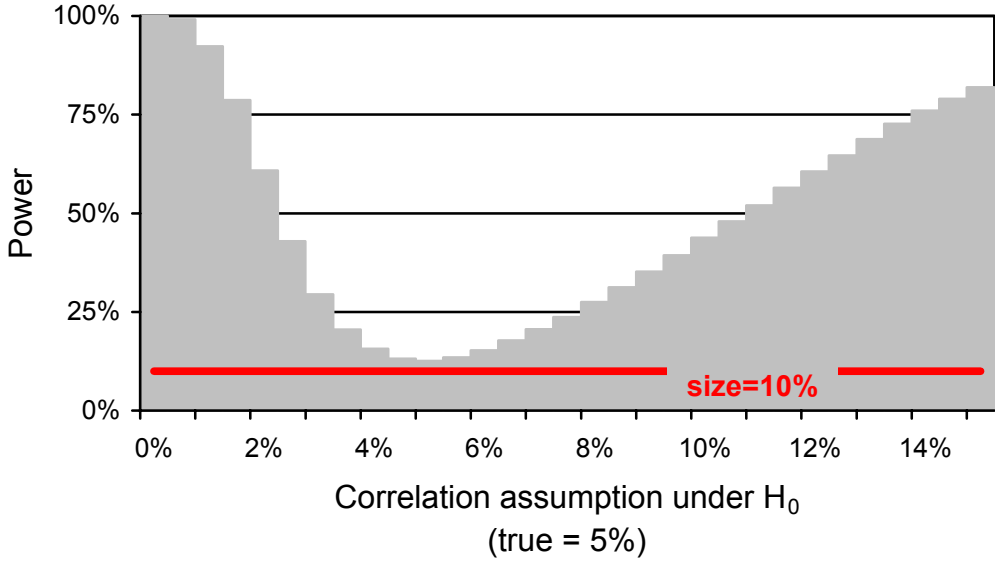


Figure 2: Power of alternative tests in base case

The power of the Berkowitz test serves as a benchmark and is identical to Figure 1. The normality test is the Doornik-Hansen test. The 'Standard' likelihood ratio test is performed on the untransformed default data. For each simulated default history, the ML estimate of the asset correlation is determined through a search procedure. The distribution of the test statistic is simulated under H_0 in order to obtain critical values.

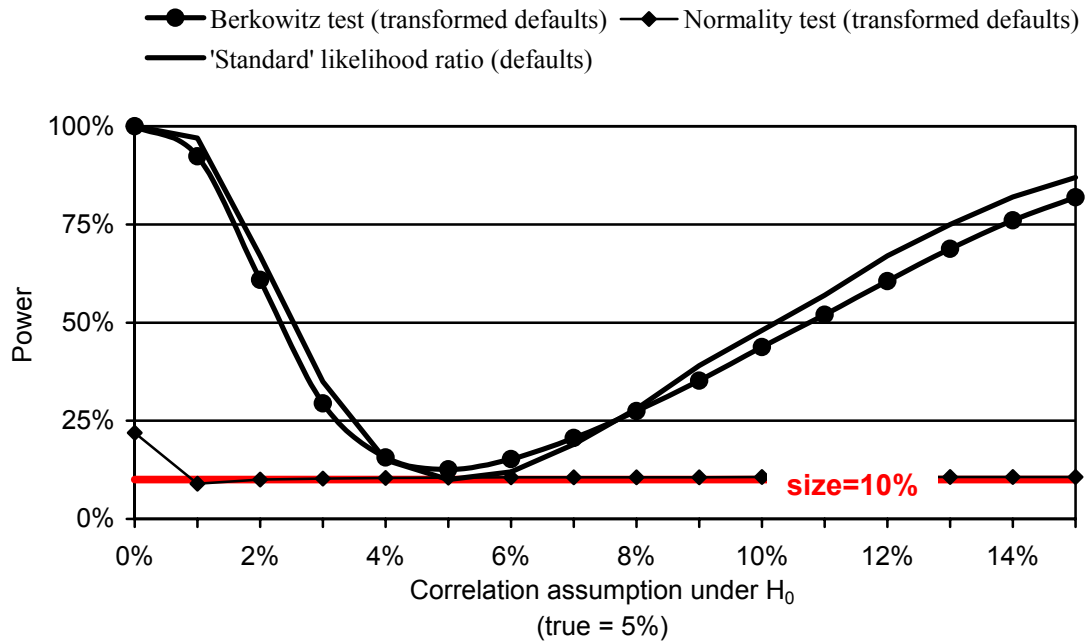


Figure 3: Power of Berkowitz test in multi-state case with systematic recovery rate risk

The Berkowitz test is applied to the heterogeneous portfolio from Table 4. The one-year transition matrix is given in Table 6, for conditional one-year ahead loan values see Table 7. Recovery rates have a mean of 0.521 (cf. Gupton, Gates, and Carty, 2000), a volatility of 0.32, and an average correlation of 2.89% (cf. Frye, 2000).

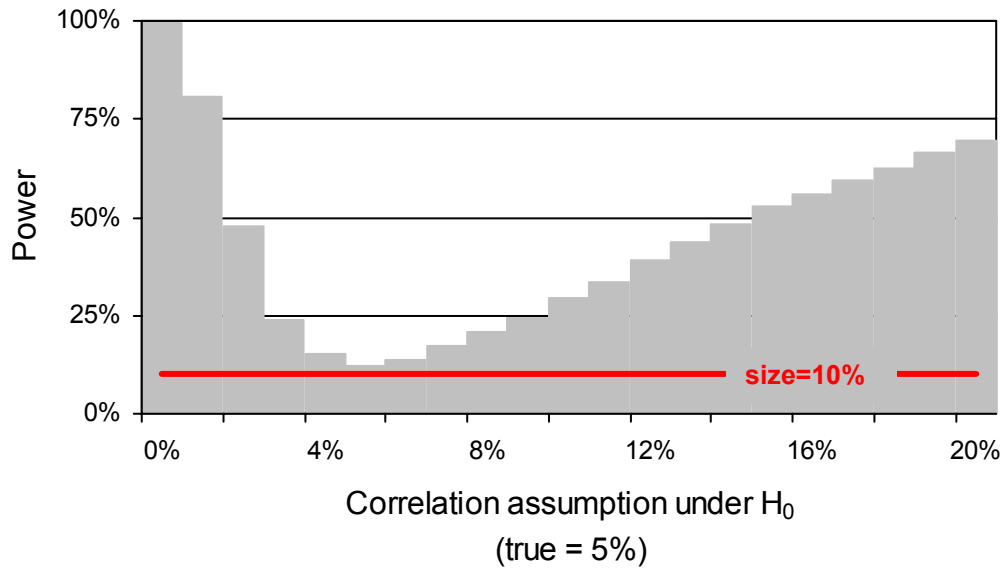


Figure 4: Power of Berkowitz test when including cross-sectional information

The setup is identical to the base case (see Table 1) except for the asset correlation within the true model. Instead of a uniform asset correlation of 5% there are two equally sized sectors with intra-sector asset correlations of 2% and 9%, respectively. The grey shaded area shows the power when the Berkowitz test is based on aggregate portfolio defaults. The dotted line depicts the power when the Berkowitz procedure is extended to assess the accuracy of sector defaults.

