Predicting Recovery Rate at the Time of Corporate Default

Jin-Chuan Duan* and Ruey-Ching Hwang†

(November 13, 2018)

Abstract

A beta regression model with point masses at the two ends is proposed for modeling recovery rate distribution that typically exhibits significant occurrences at both 0 and 100% recovery rates. We implement the model on a sample of 3,827 defaulted debts obtained from Moody's Ultimate Recovery Database. Our approach is to first extend the support of the beta distribution in both directions and then censor the part below 0 (and above 1) to create point masses at the two ends. In addition to confirming the bimodality in the recovery rate distribution, our empirical results clearly show that debt attributes known from the issuance time and industry distress level at the time of default are both significant in predicting recovery rates. Thus, the information available at the time of default can be effectively utilized to differentiate recovery rates of different debt instruments so as to avoid an often misleading application of an unconditional average recovery rate of, say 40%. A performance study based on an analysis of in-sample and out-of-sample datasets of equal size shows that our conditional recovery rate model outperforms four alternative models considered in this study.

Keywords: Beta distribution, gamma distribution, censor, obligor, default, recovery

*Duan is with the National University of Singapore (Risk Management Institute, Business School and Department of Economics). E-mail: bizdjc@nus.edu.sg.

†Hwang is with the Department of Finance, National Dong Hwa University. E-mail: rchwang@mail.ndhu.edu.tw.

1 Introduction

Credit risk has always been an important concern regardless of whether the obligor is a sovereign, corporate or a consumer. Its wide ranging impacts were acutely felt during the 2008-09 global financial crisis and the subsequent Eurozone sovereign debt turmoil. Credit risk arises from the potential of an obligor's default, and the probability of default naturally plays a critical role. No less important is the recovery rate of debt owed by the defaulting obligor. An obligor can have several debts outstanding at the time of default; for example, a bank revolver and a note of some maturity. A particular debt can also be collateralized, meaning that some assets of the obligor have been specifically pledged to service this debt. Debts by the same obligor with varying recovery rates incur different degrees of damage to the lenders. Recovery rates upon an obligor's default are naturally random, because it is impossible to ascertain beforehand the financial resources that an obligor can employ to service its debt obligations at the time of default. For credit instruments' pricing and/or managing credit portfolios, one therefore cannot do without a suitable recovery rate model reflecting individual debt attributes and market conditions at the time of default.

Recovery rate modeling has been made more critical and pressing by the Basel Capital Accords as well as the accounting profession's efforts to properly account for credit risk in financial reporting post the 2008 global financial crisis. In the case of the Basel Accords, banks adopting the advanced internal-rating-based approach are permitted to apply their own data-substantiated recovery rate models instead of using the Basel Committee's simplified recovery assumption. For financial reporting, the International Financial Standards Board has issued IFRS 9 which is scheduled to take effect in January 2018, and the Financial Accounting Standards Board, the US equivalent, has also issued a similar reporting standard known as CECL, which is expected to be in force later.

Individual debt attributes should help differentiate recovery rate distributions of different debts; for example, a collateralized debt is most likely to recover more than a non-collateralized debt issued by the same obligor. Altman and Kishore (1996) and Jankowitsch, *et al* (2014), for example, documented that seniority of debt makes an expected difference and debts of the same seniority issued by different industries face different recovery rates. It is also natural to include the degree of distress of the industry to which the defaulting obligor belongs, because the assets of the defaulting obligor will be subjected to market valuation then. This point was recognized in the literature such as Acharya, *et al* (2007). Our objective is to develop a flexible and yet practical recovery rate model for corporates by linking recovery rates to the debt attributes known from the start along with the industry distress variable at the time of default. We propose a conditional recovery rate model that in essence characterizes the loss given default distribution at the time of default, an essential component of credit risk analysis which becomes even more critical in the era of the Basel banking regulations for capital adequacy.

Our recovery rates and five associated debt attributes are obtained from Moody's Ultimate Recovery Database for 3,827 defaulted debts spanning 1990-2012. For the industry distress variable, we use the industry median one-month probability of default at the time of default provided by the Credit Research Initiative database at National University of Singapore. Schuermann (2004)

observed that recovery distributions tend to be bimodal being either very high or low. Our data sample is in agreement with this observation with over 30% of the recovery rates at 1 and close to 7% at 0, higher than the occurrence frequency at any other recovery rate. Obviously, this bimodality implies that the typical use of a 40% average recovery rate in the literature or some industry practices can be very misleading. More importantly, this bimodality suggests a need to use a conditional recovery rate distribution to reflect individual debt attributes and their varying industry's market conditions at the time of default. Simply put, the bimodality is likely a result of mixing different types of defaulted debts in the sample, which is an unconditional as opposed to conditional recovery rate distribution.

The proposed conditional recovery rate model is a type of beta regression model where the support of the beta distribution is extended in both directions so that the part below 0 or above 1 can be censored to create probability mass for the recovery rate of 0 or 1. The two shape parameters are determined by positive link functions to reflect debt attributes and the market condition at the time of default, and as a result the recovery rate distribution becomes debt specific and default time sensitive. This modeling approach is new and can be estimated by maximizing likelihood. We show that it performs better than the alternative models available in the literature. The alternative models compared to are (1) a censored gamma model of Sigrist and Stahel (2012) and Yashkir and Yashkir (2013), (2) an extended censored gamma model introduced by us in this paper, (3) a two-tailed Tobit model by Maddala (1987) and Bellotti and Crook (2012), and (4) a mixture model of two Bernoulli random variables and a beta random variable by Calabrese (2014).

Similar to Bastos (2010) and Qi and Zhao (2011), we also examined how well the fractional response model of Papke and Wooldridge (1996) performs. Note that the fractional response model is not based on a true distribution function, and thus it does not really serve the purpose of predicting conditional recovery rates. Empirically, it is also found to be less than satisfactory in describing our recovery rate sample, and thus the results are not reported to conserve space. We did not compare our approach with the recovery rate model of Altman and Kalotay (2014) because their normal mixture model requires first inverting recovery rates by a standard normal distribution function so that recovery rates are transformed from a value between 0 and 1 to a value without bound. But this transformation strictly applied is invalid for the recovery rate of 0 or 1, they therefore make a small adjustment to these two extreme recovery rates by adding or subtracting by 10^{-9}. Such an adjustment will create a bunching up of the transformed recovery rates at a large negative and positive values, making it difficult to model by a normal mixture without facing distributional degeneracy.

Our empirical results show clearly that debt attributes play a significant role in predicting recovery rate in a way consistent with economic intuition, regardless of which model is employed. Likewise, the industry distress variable is highly significant in determining recovery rate, and the conclusion is universal for all five models. In short, the conditioning information helps greatly in predicting recovery rate at the time of corporate default. Our empirical findings on relative performance among the five models are based on the graphic feature of the five competing models and a more formal analysis on 20 randomly selected pairs of in-sample and out-of-sample datasets of

equal size. For each in-sample dataset, the corresponding out-of-sample dataset is the remaining defaulted debts in the whole sample. We estimate each of the five models with an in-sample dataset and check performance on both the in-sample and its corresponding out-of-sample datasets. The analysis is then repeated 20 times to tally the performance results. Our proposed recovery rate model outperforms for both in-sample and out-of-sample datasets based on two different performance metrics. Our results also suggest that the beta distribution works better as a fundamental driver for a recovery rate model, but it needs to be modified to reflect that fact that recovery rates tend to load at the endpoints with significant probability mass. A better modification is our proposed approach of first extending the support and then censoring the parts below 0 and above 1 to create point masses.

2 A censored transformed beta regression model for recovery rate

Our censored transformed beta model (CTBM) for conditional recovery rates is a general form of beta regression model that handles a closed interval with two endpoints of positive probabilities. This feature differs from typical beta regression and is critical to the modeling of recovery rates because real data tend to load the endpoints with nontrivial probabilities. Denote by R_i the recovery rate for obligor i at the time of its default. It is defined as

$$R_i = \begin{cases} 0 & \text{if } Z_i \in (-C_l, 0] \\ Z_i & \text{if } Z_i \in (0, 1) \\ 1 & \text{if } Z_i \in [1, 1 + C_u) \end{cases} \tag{1}$$

where $C_l \geq 0$, $C_u \geq 0$, $\frac{Z_i + C_l}{1 + C_l + C_u}$ is assumed to be a beta-distributed random variable with shape parameters, $a_i > 0$ and $b_i > 0$, that are specific to a defaulting obligor. Of course, it is impossible to estimate such obligor-specific shape parameters unless some commonality across defaulting obligors is imposed. Later, a_i and b_i will be modeled by two link functions of covariates defining an obligor's characteristics and the macro-environment at the time of default. However, C_l and C_u are two constants that do not depend on the covariates.

Denote the density function of beta distribution by $\beta(z; a_i, b_i) = \frac{1}{B(a_i, b_i)} z^{a_i - 1}(1 - z)^{b_i - 1}$ where $B(a_i, b_i)$ is the beta function. It is fairly clear that R_i's support is the closed interval [0,1], but Z_i's support is the open interval $(-C_l, 1 + C_u)$ with the following density function:

$$f(z; C_l, C_u, a_i, b_i) = \frac{\beta\left(\frac{z + C_l}{1 + C_l + C_u}; a_i, b_i\right)}{1 + C_l + C_u} \tag{2}$$

Let $F(z; C_l, C_u, a_i, b_i)$ be the corresponding cumulative distribution function. The two endpoints of the recovery rate naturally have the following probabilities:

$$Prob(R_i = 0) = F(0; C_l, C_u, a_i, b_i) - F(-C_l; C_l, C_u, a_i, b_i) = \int_0^{\frac{C_l}{1 + C_l + C_u}} \beta(u; a_i, b_i) du \tag{3}$$

$$Prob(R_i = 1) = F(1 + C_u; C_l, C_u, a_i, b_i) - F(1; C_l, C_u, a_i, b_i) = \int_{\frac{1 + C_l}{1 + C_l + C_u}}^{1} \beta(u; a_i, b_i) du \tag{4}$$

For $R_i \in (0,1)$, the recovery rate density is

$$f(R_i; C_l, C_u, a_i, b_i) = \frac{\beta\left(\frac{R_i+C_l}{1+C_l+C_u}; a_i, b_i\right)}{1+C_l+C_u} \tag{5}$$

which naturally shares the same form with that of Z_i. Although C_l and $C_u \geq 0$ are two constants, the above results show that the probabilities of the two boundary points of the recovery rate, i.e., 0 and 1, still depend on the covariates through a_i and b_i.

Let $X_i(t_i)$, a k-dimensional column vector, denote the set of debt attributes corresponding to obligor i at its default time t_i, including potentially the common variables that define the macroeconomic environment at the time. One should not view $X_i(t_i)$ as representing a panel data set, because it is just the snapshot of an obligor at the time of its default. With $X_i(t_i)$, one can use some common link functions to generate a defaulting obligor's specific shape parameters that drive the beta distribution. There are infinitely many ways to specify these link functions with the only requirement that they must be positive functions. Here we assume

$$a_i(\Theta) = \ln\{1+\exp[\theta_0 + \theta_1 x_{i,1}(t_i) + \cdots + \theta_k x_{i,k}(t_i)]\} \tag{6}$$

$$b_i(\Psi) = \ln\{1+\exp[\psi_0 + \psi_1 x_{i,1}(t_i) + \cdots + \psi_k x_{i,k}(t_i)]\} \tag{7}$$

where $x_{i,j}(t_i)$ is the j-th element of $X_i(t_i)$; and $\Theta = (\theta_0, \theta_1, \cdots, \theta_k)$ and $\Psi = (\psi_0, \psi_1, \cdots, \psi_k)$ denote the parameters in equations (6) and (7), respectively. Although we use the same set of covariates to model the two link functions, they do not need to share the same covariates. Without loss of generality, one can view $X_i(t_i)$ as a union of two different sets of covariates and force some parameters to be zero. The above specification is more natural than just using the exponential-linear function of covariates because it will not be unduly influenced by abnormally large covariate value. Our specification differs from the typical beta regression which applies a link function on the mean of the beta distribution. It would be awkward to do so in our case because the two endpoints with positive probabilities do not lead to a simple analytical formula for the mean recovery rate.

CTBM is more parsimonious than the model of Calabrese (2014) who used two additional link functions to model the probabilities at the two endpoints. It is quite intuitive to think that these two additional link functions may not be needed, because a defaulting obligor's characteristics naturally determine the location and shape of recovery rate. If a debt's attributes are such that its recovery rate will be at an extreme, it will mostly likely to be just one of the extremes (i.e., 0 or 1) as opposed to being some balanced mixture of the two extremes simultaneously. Different debts can face different recovery rate extremes. It is inconceivable, however, that a single defauting debt will be confronted with two extreme recovery rates of split probabilities. Empirical studies reported in the literature often suggest that recovery rates by lumping many obligors together exhibit bimodality, but it is hard to imagine that a single debt can face a bimodal recovery rate. This being said, the shape parameters of beta distribution do change with debt attributes through the link functions, which in turn produce a variety of recovery rate distributions for different debts, including heavy concentration in one of the two extremes or even the unlikely scenario of having high probabilities at both ends.

Figure 1: Plots of the recovery rate density/probability of CTBM based on different combinations of the shape and edge parameters (a,b,C_l,C_u). In each plot, the green horizontal line is the value of 0. The blue triangle, the red plus, and the black star stand for the probabilities at the endpoints whereas the densities over the open interval are expressed by the blue dashed, the red short dashed, and the black solid curves, respectively.

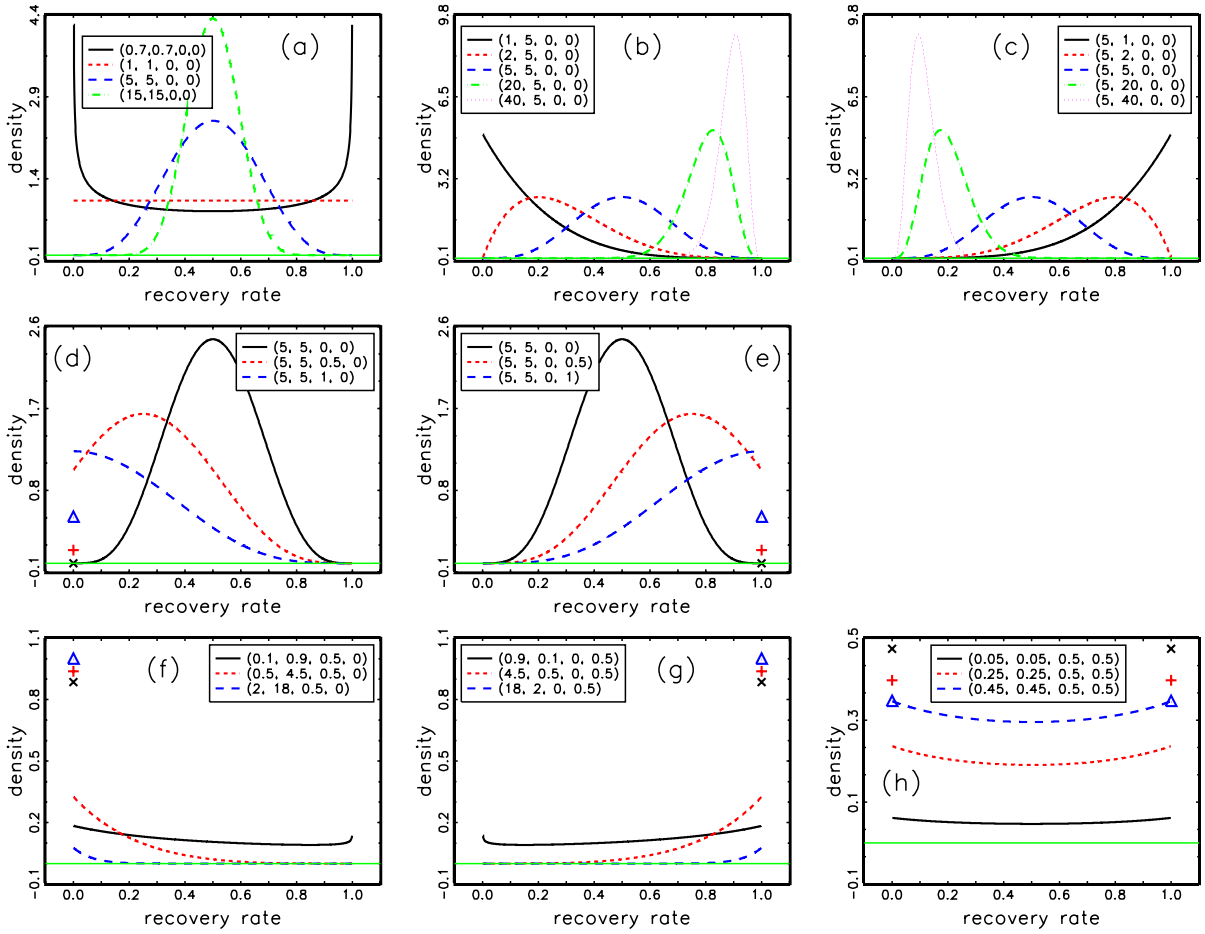


Figure 1 shows several recovery rate density/probability plots corresponding to different sets of parameters, i.e., (a, b, C_l, C_u). When a recovery rate falls strictly between 0 and 1, it is the density as in equation (5), but at each of the two endpoints the probability is based on either equation (3) or (4). As this figure reveals, the recovery rate density can display various interesting shapes, depending on the values of the shape parameters a and b and the edge parameters C_l and C_u. Plots (a)-(c) exhibit the well-known beta distribution properties. Plot (a) shows that in the case of $C_l = C_u = 0$, the density for the case of $a = b < 1$ is a symmetric U shape with the density growing without bound towards the two endpoints. The recovery rate density can become a uniform distribution when $a = b = 1$, or bell-shaped when both a and b are larger than one. Plot (b) shows the effect of increasing a, and as expected the peak of the recovery rate density moves toward the right edge. Naturally, the peak will move toward the left edge with increasing b as shown in Plot (c).

Plots (d)-(h) illustrate the effect of the edge parameters C_l and C_u in determining the endpoint probabilities. It should be noted that the density/probability function is expected to be discontinuous at the endpoints when the edge parameters are strictly positive. The endpoint probability can discretely jump up or down, depending on the range of integration defined by C_l (or C_u) and on whether the recovery density is increasing or decreasing as it approaches an endpoint. In short, three plots reveal that CTBM model is flexible enough to capture various recovery rate patterns.

Parameters, $\{C_l, C_u, \Theta, \Psi\}$, which totals $2k + 4$, can be estimated by maximizing the log-likelihood function[1], which will be a mixture of density for the recovery rate in the open interval, $(0, 1)$, and the probabilities for the recovery rates at 0 and 1, and it is in the following form:

$$\begin{aligned} & L(C_l, C_u, \Theta, \Psi; (R_i, X_i(t_i)), i = 1, 2, \cdots, n) \\ = & \sum_{i=1}^{n} 1_{\{R_i=0\}} \ln\left[\bar{F}_l(C_l, C_u, a_i(\Theta), b_i(\Psi))\right] + \sum_{i=1}^{n} 1_{\{R_i=1\}} \ln\left[\bar{F}_u(C_l, C_u, a_i(\Theta), b_i(\Psi))\right] \\ & + \sum_{i=1}^{n} 1_{\{0<R_i<1\}} \ln\left[\frac{\beta\left(\frac{R_i+C_l}{1+C_l+C_u}; a_i(\Theta), b_i(\Psi)\right)}{1 + C_l + C_u}\right] \end{aligned} \tag{8}$$

where

$$\begin{aligned} \bar{F}_l(C_l, C_u, a_i(\Theta), b_i(\Psi)) &= F(0; C_l, C_u, a_i(\Theta), b_i(\Psi)) - F(-C_l; C_l, C_u, a_i(\Theta), b_i(\Psi)) \\ \bar{F}_u(C_l, C_u, a_i(\Theta), b_i(\Psi)) &= F(1 + C_u; C_l, C_u, a_i(\Theta), b_i(\Psi)) - F(1; C_l, C_u, a_i(\Theta), b_i(\Psi)). \end{aligned}$$

3 Data on recovery rates and debt attributes

A debt instrument's default means that its issuing obligor has defaulted, which in turn implies that all other debt instruments issued by the same obligor are also in default but face varying recovery

[1]The maximum likelihood estimation program is coded in GAUSS, and the optimization is performed with the GAUSS built-in constrained optimization procedure where convergence is defined to be obtaining all elements of the gradient vector less than 10^{-5}. The initial parameter values are all taken as zeros.

rates (partial or full) specific to instruments. The same obligor could in principle defaults multiple times over some time span, and thus one could expect to see several defaulted debts by the same obligor carrying different recovery rates reflective of seniority, collateral and time of default.

Table 1: Definition of the six debt attributes for predicting the recovery rate.

Variable	Definition
Industry Distress (ID)	Industry median 1-month probability of default in basis points at the month end prior to default
Debt Cushion (DC)	Debt below/Issuer total debt; Debt below: The total of all defaulted debt that is contractually subordinate to the current instrument.
Instrument Ranking (IR)	Instruments in each default event are ranked by Moody's in relation to each other based on the structure prior to default, taking into consideration collateral and instrument type.
$IR_{Baseline}$	IR is 1
IR_1	IR is 2
IR_2	IR is 3
IR_3	IR is greater than or equal to 4
Collateral Type (CT)	
$CT_{Baseline}$	Debt has no collateral
CT_1	Debt has collateral
Instrument Type (IT)	
IT_1	Revolvers
$IT_{Baseline}$	Term loans
IT_2	Senior secured bonds
IT_3	Senior subordinated bonds
IT_4	Senior unsecured bonds
IT_5	Junior or subordinated bonds
Utility (UT)	
$UT_{Baseline}$	Issuer is not classified as a utility company
UT_1	Issuer is classified as a utility company

Our data sample comes from two sources: Moody's Ultimate Recovery Database (URD) and the Credit Research Initiative (CRI) database at the National University of Singapore. Moody's URD to which we have access covers defaulted debts by US firms in the period from 1987 to 2012, whereas the CRI database, which is freely accessible, starts from December of 1990 to now. Our final matched sample thus spans the period from December 1990 to the end of 2012, and contains 3827 defaulted debts with recovery rates. The recovery rates used in this paper are discounted recovery rates provided in Moody's URD, which are the nominal recovery rates being discounted at each instrument's prepetition interest rate. In addition to recovery rates, our sample contains five recovery-rate predictors from Moody's URD, which are Debt Cushion (DC), Instrument Ranking (IR), Collateral Type (CT), Instrument Type (IT), and Utility (UT). Beyond these five debt instrument specific variables, we employ Industry Distress (ID) as an additional predictor to reflect

the credit market condition of the industry at the time of a default. ID therefore is an time-varying industry-wide variable that is not unique to a particular defaulting debt. The definitions of the six predictors are given in Table 1. Some descriptive statistics for the recovery rate and the six predictors are provided in Table 2. Evidently, the recovery rate indeed has two point masses occurring at 0 with 6.69% and 1 with 30.36%, respectively. As Table 2 shows, each of the six predictors also covers a range of values with a reasonable distribution.

Further discussion on the five predictors in Moody's URD can be found in Altman and Kalotay (2014). ID requires further elaboration, however. It is represented by the industry median 1-month probability of default (PD) at the day of default as a means to characterize the immediate outlook on the credit market for the defaulting obligor's industry. The industry classification is according to Bloomberg which groups firms into ten industry segments. The industry median PDs are taken from the CRI database, which are computed with the PDs of individual firms in the same industry of an country/economy for each business day. Individual firm PDs are generated by the forward-intensity model of Duan, *et al* (2012), and as of the time of this writing, has been implemented on all exchange-listed firms in 109 economies around the world. The technical details on the CRI implementation can be found in the RMI-CRI Technical Report (2013). The historical series of CRI PDs are re-estimated every month to reflect data additions and revisions, and the data date of the PDs used in the paper is December 31, 2013. It is worth noting that our measure of industry distress differs from that of Acharya, *et al* (2007) by taking advantage of this newly available and more direct measure of financial distress level of an industry.

4 Estimation results and comparison with alternatives

4.1 Empirical performance of CTBM

The maximum likelihood estimation results for CTBM are presented in Table 3. It is clear from this table that industry distress (ID) is an important predictor of recovery rate. Its coefficients in shape parameter functions, $a_i(\Theta)$ and $b_i(\Psi)$, are significantly negative and positive at the 1% level, respectively. Since a lower $a_i(\Theta)$ (or a higher $b_i(\Psi)$) will lower recovery rate as shown in Figure (1), a defaulted debt in a more distressed industry (i.e., a higher industry median PD) tends to lower recovery rate.

For each of the three predictors DC, CT, and UT, its significant effect (at the 1% significance) on recovery rate goes through just one of the two shape parameters. Combining the significance results with the signs of the parameter estimates yields a conclusion that when a defaulted debt has more debt cushion, is collateralized, or is a utility firm, the recovery rate tends to be higher. These conclusions are intuitive and here are the reasons. First, DC is a capacity metric that captures beyond the rank of a debt instrument in capital structure to reflect the degree of subordination by other debts as a proportion of total claims (see Keisman and Van de Castle, 1999). Second, if a defaulted debt has collateral, i.e., CT=1, then it tends to have a higher recovery rate than those without collateral, i.e., CT=0. Third, the result on UT simply reflects that fact that the utility industry has the highest recovery rate among all industries as previously documented by authors

such as Altman and Kishore (1996), Acharya, *et al* (2007) and Qi and Zhao (2011). It is hence not a surprise to find a higher recovery rate when UT=1 as opposed to UT=0.

The parameter estimates corresponding to the indicator variables IR_1 and IR_2 for both shape parameter functions are significantly negative at the 1% level, indicating that the recovery rate of a defaulted debt in these two instrument ranking categories determined by Moody's differ from those in the category of $IR_{Baseline}$, the highest rank. A negative parameter value in the shape parameter function, $a_i(\Theta)$, suggests a lower recovery rate, but it conflicts with the effect of a negative parameter value in the other shape parameter function which implies a higher recovery rate. Therefore, an IR_1 or IR_2 debt may not have a lower recovery rate as compared to the baseline debt that receives the highest rank from Moody's. In the case of IR_3, the parameter in $a_i(\Theta)$ is significantly negative at the 1% level, but the other parameter is insignificant, suggesting that debts in the IR_3 category, the lowest rank, would face lower recovery rates.

For the instrument type, revolvers (IT_1) has a significant positive parameter at the 5% level only for the shape parameter function, $a_i(\Theta)$. As compared to term loans, i.e., the baseline case, revolvers are expected to face a higher recovery rate. In the case of bonds that senior secured (IT_2) and unsecured (IT_4), the parameter estimates are all positively significant, but their effects on recovery rate, as compared to that of term loans, are not clear because the positive value in $a_i(\Theta)$ suggesting a higher recovery rate but that conflicts with the effect of a positive parameter in $b_i(\Psi)$. For junior or subordinated bonds (IT_5), the estimated coefficients are insignificant, but the signs of the two coefficients would have indicated a lower recovery rate if they were statistically significant.

Due to the point mass at the two endpoints, i.e., 0 and 1, the expected recovery rate does not lead to a simple expression for the standard beta distribution, i.e., $E(R_i \mid X_i) \neq \frac{a_i(\Theta)}{a_i(\Theta)+b_i(\Psi)}$. Although the expected value expression does not give rise to a simple closed-form solution, its integral expression, i.e., $\int_{\frac{1+C_l}{1+C_l+C_u}}^{1} \beta(u; a_i(\Theta), b_i(\Psi))du + \int_0^1 \frac{u}{1+C_l+C_u}\beta\left(\frac{u+C_l}{1+C_l+C_u}; a_i(\Theta), b_i(\Psi)\right)du$, can be useful to our understanding on the impact of a predictor. Intuitively, if an increase (decrease) in a predictor's value is to increase (decrease) $a_i(\Psi)$ without a simultaneous increase (decrease) in $b_i(\Psi)$, then the expected recovery rate will rise. When the coefficients in the two link functions for the same predictor are opposite in sign, this is clearly the case. In general, the expected value will need to be evaluated via a numerical integration to obtain the impact of any particular predictor. Such numerical assessments can be conducted for different values of X_i using the maximum likelihood parameter estimate. The integral formula reflects the point mass at the two endpoints. It is evident that the point mass at the 100% recovery rate will, for example, increase if $a_i(\Psi)$ is increased without an offsetting increase (decrease) in $b_i(\Psi)$.

Finally, the two edge parameters – C_l and C_u – are significant at the 1% level, reflecting the need to accommodate high incidents of recovery rate at 0 and 1. In fact, C_u is much larger than C_l to reflect a higher likelihood of having complete recovery than total write-off, which according to Table 1 reflective of our sample are 30.36% and 6.69%, respectively.

CTBM performs quite well in capturing the unconditional recovery rate distribution as evident in Plot (e) of Figure (2) where the empirical frequency distribution is compared with the estimated model frequency distribution. The discussion on their construction is left in Section 4.3. Since CTBM deals with conditional recovery rate, restricting to a subsample of certain debt characteristics can be a way of assessing its performance. We pick two subsamples with the debt characteristics of $(IR_1, CT_{Baseline}, IT_4, UT_{Baseline})$ and $(IR_{Baseline}, CT_1, IT_1, UT_{Baseline})$ and apply CTBM estimated to the full data sample. CTBM's performance conditional on these debet characteristics are exhibited in Figures (3e) and (4e), which show reasonably good general performance. It is worth noting that the recovery rate distribution as shown in these plots is highly data specific. A flexible model that can factor in individual debt characteristics will be essential to the task of successfully modeling credit portfolio losses.

4.2 Four alternative recovery rate models to CTBM

We now study how well CTBM performs in comparison with four alternatives: (1) a censored gamma model (CGM) of Sigrist and Stahel (2012) and Yashkir and Yashkir (2013), (2) an extended CGM by introduced in this paper and denoted by CGM*, (3) a two-tailed Tobit model (TTM) by Maddala (1987) and Bellotti and Crook (2012), and (4) a mixture model of two Bernoulli random variables and a beta random variable (MBB) by Calabrese (2014).

With a k-dimensional predictor $\mathbf{x}$, CGM uses the gamma random variable G (with shape parameter α and scale parameter link function $\ln[1 + \exp(\beta_0 + \boldsymbol{\beta}\mathbf{x})]$) and a positive constant ξ to model recovery rates with two endpoints in the following way:

$$R = 0 \times I_{\{G \leq \xi\}} + (G - \xi) \times I_{\{G \in (\xi, 1+\xi)\}} + 1 \times I_{\{G \geq 1+\xi\}} \tag{9}$$

The model has $k+3$ parameters that need to be estimated. The link function for the scale parameter can of course be specified in many different ways. This particular choice ensures consistency with the link functions in our CTBM to facilitate meaningful empirical comparison later. The parameter estimates of CGM are provided in Table 4.

CGM* is our extension of CGM by allowing the shape parameter to also depend on $\mathbf{x}$ but still use the equation in (9) to define the recovery rate. CGM* has two link functions: $\ln[1+\exp(\alpha_0^* + \boldsymbol{\alpha}^*\mathbf{x})]$ and $\ln[1 + \exp(\beta_0^* + \boldsymbol{\beta}^*\mathbf{x})]$, and the number of parameters becomes $2k + 3$.

Note that CGM or CGM* cannot use two different cutoff values, say, $\xi_l \leq \xi_u$ because introducing them would require scaling the density in the open interval (ξ_l, ξ_u) accordingly with a factor of $1/(\xi_u - \xi_l)$. The likelihood function thus becomes unbounded when $\xi_u = \xi_l$ and cannot be properly maximized. The current approach reflected in CGM or CGM* in essence imposes the constraint of $\xi_u - \xi_l = 1$. Table 5 presents the estimation results for CGM*.

As our empirical results show later, the beta distribution used in our proposed CTBM is a better driver for recovery rates. This is not at all surprising, knowing that full recovery (i.e., $R = 1$) occurs more frequently in the data sample (see Panel A of Table 2).

Figure 2: The unconditional empirical and in-sample fitted model frequency distributions of recovery rates denoted by $h_{in}(b_j|all)$ and $\hat{h}_{in}(b_j|all)$. The sample comprises 3827 recovery rates with six prediction variables from December 1990 to 2012. Panels (a)-(e) display the results for five models: CGM*, CGM, TTM, MBB, and CTBM. The bars in green and red are the empirical and fitted model distributions, respectively. Recovery rates are divided into $m+2$ categories from left to right: 0, $\{(\frac{j-1}{m}, \frac{j}{m}]; j = 1, \cdots, m-1\}$, $(\frac{m-1}{m}, 1)$ and 1, and m is set to 20. RWSD= $\sqrt{\sum_{j=0}^{m+1}[\hat{h}_{in}(b_j|all) - h_{in}(b_j|all)]^2 h_{in}(b_j|all)}$ and WAD= $\sum_{j=0}^{m+1}|\hat{h}_{in}(b_j|all) - h_{in}(b_j|all)|h_{in}(b_j|all)$.

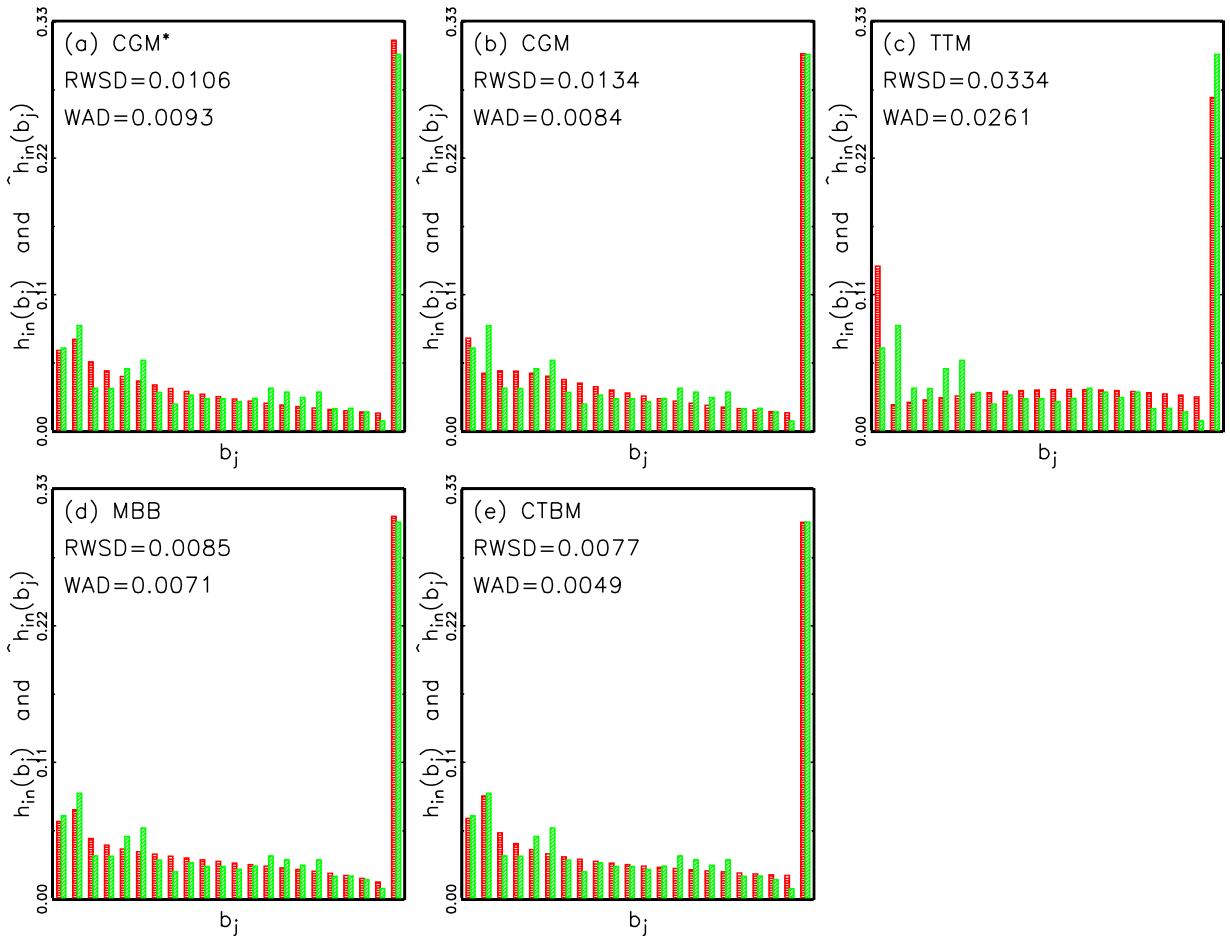


Figure 3: The conditional empirical and in-sample fitted model frequency distributions of recovery rates denoted by $h_{in}(b_j|x_0)$ and $\hat{h}_{in}(b_j|x_0)$ where x_0 is set to $(IR_1, CT_{Baseline}, IT_4, UT_{Baseline})$. The sample comprises 3827 recovery rates with six prediction variables from December 1990 to 2012, and there are 520 recovery rates with the predictor value equal to x_0. Panels (a)-(e) display the results for five models: CGM*, CGM, TTM, MBB, and CTBM. The bars in green and red are the empirical and fitted model distributions, respectively. Recovery rates are divided into $m+2$ categories from left to right: 0, $\{(\frac{j-1}{m}, \frac{j}{m}]; j = 1, \cdots, m-1\}$, $(\frac{m-1}{m}, 1)$ and 1, and m is set to 20. RWSD= $\sqrt{\sum_{j=0}^{m+1}[\hat{h}_{in}(b_j|x_0) - h_{in}(b_j|x_0)]^2 h_{in}(b_j|x_0)}$ and WAD= $\sum_{j=0}^{m+1} |\hat{h}_{in}(b_j|x_0) - h_{in}(b_j)| h_{in}(b_j|x_0)$.

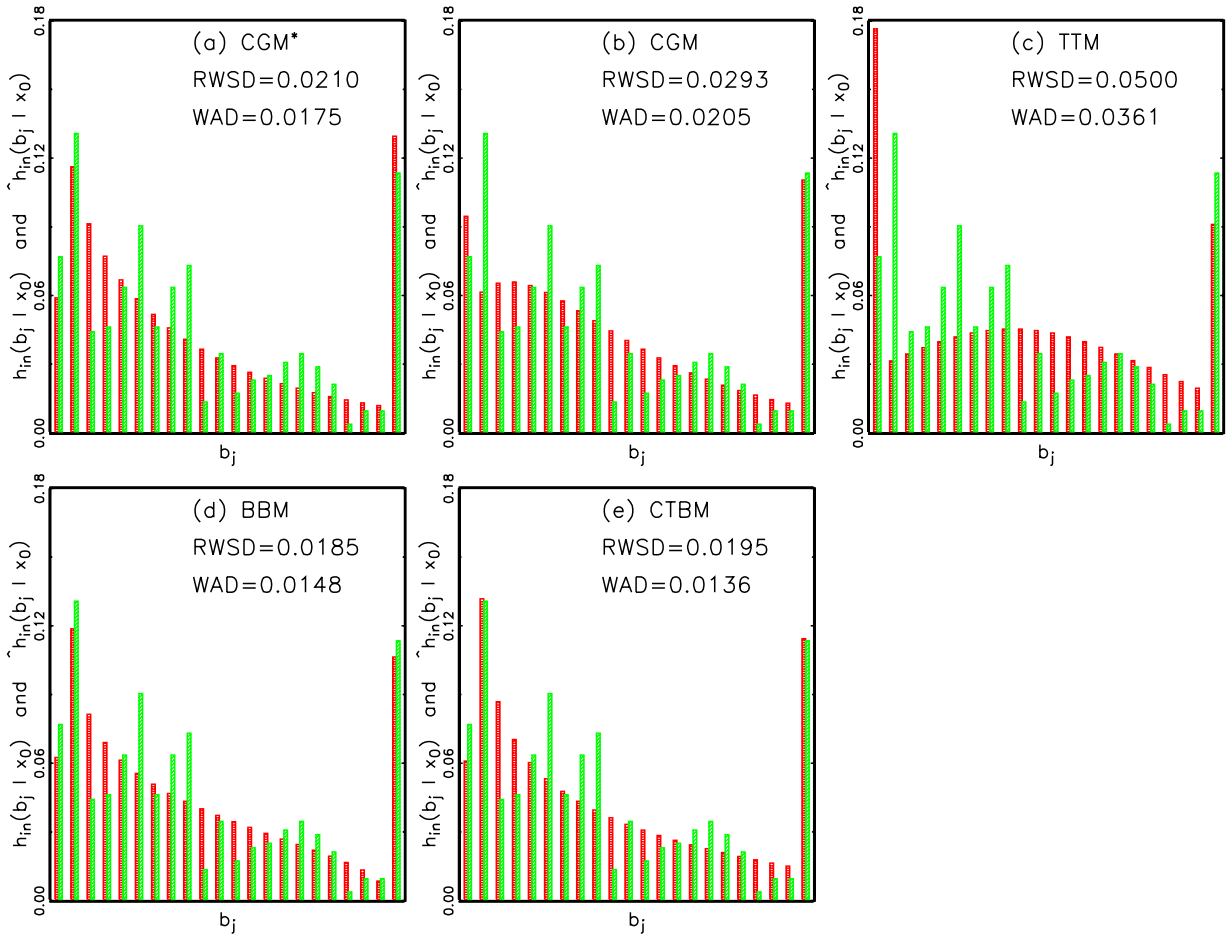
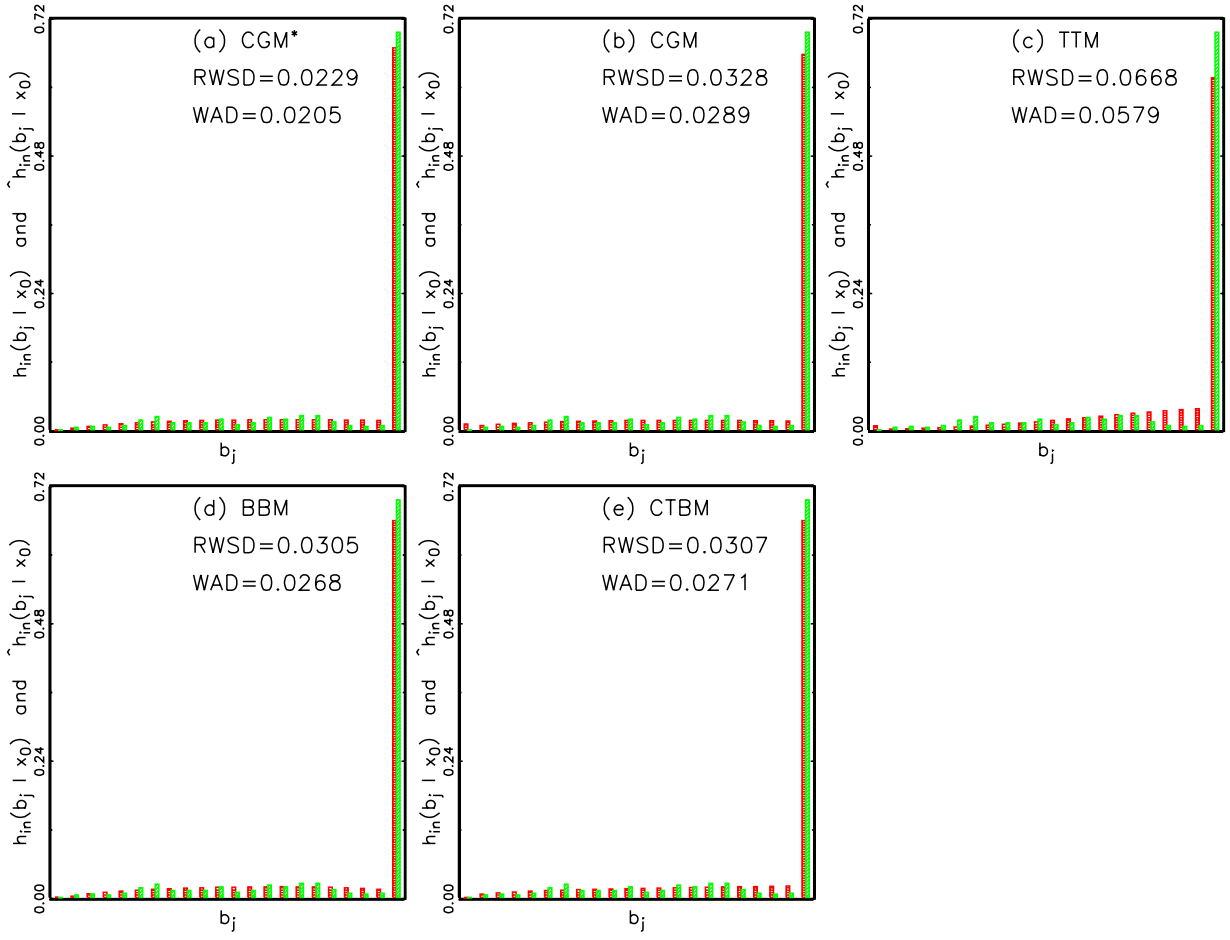


Figure 4: The conditional empirical and in-sample fitted model frequency distributions of recovery rates denoted by $h_{in}(b_j|x_0)$ and $\hat{h}_{in}(b_j|x_0)$ where x_0 is set to $(IR_{Baseline}, CT_1, IT_1, UT_{Baseline})$. The sample comprises 3827 recovery rates with six prediction variables from December 1990 to 2012, and there are 636 recovery rates with the predictor value equal to x_0. Panels (a)-(e) display the results for five models: CGM*, CGM, TTM, MBB, and CTBM. The bars in green and red are the empirical and fitted model distributions, respectively. Recovery rates are divided into $m+2$ categories from left to right: 0, $\{(\frac{j-1}{m}, \frac{j}{m}]; j = 1, \cdots, m-1\}$, $(\frac{m-1}{m}, 1)$ and 1, and m is set to 20. RWSD= $\sqrt{\sum_{j=0}^{m+1}[\hat{h}_{in}(b_j|x_0) - h_{in}(b_j|x_0)]^2 h_{in}(b_j|x_0)}$ and WAD= $\sum_{j=0}^{m+1} |\hat{h}_{in}(b_j|x_0) - h_{in}(b_j)| h_{in}(b_j|x_0)$.



TTM uses a normal random variable W with mean $\rho_0 + \boldsymbol{\rho}\mathbf{x}$ and standard deviation σ to model recovery rates with two endpoints at 0 and 1. Again, $\mathbf{x}$ is a k-dimensional recovery rate predictor. The recovery rate is modeled as

$$R = 0 \times I_{\{W \leq 0\}} + W \times I_{\{W \in (0,1)\}} + 1 \times I_{\{W \geq 1\}} \tag{10}$$

This model has $k + 2$ parameters. The estimation results for this model are presented in Table 5.

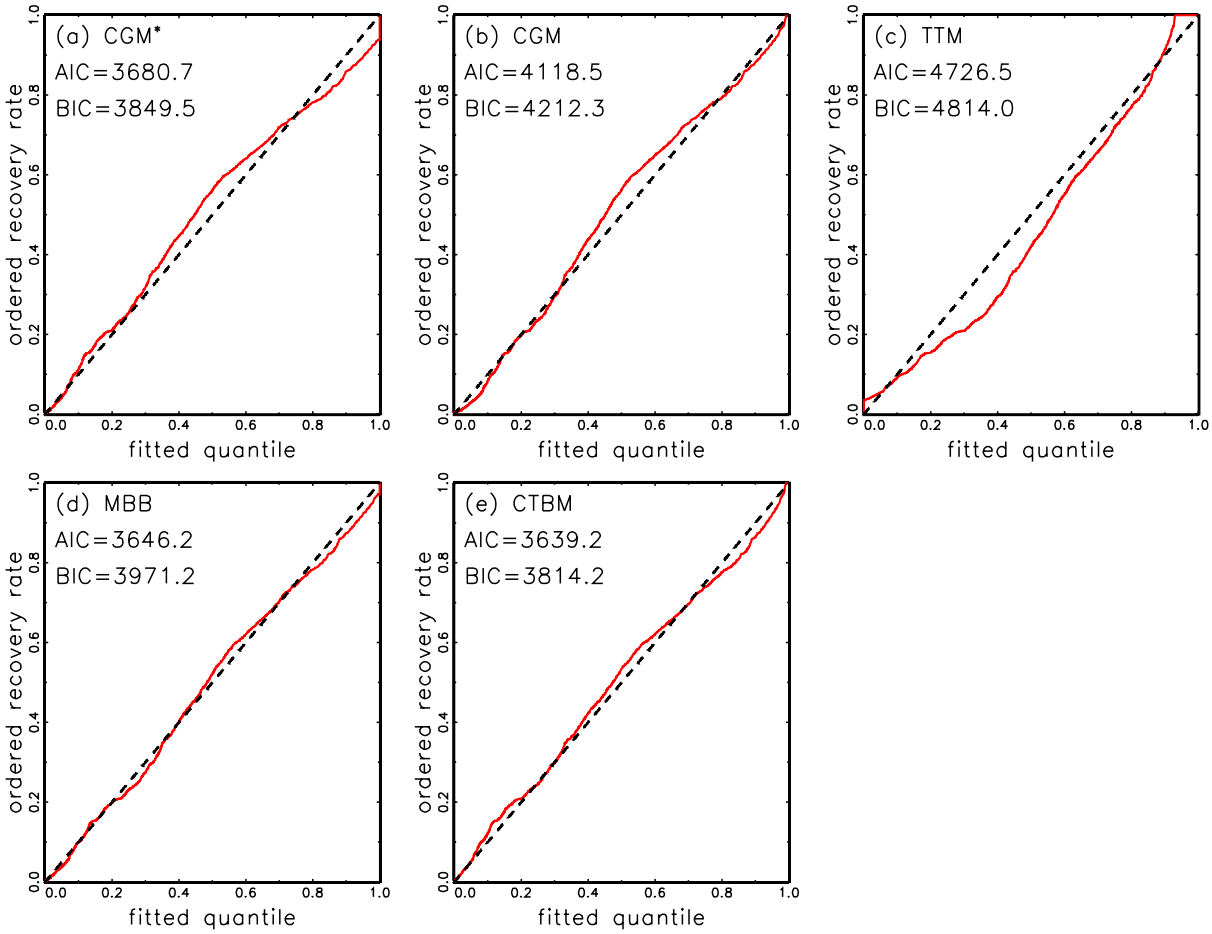
MBB first uses one Bernoulli random variable to model the occurrence of the event that the recovery rate is one of the two endpoints (0 or 1) with the probability equal to $\frac{\exp(\eta_0+\boldsymbol{\eta}\mathbf{x})}{1+\exp(\eta_0+\boldsymbol{\eta}\mathbf{x})}$. Second, it assume another Bernoulli random variable for 1, conditional on the event that an endpoint has occurred, and this conditional probability equals $\frac{\exp(\zeta_0+\boldsymbol{\zeta}\mathbf{x})}{1+\exp(\zeta_0+\boldsymbol{\zeta}\mathbf{x})}$. Finally, the conditional density for the recovery rate in the open interval $(0, 1)$ is to follow a beta distribution. This formulation is also known as the inflated beta regression by Ospina and Ferrari (2010 & 2012). Calabrese (2014) specified this beta distribution with mean $\mu = \frac{\exp(\kappa_0+\boldsymbol{\kappa}\mathbf{x})}{1+\exp(\kappa_0+\boldsymbol{\kappa}\mathbf{x})}$ so that the average recovery rate stays between 0 and 1. In addition, its variance is set to $\frac{\mu(1-\mu)}{1+\phi}$ where $\phi = \exp(-\pi_0 - \boldsymbol{\pi}\mathbf{x})$ represents the precision parameter. MBB has $4k+4$ parameters with $2k+2$ being used to define the beta distribution and $2k + 2$ to describe the probabilities of the two Bernoulli random variables.[2] Applying MBB to the same recovery rate data sample, we obtain the parameter estimates which are presented in Table 6.

Comparing CTBM with MBB is most interesting, because both models use the beta distribution as a driver. CTBM is much more parsimonious ($2k + 4$ versus $4k + 4$), however, because the probability mass for either endpoint (0 or 1) is created by simply stretching the support of the beta distribution and applying truncation. The construction of CTBM in effect bypasses the two Bernoulli random variables of MBB whose probabilities mostly likely need to depend on the debt attributes in order to work well.

Figures 2-4 exhibit the performance of CTBM discussed earlier. On the same graphs, we plot the performance of the four competing models. It is fairly clear from these plots that TTM is the worst performing model even though it is the most parsimonious among all, i.e., $k + 2$. CTBM is evidently the best performing recovery rate model. The performance of MBB and CTBM are visually comparable, but the CTBM is substantially more parsimonious ($4k+4$ versus $2k+4$). When comparing CGM with CGM*, it is clear that our extension to let the shape parameter to also depend on debt attributes improves performance, and a number of parameters in the shape parameter link function are highly significant as shown in Table 5. Even with this modification, CGM* falls short of the level of performance achieved by CTBM, suggesting perhaps non-surprisingly that the beta distribution works better for recovery modeling than does the gamma distribution.

[2]MBB defines its link functions via mean and precision of the beta distribution. An alternative way is to impose the link functions on the shape parameters as our CTBM's link functions in equations (6) and (7). The number of parameters will remain the same, and our experiment shows that MBB's empirical performance level is largely unchanged.

Figure 5: The QQ plots of the 3827 recovery rates from December 1990 to 2012 against their corresponding predicted recovery rates under five models: CGM*, CGM, TTM, MBB, and CTBM.



Model performance can also be examined by comparing QQ plots as in Figure 5. Judging from the QQ plots, the performance of CTBM and MBB are comparable. However, CTBM has fewer parameters, and this parsimony naturally places CTBM above MBB. Both AIC and BIC reported on the QQ plots indicate CTBM's clear dominance over MBB as well as other three models.

4.3 Performance study of the five models

The analyses thus far are in-sample and graphic. We now devise two performance metrics and use them to study performance both in- and out-of-sample. These two performance metrics measuring the difference between the empirical and model frequency distributions are: (1) the root weighted mean squared difference (RWSD) and (2) weighted mean absolute difference (WAD).

Recovery rates are divided into $m+2$ categories from left to right: 0, $\{(\frac{j-1}{m}, \frac{j}{m}]; j = 1, \cdots, m-1\}$,

$(\frac{m-1}{m}, 1)$ and 1, and m is set to 20 in our empirical implementation so that each category (except for 0 and 1) amounts to a 5% recovery rate range. The in-sample empirical frequency distribution can be naturally computed with the above recovery rate partition and is denoted by $h_{in}(b_j|x_0)$ where x_0 stands for the debt attribute restriction imposed on sample selection. The in-sample fitted model frequency distribution, denoted by $\hat{h}_{in}(b_j|x_0)$, is constructed by (1) estimating a model on the whole sample, (2) generating the model's distribution function for every debt in the sample meeting the debt attribute restriction indicated by x_0, (3) averaging over the fitted model frequency distributions for all debts meeting the debt attribute restriction, and (4) computing the theoretical frequency for each of the $m+2$ recovery rate categories.

The in-sample RWSD and WAD are defined as follows:

$$\begin{aligned} RWSD_{in}(x_0) &= \sqrt{\sum_{j=0}^{m+1} [\hat{h}_{in}(b_j|x_0) - h_{in}(b_j|x_0)]^2 h_{in}(b_j|x_0)}, \\ WAD_{in}(x_0) &= \sum_{j=0}^{m+1} |\hat{h}_{in}(b_j|x_0) - h_{in}(b_j)| h_{in}(b_j|x_0). \end{aligned}$$

When the entire sample is used, we will denote them by $RWSD_{in}(all)$ and $WAD_{in}(all)$. Similarly, we define $RWSD_{out}(x_0)$ and $WAD_{out}(x_0)$ by replacing $h_{in}(b_j|x_0)$ and $\hat{h}_{in}(b_j|x_0)$ with $h_{out}(b_j|x_0)$ and $\hat{h}_{out}(b_j|x_0)$.

The RWSD and WAD reported in Figure 2 are $RWSD_{in}(all)$ and $WAD_{in}(all)$ using the entire data sample of 3827 recovery rates with the models being estimated in-sample. It is evident from these two performance metrics, CTBM is the best performing model when all recovery rates are used. The same can be said when one zeros in on a sub-sample with some conditioning debt attributes as in Figure 3. But in the case of Figure 4 where a different conditioning set of debt attributes is involved, CGM* turns out to be best performing, although CTBM is not far behind. This should not be a surprise, however, because a model with the best fit overall can still be overtaken by another model for some subset of data.

In order to see how robust the above conclusions are, we now conduct a more thorough out-of-sample analysis. First, we randomly select an in-sample dataset with a size of 1914 from the entire sample of 3827 recovery rates. The remaining half of the sample will be treated as the out-of-sample dataset. Each of the five competing models is then estimated to the in-sample dataset and its $RWSD_{in}(all)$ and $WAD_{in}(all)$ are computed. The fitted model is then applied to the out-of-sample dataset to compute $RWSD_{out}(all)$ and $WAD_{out}(all)$ with 22 recovery rate categories (i.e., $m = 20$). Note that here 'all' means use all recovery rates in either the in-sample or out-of-sample dataset. Repeat the simulation, model estimation and computation of performance metrics 100 times, and then calculate the root mean squared errors (RMSE).

The RMSE of the five recovery rate models are presented in Table 7, and the results are organized under the in-sample and out-of-sample categories. The RMSE of CTBM is clearly the

smallest, implying that that CTBM proposed in this paper outperforms all other four models both in-sample and out-of-sample in terms of either performance metric (RWSD or WAD). The model that comes closest to CTBM is MBB which is far less parsimonious than CTBM. In the current implementation, MBB has 52 (i.e., $4 \times 12 + 4$) parameters whereas CTBM only involves 28 (i.e., $2 \times 12 + 4$) parameters. Although the out-of-sample performance of CTBM deteriorates somewhat, its RMSE in RWSD (or WAD) remains reasonably close to that of its in-sample counterpart.

We also conduct the in-sample and out-of-sample analysis over time. We divide the sample into two halves according to their default times. The first half is taken as the in-sample data which contains 1914 debts whose defaults occurred earlier in the sample whereas the second half, the out-of-sample data, contains the remaining 1913 defaulted debts. Similar to the random out-of-sample analysis, the results in Table 8 based on the $RWSD$ and WAD of the five models are clearly supportive of CTBM, our proposed model.

5 Conclusion

This paper presents a new beta regression model for recovery rates that is constructed by first extending the support of the beta distribution beyond $(0, 1)$ and then censoring the parts below 0 and above 1 to create point masses for the recovery rates at 0 and 1. This approach is shown to outperform four alternative models considered in this paper which relies on the gamma, beta or normal distribution as the basic driver. The performance study is based on an analysis of 20 randomly selected pairs of in-sample and out-of-sample datasets of equal size created from a sample of 3,827 defaulted debts obtained from Moody's Ultimate Recovery Database complemented by an industry distress measure constructed from the Credit Research Initiative Database at the National University of Singapore. The results indicates that the beta distribution is a better recovery rate driver, and extending/censoring the support is a better way to accommodate probability mass for recovery rate at 0 and 1.

Our empirical results also show clearly that debt attributes known from the issuance time and industry distress level at the time of default are both significant in predicting recovery rate. The typically observed bimodality in recovery rates fails to account for the available conditioning information. It is therefore critical to differentiate between conditional and unconditional recovery rate distributions. Bi-modality in recovery rate alone makes a straightforward use of an average recovery rate of, say 40%, a questionable practice. The fact that conditional distribution differs significantly from the unconditional one reinforces the need to factor in the conditioning information that is available at the time of corporate default. A better credit risk management tool can be created by suitably coupling this conditional recovery rate model with a good corporate default prediction model.

References

[1] Acharya, V., S.T. Bharath and A. Srinivasan, 2007, Does industry-wide distress affect defaulted firms? Evidence from creditor recoveries, *Journal of Financial Economics* 85, 787-821.

[2] Altman, E. and E. Kalotay, 2014, Ultimate recovery rate mixtures, *Journal of Banking and Finance* 40, 116-129.

[3] Altman, E. and V. Kishore, 1996, Almost everything you wanted to know about recoveries on defaulted bonds, *Financial Analysts Journal* 52, 57-64.

[4] Bastos, J.A., 2010, Forecasting bank loans loss-given-default, *Journal of Banking and Finance* 34, 25102517.

[5] Bellotti, T. and J. Crook, 2012, Loss given default models incorporating macroeconomic variables for credit cards, *International Journal of Forecasting* 28, 171-182.

[6] Calabrese, R., 2014, Predicting bank loan recovery rates with mixed continuous-discrete model, *Applied Stochastic Models in Business and Industry* 30, 99-114.

[7] Duan, J.-C., J. Sun and T. Wang, 2012, Multiperiod corporate default prediction – a forward intensity approach, *Journal of Econometrics* 170(1), 191-209.,

[8] Jankowitsch, R., F. Nagler and M.G. Subrahmanyam, 2014, The determinants of recovery rates in the US corporate bond market, *Journal of Financial Economics*, forthcoming.

[9] Keisman, D. and K. Van de Castle, 1999, Recovering your money: insights into losses from defaults, Standard and Poor?s Credit Week, 16-34.

[10] Maddala, G.S., 1987, Limited-dependent and Qualitative Variables in Econometrics, Cambridge University Press, New York.

[11] Ospina, R. and S.L.P. Ferrari, 2010, Inflated beta distributions, *Statistical Papers* 51, 111-126.

[12] Ospina, R. and S.L.P. Ferrari, 2012, A general class of zero-or-one inflated beta regression models, *Computational Statistics and Data Analysis* 56, 1609-1623.

[13] Papke, L.E. and J.M. Wooldridge, 1996, Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics* 11, 619632.

[14] Qi, M. and X. Zhao, 2011, Comparison of modeling methods for loss given default. *Journal of Banking and Finance* 35, 2842-2855.

[15] RMI-CRI Technical Report (Version 2013 Update 2b), 2013, Risk Management Institute, National University of Singapore.

[16] Schuermann, T., 2004, What do we know about loss given default?, in Credit Risk Models and Management edited by D. Shimko, Risk Books.

[17] Sigrist, F. and W.A. Stahel, 2012, Using the censored gamma distribution for modeling fractional response variables with an application to loss given default, Available at http://arxiv.org/pdf/1011.1796v5.pdf.

[18] Yashkir, O. and Y. Yashkir, 2013, Loss given default modeling: a comparative analysis, *Journal of Risk Model Validation* 7, 25-59.

Table 2: Distributions of recovery rates and each of the six predictors. The sample contains 3827 recovery rates and the values of the six predictors from December 1990 to 2012. The six predictors are ID, DC, IR, CT, IT, and UT with definitions in Table 1.

Variable	Mean	Std	5%	25%	50%	75%	95%
Panel A: Recovery rate							
$R_i \in [0,1]$	0.551	0.385	0.000	0.178	0.581	1.000	1.000
($R_i = 1$ with 30.36%; $R_i = 0$ with 6.69%)							
Panel B: ID							
Low ID (ID$\leq$0.359bps)	0.652	0.360	0.000	0.376	0.742	1.000	1.000
Median ID (0.359bps<ID<1.240bps)	0.566	0.378	0.000	0.205	0.601	1.000	1.000
High ID (ID$\geq$1.240bps)	0.434	0.385	0.000	0.075	0.266	0.834	1.000
Panel C: DC							
Low DC (DC=0)	0.403	0.359	0.000	0.057	0.293	0.724	1.000
Median DC (0<DC<0.339)	0.456	0.342	0.001	0.137	0.425	0.750	1.000
High DC (0.339$\leq$DC$\leq$1)	0.813	0.300	0.153	0.679	1.000	1.000	1.000
Panel D: IR							
$\text{IR}_{\text{Baseline}}$	0.737	0.324	0.101	0.494	0.960	1.000	1.000
IR_1	0.417	0.356	0.000	0.102	0.300	0.725	1.000
IR_2	0.300	0.344	0.000	0.002	0.131	0.536	1.000
IR_3	0.276	0.331	0.000	0.001	0.103	0.617	0.919
Panel E: CT							
$\text{CT}_{Baseline}$	0.374	0.354	0.000	0.036	0.246	0.671	1.000
CT_1	0.730	0.328	0.114	0.455	0.947	1.000	1.000
Panel F: IT							
$\text{IT}_{\text{Baseline}}$	0.708	0.344	0.022	0.435	0.839	1.000	1.000
IT_1	0.823	0.282	0.208	0.690	1.000	1.000	1.000
IT_2	0.593	0.338	0.103	0.209	0.574	1.000	1.000
IT_3	0.257	0.301	0.000	0.010	0.124	0.441	0.908
IT_4	0.441	0.355	0.000	0.106	0.365	0.748	1.000
IT_5	0.242	0.313	0.000	0.000	0.106	0.377	1.000
Panel G: UT							
$\text{UT}_{\text{Baseline}}$	0.535	0.384	0.000	0.162	0.547	1.000	1.000
UT_1	0.753	0.342	0.038	0.491	1.000	1.000	1.000

Table 3: Parameter estimates of CTBM. The sample contains 3827 recovery rates and six predictors from December 1990 to 2012. The p-values are based on the Wald chi-squared test.

	Θ		Ψ	
Variable	Estimate	p-value	Estimate	p-value
Intercept	0.187	0.368	1.983**	0.000
ID	-0.0530**	0.000	0.0798**	0.003
DC	-0.188	0.371	-3.788**	0.000
IR_1	-0.765**	0.000	-0.599**	0.000
IR_2	-1.291**	0.000	-0.971**	0.000
IR_3	-1.206**	0.000	-0.306	0.371
CT	0.648**	0.000	-0.129	0.671
IT_1	0.371*	0.023	-0.225	0.263
IT_2	1.144**	0.000	1.815**	0.000
IT_3	0.207	0.284	1.191**	0.002
IT_4	0.577**	0.001	0.685*	0.031
IT_5	-0.290	0.135	0.237	0.531
UT	0.100	0.567	-1.878**	0.000
C_l	0.0089**	0.000		
C_u	0.6918**	0.000		

Note: '**' and '*' indicate significance at the 1% and 5% level, respectively.

Table 4: Parameter estimates of CGM and TTM. The sample contains 3827 recovery rates and six predictors from December 1990 to 2012. The p-values are based on the Wald chi-squared test.

	β in CGM		ρ in TTM	
Variable	Estimate	p-value	Estimate	p-value
Intercept	-1.025**	0.000	0.424**	0.000
ID	-0.0366**	0.000	-0.02167**	0.000
DC	1.753**	0.000	0.733**	0.000
IR_1	-0.198**	0.000	-0.090**	0.000
IR_2	-0.311**	0.000	-0.185**	0.000
IR_3	-0.562**	0.000	-0.229**	0.000
CT	0.406**	0.000	0.199**	0.000
IT_1	0.423**	0.000	0.156**	0.000
IT_2	-0.162*	0.024	-0.038	0.176
IT_3	-0.258*	0.017	-0.103*	0.018
IT_4	0.094	0.364	0.051	0.195
IT_5	-0.274*	0.013	-0.126**	0.005
UT	0.999**	0.000	0.373**	0.000
α	1.8606**	0.000		
ξ	0.1279**	0.000		
σ			0.4169**	0.000

Note: '**' and '*' indicate significance at the 1% and 5% level, respectively.

Table 5: Parameter estimates of CGM*. The sample contains 3827 recovery rates and six predictors from December 1990 to 2012. The p-values are based on the Wald chi-squared test.

	α^*		β^*	
Variable	Estimate	p-value	Estimate	p-value
Intercept	0.284*	0.048	0.020	0.878
ID	-0.0460**	0.000	-0.0357**	0.000
DC	0.467*	0.015	1.645**	0.000
IR_1	-0.688**	0.000	0328**	0.002
IR_2	-1.349**	0.000	0.841**	0.000
IR_3	-1.189**	0.000	0.309	0.150
CT	0.869**	0.000	-0.359	0.118
IT_1	0.485**	0.007	-0.037	0.847
IT_2	1.595**	0.000	-1.300**	0.000
IT_3	0.344*	0.034	-0.934**	0.000
IT_4	0.745**	0.000	-753**	0.000
IT_5	-0.061	0.731	-0.522*	0.011
UT	0.105	0.585	1.134**	0.000
ξ^*	0.0167**	0.000		

Note: '**' and '*' indicate significance at the 1% and 5% level, respectively.

Table 6: Parameter estimates of MBB. The sample contains 3827 recovery rates and six predictors from December 1990 to 2012. The p-values are based on the Wald chi-squared test.

	η		ζ		κ		π	
Variable	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
Intercept	-1.439**	0.000	2.173**	0.000	-0.476**	0.000	-0.818**	0.000
ID	-0.0254	0.076	-0.0319	0.398	-0.0802**	0.000	-0.0532**	0.000
DC	2.956**	0.000	2.844**	0.000	0.707**	0.000	0.394**	0.004
IR_1	0.191	0.089	-2.638**	0.000	-0.213**	0.000	0.029	0.666
IR_2	0.526**	0.000	-3.190**	0.000	-0.321**	0.000	0.293**	0.002
IR_3	0.467*	0.017	-4.474**	0.000	-0.255*	0.021	0.243*	0.043
CT	0.213	0.260	1.049*	0.032	0.358**	0.000	-0.196	0.100
IT_1	0.469**	0.000	1.142	0.054	0.204*	0.014	-0.125	0.216
IT_2	-0.873**	0.000	2.089	0.051	0.218**	0.006	-0.227*	0.020
IT_3	-0.441	0.059	-0.719	0.188	-0.189	0.156	0.088	0.548
IT_4	-0.647**	0.002	0.605	0.207	0.095	0.419	0.032	0.812
IT_5	0.332	0.147	-1.043*	0.040	-0.196	0.163	-0.053	0.734
UT	1.458**	0.000	2.504**	0.000	0.360**	0.000	0.268*	0.016

Note: '**' and '*' indicate significance at the 1% and 5% level, respectively.

Table 7: The performance of the five competing models for the recovery rate distribution over the 100 pairs of randomly selected in-sample and out-of-sample datasets. The in-sample dataset with a size of 1914 is randomly selected from the entire sample of 3827 recovery rates. The remainder is treated as the out-of-sample dataset whose size is 1913. Each of the five models is estimated using the in-sample dataset and then applied to the out-of-sample dataset. The in-sample root mean squared errors (RMSE) of $RWSD$ and WAD are presented along with those from the out-of-sample analysis.

	In-sample RMSE		Out-of-sample RMSE	
	RWSD	WAD	RWSD	WAD
CGM*	0.0107**	0.0093**	0.0129**	0.0108**
CGM	0.0134**	0.0088**	0.0150**	0.0111**
TTM	0.0334**	0.0263**	0.0338**	0.0262**
MBB	0.0087**	0.0072**	0.0111**	0.0093**
CTBM	0.0079	0.0052	0.0104	0.0084

Note: '**' indicates significance of a paired t test at the 1% level. The null hypothesis is that the mean performance metric (RWSD or WAD) under a model is smaller than or equal to that of the CTBM.

Table 8: The performance of the five competing models for the recovery rate distribution in-sample and out-of-sample over time. The sample period (December 1990 to 2012) is partitioned into two halves with the first half (in-sample) containing 1914 debts with earlier default dates and the remaining 1913 debts are treated as out-of-sample. Each of the five models is estimated using the in-sample dataset and then applied to the out-of-sample dataset. The $RWSD$ and WAD of the five models are presented.

	In-sample performance		Out-of-sample performance	
	RWSD	WAD	RWSD	WAD
CGM*	0.0102	0.0075	0.0250	0.0226
CGM	0.0139	0.0093	0.0253	0.0230
TTM	0.0470	0.0377	0.0572	0.0411
MBB	0.0102	0.0080	0.0215	0.0178
CTBM	0.0092	0.0062	0.0186	0.0160