Evaluating credit risk models using loss density forecasts

Hergen Frerichs
Chair of Banking and Finance, University of Frankfurt (Main), PO Box 11 19 32, 60054 Frankfurt (Main), Germany.

Gunter Löffler
Chair of Banking and Finance, University of Frankfurt (Main), PO Box 11 19 32, 60054 Frankfurt (Main), Germany.

The evaluation of credit portfolio risk models is an important issue for both banks and regulators. It is impeded by the scarcity of credit events, long forecast horizons, and data limitations. To make efficient use of available information, the evaluation can be based on a model's density forecasts, instead of examining only the accuracy of point forecasts such as value-at-risk. We suggest the Berkowitz (2001) procedure, which relies on standard likelihood ratio tests performed on transformed loss data. We simulate the power of this approach to detect misspecified parameters in asset value models, focusing on asset correlations. Monte Carlo simulations show that a loss history of ten years can be sufficient to resolve uncertainties currently present in credit risk modeling. The power is better for two-state models than for multi-state models, and it can be improved by incorporating cross-sectional information.

1 Introduction

Portfolio credit risk models quantify potential losses (and gains) from holding a portfolio of risky debt.[1] Their outcome is a probability distribution for the value effects of credit-related events, which usually has a one-year horizon. Some models restrict the analysis to losses from defaults, while others include the effects of credit quality changes. In the literature on these models, it is customary to refer to the difficulties of evaluating their quality. 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

We wish to thank Ron Anderson, Wolfgang Bühler, Philippe Jorion, Jan Pieter Krahnen, Thilo Liebig, Ludger Overbeck, Peter Raupach, Mark Wahrenburg, an anonymous referee, 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.

short to produce reliable results. To date, there is no study that clarifies to what extent these doubts are justified.

We show through simulations that the statistical power of validation procedures can be satisfactory if they assess the adequacy of the complete credit portfolio loss distribution, instead of examining the accuracy of individual point forecasts such as value-at-risk. For this purpose, we recommend the 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 annual 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 borrowers' asset values. According to the Basel Committee on Banking Supervision (2001a), 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. The result is robust to the number of borrowers, the choice of the significance level, and to heterogeneity and noise in default probabilities. If the asset correlation is misspecified within a multi-state model which includes migration and recovery risk, the power is lower than in the case of two-state models which forecast only the risk of default.

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 can again be above 90%. Default probability estimates that lead to similar errors in value-at-risk figures as incorrect asset correlations or distributions are identified with a larger probability because they also lead to false predictions of mean default rates. An incorrect assumption about the autocorrelation in the time series of the systematic factor is not easily detected.

To exploit information contained in the cross-section of credit losses, the Berkowitz procedure can be adapted to jointly test the validity of predictions for subportfolio 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. 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%. 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. One could argue that such a misspecification is irrelevant for practical purposes, because the aggregate credit risk prediction is almost correct. The argument would neglect the fact that credit portfolios change through time, while model adequacy is assessed using several years of historical data. Consider the following uses of the model described above: a bank allocating capital based on estimated credit portfolio risk, investors monitoring the risk of a managed corporate bond portfolio, or, in some future time, regulators setting capital requirements according to estimated credit portfolio risk. Once the model described above is approved for any of these purposes, credit managers have incentives to increase exposures to the second sector whose default probability is underestimated by the model. Estimated credit risk would remain constant, while expected returns, in the presence of a default risk premium, would rise.

Papers that either empirically evaluate credit portfolio risk models or theoretically develop statistical evaluation methods are rare. The only empirical paper is Nickel, Perraudin and Varotto (2001), who use two different credit risk models to predict the credit value-at-risk of a large portfolio of dollar-denominated eurobonds. The authors compare the expected and the realized number of value-at-risk violations, but do not conduct a formal test of the models' validity. The only theoretical paper is Lopez and Saidenberg (2001), who propose cross-sectional resampling techniques in order to make more efficient use of available data. In a former version of this paper (Frerichs and Löffler, 2002), we point out a difficulty of tests that are based on resampled portfolios. Credit losses in resampled portfolios will be cross-sectionally dependent, which, if unaccounted for, can bias statistical inference. Other credit risk papers, like Carey (1998) and Carey (2001), simulate credit loss distributions based on empirical data, and examine which portfolio characteristics are important to credit value-at-risk. As credit data are hardly available, other studies employ stylized portfolios to analyze the variance of risk measures across credit risk models (Gordy, 2000), and portfolio types (Gordy, 2000, Kiesel, Perraudin and Taylor, 2001). The Berkowitz (2001) test procedure has been implemented in a couple of studies in fields other than credit risk. Clements and Smith (2000) compare three different validation techniques for models to forecast macroeconomic variables: the approach of Diebold, Gunter and Tay (1998), Berkowitz (2001) and a normality test recommended by Doornik and Hansen (1994).[2] The authors suspect that the Berkowitz (2001) test and the normality test might be sensitive to outlier observations. De Gooijer and Zerom (2000), however, in another comparative study of prediction evaluation criteria for models to forecast interest rates cannot confirm this conjecture.

The paper is organized as follows. Section 2 describes the framework for the evaluation of test procedures. Section 3 assesses the power of the Berkowitz (2001) test procedure using Monte Carlo simulations. Section 4 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 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, ie, the size of the test. If the H_0-model is incorrect, the rejection frequency, ie, the power of the test, should be as large as possible.

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 (2001a) as well as industry models such as CreditMetrics and KMV PortfolioManager.[3]

We examine two variants, which differ in their complexity:

(i) Initially, 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.[4]

(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 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.[5] 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 $\tilde{Z}$ and idiosyncratic factors $\tilde{\varepsilon}_i$:[6]

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

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 $p_i \mid Z$ equals

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

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.[7] 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. The assignment of bank internal ratings is based on the current default probability, usually measured over a one-year horizon (Carey and Hrycay, 2001). 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 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-state model, all we need in addition is the individual default probabilities p_i. To model the full loss distributions, we assume G rating categories; the last category G 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. Forward rates are fixed as we ignore market risk. 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

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

where m_R is the mean recovery rate assumed for the loans.[8] $\tilde{Z}$ 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.

In the following, we describe the parameters used for most of the analyses; we refer to this set of assumptions as the base case. Base case assumptions are summarized in Table 1. They will be varied to check the robustness of the results. We consider two-state models with zero recovery in the case of default. Portfolios are homogeneous in terms of default probability, asset correlations, and loan size. We assume that the available data sets comprise $T = 10$ years of annual data on the number of defaults within homogeneous portfolios of $N = 10{,}000$ borrowers. The common factor is assumed to be serially uncorrelated, and asset values follow a standard normal distribution. We choose an

TABLE 1 Base case setup

Parameter	Value
Number of possible states	2
Recovery in case of default	0
Number of borrowers (N) in portfolio	10,000
Constant unconditional 1-year default probability (p)	1%
Uniform asset correlation in true data-generating model (w^2)	5%
Asset value distribution	$N(0, 1)$
Serial correlation of systematic factor	None
Forecast horizon (years)	1
Length of credit loss history (T years)	10
Test size / Type-I error	10%

unconditional default probability of 1% for each obligor and a uniform asset correlation of $w^2 = 5\%$ for all pairs of borrowers. Both values are consistent with a random effects probit analysis of Standard & Poor's rating data from 1982–1999, which yields unconditional default probability estimates of 1.2% (1.1%) and uniform asset correlation estimates of 3.9% (6.0%) for all issuers (BB issuers).[9]

In H0 models, either the asset correlation, or the asset value distribution, or default probabilities are changed from the base case. For each H_0 model, we first determine the predicted loss distribution using $K = 1{,}000{,}000$ random loss scenarios. Based on a simulated T-year loss history, we calculate a test statistic for the adequacy of the H_0 model. A test's power is assessed based on $S = 10{,}000$ tests. For each test, we generate an independent T-year loss history from the true credit risk model.

To summarize, a simulation exercise that determines the probability of rejecting a specific credit risk model is structured as follows:

(1) Specify portfolio size and portfolio structure.

(2) Specify data availability, ie, the number of years for which credit losses are observed ($= T$).

(3) Specify the true, data generating credit risk model. Includes full parameterization for the portfolio from (1).

(4) Specify the H_0 model, ie, the credit risk model whose validity is to be tested. Includes full parameterization for the portfolio from (1). Use Monte Carlo simulation ($K = 1{,}000{,}000$ trials) to obtain the loss distribution predicted by this model for the portfolio structure chosen in (1).

(5) Based on (1) to (3), randomly generate a data set comprising T years of annual credit losses, that is, each observation of the loss history is drawn from the true, data generating credit risk model.

(6) Conduct the statistical test for the validity of the H_0 model given the loss history from (5).

(7) Repeat steps (5) to (6) $S = 10{,}000$ times. The test's power is the relative frequency with which the H_0 model is rejected in the course of these S trials.

Step (7) requires critical values of the test statistic. In most cases we directly refer the statistic to its asymptotic distribution. For the tests using cross-sectional information, asymptotic values are inappropriate, which is why we simulate critical values as described in Section 3.3.

3 Evaluating credit risk models based on the entire distribution

In this section, we apply the Berkowitz (2001) procedure to the evaluation of portfolio credit risk models. The evaluation is based on a model's density forecasts, not on the accuracy of individual point forecasts such as value-at-risk.[10] Specifically, the observed loss history is transformed such that one obtains a series of standard normally distributed variables if the risk model is correct. 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

$$x_t = \hat{F}(y_t) = \int_{-\infty}^{y_t} \hat{f}(u)\,\mathrm{d}u \tag{4}$$

where y_t are observed losses and $\hat{f}(u)$ is the forecasted probability of a loss of u.[11] 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 $\Phi(\cdot)$

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

resulting in a series of transformed observations z_t which is iid $N(0,1)$ if the predicted distribution function is correct.[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 credit loss data in order to assess their power.[13]

3.1 Two-state models

3.1.1 Alternative models differ in asset correlation assumption

In this section, we compare asset value models with one systematic factor and a uniform mutual asset correlation (all parameters as in Table 1). 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 series of observed credit losses z_t (which was introduced in the last section):

$$\log L = -\frac{T}{2}\log 2\pi - \frac{T}{2}\log\sigma^2 - \sum_{t=1}^{T}\frac{(z_t-\mu)^2}{2\sigma^2} \tag{6}$$

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

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

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

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

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

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, as well as in two variations in which the true asset correlation is 10% or 20% instead of 5% as in the base case. As the asset correlation of the null hypothesis moves away from the true value, the test's power increases the faster, the lower the true asset correlation is. If the true asset correlation equals 5%, the power is larger than 50% if the assumed correlation is below 2.5% or above 10.5%. For an asset correlation equalling 10% (20%), the corresponding values are 5% (11%) and 19% (32%). If the null hypothesis posits a zero asset correlation, it is rejected in 100% of all cases.

If the null hypothesis coincides with the true model, the power is slightly higher than the nominal significance level of 10%. Due to the small sample size,

FIGURE 1 Power of Berkowitz test depending on the true asset correlation.

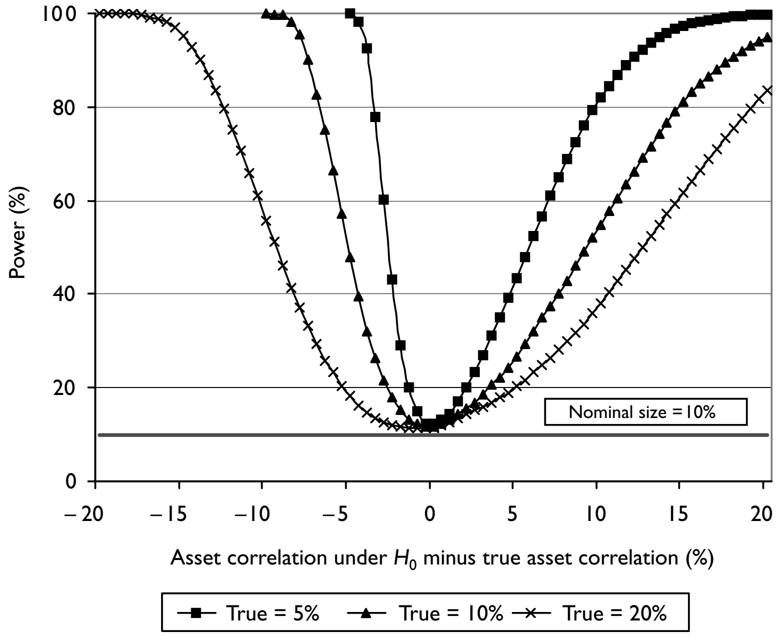


TABLE 2 Power of Berkowitz test if H_0 varies in asset correlation

H_0		Power in base case (%)	Power in variations of base case (%)					
Correlation (%)	99% quantile of default distribution (base case)		Size = 5% (vs. 10%)	1,000 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	99.9	99.8	87.3	100	99.7	99.6	99.8
1	181	91.2	87.3	54.5	88.8	72.4	89.7	91.0
2	221	60.2	50.2	28.4	54.3	42.0	55.0	56.8
3	257	29.2	19.6	16.1	25.9	23.5	26.6	27.6
4	289	15.0	8.4	12.9	14.1	16.5	14.2	14.7
5% = true	321	12.1	6.1	13.7	12.4	15.3	11.7	12.5
6	352	14.4	8.0	16.6	14.8	16.3	14.7	14.4
7	382	19.8	11.7	21.2	19.7	18.7	20.2	19.1
8	411	26.8	16.4	27.2	27.0	21.8	27.4	25.2
9	440	34.8	22.3	34.2	34.8	25.7	35.5	32.6
10	468	43.2	29.2	41.3	42.7	29.9	44.7	39.7
11	497	52.2	36.6	48.3	51.1	34.0	53.5	47.1
12	526	60.8	44.5	55.3	60.0	38.8	61.8	54.9
13	554	68.5	52.6	61.6	67.9	43.4	69.8	62.2
14	582	75.7	60.3	68.2	75.1	48.3	76.6	69.0
15	610	81.8	67.6	73.8	80.8	52.7	82.4	74.8
16	639	86.7	74.1	78.7	86.0	57.4	87.2	80.2
17	667	90.5	80.3	82.8	89.8	61.8	90.6	84.4
18	695	93.5	84.9	86.5	92.8	65.9	93.5	88.2
19	724	95.7	88.9	89.7	95.0	69.9	95.4	90.9
20	753	97.1	92.3	92.3	96.6	73.7	96.9	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 3) that is similar to the high quality credit portfolio in Gordy (2000).

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 for the true asset correlation of 5% depicted in Figure 1 are also shown in column three of Table 2, 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. In our two-state setting with zero recovery rate, the 99% quantile of the loss distribution corresponds to the number of defaults that is exceeded only in 1% of all trials. Columns 4–9 of Table 2 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 probability is 0.5% instead of 1%
- ❑ the portfolio is heterogeneous in terms of default probabilities. Rather than assuming a uniform default probability of 1% we split the portfolio into seven rating classes (Table 3). 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 probability 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 loss history is five instead of ten. With heterogeneous default probabilities, the power decreases modestly.

TABLE 3 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

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 (2001a) favors an asset correlation of 20%, calibration exercises (Gordy, 2000, and Hamerle, Liebig and Rösch, 2002) typically lead to much lower correlation estimates.[14] Often, the estimates are smaller than 5%. In Table 2, 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 3). 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 2). 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.

3.1.2 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 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, and ten observations). The simulated power is presented in Table 4.

When comparing the power to the previous results, it is illustrative to compare null hypotheses that produce similar errors in predicting extreme losses, eg, 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 2), 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. One piece of evidence against this assumption is presented in Gordy and Heitfield (2001) and Löffler (2002), who fit Merton-style structural models to empirical rating transitions and find that these are best replicated by fat-tailed asset value distributions.

A general specification that allows for different degrees of tail-thickness 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 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. An example shall help to assess the power statistics shown in Table 4. Fitting a *t*-distribution to empirical rating tran-

TABLE 4 Power of Berkowitz test if H_0 varies in default probability or asset value distribution

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 H0**	**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.

sition matrices provided by KMV, Moody's and Standard & Poor's, Löffler (2002) obtains degrees of freedom parameters that are always below 11. This could lead a risk manager to favor a *t*-distribution with 10 (or less) degrees of freedom. If the normal assumption is correct, and there are ten years of credit data to check whether a *t*-distribution with 10 degrees of freedom is appropriate, the power is 100% Conclusions do not change when we look at the opposite case in which the true asset value distribution is a *t*-distribution. 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).

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

$$\tilde{Z}_t = 0.5\tilde{Z}_{t-1} + 0.866\tilde{u}_t, \quad \tilde{u}_t \sim N(0,1), \quad \tilde{Z}_1 \sim N(0,1) \tag{10}$$

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 (2001), we augment the density function for the transformed losses z_t by allowing them to follow a first-order autoregressive process:

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

As the estimator for the autocorrelation coefficient r 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 $\rho = -0.114.6$. 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 H0 posits an asset correlation of 10% (true being 5%), the power is 43% 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.

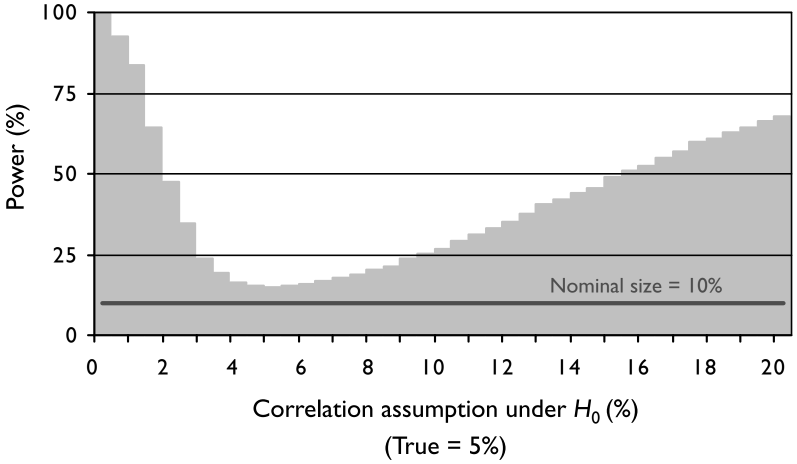
3.2 Multi-state models

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 3). The probabilities of rating transition are taken from Lando and Skodeberg (2002, Table 3). As described in Lando and Skodeberg, the estimates are based on continuous rating data from Moody's over the period 1988–98.[17] 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 were taken from the CreditMetrics web site on April 8, 2002.[18] 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 (four-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. 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 2 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 68% if the asset correlation under the null hypothesis is set to 20%. Comparing these results with Table 2, 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, an asset correlation of 20% leads to an unexpected loss that is 1.7 times

FIGURE 2 Power of Berkowitz test in multi-state case with systematic recovery.



The Berkowitz test is applied to the heterogeneous portfolio from Table 3. The one-year transition matrix is taken from Lando and Skodeberg (2002, Table 3). The probability mass of the not-rated category has been apportioned to the other rating classes according to their probability mass. Conditional one-year ahead loan values are calculated based on yields on US treasuries and yield spreads for US corporates taken from the CreditMetrics web site on 8 April 2002. All loans mature in five years. 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. 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).

higher than the unexpected loss that is obtained with a 5% asset correlation. In the two-state base case (see Section 3.1.1) 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.

3.3 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:

$$\begin{aligned} &\Delta \tilde{A}_i = w_i \tilde{Z} + \sqrt{1 - w_i^2}\, \tilde{\varepsilon}_i, \\ &w_i^2 = 0.02 \ \text{ for } \ i \in \text{sector } 1,\ w_i^2 = 0.09 \ \text{ for } \ i \in \text{sector } 2 \end{aligned} \qquad (12)$$

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 that posits a uniform asset correlation of 5%. With a size of 10%, the power is only 12% (Figure 3). 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:

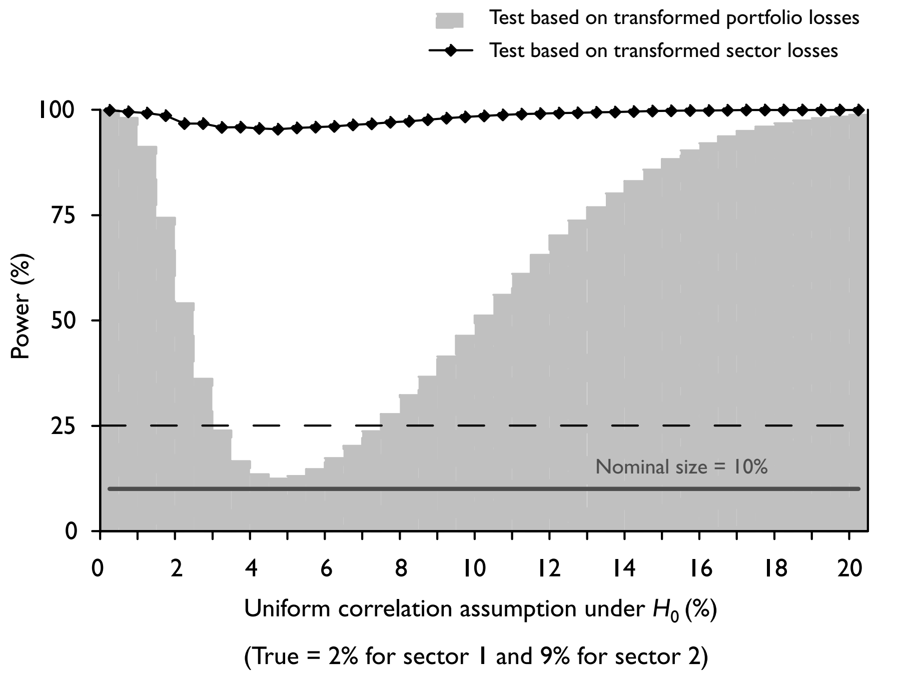
$$\log L = -T\log 2\pi - T\log\sigma_1 - T\log\sigma_2 - \frac{T}{2}\log(1-\rho_{12}^2) - \frac{1}{2(1-\rho_{12}^2)}\sum_{t=1}^{T}\left[\left(\frac{z_{t1}-\mu_1}{\sigma_1}\right)^2 - 2\rho_{12}\left(\frac{z_{t1}-\mu_1}{\sigma_1}\right)\left(\frac{z_{t2}-\mu_2}{\sigma_2}\right) + \left(\frac{z_{t2}-\mu_2}{\sigma_2}\right)^2\right] \qquad (13)$$

We obtain maximum likelihood estimators for the parameters μ_1, μ_2, σ_1^2, σ_2^2 and ρ_{12} and construct a likelihood ratio statistic to jointly test the restrictions $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$. Due to the short time series, the statistic is not exactly chi-squared distributed with four degrees of freedom. Therefore, we simulate the critical value corresponding to a test size of 10%.[19]

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 95.7% 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 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. If one wants to test whether a model is too parsimonious (as in the example) one would split the portfolio into subsets one believes to be different. Similarly, the choice will arise naturally once evaluators have defined a bench-

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. Due to the short time series, the likelihood ratio test statistic is not chi-squared distributed, so we simulated the critical value corresponding to a size of 10%.

mark model for evaluation purposes. In such a case, evaluators would determine the portfolio-split such that the 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 above or below the median difference. If a model departs from the benchmark in several dimensions, the split would aim at maximizing differences in predicted subportfolio risk.

By extending the bivariate likelihood (13) 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 (2000), 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.[20] 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.

4 Concluding remarks

We have described procedures for evaluating parameterizations of asset value models. Because of the structural similarities of current generation credit risk models, we are confident that our results apply to other credit risk models as well. 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%. In particular, we showed that misspecifications in asset correlations can be identified by the Berkowitz procedure. Results are largely robust to portfolio size and composition. However, the test's power is significantly better for two-state models than for multi-state models. Whereas the test's power in identifying a misspecified asset value distribution is considerable, incorrect assumptions about the autocorrelation of the systematic factor are not easily detected.

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, 2001a, b). Since many banks have insufficient records of credit losses, such a validation cannot yet be performed separately for each bank. However, the data that is at hand can be used to check whether the assumptions are adequate on average. In addition, regulators could encourage banks with sufficient loss records to test whether the Basel assumptions are consistent with them. If not, these banks could be allowed to change the parameters which determine capital requirements. To perform such an exercise, as few as five years of data can be sufficient. If a bank correctly specifies the asset correlation to be 5%, whereas Basel prescribes a value of 15%, for example, simulations suggest that the Basel value could be rejected with a probability larger than 50%. Of course, the application of evaluation procedures is not restricted to the regulatory domain. Many banks allocate economic capital based on credit portfolio risk models. Evaluation procedures described in this paper can help to confirm, or improve the chosen model specification. Using Monte Carlo simulations, a bank can assess the power of the tests when applied to its specific data set, and then decide how much weight the results should receive in the specification process.

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 years of default data 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.

[1] A useful summary of available credit risk models is given in Crouhy, Galai and Mark (2000).

[2] 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.

[3] 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.

[4] Basel Committee on Banking Supervision (2001a), S.36.

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

[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. Burgisser, Kurth, and Wagner (2001) show how systematic recovery risk can be modeled in the CreditRisk+ framework.

9. Hamerle, Liebig, and Rösch (2002)

10. Simple quantile tests 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.

11. We use the term "observed losses" to denote the dollar amount of losses. In a two-state credit risk model, this amount coincides with the number of defaults provided all exposures equal one dollar and recovery is zero. Equivalently, one could set y_t equal to percentage portfolio losses.

12. See Berkowitz (2001) for a proof.

13. 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 applied the Doornik and Hansen (1994) normality test, a test based on transformed statistics of skewness and kurtosis to improve small-sample performance, to the transformed series z_t, and simulated its power. In the base case, the power barely exceeded the size if the null hypotheses were defined by choosing an asset correlation from the interval [0%, 20%].

14. Recently, the Basel Committee on Banking Supervision (2001b) proposed to use asset correlations between 10% and 20%.

15. For example, the H0 default probabilities for obligors rated BB are 0.53% or 1.59% instead of 1.06%.

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.

17. We shifted the probability mass of the not-rated category to the other rating classes according to their probability mass.

18. The data are available on request.

19. We set the H_0 model equal to the true model, simulate 1,000,000 likelihood ratio tests, and choose the critical value to be the 90% quantile of the simulated test statistics.

20. 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 below 10^{-300}. The probability of obtaining an even mixture of sectors is 2%.

REFERENCES

Andrews, D. W. K. (1993). Exactly median-unbiased estimation of first order autoregressive/unit root models. *Econometrica* **61**, 139–65.

Basel Committee on Banking Supervision, (2001a). The internal ratings-based approach, Basel.

Basel Committee on Banking Supervision, (2001b). Potential modifications to the committee's proposals, Press release, 5 November.

Belkin, B., Suchower, S., and 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., and 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,** 465–74.

Burgisser, P., Kurth, A., and Wagner, A. (2001). Incorporating severity variations into credit risk. *Journal of Risk* **3**, 50–76.

Carey, M. (1998). Credit risk in private debt portfolios. *Journal of Finance* **53**, 1363–87.

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., and Hrycay, M. (2001). Parameterizing credit risk models with rating data, *Journal of Banking and Finance* **25**, 197–270.

Clements, M. P., and Smith, J. (2000, Evaluating the forecast densities of linear and non-linear models: Applications to output growth and unemployment. *Journal of Forecasting* **19**, 255–76.

Crouhy, M., Galai, D., and Mark, R. (2000). A comparative analysis of current credit risk models. *Journal of Banking and Finance* **24**, 59–117.

De Gooijer, J. G., and Zerom, D. (2000). Kernel-based multistep-ahead predictions of the US short-term interest rate. *Journal of Forecasting* **19**, 335–53.

Diebold, F. X., Gunther, T.A. and Tay, A.S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review* **39**, 863–83.

Doornik, J. A., and 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.

Frerichs, H., Löffler, G. (2002). Evaluating credit risk models: A critique and a proposal. Working paper, University of Frankfurt (Main).

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**, 119–49.

Gordy, M., and Heitfield, E. (2001). Of Moody's and Merton: a structural model of bond rating transitions. Working paper, Board of Governors of the Federal Reserve System.

Gupton, G. M., Finger, C.C., and Bhatia, M. (1997). CreditMetrics – Technical document, New York.

Gupton, G. M., Gates, D., and Carty, L.V. (2000). Bank loan loss given default, Moody's Investors Service, Global Credit Research, November.

Hamerle, A., Liebig, T., and Rösch, D. (2002). Credit risk factor modeling and the Basel II IRB approach, Working paper, University of Regensburg, Deutsche Bundesbank.

Kiesel, R., Perraudin, W., and Taylor, A. (2001). The structure of credit risk. *Journal of Risk*, forthcoming.

Koyluoglu, H. U., and Hickman, A. (1998). Reconcilable differences. *Risk* **11**(10), 56–62.

Lando, D., and Skodeberg, T. (2002). Analyzing rating transitions and rating drift with continuous observations. *Journal of Banking and Finance* **26**, 423–44.

Löffler, G. (2002). Implied asset value distributions. Working paper, University of Frankfurt (Main).

Lopez, J. A., and Saidenberg, M. R. (2000). Evaluating credit risk models. *Journal of Banking and Finance* **24**, 151–65.

Lucas, A., Klassen, P., Spreij, P., and Straetmans, S. (2001). An analytic approach to credit risk of large corporate bond and loan portfolios. *Journal of Banking and Finance* **25**, 1635–64.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* **29**, 449–70.

Nickel, P., Perraudin, W., and 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* **B11**, 68–84.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics* **23**, 470–2.