Risk books

The Basel Handbook 2nd edition

A Guide for Financial Practitioners

Edited by Michael Ong

Foreword by Jörg Hashagen

Published in association with

KPMG

9

Advanced Credit Model Performance Testing to Meet Basel Requirements: How Things Have Changed!

Donald R. van Deventer, Li Li, Xiaoming Wang

Kamakura Corporation and University of Hawaii

INTRODUCTION

In Geneva in December 2002, Robert Merton was asked for his advice on the use by financial institutions of the credit risk model he created in 1974 to manage credit risk. After a long pause, Professor Merton replied, "Well, the first thing you have to remember is that the model is 28 years old."[1] In the first edition of this Handbook, we puzzled over the persistence of the usage of Merton model by major financial institutions 30 years after its publication. The Merton technology's obvious influence on the early versions of the then-proposed New Capital Accord (henceforth, Basel II) by the Basel Committee on Banking Supervision is a testimony to the powerful intuitive appeal of the model.

At the time, this was a concern to us because it seemed striking that the 1974 Merton technology, with modest extensions,[2] was still regarded by so many as the state of the art. We asked whether the continuing prominence of the Merton model was a rare example of an intellectual breakthrough that has stood an extraordinary test of time, or whether we should be more worried, as Merton said, that the model is 28 years old. The purpose of this chapter is to provide a framework for answering that question and to provide some concrete answers. Since the first edition of this chapter, there is now a wide consensus on the proper regime for testing credit models and a firmly established result: the Merton technology's attractive intuition has not been as successful as even the most naïve single

variable credit models in predicting default. Nearly all studies on large default databases share this conclusion. The result has been a rapid advance in credit risk technology now that credit model testing has shown how much room there is for improvement in modelling.

Van Deventer and Imai (2003) note that Basel II requires that banks must prove to their regulatory supervisors that the credit models they use perform "consistently and meaningfully".[3] Typically, the only institutions who have the capability to assemble these kinds of databases are extremely large financial institutions and commercial vendors of default probabilities. Prior to the commercialisation of default probabilities by Moody's KMV, studies of default were based on a very small number of defaulting observations. Falkenstein and Boral (2000) cite academic papers by Altman (1968) (33 defaults), Altman (1977) (53 defaults), and Blum (1974) (115 defaults) to illustrate the relatively small sample sizes used to draw inferences about bankruptcy probabilities prior to 1999. By way of contrast, Kamakura Corporation's commercial default database includes 1.4 million total observations, including 1,747 failed company observations. Its research database, which spans a longer period, contains more than 2,500 failed companies.

For major financial institutions that have incurred the expense of a large default database, the results of model testing are highly valuable and represent a significant competitive advantage over other financial institutions who do not have the results of credit-model performance tests. For example, there is a large community of arbitrage investors actively trading against users of the Merton default probabilities when the arbitrage investors perceive the signals sent by the Merton model to be incorrect. More and more financial institutions have taken the time to assemble such databases due to the Basel II requirements, and the results have been surprising to many institutions who had accepted the Merton technology on faith, without any independent testing.

Why would any financial institution adopt a third-party default model "on faith" without performing independent tests? Until recently,[4] commercial vendors of default probabilities offered a single default probability model. This presented a dilemma for potential consumers of commercial default probabilities. A vendor of a single type of credit model has two reasons not to publish quantitative tests of performance. The first reason is that the tests may

prove that the model is inferior and ultimately may adversely affect the vendor's commercial prospects. Perhaps for this reason, most vendors require clients to sign licence agreements that forbid the clients from publicising any results of the vendor's model performance. The second reason is more subtle. Even if quantitative performance tests are good, the fact that the vendor offers only one model means that the vendor's tests will be perceived by many as biased in favour of the model that the vendor offers. Under the pressure of Basel II, however, even single-model vendors are seeing their products subjected to intensive testing by both existing clients and potential new clients who no longer accept promises of performance without proof.

Four former employees of Moody's Investors Service have set the standard for quantitative model test disclosure in a series of papers: Andrew Boral, Eric Falkenstein, Sean Keenan and Jorge Sobehart. The authors respect the important contributions of Boral, Falkenstein, Keenan and Sobehart to the integrity of the default probability generation and testing process.

The need for such tests is reflected in the frequently heard comments of default probability users who display a naïveté with respect to credit models that will ultimately result in their failure to meet the credit model testing requirements of Basel II. We present some samples in Panel 1 that illustrate the need for better understanding of credit model testing.

PANEL: MISUNDERSTANDINGS ABOUT CREDIT-MODEL TESTING

A commonly heard comment on credit-model performance goes like this: "I like model A because it showed a better early warning of the default of companies X, Y and Z."

Many users of default probabilities make two critical mistakes in assessing default probability model performance. They choose a very small sample (in this case three companies) to assess model performance and use a naïve criterion for good performance. Assessing model performance on only three companies or 50 or even 100 in a universe of 8,000–10,000 in the total universe of US corporates needlessly exposes the banker to: (a) an incorrect conclusion just because of the noise in the small sample; and (b) the risk of data mining by the default probability vendor, who (like a magician doing card tricks) can steer

the banker to the three or 50 or 100 examples that show the model in the best light. A test of the whole sample eliminates these risks. Bankers should demand this of both internal models and models purchased from third parties.

The second problem this banker's quote has is the performance criteria. The implications of his comment are twofold.

- ❑ I can ignore all false predictions of default and give them zero weight in my decision.
- ❑ If model A has higher default probabilities than model B on a troubled credit, then model A must be better than model B.

Both of these implications should be grounds for a failing grade by banking supervisors. The first comment, ignoring all false positives, is sometimes justified by saying, "I sold Company A's bonds when its default probabilities hit 20% and saved my bank from a loss of US$1.7 million, and I don't care if other companies that don't default have 20% default probabilities because I would never buy a bond with a 20% default probability anyway." Why, then, did the bank have the bond of company A in its portfolio? And what about the bonds that were sold when default probabilities rose, only to have the bank miss out on gains in the bond's price that occurred after the sale? Without knowledge of the gains avoided, as well as the losses avoided, the banker has shown a striking "selection bias" in favour of the model he is currently using. This selection bias will result in any model being judged good by a true believer. We give some examples below.

The second implication exposes the banker and the vendor to a temptation that can be detected by the tests we discuss below: the vendor can make any model show "better early warning" than any other model simply by raising the default probabilities. If the vendor of model B wants to win this banker's business, all he has to do is multiply all of his default probabilities by 6 or add an arbitrary scale factor to make his default probabilities higher than Model A. The banker making this quote would not be able to detect this moral hazard because he does not use the testing regime mentioned below. There has been strong confirmation from testing by major financial institutions that inflated default probabilities have, in fact, been very common and that the bias has persisted over more than a decade.

Eric Falkenstein and Andrew Boral (2000, p 46) of Moody's Investors Service address this issue directly:

> Some vendors have been known to generate very high default rates, and we would suggest the following test to assess those predictions. First, take a set of historical data and group it into 50 equally populated buckets (using percentile breakpoints of 2%, 4%, ..., 100%). Then consider the mean default prediction on the x-axis with the actual, subsequent bad rate on the y-axis. More often than not, models will have a relation

that is somewhat less than 45% (ie, slope ,1), especially at these very high risk groupings. This implies that the model purports more power than it actually has, and more importantly, it is mis-calibrated and should be adjusted.

We present below a second type of test to detect this kind of bias in credit modelling. If a model has a bias to levels higher than actual default rates, it is inappropriate for Basel II use because it will be inaccurate for pricing, hedging, valuation, and portfolio loss simulation.

Another typical comment illustrates a similar point of view that is inconsistent with Basel II compliance in credit modelling: "That credit model vendor is very popular because they have correctly predicted 10,000 of the last 10,500 small business defaults."

Again, this comment ignores false predictions of default and assigns zero costs to them. If any banker truly had that orientation, the Basel II credit supervision process will root them out with a vengeance because the authors hereby propose a credit model at zero cost that outperforms the commercial model referred to above:

100% accurate prediction of small-business defaults: default probability for all small businesses is 100%

This naïve model correctly predicts 10,500 of the last 10,500 defaults. It is free in the sense that assigning a 100% default probability to everyone requires no expense or third-party vendor, since anyone can say the default probability for everyone is 100%. And, as with the banker cited above, it is consistent with a zero weight on the prediction of false positives. When pressed, most financial institutions admit that false positives are important. As one major financial institution comments, one model "correctly predicted 1,000 of the last three defaults".

Once this is admitted, there is a reasonable basis for testing credit models.

THE TWO COMPONENTS OF CREDIT-MODEL PERFORMANCE

Basel II requires that financial institutions have the capability to test credit model performance and internal ratings to ensure that they consistently and meaningfully measure credit risk. There are two principal measures of in-sample credit risk model performance. The first is a measure of the correctness of the ordinal ranking of the companies by riskiness. For this measure, we use the so-called receiver operating characteristics (ROC) accuracy ratio,

whose calculation is reviewed briefly in the next section. For many years, there was confusion over a similar concept that Sobehart labelled the "cumulative accuracy profile". It is now widely acknowledged that the "CAP" is simply a linear transformation of the standard ROC accuracy ratio and that it adds no incremental information. A May 31, 2006 Google search found 2,110,000 Web pages in response to a joint search on "ROC" and "accuracy". A search on "cumulative accuracy profile" turned up only 150 Web-page citations. For this reason we concentrate on the standard measure, the ROC accuracy ratio, in this chapter.

The second in-sample credit model test is a measure of the consistency of the predicted default probability with the actual default probability, which Falkenstein and Boral (2000) call "calibration". This test is necessary to ensure the accuracy of the model for pricing, hedging, valuation and portfolio simulation. Just as important, it is necessary to detect a tendency for a model to bias default probabilities to the high side as Falkenstein and Boral note, which overstates the predictive power of a model by the naïve criteria of the first quote in the introduction. The consistency of actual and expected defaults over time can also be measured in a test advocated by Donald R. van Deventer and Xiaoming Wang as discussed in van Deventer and Imai (2003).

We discuss each of these in-sample tests in turn in the next two sections and present results for four types of credit model on a common database of 1.4 million monthly observations for North American companies from 1990 to October 2004 maintained by Kamakura Corporation as part of its Kamakura Risk Information Services default-probability database product. Space does not permit a discussion of out-of-sample predictive performance, but the results are similar. Going five years out of sample on a database of 15 years in length leaves analysts in the US looking at an almost linear upward trend in defaults five years forward, ending (as of this writing) at the peak of the last recession. It is not very challenging to predict this straight line, even with naïve models. Once "five years forward" default totals have declined to show more cyclicality, out-of-sample tests will be more revealing than they are in 2006.

As of this writing, Kamakura Corporation reports default probabilities for four types of model.

Reduced form credit models

The "reduced form" approach to default probability modelling allows one to derive default probabilities from bond prices, credit derivatives prices or historical default data. When estimating reduced form default probabilities from history, the state of the art technique is logistic regression. See Hosmer and Lemeshow (2000) for an extensive review of this important technique. The Kamakura models are labelled the "Jarrow–Chava" approach, based on early work in this technology by Robert Jarrow and Sudheer Chava. Kamakura labels its default models with version numbers, since the Basel II regulations require at least an annual update of the models. In this chapter, we report on the accuracy of two versions of the "KDP-jc" (the Kamakura Default Probability using the Jarrow–Chava approach):

- KDP-jc, Version 4.1, which was released commercially in January 2006. This fourth-generation reduced form model uses seven financial ratios, three macroeconomic variables and many inputs from the time series of the public firm's stock price to predict default.
- KDP-jc, Version 3.0, is the third generation of the reduced form technology from Kamakura. KDP-jc Version 3.0 uses a smaller number of explanatory variables than version 4.1 but it has the same balance of financial ratio inputs, macro-factor inputs, and equity-related inputs.

The exact inputs, coefficients and model test results are made available by Kamakura in its Kamakura Risk Information Services Technical Guide, Version 4.1, by Jarrow, Mesler and van Deventer (February 2006).

Structural credit models

- KDP-ms, Kamakura Default Probabilities (Merton-structural) using the "best" Merton structural model approach with proprietary mapping to actual default experience by Kamakura.

Hybrid credit models

- KDP-jm, Kamakura Default Probabilities (Jarrow–Merton) combining the Jarrow–Chava and Merton approaches in a hybrid

model within the logistic-regression framework. The KDP-ms Merton default probability is added as an additional explanatory variable to the Jarrow–Chava variables in KDP-jc to form KDP-jm. The version we report on here is version 4.1.

We present the performance results for each model in the ordinal ranking of companies by riskiness in the next section.

Measuring ordinal ranking of companies by credit risk

The standard statistic for measuring the ordinal ranking of companies by credit riskiness is the ROC accuracy ratio. The ROC curve was originally developed in order to measure the signal-to-noise ratio in radio receivers. The ROC curve has become increasingly popular as a measure of model performance in fields ranging from medicine to finance. It is typically used to measure the performance of a model that is used to predict which of two states will occur (sick or not sick, defaulted or not defaulted and so on). Van Deventer and Imai (2003) go into extensive detail on the meaning and derivation of the ROC accuracy ratio, which is a quantitative measure of model performance. Model testing is also reviewed in detail in van Deventer, Imai and Mesler (2004).

In short, the ROC accuracy ratio is derived in the following way:

- Calculate the theoretical default probability for the entire universe of companies in a historical database that includes both defaulted and non-defaulted companies.
- Form all possible pairs of companies such that the pair includes one defaulted "company" and one non-defaulted "company". To be very precise, one pair would be the December 2001 defaulted observation for Enron and the October 1987 observation for General Motors, which did not default in that month. Another pair would include defaulted Enron, December 2001, and non-defaulted Enron, November 2001, and so on.
- If the default probability technology correctly rates the defaulted company as more risky, we award one point to the pair.
- If the default probability technology results in a tie, we give half a point.
- If the default probability technology is incorrect, we give zero points.

- We then add up all the points for all of the pairs, and divide by the number of pairs.[5]

The results are intuitive and extremely clear on model rankings:

- A perfect model scores 1.00 or 100% accuracy, ranking every single one of the defaulting companies as more risky than every non-defaulting company.
- A worthless model scores 0.50 or 50%, because this is a score that could be achieved by flipping a coin.
- A score in the 90% range is extremely good on most datasets.
- A score in the 80% range is very good on most datasets, and so on.

Van Deventer and Imai provide worked examples to illustrate the application of the ROC accuracy ratio technique. The ROC accuracy ratio can be equivalently summarised in one sentence: "It is the average percentile ranking of the defaulting companies in the universe of non-defaulting observations."

We turn now to a supplement to the standard ROC accuracy ratio that measures the accuracy of default N periods ahead, conditional on the company's surviving the intervening N-1 periods.

THE PREDICTIVE ROC ACCURACY RATIO: MEASURING THE ACCURACY OF FORWARD DEFAULT PROBABILITIES

The standard ROC accuracy ratio has the periodicity of the data used and the length of the default flag used. For instance, calculating the ROC ratio for quarterly data, where the default flag is set to 1 if default occurs in the forthcoming quarter, produces an ROC accuracy ratio that is the cumulative accuracy for three months. If we are using monthly data and a monthly default flag, it is the cumulative accuracy for one month. Similarly, an annual database with an annual default flag produces the cumulative ROC accuracy ratio for a year.

Very frequently, it is important to create the full term structure of default probabilities out to N periods from a historical dataset. In the case of the KRIS default-probability service, Kamakura produces a five-year term structure of default probabilities from its 1.4-million-observation default database, which is based on monthly data. To create the five-year term structure, Kamakura estimates 60 logistic regressions:

(1) Logistic Regression 1, which predicts default in Month 1;
(2) Logistic Regression 2, which gives the probability of default in Month 2 conditional on the company's surviving the first month;
(3) Logistic Regression 3, which gives the probability of default in Month 3 conditional on the company's surviving the first two months ... and so on until the 60th month.

Associated with each of these logistic regressions is a forward or predictive ROC accuracy ratio, which measures the accuracy of your ability to predict default in Month N conditional on survival for the first N − 1 periods.

It is very important to recognise that the 12-month forward ROC accuracy ratio conditional on surviving the first 11 months is *not* the same as a one-year ROC ratio for a regression, which allows default in any month from Month 1 to 12. The former calculation is much, much more difficult. The two accuracy ratios are mathematically linked in a complex way, but it is a serious error in analysis to assume they are equivalent.

In the following sections, we report on both the standard ROC accuracy ratio and the predictive or forward ROC accuracy ratios out to 60 months.

ROC ACCURACY RATIO AND PREDICTIVE ACCURACY RESULTS ON THE KRIS DATABASE

Chapter 6 in van Deventer and Imai (2003) summarises the database compiled by Jarrow and Chava (2002a, 2002b). The database includes monthly observations on all listed companies in the US from 1963 to 1998. Chapter 19 of van Deventer, Imai and Mesler's *Advanced Financial Risk Management* (2004) reports similar results for the commercial database used by the Kamakura Risk Information Services Version 3.0, which were estimated in December 2003. In this section, we report the most recent results for the KRIS 4.1 database, estimation released to clients in January 2006. The gross number of monthly observations in this North American database is more than 1.4 million. There were 1,747 company "failures" during this period, a definition that by necessity is broader than pure default since most public companies in North America are not issuers of public debt. This database is sold commercially by Kamakura Corporation as part of its Kamakura Risk Information Services product line.

Other researchers have used annual data, including the work of Sobehart and colleagues noted above. Annual data have been used in the past because of the researchers' interest in long-term bankruptcy prediction. One of the purposes of this chapter is to show how monthly data can be used for exactly the same purpose, and to show how accuracy changes as the prediction period grows longer. The authors believe that only monthly databases correctly capture the impact of common macroeconomic factors on default.

The basic Jarrow–Chava model uses logistic-regression technology to combine equity market and accounting data for default-probability prediction. Jarrow and Chava provide a framework that shows that these default-probability estimates are consistent with the best practice reduced form credit models of Jarrow (2001) and Duffie and Singleton (1999). They estimate the probability of default based on five basic variables:

- net income to total assets ratio;
- total liabilities to total assets ratio;
- relative size, ratio of total firm equity market value divided by total NYSE and AMEX equity value;
- excess return, the monthly return on the firm minus the monthly value-weighted CRSP NYSE/Amex index return; and
- stock's volatility of previous month's daily prices.

Versions 4.1 and 3.0 of Jarrow–Chava reduced form models from the KRIS default-probability service use financial ratios, macroeconomic factors and statistics from the company's stock price behaviour as predictive variables.

THE PREDICTIVE CAPABILITY OF THE JARROW–CHAVA REDUCED-FORM MODEL DEFAULT PROBABILITIES

Table 1 shows that the ROC accuracy ratios for the reduced-form Jarrow–Chava version 4.1 are very high for the entire universe of public companies. The accuracy ratio is 95.54% for version 4.1 of the Jarrow–Chava model and slightly lower at 95.45% for the Jarrow–Merton hybrid model, where the Kamakura Merton default probability is added as an additional explanatory variable. The Merton model, by contrast, scores an accuracy ratio of only 82.59%.

Table 1 ROC accuracy ratios for KRIS versions 1.0, 2.2, 3.0 and 4.1 on original and most recent default databases

	Release date	Default database used for testing	KDP-jc Jarrow Chava reduced form model	KDP-jm Jarrow Merton hybrid model	KDP-ms Merton structural model
Version 4.1	January 9, 2006	Version 4.1 data base	0.9554	0.9545	0.8259
Version 3.0		Version 4.1 data base	0.9418		
Version 2.2		Version 4.1 data base	0.9448		
Version 1.0		Version 4.1 data base	0.9114		
Version 3.0	December 1, 2003	Version 3.0 data base	0.9362	0.9183	0.8342
Version 2.2	June 26, 2003	Version 2.2 data base	0.9573	0.9625	0.8149
Version 1.0	October 31, 2002	Version 1.0 data base	0.8919	0.9329	

This very low ratio has been independently confirmed by Jorge Sobehart *et al*, formerly of Moody's, and by Bohn *et al* (2005), while Bohn was still director of research at Moody's/KMV.

Recent papers by Campbell, Hilscher and Szilagyi (2005) and Bharath and Shumway (2004) are very consistent with these findings on the superiority of the reduced-form approach in default prediction over the Merton approach. For those conducting independent tests of this performance differential, there is a standard two-step test for Merton model performance, which is nicely summarised in the Bharath and Shumway paper.

Test 1: Is the Merton model perfect?

Bharath and Shumway (2004) ask whether the Merton model alone is a "sufficient statistic" for predicting default. That is, is the Merton model so perfect that no other variables in a logistic regression can add additional explanatory power.

There is a very simple test for this. First, one builds the best logistic regression model one can build. Next, one adds the Merton model as an additional explanatory variable to create a hybrid model. How does one test whether the Merton model is perfect? If the Merton model is perfect, the statistical significance of all of the other variables in the logistic regression should drop to "not significant". Shumway and Bharath, Campbell, Hilscher and Szilagyi (2005), and Kamakura's hybrid model all reach the same conclusion: we strongly reject the hypothesis that the Merton model alone is perfect. In the case of the Kamakura hybrid model, we use 10 inputs to create 31 variables in the logistic regression for the Jarrow–Chava model. When we add the Merton model as an additional variable, 26 of the 31 variables remain statistically significant. In fact, we believe it is impossible for a sophisticated econometrician to conclude that the Merton model cannot be improved on. In fact, almost any additional financial ratio a good credit analyst would look at will improve the predictive power of a Merton default probability.

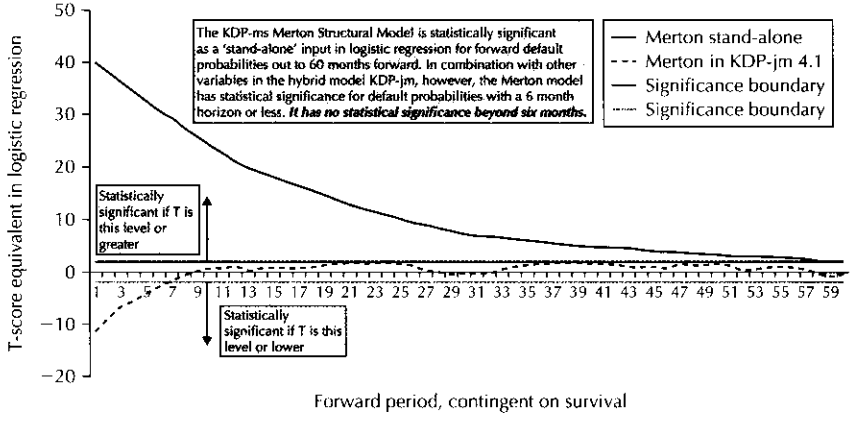
Test 2: Does the Merton model have incremental explanatory power?

Another important test one can employ when evaluating the Merton model is to determine whether or not it adds incremental explanatory power when added as an additional variable in a hybrid model. Shumway and Bharath conclude that the answer to this question is "no" for a large US public-firm dataset. Campbell, Hilscher and Szilagyi conclude that the Merton variable adds modest explanatory power.

The results of this test for the KRIS hybrid model are mixed and can be summarised in Figure 1.

The dark-solid line shows the t-score equivalent of the Merton default probability when that probability is used as the sole explanatory variable in a series of logistic-regression models. The t-score for the Merton default probability is near 40 for a one-month default prediction, so it is highly statistically significant. The second data point is the t-score equivalent for the Merton model when predicting default one month forward, ie, when predicting default in Month 2 conditional on the company's surviving Month 1. This is of course a much harder prediction, so the statistical significance

Figure 1 Statistical significance of the merton default probability as stand-alone and hybrid model input in KRIS version 4.1



declines as one goes out 60 months. Still, at first glance the statistical significance of the Merton default probability remains above 2 standard deviations out to 60 months. At first glance, this seems to confirm that the Merton model is important.

That conclusion is not correct, however, because a very large number of financial ratios are equally capable and in many cases *more* capable of predicting default as a single input to a logistic regression. The more appropriate test, as noted by Shumway and Bharath, and Campbell, Hilscher and Szilagyi, is to look at the statistical significance of the Merton model in the presence of other well-selected financial ratios and other inputs. The dotted line in Figure 1 shows the t-score equivalent for the Merton default probability in the Kamakura hybrid model (which has 10 inputs, which are converted to 31 distinct inputs to the logistic regression). As the graph shows, the sign on the Merton default probability changes from positive to negative because of the high correlation of the Merton default probability with other inputs. This, by itself, is not a concern because even in the presence of correlation the model coefficients remain unbiased. The Merton default probability for a one-month time horizon is statistically significant with a t-score equivalent of about minus 11. From Month 7 onward, however, the Merton model has no statistical significance as a predictor of default out to 60 months.

The results above are for the North American universe. A similar test for Japan reaches the conclusion that the Merton model has no statistical significance at any time horizon. We encourage readers to repeat this testing process on the datasets that are most meaningful to the task at hand.

MAPPING THE MERTON MODEL TO ACTUAL DEFAULTS AND THE IMPACT ON ROC ACCURACY RATIO

In comparing different versions of the Merton credit model, one of the key issues is the mapping of theoretical default probabilities to actual default experience. Different users of the Merton model do this in varying ways, but almost all users of the Merton model have one characteristic in common: The theoretical Merton default probabilities are mapped to actual default experience in a way that changes the absolute level of the default probabilities but not the ordinal ranking of the companies in the universe.

This has a very important implication for the ROC accuracy ratios of the Merton model:

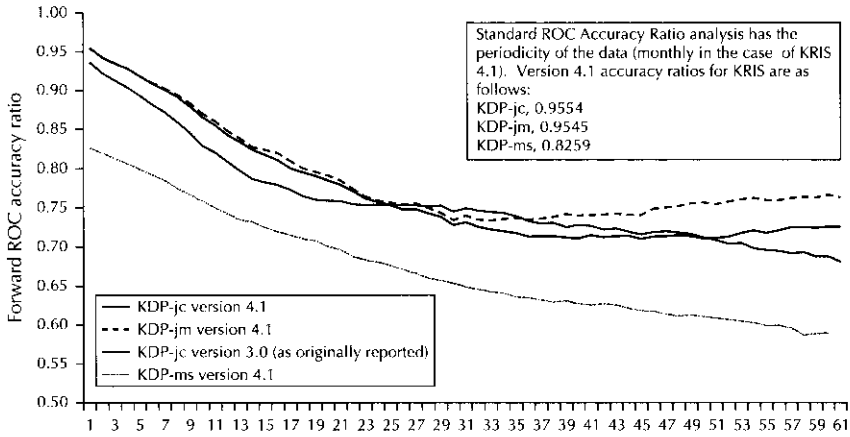
> ***The methodology used for mapping theoretical default probabilities to actual default experience will not change the ROC accuracy ratio for the Merton model if it preserves the ordinal ranking of companies by riskiness.***

Stating it more simply, the ROC accuracy ratio measures only ordinal accuracy, not calibration or consistency between expected and actual defaults. This point is obvious but often overlooked by naïve users, who believe that the mapping methodology improves accuracy. As far as the ROC accuracy ratio goes, this belief is without foundation. The only benefit to a better mapping technology is shown below in measuring the consistency between actual and expected defaults. This is a different issue from measuring the accuracy of the ordinal ranking of companies.

PREDICTIVE ROC ACCURACY RATIOS

Kamakura Risk Information Services provides a term structure of default probabilities out to 60 months, which is actually constructed from 60 logistic regressions, which predict not only the probability of default in the next month but also the probability of default in month N conditional on survival to month N–1. From

Figure 2 KRIS version 4.1 forward ROC accuracy ratios, contingent on survival to month N, for 1 to 61 months forward



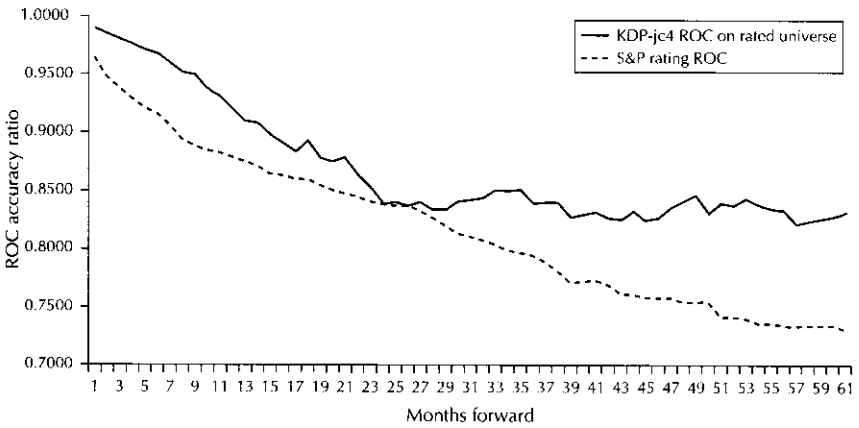
these 60 spot and forward default probabilities, the full term structure of default is produced. Figure 2 shows the spot and forward ROC accuracy ratios for the KRIS Version 4.1 Jarrow–Chava reduced-form model, the Version 4.1 hybrid model, the Version 3.0 Jarrow–Chava model and the Version 4.1 Merton model.

The ROC accuracy ratio on a forward basis for all three of the reduced-form models is dramatically higher than the forward ROC accuracy ratios for the Merton model at all time horizons. This comprehensive testing of model performance is essential. Model testing for a standard one-year period is inadequate for a complete understanding of credit-model performance.

How do reduced form models compare to ratings in predictive performance?

The same kind of test regime can be used to compare quantitative model performance with internal ratings or agency ratios. We demonstrate the methodology in this section using public ratings of one of the two US rating agencies. The first and most important point to note is that the rated universe is much easier to model than the full universe of public companies in North America because the

Figure 3 ROC accuracy ratios for 60 months forward, contingent on survival, for S&P ratings and Kamkura Jarrow–Chava KDP-jc4 version 4.1 explanatory variables weighted for rated universe



companies are much bigger, and there is much less volatility in both stock prices and accounting figures than there is in the non-rated universe. In the test results below, we have re-estimated the weightings in the Jarrow–Chava 4.1 model to those that best fit the rated universe. This has only a very minor impact on the performance differential reported below. Another important thing to note is that the rated universe is much smaller than the full public firm universe. The tests below are based on 230,000 monthly observations, not the 1.4 million observations on the full public universe. There are also only 100 defaulting observations compared with 1,747 defaults in the full public universe.

The spot and forward ROC accuracy ratios for the reduced form model *versus* agency ratings are shown in Figure 3.

The dark-solid line is the ROC accuracy ratio on a spot and forward basis for the Jarrow–Chava Version 4.1 model weighted for the rated universe. The accuracy ratio is 99.00, an extremely high level, on a spot (the standard) basis. The dotted line represents the spot and forward basis for agency ratings. The accuracy ratio for ratings on a spot basis is 96.44. In order to make a fair comparison, one has to be careful about the default analysis in many statistical packages. If one models agency ratings with one dummy variable

for AAA, another for AA+ and so on, standard statistical software normally drops observations where default or non-default is predicted perfectly (ie, if there are no defaults within a month on any AAA-rated companies) and analyses only the remaining (and more difficult data), reporting an ROC accuracy ratio only on the hardest part of the dataset. To avoid this, the proper procedure is to run a logistic regression on one ordinal variable that is 1, if the rating is AAA; 2, if it is AA+; 3, if it is AA and so on. This will prevent any observations from being dropped and will result in an accuracy ratio for the full rated universe. Figure 3 was generated using this basis. Because this ordinal variable preserves the ranking of companies by riskiness as reflected by their ratings, it does not bias the accuracy ratio up or down.

PERFORMANCE OF NAÏVE MODELS AS A PERFORMANCE BENCHMARK FOR THE MERTON MODEL

As noted in the introduction, often a naïve model outperforms a seemingly more elegant model. It is very important that users of credit models test performance of their "favourite" model *versus* naïve models: both better to assess the accuracy of their favourite model and to determine what amount of financial resources should be invested in the favourite model.

As we saw in earlier sections:

- no mapping of the Merton theoretical default probabilities to actual defaults changes the ROC accuracy ratio; and
- testing the accuracy of ordinal ranking of companies by riskiness should be based on a common historical default database and measured using the ROC accuracy ratio.

In model testing for its Version 4.1 default probability release, Kamakura tested its three credit models listed in the previous section against a large number of financial ratios used as a single variable predictor of credit risk. We fitted a logistic regression to each of the financial ratios, one at a time, and measured the ROC accuracy ratios from the resulting logistic regressions. The results are well known to sophisticated users of the Merton and other credit models, but have never been previously tested.

- Four of the financial ratios tested had ROC accuracy ratios superior to the accuracy ratio for the Merton model.
- Both the KDP-jc advanced Jarrow–Chava model and the KDP-jm Jarrow–Merton hybrid model had ROC accuracy ratios far superior to any financial ratio on a standalone basis

This is a striking conclusion, but it is well known in the industry and consistent with Robert Merton's concern about the age of the Merton model. Needless to say, several of the best-performing financial ratios are available free on popular financial Web sites, which call into question the wisdom of a large investment in a competing credit measure that is less successful. Users of popular credit models and the Basel Committee on Banking Supervision clearly need to be aware of performance *versus* naïve credit models such as those we have examined here.

We turn now to another measure of model performance.

CONSISTENCY OF ESTIMATED AND ACTUAL DEFAULTS

Falkenstein and Boral (2000) correctly emphasise the need to do more than measure the correctness of the ordinal ranking of companies by riskiness. One needs to determine whether a model is correctly "calibrated", in the words of Falkenstein and Boral as to whether the model has default probabilities that are biased, high or low. As noted in the example in the introduction, a naïve user of credit models can be convinced a model has superior performance just because it gives higher default probabilities for some subset of a sample. A test of consistency between actual and expected defaults is needed to see whether this difference in default probability levels is consistent with actual default experience or just an ad hoc adjustment or noise.

A simple example is enough to show why this comparison of actual and expected defaults has to be done period by period, not just over the sample as a whole.

Consider the following example:

- Assume we know the actual average probability of default for all listed companies in North America from 1963 to 2004 and that all companies in North America have this probability of default.
- Assume that this default probability is constant over 1963–2004 (an assumption common to many CDO modelling approaches).

- Assume that there is no correlation between the default probabilities of any two companies (another common assumption in CDO modelling).

How consistent would the actual number of defaults and the expected number of defaults has been given these assumptions?

We take the following steps:

(1) Based on the number of companies that are listed in North America at the start of each year, we calculate the confidence intervals on the high and low numbers of default that should occur in that year if our assumption that there is zero correlation is true.
(2) We then compare the actual number of defaults to our confidence interval.
(3) When we do this analysis, we know the following:

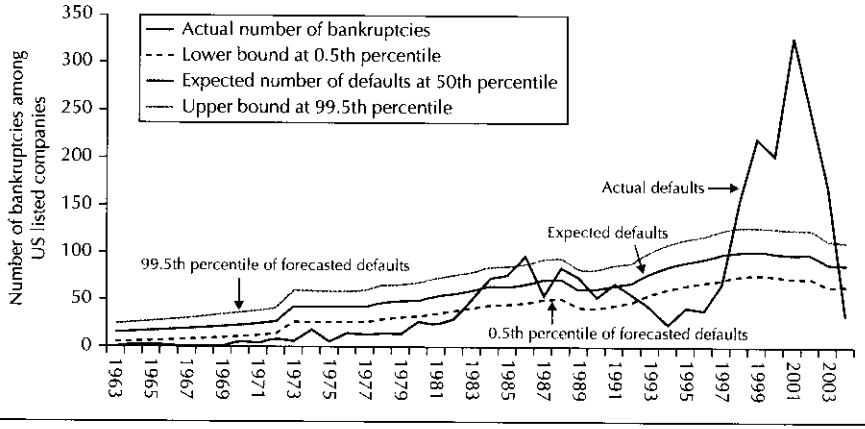
   (a) over the entire period 1963–2004, our expected number of defaults will exactly match the North America total (which is a much better performance than we would get in forecasting CDO defaults);
   (b) we can calculate the 99.5% percentile for the number of defaults;
   (c) we can calculate the 0.5% percentile for the number of defaults; and
   (d) if our credit modelling assumptions are good, we will have a high degree of consistency between actual and expected defaults, falling out of the confidence interval only 1% of the time.

Figure 4 shows the results of the forecasting exercise described above.

As the graph shows, even though we correctly forecast the 1,120 bankruptcies that occurred in North America over the 1963–2004 period, we were dramatically wrong on timing and our assumption that there is no correlation among listed companies in North America seems to be seriously wrong. Over 33 years of the 42-year period, we are at or below the 0.5% percentile level or over the 99.5% percentile level when it comes to actual number of bankruptcies – we are out of the 99% range of probability more than 75% of the time.

Even though on average we predicted exactly the right number of defaults over the 42 years, we were dramatically wrong on

Figure 4 Actual North American coporate failures *versus* 99% confidence interval assuming actual average annual default probability of 1.19% and no correlation in default, 1963–2004 actual defaults were outside 99% confidence interval 33 of 42 years



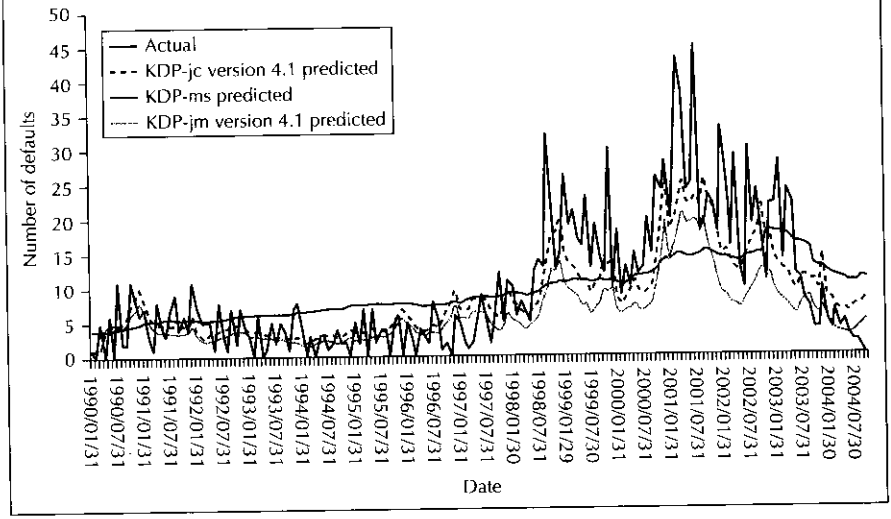
timing. This shows that model calibration over time, not just over a "one-period" sample, is very important. We turn to a methodology for doing that in the next section.

A QUANTITATIVE TEST OF CONSISTENCY BETWEEN EXPECTED AND ACTUAL DEFAULTS

Along with the ROC accuracy ratio, one of the key measures of model performance is the consistency between the expected number of defaults (according to a given credit model) and the actual number of defaults that results. Robert Jarrow has established that, conditional on the values of the Jarrow–Chava input variables, the default probabilities of each company are independent.[6] A similar argument applies for the Merton model conditional on the values of company assets being given at any point in time. Figure 5 uses this fact to compare the expected number of defaults with the actual number of defaults for all three models produced by Kamakura for the North America universe.

The dark-solid line shows the actual number of defaults in the North America universe from 1990 to 2004. The dramatic rise and fall and rise in the number of defaults is strong evidence of correlation among default probabilities due to common dependence

Figure 5 Actual *versus* predicted defaults, North America, 1990–2004 for KRIS version 4.1



on macro factors driving default, as explained in van Deventer and Imai (2003). The dotted line is the expected number of defaults according to the KDP-jc Jarrow–Chava version 4.1 reduced form default probability. It does a good job of capturing the rise and fall and rise in defaults, explaining 66% of the variation of defaults over time. If graphed on the basis of the average default probability in the universe, the results would show the well-known autocorrelation in the observed default rate.

The fairly straight gray line is the expected number of defaults in the KDP-jm Jarrow–Merton hybrid model. This model also captures about 64% of the credit cycle.

The KDP-ms Merton structural model's expected defaults is the grey line, which tends not to capture the peaks and valleys of defaults over the credit cycle. Expected defaults are nearly a straight line over the 1990–2004 period.

Clearly, the Merton model misses the peaks and valleys of the credit cycle. This is counter-intuitive to many, who argue that stock prices will be low when times are bad and default rates are high. The truth is somewhat different, as confirmed by logistic regression that links stock price indices and default rates. Stock

prices are leading indicators of good times and bad times, as many researchers over the last four decades have found. When times are at their worst and defaults at their peak, stock indices will have already risen in anticipation of good times ahead. The US economy in mid-2003 was a perfect example, with defaults at an all-time high and stock price indices at their highest level in 18 months.

The reduced form models capture this timing difference, but the Merton model does not. As a result, it is "out of synch" by about half of a credit cycle, which is why it misses the peaks and valleys of default. We measure this quantitatively in the next section.

RANKING THE MODELS BY THEIR POWER TO PREDICT THE ACTUAL NUMBER OF DEFAULTS

It is helpful to quantify the models' ability to predict the actual number of defaults. Most ability measures come from running a regression that predicts the actual number of defaults as a function of the expected number of defaults derived from the model.

We can run the regression, Actual Defaults = A + B (Expected Defaults), for each model.

The reason for running this model is to quantify the stability of its errors. If the predicted number of defaults is always too low by 10 defaults, the regression parameter A will adjust to pick this up. If the model's bias is a consistent proportion of the total, then the coefficient B will pick it up. If the errors of the model are unpredictable, the adjusted R^2 of the regression will measure that precisely.

A superior model has the following characteristics:

- It explains a higher percentage of the variation in actual defaults and the adjusted R^2 will be higher.
- It has a higher t-score on the expected number of defaults, the explanatory variable in the regression that predicts actual defaults as a function of expected defaults.
- It has a lower standard deviation of the difference between the actual defaults and predicted defaults that come from this regression analysis.
- It has a coefficient of expected defaults closest to one.

These measures of significance are mathematically related.

Table 2 Deventer ability to explain defaults over the business cycle

Actual Defaults Measured as a Function of Predicted Defaults by the Model
Actual Defaults =A+B* Expected Defaults

Adjusted R^2 of the Regression of Actual Defaults as a Function of Predicted Defaults
Ranked from Best to Worst

Version	Model	Adjusted R^2 Data Set
Version 4.1	KDP-jc Jarrow Chava Model	66.29% Version 4.0, 1990–2004, Monthly
Version 4.1	KDP-jm Jarrow Merton Hybrid Model	64.30% Version 4.0, 1990–2004, Monthly
Version 4.1	KDP-ms Merton Structural Model	47.37% Version 4.0, 1990–2004, Monthly
Version 3.0	KDP-jc Jarrow Chava Model	63.80% Version 3 0 1989–2003, Monthly
Version 3.0	KDP-ms Merton Structural Model	61.03% Version 3.0. 1989–2003, Monthly
Version 3.0	KDP-jm Jarrow Merton Hybrid Model	36.25% Version 3.0. 1989–2003, Monthly
Version 2.2	KDP-jc Jarrow Chava Model	43.48% Version 2 2 1989–2002, Annual
Version 2.2	KDP-jm Jarrow Merton Hybrid Model	34.96% Version 2.2, 1989–2002, Annual
Version 2.2	KDP-ms Merton Structural Model	24.75% Version 2.2 1989–2002. Annual

EXPLANATORY POWER

The adjusted R^2 in a linear regression is a popular measure of the explanatory power of a linear regression. As shown in the table below, the KDP-jc Jarrow–Chava model explains more of the variation in actual defaults. The results are based on the regression model we discussed above: Actual Defaults = A + B (Expected Defaults), where A and B are the coefficients determined by the linear regression.

The KDP-jc model explains 66.29% of the variation in actual defaults, compared with 64.30% and 47.37% for the KDP-jm hybrid and KDP-ms Merton structural models. This advantage has persisted over three generations of the model's development, and explanatory power has risen as the business cycle became more pronounced after the recent 2001–03 recession.

IMPLICATIONS OF MODEL TESTING FOR BASEL COMPLIANCE AND PRACTICAL USE OF CREDIT MODELS

As Robert Merton pointed out in the opening quotation, the Merton model is more than 28 years old – in fact, it is now 32 years old at the time of writing. It remains popular in industry and regulatory circles, but for the first time there is a well-established scientific basis for measuring the performance of credit models on the same historical default data in two key dimensions:

- ordinal ranking of firms by credit riskiness; and
- consistency of actual and expected defaults.

Before such tests became available, many well-intentioned bankers were unable to assess correctly credit model performance because of a lack of data and, as noted by Falkenstein and Boral, a tendency for some models to produce higher default probabilities than actual default experience could justify.

This bias is harmful in two respects. Unless the tests outlined in this chapter are performed, it can result in an inaccurate ranking of model performance. More importantly, if this bias is not detected, all calculations using such a model would produce inaccurate pricing, valuation, hedging and portfolio loss simulation. This compounding of effects is contrary to the principles laid out in Basel II, even though the Basel Committee clearly had the legacy of the older Merton model in mind when drafting its proposals.

A scientific approach to testing multiple models reveals that reduced form and hybrid models offer superior performance by both the criteria listed above. This finding has been confirmed by researchers such as Campbell, Hilscher and Szilagyi (2005), Bharath and Shumway (2004), and many large financial institutions in Europe and the US. Furthermore, at least four financial ratios are more accurate in ranking companies by riskiness than the Merton model for every monotonic mapping of theoretical default probabilities to actual default experience.

Practical bankers and skilled regulators need accurate model test results to generate value-added for their shareholders and stakeholders respectively.

* The authors would like to thank their colleagues at Kamakura Corporation for many helpful comments, particularly Mark Mesler, who provided the data warehouse on which the model performance tests within are based. Robert Jarrow's comments and advice have been invaluable over the 11 years of his association with Kamakura Corporation, including his comments on this chapter. Private conversations with Eric Falkenstein and Jorge Sobehart, both formerly of Moody's Investors Service, have also provided many insights. The authors would also like to thank seminar participants at the following regulatory authorities for helpful conversations regarding model testing in a Basel II context: Board of Governors of the Federal Reserve System, Federal Reserve Bank of New York, Office of the Comptroller of the Currency, Australia Prudential Regulatory Authority, Bank of England, FSA of the United Kingdom, FSA of Sweden, FSA of Japan, Bank of Japan, Banca d'Italia, Hong Kong Monetary Authority and the Bank of Israel. The authors alone are responsible for any errors which may remain in what follows.

1 Professor Merton's comments were made at a major risk-management conference before an audience of approximately 400 risk-management experts.

2 An example is the extension of the Merton credit model to incorporate random interest rates by Shimko, Tejima and van Deventer (1993), who combined other Robert Merton insights from the 1970s with his credit model. Other examples are minor modifications of the Merton credit model to allow for early hitting of the default barrier and various methodologies for the estimation of the volatility of company assets in the model.

3 See Section 302, p 55, of Basel Committee on Banking Supervision (2001). Similar language is found in the final version of Basel Committee on Banking Supervision (2004).

4 Kamakura Corporation launched the first multiple-models default service, Kamakura Risk Information Services, in November 2002.

5 This calculation can involve a very large number of pairs. The current commercial database at Kamakura Corporation involves the comparison of 1.4 billion pairs of observations, but on a modern personal computer, as of this writing, processing time for the exact calculation is slightly over one hour. Processing time using a very close approximation is less than one minute.

6 They are independent because the macro factors in the logistic regression are the sole drivers of correlation among default probabilities. See Jarrow, Lando and Yu (2003) for proof of this point.

REFERENCES

Altman, E., 1968, "Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy", *Journal of Finance* 23.

Altman, E., R. Haldeman, and P. Narayanan, 1977, "ZETA Analysis: A New Model to Identify Bankruptcy Risk of Corporations", *Journal of Banking and Finance*, pp 29–55.

Basel Committee on Banking Supervision, 2001a, "Consultative Document: The New Capital Accords", monograph, Bank for International Settlements, Basel.

Basel Committee on Banking Supervision, 2001b, "New Basel Capital Accord", May.

Basel Committee on Banking Supervision, 2004a, "International Convergence of Capital Measurement and Capital Standards: A Revised Framework", monograph, Bank for International Settlements, Basel.

Basel Committee on Banking Supervision, 2004b, "International Convergence of Capital Measurement and Capital Standards: A Revised Framework", June.

Bharath, S. and T. Shumway, 2004, "Forecasting Default with the KMV–Merton Model", University of Michigan and Stanford University Graduate School of Business, December.

Black, F. and M. Scholes, 1973, "The Pricing of Options and Corporate Liabilities", *Journal of Political Economy*, 81, pp 399–418.

Blum, M., 1974, "Failing Company Discriminant Analysis", *Journal of Accounting Research*, Spring.

Bohn, J., N. Arora and I. Korablen, 2005, Power and Level Validation of the EDF™ Credit Measure in North America, Moody's KMV memo.

Campbell, J., J. Hilscher, and J. Szilagyi, 2005, "In Search of Distress Risk", Harvard University working paper.

Duffie, D. and K. Singleton, 1999, "Modelling Term Structures of Defaultable Bonds". *Review of Financial Studies*, **12**(4) pp 197–226.

Falkenstein, E. and A. Boral, 2000, "RiskCalc for Private Companies: Moody's Default Model", Moody's Investors Service memorandum.

Hosmer, D. W. and S. Lemeshow, 2000, *Applied Logistic Regression* (New York: John Wiley & Sons).

Jarrow, R., 2001, "Default Parameter Estimation Using Market Prices", *Financial Analysts Journal*, September/October.

Jarrow, R. and S. Chava, 2002a, "Bankruptcy Prediction with Industry Effects", working paper, Cornell University.

Jarrow, R. and S. Chava, 2002b, "A Comparison of Explicit *versus* Implicit Estimates of Default Probabilities", working paper, Cornell University.

Jarrow, R., D. Lando, and F. Yu, 2003, "Default Risk and Diversification: Theory and Applications", working paper, Cornell University.

Jarrow, R., D. van Deventer, and X. Wang, 2002, "A Robust Test of Merton's Structural Model for Credit Risk", *Journal of Risk*, forthcoming.

Merton, R. C., 1974, "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates", *Journal of Finance*, 29, pp 449–470.

Shimko, D., H. Tejima, and D. van Deventer, 1993, "The Pricing of Risky Debt when Interest Rates are Stochastic", *Journal of Fixed Income*, September, pp 58–66.

Shumway, T., 2001, "Forecasting Bankruptcy More Accurately: A Simple Hazard Model", *Journal of Business*, **74**(1).

Sobehart, J., S. Keenan, and R. Stein, 2000, "Validation Methodologies for Default Risk Models", *Credit*, May, pp 51–56.

Van Deventer, D. and K. Imai, 1996, *Financial Risk Analytics: A Term Structure Model Approach for Banking, Insurance, and Investment Management* (New York: McGraw Hill).

Van Deventer, D. and K. Imai, 2003, *Credit Risk Models and the Basel Accords: The Merton Model and Reduced Form Models* (New York: John Wiley & Sons).

Van Deventer, D. K. Imai, and M. Mesler, 2004, *Advanced Financial Risk Management: Tools and Techniques for Integrated Credit Risk and Interest Rate Risk Management* (New York: John Wiley & Sons).

Van Deventer, D. and X. Wang, 2002, "Basel II and Lessons from Enron: The Consistency of the Merton Credit Model with Observable Credit Spreads and Equity Prices", working paper, Kamakura Corporation.

Van Deventer, D. and X. Wang, 2003, "Measuring Predictive Capability of Credit Models Under the Basel Capital Accords: Conseco and Results from the United States, 1963–1998", working paper, Kamakura Corporation.