Evaluating credit risk models: A critique and a new proposal✧

Hergen Frerichs*

Gunter Löffler

University of Frankfurt (Main)

February 14, 2001

Abstract

Evaluating the quality of credit portfolio risk models is an important question for both banks and regulators. Lopez and Saidenberg (2000) suggest cross-sectional resampling techniques in order to make efficient use of available data and to produce measures of forecast accuracy. We first show that their proposal disregards cross-sectional dependence in simulated subportfolios, which renders standard statistical inference invalid. We proceed by suggesting another evaluation methodology which draws on the concept of likelihood ratio tests. Specifically, we compare the predictive quality of alternative models by comparing the probabilities that observed data have been generated by these models. The distribution of the test statistic can be derived through Monte Carlo simulation. To exploit differences in cross-sectional predictions of alternative models, the test can be based on a linear combination of subportfolio statistics. In the construction of the test, the weight of a subportfolio depends on the difference in the loss distributions which alternative models predict for this particular portfolio. This makes efficient use of the data, and reduces computational burden. Monte Carlo simulations suggest that the power of the tests is satisfactory.

Key words: credit risk, backtesting, model validation, bank regulation

JEL classification: G2; G28; C52

*Corresponding author: Hergen Frerichs, Chair of Banking and Finance, University of Frankfurt (Main), P.O. Box 11 19 32, 60054 Frankfurt (Main), Germany. Phone: ++49-69-79828959, facsimile: ++49-69-79822143, e-mail: frerichs@wiwi.uni-frankfurt.de.

✧ The paper is part of a joint research project with Deutsche Bundesbank on modeling credit portfolio risk. We wish to thank Jan Pieter Krahnen, Thilo Liebig, Ludger Overbeck, Peter Raupach, Mark Wahrenburg and seminar participants at the University of Frankfurt (Main) for helpful comments.

1. Introduction

In the literature on credit risk, it is commonplace to refer to the problems of evaluating the quality of credit portfolio risk models. Several years after the first models have been proposed, there is only one paper which empirically tests their predictive ability (Nickell, Perraudin and Varotto, 1999). One explanation for this scarcity of research are concerns that evaluation procedures borrowed from the literature on market risk models have little power when applied to credit portfolios.

To make efficient use of available data, Lopez and Saidenberg (2000) propose evaluation methods for credit portfolio risk models based on cross-sectional simulation. In the construction of the tests they assume that prediction errors of portfolios resampled from the loss experience of one year are independent. As we demonstrate in this paper, this will not be the case in a typical setting. If the economy moves into recession, for instance, credit losses will be above average both in the entire sample and in randomly drawn subsamples. This renders standard statistical inference invalid.

The major part of this paper is thus devoted to the development of reliable and powerful test procedures for the evaluation of credit risk models. We propose to compare the predictive quality of alternative models by comparing the probabilities that observed data have been generated by these models. The test statistic is based on a ratio of likelihoods, even though it is not a likelihood ratio test in the traditional sense. Its distribution can be derived using Monte Carlo simulation. To exploit differences in cross-sectional predictions of alternative models, the likelihood ratio criterion is applied to subportfolios, and the test statistic is formed as a linear combination of the subportfolio ratios.

Unlike Lopez and Saidenberg (2000), we do not recommend to resample these portfolios randomly from the entire portfolio. Rather, the weight of a subportfolio depends on the difference in the loss distributions which alternative models predict for this particular portfolio. This makes efficient use of the data and reduces computational burden. To gain intuition, consider a portfolio whose obligors belong to one of two sectors. The correct default probability is 1% for the first, and 3% for the second sector. 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 entire portfolio or random subsets, the inadequacy of the

model will not be revealed as the expected default rate will be 2% in either case.

We examine the power of our proposed test procedures using Monte Carlo simulations. The power appears to be satisfactory. With 10 years of data on annual defaults, for example, it is similar to a market risk setting where two risk models are compared based on 60 observations. For ease of exposition, the simulations are conducted for simple two-state credit risk models which ignore migration and recovery risk. However, the tests can easily be applied to more complex models, and they do not require that the models under comparison are nested.

Our test procedure follows recent papers from the market risk literature in that tests are based on the entire portfolio distribution rather than on quantiles (e.g. Berkowitz, 1999, or Crnkovic and Drachman, 1996). Related papers include Carey and Hrycay (2001), who discuss various resampling strategies to construct expected loss distributions from a default history. Sobehart, Keenan and Stein (2000) proposes techniques for assessing the quality of individual default rate estimates, an important input to credit risk models. Löffler (2000) quantifies the effects of input parameter uncertainty on the reliability of credit risk models. A useful summary of available credit risk models is given in Crouhy, Galai and Mark (2000). To the best of our knowledge, our proposed approach is novel in two ways: first, in the use of likelihoods to evaluate credit portfolio risk models, and second, in the judicious use of subportfolios to increase the tests' power.

The paper is organized as follows. Section 2 describes the methodology for the evaluation of test procedures. Section 3 discusses the tests proposed by Lopez and Saidenberg (2000). Section 4 presents our proposal and assesses its power. Section 5 contains sensitivity analyses, and Section 6 concludes.

2. Methodology for the evaluation of test procedures

A natural way for evaluating the performance of test procedures is to employ a Monte Carlo study. We simulate a large number (10,000) of random default histories which are all generated by one specific credit portfolio risk model. We then state the null hypothesis that the default history is governed by some model specification, choose a significance level, and apply a statistical test separately for each simulated default 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 in the simulations should be equal to the chosen significance level. If the H_0-model is incorrect, the rejection frequency, i.e. the power of the test, should be as large as possible.

We consider credit risk models whose output is a distribution of the expected number of defaults within a portfolio. Thus, we neglect both migration risk and recovery rate uncertainty, which renders a discrete distribution of expected and realized outcomes. The framework we apply is similar to a two-state version of CreditMetrics[1]. Default correlations are modeled based on the asset value model of Merton (1974). There, a firm defaults if its asset value falls below a critical level defined by the value of liabilities. Correlations of changes of asset values thus translate into default correlations.

In a two-state world, the credit portfolio risk models which have been developed in the last few years are similar in the underlying structure and produce nearly identical outputs when parameterized appropriately.[2] For this reason, we are confident that our results are of general interest even though we examine only one class of credit portfolio risk models. In addition, the test procedures put forward in this paper can also be used in more complex settings, e.g. when migration risk is added.

To capture asset correlations, asset value changes ΔA_i are modeled through a *K*-factor model:

$$\Delta A_i = \sum_{k=1}^{K} w_{i,k} Z_k + \sqrt{1-\sum_{k=1}^{K} w_{i,k}^2}\,\varepsilon_i\,, \qquad (1)$$

with Z_k denoting the systematic factors, and ε_i the idiosyncratic risk of obligor *i*.[3] The Z_k and ε_i are assumed to be normally distributed with zero mean and a covariance matrix equal to the identity matrix. As a consequence, asset value changes ΔA_i also follow a standard normal distribution.

The factor sensitivities w_i determine asset correlations. In the simplest form of (1), a one-factor model (*K=1*) with constant factor sensitivities $w_i=w$, the mutual asset correlation is equal to w^2. Default correlations can be calculated via the bivariate

[1] Cf. JP Morgan (1997) for a general description of CreditMetrics.
[2] Cf. Finger (1998), Koyluoglu and Hickman (1998), Gordy (2000), and Wahrenburg and Niethen (2000).
[3] Cf. Finger (1999), Koyluoglu and Hickman (1998), and Belkin, Suchower and Forest (1998b) for

normal distribution. A borrower defaults whenever $\Delta A_i < \Phi^{-1}(p_i)$, where p_i is the unconditional default probability and Φ denotes the cumulative standard normal distribution function. Given the realizations of the systematic factors Z_k, the conditional default probability $p_i|Z_k$ equals

$$p_i \mid Z_k = \Phi\left[\frac{\Phi^{-1}(p_i) - \sum_{k=1}^{K} w_{i,k} Z_k}{\sqrt{1 - \sum_{k=1}^{K} w_{i,k}^2}}\right] . \qquad (2)$$

Expected loss distributions are generated through a Monte Carlo simulation involving 1,000,000 scenarios for the number of defaults in the portfolio.[4] To avoid zero probabilities we fit exponential trends to the number of defaults, separately for both tails of these discrete distributions. For the left tail of the distribution, the regression parameters are estimated over an interval starting at the first point on which at least 10 random scenarios fell, and ending at the first point with at least 200 scenarios. All numbers of defaults with less than 10 scenarios are replaced by the predicted trend. For the right tail of the distribution, we proceed in the same manner. In the parameterizations used in this paper, the R^2 of the trend regressions is always larger than 85%.

In the base case, we assume that evaluators of credit risk models observe 10 years of annual data on homogeneous portfolios of 10,000 borrowers. The portfolios are homogeneous in terms of constant model parameters and portfolio composition across time. Portfolio weights are irrelevant as the evaluators only examine the number of defaults. Throughout the paper, we set the unconditional annual default probability equal to 1% for each obligor. In the base case, we also assume that there is no serial correlation of defaults across time. The justification is that a credit portfolio risk model should eliminate any serial correlation by conditioning on current information.

---

applications of this model.

[4] If unconditional default probabilities are equal across obligors, 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. For descriptive statistics of simulated expected loss distributions see appendix.

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 credit loss distribution is forecasted and compared with the subportfolio's observed number of defaults. In a sense, the number of observations available for model evaluation is thus multiplied by the factor *R*.

Lopez and Saidenberg propose tests for the predicted expected loss, predicted quantiles and the predicted entire credit loss distribution. In the derivation of the tests, the authors assume that observed prediction errors are independent and that a model's accuracy can be examined using standard testing procedures. As it turns out, this assumption is invalid.

The lack of independence of subportfolio defaults results from the fact that any sub-portfolio defaults drawn from a one-year default experience mirror the realization of the systematic factors in this specific year. In a bad year, subportfolios will have a relatively high number of defaults on average. In a good year, there will typically be few defaults in the resampled portfolios. One consequence of this dependence is that there will be cases where there is no violation of the predicted value at risk across all subportfolios pertaining to one year.[5]

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. We proceed as follows:

1. Simulate 10,000 10-year default histories based on a one-factor asset value model with an unconditional default probability of 1% and an asset correlation of

[5] 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 0.01 and a zero default correlation. If the chosen subportfolio size is 500, the 90% quantile for subportfolio defaults is 10. With a probability of 46%, however, the overall number of defaults in the entire portfolio is 9 or less. That is, in 49% of all cases you would not observe a violation of the 90% quantile in any of the random subportfolios.

5% (i.e. $w_i^2 = 0.05$ for all *i*).

2. Draw 1,000 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 borowers, respectively.
3. Simulate expected credit loss distributions of several credit risk models, one of which is equal to the one used for generating the default histories. While keeping the unconditional default rate constant at 1%, we vary the assumed asset correlations (1%, 2.5%, 5%, 7.5%, and 20%).
4. Implement one of the test statistics proposed by Lopez and Saidenberg and calculate the rejection frequencies for all models in question.

We concentrate on the quantile test proposed by Lopez and Saidenberg.[6] Under the assumption that the predicted quantiles are accurate and observed violations of the quantiles are independent, these violations are draws from a binomial distribution. Whether or not the percentage of observed violations $\hat{\alpha}$ is equal to the chosen confidence level α can be tested using the likelihood ratio statistic

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

where *y* is the number of violations across the *T*R* subportfolios. In the simulations, we examine the performance of a test using the 90% quantile, using a significance level of 10%.

As a preliminary analysis, we compute the frequency of 90%-quantile violations across all simulated subportfolios. As we generate 10,000 10-year default histories and draw 1,000 subportfolios for each year, there are 100,000,000 subportfolios in our sample. The results are summarized in Table 1.

[6] Lopez / Saidenberg (2000), p. 160.

Table 1: Simulated frequency of 90%-quantile violations (correct correlation assumption is 5%)

Correlation Assumption	Frequency of 90%-quantile violations for a subportfolio size of		
	2,000	5,000	8,000
1.0%	19.1%	20.5%	20.7%
2.5%	13.7%	14.6%	14.7%
5.0%	10.0%	9.8%	9.9%
7.5%	7.2%	7.3%	7.2%
20.0%	3.5%	3.2%	3.2%

Since the test is based on the 90%-quantile we expect the frequency of quantile violations to be 10% for the correct model, which has an asset correlation of 5%. This is indeed the case. (The small deviations from 10% are due to simulation error.) Falsely assuming lower correlations leads to a higher frequency of violations, and vice versa. The results demonstrate that the chosen models differ considerably from each other. The subportfolio size does not affect the average frequency of quantile violations.

We now calculate the likelihood ratio test statistic (3) separately for each of our 10,000 histories. The test results are summarized in Table 2. The frequency with which the null hypothesis is rejected decreases with increasing risk (see column three). With a subportfolio size of 20%, the rejection frequency equals 95.8% for a credit risk model with 1% correlation. Viewed in isolation, this would indicate that the test is powerful. For a null hypothesis of 20% correlation, the rejection frequency equals 69.6% which still is a satisfactory result. Disturbing, however, is the rejection frequency of 88.1% for the 5%-correlation model. As this is the model which has generated the histories, the figure should be 10% which is the chosen size of the test. The result suggests that the test is of little use since the probability that the true model will be rejected is, first, much higher than the assumed size and, second, higher than the probability of rejecting a model with a false correlation assumption.

A second problem of the quantile test is the large number of scenarios where there are no violations at all (see column 4). In these cases the test statistic is not defined. The number of scenarios without any violations rises as either the correlation parameter is increased relative to the correlation of the default history or a larger subportfolio size is used. A higher correlation implies higher risk and a wider credit loss distribution. Thus, the probability of extreme events is higher as well. The

Table 2: Simulated performance of the Lopez and Saidenberg quantile test (correct correlation = 5%, significance level of quantile test = 10%)

Subportfolio size	H_0 (Correlation)	Rejection frequency of H_0	Runs with no violation	Runs with LR(False)>LR(True)
2,000	1.0%	95.8%	0.2%	65.8%
	2.5%	91.6%	1.2%	54.9%
	5.0%	88.1%	2.4%	
	7.5%	85.6%	4.2%	59.0%
	20.0%	69.6%	23.2%	72.9%
5,000	1.0%	89.6%	2.0%	64.8%
	2.5%	78.4%	6.6%	53.0%
	5.0%	65.7%	15.4%	
	7.5%	57.6%	23.2%	43.4%
	20.0%	37.7%	49.2%	60.0%
8,000	1.0%	81.2%	4.2%	60.4%
	2.5%	64.0%	12.3%	45.0%
	5.0%	48.5%	24.2%	
	7.5%	39.3%	33.7%	29.8%
	20.0%	20.5%	61.5%	50.3%

second pattern arises because, for a given realized default rate, the probability of drawing subportfolios with a high default frequency goes down when the subportfolio size is increased. This results from the fact that the number of defaults in the resampled subportfolios follows a hypergeometric distribution.

The last column of Table 2 will be interesting when comparing the Lopez and Saidenberg quantile test with the test statistic which we will propose in the next section. It displays the frequency that the likelihood ratio statistic $LR(\alpha)$ of the false model is higher than that of the true one. With a subportfolio size of 20% the frequency is always above 50% and increases with the distance to the true model. When the subportfolio size is increased, the values drop below 50% for some correlation assumptions. In the next section, it will be shown that the corresponding figures for our test statistic are better.

To sum up, as the *R* subportfolios are not cross-sectionally independent the standard test statistics proposed by Lopez and Saidenberg cannot be used. This also holds if

one tested the complete loss distribution instead of the 90%-quantile, or ran Mincer-Zarnowitz regressions to examine unbiasedness of the forecasted credit losses. Each of the tests proposed by Lopez and Saidenberg requires independent draws.

One might think of modifying the test procedure by drawing subportfolios with replacement. This would mitigate some of the problems, but the contemporaneous dependence of subportfolio losses would remain. Another possible solution would be to remove the dependence by conditioning the forecasts of subportfolio losses 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 constant factor sensitivity *w*, and which differ only in the assumed sensitivity *w*. Another remedy would be to simulate the distribution of the test statistics in the presence of cross-sectional dependence. This is one element of the alternative test procedures which we will present in the next section.

4. Evaluating models based on likelihoods

4.1 Outline of the methodology

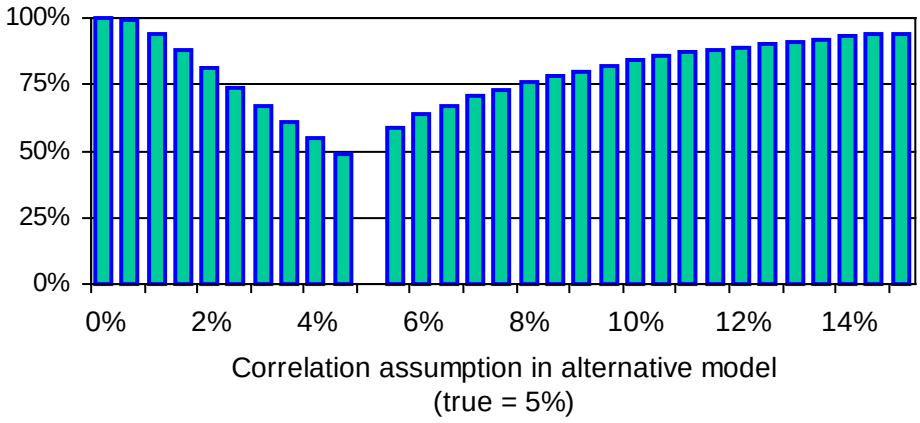
The basic idea of our approach is to examine the likelihood that a given default history has been generated by a particular credit portfolio risk model. Consider a history of *T* annual observations on the number of defaults in a portfolio. Call this empirical distribution *h(t)*. Credit risk model A produces an expected distribution of the number of annual defaults. For each possible number of defaults *x*, it specifies $p_A(x,t)$, the probability of the number of defaults being equal to *x* in year *t*. The probability of observing a default history *h(t)* under model *A* is thus

$$L_A = \prod_{t=1}^{T} p_A(h(t),t) \qquad (4)$$

This holds true if the annual losses are serially independent or any existing dependence is reflected in the model's prediction of losses.[7]

A first insight into the ability of this likelihood criterion to discriminate between different specifications of credit portfolio risk models can be given by a simple

Figure 1: Simulated frequency of observing likelihood (true model) > likelihood (alternative model) for a one-factor model with uniform sensitivities



simulation experiment:

1. Generate 10,000 10-year default histories based on a one-factor asset value model with an unconditional default probability of 1% and an asset correlation of 5% (i.e. $w_i^2 = 0.05$ for all *i*).

2. For each of the 10,000 histories, calculate likelihoods for different correlation assumptions while keeping the default probability constant.

3. Compute the frequency that the likelihood of the 5%-correlation model is larger than that of an alternative model.

Figure 1 shows this frequency for various correlation assumptions. In several of the mutual comparisons, the figure is higher than 75%. When setting a correlation assumption of 0% against the correct 5%, the observed likelihood of the latter is in all simulations larger than that of the former. These results indicate that the likelihood criterion is useful for discriminating between alternative models. Remember that we computed similar statistics for the Lopez and Saidenberg quantile test. There, the frequency figures for a subportfolio size of 2,000 ranged from 54.9% for the 2.5%-correlation assumption to 72.9% for the 20%-correlation assumption. This compares to 74% for the 2.5% correlation and 94% for the 15% correlation in Figure 1.

While there is some information in a simple comparison of likelihoods, it is not yet a test statistic for which we can assess the power given some significance level of the

[7] This assumption is tested in section 5.

test. We therefore use the likelihood criterion as part of a test statistic for the comparison of two competing credit portfolio risk models A and B. Specifically, we construct the following likelihood ratio statistic:

$$\lambda = -2\log\left(\frac{L_B}{L_A}\right), \tag{5}$$

where L_A is the observed likelihood of model A.

In standard statistics, maximum likelihood estimation involves setting up a likelihood function and maximizing it for a set of parameters. The likelihood ratio test is used for hypothesis testing in this setting. Imposing a valid restriction on the unrestricted maximum likelihood should not lead to a large reduction in the log-likelihood function. Since the unrestricted likelihood must always be larger than the restricted likelihood and both likelihoods cannot be negative, the likelihood ratio will lie between 0 and 1. Under regularity conditions, the large sample distribution of the likelihood ratio test statistic λ is chi-squared with degrees of freedom equal to the number of restrictions imposed.[8]

In our setting, there are two sensible ways of using a likelihood ratio test:

1. We perform an optimization within a specified class of credit portfolio risk models in order to obtain the model parameters which maximize the likelihood (4). Then we test restrictions on this optimum using a likelihood ratio test. For example, we might have another default correlation estimate derived from equity correlations (as is done in CreditMetrics) and would like to test the null hypothesis that this estimate is true given a default history.
2. We have two sets of parameters specified separately from the default history, and would like to use the associated likelihood to discriminate between these parameterizations. For example, it might be the case that we infer different default correlations from equity data and, alternatively, from default rate volatilities (as is done in CreditRisk$^+$).[9] Even if one agreed on using equity correlations to infer

[8] Cf. Greene (2000), p. 152

[9] Wahrenburg and Niethen (2000) compare Value-at-risk estimates of CreditMetrics and CreditRisk$^+$ in a two-state setting (default / no default) for the German construction industry. Default correlations are estimated from equity correlations for CreditMetrics, while they are inferred from default volatilities for CreditRisk$^+$. The 5%-value-at-risk estimates differ by a factor up to 3 depending on the number of obligors. Cf. also Crouhy, Galai and Mark (2000) and Koyluoglu, Bangia, and Garside (1999).

Table 3: Model specifications

	Model A (taken to be the true model)	Model B (representing the H_0-hypothesis)
Case 1	$\Delta A_i = vZ + \sqrt{1-v^2}\,\varepsilon_i$	$\Delta A_i = wZ + \sqrt{1-w^2}\,\varepsilon_i$
Case 2	$\Delta A_i = v_i Z + \sqrt{1-v_i^2}\,\varepsilon_i$, $v_i = v'$ for $i \le 5000$, $v_i = v''$ for $i > 5000$	$\Delta A_i = w\,Z + \sqrt{1-w^2}\,\varepsilon_i$,
Case 3	$\Delta A_i = vZ_1 + \sqrt{1-v^2}\,\varepsilon_i$	$\Delta A_i = \sum_{k=2}^{3} w_{ik} Z_k + \sqrt{1-\sum_{k=2}^{3} w_{ik}^2}\,\varepsilon_i$, $w_{i1} = w \cdot I_{i\le 5000}$, $w_{i2} = w \cdot I_{i>5000}$

default correlations, the use of different time periods or return frequencies could produce conflicting parameter estimates. When conducting such mutual comparisons, it is useful but not necessary to start with an a-priori assumption on which of two rival models is the better one. An obvious criterion for this decision is the observed likelihood, i.e. one would test the null hypothesis that the model with the smaller likelihood is correct. For such a test, it is not necessary that the two alternative models are nested in each other.

An important characteristic of many credit risk portfolio models is that the models' parameters are not estimated using a history of the variable which is to be predicted. Credit risk models predict the distribution of losses, while the parameters are estimated, for example, based on equity returns (in the case of CreditMetrics), or equity prices and volatilities (in the case of the KMV model). Thus, we believe that evaluators of credit risk models often face a situation in which they have to choose between alternative a-priori specifications. In addition, the models under analysis may exhibit structural differences which complicate the standard testing procedure. For these reasons, we focus on the second way of constructing a test, and examine the first only in the course of the sensitivity analyses in section 5.

If we do not maximize the likelihood function in the first place, the test statistic, even though it is a ratio of likelihoods, does not belong to the class of likelihood ratio tests used in standard statistics. In consequence, the likelihood ratio statistic will not be

distributed as a chi-squared variable.[10] However, we can simulate the distribution of the test statistic under the null hypothesis and then work with this statistic as we do with standard ones. In particular, we can freely choose a significance level, and assess the power of the test.

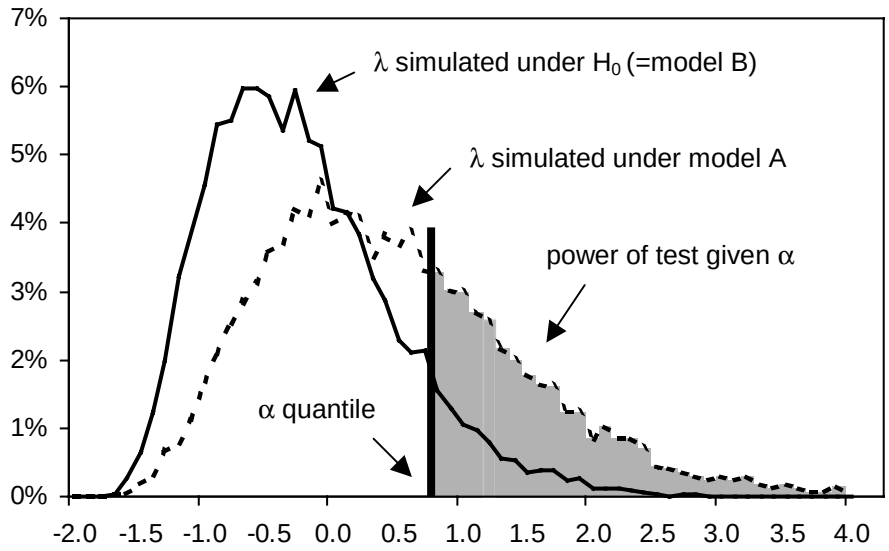
In the base case setting, the construction of the test involves two steps (cf. Figure 2):

1. Set up the null hypothesis: parameters = parameter set of model B
2. Calculate the likelihood ratio statistic for 10,000 scenarios of 10-year histories under H_0 (= Model B), and compute the (1-α)-quantile of this distribution. In an application of the test, the null hypothesis is rejected when the test statistic exceeds this quantile. Thus, the type I error, which is the probability of rejecting the null hypothesis when it is true, equals α.

The power of the test can be assessed as follows:

3. Calculate the likelihood ratio statistic (5) for 10,000 scenarios of 10-year histories under model A and compute the frequency that the test statistic under model A exceeds the (1-α)-quantile simulated in step 2. This is the power of the test if model A is correct. The type II error, which indicates the likelihood of not rejecting the null hypothesis when it is false, is equal to one minus the power.

Figure 2: Illustration of the likelihood ratio test statistic



[10] Cf. Greene (2000), p. 153.

4.2 Test specifications and their power

In this section, we will suggest variants of the likelihood ratio test introduced above, and determine their power for selected specifications of factor models (see Table 3). In each case, we will take alternative A to be the correct model, and test the hypothesis that model B is correct.

We start by comparing two one-factor models with different sensitivities towards the common risk factor (Case 1). We continue by allowing two different sensitivities towards the common risk factor (Case 2). Here, cross-sectional differences between the two rival models provide additional information. Finally, a one-factor model is compared with a two-factor model, so that there are three different risk factors (Case 3). Again, there is information in the cross-section.[11]

Throughout, we do not test alternative assumptions on unconditional default rates, but only assumptions on correlation parameters. The applicability of our results is not affected by this restriction. In practice, there will most probably be more than one combination of default rates and correlation assumptions which results in a default history with an equally high probability, especially in large inhomogeneous portfolios with many risk factors.

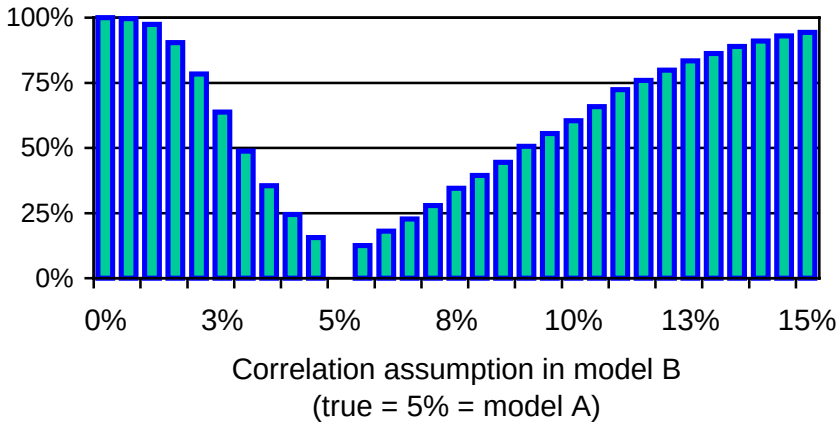
Case 1: The base case

We start by assuming that the observed default history is best explained by a one-factor asset value model with an unconditional default rate of 1% and a constant mutual asset correlation of 5% (model A). We test different null hypotheses by changing the correlation parameter (model B). The test statistic is the likelihood ratio given in (5).

Figure 3 shows the simulated power of our test statistic in the base case, i.e. the test is based on a 10-year history of a portfolio with 10,000 borrowers. We choose a 10% (one-sided) significance level for the test. 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 3% or above 9%.

[11] For descriptive statistics of simulated expected loss distributions see appendix.

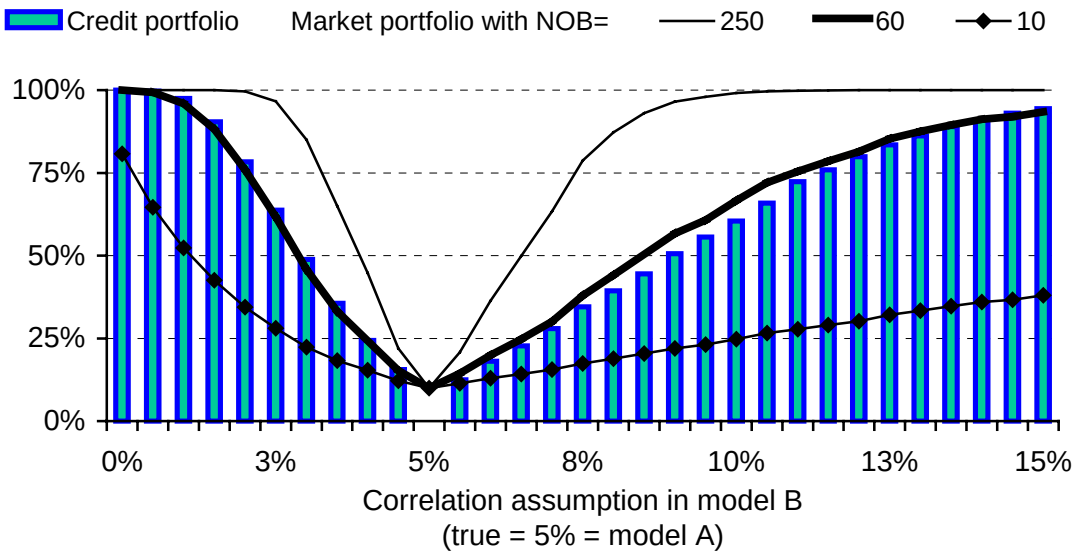
Figure 3: Power of test statistic in Case 1 (size of test=10%)



It is instructive to compare our test statistic's power to the power of an analogous test in a market risk setup. Consider the situation in which we compare two market risk models (model A and model B) whose expected loss distributions are normal with zero mean and 90%-quantiles (10% Value at Risk) equal to those of the credit risk models used in constructing Figure 3. Then we perform a test that Model B correctly predicts the risk of the portfolio given that the portfolio returns are generated by model A. The most powerful test will be based on the sample portfolio variance, and will use the fact that the sample variance divided by the true variance follows a chi-squared distribution. (In the test we assume that the mean is known.) Figure 4 shows the result of this comparison, using a one-sided test and varying the number of observations available for the test. As can be seen, our test statistic's power is similar to a market risk setup in which a test is based on 60 observations. While the power with 250 observations is far better, it is considerably lower with 10 observations - even though our test also uses only 10 observations.[12]

[12] This can be explained by noting that the normal distribution is the distribution with maximum entropy. For a given number of observations, it is thus easier to discriminate between different credit models as more distributional assumptions are imposed.

Figure 4: Power of discriminating credit portfolio models in Case 1 and comparable normally distributed portfolios (size of test = 10%)

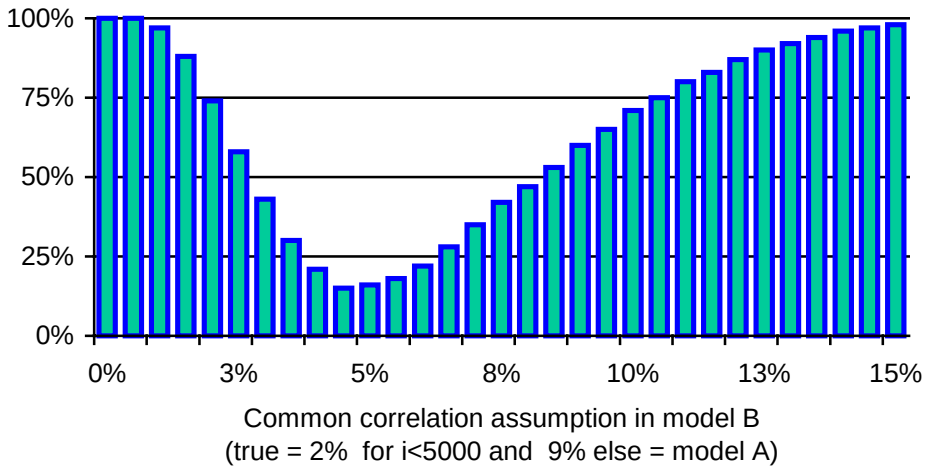


Case 2: Exploiting unconditional cross-sectional differences

So far, we 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 two models which exhibit an identical distribution of overall defaults, yet differ in the prediction about the number of defaults in subsets of the portfolio. The likelihood criterion would not be able to discriminate between the two models because the likelihoods based on the total portfolio defaults would be the same.

As an illustration we analyze a portfolio whose defaults are driven by a one-factor asset value model with an unconditional default rate of 1%. One half of the obligors belongs to a sector with an intra-sector correlation of 2%, the other half to a sector with an intra-sector correlation of 9%. (The inter-sector correlation is 4.2%). If we take this setup to be model A and compare it with an alternative one-factor model B which assumes a common factor sensitivity, the power curve, which is shown in Figure 5, looks very similar to the one in which model A is a one-factor model with 5% correlation (cf. Figure 3). This is due to the fact that the aggregate expected loss distributions of these two A-models are nearly identical in their first and second moments and do not differ much in higher moments, even though the sector portfolio distributions differ. As a consequence, if the null hypothesis posits a uniform

Figure 5: Power of test statistic in Case 2 using the likelihood ratio for aggregate defaults (size of test = 10%)



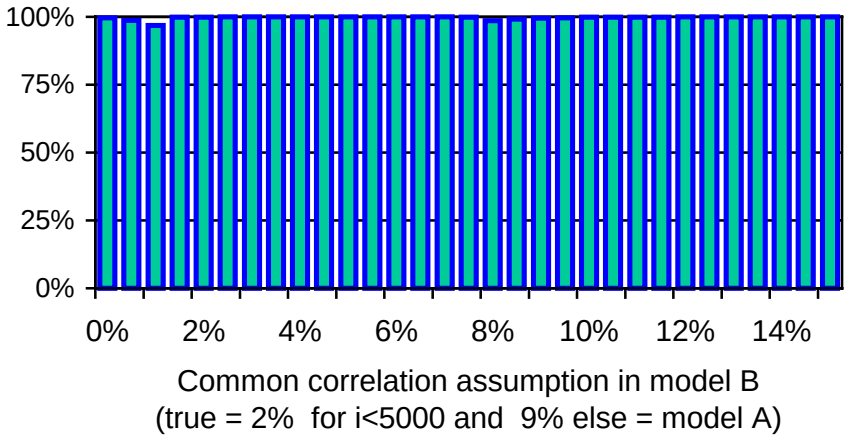
correlation of 5% the power is only 16%, little more than the size of the test.

One possible way of exploiting the cross-sectional information is to utilize the idea of Lopez and Saidenberg and apply the likelihood-criterion (5) to randomly drawn portfolio subsets. This would not make efficient use of the information, though. The main disadvantage of drawing the subportfolios randomly is that we hardly ever get extreme subportfolio compositions. Consider the present setting, i.e. Case 2 in which a large portfolio consists of two sectors, each of which represents half of the portfolio. 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 companies of one sector is extremely low.[13] If the null hypothesis is a 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 are the two models under comparison. 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

[13] 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 an even mixture of sectors is 2%.

Figure 6: Power of test statistic in Case 2 using composite likelihood ratios for aggregate defaults and two subportolios (size of test = 10%)



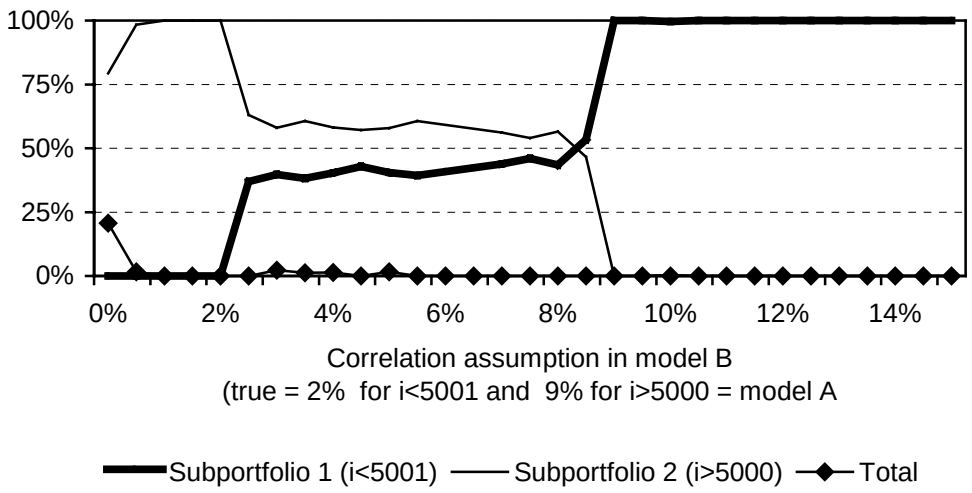
subportfolios and calculate the test statistic for each of them. Thus, if we have a two-sector portfolio and assume that companies of these two sectors have different sensitivities towards a common systematic factor, we form two subportfolios consisting of just one sector and proceed as though we were to test models on two different portfolios. This produces an aggregation problem. We propose to calculate the likelihood ratio test statistic for the entire portfolio (λ_{Total}) as well as for the extreme subportfolios ($\lambda_{Sub-port}$), and then use numerical procedures to determine the optimal weight for each of these figures:

$$\lambda = a_1\lambda_{Total} + a_2\lambda_{Sub-port(1)} + a_3\lambda_{Sub-port(2)} \qquad (7)$$

We optimize the weights a_i by maximizing the frequency that the likelihood of model A is larger than the likelihood of model B under 10,000 simulated histories of model A. Afterwards we proceed as in the base case using the composite likelihood ratio test statistic (7) instead of the simple statistic (5). The histories we use to assess power are different from the ones used for optimizing the weights.

Applying this methodology to our example of a one-factor model A with two intra-sector correlations of 2% and 9%, we are able to enhance the power of the test. The power, depicted in Figure 6, is close to 100% for every specification of model B. The reason for this substantial improvement is that the two correlation parameters of model A are sufficiently different from each other. When we set a model with 1% correlation against model A, for instance, correlation assumptions in the first subportfolio are 1% and 2%, respectively, yielding a low power. Yet, in the second

Figure 7: Standardized optimal weights for the composite test statistic in Case 2 ($\lambda=a_1\lambda_{Total}+ a_2\lambda_{Sub\text{-}portfolio\ 1}+ a_3\lambda_{Sub\text{-}portfolio\ 2}$; size = 10%)



subportfolio we compare two models with 1% and 9% correlation, respectively. Here, the power is large. After optimizing the weights of the two subportfolio statistics, we will completely rely on the second subportfolio for the purpose of discriminating between the two models.

Figure 7 shows the weights of the three components of the composite test statistic which yielded the result shown in Figure 6. The optimal weights are multiplied by the standard deviation of the respective components in order to facilitate interpretation. The pattern of the optimal weights is plausible. For example, when the correlation assumption of model B is equal to or less than 2%, the information value of the subportfolio with a model A correlation of 9% is clearly dominant to the subportfolio with a model A correlation of 2%. For this reason, the weight of the latter is equal to zero. The opposite holds when the model B correlation is above 9%. In between, the weight of the first subportfolio (model A correlation = 2%) is increasing. The optimal weight of the total portfolio statistic always is close to zero, meaning that aggregate defaults do not contain useful information in addition to the subportfolios. One might expect the total statistic to receive a larger weight as the larger portfolio size reduces noise in the observed defaults. In the chosen setting, however, this effect is small. As will be shown in section 5, reducing the number of borrowers from 10,000 to 5,000 leads only to a minor decrease of power.

Case 3: Exploiting conditional cross-sectional differences

Consider the case that model A is a one-factor model with a constant sensitivity v across all obligors and that model B is a two-factor model with constant sensitivities $w_{i1} = w \cdot I_{i \le 5000}$ and $w_{i2} = w \cdot I_{i>5000}$. Thus, model B does not have a common factor.

If v equals w, factor sensitivities are identical for both models. As a result, the unconditional sector portfolio distributions are also identical. If we calculated subportfolio test statistics as before, we would compare two models with a correlation of w^2, and therefore could not extract any information from the statistic. The picture changes as soon as we condition the likelihood criterion on the other sector's default information. For model A, conditioning provides worthwhile information, because both sectors depend on a common risk factor. For model B, conditioning does not change the predicted loss distribution because the two sectors are not subject to a common factor.

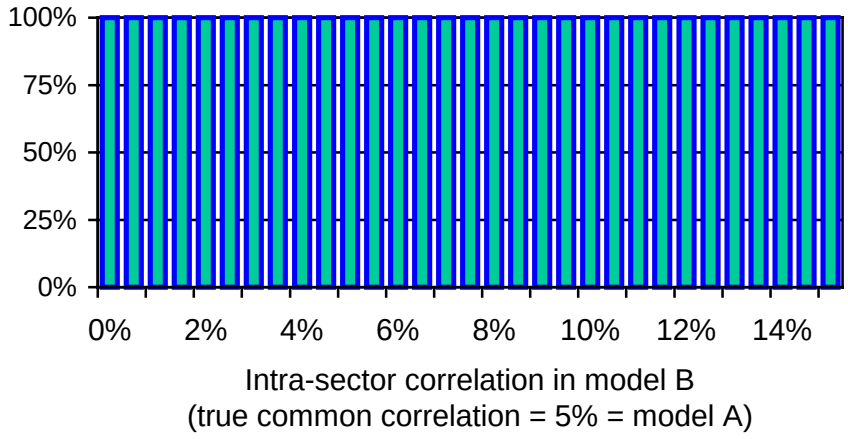
To exploit this difference, we use conditional probabilities to compute the likelihood ratio statistic. In equation (4) we multiplied the probabilities $p_A(x,t)$, the probability of the number of defaults being equal to x in year t under credit risk model A for $t=1, ..., T$. Now we use the conditional probability $p_A(x_1,t| x_2)$ of observing x_1 defaults in sector 1 given that x_2 borrowers have defaulted in sector 2. In the framework of this paper, this conditional probability is the binomial density of the number of defaults in sector 1 being equal to x_1 in year t when the subportfolio size equals y in both sectors, and the conditional default probability equals x_2/y, That is, the conditional default probability is equal to the observed default rate in sector 2. The conditional likelihood of observing a default history $h_1(t)$ for subportfolio 1 under model A is:

$$L_A^C = \prod_{t=1}^{T} p_A(h_1(t), t \mid h_2(t)), \tag{8}$$

and the test statistic for a subportfolio is $\lambda^C = -2\log\left(L_B^C / L_A^C\right)$. Similar to the previous section, we aggregate the total, unconditional likelihood ratio and the conditional subportfolio ratios:

$$\lambda = a_1\lambda_{Total} + a_2\lambda^C_{Sub-port(1)} + a_3\lambda^C_{Sub-port(2)} \tag{9}$$

Figure 8: Power of test statistic in Case 3 using composite likelihood ratios for aggregate defaults and two subportolios (size of test = 10%)



The weights are chosen to maximize the probability that the likelihood of model A is larger than the likelihood of model B under the simulated history of model A.

Figure 8 shows that the use of conditional likelihoods can be very useful for discriminating between models. We vary the factor sensitivities in model B, and hold the correlation in model A constant at v^2=5%. In each case, the power of the aggregate test statistic (9) is 100%.

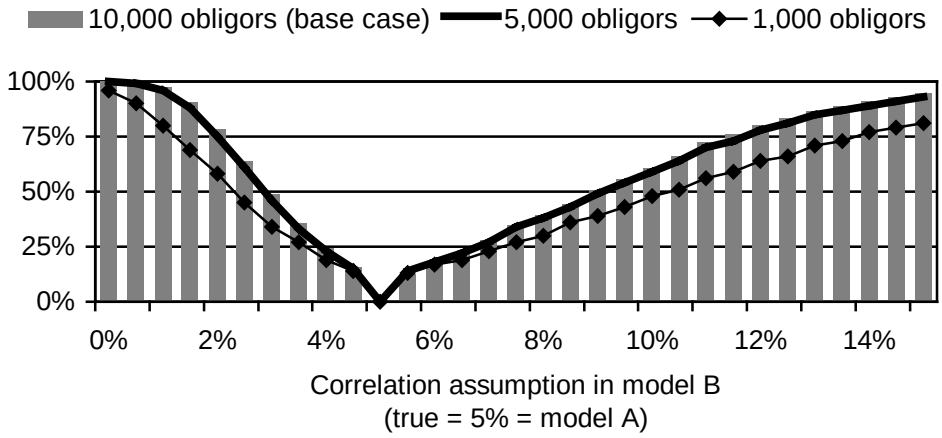
4.3 Summary of the test procedure

In general, we propose the following procedure for comparing two credit portfolio risk models, A and B. Examine the models to determine subportfolios where differences in the unconditional or conditional loss distribution are large. The test statistic aggregates likelihood ratios of the two models, $\lambda = -2\log\left(L_B / L_A\right)$, computed separately for the total portfolio and the chosen subportfolios, and using unconditional or conditional likelihoods:

$$\lambda = a_1\lambda_{Total} + a_2\lambda_{Sub-port(1)} + a_3\lambda_{Sub-port(2)} + \ldots + a_m\lambda^C_{Sub-port(m)} + a_{m+1}\lambda^C_{Sub-port(m+1)} + \ldots + a_M\lambda^C_{Sub-port(M)} \quad (10)$$

The weights are chosen to maximize the probability that the likelihood of model A is larger than the likelihood of model B under the simulated history of model A. Use the optimized weights to simulate the distribution of λ under model B, which is the null hypothesis of the test. Choose a significance level α. For a given data set, estimate λ. The hypothesis is rejected if the empirical λ exceeds the $(1-\alpha)$ quantile of the

Figure 9: Power of test statistic in Case 1 for default histories with different numbers of obligors (size = 10%)



simulated distribution.

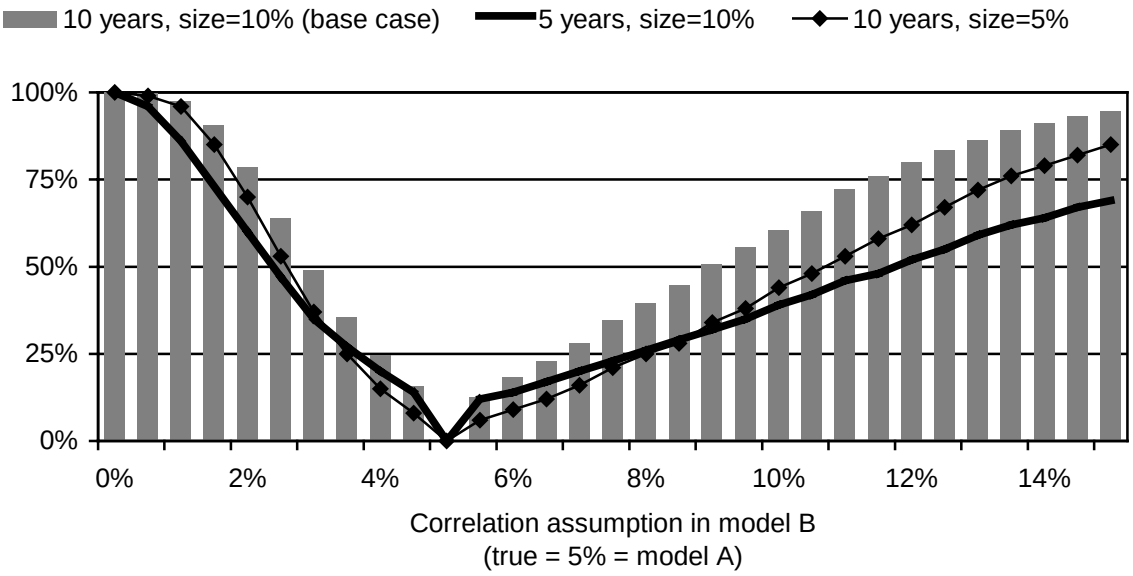
In the first step, the description of the test procedure is vague in the sense that there is no precise recommendation on the nature and number of the chosen subportfolios. In principle, one could choose all possible subportfolios or a random subset thereof, and optimize their weights in the aggregate test statistic. However, the associated computing time is likely to be prohibitive. We do not regard this as a serious limitation of the test procedure as the choice of subportfolios will be evident in many cases. In addition, choosing 'wrong' subportfolios will not decrease the power relative to the benchmark test based solely on the aggregate portfolio distribution.

5 Additional analyses

In the previous section, we illustrated the power of likelihood ratio tests for various credit portfolio risk models which differed in their correlation assumptions. To put these results into perspective, we now vary the general setting of these tests, which was held constant. Specifically, we estimate the power of the test in case 1, where two one-factor models with different sensitivities are set against each other. The following results should thus be compared to Figure 3, which shows the probability of rejecting a model with correlation *x*% when the true model has 5% correlation.

We start by assessing the power when the portfolio under analysis contains only 1,000 or 5,000 borrowers instead of 10,000. Figure 9 shows that the power is lower when the portfolio size is reduced, as should be expected. However, the loss of

Figure 10: Power of the test statistic in case 1 varying the size of the test and the number of years in the default history



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.

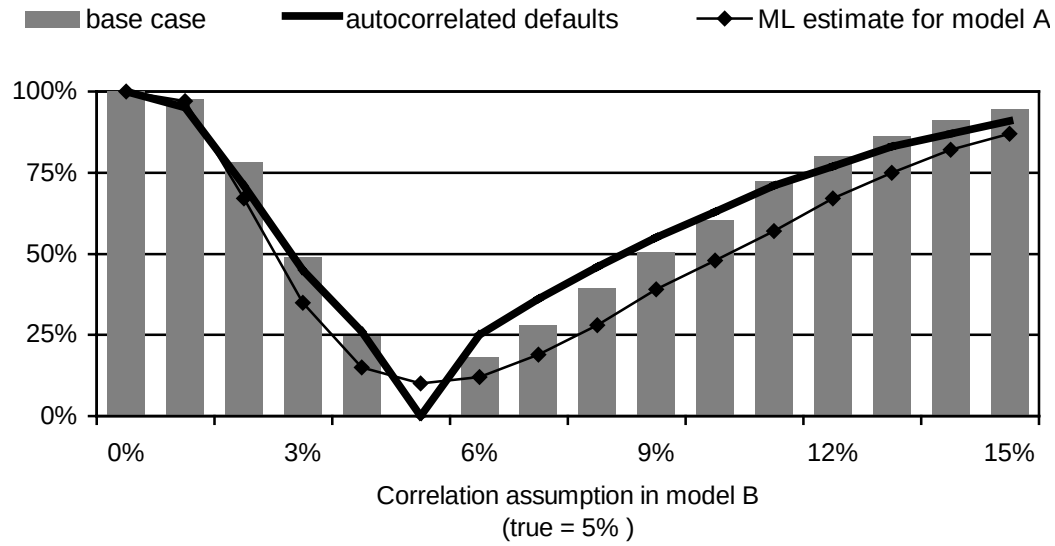
The power also decreases when we lower the size of the test, or reduce the number of years in the default history. Figure 10 compares the power in the base case to the one when the chosen size of the test is 5% instead of 10%, or when the number of years in the observed default history is 5 instead of 10. In general, reducing the number of years has a larger effect. Still, even with only 5 years of data the power appears to be satisfactory.

The final two variations assess the importance of information we presupposed in the construction of the test statistic. We assumed that the prediction errors of the various credit risk models are independent across time. This is a valid assumption if there is no serial dependence in defaults, or credit risk models efficiently use available information in the prediction of defaults. If these conditions are not met, the performance of the tests might deteriorate. We therefore modify the base case by introducing autocorrelation into the time series of the systematic factor Z. In simulating the default histories, we use the following autoregressive process for Z_t:

$$Z_t = 0.5\, Z_{t-1} + 0.866\, u_t, \quad u \sim N(0,1), \; Z_1 \sim N(0,1) \qquad (11)$$

The choice of 0.5 for the autocorrelation coefficient is based on the study of Belkin,

Figure 11: Power of the test statistic in case 1 in the presence of autocorrelation and absent knowledge of the true model (size of test=10%)



Suchower and Forest (1998a), who estimate such a process using annual transition matrices and obtain a autocorrelation coefficient of 0.46. We then test the power of the test as before, assuming that both models A and B do not use the time series information to predict next year's factor realization.

Another piece of information we presupposed was that the rival model A is the true one. In a practical setting, this knowledge will not be given. We thus follow the first way of setting up a likelihood test which we described in section 4.1. Instead of comparing model B to model A, the true model, we estimate the parameters of the model based on maximum likelihood, and test whether model B is a valid restriction. Thus, we construct a standard likelihood ratio test. As before, we simulate its distribution because we cannot rely on the test being distributed as a chi-squared variable. Maximization is done through a simple search procedure in which we evaluate the likelihood for each correlation assumption in the interval [0%, 1%,....., 50%].

The power curves under these two variations are shown in Figure 11. If both the models under analysis and the test neglects serial correlation in defaults, the power does not change significantly relative to the base case. This suggests that autocorrelation does not pose a major problem for the test proposed in this paper. If the evaluator does not know the true model, the power goes down as well, but only modestly. To sum up, the power calculations of section 4 do not appear to depend crucially on the chosen setting.

6 Concluding remarks

We have described a procedure for evaluating rival credit risk models. In essence, we test whether one model is more likely to have generated the data than another alternative model. Monte Carlo simulations show that the power of the tests is satisfactory. With ten years of annual data, for example, it is similar to a market risk setting where two risk models are compared based on 60 observations.

A test should meet other criteria than a large power, for instance ease of implementation and general applicability. The use of the test involves simulations, but the associated computing time is likely to be smaller than when implementing the proposal of Lopez and Saidenberg (2000). 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 losses, we propose to aggregate the test statistics of judiciously chosen subportfolios. Thus, there is no general rule for the design of the test. We do not regard this as a serious shortcoming as the choice of the most important subportfolios will be straightforward in many cases. Note, too, that the test procedures can easily be applied to models which include migration and recovery risk, and that the models need not be nested.

Another possible criticism is that the test is based on the entire range of the distribution, whereas risk managers and regulators are mainly concerned about the probability of extreme events. 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 simulation procedures cannot overcome. Second, differences in the tails of two distributions will 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.

References

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

Crnkovic, C., Drachman, J., 1996. Quality control. Risk 9, No 9, 138-143.

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

Finger, C.C., 1999. Conditional Approaches for CreditMetrics Portfolio Distributions. CreditMetrics Monitor, First Quarter, 14-33

Finger, C.C., 1998. Sticks and stones. Working paper, The RiskMetrics Group, New York.

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

Greene, W.H., 2000. Econometric analysis. 4th International Edition, Upper Saddle River, New Jersey.

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

Koyluoglu, H.U., Bangia, A., Garside, T., 1999. Devil in the parameters. Working Paper, Oliver, Wyman & Company, New York

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

Löffler, G., 2000. The effects of estimation error on measures of portfolio credit risk. Working paper, University of Frankfurt (Main).

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., 1999. Ratings- versus equity-based credit risk modeling: An empirical analysis. Working paper, Bank of England.

Sobehart, J.R., Keenan, S.C., Stein, R.M., 2000. Benchmarking quantitative default risk models: A validation methodology. Moody's Investors Service, New York.

Wahrenburg, M., Niethen, S., 2000. Vergleichende Analyse alternativer Kreditrisikomodelle. Kredit und Kapital, No 2, S. 235 - 257. (With English summary.)

Appendix

A1: Descriptive statistics for expected percentage defaults (10,000 obligors; 1% default rate)

Factor loadings			Standard deviation	Quantiles		
W_1	W_2	W_3		90^{th}	95^{th}	99^{th}
0%	.	.	0.10%	1.13%	1.16%	1.23%
1%	.	.	0.29%	1.38%	1.52%	1.81%
2%	.	.	0.40%	1.53%	1.74%	2.21%
3%	.	.	0.49%	1.65%	1.93%	2.57%
4%	.	.	0.57%	1.74%	2.09%	2.90%
5%	.	.	0.64%	1.83%	2.24%	3.21%
6%	.	.	0.72%	1.91%	2.38%	3.52%
7%	.	.	0.78%	1.97%	2.50%	3.83%
8%	.	.	0.85%	2.04%	2.63%	4.12%
9%	.	.	0.91%	2.10%	2.75%	4.38%
10%	.	.	0.97%	2.15%	2.86%	4.70%
11%	.	.	1.02%	2.20%	2.96%	4.98%
12%	.	.	1.08%	2.24%	3.06%	5.25%
13%	.	.	1.14%	2.29%	3.17%	5.56%
14%	.	.	1.20%	2.33%	3.26%	5.82%
15%	.	.	1.26%	2.36%	3.36%	6.11%
2% / 9%*	.	.	0.65%	1.81%	2.24%	3.30%
.	0%	0%	0.10%	1.13%	1.16%	1.23%
.	1%	1%	0.22%	1.28%	1.38%	1.57%
.	2%	2%	0.29%	1.39%	1.53%	1.82%
.	3%	3%	0.35%	1.47%	1.65%	2.04%
.	4%	4%	0.41%	1.54%	1.76%	2.26%
.	5%	5%	0.46%	1.61%	1.87%	2.46%
.	6%	6%	0.51%	1.67%	1.97%	2.65%
.	7%	7%	0.56%	1.72%	2.06%	2.84%
.	8%	8%	0.60%	1.78%	2.15%	3.03%
.	9%	9%	0.65%	1.83%	2.24%	3.23%
.	10%	10%	0.69%	1.87%	2.32%	3.41%
.	11%	11%	0.73%	1.92%	2.40%	3.59%
.	12%	12%	0.77%	1.96%	2.49%	3.79%
.	13%	13%	0.81%	2.00%	2.56%	3.97%
.	14%	14%	0.85%	2.04%	2.64%	4.17%
.	15%	15%	0.89%	2.08%	2.72%	4.35%

* $w_{1i}= 2\% \cdot I_{i\leq 5000,}$ $w_{1i}= 9\% \cdot I_{i>5000}$ (equals nearly $w_{1i}=5\%$ for all i)

A2: Descriptive statistics for expected percentage defaults (5,000 obligors; 1% default rate)

Factor loadings W_1	Standard deviation	Quantiles 90^{th}	95^{th}	99^{th}
0%	0.14%	1.18%	1.24%	1.32%
1%	0.30%	1.40%	1.54%	1.86%
2%	0.41%	1.54%	1.76%	2.24%
3%	0.50%	1.66%	1.94%	2.58%
4%	0.58%	1.76%	2.10%	2.92%
5%	0.65%	1.84%	2.26%	3.24%
6%	0.72%	1.92%	2.38%	3.52%
7%	0.79%	1.98%	2.52%	3.84%
8%	0.85%	2.04%	2.64%	4.14%
9%	0.91%	2.10%	2.76%	4.42%
10%	0.97%	2.16%	2.86%	4.70%
11%	1.03%	2.20%	2.98%	5.00%
12%	1.09%	2.26%	3.08%	5.28%
13%	1.15%	2.30%	3.18%	5.58%
14%	1.21%	2.34%	3.26%	5.86%
15%	1.27%	2.38%	3.36%	6.14%