Evaluating credit risk models: A critique and a proposal

Hergen Frerichs [a,*]

Gunter Löffler [a]

University of Frankfurt (Main)

First Version: February 15, 2001
This version: October 9, 2001

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 renders standard statistical inference invalid. We proceed by suggesting the Berkowitz (1999) procedure, which relies on standard likelihood ratio tests performed on transformed default data. We simulate the power of this approach in various settings including one in which the test is extended to incorporate cross-sectional information. To compare the predictive ability of alternative models, we propose to use either Bonferroni bounds or the likelihood-ratio of the two models. Monte Carlo simulations show that a default history of ten years can be sufficient to resolve uncertainties currently present in credit risk modeling.

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

JEL classification: G2; G28; C52

[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

Acknowledgements

The paper is part of a joint research project with Deutsche Bundesbank on modeling credit portfolio risk. We wish to thank Ron Anderson, Jan Pieter Krahnen, Thilo Liebig, Ludger Overbeck, Peter Raupach, Mark Wahrenburg and participants of the brown bag seminar at the University of Frankfurt (Main), the 2001 meeting of the European Financial Management Association and the 2001 research conference of the International Association of Financial Engineers for helpful comments.

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 which 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.

The following example illustrates the validation problem: in the current supervisory backtesting framework[1], internal market risk models can be sanctioned if there are more than four violations of the 99% VaR at Risk (VaR) over the preceding year. Under the assumption that violations are binomially distributed (using a success probability of 99% and 250 days), one out of nine models might be sanctioned even though it is correct. Consider another model which underestimates the VaR by 12% because it mistakenly sets the 99% VaR equal to the 98%-quantile of the loss distribution. With a probability of 44%, this misspecified model will not be sanctioned under the current rule. What happens if we base the test on only ten data points instead? For the true model, the probability of observing no violations is now 90%. If regulators took a single violation to be sufficient for sanctioning, the error of sanctioning the true model would be similar to the former setting. But the probability of not sanctioning the misspecified model would increase to 82%. This shows that the power of the regulatory market risk backtest drops sharply when applied to short histories typical for credit risk data.

In order to overcome the lack of credit data on the time dimension Lopez and Saidenberg (2000) propose to evaluate credit portfolio risk models based on cross-sectional simulation. Given a data set of N loans over T years, the idea is to resample without replacement, for each of the T years, a large number of subportfolios containing a fixed number of loans ($<N$). The credit risk of each subportfolio is predicted and compared with the actual portfolio return. Lopez and

[1] Basel Committee on Banking Supervision (1996 a, b)

Saidenberg propose tests for the accuracy of the predicted loss distribution's mean and quantiles and of the complete distribution. In the construction of the tests they 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, defaults will be above average both in the entire sample and in randomly drawn subsamples, which renders standard statistical inference invalid.

Subsequently, we show that backtesting credit portfolio risk models based on a default history of only ten years is possible if we use the information of the complete default distribution. For this purpose, we recommend using Berkowitz' (1999) 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. Berkowitz proposes a test of independence and a test of zero mean, unit variance and independence against a first-order autoregressive structure. 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 even ten observations are sufficient to detect misspecifications in credit risk models.

This can be illustrated through the following example: in a CreditMetrics[2] type model, the value chosen for the asset correlation is crucial for the results because it drives default correlations. 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 which vary between 1.5% and 12.5%. With ten years of data on annual defaults, a true correlation of 5% and a significance level of 10%, the probability of rejecting a correlation assumption of 20% equals 97%.

We follow Lopez and Saidenberg in trying to exploit information contained in the cross-section of defaults. However, we argue that random sampling will often fail to make efficient use of this information. Tests can rather be based on judiciously

[2] Cf. J.P. Morgan (1997) for a general description of CreditMetrics.

chosen subportfolios. To gain an intuition for our 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 a risk analyst uses the default experience of this portfolio to evaluate a model, which 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. By examining the default experience of single sector subportfolios, we are in a much better position to identify the inadequacy of the model. If the number of subportfolios is not too large, the Berkowitz procedure can be adapted to jointly test the validity of predictions for subportfolio defaults.

If the aim of the evaluation is to choose among alternative credit risk models, Lopez and Saidenberg propose to use Bonferroni bounds. As an alternative, we draw on Bayesian statistics and suggest to compare the likelihood ratio of two models, that is, examine which model is more likely to have generated the data.[3] Usually, such an analysis does not involve testing whether alternative models are statistically different from each other. Since we feel that many practitioners and regulators will want to base their decisions on the usual concepts of significance, we take up a proposal by Good (1957). The likelihood ratio is taken as a statistic, whose distribution is obtained through Monte Carlo simulation. This allows tests of the following form: What is the confidence that model *A* provides a better fit to the data than model *B*?

Nickel, Perraudin and Varotto (2001) 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

[3] Kon (1984), for example, uses odds ratios to examine whether stock returns are best described by a t-distribution or by a mixture of normal distributions.

measures vary across different portfolio types. Crouhy, Galai and Mark (2000), Gordy (2000) and Wahrenburg and Niethen (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)[4], Berkowitz (1999) and a normality test recommended by Doornik and Hansen (1994). The authors suspect that the Berkowitz (1999) test and the normality test might be sensitive to outlier observations. De Gooijer and Zerom (2000), however, cannot confirm this conjecture.

Bedendo and Hodges (2001) compare the power of multivariate goodness-of-fit tests based on the empirical characteristic function. The focus of the authors is on testing market risk models with potentially hundreds of risk factors. The Berkowitz (1999) approach does not lend itself easily to multidimensional tests. In a credit risk setting, where the number of risk factors is typically small, this does not seem to be a major disadvantage.

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 random default histories which are all generated by one specific credit portfolio risk model. We then state the null

[4] 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.

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.

The framework we apply is similar to a two-state version of CreditMetrics. Without loss of generality, we neglect both migration risk and recovery rate uncertainty. In consequence, the output of a credit risk model is a discrete distribution of the expected number of defaults within a portfolio, and portfolio weights are irrelevant. Default correlations are modeled based on correlated 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 default correlations.

In a two-state world, various credit portfolio risk models like CreditRisk$^+$, CreditMetrics, KMV PortfolioManager or CreditPortfolioView are similar in the underlying structure and produce almost identical outputs when parameterized consistently.[5] For this reason, we conjecture that our results are applicable to a broad range of credit risk models although we examine only one class of models. In addition, the test procedures put forward in this paper can be directly applied in more complex settings, e.g. when migration risk is added. Even though we restrict the analysis to one particular class of portfolio credit risk models, we will nevertheless speak of various 'models' which we are going to evaluate. In the following, the term 'models' will thus refer to different parameterizations of the basic latent variable approach.

[5] Cf. Finger (1998), Koyluoglu and Hickman (1998), Gordy (2000), and Wahrenburg and Niethen (2000).

In the simplest setup of this framework asset value changes $\Delta\widetilde{A}_i$ depend on a systematic factor $\widetilde{Z}$ (e.g. the growth rate of the economy) and idiosyncratic factors $\widetilde{\varepsilon}_i$:[6]

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

where $\widetilde{Z}$ and $\widetilde{\varepsilon}_i$ are iid N(0,1). The term $\sqrt{1-w_i^2}$ causes the asset value change $\Delta\widetilde{A}_i$ to be standard normally distributed. A borrower defaults whenever $\Delta\widetilde{A}_i < \Phi^{-1}(p_i)$, where p_i is the unconditional default probability and Φ denotes the cumulative standard normal distribution function. 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.[7]

Since Gordy (2000) and Frey, McNeil and Nyfeler (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 that choosing the shape of the distribution is one step in parameterizing the asset value model (1).

For a given realization of the systematic factor Z the conditional default probability $p_i|Z$ equals

$$\mathrm{p_i} \mid \mathrm{Z} = \mathrm{Prob}\left(\varepsilon_i \le \frac{\Phi^{-1}(p_i) - w_i Z}{\sqrt{1-w_i{}^2}}\right) = \Phi\left[\frac{\Phi^{-1}(p_i) - w_i Z}{\sqrt{1-w_i^2}}\right] \qquad (2)$$

The default distributions of this credit risk model can be easily derived using Monte Carlo simulations.[8] We conduct Monte Carlo simulations with 1,000,000 trials to ensure accurate results.

---

[6] The extension to a multi-factor model is straightforward.

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

[8] If the portfolio is homogeneous, a quick way to perform the simulations is i) draw N(0,I)-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%).

In the base case, we assume that evaluators of credit risk models observe ten years of annual data on homogeneous portfolios of 10,000 borrowers. We set the unconditional annual default probability equal to 1% for each obligor, and assume that there is no serial correlation of defaults across time. Asset values follow a standard normal distribution.

The asset correlation parameter is the only one which is varied in the base case. In the true model, which underlies 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 $w^2 \in [0\%, 20\%]$.

The power of our tests will be calculated based on 10,000 independent 10-year default histories which will be simulated using 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 assumptions are summarized in Table 1. They will be varied in section 4.1 to check the robustness of our simulation 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* loans, a large number *R* of subportfolios is randomly drawn for each year *t* in *T*. In drawing the borrowers for a particular subportfolio, Lopez and Saidenberg suggest to draw without replacement. They also recommend to draw 'large' subportfolios, but do not discuss this issue in detail. For each subportfolio, the loss distribution is forecasted and compared with observed subportfolio losses. In a sense, the number of observations available for model evaluation is thus multiplied by the factor *R*.

The following example illustrates this procedure: for a credit portfolio of 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) with 1,000 borrowers each. We count the number of defaults in each subportfolio:

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

If overall portfolio defaults are high, as in year one, the number of defaults in resampled subportfolios will be high as well. Similarly, the low number of portfolio defaults in year ten shows in subportfolios drawn from that year. If the VaR were equal to 15 defaults we would record many violations in the first year and few, if any, in the last. Obviously, defaults in the 1,000 subportfolios resampled from one year's default experience are not independent, so that standard testing procedures cannot be used.[9]

In the following, we conduct simulations to demonstrate that the lack of independence can severely affect the performance of the test statistics proposed by Lopez and Saidenberg. For this purpose, we implement the quantile test the authors propose (Lopez and Saidenberg (2000), p. 160).

Under the assumption that the predicted quantiles are accurate and observed violations of the quantiles are independent, these violations are draws from a

[9] In fact, cross-sectional dependence arises even when we resample from a portfolio with zero default correlation. Consider a homogenous portfolio with 1,000 obligors, a default probability of 1% and a zero default correlation. If the chosen subportfolio size is 500, the 90% quantile for subportfolio defaults is ten. With a probability of 46%, however, the overall number of defaults in the entire portfolio is nine or less; in these cases, one would not observe a violation of the 90% quantile in any of the random subportfolios.

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

$$LR(\alpha) = 2\left[\log\left(\hat{\alpha}^{y}(1-\hat{\alpha})^{T\bullet R-y}\right) - \log\left(\alpha^{y}(1-\alpha)^{T\bullet R-y}\right)\right], \qquad (3)$$

where *y* is the number of violations across the *T*R* subportfolios.

In the simulations, we apply this test to validate asset value models that differ only in their asset correlation w^2 (all parameters as in Table 1). For the test, we use the 90%-quantile and a significance level of 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. (We do this for three different subportfolio sizes of 2,000, 5,000 and 8,000 borrowers, respectively.) The borrowers included in a subportfolio are drawn without replacement.
3. Implement the LR-test (3) for a specific credit risk model by calculating the number of violations of the predicted 90% quantile of defaults..
4. Repeat steps 1. - 3. 10,000 times.

The results are summarized in Table 2.[10] 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 73% and 91%. The intuition for the results is that, by assuming independence across simulated subportfolios, the test overestimates the amount of information contained in the data. In consequence, the test is biased towards rejection. The fact that the rejection frequency decreases with increasing subportfolio size is due to the resampling procedure. As we draw without replacement, subportfolio defaults are hypergeometrically distributed, and the variation of the number of defaults across subportfolios goes down with increasing subportfolio size. In the extreme case of a subportfolio size of 100% there is no variation any more.

[10] 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^{-311}.)

The decreasing variation leads to lower rejection frequencies.

We conclude that the test procedure proposed by Lopez and Saidenberg (2000) is inaccurate. As the *R* subportfolios are not cross-sectionally independent the standard test statistics proposed by the authors cannot be used. This also holds if one tested the complete default distribution instead of the 90%-quantile, or ran Mincer-Zarnowitz regressions to examine unbiasedness of the forecasted number of defaults. Each of the tests proposed by Lopez and Saidenberg requires independent draws.

One might think of modifying the test procedure by conditioning the forecasts of subportfolio defaults on the default experience of those borrowers which 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 not be possible to discriminate between models which posit that asset values are driven by one factor with a uniform factor sensitivity w but which differ in the value assumed for 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 which the tests are meant to exploit loses importance.

4. Evaluating credit risk models based on the entire forecast 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 default distribution.[11]

We recommend the Berkowitz (1999) test procedure. In this approach, the default 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 (1999) applies a simple twist to the so-called Rosenblatt (1952) transformation of observed data. First, the estimated cumulative distribution function $\hat{F}(.)$ is applied to the observed number of defaults

$$x_t = \hat{F}(y_t) = \int_{-\infty}^{y_t} \hat{f}(u)du \,, \qquad (4)$$

where y_t is the ex post number of defaults and $\hat{f}(u)$ is the forecasted probability of *u* defaults. If the estimated default distribution is equal to the true one, the transformed variable x_t is iid U(0,1), where U(.) denotes the uniform distribution.

In a second step, Berkowitz suggests to apply another transformation using the inverse of the standard normal distribution function Φ:

$$z_t = \Phi^{-1}(x_t) \qquad (5)$$

If the predicted distribution function is correct, the transformed observations z_t are iid N(0,1).[12] 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 default data in order to assess their power. We investigate two cases: in the first case we only use aggregate portfolio defaults across time for our tests. In the second we extend the analysis to include information inherent in subportfolio defaults. Finally, we treat the problem of model comparison.

---

[11] Simple quantile tests as in (3) 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.

[12] See Berkowitz (1999) for a proof.

4.1 Using only aggregate portfolio defaults

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:

$$\log \mathrm{L} = -\frac{1}{2}\log 2\pi - \frac{\mathrm{T}}{2}\log \sigma^2 - \sum_{\mathrm{t}=1}^{\mathrm{T}}\left(\frac{(\mathrm{z_t}-\mu)^2}{2\sigma^2}\right), \qquad (6)$$

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

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

$$\hat{\sigma}^2_{ML} = \frac{\sum (z_t - \hat{\mu}_{ML})^2}{T}, \qquad (7)$$

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

$$\lambda = 2\left[\log L(\mu = \hat{\mu}_{ML}, \sigma^2 = \hat{\sigma}^2_{ML}) - \log L(\mu = 0, \sigma^2 = 1)\right] \qquad (8)$$

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

Figure 1 shows the simulated power of our test statistic in the base case. If the false model posits a zero default correlation, the null hypothesis is rejected in 100% of all cases. For models which 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%.

When using an alternative correlation assumption of 5%, which coincides with the true model, the power equals 12%, which is slightly higher than the expected value of 10%. Due to the small size, the test statistic is not exactly chi-squared distributed. The inaccuracy seems to be small, and is probably negligible for 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 which puts them into perspective. In column two we list the 99% quantiles of the default distribution under the various null hypotheses to illustrate how different these distributions are from the true model.

Let us compare our results with the regulatory market risk backtesting. Recall that models can be sanctioned if there are more than four violations of the predicted VaR. With 250 observations, the probability that the true model is sanctioned equals 11%. A model underestimating the VaR by 12% will be sanctioned with a probability of 56%. In Table 3, the 99% quantile of the true model with an asset correlation of 5% equals 321 defaults while an asset correlation of 3.5% leads to a 99% quantile of 273 defaults, that is, underestimates the VaR by 15%. The power of the LR-test equals 21% for this model which is significantly lower than the power of the market risk test. A comparable power is only achieved with an asset correlation of 2% (power = 61%). For this scenario, the 99% quantile equals 221 defaults which underestimates the VaR by 31%.

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. This is due to the fact that differences in asset correlations matter less in a heterogeneous portfolio which is ceteris paribus less risky than its homogeneous counterpart.

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 or Wahrenburg and Niethen, 2000) typically lead to much lower correlation estimates.[13] 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.[14] 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 which diverged from the true

[13] The asset correlations are calibrated to match the observed default rate volatility.

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 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 which 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 defaults. Gordy (2000) 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):

$$\Delta\tilde{A}'_i = \sqrt{\frac{\nu}{\tilde{w}}}\,\Delta\tilde{A}_i\,, \quad \tilde{w} \sim \chi^2(\nu)\,, \tag{9}$$

where ν denotes the degrees of freedom assumed for the t-distribution. The

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

distribution approaches the normal as ν approaches infinity. A borrower defaults when $\Delta\widetilde{A}'_i < t_\upsilon^{-1}(p)$, 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 ν = ∞ to describe the true model, and vary the degrees of freedom assumed under the null hypothesis.[15] 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%.

Finally, we modify the base case by introducing autocorrelation into the time series of the systematic factor $\widetilde{Z}$. In simulating the default histories, we use the following autoregressive process for $\widetilde{Z}_t$:

$$\widetilde{Z}_t = 0{,}5\widetilde{Z}_{t-1} + 0{,}866\widetilde{u}_t, \quad \widetilde{u}_t \sim N(0,1),\ \widetilde{Z}_1 \sim N(0,1) \qquad (10)$$

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 (1999), we augment the density function for the transformed defaults z_t by allowing them to follow a first-order autoregressive process:

[15] Conclusions do not change when we look at the opposite case that the true asset value distribution is a t-distribution and we test alternative hypotheses whose underlying distribution is normal.

$$\log L = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\left[\frac{\sigma^2}{(1-\rho^2)}\right] - \frac{(z_1 - \mu/(1-\rho))^2}{2\sigma^2/(1-\rho^2)} - \frac{T-1}{2}\log\sigma^2 - \sum_{t=2}^{T}\left(\frac{(z_t - \mu - \rho z_{t-1})^2}{2\sigma^2}\right) \quad (11)$$

Having obtained maximum likelihood estimators for the three parameters μ, σ^2 and ρ, it is tempting to construct a likelihood ratio statistic for the H_0 restrictions $\mu=0$, $\sigma^2=1$ and $\rho=0$. This would neglect the fact that the estimator for the autocorrelation coefficient ρ is downward biased in small samples (cf. Quenouille, 1949 or Andrews, 1993). Monte Carlo simulations show that, 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 the base case, produces the following result: if the factor is governed by the process described in (10), 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 which exclude serial correlation.

4.1.3 Alternative tests

Under the null hypothesis, the transformed variables should be standard normally distributed. Following Berkowitz (1999), however, we only tested whether they have mean zero and unit variance. One could presume that the power of the test could be

[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

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 which will typically not be feasible, but provides a useful benchmark in the stylized 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*:

$$\hat{w}_{ML} = \arg\max_{w} \sum_{t=1}^{T} \log[f_w(u_t)], \qquad (12)$$

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%,....., 50%]. This estimate can be used to construct a standard likelihood ratio test against a specific H_0.

$$\lambda_{Alt} = 2\left[\log L_{\hat{w}_{ML}} - \log L_{w_{H0}}\right] = 2\left[\sum_{t=1}^{T} \log f_{\hat{w}_{ML}}(u_t) - \sum_{t=1}^{T} \log f_{w_{H0}}(u_t)\right] \qquad (13)$$

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 testing the transformed variables for a mean of zero and a variance of one. The

---

does not change the results significantly.

[17] Typically, the number of free parameters will be too large to estimate them based on aggregate portfolio data.

evidence supports the view that we do not lose significant information by (i) transforming the default data and (ii) testing only a subset of the restrictions which the transformed data should obey if the null hypothesis is correct.

4.2 Testing cross-sectional predictions

Up until now, we have used only the aggregate annual defaults to construct a test. This will not be efficient if there is additional information in the cross-section of the data. Consider evaluating a model which assumes a uniform asset correlation across obligors. Using the test procedure described above, the evaluator cannot reject the validity of the model. However, she conjectures that the true correlations differ across obligors. How could one test this conjecture?

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

$$\Delta\widetilde{A}_i = w_i\widetilde{Z} + \sqrt{1-w_i^2}\,\widetilde{\varepsilon}_i,\ w_i^2 = 0.02 \text{ for } i \in \text{sector 1},$$
$$w_i^2 = 0.09 \text{ for } i \in \text{sector 2}. \qquad (14)$$

We simulate 10-year default histories using this two-sector model and use the Berkowitz test (8) to check whether we can reject a model which posits a uniform asset correlation of 5%. With a size of 10%, the power is only 16% (Figure 3). This result is due to the fact that the aggregate expected default distributions of the true model and the null hypothesis are almost identical, even though the sector portfolio distributions differ.

One possible way of exploiting the cross-sectional information is to utilize the idea of Lopez and Saidenberg and apply the test to randomly drawn portfolio subsets. This would not make efficient use of the information, though. The main disadvantage of drawing the random subportfolios is that we hardly ever get extreme subportfolio compositions. If we draw a large number of reasonably large subportfolios (say, with 2,000 borrowers each), the probability that we obtain at least one subportfolio which consists only of borrowers of one sector is extremely low.[18] If the null hypothesis is a

[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^{-314}. By contrast, the probability of obtaining an even mixture of

common correlation of 5%, it is these extreme portfolio compositions which have the greatest informational value for our purpose. The more evenly mixed a subportfolio is, the more similar the H_0 model is to the true one. Even if we obtain some extreme portfolio compositions through resampling, their informational value will be lost by averaging across all subportfolios. As a consequence, randomly drawing subportfolios is unlikely to yield a significant increase of power.

A more efficient way of tackling the problem is to divide the portfolio into the extreme subportfolios and calculate the test statistic for each of them. Thus, if we have a two-sector portfolio and assume that borrowers of these two sectors have different sensitivities towards a common systematic factor, we form two subportfolios consisting of just one sector and proceed as though we were to test models on 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, the variables will be contemporaneously correlated. Under the null, they follow a bivariate standard normal distribution, which has the following likelihood:

$$\log L = -T\log 2\pi - T\log\sigma_1 - T\log\sigma_2 - \frac{T}{2}\log(1-\rho_{12}^2)$$
$$-\frac{1}{2(1-\rho_{12}^2)}\sum_{t=1}^{T}\left[\left(\frac{z_{t1}-\mu_1}{\sigma_1}\right)^2 - 2\rho_{12}\left(\frac{z_{t1}-\mu_1}{\sigma_1}\right)\left(\frac{z_{t2}-\mu_2}{\sigma_2}\right)+\left(\frac{z_{t2}-\mu_2}{\sigma_2}\right)^2\right] \qquad (15)$$

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

---

sectors is 2%.

assumption about the value of the uniform asset correlation. The results are shown in Figure 3. 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. Typically, one will want to test whether models are too parsimonious (as in the example) or too complex. In the former case, evaluators would split a portfolio into sectors they believe to be different. In the latter case, one would examine portfolios for which the model predicts large differences. If a model assumes, for instance, that individual default probabilities vary from 0% to 2%, while evaluators assume that they are uniform at 1%, one could separate borrowers according to whether the model predicts a default rate of less than 1% or larger than 1%, respectively.

By extending the bivariate likelihood (15) 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 ($M \times T$).

4.3 Model comparisons

So far we have tested whether one particular model is consistent with the default data. Another evaluation objective can be to decide whether one model provides a significant improvement against an alternative one.

In such a setting, the Berkowitz (1999) test can be used to separately evaluate each model under consideration. The problem is that the type-I-errors might add up. If we test each model using a significance level of 10%, then the type-I error of the joint test will lie between 10% and 20%. Lopez and Saidenberg (2000) suggest using Bonferroni bounds to test whether two models are equally accurate. If the size of the test is to be bounded above by γ, one separately evaluates the accuracy of the models using a size of $\gamma/2$. If the validity of just one model is rejected, the other one can be said to be more accurate.

For the same setting, we assess the power of an alternative procedure. Consider the following likelihood ratio for two models A and B:

$$\lambda_{MC} = \log\left(\frac{L_A}{L_B}\right) = \log\left(\frac{\prod_{t=1}^{T} f_A(y_t)}{\prod_{t=1}^{T} f_B(y_t)}\right) \qquad (16)$$

where $f_A(y)$ denotes the density which model A assigns to a loss of y. If λ_{MC} is greater than zero, model A is more likely to have generated the data than model B. Such a comparison already gives an indication on which model to choose. If one wants to conduct a standard statistical test whether one model is significantly more accurate than the other, one cannot appeal to the chi-squared distribution. The λ_{MC} is not a likelihood ratio statistic in the usual sense, because it does not result from imposing a restriction on maximum likelihood estimates. However, we can nevertheless treat λ_{MC} as a statistic and simulate its distribution (cf. Good, 1957). If we want to test whether model A is significantly more accurate than model B, this would involve the following steps:

1. Set up the hypotheses H_0: model B is at least as accurate as model A
   H_1: model A is more accurate than model B.

Since we take the models' likelihood as a criterion for accuracy, we do not reject the null if $\lambda_{MC} = \log(L_A / L_B) \leq 0$. Else:

2. Calculate λ_{MC} for a large number of histories generated under model B, and compute the (1-α)-quantile of this distribution. The sample size used for the random histories is equal to the one of the actual data available to the evaluator.
3. Compute the value of the λ_{MC} statistic using the actual default history, and decide whether or not to reject the null hypothesis by comparing it to the simulated (1-α) quantile.

We simulate the power of this test as well as that of the Bonferroni test for the base case setup. We examine the polar case in which one of the models (model A) is the true one. For the Bonferroni test, the power is the probability that the Berkowitz test (8), while not rejecting model A, does reject the alternative model B (at a significance level of 5%). The power of the likelihood ratio test is the probability that the

statistic λ_{MC} is positive and larger than the 90% quantile of λ_{MC} simulated under model B.

The results are depicted in Figure 4. The power of the likelihood ratio test is generally larger than that of the Bonferroni bounds test which is not surprising as the Bonferroni test is conservative. When testing the true model against another one which posits an asset correlation of 10%, for instance, the Bonferroni test detects the correct model in 24% of all cases, compared to a power of 60% when applying the likelihood ratio λ_{MC}.

We have thus presented a relatively powerful and simple tool for comparing alternative model specifications. Compared to the Berkowitz procedure described in section 4.1.1, the test involves only one more step, i.e., simulating the distribution of the test statistic. Unlike the Berkowitz test, the procedure does not easily lend itself to testing cross-sectional predictions. A possible solution would be to aggregate the statistic (16) across subportfolios and base a decision on the simulated distribution of the aggregate statistic.[19]

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, for example, some of the questions currently debated by credit risk managers can be resolved with a probability larger than 90%.

A test should meet other criteria than a large power, for instance ease of implementation and general applicability. The tests are computationally simple. In most cases, they require only the predicted cumulative distribution of defaults and some elementary transformations. The simplest form of the test, which is based only on aggregate defaults, provides a benchmark which 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. Thus, there is no general rule for the design of the test. We do not regard this as a serious

[19] We have examined this possibility in an earlier version of the paper but do not explore it here because (i) it is computationally expensive and (ii) its general applicability is difficult to establish.

shortcoming because the choice of the most important subportfolios will be straightforward in many cases. Note, too, that the test procedures can directly be applied to models which include migration and recovery risk. Model comparisons can be based on a likelihood-ratio; for this test, alternative models need not be nested.

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.[20] There are two arguments against focusing on the right tail of the distribution when constructing a test. First, we observe only few of these rare events in the data, a problem even sophisticated procedures are unlikely to overcome. 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 asset value distribution in a CreditMetrics type model. 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.

[20] Diebold, Schuermann and Stroughair (1998)

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, 1996a, Overview of the amendment to the capital accord to incorporate market risks (January 1996, updated to April 1998), Basel.

Basel Committee on Banking Supervision, 1996b, Supervisory framework for the use of “backtesting” in conjunction with the internal models approach to market risk capital requirements, Basel.

Basel Committee on Banking Supervision, 2001, Overview of the new Basel capital accord, Basel.

Bedendo, M., Hodges, S.D., 2001, Multivariate distributional tests in risk management: an empirical characteristic function approach. Working paper, University of Warwick.

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., 1999, Evaluating the forecasts of risk models. Working paper, Federal Reserve Board.

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., 1998, Credit risk in private debt portfolios. Journal of Finance 53, 1363-1387.

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.

Diebold, F.X., Schuermann, T., Stroughair, J.D., 1998, Pitfalls and opportunities in the use of extreme value theory in risk management. Working paper, The Wharton School, University of Pennsylvania, No 98-10.

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.

Good, I.J., 1957, Saddle-point methods for the multinomial distribution. Annals of Mathematical Statistics 28, 861-881.

Gordy, M., 2000, A comparative anatomy of credit risk models. Journal of Banking and Finance 24, 119-149.

J.P. Morgan, 1997, CreditMetrics – Technical document, New York.

Kiesel, R., Perraudin, W., Taylor, A. 2001, The structure of credit risk. Working Paper, Birkbeck College, Bank of England.

Kon, S. 1984, Models of stock returns - a comparison, Journal of Finance 39, 147-165.

Koyluoglu, H.U., Hickman, A., 1998, Reconcilable differences. Risk 11, No 10, 56-62.

Lopez, J.A., Saidenberg, M.R., 2000, Evaluating credit risk models. Journal of Banking and Finance 24, 151-165.

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.

Wahrenburg, M., Niethen, S., 2000, Vergleichende Analyse alternativer Kreditrisikomodelle. Kredit und Kapital 33, 235-257. (With English summary.)

Table 1: Base case setup

Parameter	Value
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 in alternative models (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 default distributions	1,000,000

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

Subportfolio size	H_0 (Correlation)	Rejection frequency of H_0
2,000	1.0%	96.0%
	2.5%	92.8%
	5.0% = true	**90.5%**
	7.5%	89.8%
	20.0%	92.8%
5,000	1.0%	91.6%
	2.5%	85.0%
	5.0% = true	**81.1%**
	7.5%	80.8%
	20.0%	86.9%
8,000	1.0%	85.4%
	2.5%	76.3%
	5.0% = true	**72.7%**
	7.5%	73.0%
	20.0%	82.0%

Lopez and Saidenberg (2000) test procedure implemented for the base case (see Table 1). For each simulated default history and each year, 1,000 subportfolios of varying size (2,000, 5,000, 8,000) are drawn randomly without replacement. A likelihood ratio test is performed for each scenario to test the null hypothesis that the observed number of 90%-quantile violations is equal to the expected number.

Table 3: Simulated power of Berkowitz test

H_0			Power in variations of base case					
Correlation	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%	351	15.3%	8.4%	16.9%	15.0%	17.3%	15.2%	14.4%
7%	381	20.6%	12.1%	21.7%	19.9%	19.9%	20.2%	19.4%
8%	412	27.5%	16.9%	27.1%	26.5%	23.2%	27.0%	25.7%
9%	440	35.2%	22.7%	33.6%	34.0%	26.9%	34.8%	32.8%
10%	468	43.8%	29.5%	41.2%	42.4%	31.1%	43.5%	40.3%
11%	493	52.0%	36.7%	48.1%	50.7%	35.2%	51.9%	47.7%
12%	521	60.6%	44.6%	55.6%	59.8%	39.2%	61.0%	55.3%
13%	553	68.8%	52.6%	62.0%	67.0%	43.3%	68.9%	63.0%
14%	580	76.0%	60.6%	67.8%	73.9%	47.8%	76.0%	69.4%
15%	607	81.9%	67.8%	73.3%	79.5%	52.4%	81.7%	75.3%
20%	739	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) which 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

Varying default probabilities under H_0			*Varying the asset value distribution under H_0*		
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 in both scenarios is equal to the base case setting (see Table 1). We modify the base case by varying the unconditional default probability under H_0 on the left and the type of the asset value distribution under H_0 on the right instead of varying the asset correlation.

Figure 1: Power of Berkowitz test in base case

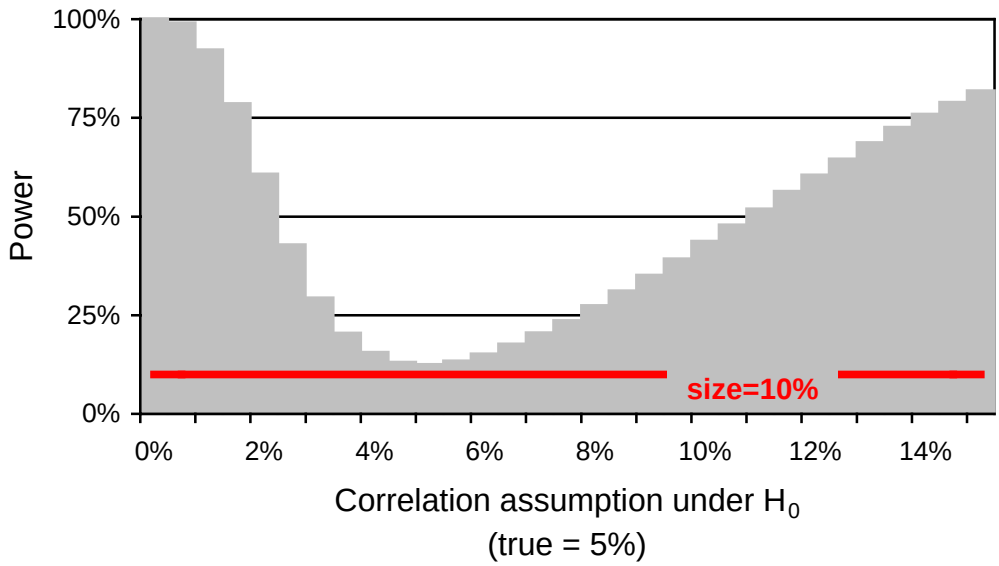

Figure 2: Power of alternative tests in base case

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

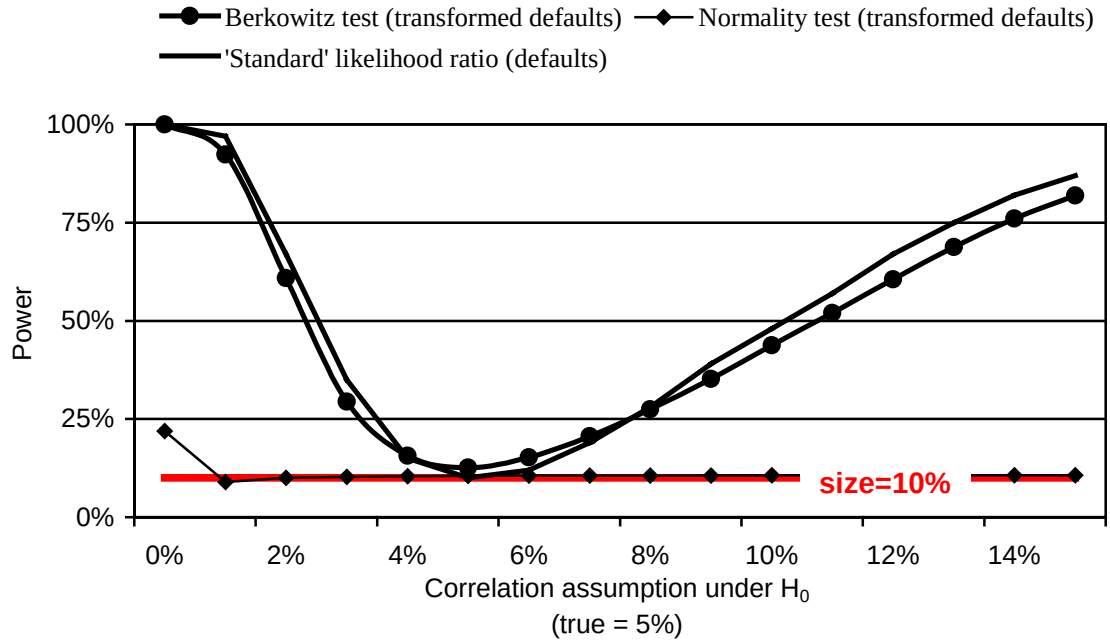

Figure 3: Power of Berkowitz test when including cross-sectional information

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.

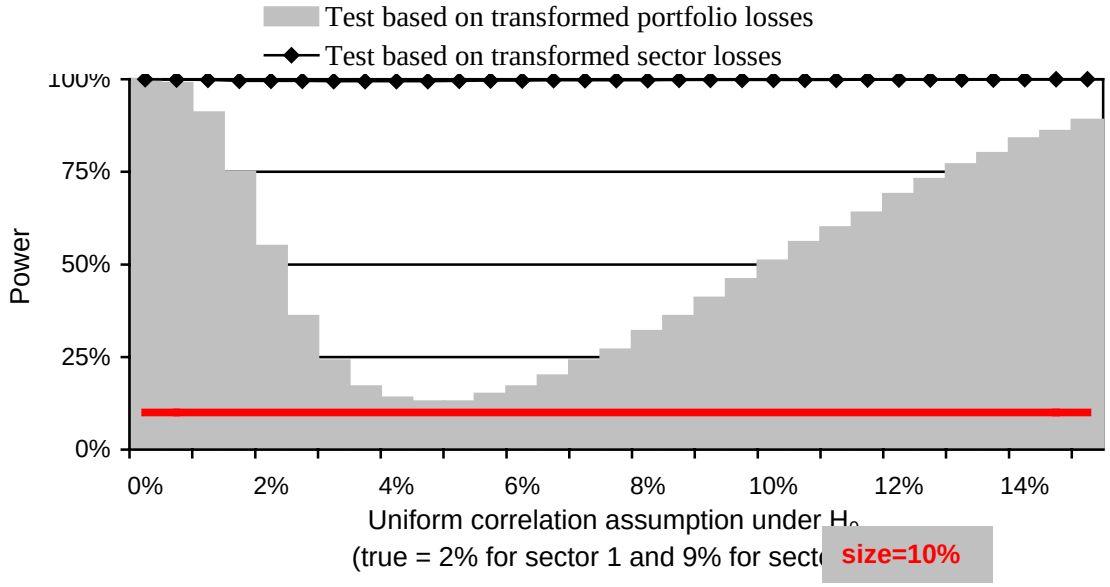

Figure 4: Power of model comparisons

Tests are performed for the base case (see Table 1). To derive the power of the Bonferroni bounds, we apply the Berkowitz test to both models under comparison (A,B) using a significance level of 5% instead of 10%. The power of the likelihood ratio test is the probability that the likelihood ratio $\log(L_A/L_B)$ is positive and larger than the 90% quantile of $\log(L_A/L_B)$ simulated under model B.