Evaluating Credit Risk Models

Jose A. Lopez

Economic Research Department
Federal Reserve Bank of San Francisco
101 Market Street
San Francisco, CA 94105-1530
Phone: (415) 977-3894
Fax: (415) 974-2168
jose.a.lopez@sf.frb.org

Marc R. Saidenberg

Research and Market Analysis Group
Federal Reserve Bank of New York
33 Liberty Street
New York, NY 10045
Phone: (212) 720-5958
Fax: (212) 720-8363
marc.saidenberg@ny.frb.org

Draft Date: June 30, 1999

Acknowledgments: The views expressed here are those of the authors and not necessarily those of the Federal Reserve Bank of New York, the Federal Reserve Bank of San Francisco or the Federal Reserve System. We thank Beverly Hirtle, William Perraudin, Judy Peng, Anthony Saunders, Philip Strahan, and participants at the Bank of England's conference on "Credit Risk Modelling and the Regulatory Implications" for their comments and suggestions.

Evaluating Credit Risk Models

Abstract

Over the past decade, commercial banks have devoted many resources to developing internal models to better quantify their financial risks and assign economic capital. These efforts have been recognized and encouraged by bank regulators. Recently, banks have extended these efforts into the field of credit risk modeling. However, an important question for both banks and their regulators is evaluating the accuracy of a model's forecasts of credit losses, especially given the small number of available forecasts due to their typically long planning horizons. Using a panel data approach, we propose evaluation methods for credit risk models based on cross-sectional simulation. Specifically, models are evaluated not only on their forecasts over time, but also on their forecasts at a given point in time for simulated credit portfolios. Once the forecasts corresponding to these portfolios are generated, they can be evaluated using various statistical methods.

I. Introduction

Over the past decade, banks have devoted many resources to developing internal risk models for the purpose of better quantifying the financial risks they face and assigning the necessary economic capital. These efforts have been recognized and encouraged by bank regulators. For example, the 1997 Market Risk Amendment (MRA) to the Basle Capital Accord formally incorporates banks' internal, market risk models into regulatory capital calculations. That is, the regulatory capital requirements for banks' market risk exposures are explicitly a function of the banks' own value-at-risk (VaR) estimates. A key component in the implementation of the MRA was the development of standards, such as for model validation, that must be satisfied in order for banks' models to be used for regulatory capital purposes.

Recently, there has been a flurry of developments in the field of credit risk modeling, as evidenced by the public release of such models by a number of financial institutions; see J.P. Morgan (1998) and Credit Suisse Financial Products (1997) for examples. Credit risk is defined as the degree of value fluctuations in debt instruments and derivatives due to changes in the underlying credit quality of borrowers and counterparties. Recent proposals, such as by the International Swap Dealers Association (ISDA, 1998) and the Institute of International Finance Working Group on Capital Adequacy (IIF, 1998), argue that credit risk models should also be used to formally determine risk-adjusted, regulatory capital requirements. However, the development of the corresponding regulatory standards for credit risk models is much more challenging than for market risk models.

Specifically, a major impediment to model validation (or "backtesting" as it is popularly known) is the small number of forecasts available with which to evaluate a model's forecast

accuracy. That is, while VaR models for daily, market risk calculations generate about 250 forecasts in one year, credit risk models can generally produce only one forecast per year due to their longer planning horizons. Obviously, it would take a very long time to produce sufficient observations for reasonable tests of forecast accuracy for these models. In addition, due to the nature of credit risk data, only a limited amount of historical data on credit losses is available and certainly not enough to span several macroeconomic or credit cycles. These data limitations create a serious difficulty for users' own validation of credit risk models and for validation by third-parties, such as external auditors or bank regulators.

Using a panel data approach, we propose in this paper several evaluation methods for credit risk models based on cross-sectional simulation techniques that make the most use of the available data. Specifically, models are evaluated not only on their forecasts over time, but also on their forecasts at a given point in time for simulated credit portfolios. Once a model's credit loss forecasts corresponding to these portfolios are generated, they can be evaluated using a variety of statistical tools, such as the binomial method commonly used for evaluating VaR models and currently embodied in the MRA. Note that, since simulated data are used, the number of forecasts and observed outcomes can be made to be as large as necessary.

Although this resampling approach cannot avoid the limited number of years of available data on credit defaults and rating migrations, it does provide quantifiable measures of forecast accuracy that can be used for model validation, both for a given model and across models. These evaluation methods could be used by credit portfolio managers to choose among credit risk models as well as to examine the robustness of specific model assumptions and parameters. Supervisors could use these methods to monitor the performance of banks' credit risk

management systems, either alone or relative to peer group performance.

The paper is organized as follows. Section II provides a general description of credit risk models and highlights two main difficulties with conducting model validation: the lack of credit performance data over a sufficiently long time period and uncertainty about which statistical methods to use in evaluating the models' forecasts. Section III presents the proposed evaluation methodology; i.e., the cross-sectional simulation approach and various statistical tools for forecast evaluation. Section IV concludes with a summary and discussion of future research.

II. General Issues in Credit Risk Modeling

The field of credit risk modeling has developed rapidly over the past few years to become a key component in the risk management systems at financial institutions.[1] In fact, several financial institutions and consulting firms are actively marketing their credit risk models to other institutions. In essence, such models permit the user to measure the credit risk present in their asset portfolios. (Note that such models generally do not measure market-based risk factors, such as interest rate risk.) This information can be directly incorporated into many components of the user's credit portfolio management, such as pricing loans, setting concentration limits and measuring risk-adjusted profitability.

As summarized by the Federal Reserve System Task Force on Internal Credit Risk Models (FRSTF, 1998) and the Basle Committee on Banking Supervision (BCBS, 1999), there exists a wide variety of credit risk models that differ in their fundamental assumptions, such as their definition of credit losses; i.e., default models define credit losses as loan defaults, while

[1] See Altman and Saunders (1997) for a survey of developments over the past twenty years.

mark-to-market or multi-state models define credit losses as ratings migrations of any magnitude. However, the common purpose of these models is to forecast the probability distribution function of losses that may arise from a bank's credit portfolio.[2] Such loss distributions are generally not symmetric. Since credit defaults or rating changes are not common events and since debt instruments have set payments that cap possible returns, the loss distribution is generally skewed toward zero with a long right-hand tail.

Although an institution may not use the entire loss distribution for decision-making purposes, credit risk models typically characterize the full distribution. A credit risk model's loss distribution is based on two components: the multivariate distribution of the credit losses on all the credits in its portfolio and a weighting vector that characterizes its holdings of these credits. Let N represent the number of credits in a bank's portfolio, and let A_t, an (Nx1) vector, represent the present discounted value of these credits at time t. If the bank's holdings of these credits is denoted as the (Nx1) vector w_b, then the value of bank b's credit portfolio at time t is $P_{bt} = w_b'A_t$. Once P_{bt} has been established, the object of interest is ΔP_{bt+1}, the change in the value of the credit portfolio from time t to time t+1, which is a function of the change in the value of the individual credits; i.e.,

$$\Delta P_{bt+1} = P_{bt+1} - P_{bt} = w_b'A_{t+1} - w_b'A_t = w_b'\left(A_{t+1} - A_t\right) = w_b'\Delta A_{t+1}.$$

A credit risk model, say model m, is characterized by its forecast of ΔP_{bt+1} over a specified horizon, which is commonly set to one year. That is, the model generates a forecast $\hat{F}_m\left(\Delta P_{bt+1}\right)$ of the cumulative distribution function of portfolio losses based on the portfolio weights w_b and

[2] Note that some credit risk models directly forecast the entire loss distribution, while others assume a parametric form for the distribution and forecast its relevant parameters. In this paper, we refer more generally to the output of these approaches as forecasted distributions.

the distribution function of the (Nx1) random variable ΔA_{t+1}.[3]

This ability to measure credit risk clearly has the potential to greatly improve banks' risk management capabilities. With the forecasted credit loss distribution in hand, the user can decide how best to manage the credit risk in a portfolio, such as by setting aside the appropriate loan loss reserves or by selling loans to reduce risk. Such developments in credit risk management have led to suggestions, such as by ISDA (1998) and IIF (1998), that bank regulators permit, as an extension to risk-based capital standards, the use of credit risk models for determining the regulatory capital to be held against credit losses. Currently, under the Basle Capital Accord, regulated banks must hold 8% capital against their risk-weighted assets, where the weights are determined according to very broad criteria. For example, all corporate loans receive a 100% weight, such that banks must hold 8% capital against such loans. Proponents of credit risk models for regulatory capital purposes argue that the models could be used to create risk-weightings more closely aligned with actual credit risks and to capture the effects of portfolio diversification. These models could then be used to set credit risk capital requirements in the same way that VaR models are used to set market risk capital requirements under the MRA.

However, as discussed by FRSTF (1998) and BCBS (1999), two sets of important issues must be addressed before credit risk models can be used in determining risk-based capital requirements. The first set of issues corresponds to the quality of the inputs to these models, such as accurately measuring the amount of exposure to any given credit and maintaining the internal consistency of the chosen credit rating standard. For example, Treacy and Carey (1998)

[3] For simplicity we have assumed that a bank's exposures or portfolio weights are known. If, as in some credit risk models, exposures have a random component, then the object of interest, $\hat{F}_m(\Delta P_{bt+1})$, does not change, but the distribution of the weights, w_b, must also be considered.

discuss some of the difficulties in creating and maintaining internal ratings systems. Although such issues are challenging, they can be addressed by various qualitative monitoring procedures, both internal and external.

The second set of issues regarding model specification and validation are much more difficult to address, however. The most challenging aspect of credit risk modeling is the construction of the distribution function of the (Nx1) random variable ΔA_{t+1}. Changes in the value of these credits will be due to a variety of factors, such as changes in individual loans' credit status, general movements in market credit spreads and correlations between portfolio assets. Thus, in general, a variety of modeling assumptions and parameter values are involved in the construction of a credit risk model's forecasted distribution. However, testing the validity of these model components is limited, mainly because the historical data available on the performance of different types of credits generally do not span sufficiently long time periods.

To evaluate $\hat{F}_m\left(\Delta P_{bt+1}\right)$ or any other distribution forecast, a relatively large number of observations is needed. This evaluation issue is less of a concern for evaluating distribution forecasts generated by VaR models. Such models generate daily forecasts of market portfolio changes, which can then be evaluated relative to the actual portfolio changes. As specified in the MRA, 250 observations (about one year) are currently used in the regulatory evaluation of such models; see Lopez (1999a,b) for a further discussion of the evaluation of VaR models. However, it is difficult to gather a large number of observed credit losses with which to evaluate credit risk models, because these models have much longer forecast horizons, typically one year. Thus, one year generates only one observation of credit losses**.** To literally replicate the evaluation procedure specified in the MRA**,** it would take an unrealistic 250 years of observations to gather

the number specified in the MRA. Even if credit risk models of shorter horizons (say, one month) were used for regulatory evaluations, it would still take over 20 years to gather the specified number of observations.[4]

It is this limited ability to validate credit risk models' forecasted loss distributions with actual credit losses observed over multiple credit cycles that causes severe problems in using them for determining risk-based capital requirements. In this paper, using a panel data approach, we present simulation-based evaluation methods for credit risk models that address this regulatory concern more directly than has been done to date. Our objective is to develop evaluation methods that provide quantifiable performance measures for credit risk models, even in light of the short history of available credit risk data. Although the lack of data is an insurmountable issue, our proposed methods make the most use of the available data for backtesting purposes. Of course, as more data becomes available, these methods become even more useful. In addition, several of the statistical tools employed in these evaluation methods are familiar to regulators, since they are commonly used in the evaluation of VaR models. In the following section, both the intuition and the mathematical details of these methods are described.

III. Evaluation Methodology Based on Simulated Credit Portfolios

As mentioned, the data limitations for evaluating credit risk models are considerable. In terms of a panel dataset, credit data is generally plentiful in the cross-sectional dimension (i.e., N is usually large since many credits are available for study), but scarce in the time dimension.

[4] Note that evaluations of models based on different forecast horizons are not directly comparable. That is, the statistical properties of monthly and annual data can be sufficiently different to prevent inference for annual data to be drawn from an evaluation conducted with monthly data.

This limitation has led the users of credit risk models to construct alternative methods for validating these models.

For example, credit risk models have been evaluated using "stress testing". For this method, a model's performance is evaluated with respect to event scenarios, whether artificially constructed or based on historical outcomes. One possible "stress" scenario would be the simultaneous default of several sovereign borrowers. Once the scenarios are specified, the model's forecasts under these circumstances are examined to see if they intuitively make sense. Although such a practice may provide a consistency check regarding the model's various assumptions, these scenarios generally do not occur.

Recently, researchers have begun comparing the forecasts from different credit risk models given similar assumptions about their underlying parameters; see, for example, Crouhy and Mark (1998), Gordy (1998), and Koyluoglu and Hickman (1998). However, as per the evaluation of VaR models, the ability to compare a credit risk model's forecasts to actually observed outcomes is more desirable. For example, Nickell, Perraudin, and Varotto (1998) use actual prices for a set of publicly traded bonds to compare the performance of credit loss forecasts from two types of credit risk models. In this paper, we present evaluation methods that specifically focus on quantitative comparisons between credit loss distribution forecasts and observed credit losses.

A. Intuition from time-series analysis

As outlined in Granger and Huang (1997), methods commonly used for model specification and forecast evaluation in time-series analysis can be adapted for use with panel-

data analysis, such as credit risk modeling. The general idea behind forecast evaluation in time-series analysis is to test whether a series of out-of-sample forecasts (i.e., forecasts of observed data not used to estimate the model) exhibit properties characteristic of accurate forecasts. For example, an important characteristic of point forecasts is that their errors (i.e., the differences between the forecasted and observed amounts) be independent of each other, which is an outcome of properly specified models. See Diebold and Lopez (1996) for further discussion.

This idea can be extended to the cross-sectional element of panel data analysis. In any given year, out-of-sample predictions for cross-sectional observations not used to estimate the model can be used to evaluate its accuracy. As long as these additional out-of-sample observations are drawn independently from the cross-sectional sample population, the observed prediction errors should be independent. Standard tests for the properties of optimal predictions can then be used to test the cross-sectional model's accuracy.

For evaluating credit risk models, we propose to use simulation methods to generate the additional observations of credit portfolio losses needed for model evaluation. That is, the models in question can be used to forecast the corresponding loss distributions for the simulated portfolios, and these forecasts and corresponding observed losses can then be used to evaluate the accuracy of the models. Since simulation techniques are used, we can generate as many observations as we might need. The simulation method used here to generate these additional credit portfolios is simply resampling with replacement from the original panel dataset of credits; see Carey (1998) for a related methodology in which credit portfolios are randomly drawn from a dataset to construct a nonparametric distribution of credit losses for specific portfolio strategies.

Consider a credit dataset that spans T years of data for N assets.[5] In any given year t, let $\rho \in (0,1)$ denote the percentage of credits to be included in the resampled portfolios. (Note that the choice of ρ is up to the discretion of the forecast evaluator, but should generally be large.) We can now generate the (Nx1) vector w_i, which is the set of portfolio weights for resampled portfolio i, by generating N independent draws from the uniform distribution over the interval [0,1]. For each draw above ρ, the associated credit is assigned a weight of zero and is not included in the resampled portfolio. For each draw below ρ, the associated credit is assigned a weight of one and is included in the resampled portfolio.[6] We would expect a resampled portfolio to contain (ρ * N) credits, on average, from the original set of N credits.[7]

Let $P_{it+1} = w_i' A_{t+1}$ denote the value of resampled credit portfolio i at time t, such that $\Delta P_{it+1} = w_i' \Delta A_{t+1}$. Credit model m can be used to generate the corresponding forecast $\hat{F}_m(\Delta P_{it+1})$ of the cumulative distribution function of the random variable ΔP_{it+1}. For each of the T years, we resample with replacement R times (i.e., i = 1,...,R), where R is a large number (say, 1000).[8] Doing so, we have (T * R) forecasted loss distributions upon which to evaluate the

---

[5] In practice, two types of credits are commonly analyzed using credit risk models. The first type are basic debt instruments, such as loans or debentures. The second type is the expected future exposures arising from derivative positions; i.e., forecasts of the expected size of the credit exposure. In this paper, we take such expected future exposures as inputs to the credit risk models.

[6] Note that for the purposes of model validation, the forecast evaluation should not be greatly affected by using a simple, binary weighting scheme. Of course, other weighting schemes could be used, if so desired. Certainly, the weighting scheme is of major importance when credit risk models are actually being used for risk management purposes.

[7] Note that only unique credit portfolios should be used; if a duplicate were to be drawn, it should be discarded and replaced with another.

[8] The number of possible different portfolios that could be resampled from a set of N credits using a binary weighting scheme is 2^N. The choice of the number of simulations, denoted as R, should be less than that number. The 1,000 simulations suggested here is appropriate for occasions where $N \geq 10$; i.e., $1000 < 1024 = 2^{10}$. Additional research is needed to fully understand the impact of the R, N, and ρ resampling parameters on the

accuracy and performance of model m, as opposed to just T forecasts based on the original credit portfolio.[9]

Given the data limitations discussed, the T available years of credit data for model evaluation may not span a macroeconomic or a credit cycle, not to mention the larger number of such cycles that would be ideally available. Although the proposed simulation method does make the most use of the data available, evaluation results based on just one or a few years of data must obviously be interpreted with care since they reflect the macroeconomic conditions prevalent at that time. As more years of data become available, the resampling of credit portfolios under different economic conditions provides for a sterner and more extensive evaluation of a credit model's forecast accuracy.

B. Evaluation Methods for a Credit Risk Model

Having generated (T * R) resampled credit portfolios, we can generate the corresponding predicted cumulative density functions of credit losses, denoted $\hat{F}_m\left(\Delta P_{it+1}\right)$ with i=1,...,R and t=1,...,T, for model m. To evaluate the model's forecast accuracy, we can examine the properties of certain characteristics of these distributions, such as their means. Below, we focus on three hypothesis testing methods, which are also used for evaluating VaR models.

---

evaluation results.

[9] Alternatively, we could randomly choose both the year and the portfolio weights with replacement a total of (T*R) times to generate the same number of simulated portfolios. If the years are chosen from a uniform distribution, then we would have, on average, R simulated portfolios each year. However, in our discussion, we focus on simulating portfolios within each year.

(i). Evaluating forecasts of expected loss

We can evaluate the accuracy of the model's predicted expected losses by comparing them to the actual observed losses on the resampled portfolios. The predicted expected loss in year t for portfolio i at time t+1, denoted $\hat{\mu}_{mit}$, is simply the expected value of $\hat{F}_m(\Delta P_{it+1})$. If the specified credit risk model is accurate, then the difference between the predicted expected losses and the observed losses should be zero on average across the full simulated sample. Let $L_{it+1} = w_i' \Delta A_{t+1}$ denote the observed loss on resampled portfolio i in year t+1. Note that L_{it+1} can be defined in terms of either losses due to credit defaults or credit migrations.[10]

The null hypothesis that the prediction errors $e_{mit+1} = L_{it+1} - \hat{\mu}_{mit}$ have mean zero implies that the model is accurate. A large number of test statistics are available to test this hypothesis. For example, we can use the intuition of Mincer-Zarnowitz regressions to test this property. Specifically, for the regression $L_{it+1} = \alpha + \beta \hat{\mu}_{mit} + \eta_{it+1}$ (where η_{it+1} is an error term), it should be the case that $\alpha=0$ and $\beta=1$, which implies that the mean of the prediction error is zero.

(ii). Evaluating forecasted critical values

A commonly-used output from credit risk models is a specified quantile or critical value of the forecasted loss distribution; i.e. the dollar amount of losses that will not be exceeded with a given probability. The evaluation of these forecasted critical values, denoted $\hat{CV}_m(\alpha, \Delta P_{it+1})$ for the upper α% critical value from $\hat{F}_m(\Delta P_{it+1})$ with i=1,...,R and t=1,...,T, can be conducted

[10] As mentioned previously, credit losses are due to either defaults or rating migrations (for example, a credit downgrade from AAA to A). Defaults and the associated losses can readily be observed, but rating migrations and corresponding losses may not be. In this paper, we assume that such losses are observable. If they are not or are the outcomes of pricing models, additional uncertainties are introduced into the credit model evaluations, as further discussed in Section D.

along the lines of the evaluation of VaR estimates from market risk models. Nickell, Perraudin, and Varotto (1998) present comparative evidence of this type. Specifically, we can use the binomial method to evaluate these forecasted critical values.

Under the assumption that these forecasted critical values are accurate, the exceptions -- occasions when actual credit losses exceed the forecasted critical values -- can be modeled as draws from an independent binomial random variable with a probability of occurrence equal to the specified α percent. We can test whether the percentage of observed exceptions, denoted $\hat{\alpha}$, equals α using the likelihood ratio statistic

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

where y is the number of exceptions in the sample. This LR(α) statistic is asymptotically distributed as a $\chi^2_{(1)}$ random variable, and the null hypothesis that $\alpha = \hat{\alpha}$ can be rejected at the five percent level if LR(α) > 6.64. As noted by Kupiec (1995) and Lopez (1999a,b), the power of such tests can be quite low in small samples. However, since we can control the sample size by increasing R within certain limits, this should not be such a concern.

(iii). Evaluating forecasted loss distributions

Given that credit risk models generate full distributions of credit losses, it is reasonable to evaluate the accuracy of the (T * R) distribution forecasts themselves. In the context of evaluating VaR models, Crnkovic & Drachman (1996) propose specific tests for evaluating forecasted distributions; see Diebold, Gunther & Tay (1997) and Berkowitz (1999) for further discussion. The object of interest for such hypothesis tests is the observed quantile q_{mit+1}, which is the quantile under $\hat{F}_m\left(\Delta P_{it+1}\right)$ corresponding to the observed credit loss L_{it+1}; i.e.,

$q_{mit+1} = \hat{F}_m\left(L_{it+1}\right)$. These hypothesis tests examine whether the observed quantiles derived under a model's distribution forecasts exhibit the properties of observed quantiles from accurate distribution forecasts. Specifically, since the quantiles of independent draws from a distribution are uniformly distributed over the unit interval, the observed quantiles under model m should also be independent and uniformly distributed. These two properties are typically tested separately.

C. Evaluation Methods Across Credit Risk Models

As suggested by ISDA (1998), there exist credit risk models of different levels of sophistication. Users may, over time, wish to upgrade the model they use or simply change an assumption in their model. The simulation method described above can be used to make meaningful comparisons on the relative performance between credit risk models.[11] As before, we generate the resampled portfolios and the corresponding, forecasted loss distributions from both models, denoted $\hat{F}_{1it}$ and $\hat{F}_{2it}$, for i= 1,...,R and t=1,...,T. Here again, we can compare various elements of these distributions, such as their expected losses (or means) and specific tail regions, using slightly different statistical tools.

(i). Evaluating forecasts of expected loss

To evaluate the two competing sets of expected loss forecasts, we can examine the two

[11] Note that comparisons between default models and multi-state models based credit migrations are complicated by their different definitions of credit losses. The comparisons of forecast accuracy can be conducted since both types of models generate credit loss distributions, but the results may not be as meaningful as comparisons across models with the same loss definition.

sets of forecast errors between the realized and expected losses, denoted $e_{1it+1} = L_{it+1} - \hat{\mu}_{1it}$ and $e_{2it+1} = L_{it+1} - \hat{\mu}_{2it}$. As described by Granger and Huang (1997), a number of tests are available for evaluating these forecast errors.

(a). The count method

Let e^2_{1it+1} and e^2_{2it+1} denote the squared forecast errors for resampled portfolio i. The null hypothesis that the two credit risk models are equally accurate across the R resampled portfolios in a given year is expressed as H_0: $E\left[e_1^2\right] = E\left[e_2^2\right]$. A simple test of this hypothesis is to count the number of times that e^2_{1it+1} is greater than e^2_{2it+1}, which we denote as p_1. If the null hypothesis is correct, then the random variable p_1/R should have a normal distribution with a mean of 0.5 and a variance of 1/(4*R). We can reject the null hypothesis at the five percent level if the observed value of p_1/R lies outside the range $[0.5\text{-}R^{-1/2}, 0.5+R^{-1/2}]$. Although this is not particularly powerful test, it is robust to any covariance between the two forecast errors as well as any heteroskedasticity of the individual errors.

(b). Sum/difference regressions

A second test for evaluating the models' predicted expected losses is based on analyzing the sum and difference of the forecast errors. For a given year, let S_{it+1} be the sum of the e_{1it+1} and e_{2it+1} errors, and let D_{it+1} be their difference. The regression $S_{it+1} = \alpha + \beta D_{it+1} + \eta_{it+1}$ is then run, and the null hypothesis that $\beta = 0$ is examined using the t-test based on White's robust standard errors. If we reject the null hypothesis, then the two models are not equally accurate, and the model with the lower forecast error variance should be considered to be more accurate.

(c). Analysis under a general loss function

In the discussion of the count method above, the quadratic loss function $f(x) = x^2$ is implied. Diebold and Mariano (1995) suggest testing the same null hypothesis under the general loss function g(x); i.e., H_0: $= E\left[g\left(e_1\right)\right] = E\left[g\left(e_2\right)\right]$. They propose various asymptotic and finite-sample statistics to test this null hypothesis. If we reject the null hypothesis, then the two models generate forecasts of differing quality, and we should select the model whose forecasts have the lowest value under the loss function g(x).

(ii). Evaluating forecasted critical values

The evaluation criteria described in the section above focuses specifically on the forecasted expected losses under two competing models. However, the relative performance of other aspects of the distribution forecasts, such as their forecasted α% critical values, may also be of interest. The binomial test used to evaluate the critical value forecasts from one model's forecasted distributions can be adapted to examine the performance of the two competing forecasts. Specifically, a Bonferroni bounds test with size bounded above by k% can be conducted to test the null hypothesis that the forecasted α% critical values from each model are accurate and provide the expected α% coverage. For this test, we conduct binomial tests individually for each set of forecasts with the smaller size of k/2%. The null hypothesis that the two models are equally accurate is rejected if either set of forecasts rejects the binomial null hypothesis. If the null hypothesis is rejected by just one set of forecasts, then the other set can be said to be more accurate. Note that this multi-model testing is conservative, even asymptotically.

(iii). Evaluating forecasted loss distributions

Similarly, Bonferroni bounds tests based on the hypothesis tests proposed by Crnkovic & Drachman (1996), Diebold *et al.* (1997) and Berkowitz (1999) can be constructed for comparing two sets of forecasted loss distributions. See Diebold, Hahn and Tay (1998) for further discussion of multivariate density forecast evaluation.

D. Limitations to the Proposed Evaluation Methodology

The proposed simulation approach permits the comparison of a model's forecasted credit loss distributions to actually observed outcomes, as in the standard backtests performed for VaR estimates. However, a few important limitations must be kept in mind when using this approach.

First, changes in the value of a credit portfolio are mainly driven by three factors: changes in credit quality; changes in the value of credits of given quality (possibly due to changes in credit spreads); and changes in the portfolio weights of the credits. The first two elements, changes in credit quality and valuation, are captured in this methodology by ΔA_{t+1}, the changes in the value of the credits. Thus, model performance can be evaluated along these dimensions using the resampling approach outlined here, with one notable exception. Many credit instruments, especially potential future exposures, do not have observable market prices and must instead be "marked-to-model" or priced according to a valuation model. This process introduces a source of potential error not captured by this methodology. As to the third element, this methodology generally takes actual and simulated portfolio weights as fixed parameters over the planning horizon; thus, changes in portfolio weights are not directly addressed by this methodology.

Second, credit risk models are essentially panel data models used for datasets where the

number of assets N is much greater than the number of years T. This limited amount of credit default and migration data over time complicates the forecast evaluation of these models because the results are inextricably tied to the prevailing macroeconomic and credit conditions over the T year period. Our approach permits model validation using quantitative measures of model performance, but these measures must obviously be interpreted with care if T is small and does not span a business or credit cycle. As more years of data become available, the resampling of credit portfolios under different economic conditions provides for more extensive evaluation of a credit model's accuracy.

As data for the time dimension becomes available, a credit risk model's performance can be evaluated according to both dimensions using these criteria. For example, resampled portfolios can be randomly constructed for a randomly selected year, and the model's performance could be examined as described. This approach could be interpreted as evaluating a model's conditional performance. Alternatively, resampled portfolios and the forecasted outcomes could be examined across the available time period. As more years of data are available, we can be more confident that we are evaluating a model's unconditional performance. Further study is needed to understand the properties of such evaluations.

Finally, as presented, this resampling approach is wedded to the original set of N credits. Since institutions actively manage their existing credit portfolios and originate new credits, this limitation essentially causes the evaluation of the model's performance to be static. Although static diagnostic tools may be useful under some circumstances, the implications for credit risk models used by firms with dynamic credit exposure are less clear. In such cases, it is reasonable to sample for each year t from the N_t credits held at the end of that year; this generalization is

straightforward.

IV. Conclusions

In general, the evaluation of credit risk models will always be more difficult than market risk models because of their underlying time horizons. The evaluations of the distribution forecasts generated by both of these types of models require a relatively large number of forecasts and observed outcomes. Certainly, the daily horizon underpinning market models guarantees a steady stream of observations over which to evaluate forecasts. However, the yearly horizon commonly used for credit risk models does not. Thus, qualitative methods, such as stress-testing and sensitivity analysis, will always be important in the evaluation of credit risk models.

In this paper, we propose evaluation methods based on statistical resampling that can provide quantitative measures of model accuracy for credit risk models. These methods provide performance evaluation in a cross-sectional environment. The proposed statistical tools are relatively simple; are well known in the forecast evaluation and risk management literatures; and are general enough to be used on any type of credit risk model. Although important caveats must be attached to any inference drawn from this type of evaluation, at least some is now available where previously there was little.

Several aspects of the proposed evaluation methodology require additional research. For example, the impact of specific parameters, such as the number of credits to be included in a simulated portfolio and the nature of the simulated portfolio's weights, must be better understood. However, most of the future research in this area should be on actual comparisons of credit risk models over various credit datasets.

References

Altman, E.I. and Saunders, A., 1997. "Credit Risk Measurement: Developments over the Last Twenty Years," *Journal of Banking and Finance*, 21, 1721-1742.

Basle Committee on Banking Supervision, 1999. "Credit Risk Modelling: Current Practices and Applications," Basle Committee on Banking Supervision, Basle. (http://www.bis.org/press/index.htm)

Berkowitz, J., 1999. "Evaluating the Forecasts of Risk Models," Manuscript, Trading Risk Analysis Group, Federal Reserve Board of Governors.

Carey, M., 1998. "Credit Risk in Private Debt Portfolios," *Journal of Finance*, 53, 1363-1388.

Credit Suisse Financial Products, 1997. *CreditRisk+: A Credit Risk Management Framework.* (http://www.csfp.co.uk/csfpfod/html/csfp_10.htm).

Crnkovic, C. and Drachman, J., 1996. "Quality Control," *Risk*, 9, 139-143.

Crouhy, M. and Mark, R., 1998. "A Comparative Analysis of Current Credit Risk Models," Manuscript, Conference on Credit Risk Modelling and Regulatory Implications.

Diebold, F.X., Gunther, T.A. and Tay, A.S., 1997. "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review*, 39, 863-883.

Diebold, F.X., Hahn, J. and Tay, A.S., 1998. "Real-Time Multivariate Density Forecast Evaluation and Calibration: Monitoring the Risk of High-Frequency Returns on Foreign Exchange," Manuscript, Department of Economic, University of Pennsylvania.

Diebold, F.X. and Lopez, J.A., 1996. "Forecast Evaluation and Combination," in Maddala, G.S. and Rao, C.R., eds., *Handbook of Statistics, Volume 14: Statistical Methods in Finance*, 241-268. Amsterdam: North-Holland.

Diebold, F.X. and Mariano, R., 1995. "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253-264.

Federal Reserve System Task Force on Internal Credit Risk Models, 1998. "Credit Risk Models at Major U.S. Banking Institutions: Current State of the Art and Implications for Assessments of Capital Adequacy." Manuscript, Board of Governors of the Federal Reserve System. (http://www.federalreserve.gov:80/boarddocs/press/General/1998/19980529/study.pdf)

Gordy, M.B., 1998. "A Comparative Anatomy of Credit Risk Models," Manuscript, Conference on Credit Risk Modelling and Regulatory Implications.

Granger, C.W.J. and Huang, L.-L., 1997. "Evaluation of Panel Data Models: Some Suggestions from Time Series," Discussion Paper 97-10, Department of Economics, University of California, San Diego.

International Swaps and Derivatives Association, 1998. *Credit Risk and Regulatory Capital.* (http://www.isda.org/crsk0398.pdf).

The Institute of International Finance Working Group on Capital Adequacy, 1998. "Report of the Working Group on Capital Adequacy – Recommendations for Revising the Regulatory Capital Rules for Credit Risk" The Institute of International Finance, Inc.

J.P. Morgan, 1998. *CreditMetrics - Technical Document.* (http://riskmetrics.com/cm/pubs/CMTD1.pdf)

Koyluoglu, H.U. and Hickman, A., 1998. "A Generalized Framework for Credit Risk Portfolio Models," Manuscript, Oliver Wyman & Company.

Kupiec, P., 1995. "Techniques for Verifying the Accuracy of Risk Measurement Models," *Journal of Derivatives*, 3, 73-84.

Lopez, J.A., 1999a. "Regulatory Evaluation of Value-at-Risk Models," *Journal of Risk*, forthcoming.

Lopez, J.A., 1999b. "Methods for Evaluating Value-at-Risk Estimates," *Federal Reserve Bank of San Francisco Economic Review*, forthcoming.

Nickell, P., Perraudin, W., and Varotto, S., 1998. "Ratings- Versus Equity-Based Credit Risk Modelling: An Empirical Analysis." Manuscript, Conference on Credit Risk Modelling and Regulatory Implications.

Treacy W.F. and Carey, M..,1998. "Credit Risk Rating at Large U.S. Banks," *Federal Reserve Bulletin*, 897-921.