HONG KONG INSTITUTE FOR MONETARY RESEARCH

DISCRIMINATORY POWER AND PREDICTIONS OF DEFAULTS OF STRUCTURAL CREDIT RISK MODELS

T. C. Wong, C. H. Hui and C. F. Lo

HKIMR Working Paper No.34/2009

November 2009

Discriminatory Power and Predictions of Defaults of Structural Credit Risk Models

T. C. Wong*

Hong Kong Monetary Authority

and

C. H. Hui*

Hong Kong Monetary Authority

Hong Kong Institute for Monetary Research

and

C. F. Lo**

The Chinese University of Hong Kong

Hong Kong Institute for Monetary Research

November 2009

Abstract

This paper studies the discriminatory power and calibration quality of the structural credit risk models under the "exogenous default boundary" approach including those proposed by Longstaff and Schwartz (1995) and Collin-Dufresne and Goldstein (2001), and "endogenous default boundary" approach in Leland and Toft (1996) based on 2,050 non-financial companies in 46 economies during the period 1998 to 2005. Their discriminatory power in terms of differentiating defaulting and non-defaulting companies is adequate and the differences among them are not material. In addition, the calibration quality of the three models is similar, although limited evidence is found that the Longstaff and Schwartz model marginally outperforms the others in some subsamples. Overall, no significant difference in the capability of measuring credit risk between the "exogenous default boundary" and "endogenous default boundary" approaches is found.

Keywords: Default Probabilities, Credit Risk Models

JEL Classification: C60, G13, G28

---

* Research Department, Hong Kong Monetary Authority, 55/F, Two International Finance Centre, 8, Finance Street, Central, Hong Kong, China. Email: chhui@hkma.gov.hk Phone: (852) 2878 1485 Fax: (852) 2878 2485

** Institute of Theoretical Physics and Department of Physics, The Chinese University of Hong Kong, and Hong Kong Institute for Monetary Research. Email: cflo@phy.cuhk.edu.hk

1. Introduction

This paper compares the discriminatory power and calibration quality of three popular structural credit risk models, including Longstaff and Schwartz (1995), Collin-Dufresne and Goldstein (2001), and Leland and Toft (1996) (referred to as the LS, CG and LT models respectively). The study contributes to the literature in the following three areas. First, since the models cover two streams of the structural credit risk models, the "exogenous default boundary" approach (as represented by the CG and LS models) and the "endogenous default boundary" approach (as represented by the LT model), the comparisons could shed light on the differences in default predication power between these two streams. Secondly, although economic intuitions generally suggest that models with more parameters are easier to capture main economic characteristics of corporate behaviour in the real world and thus higher default prediction power, empirical evidence in this area is relatively scant. This study could fill this gap in the literature, as the three models represent different degrees of complexity in terms of the number of model parameters. Thirdly, since our analysis mainly follows the methodology by Basel (2005) in validating credit risk systems under Basel II, the empirical findings could enrich our understanding of the comparative performance of the structural models in the context of the Basel II model validations, in particular the internal rating-based approach for large corporate portfolios.

Black and Scholes (1973) and Merton (1974) have been the pioneers in the development of the structural models for credit risk of corporates using a contingent-claims framework. They treat default risk equivalent to a European put option on a firm's asset value and the firm's liability is the option strike. To cope with the possibility of early default before bond maturity, Black and Cox (1976) assume an exogenous default-triggering level for the firm's assets whereby default can occur at any time.

Extensions of structural models have long been a core interest of academics and market practitioners. The LS model extends the Black-Cox model to incorporate interest rate risk explicitly into the analysis and also open a new research avenue in that the evolution of the firm's capital structure is no longer tied up with the payoffs of any individual claim on the firm's assets. This feature allows the default boundary to be independent of the state-contingent payoffs of the claim under consideration. Default occurs when the firm asset value falls below an exogenous default boundary. The constant default boundary in the LS model corresponds to the total amount of debts issued by the firm, that is kept constant over time. The LS model therefore predicts that the expected leverage ratio will decline exponentially over time.

Some empirical findings suggest that companies tend to gradually adjust their capital structures toward a target level of leverage.[1] This means that a firm adjusts its outstanding debt in response to changes in its firm value in order to achieve a target level of leverage. These findings call for the stationary-leverage model for pricing corporate bonds, which is the CG model. The model considers a mean-reverting liability

1 See Marsh (1982), Jalilvand and Harris (1984), Auerbach (1985) and Opler and Titman (1995)

that is an exogenous default boundary. This assumption makes the leverage ratio approach towards a constant target liability-to-asset (i.e. leverage) ratio over time. The long-term target ratio is observed to be close to the average leverage ratio of BBB-rated firms. The CG model helps reconcile some predictions of credit spreads with empirical observations. These include credit spreads that are larger for low-leverage firms and less sensitive to changes in firm value, and upward sloping term structures of credit spreads of speculative-grade bonds.

In contrast to an exogenous default boundary, the LT model considers an endogenous-boundary model in which the firm issues debt of arbitrary maturity. The LT model specifies the default boundary as a function of the expected return and volatility of asset value, the risk free rate of interest, leverage, debt maturity, and default costs. To capture the idea of a long-term or "permanent" capital structure, the model presumes that debt is continuously rolled over. This structure assures that total outstanding principal and coupon payments, as well as average debt maturity, remain constant over time, even though each individual bond has a life that shortens with time. This stationary capital structure implies that the optimal default boundary remains constant through time, although its level now depends upon the maturity of debt issued as well as the other parameters of the model.

Despite significant theoretical development of structural models in the past decades, empirical comparisons are rather limited. Indeed, only a few articles implement a structural model to evaluate its ability to predict prices or spreads (see Eom *et al*. (2004) and the reference therein). Eom *et al*. (2004) show that structural models (including the LS, CG and LT models) predict spreads which are too high on average, suggesting that the models in generally could not achieve a high level of accuracy. Few empirical studies have been conducted on the relationship between actual default rates and theoretical default probabilities (PDs) calculated from these models. Leland (2004) finds that PDs generated from the LS and LT models fit the term structures of actual default rates provided by Moody's (1998) for longer time horizons quite well for reasonable parameters with proper calibrations.[2] Hui *et al*. (2005) show that a generalised "exogenous default boundary" is capable of generating term structures of PDs which are consistent with the term structures of actual default rates of credit ratings of BBB and below, in particular at longer time horizons.

In this paper, the discriminatory power and calibration quality of the three models are assessed using a large dataset from the Credit Monitor of Moody's KMV. As the analysis involves comparisons of PDs against the real-world default frequencies of companies in the dataset, all assessments are based on the real-world PDs, which are derived by applying a simple calibration method on the "theoretical" PDs[3] generated from the models so that the comparisons are meaningful. Since the dataset also contains the

[2] The predicted PDs are too low for short maturities. The problem of downward-biased PDs at short maturities is however common to all contingent-claims credit risk models which assume continuous dynamics.

[3] Theoretical PDs, in this study refer to the PDs generated from the structural models based on some reasonable parameter values which will be discussed in subsection 3.1.

companies' Expected Default Frequencies (EDFs), which are the 1-year real-world PDs from the KMV's structural model, the performance of the three models will be benchmarked to the EDFs where appropriate to provide further insight on the models' performance.[4] A brief introduction on the calculations of the EDF by the KMV is presented in the Appendix.

The remainder of the paper is organised as follows. In the following section we illustrate the data used for the study. The empirical results are presented in section 3. The final section summarises and discusses the findings.

2. Data and Model Parameters

The Credit Monitor of Moody's KMV[5] dataset for the analysis consists of 8,486 year-end observations from 2,050 publicly listed non-financial companies[6] in 46 economies covering the period 1998 to 2005.[7] Only those companies with S&P's credit ratings are included in the analysis.[8] Table 1 presents the distribution of S&P's ratings (AAA to CCC+ or below) of the samples. The distribution of the samples across industries and economies are presented in Tables 2 and 3 respectively. Major model inputs, including the asset volatility, the default point (firm liability), and the market asset value of companies[9] are directly extracted from the dataset and apply identically to the three models to derive the PDs of companies. This ensures the PDs from different models are comparable.

Regarding the value of model parameters, we mainly follow the literature. The predefined default-triggering level of the leverage ratio is set to be one for the CG and LS models, which is also adopted in Collin-Dufresne and Goldstein (2001). In the CG model, the mean-reverting parameter for the leverage ratio is set to be 0.1 that is estimated by Fama and French (2002) who investigate the universe of firms. The target leverage ratio is assumed to be 0.315 which is the average leverage ratio of BBB-rated firms

4 The EDF is determined by mapping the distance-to-default (measured by the difference between a firm's asset value and a default threshold in terms of the firm's asset volatility), which is a default risk indicator generated from a structural model, to actual default rates of a large proprietary dataset of companies.

5 According to Moody's KMV, Credit Monitor is a software tool that helps monitor and manage credit risk of corporate obligors. Credit Monitor reports and calculates the PDs for corporate obligors for a term of one to five years. While private firm models are available in Credit Monitor, all data used in this study are extracted from the public firm model, namely the EDF™ Public Firm Model.

6 All utility companies are excluded in this study.

7 Among the 46 economies, 31 of them are developed economies (i.e., high-income economies defined by the World Bank) which share about 97% of the samples. In terms of geographical regions of the economies, 69% of the sample is from North America, 16% is from Asia-Pacific region, and 14% is from Western Europe.

8 This is mainly due to that the definition of default adopted in this study is based on the firms' S&P's ratings.

9 According to Moody's KMV, the default point is the point to which a firm's asset value must fall before the firm defaults. It is approximately equal to the total amount of short-term liabilities, plus half of the long-term liabilities. However, exact definition varies across industries. Asset values refer to the underlying economic assets of firms instead of the book value reported on their balance sheets. For a public firm, its asset value is estimated from its equity market value, equity volatility and liability structure (see the Appendix in this paper for details).

reported by S&P's (2001). This is consistent with the empirical finding by Collin-Dufresne and Goldstein (2001) that the long-term target ratio is close to the average leverage ratio of BBB-rated firms. Following Leland (2004), the corporate tax rate, maturity of the bonds, and fractional default costs in the LT model are set to be 0.15, 10 years and 0.3 respectively. For the debt coupon, the value for each sample is required to be solved numerically that allows the aggregate market values of the bonds equal to their aggregate par values, where the aggregate market values of bonds are given in equation (3) in Leland and Toft (1996).

A set of common model parameters is required to be specified. They are the risk-free interest rate, the asset risk premium, and the asset payout rate. The values of the parameters are based on those used in the "base case" in Leland (2004). The interest rate is set to be constant 5% which falls close to the historical average Treasury rate during the period 1996-2006.[10] The asset risk premium is set to be 4%. This value is consistent with an equity premium of about 6% when the average firm has about 35% leverage, which is close to the average leverage ratio of all S&P 500 companies.[11] The asset payout rate is 6% for all companies, as assumed by Huang and Huang (2002).[12]

3. Empirical Results

3.1 The Derivation of Real-World Default Probabilities

Using the parameters specified in the pervious section, the theoretical 1-year PDs of firms can be computed from the structural models. The theoretical PDs are calibrated to the real-world PDs so that the comparisons of PDs against the real-world realised default frequencies are meaningful and consistent. A simple calibration method is adopted to obtain the real-world PDs. For each sample at time *t*, we define the real-world PD for firm *i* at time *t* as

$$PD_{i,t}^{R} = \min[0.0002 + \alpha_t PD_{i,t}^{Q}, 0.2] \quad (1)$$

where $PD_{i,t}^{Q}$ is the corresponding theoretical PD, and α_t is a scaling factor which is obtained by

10 The original model specifications of the LS and CG models allow stochastic interest rates. A negative correlation between assets and interest rates, as seemed to accord with empirical evidence, reduces default risk, but the effect is very small (see Longstaff and Schwartz (1995)).

11 The value of equity premium is consistent with the regression results in Bhandari (1988).

12 The value of the asset payout rate is roughly the weighted average between the average historical dividend yields which is about 4% reported by Ibbotson Associates (2002) and the average historical coupon rate which is about 9% during the period 1973 to 1998, with the weight is set to 35% which is close to the average leverage ratio of all S&P's 500 companies.

$$\underset{\alpha_t}{Min}\ (\frac{\sum_{i=1}^{N_t}\min(0.0002+\alpha_t PD_{i,t}^{Q},0.2)}{N_t}-\overline{EDF_t})^2 \quad (2)$$

where N_t is the number of firms at *t*, and $\overline{EDF_t}$ is the average KMV's EDF of all firms in *t*. In essence, we find a firm-invariant scaling factor of the theoretical PDs such that the average PDs after the calibration (i.e. $PD_{i,t}^{R}$) equals to the average EDF. To be consistent with the KMV's EDF, a lower bound of 0.0002 and an upper bound of 0.2 are imposed on the real-world PDs. It is worth mentioning that the calibration method only requires the average EDF of the portfolio, but not individual firms' EDFs.[13]

While an individual $PD_{i,t}^{R}$ for each company can be derived, a common risk management practice is to pool companies with similar default risk into one rating class, and all companies in that rating class are assumed to share with a single real-world PD, $\overline{PD}_{i,t}^{R}$, which is computed as the average $PD_{i,t}^{R}$ of companies in that rating class. Following this practice, this study assumes a credit system of ten rating classes and companies at every point in time are evenly distributed in the ten classes according to their $PD_{i,t}^{R}$. All assessments are based on $\overline{PD}_{i,t}^{R}$ rather than $PD_{i,t}^{R}$ unless otherwise stated.

3.2 Discriminatory Power under Conventional Performance Measures

In this subsection, two conventional measures, the accuracy ratio (AR) and the area under the receiver operating characteristic curve (AUROC), are adopted to assess the models' discriminatory power, the ability to distinguish *ex-ante* between realised defaulting and non-defaulting firms. These two statistics are popular in the literature on credit risk model validations (Engelmann *et al*., 2003). It should be noted that, however, this conventional validation technique has its limitation, which has been separately discussed in Hamerle *et al*. (2003), Blochwitz *et al*. (2005) and Lingo and Winkler (2008). In essence, they point out that the AR and AUROC values are dependent on the realised default events from empirical data which are stochastic in nature, implying that the AR and AUROC themselves are also stochastic. Therefore, a high (low) value of realised AR may not be sufficient to conclude that a model has high (low) discriminatory power. In a validation context, they suggest that the expected ARs using the *ex ante* PDs should be derived and compare to the realised ARs. Significant deviations between the expected and

13 In essence, the calibration method adopted in this study is to calibrate the theoretical PDs of firms to the average EDF of the portfolio (i.e. calibrate to the "expected" default probabilities of the portfolio). Assessments based on this calibration method could shed light on the differences in the discriminatory power between the structural models and EDFs given that all structural models are with the same calibration quality as the EDFs portfolio-wise. However, it should be noted that, in practice, it is more common to calibrate the theoretical PDs to the average historical default rates of the portfolio. One feasible way to do so is to construct a rating system based on the firms' S&P's credit ratings and to calibrate the theoretical PDs to the actual default rate in each S&P's rating grade. This alternative calibration method is more practical for those banks cannot access information on the average EDF of their loan portfolios. Thanks to the anonymous referee for pointing out this practical issue and the suggested solution.

realised values of ARs indicate that the PDs generated from the model in question may be miscalibrated. This suggests that the validations of discriminatory power and that of calibration quality are interrelated. Following the literature, comparisons between realised and expected ARs are also adopted to validate the models in this study. These will be discussed in the next subsection.

Despite the limitation of the conventional validation method using realised ARs, comparisons of realised AR and AUROC values between the models are still meaningful in a special case where the statistics are derived from the same portfolio at the same point in time (see Blochwitz *et al*. (2005) and Stein (2007)). All comparisons of the statistics in this subsection are on this basis.

We derive the AR and AUROC from the receiver operating characteristic (ROC) curve. The construction of the ROC curves requires two important statistics, the hit rate and the false alarm rate. The calculation of these two statistics requires a specification of a PD threshold (i.e. *PD**) such that a company is predicted as a defaulter if its PD is higher than *PD** and is predicted as a non-defaulter otherwise.

The hit rate of *PD** is defined as

$$HR(PD^*) = \frac{H(PD^*)}{N_B} \tag{3}$$

where $H(PD^*)$ is the number of actual defaulters having their PD estimates higher than the threshold *PD** (i.e. the number of defaulters predicted correctly) and N_B is the total number of actual defaulters in the sample.

The false alarm rate of *PD** is defined as

$$FAR(PD^*) = \frac{F(PD^*)}{N_G} \tag{4}$$

where $F(PD^*)$ is the number of false alarms that is the number of actual non-defaulters having their PD estimates higher than the threshold *PD** (i.e. the number of actual non-defaulters that were wrongly predicted as defaulters). N_G is the total number of non-defaulters in the aggregate sample.

We set every possible value of the 1-year PD as *PD** and calculate the corresponding $HR(PD^*)$ and $FAR(PD^*)$. ROC curve is obtained by plotting $FAR(PD^*)$ in the x-axis against $HR(PD^*)$ in the y-axis, with $FAR(PD^*)$ being sorted by *PD** in descending order.

A model's performance is the better the steeper the ROC curve is at the left end and the closer the ROC curve's position is to the point (0,1). This means that the model is the better, the larger AUROC is. The discriminatory power of a credit risk model can be evaluated by either the AUROC or AR which is defined as

$$AR = 2\int_0^1 HR(FAR)d(FAR) - 1 \quad (5)$$

where $\int_0^1 HR(FAR)d(FAR)$ is the AUROC. Engelmann *et al*. (2003) prove that there is a one-to-one relation between the AR and the AUROC and either one implies another. Therefore, these two performance statistics contain the same information. The AR is 0 for a random model without discriminatory power and it is 1.0 for a perfect model. Statistical significances of the differences in discriminatory power of the models can be evaluated by the non-parametric method proposed by Delong et at. (1998) using the AUROC.

In our discriminatory analysis, we set the horizon of default prediction as one year. As the dataset covers the period December 1998 to December 2005, it consists of seven non-overlapping one-year windows. The first window starts in December 1998 and ends in December 1999 (i.e. the 1999 window); the last window starts in December 2004 and ends in December 2005 (i.e. the 2005 window). For each one-year window, we define a static portfolio consisting of all non-defaulting firms at the beginning of window. For each firm within the portfolio, the 1-year $\overline{PD}_{i,t}^{R}$ using the most updated information available at the beginning of the window is computed with the calibration method described in subsection 3.1. Default status of each firm is then observed at the end of the 1-year window. A firm is classified as a *defaulter* if at least one of the following events has been triggered within the one-year window: (a) the company receives "SD" or "D" for its long-term issuer credit rating from S&P's; (b) Moody's reports that the company defaults; and (c) the company is delisted because of bankruptcy. The default date is defined as the earliest date of the company triggering any one of these three events. The realised defaults together with the $\overline{PD}_{i,t}^{R}$ facilitate the construction of the ROC curves, and thus the calculation of the AR and AUROC.

Table 4 presents the realised ARs of the LS, CG and LT models for the seven time windows. Those of KMV's EDFs are also provided to serve as a benchmark. The calculation of the discriminatory power statistics for EDFs follow the exact procedure of the structural models as described in the last paragraph of subsection 3.1.[14] This facilitates fair comparisons of the discriminatory power statistics between the

[14] Specifically, we assume a credit system of ten rating classes and companies at every point in time are evenly distributed in the ten classes according to their EDFs. The resulting average EDFs in the ten rating classes are used to calculate the discriminatory power statistics.

structural models and EDFs. The results show that the models perform considerably better than a random model and have adequate discriminatory power of ranking credit risk of the companies, as their respective ARs are significantly larger than zero (i.e. the AR value of a random model). While the three models are generally outperformed by the KMV model in all time windows when comparing their AR estimates[15], the outperformance of the KMV model becomes less obvious when statistical significances are taken into consideration. In fact, using the non-parametric method by Delong *et al*. (1998), which is advocated by Engelmann *et al*. (2003) in validating credit models, it is found that the KMV model only outperforms the CG and LT models in two windows and the LS model in one window at the 5% significance level (Table 5).

Comparing the three structural models themselves, it appears that no model could consistently outperform the others statistically. Reflecting this, the null hypothesis of same AUROC (i.e. same discriminatory power) cannot be rejected at the 5% significance level for almost all pairs of the models in all windows, except in the 2002 window where the CG and LS models outperform the LT model statistically. In fact, stronger empirical evidence of similar discriminatory power of the three models is found when theoretical PDs of the three models are compared (Table 6).

Based on the empirical findings so far, it appears that the discriminatory power is less dependent on the choice of structural models. This is largely due to the fact that while each structural model has its own unique characteristics and parameters affecting the PD estimates of firms, all structural models structurally share with one common feature in that PD estimates are monotonically increasing with two major model inputs, the leverage ratio and the asset volatility. This common feature leads different structural models to produce similar risk ranks for identical set of firms in general and thus comparable discriminatory power. In fact, for the models under consideration, their 1-year PDs (regardless whether the real-world or theoretical PDs being compared) exhibit very high rank-order correlations (Table 7), implying similar ARs of the models.

3.3 The Calibration Quality of the Models

The calibration quality is another measure to assess credit risk models by examining how accurate the models' PDs are in predicting the actual default rates. The aim of this subsection is to assess the calibration quality based on the $\overline{PD}_{i,t}^{R}$ produced by the three models.

As mentioned in subsection 3.2, recent research by Hamerle *et al*. (2003), Blochwitz *et al*. (2005) and Lingo and Winkler (2008) proposed an "expected AR approach" to assess the calibration quality by comparing the expected and realised ARs of a model where the expected AR is defined as the AR value that the *ex post* PDs (i.e. empirical default rates) are identical to the *ex ante* PDs generated from the

15 Except in the 2002 window where the KMV model is outperformed by the LS model.

model (i.e. perfect calibration). To facilitate comparisons between the realised and expected ARs, we derive the expected AR for each model in each time window using equation (11) in Lingo and Winkler (2008). Figure 1 shows the realised ARs and the corresponding 95% asymptotic confidence intervals calculated based on the method by Goodman and Kruskal (1979), as well as their expected ARs. Panels A, C and E (i.e. on the left-hand side) show the result based on $\overline{PD}_{i,t}^{R}$ of the CG, LS and LT models respectively, while the corresponding results based on the model's $PD_{i,t}^{Q}$ (i.e. theoretical PDs) are shown on the panels (B, D and F) on the right-hand side. There are two significant empirical findings revealed from Figure 1. First, no explicit evidence on miscalibration of the models' $\overline{PD}_{i,t}^{R}$ is found, as all expected ARs fall within the 95% confidence intervals of the AR estimates. In contrast, the theoretical PDs of the models (i.e. $PD_{i,t}^{Q}$) are very unlikely to achieve reasonable calibration quality despite some reasonable parameter values being used, as they in general exhibit significant deviations between the expected and realised ARs in some time windows. This is largely consistent with the fact that the theoretical PDs of structural models in general underestimate the real-world PDs in a short-horizon. In addition, while the calibration method we adopt is rather simple, the improvement in the calibration quality of the structural models is substantial and such improvement seems to be not significantly dependent on the choice of the structural models.

We further analyse the comparative performance of the three models using two conventional measures, the Brier Score (BS) and the geometric mean probability (GMP). It is worth mentioning at the outset that unlike the "expected AR approach", these measures are more sensitive to the overall default rate of the portfolio. The analysis below at best could then only evaluate the comparative performance of the models in our dataset.

We first give the definitions of the two measures. The *BS* is developed by Brier (1950) which is defined as

$$BS = \frac{1}{N}\sum_{i=1}^{N}(PD_i - \pi_i)^2 \tag{6}$$

where PD_i is the 1-year PD of the company *i* in a portfolio which contains *N* companies. π_i is a binary variable for the company *i* which is defined as 1 if the company defaults, and 0 otherwise. By definition, the *BS* statistic is the mean square error of default forecasts in which greater discrepancies between realised outcomes and forecasts are penalised using a quadratic function (see Rauhmeier and Schuele (2005)). The higher the calibration quality, the smaller the *BS* is. The *BS* statistic ranges from zero to one and zero indicates perfect calibration.

The *GMP* is defined as

$$GMP = \exp(\frac{1}{N}\sum_{i=1}^{N}[\ln(PD_i)(\pi_i) + \ln(1-PD_i)(1-\pi_i)]) \quad (7)$$

which ranges from 0 to 1, with 1 indicating perfect calibration.

Tables 8 presents the differences in the *BS* for each pair of the models (denoted by $\Delta BS_{i,j;t} = BS_{i,t} - BS_{j,t}$ where $BS_{i,t}$ refers to the *BS* for model *i* in window *t*) and their 95% confidence levels (i.e. figures within parentheses), which are derived from a nonparametric bootstrapping method.[16] From the table, the null hypothesis of no difference in the *BS* between models *i* and *j* in *t* (i.e. $H_0 : \Delta BS_{i,j;t} = BS_{i,t} - BS_{j,t} = 0$; same calibration quality)can be tested. Specifically, the null hypothesis can be rejected at the 5% significance level whenever zero falls outside the 95% confidence interval. Model *i* is said to outperform (underperform) model *j* in *t* in terms of the *BS* if the upper (lower) bound of the confidence interval is negative (positive).

Similar hypothesis testing based on the *GMP* can be conducted using the statistics given by Table 9 (i.e. $H_0 : \Delta GMP_{i,j;t} = GMP_{i,t} - GMP_{j,t} = 0$ where $GMP_{i,t}$ denotes the *GMP* for model *i* in window *t*). The only difference is that model *i* is said to outperform (underperform) model *j* in *t* in terms of the *GMP* if the lower (upper) bound of the confidence interval is positive (negative).

From Table 8, it appears that no model consistently outperforms the others, as the null hypothesis of no difference in the *BS* cannot be rejected at the 5% significance level in nearly all cases. The only exception is that the LS model is found to outperform the CG model in 2002. However, the statistical results based on the *GMP* in Table 9 show limited evidence that the LS model outperforms the CG and LT models in some subsamples. Focusing on the two “exogenous default boundary” models, the LS model outperforms the CG model in three out of seven windows, suggesting that the assumption of a constant target leverage ratio of companies (in the CG model) may not give sufficient close description of the dynamics of the leverage ratios of non-financial firms. Similarly, the LS model is found to outperform the LT model in two windows statistically. This may indicate that the consideration of “endogenous default boundary” (as in the LT model) may not improve the calibration quality. Nevertheless, it is difficult to conclude that the LS model could outperform the two models consistently over time, as there is even larger number of

---

16 A similar method is also adopted by Güttler (2005). We randomly draw *N* companies with replacement from the original sample in time *t*, where *N* is the number of companies in the original sample in *t*. The process is repeated by *B* times such that *B* bootstrap samples of size *N* each are created. $\Delta BS_{i,j;t}$ is calculated for each bootstrap sample *b* (where *b* = 1,…, *B*) and the resultant statistic is denoted by $\Delta BS^b_{i,j;t}$. *B* is set to be 5,000 to give a reliable estimate. The distributional properties of $\Delta BS_{i,j;t}$ are revealed from the vector $\{\Delta BS^1_{i,j;t},...,\Delta BS^b_{i,j;t},...,\Delta BS^B_{i,j;t}\}$. The 95% confidence interval is defined as the values covered by the 2.5 and 97.5 percentiles of the vector. For details about the bootstrap methods, see Elfron and Tibshirani (1993).

cases in the seven windows that the null hypothesis of no difference in *GMP* cannot be rejected (i.e. 4 for the CG model against the LS model, and 5 for the LS model against the LT model).

4. Conclusion

Using a large dataset of listed non-financial companies in 46 economies during the period 1998 to 2005, this paper compares discriminatory power and calibration quality of three well-known structural models. Empirical evidence suggests that the models have adequate discriminatory power and the differences between them are not material. For the assessments of calibration quality using the “expected AR approach”, while the theoretical PDs from the models are found to be miscalibrated even reasonable values of model parameters being used, the real-world PDs based on a simple calibration method could improve the calibration quality significantly. And the improvements are found not significantly dependent on the choice of the structural models.

Overall, we do not find any model can consistently outperform the others over time in terms of discriminatory power and calibration quality. However, evidence based on the *GMP* statistics suggests that the LS model marginally outperforms the LT and CG models in some subsamples. Despite it is rather limited evidence, the outperformance of the LS model relative to the LT model may suggest that structural models incorporating “endogenous default boundary” or/and with more parameters may not improve default prediction power significantly.

Similarly, the outperformance of the LS model relative to the CG model in some subsamples may suggest that the assumption of a constant target leverage ratio cannot describe the dynamics of the leverage ratios of non-financial firms adequately and thus need to be refined. In view of recent empirical findings and theoretical studies that the dynamics of the leverage ratio of non-financial companies are mean reverting with a time-varying target ratio (see Roberts (2002), Childs *et al*. (2005), Hennessy and Whited (2005), Titman and Tsyplakov (2005), Hui *et al*. (2006), Flannery and Rangan (2006), Flannery *et al*. (2008) and Liu (2009)), empirical comparisons between “exogenous default boundary” models with a time-varying target leverage ratio and other structural models would give a clearer picture on the dynamics of financial structure of companies that affect the accuracy of the model on default risk estimations.

References

Auerbach, A. J. (1985), *Real Determinants of Corporate Leverages, Corporate Capital Structures in the United States*, Chicago: University of Chicago Press.

Basel Committee on Banking Supervision (2005), “Studies on the Validation of Internal Rating Systems,” Working Paper No.14.

Bhandari, C. (1998), “Debt/Equity Ratio and Expected Common Stock Returns: Empirical Evidence,” *Journal of Finance*, 43: 507-28.

Black, F. and J. Cox (1976), “Valuing Corporate Securities: Some Effects of Bond Indenture Provisions,” *Journal of Finance*, 35: 1223-34.

Black, F. and M. Scholes (1973), “The Pricing of Options and Corporate Liability,” *Journal of Political Economy*, 81: 637-54.

Blochwitz, S., A. Hamerle, S. Hohl, R. Rauhmeier and D. Rösch (2005), “Myth and Reality of Discriminatory Power for Rating Systems,” *Wilmott Magazine*: 2-6.

Brier, G. (1950), “Verification of Forecast Expressed in Terms of Probability,” *Monthly Weather Review*, 78: 1-3.

Childs, P. D., D. C. Mauer and S. H. Ott (2005), “Interactions of Corporate Financing and Investment Decisions: the Effects of Agency Conflicts,” *Journal of Financial Economics*, 76: 667-90.

Collin-Dufresene, P. and R. S. Goldstein (2001), “Do Credit Spreads Reflect Stationary Leverage Ratio?” *Journal of Finance*, 56: 1929-57.

Delong, E. R., D. M. Delong and D. L. Clarke-Pearson (1988), “Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach,” *Biometrics*, 44: 837-45.

Efron, B. and R. J. Tibshirani (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

Engelmann, B., E. Hayden and D. Tasche (2003), “Testing for Rating Accuracy,” *Risk*, 16: 82-6.

Eom, Y., J. Helwege and J. Hung (2004), “Structural Models of Corporate Bond Pricing an Empirical Analysis,” *Review of Financial Studies*, 17(2): 499-544.

Fama, E. and K. R. French (2002), “Testing Tradeoff and Pecking Order Predications About Dividends and Debt,” *Review of Financial Studies*, 15: 1-33.

Flannery, J. F. and K. P. Rangan (2006), “Partial Adjustment toward Target Capital Structures,” *Journal of Financial Economics*, 79: 469-506.

Flannery, M. J., S. Nikolova and Ozde Oztekin (2008), “Leverage Expectations and Bond Credit Spreads,” available at SSRN: http://ssrn.com/abstract=1308172.

Goodman, L. and W. H. Kruskal (1979), *Measures Association for Cross Classifications*, New York: Springer.

Güttler, A. (2005), “Using a Bootstrap Approach to Rate the Raters,” *Financial Markets and Portfolio Management*, 19: 277-95.

Hamerle, A., R. Rauhmeier and D. Rösch (2003), “Uses and Misuses of Measures for Credit Rating Accuracy,” Universität Regensburg Working Paper.

Hennessy, C. A. and T. M. Whited (2005), “Debt Dynamics,” *Journal of Finance*, 60: 1129-65.

Huang, J. Z. and M. Huang (2002), “How Much of the Corporate-Treasury Yield Spread is due to Credit Risk? Results from a New Calibration Approach,” Stanford University Working paper.

Hui, C. H, T. C. Wong, C. F. Lo and M. X. Huang (2005), “Benchmarking Model of Default Probabilities of Listed Companies,” *Journal of Fixed Income*, 15(2): 76-86.

Hui, C. H., C. F. Lo and M. X. Huang (2006), “Are Corporates' Target Leverage Ratios Time-Dependent?” *International Review of Financial Analysis*, 15: 220-36.

Ibbotson Associates (2002), *Stocks, Bonds, Bills and Inflation, 2002 Yearbook.*

Jalivand, A. and R. S. Harris (1984), “Corporate Behaviour in Adjusting to Capital Structure and Dividend Targets: An Econometric Investigation,” *Journal of Finance*, 39: 611-25.

Kealhofer, S. (1998), “Portfolio Management of Default Risk,” Net Exposure 1 (2).

Kealhofer, S. (2003), "Quantifying Credit Risk I: Default Prediction," *Financial Analysts Journal*, 59(1): 30-44.

Leland, H. E. and K. Toft (1996), "Optimal Capital Structure, Endogenous Bankruptcy, and the Term Structure of Credit Spreads," *Journal of Finance*, 51: 987-1019.

Leland, H. E. (2004), "Predictions of Expected Default Frequencies in Structural Models of Debt," *Journal of Investment Management*, 2(2): 5-21.

Lingo, M. and G. Winkler (2008), "Discriminatory Power – An Obsolete Validation Criterion?" *Journal of Risk Model Validation*, 2(1): 45-71.

Liu, L. X. (2009), "Historical Market-to-Book in a Partial-Adjustment Model of Leverage," *Journal of Corporate Finance*, forthcoming.

Longstaff, F. and E. Schwartz (1995), "A Simple Approach to Valuing Risky Fixed and Floating Rate Debt," *Journal of Finance*, 50: 789-819.

Marsh, P. (1982), "The Choice between Equity and Debt: An Empirical Study," *Journal of Finance*, 37: 121-44.

Merton, R. C. (1974), "On the Pricing of Corporate Debt: the Risk Structure of Interest Rates," *Journal of Finance*, 29: 449-70.

Moody's Investors Services (1998), "Historical Default Rates of Corporate Bond Issuers, 1920-1997,"

Opler T. and S. Titman (1995), "The Debt-Equity Choice," Ohio State University and Boston College Working Paper.

Rauhmeier, R. and R. Scheule (2005), "Rating Properties and their Implications for Basel II Capital," *Risk*: 78-81.

Robert, M. (2002), "The Dynamics of Capital Structure: An Empirical Analysis of a Partially Observable System," Duke University Working Paper.

Standard & Poor's (2001), "Adjusted Key U.S. Industrial Financial Ratios,"

Stein, R. M. (2007), "Benchmarking Default Prediction Models: Pitfalls and Remedies in Model Validation," *Journal of Risk Model Validation*, 1(1): 77-113.

Titman, S. and S. Tsyplakov (2005), "A Dynamic Model of Optimal Capital Structure," University of Texas at Austin Working Paper.

Vasicek, O. A. (1997), "Credit Valuation," Net Exposure 1 (1).

Table 1. Distribution of S&P's Ratings of the Samples

S&P's ratings	Numbers of sample companies	Percentage share (%)
AAA	45	0.5
AA+	34	0.4
AA	119	1.4
AA-	229	2.7
A+	364	4.3
A	655	7.7
A-	608	7.2
BBB+	814	9.6
BBB	1,026	12.1
BBB-	832	9.8
BB+	591	7.0
BB	738	8.7
BB-	910	10.7
B+	787	9.3
B	394	4.6
B-	217	2.6
CCC+ or below	123	1.4
Total	8,486	100.0

Table 2. Distribution of Industries of the Samples

Industry	Numbers of sample companies	Percentage share (%)
Basic Materials	979	11.5
Communications	1,188	14.0
Consumer, Cyclical	1,735	20.4
Consumer, Non-cyclical	1,604	18.9
Energy	754	8.9
Industrial	1,725	20.3
Technology	501	5.9

Table 3. Distribution of Economies of the Samples

Economy	Number of samples	% share	Economy	Number of samples	% share
Argentina	37	0.44	Italy	32	0.38
Australia	231	2.72	Japan	932	10.98
Austria	6	0.07	South Korea	25	0.29
Bahamas	10	0.12	Luxembourg	15	0.18
Belgium	5	0.06	Mexico	84	0.99
Bermuda	35	0.41	Netherlands	102	1.20
Brazil	20	0.24	Netherlands Antilles	2	0.02
Canada	370	4.36	New Zealand	47	0.55
Cayman Islands	10	0.12	Norway	32	0.38
Chile	35	0.41	Philippines	21	0.25
China	6	0.07	Poland	2	0.02
Denmark	7	0.08	Portugal	9	0.11
Dominican republic	5	0.06	Russia	27	0.32
Finland	42	0.49	Singapore	26	0.31
France	188	2.22	South Africa	9	0.11
Germany	153	1.80	Spain	31	0.37
Greece	16	0.19	Sweden	14	0.16
Hong Kong, China	29	0.34	Switzerland	57	0.67
Hungary	2	0.02	Taiwan, province of China	26	0.31
India	1	0.01	Thailand	15	0.18
Indonesia	23	0.27	Turkey	4	0.05
Ireland	16	0.19	UK	463	5.46
Israel	24	0.28	USA	5,240	61.75
			Total	**8,486**	**100%**

Table 4. Realised Accuracy Ratios of CG, LT, and CG Models in the Seven Time Windows

Time Windows	CG model	LS model	LT model	KMV model
1999	0.8299	0.8299	0.8302	0.8378
2000	0.8247	0.8553	0.8451	0.8664
2001	0.7138	0.7256	0.7256	0.7831
2002	0.8069	0.8197	0.7554	0.8142
2003	0.7687	0.8160	0.7890	0.8171
2004	0.7360	0.7504	0.7475	0.7946
2005	0.8685	0.8685	0.8685	0.8693

Table 5. Statistical Analysis of the Differences in the Discriminatory Power between the Models Based on the Real-World PDs.

Model A	CG			LS		LT
Model B	LS	LT	KMV	LT	KMV	KMV
Time Windows	$AUROC_A$- $AUROC_B$					
1999	0.0000	-0.0001	-0.0040	-0.0001	-0.0040	-0.0038
2000	-0.0153 *	-0.0102	-0.0208	0.0051	-0.0056	-0.0107
2001	-0.0059	-0.0059	-0.0347 ***	0.0000	-0.0288 ***	-0.0288 ***
2002	-0.0064	0.0258 ***	-0.0036	0.0321 ***	0.0027	-0.0294 ***
2003	-0.0236 *	-0.0101	-0.0242 *	0.0135	-0.006	-0.0141
2004	-0.0072	-0.0058	-0.0293 **	0.0014	-0.0221 *	-0.0235 *
2005	0.0000	0.0000	-0.0004	0.0000	-0.0004	-0.0004

Notes:
(1) $AUROC_A$- $AUROC_B$ refers to the AUROC of model A minus the AUROC of model B.
(2) *, **, and *** denote statistical significances at the 10%, 5%, and 1% levels, respectively.
(3) Statistic significances of $AUROC_A$- $AUROC_B$ are determined using the method proposed by Delong *et al*. (1988).

Table 6. Statistical Analysis of the Differences in the Discriminatory Power between the Models Based on the Theoretical PDs.

Model A	CG			LS		LT
Model B	LS	LT	KMV	LT	KMV	KMV
Time Windows	$AUROC_A$- $AUROC_B$					
1999	0.0000	0.0119	0.0000	0.0119	0.0000	-0.0119
2000	-0.0106	0.0097	-0.0310 *	0.0204	-0.0204	-0.0408
2001	-0.0154	0.0142	-0.0337 **	0.0296 *	-0.0183	-0.0479 **
2002	-0.0097 *	0.0096	-0.0197 *	0.0193	-0.0100	-0.0293
2003	-0.0135	0.0101	-0.0141	0.0236	-0.0006	-0.0242
2004	-0.0577	-0.0361	-0.0870 *	0.0216	-0.0293	-0.0509
2005	0.0000	0.0000	-0.0475	0.0000	-0.0475	-0.0475

Notes:
(4) $AUROC_A$- $AUROC_B$ refers to the AUROC of model A minus the AUROC of model B.
(5) *, **, and *** denote statistical significances at the 10%, 5%, and 1% levels, respectively.
(6) Statistic significances of $AUROC_A$- $AUROC_B$ are determined using the method proposed by Delong *et al*. (1988).

Table 7. Rank-Order Correlations between the PDs Generated from the Models

Theoretical PDs, $PD_{i,t}^{Q}$

	CG model	LS model	LT model
LS model	0.9962		
LT model	0.9971	0.9994	
KMV model	0.9881	0.9970	0.9917

Real-world PDs, $\overline{PD}_{i,t}^{R}$

	CG model	LS model	LT model
LS model	0.9888		
LT model	0.9821	0.9887	
KMV model	0.9775	0.9884	0.9793

Note: The rank-order correlations are calculated using the full sample of 8,486 non-financial firms.

Table 8. Statistical Comparisons of the Calibration Quality between the Models Based on the Brier Score (BS).

Model *i*	CG		LS
Model *j*	LS	LT	LT
Time windows, *t*	$\Delta BS_{i,j;t} = BS_{i,t} - BS_{j,t}$		
1999	-0.00001 (-0.00022, 0.00022)	0.00004 (-0.00021, 0.00029)	0.00002 (-0.00017, 0.00022)
2000	0.00013 (-0.00013, 0.00041)	0.00000 (-0.00028, 0.00029)	0.00002 (-0.00014, 0.00015)
2001	0.00031 (-0.00006, 0.00105)	0.00028 (-0.00010, 0.00102)	-0.00080 (-0.00205, 0.00054)
2002	0.00053 * (0.00007, 0.00126)	-0.00065 (-0.00184, 0.00046)	-0.00116 (-0.00244, 0.00005)
2003	-0.00017 (-0.00095, 0.00059)	0.00016 (-0.00011, 0.00071)	0.00025 (-0.00064, 0.00127)
2004	0.00016 (-0.00002, 0.00059)	0.00015 (-0.00004, 0.00059)	-0.00006 (-0.00019, 0.00009)
2005	0.00000 (-0.00003, 0.00003)	0.00000 (-0.00003, 0.00003)	0.00013 (-0.00004, 0.00036)

Notes:
(1): Figures in the parentheses are the 95% confidence intervals.
(2): * denotes that the null hypothesis of $\Delta BS_{i,j;t} = BS_{i,t} - BS_{j,t} = 0$ can be rejected at the 5% significance level.

Table 9. Statistical Comparisons of the Calibration Quality between the Models Based on the Geometric Mean Probability (GMP).

Model *i*	CG		LS
Model *j*	LS	LT	LT
Time windows, *t*	$\Delta GMP_{i,j;t} = GMP_{i,t} - GMP_{j,t}$		
1999	-0.00026 (-0.00163, 0.00093)	0.00076 (-0.00102, 0.00380)	0.00101 (-0.00098, 0.00496)
2000	-0.00494 * (-0.01296, -0.00010)	-0.00145 (-0.00735, 0.00223)	0.00349 (-0.00048, 0.01308)
2001	-0.00525 * (-0.01359, -0.00053)	-0.00148 (-0.00864, 0.00333)	0.00377 * (0.00070, 0.00799)
2002	-0.00438 * (-0.01037, -0.00045)	0.00365 (-0.00472, 0.01042)	0.00802 * (0.00296, 0.01345)
2003	-0.00285 (-0.00892, 0.00136)	-0.00206 (-0.00869, 0.00156)	0.00079 (-0.00272, 0.00483)
2004	-0.00300 (-0.01097, 0.00017)	-0.00229 (-0.00936, 0.00154)	0.00072 (-0.00034, 0.00339)
2005	-0.00007 (-0.00050, 0.00038)	0.00023 (-0.00029, 0.00096)	0.00030 (-0.00010, 0.00117)

Notes:
(1): Figures in the parentheses are the 95% confidence intervals.
(2): * denotes that the null hypothesis of $\Delta GMP_{i,j;t} = GMP_{i,t} - GMP_{j,t} = 0$ can be rejected at the 5% significance level.

Figure 1. The Analysis of Realised and Expected ARs of the Structural Models

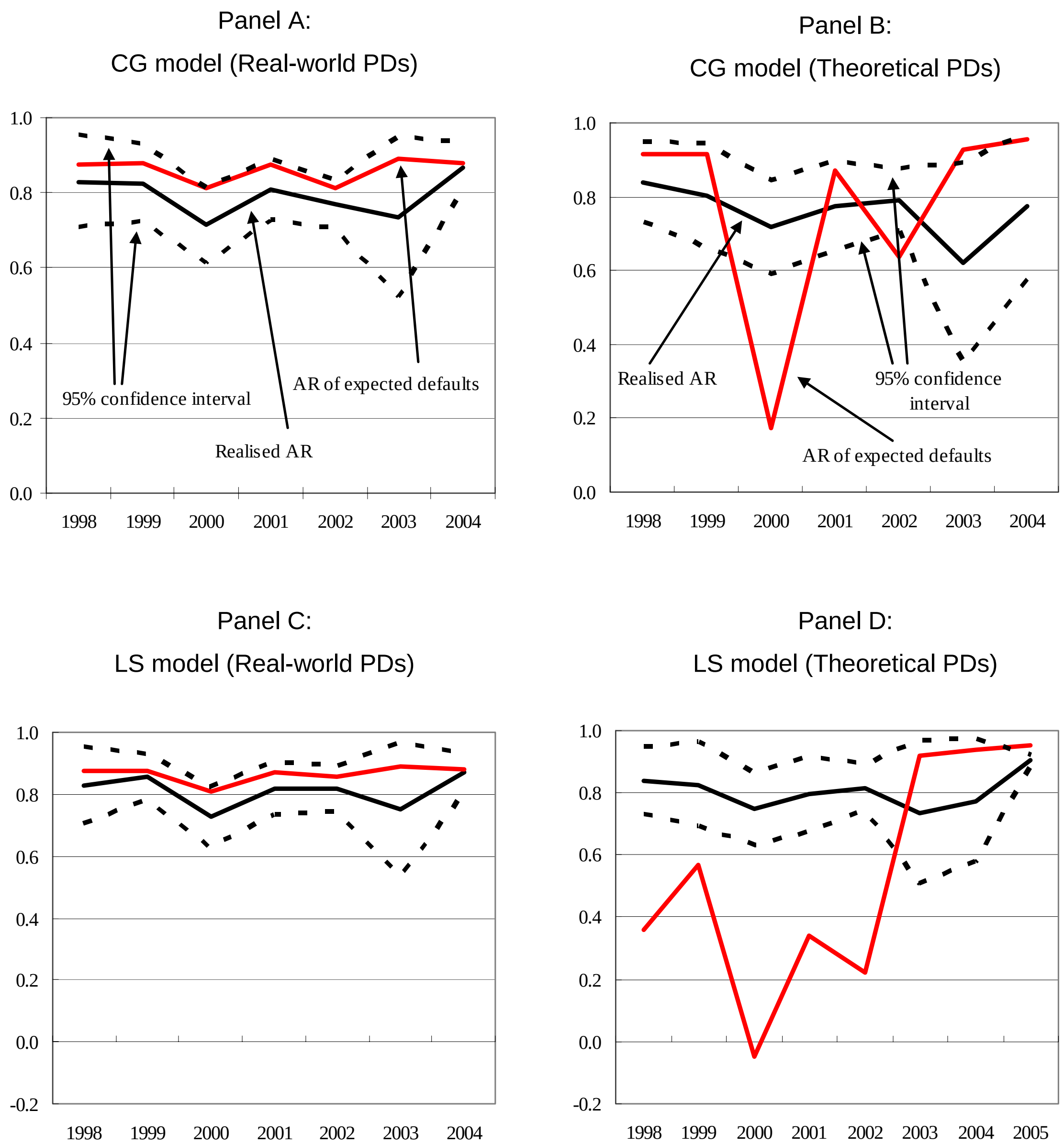
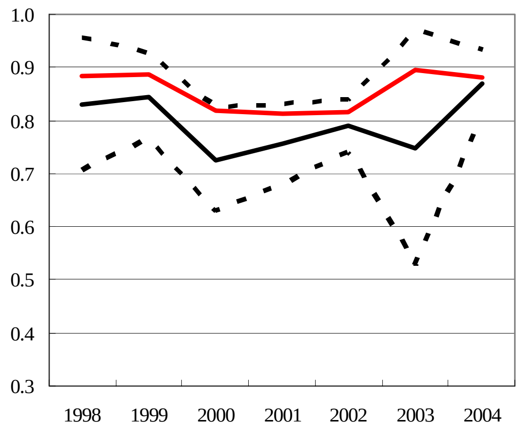
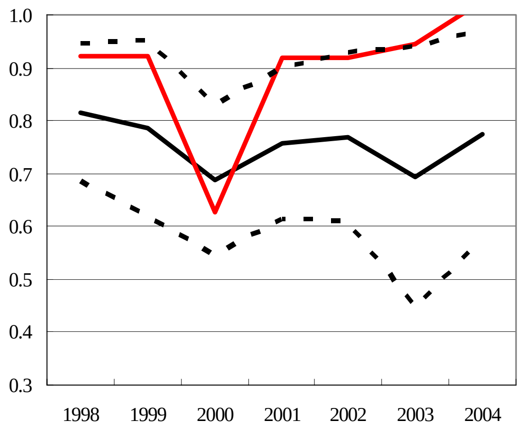


Figure 1. The Analysis of Realised and Expected ARs of the Structural Models (Continued)

Panel E:
LT model (Real-world PDs)



Panel F:
LT model (Theoretical PDs)



Appendix

The KMV model produces a PD for each firm at any given point in time. To calculate the PD, the model consists of the following procedures: estimation of the market value and volatility of the company's asset; calculation of the distance-to-default; and scaling of the distance-to-default to actual PD using a proprietary default dataset.

The KMV model estimates the market value of a company's asset by applying the Merton model.[17] The KMV model makes two assumptions. The first is that the total value of a firm is assumed to follow geometric Brownian motion,

$$dV = \mu V dt + \sigma_V V dz_V \qquad \text{(A1)}$$

where *V* is the market value of the firm's assets, μis the expected continuously compounded return on *V*, σ_V is the volatility of firm's asset value and dz_V is a standard Weiner process. The second assumption of the KMV model is that that the capital structure of the firm is only composed of equity, short-term debt which is considered equivalent to cash, long-term debt and convertible preferred shares. With these simplifying assumptions it is then possible to derive analytical solutions for the value of equity *E*, and its volatility σ_E :

$$E = f(V, \sigma_V, K, c, r) \qquad \text{(A2)}$$

$$\sigma_E = g(V, \sigma_V, K, c, r) \qquad \text{(A3)}$$

where *K* denotes the leverage ratio in the capital structure, *c* is the average coupon paid on the long-term debt and *r* the risk-free interest rate.[18, 19]

---

17 See Vasicek (1997) and Kealhofer (1998, 2003).

18 In the simple Merton's framework, where the firm is financed only by equity and a zero coupon debt, equity is a call option on the assets of the firm with striking price being the face value of the debt and maturity being the redemption date of the bond. The equity value of a firm satisfies

$$E = VN(d_1) - e^{-rT} KN(d_2)$$

where, d_1 is given by

$$d_1 = \frac{\ln(V/K) + (r + \sigma_V^2/2)T}{\sigma_V \sqrt{T}}$$

$d_2 = d_1 - \sigma_V \sqrt{T}$ and *T* is the time-to-maturity of the debt.

19 It can be shown that $\sigma_E = \eta_{E,V}\, \sigma_V$ where $\eta_{E,V}$ denotes the elasticity of equity to asset value, i.e. $\eta_{E,V} = (V/E)(\partial E / \partial V)$.

The KMV model estimates σ_E from market data (i.e. from either historical stock returns data or from option implied volatility data). An iterative technique is used to simultaneously solve equations (A2) and (A3) numerically for values of *V* and σ_V.[20]

Using the values of *V* and σ_V, the KMV model computes an index called “distance-to-default” (DD). DD is the number of standard deviations between the mean of the distribution of the asset value, and a critical threshold, the “default point”, set at the par value of current liabilities including short term debt to be serviced over the time horizon, plus half the long-term debt. The default point *F* is based on KMV’s observations from a large sample firms that default when the asset value reaches a level somewhere between the value of total liabilities and the value of short-term debt. DD can be calculated as:

$$DD = \frac{\ln(V/F) + (\mu - \sigma_V^2/2)T}{\sigma_V \sqrt{T}} \tag{A4}$$

where μ is an estimate of the expected annual return of the firm’s assets, and *T* is a forecasting horizon.

Based on historical information on a large sample of firms, the DD can be mapped to the corresponding implied PD for a given time horizon.[21] This implied PD is the EDF of the firm.[22]

---

[20] Vasicek (1997) notes that the numerical technique is complex due to the complexity of the boundary conditions attached to the various liabilities.

[21] In addition to the empirical approach of the DD-to-EDF mapping, there are some other notable differences between the Merton model and the current version of the KMV model in practice. According to Moody’s KMV’s information, these include: (1) the KMV model adopts a combined approach to estimate asset volatility of a firm in which the estimate is a weighted average of empirical asset volatility and modelled asset volatility. The former is estimated from historical time series of asset returns, while the latter estimates volatility based on firms’ size, income, profitability, industry and geographical region; (2) the model also allows default to occur at or before maturity, while the original Merton model allows default to occur only at maturity; and (3) the model is capable of dealing with a broader class of financial liabilities including preferred stocks and convertible bonds.

[22] The probability below the default point is *N*(-*DD*) which is the EDF in the simple Merton’s framework.