European Journal of Operational Research 240 (2015) 528–538

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

# Support vector regression for loss given default modelling

Xiao Yao *, Jonathan Crook, Galina Andreeva

*Credit Research Centre, The University of Edinburgh Business School, 29 Buccleuch Place, Edinburgh EH8 9JS, UK*

ARTICLE INFO

*Article history:*
Received 18 July 2013
Accepted 27 June 2014
Available online 5 July 2014

*Keywords:*
Support vector regression
Loss given default
Recovery rate
Credit risk modelling

ABSTRACT

Loss given default modelling has become crucially important for banks due to the requirement that they comply with the Basel Accords and to their internal computations of economic capital. In this paper, support vector regression (SVR) techniques are applied to predict loss given default of corporate bonds, where improvements are proposed to increase prediction accuracy by modifying the SVR algorithm to account for heterogeneity of bond seniorities. We compare the predictions from SVR techniques with thirteen other algorithms. Our paper has three important results. First, at an aggregated level, the proposed improved versions of support vector regression techniques outperform other methods significantly. Second, at a segmented level, by bond seniority, least square support vector regression demonstrates significantly better predictive abilities compared with the other statistical models. Third, standard transformations of loss given default do not improve prediction accuracy. Overall our empirical results show that support vector regression techniques are a promising technique for banks to use to predict loss given default.



## 1. Introduction

The introduction of the Basel II and Basel III Accords (BIS, 2005a, 2005b, 2011) requires that banks in the G20 countries hold specified amounts of capital to reduce the chance of their insolvency. The amount of capital required under the Internal Rating Based (IRB) advanced approach is based on the calculation of the proportions of defaulted loans that the bank will never recover, termed Loss Given Default (LGD). Similarly the proportion has been recovered can be defined as recovery rate (RR) equals to one minus LGD. Yet compared with the extensive research on modelling the probability of default, there is relatively little research on LGD, and that which has been published shows very poor predictive accuracy. In this paper we present improved support vector regression (SVR) models that give substantial increases in predictive accuracy compared with previously published methods.

Two types of predictive models have been applied in the empirical literature: parametric and non-parametric. Among the parametric models the most popular are linear regression models that have shown robustness and effectiveness in LGD prediction and explanation. Acharya, Bharath, and Srinivasan (2007) conclude from including the industry distress dummies into a linear regression model that industry distress conditions have negative effects on the RR of defaulted firms' debts. Their results suggest RR falls during distress periods due to both the downward trend in asset values and liquidity constraints. Qi and Yang (2009) in a study of LGD of residential mortgages demonstrate that LGD can be explained by linear regression that includes debt characteristics, with loan-to-value playing the single most important role. These results are confirmed by Khieu, Mullineaux, and Yi (2012) who estimate RR of bank loans with loan characteristics, borrower characteristics and macroeconomic conditions. They suggest loan characteristics are more significant determinants of RR than the other factors. Leow, Mues, and Thomas (2013) investigate the role of macroeconomic variables in two retail loans data sets. They find that the inclusion of macroeconomic variables can improve the prediction of residential mortgage LGD but bring little improvements for personal loan LGD.

Empirical LGD distributions are often bi-modal and usually bounded between [0,1], suggesting that a linear regression model might fit poorly. Therefore, in order to improve the fit and predictive accuracy of the model, various transformations of LGD have been tried prior to the modelling stage. Gupton and Stein (2002) propose to transform the distribution of LGD into a normal distribution by a beta distribution function and then to model the transformed target with nine factors. They conduct extensive validation studies showing that such beta transformed linear regression gives better predictions than historical average methods. Another attractive alternative to linear regression is a generalized linear model such as a fractional response model. Dermine and Neto De Carvalho (2006) employ a complementary log–log model to predict

* Corresponding author.
*E-mail address:* X.YAO-2@sms.ed.ac.uk (X. Yao).

http://dx.doi.org/10.1016/j.ejor.2014.06.043
0377-2217/

the cumulative RR of corporate loans from a Portuguese bank and report the $R^2$ as 0.13 for the 12-month prediction. Jacobs and Karagozoglu (2011) propose a beta-link generalized linear model to estimate LGD at firm and instrument levels jointly and report a significant improvement in terms of both in-sample and out-of-sample performances. Leow and Mues (2011) investigate a two-stage model to predict the LGD of UK residential mortgage loans with a combination of a probability of repossession model and a haircut model (a model that predicts a proportion of lost value for a repossessed property). This study suggests that such a two-stage modelling approach works better than a single-stage model. Calabrese (2010) applies an inflated beta regression model to predict RR of loans from The Bank of Italy where the dependent variable is assumed as a mixture of a continuous beta distribution on (0,1) and a discrete Bernoulli distribution to model the probability mass at the boundaries 0 and 1. This study shows that the out-of-sample prediction of the inflated beta regression model outperforms fractional response regression models in terms of both MSE and MAE. Bellotti and Crook (2012) benchmark a number of different transformations and algorithms to predict the LGD for a credit cards data set. Surprisingly, they find that linear regression (OLS) with no variable transformations gives greater predictive accuracy.

Although parametric models are simple to implement and easy to explain, past research reports rather poor predictions of LGD, and generalized linear regression models do not achieve significant improvements compared with linear regression. Zhang and Thomas (2010) compare both linear regression and survival regression for modelling RR of personal loans from a UK bank, and report the out-of-sample $R^2$ as low as 0.0904 for linear regression, and the parametric survival models exhibit even poorer predictions. It is also surprisingly interesting to see that given the versatility of the distribution allowed in the Cox approach, the predictive accuracies can still not be improved compared with linear regression model. Similar evidence provided by Bellotti and Crook (2012) show the model fit of simple linear regression to be rather weak with $R^2$ of 0.1428, and still the predictions of this model outperform the other ones including logit and probit models.

In contrast, non-parametric methods provide much more flexibility in modelling LGD, although literature on this topic is not as extensive as for parametric models. One of the major advantages of non-parametric methods is that they do not assume a specific distribution for LGD. Unlike parametric models which imply a specific form of the LGD distribution, non-parametric methods do not make any prior assumptions when fitting a regression model. This often leads to a better performance compared with parametric techniques, as reported by previous research. For example, Bastos (2010) compares parametric fractional response regression and a non-parametric regression tree model to forecast bank loans RR and finds that the latter is superior. More strong evidence comes from Qi and Zhao (2011) who compare six modelling methods including four parametric statistical models and two data mining techniques (decision trees and neural networks) for a mixed portfolio of bonds and loans. They find non-parametric methods perform significantly better than other parametric methods in terms of both model fit and prediction accuracy. Tong, Mues, and Thomas (2013) develop a zero adjusted gamma model to predict LGD of a UK bank where the non-parametric smoothing splines are incorporated into the predictor of a mixture gamma distribution. The findings show that such a semi-parametric formulation gives favorable out-of-sample predictions compared with the traditional linear regression.

This study focuses on another promising non-parametric data mining technique: support vector machines (SVM) and application to LGD modelling. SVM was first studied by Vapnik (1995, 1998) and are widely applied in engineering, bioinformatics and decision sciences. Previous research has revealed that SVM can not only handle non-linear problems well, but also avoid the over-fitting problem that is common in neural networks based on the principle of structural risk minimization. SVM models have been widely applied in credit risk modelling as a tool to solve classification problems such as in credit scoring, i.e. to classify credit applicants into 'Good' or 'Bad' risks. On the other hand, support vector regression (SVR) adapted to regression problems has been developed and effectively applied to non-linear regression and to time series prediction problems. However, until now only one published paper, by Loterman, Brown, Martens, Mues, and Baesens (2011), has investigated the application of SVR to LGD modelling. They conduct a comprehensive benchmarking study on six retail loan data sets with 24 techniques, some of which are two-stage models including both linear and non-linear techniques, and they find that non-linear techniques including neural networks and SVR models consistently outperform other traditional linear methods. But they do not make any further improvements on SVR models.

Our paper makes three distinct contributions based on the analysis of the RR of corporate bonds. First, the predictive performance of RR is modelled by using different intercepts or dummy variables to explain the unobservable heterogeneity of different bond seniorities. Second, SVR models are applied to losses from corporate bonds for the first time. In addition, the dataset comprises a longer time series of observations than previous studies and uses a more comprehensive set of predictor variables, including the debt characteristic, the accounting ratios from obligors' financial statements. Macroeconomic factors are also included to allow for any possible systematic differences in LGD over time. Third, the paper investigates whether transforming LGD values using a logistic or beta transformation prior to analysis can improve SVR model fitting and prediction accuracy. The results show that all SVR models substantially outperform other statistical models in terms of both model fit and out-of-sample predictive accuracy, and we find that the robustness of SVR models is comparable to that of statistical models. However, a logistic or beta transformation prior to modelling does not provide any improvement in prediction.

The rest of the paper is organized as follows. Section 2 presents the models, the data used in this research is described in Section 3. Section 4 discusses the results and conclusions are drawn in Section 5.

## 2. Models

In this section both parametric regression and SVR models are presented and the proposed SVR models are elaborated in more detail. Note that in line with literature and our data the target variable is RR instead of LGD.

### 2.1. Linear regression

Previous empirical research shows that linear regression models appear to be of comparable predictive accuracy as other more complicated statistical models (Bellotti & Crook, 2012; Qi & Zhao, 2011) even though they have the potential risk to make predictions out of the range between 0 and 1. Consider a dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ with the covariates $\mathbf{x}_i \in R^m$ which is $m$-dimensional and the related dependent variable is $y_i \in R$, and $\boldsymbol{\beta}$ denotes a vector of population parameters. The linear regression model is given as

$$\begin{aligned} y_i &= \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma^2), \end{aligned} \tag{1}$$

Maximum likelihood methods can be applied to estimate the parameters.

### 2.2. Fractional response regression

Fractional response regression is defined by Papke and Wooldridge (1996) and has been widely applied in RR modelling (Dermine & Neto De Carvalho, 2006; Bastos, 2010; Bellotti & Crook, 2012; Khieu et al., 2012). In this model, the dependent variable is bounded between 0 and 1 by imposing a link function. The model is defined as

$$E(y_i|\mathbf{x}_i) = G(\boldsymbol{\beta}^T\mathbf{x}_i), \tag{2}$$

where $G(\bullet)$ denotes some link function such as a logistic transformation function or a complementary log–log function such as:

$$\begin{aligned} G(\boldsymbol{\beta}^T\mathbf{x}_i) &= \exp(\boldsymbol{\beta}^T\mathbf{x}_i)/(1+\exp(\boldsymbol{\beta}^T\mathbf{x}_i)) \\ G(\boldsymbol{\beta}^T\mathbf{x}_i) &= \exp(-\exp(-\boldsymbol{\beta}^T\mathbf{x}_i)), \end{aligned} \tag{3}$$

and the quasi maximum likelihood function can be written as follows

$$\log L = y_i \log(G(\boldsymbol{\beta}^T\mathbf{x}_i)) + (1-y_i)\log(1-G(\boldsymbol{\beta}^T\mathbf{x}_i)). \tag{4}$$

### 2.3. Support vector regression

In the following we present three support vector regression models. The first one is least squares support vector regression (LS-SVR) proposed by Suykens and Vandewalle (1999) and Suykens et al. (2002). Two improved models are proposed based on LS-SVR.

#### 2.3.1. Least squares support vector regression

Consider the dataset given in Section 2.1. The LS-SVR is defined based on the quadratic loss function such as

$$\begin{aligned} \min \quad & J(\mathbf{w}, b; u_i) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{N} u_i^2 \\ \text{s.t.} \quad & y_i = \mathbf{w}^T\varphi(\mathbf{x}_i) + b + u_i, \quad i = 1, \ldots, N, \end{aligned} \tag{5}$$

where $\mathbf{w}$ denotes the parameter vector of the associated covariates and $b$ is the intercept. Notice that the error terms $u_i^2$ are scaled by a regularized parameter $C$, and $\varphi(\mathbf{x}_i)$ denotes the kernel function that maps the data from original data space to a higher dimensional space. This model is solved by its dual form problem which can be derived from a Lagrangian function such as

$$L(\mathbf{w}, b, u_i; \alpha_i) = J(\mathbf{w}, u_i) - \sum_{i=1}^{N}\alpha_i(\mathbf{w}^T\varphi(\mathbf{x}_i) + b + u_i - y_i),$$

where $\alpha_i$ is the Lagrangian multiplier. Based on the KKT condition, the solution of the dual form is equivalent to solving the following linear equation systems

$$\begin{pmatrix} 0 & \mathbf{e}^T \\ \mathbf{e} & \overline{\mathbf{K}} \end{pmatrix}\begin{pmatrix} b \\ \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{y} \end{pmatrix}, \tag{6}$$

where $\mathbf{e} = (\underbrace{1, \ldots, 1}_{1\times N})^T$, $\mathbf{y} = (y_1, \ldots, y_N)^T$, $\alpha = (\alpha_1, \ldots, \alpha_N)^T$, $\overline{\mathbf{K}} = \mathbf{K} + \frac{1}{C}\mathbf{I}$, where $\mathbf{K}$ is the kernel matrix with $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)\cdot\varphi(\mathbf{x}_j)$ and $\mathbf{I}$ is the identity matrix. The closed form solution is obtained as

$$\begin{cases} \boldsymbol{\alpha}^* = \overline{\mathbf{K}}^{-1}(\mathbf{y} - b^*\mathbf{e}) \\ b^* = \frac{\mathbf{e}^T\overline{\mathbf{K}}^{-1}\mathbf{y}}{\mathbf{e}^T\overline{\mathbf{K}}^{-1}\mathbf{e}} \end{cases}, \tag{7}$$

Finally the estimated regression model can be written as

$$g(\mathbf{x}) = \sum_i \alpha_i^*\mathbf{K}(\mathbf{x}_i, \mathbf{x}) + b^*. \tag{8}$$

#### 2.3.2. Least squares support vector regression with different intercepts

Now we consider extending LS-SVR by introducing heterogeneity for different groups. In this model we assume that observations in the same group have an unobserved homogeneity that can be represented by intercepts. Now consider a clustered cross sectional data set such as $D = \{(\mathbf{x}_{kj}, y_{kj})\}, j = 1, \ldots, p_k, k = 1, \ldots, M$ where $\mathbf{x}_{kj}$ denotes the covariates of the $j$th sample in the $k$th group, and $p_k$ is the number of individuals in this group. The total number of cases in the whole dataset is $p_1 + p_2 + \cdots + p_M = N$, where $M$ indicates the total number of groups in this dataset. The least squares SVR model with different intercepts can be constructed as follows

$$\begin{aligned} \min \quad & J(\mathbf{w}, b_k; u_{kj}) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2}\sum_{k=1}^{M} b_k^2 + \frac{C}{2}\sum_{k=1}^{M}\sum_{j=1}^{p_k} u_{kj}^2 \\ \text{s.t.} \quad & y_{kj} = \mathbf{w}^T\varphi(\mathbf{x}_{kj}) + b_k + u_{kj} \\ & k = 1, \ldots, M \quad j = 1, \ldots, p_k. \end{aligned} \tag{9}$$

Notice that $b_k$ is a group specific intercept. With such specifications this model is able to predict the out-of-sample individuals. The Lagrangian function of model (9) can be written as

$$\begin{aligned} L(\mathbf{w}, b_k, u_{kj}; \alpha_{kj}) = {} & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2}\sum_{k=1}^{M} b_k^2 + \frac{C}{2}\sum_{k=1}^{M}\sum_{j=1}^{p_k} u_{kj}^2 \\ & - \sum_{k=1}^{M}\sum_{j=1}^{p_k}\alpha_{kj}(\mathbf{w}^T\varphi(\mathbf{x}_{kj}) + b_k + u_{kj} - y_{kj}). \end{aligned}$$

The KKT conditions are

$$\begin{cases} \frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_k\sum_j \alpha_{kj}\varphi(\mathbf{x}_{kj}) = 0 \Rightarrow \mathbf{w} = \sum_k\sum_j \alpha_{kj}\varphi(\mathbf{x}_{kj}) \\ \frac{\partial L}{\partial b_k} = b_k - \sum_j \alpha_{kj} = 0 \Rightarrow b_k = \sum_j \alpha_{kj} \\ \frac{\partial L}{\partial u_{kj}} = Cu_{kj} - \alpha_{kj} = 0 \Rightarrow u_{kj} = \frac{\alpha_{kj}}{C} \end{cases}. \tag{10}$$

Then the dual form problem is given as

$$\min \ \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{W}\boldsymbol{\alpha} + \frac{1}{2C}\boldsymbol{\alpha}^T\boldsymbol{\alpha} - \mathbf{y}^T\boldsymbol{\alpha}. \tag{11}$$

Here $\mathbf{W}$ is a block diagonal matrix defined as $\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 & & & \\ & \mathbf{W}_2 & & \\ & & \ddots & \\ & & & \mathbf{W}_M \end{pmatrix}$, and each $\mathbf{W}_k$ is a $p_k \times p_k$ matrix with all elements equal to 1. To solve for the optimal solution it is only necessary to solve the following linear system by taking the partial derivative of model (11) with respect to $\boldsymbol{\alpha}$

$$\left(\mathbf{K} + \mathbf{W} + \frac{1}{C}\mathbf{I}\right)\boldsymbol{\alpha} = \mathbf{y}, \tag{12}$$

where $\mathbf{I}$ denotes a $N \times N$ identity matrix, and $\mathbf{K}$ is defined as above. Denoting the solution of the above equation as $\boldsymbol{\alpha}^*$, the optimal solution $(\mathbf{w}^*, b_k^*)$ for Eq. (9) is obtained as

$$\begin{aligned} \mathbf{w}^* &= \sum_k\sum_j \alpha_{kj}^*\varphi(\mathbf{x}_{kj}) \\ b_k^* &= \sum_j \alpha_{kj}^*. \end{aligned} \tag{13}$$

#### 2.3.3. Semi-parametric least squares support vector regression

This section presents a semi-parametric model where dummy variables are applied to denote the unobservable heterogeneity of the seniorities of bonds. In this semi-parametric model, we assume dummy variables influence the dependent variable linearly while other variables are still equipped with kernel functions such that

$$\begin{aligned} \min \quad & J(\mathbf{w}, b_k; u_{kj}) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} + \frac{1}{2}b^2 + \frac{C}{2}\sum_{k=1}^{M}\sum_{j=1}^{p_k}u_{kj}^2 \\ \text{s.t.} \quad & y_{kj} = \mathbf{w}^T\varphi(\mathbf{x}_{kj}) + \boldsymbol{\beta}^T\mathbf{z}_{kj} + b + u_{kj} \\ & k = 1,\ldots,M \quad j = 1,\ldots,p_k, \end{aligned} \tag{14}$$

where $\mathbf{z}_{kj}$ is a vector consisting of the dummy variables and $\boldsymbol{\beta}$ is the vector of the corresponding parameters. Here $\boldsymbol{\beta}$ is treated as a vector of fixed effects with respect to the group specific variables while $b_k$ are replaced by a common intercept $b$ as in model (5). The Lagrangian function and KKT conditions are as above

$$\begin{aligned} L(\mathbf{w}, b, u_{kj}; \alpha_{kj}) = & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} + \frac{1}{2}b^2 + \frac{C}{2}\sum_{k=1}^{M}\sum_{j=1}^{p_k}u_{kj}^2 \\ & - \sum_{k=1}^{M}\sum_{j=1}^{p_k}\alpha_{kj}(\mathbf{w}^T\varphi(\mathbf{x}_{kj}) + \boldsymbol{\beta}^T\mathbf{z}_{kj} + b + u_{kj} - y_{kj}), \end{aligned}$$

and

$$\begin{cases} \frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_k\sum_j\alpha_{kj}\varphi(\mathbf{x}_{kj}) = 0 \Rightarrow \mathbf{w} = \sum_k\sum_j\alpha_{kj}\varphi(\mathbf{x}_{kj}) \\ \frac{\partial L}{\partial \boldsymbol{\beta}} = \boldsymbol{\beta} - \sum_k\sum_j\alpha_{kj}\mathbf{z}_{kj} = 0 \Rightarrow \boldsymbol{\beta} = \sum_k\sum_j\alpha_{kj}\mathbf{z}_{kj} \\ \frac{\partial L}{\partial b} = b - \sum_k\sum_j\alpha_{kj} = 0 \Rightarrow b = \sum_k\sum_j\alpha_{kj} \\ \frac{\partial L}{\partial u_{kj}} = Cu_{kj} - \alpha_{kj} = 0 \Rightarrow u_{kj} = \frac{\alpha_{kj}}{C} \end{cases}. \tag{15}$$

The dual form is

$$\min \ \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} + \frac{1}{2}\alpha^T\mathbf{Z}\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{V}\boldsymbol{\alpha} + \frac{1}{2C}\boldsymbol{\alpha}^T\boldsymbol{\alpha} - \mathbf{y}^T\boldsymbol{\alpha}, \tag{16}$$

where $\mathbf{Z}_{ij} = \mathbf{z}_{ki}^T\mathbf{z}_{kj}$, and $\mathbf{V}$ is a $N \times N$ matrix with all elements equal to 1. All the other notations are the same as model (5). Model (16) can be solved with the same procedure as above and the linear equation systems can be obtained as follows

$$\left(\mathbf{K} + \mathbf{Z} + \mathbf{V} + \frac{1}{C}\mathbf{I}\right)\boldsymbol{\alpha} = \mathbf{y}. \tag{17}$$

The solution for $\mathbf{w}$ and $\boldsymbol{\beta}$ can be derived as

$$\begin{aligned} \mathbf{w}^* &= \sum_k\sum_j\alpha_{kj}^*\varphi(\mathbf{x}_{kj}) \\ \boldsymbol{\beta}^* &= \sum_k\sum_j\alpha_{kj}^*\mathbf{z}_{kj} \\ b^* &= \sum_k\sum_j\alpha_{kj}^*. \end{aligned} \tag{18}$$

### 2.4. Two-stage model

A two-stage model is proposed by Bellotti and Crook (2012) to predict the LGD of credit cards from a UK retail bank. They first split the LGD into three classes including LGD equal to 0 or 1 and $0 < \text{LGD} < 1$ by a decision tree, and then estimate the LGD belongs to the interval $(0,1)$ by an ordinary linear regression.

### 2.5. Transformations

Two different transformations are employed in this study; one is a logistic transformation defined as follows

$$y_{logit} = \ln\left(\frac{y}{1-y}\right), \tag{19}$$

and the other is a beta transformation that is well recognized in LGD modelling since it was proposed in Gupton and Stein's seminal paper (Gupton & Stein, 2002). The beta distribution is defined within the interval $(0,1)$ as follows

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}y^{p-1}(1-y)^{q-1}, \quad p > 0, \quad q > 0, \tag{20}$$

where $p$ and $q$ are two parameters that control the shape of distribution. Following from the idea of Moody's LossCalc model, the transformed dependent variable becomes

$$y_{beta} = N^{-1}(Beta(y; p, q)), \tag{21}$$

where $N^{-1}(\cdot)$ denotes the inverse cumulative normal distribution. We examine the applications of these two different transformation methods to all the SVR models in the following empirical study.

## 3. Data

The source of data is Moody's Ultimate Recovery Database (MURD) which contains more than 6000 default debt instruments including bonds, loans and revolvers issued by more than 1700 American companies. Here the focus is on corporate bonds that are categorized into five types: senior secured, senior unsecured, senior subordinated, subordinated and junior subordinated. The sample has 1413 observations that range from 1985 to 2012. We follow Qi and Zhao (2011) to adopt the preferred method recommended by Moody's analysts as the ultimate RR of each instrument (Moody's Analytics, 2012).[1] Table 1 describes the frequency and the percentage of each seniority as well as corresponding mean values of RR. It is clear that the mean RR tends to be higher for more senior bonds. Subordinated bonds have low frequencies (especially junior subordinated bonds) and this may affect the quality of estimation. Therefore, junior subordinated, subordinated and senior subordinated bonds are merged together, and are referred to as "Subordinated bonds". Fig. 1 shows the distribution of RR for all observations. Clearly the distribution is highly skewed between 0 and 0.2, indicating that a large proportion of observations have a RR lying in this interval.

According to Resti and Sironi (2007) the drivers of RR can be categorized into five classes: characteristics of the exposure, characteristics of the borrower, bank's internal factors and macroeconomic factors. For the exposure characteristics we only select the variables with enough non-missing values. We select the accounting and macroeconomic characteristics that are commonly used in the literature are available in our dataset (Acharya et al., 2007; Khieu et al., 2012; Qi & Zhao, 2011). The accounting information of the borrowers is incorporated by using the ticker (unique identification) of each obligor to match the bond information from MURD with financial statements of the corresponding companies in Compustat.[2] Macroeconomic variables are included to capture economic cyclical effects while bank internal factors are unavailable in this study. The summarized statistics of all variables are listed in Table 2.

*Collateral_rank* refers to the relative importance of the collateral. A higher collateral rank of an instrument is expected to be associated with a higher RR. *Percent_Above* indicates the percentage of debt of an obligor that is more senior than the current instrument. It is expected to have a negative effect on RR, because a high value of *Percent_Above* means the current instrument has to

[1] There are three methods provided in MURD to calculate RR for each instrument. (1) Discount_Settlement_Total: The nominal settlement recovery amount discounted back from each settlement instrument's trading date to the last date cash paid of the individual defaulted instruments, using the defaulted instrument's effective interest rate. (2) Discount_Liquidity_Total: The nominal liquidity recovery total discounted back from each settlement instrument's trading date to the last date cash paid of the individual defaulted instruments, using the defaulted instrument's effective interest rate. (3) Discount_Trading_Price: The trading price nominal recovery value discounted from the trading date to the instrument's last date cash paid using the effective interest rate of the pre-defaulted instrument.

[2] Compustat Database is integrated into Wharton Research Data Services.

**Table 1**
Summarized statistics of recovery rates by seniority.

| Seniority | No. | Percentage | Mean |
|---|---|---|---|
| Junior subordinated | 28 | 1.98 | 0.1628 |
| Subordinated | 174 | 12.31 | 0.3122 |
| Senior subordinated | 198 | 14.01 | 0.3065 |
| Senior unsecured | 681 | 48.20 | 0.5099 |
| Senior secured | 332 | 23.50 | 0.6287 |
| Total | 1413 | 100 | 0.4781 |

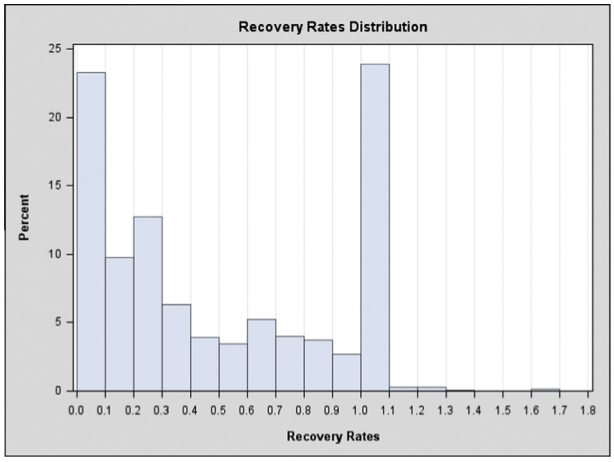


**Fig. 1.** Distribution of recovery rates.

wait for the complete recovery of the debt of senior instruments before it can be recovered. The *Orginal_Amount* denotes the face amount of the instrument when it was issued. There is no agreement on the effects of this variable on RR based on previous research. Dermine and Neto De Carvalho (2006) find that it negatively affects the RR, but Acharya et al. (2007) suggest that larger loan volume means the obligor has greater bargaining power in the bankruptcy proceedings that will result in a higher RR.

For the accounting variables, we consider both profitability and solvency characteristics of the obligors (Acharya et al., 2007). It is expected that with higher profitability and solvency, the obligor should have higher RR of its corresponding instruments. *EBITDA* is used to represent the profitability, and solvency is described by the remaining accounting variables. Notice that *Leverage* and *Debt_Ratio* are expected to have negative effects on RR because higher values indicate weak solvency of the company. All the accounting variables are lagged one year before the instrument default date.

Macroeconomic variables are considered to reflect the economic cycle over the period covered in this dataset. Because all the obligors are US companies, we only select macroeconomic variables closely related with the US economy. *GDP* and *Unemployment_Rate* are used as coincident indicators to explain the overall economic condition. *S&P_Return* is the Standard and Poor's 500 index annual return denoting the market performance for each year. The three months Treasury bill rate *T_Bill* is taken as a risk-free interest rate. All the macroeconomic variables are lagged one year before default date. In summary, the regression model can be presented as follows

$$\begin{aligned}\text{Recovery Rate}_{i,t} = {} & \text{Intercept} + \text{Recovery Characteristics}_{\boldsymbol{i}} \\ & + \text{Accounting Variables}_{i,t-1} \\ & + \text{Macroeconomic Variables}_{t-1}\end{aligned}$$

where subscript $i$ denotes the $i$th instrument and $t$ is the related default year.

## 4. Model specification and empirical results

The empirical experiment in this paper is presented in two subsections: firstly for all seniorities pooled together and secondly, models for individual seniority are presented separately.

### 4.1. Aggregated models

The aggregated sample is split into training and testing sets, with a stratified sampling method in order to keep the same proportions of different bond seniorities in both the training and test sets. For each stratum seventy percent of observations are randomly drawn as a training set and the remaining thirty percent of observations are left as a testing set. The procedure is repeated 100 times as cross-validation (with new random samples drawn each time) to ensure the robustness of the results. The regularized and the kernel parameters of SVR models are selected based on the principle of design of experiment proposed by Staelin (2003) and the out-of-sample prediction results on the testing sets are reported. To compare the performance of these models, pair-wise $t$-tests are used to examine the differences in the mean values of different models and $t$ values with corresponding significance levels are presented.

The models presented in Section 2 have been fitted to the aggregated training samples. For parametric regression techniques these include linear regression, fractional response regression where the

**Table 2**
List of variables and their explanations.

| | |
|---|---|
| *Recovery characteristics* | |
| Collateral_Rank | Instruments are ranked related to each other based on the structure prior to default, taking into consideration collateral and instrument type |
| Percent_Above | Percentage of debt which is contractually senior to the current instrument |
| Original_Amount | Total original or face amount of the relevant instrument |
| *Accounting variables* | |
| Total_Asset | Total assets of the obligor |
| EBITDA | Earnings before interest |
| Leverage | Ratio of total debt and total assets |
| Debt_Ratio | Ratio of current liabilities and long term debt |
| Netval_Share | Book value per share |
| Asset_Tangibility | Ratio between intangible assets and tangible assets |
| Quick_Ratio | Sum of cash and short-term investment and total receivables divided by the current liabilities |
| *Macroeconomic variables* | |
| GDP | Annual GDP of USA |
| S&P_Return | Annual return of S&P 500 index |
| T_Bill | Three months annual treasury bill rate of USA |
| Unemployment_Rate | US annual unemployment rate |

logistic link function is adopted, linear regression with a beta transformation and a two-stage regression model. For the two-stage model, the observations with RR = 0 and RR > 0 are first classified by logistic regression, and then the cases with RR > 0 are further separated such that RR = 1 and 0 < RR < 1, and finally an OLS regression is applied to values of RR in the interval (0,1). SVR techniques include least squares support vector regression (LS_SVR, Eq. (5)), least squares support vector regression with different intercepts (LS_SVR_DI, Eq. (9)) and semi-parametric least squares support vector regression (Semi_LS_SVR, Eq. (14)). Two different transformation methods are applied to all these three SVR models. The abbreviation names are used in the following description for convenience. For example, Beta_LS_SVR denotes the least squares support vector regression with a beta transformation on RR, and Log_Semi_LS_SVR means semi-parametric least squares support vector regression with a logistic transformation.

### 4.2. Segmented models

RR varies with respect to different bond seniorities, and to control for this and to check if segmentation affects the predictive performance of models, the aggregated dataset has been split into three subsets, as described in Section 3. Since the models are fitted to each subset separately, LS_SVR_DI and Semi_LS_SVR including related models with transformed RR are not evaluated here. All the parametric models and LS_SVR models with both original and transformed RR are applied to instrument segments. The same procedure of cross-validation is followed as described in the previous section. Similarly, the pair wise $t$-tests are also employed.

In this experiment variables including *Total Assets*, *Original Amount* of the instrument and *GDP* are subjected to a log transformation to scale the variable into an appropriate range. The outliers of each numeric variable defined as either larger than 99*th* or smaller than 1*st* percentile are replaced by the corresponding median value. Two different performance measurements are selected including root mean squared errors (RMSE), mean absolute errors (MAE) and $R$ square ($R^2$) defined as follows

$$\begin{aligned} \mathrm{RMSE} &= \sqrt{\frac{1}{n}\sum_i (r_i - \hat{r}_i)^2} \\ \mathrm{MAE} &= \frac{1}{n}\sum_i |r_i - \hat{r}_i| \\ \mathrm{R}^2 &= 1 - \frac{\sum_i (r_i - \hat{r}_i)^2}{\sum_i (r_i - \bar{r})^2}, \quad \bar{r} = \frac{1}{N}\sum_i r_i. \end{aligned} \tag{22}$$

All support vector models adopt the RBF kernels defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right).$$

where $\sigma$ is the scale parameter of the kernel and is tuned in cross-validation.

In this paper all models are implemented in SAS 9.2 (SAS Inc., 2009), where linear regression and fractional response models are fitted in SAS PROC REG and NLMIXED respectively, and all SVR models are programmed in SAS PROC IML.

### 4.3. Experimental results

#### 4.3.1. Results of aggregated models

Table 3 shows the results of cross-validation including the out-of-sample mean values and standard deviations (to indicate robustness) of all the performance metrics as well as the corresponding tuned parameters for SVR models. Table 4 gives the $t$-values of pair wise $t$-tests of differences between the mean values of RMSE, MAE and $R^2$ between each pair of methods.

**Table 3**
Cross validation results of aggregated models.

| Models | $C$ | $\sigma$ | RMSE | RMSE_sd | MAE | MAE_sd | $R^2$ | $R^2$_sd |
|---|---|---|---|---|---|---|---|---|
| M1 | – | – | 0.3258 | 0.0100 | 0.2678 | 0.0066 | 0.3044 | 0.0401 |
| M2 | – | – | 0.3931 | 0.0129 | 0.2761 | 0.0116 | 0.0137 | 0.0733 |
| M3 | – | – | 0.3193 | 0.0023 | 0.2628 | 0.0027 | 0.3263 | 0.0367 |
| M4 | – | – | 0.3343 | 0.0104 | 0.2628 | 0.0082 | 0.2673 | 0.0473 |
| M5 | 10 | 5 | 0.2357 | 0.0128 | 0.1455 | 0.0097 | 0.6353 | 0.0374 |
| M6 | 10 | 5 | 0.2165 | 0.0085 | 0.1302 | 0.0070 | 0.6920 | 0.0322 |
| M7 | 10 | 2 | 0.2136 | 0.0106 | 0.1375 | 0.0079 | 0.7006 | 0.0276 |
| M8 | 10 | 2 | 0.3021 | 0.0206 | 0.1817 | 0.0148 | 0.3999 | 0.0798 |
| M9 | 10 | 2 | 0.2762 | 0.0168 | 0.1508 | 0.0115 | 0.4986 | 0.0637 |
| M10 | 10 | 2 | 0.2726 | 0.0155 | 0.1531 | 0.0111 | 0.5116 | 0.0574 |
| M11 | 10 | 2 | 0.2442 | 0.0120 | 0.1486 | 0.0103 | 0.6100 | 0.0355 |
| M12 | 10 | 5 | 0.2491 | 0.0146 | 0.1351 | 0.0098 | 0.5921 | 0.0483 |
| M13 | 10 | 2 | 0.2402 | 0.0132 | 0.1333 | 0.0088 | 0.6210 | 0.0427 |

M1: Linear Regression; M2: Linear Regression with a Beta Transformation; M3: Fractional Response Regression; M4: Two-stage Model; M5: Least Squared Support Vector Regression; M6: Least Squared Support Vector Regression with Different Intercepts; M7: Semi-Parametric Least Squared Support Vector Regression; M8: Least Squared Support Vector Regression with a Logistic Transformation; M9: Least Squared Support Vector Regression with Different Intercepts with a Logistic Transformation; M10: Semi-Parametric Least Squared Support Vector Regression with a Logistic Transformation; M11: Least Squared Support Vector Regression with a Beta Transformation; M12: Least Squared Support Vector Regression with Different Intercepts with a beta Transformation; M13: Semi-Parametric Least Squared Support Vector Regression with a Beta Transformation.

From Tables 3 and 4 it is clear to see that all the SVR models (models M5–M13) outperform the statistical models (models M1–M4).[3] Among the statistical models, linear regression and fractional response models present similar predictive accuracy, but two-stage models appear to give less accurate predictions. Linear regression with a beta transformation shows inferior predictive performances compared with linear regression without transformation. We assume this is because the RR is not well fitted by the beta distribution so that information is lost during such transformation. It also shows that SVR models with logistic transformation give worse predictions than with a Beta transformation. Notice that the SVR models have comparable standard deviations with statistical models, showing their similar robustness.

More specifically, both LS_SVR_DI (M6) and Semi_LS_SVR (M7) outperform LS_SVR (M5) even though LS_SVR performs much better than other statistical techniques in terms of all three performance metrics. Semi_LS_SVR obtains the best performance on RMSE and $R^2$, and Panel A in Table 4 shows that such an improvement is significantly better than the other methods except for LS_SVR_DI. Panel B in Table 4 shows that LS_SVR_DI gives a significantly better performance in terms of MAE than any other model except Beta_LS_SVR_DI (M12) and Beta_Semi_LS_SVR (M13). Similarly Panel C in Table 4 provides strong evidences that LS_SVR_DI and Semi_LS_SVR yield higher $R^2$ than the others. This confirms that the models proposed in this paper are in general more accurate at predicting LGD for bonds than other established methods. Notice also from Table 3 that transformations of the dependent variable do not increase the accuracy of SVR methods. For example, Beta_LS_SVR (M11) does not lead to reduced RMSE compared to LS_SVR and a logistic transformation reduces accuracy further. Among the statistical models the fractional response model (M3) appears to be better on RMSE compared with linear regression (M1) and a two-stage model (M4) at the 5% confidence level. Surprisingly linear regression with a beta transformation (M2) gives the worst performances. In summary, all SVR models result in significantly lower errors in the test set compared with statistical models. LS_SVR_DI and Semi_LS_SVR present the highest levels

[3] All the models are labelled. See the descriptions in Table 3.

**Table 4**
Paired $t$-test for comparisons of RMSE, MAE and $R^2$ for aggregated models.

| Models | M1 | M2 | M3 | M4 | M5 | M6 | M7 | M8 | M9 | M10 | M11 | M12 | M13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Panel A. RMSE* | | | | | | | | | | | | | |
| M1 | – | | | | | | | | | | | | |
| M2 | 14.8666** | – | | | | | | | | | | | |
| M3 | −2.2840* | −20.3069** | – | | | | | | | | | | |
| M4 | 2.1242 | −12.7945** | 5.0776** | – | | | | | | | | | |
| M5 | −19.9998** | −31.2288** | −23.1776** | −21.5558** | – | | | | | | | | |
| M6 | −30.0270** | −41.2166** | −42.0922** | −31.6218** | −4.5054** | – | | | | | | | |
| M7 | −27.7606** | −38.7626** | −35.1359** | −29.3059** | −4.7946** | −0.7696 | – | | | | | | |
| M8 | −3.7317** | −13.4991** | −2.9919* | −5.0311** | 9.8714** | 13.8496** | 13.7734** | – | | | | | |
| M9 | −9.1471** | −19.8991** | −9.1645** | −10.6021** | 6.9139** | 11.4326** | 11.3623** | −3.5131** | – | | | | |
| M10 | −10.3988** | −21.5448** | −10.7455** | −11.9182** | 6.6185** | 11.4422** | 11.3286** | −4.1258** | −0.5679 | – | | | |
| M11 | −18.8351** | −30.4718** | −22.1614** | −20.4578** | 1.7467 | 6.7916** | 6.8908** | −8.7567** | −5.5885** | −5.2238** | – | | |
| M12 | −15.6273** | −26.6494** | −17.1251** | −17.1373** | 2.4883* | 6.9575** | 7.0943** | −7.5683** | −4.3900** | −3.9792** | 0.9348 | – | |
| M13 | −18.6372** | −29.8693** | −21.2853** | −20.1897** | 0.8824 | 5.4428** | 5.6652** | −9.1221** | −6.0752** | −5.7380** | −0.8085 | −1.6304 | – |
| *Panel B. MAE* | | | | | | | | | | | | | |
| M1 | – | | | | | | | | | | | | |
| M2 | 2.2423* | – | | | | | | | | | | | |
| M3 | −2.5281* | −4.0263** | – | | | | | | | | | | |
| M4 | −1.7127 | −3.3757** | 0.0000** | – | | | | | | | | | |
| M5 | −37.5846** | −31.1408** | −42.0043** | −33.2975** | – | | | | | | | | |
| M6 | −51.5678** | −38.8274** | −63.7235** | −44.3443** | −4.6117** | – | | | | | | | |
| M7 | −45.6378** | −35.6070** | −54.1136** | −39.6768** | −2.3057* | 2.4936* | – | | | | | | |
| M8 | −19.1570** | −18.1004** | −19.4367** | −17.2821** | 7.3759** | 11.3417** | 9.4993** | – | | | | | |
| M9 | −31.8153** | −27.6581** | −34.1854** | −28.5910** | 1.2702 | 5.5170** | 3.4370** | −5.9443** | – | | | | |
| M10 | −32.0240** | −27.6223** | −34.6237** | −28.6608** | 1.8589 | 6.2918** | 4.1284** | −5.5740** | 0.5188 | – | | | |
| M11 | −35.1325** | −29.6339 | −38.6696** | −31.2753** | 0.7900 | 5.3272** | 3.0832** | −6.6187** | −0.5138 | −1.0715 | – | | |
| M12 | −40.4949** | −33.4781 | −45.2949** | −36.0326** | −2.7194* | 1.4670 | −0.6874 | −9.4656** | −3.7465** | −4.3830** | −3.4237** | – | |
| M13 | −44.0861** | −35.3616 | −50.7251** | −38.8184** | −3.3586** | 0.9940 | −1.2805 | −10.1349** | −4.3573** | −5.0398** | −4.0720** | −0.4927 | – |
| *Panel C. $R^2$* | | | | | | | | | | | | | |
| M1 | – | | | | | | | | | | | | |
| M2 | −12.5447** | – | | | | | | | | | | | |
| M3 | 1.4526 | 13.7494** | – | | | | | | | | | | |
| M4 | −2.1572 | 10.4815** | −3.5533** | – | | | | | | | | | |
| M5 | 21.7580** | 27.2355** | 21.2622** | 22.0041** | – | | | | | | | | |
| M6 | 27.1741** | 30.5474** | 27.0065** | 26.7612** | 4.1424** | – | | | | | | | |
| M7 | 29.3449** | 31.6206** | 29.3893** | 28.5278** | 5.0653** | 0.7311 | – | | | | | | |
| M8 | 3.8555** | 12.8509** | 3.0212* | 5.1538** | −9.6307** | −12.2390** | −12.8400** | – | | | | | |
| M9 | 9.3024** | 18.0034** | 8.4504** | 10.5111** | −6.6724** | −9.7696** | −10.4912** | 3.4853** | – | | | | |
| M10 | 10.6694** | 19.2825** | 9.8064** | 11.8427** | −6.5102** | −9.8829** | −10.6993** | 4.0971** | 0.5466 | – | | | |
| M11 | 20.5739** | 26.3984** | 20.0331** | 20.8932** | −1.7690 | −6.1687** | −7.2645** | 8.6733** | 5.5079** | 5.2568** | – | | |
| M12 | 16.5239** | 23.7570** | 15.7985** | 17.3229** | −2.5498* | −6.2050** | −7.0323** | 7.4292** | 4.2171** | 3.8690** | −1.0767 | – | |
| M13 | 19.4874** | 25.8121** | 18.8716** | 20.0130** | −0.9083 | −4.7867** | −5.6448** | 8.8081** | 5.7548** | 5.5136** | 0.7142 | 1.6163 | – |

Values are paired $t$ statistics where a positive value means the accuracy statistic for the model on the horizontal axis is better than that for the model on the vertical axis, and vice versa. M1: Linear Regression; M2: Linear Regression with a Beta Transformation; M3: Fractional Response Regression; M4: Two-stage Model; M5: Least Squared Support Vector Regression; M6: Least Squared Support Vector Regression with Different Intercepts; M7: Semi-Parametric Least Squared Support Vector Regression; M8: Least Squared Support Vector Regression with a Logistic Transformation; M9: Least Squared Support Vector Regression with Different Intercepts with a Logistic Transformation; M10: Semi-Parametric Least Squared Support Vector Regression with a Logistic Transformation; M11: Least Squared Support Vector Regression with a Beta Transformation; M12: Least Squared Support Vector Regression with Different Intercepts with a beta Transformation; M13: Semi-Parametric Least Squared Support Vector Regression with a Beta Transformation.

* 5% Significance level.

** 1% Significance level.

**Table 5**
Cross validation results of segmented models.

| Models | $C$ | $\sigma$ | RMSE | RMSE_sd | MAE | MAE_sd | $R^2$ | $R^2$_sd |
|---|---|---|---|---|---|---|---|---|
| *Panel A. Senior secured bonds* | | | | | | | | |
| M1 | – | – | 0.2848 | 0.0361 | 0.2064 | 0.0151 | 0.3910 | 0.1595 |
| M2 | – | – | 0.3692 | 0.0387 | 0.2074 | 0.0869 | 0.0178 | 0.0243 |
| M3 | – | – | 0.1973 | 0.0044 | 0.1401 | 0.0039 | 0.5423 | 0.0778 |
| M4 | – | – | 0.2656 | 0.0283 | 0.1776 | 0.0175 | 0.4872 | 0.0561 |
| M5 | 10 | 5 | 0.2050 | 0.0202 | 0.1144 | 0.0158 | 0.6866 | 0.0604 |
| M8 | 10 | 2 | 0.2953 | 0.0296 | 0.1463 | 0.0211 | 0.3493 | 0.1263 |
| M11 | 10 | 2 | 0.2433 | 0.0247 | 0.1174 | 0.0138 | 0.5324 | 0.0970 |
| *Panel B. Senior unsecured bonds* | | | | | | | | |
| M1 | – | – | 0.2977 | 0.0128 | 0.2381 | 0.0093 | 0.3856 | 0.0607 |
| M2 | – | – | 0.3646 | 0.0176 | 0.2546 | 0.0172 | 0.0770 | 0.1088 |
| M3 | – | – | 0.2773 | 0.0032 | 0.2230 | 0.0045 | 0.4053 | 0.0516 |
| M4 | – | – | 0.3049 | 0.0136 | 0.2297 | 0.0115 | 0.3551 | 0.0681 |
| M5 | 10 | 5 | 0.2098 | 0.0146 | 0.1218 | 0.0111 | 0.6946 | 0.0415 |
| M8 | 10 | 2 | 0.2716 | 0.0305 | 0.1501 | 0.0198 | 0.4839 | 0.1143 |
| M11 | 10 | 2 | 0.2159 | 0.0143 | 0.1224 | 0.0121 | 0.6766 | 0.0417 |
| *Panel C. Subordinated bonds* | | | | | | | | |
| M1 | – | – | 0.3455 | 0.0223 | 0.2683 | 0.0146 | 0.0778 | 0.0906 |
| M2 | – | – | 0.3954 | 0.0280 | 0.2714 | 0.0234 | 0.2111 | 0.1625 |
| M3 | – | – | 0.3169 | 0.0061 | 0.2523 | 0.0075 | 0.0925 | 0.0782 |
| M4 | – | – | 0.3471 | 0.0224 | 0.2675 | 0.0155 | 0.0683 | 0.1047 |
| M5 | 10 | 5 | 0.2719 | 0.0245 | 0.1917 | 0.0150 | 0.4275 | 0.0846 |
| M8 | 10 | 2 | 0.3349 | 0.0389 | 0.1966 | 0.0262 | 0.1316 | 0.1556 |
| M11 | 10 | 2 | 0.3026 | 0.0358 | 0.1839 | 0.0214 | 0.2916 | 0.1277 |

M1: Linear Regression; M2: Linear Regression with a Beta Transformation; M3: Fractional Response Regression; M4: Two-stage Model; M5: Least Squared Support Vector Regression; M8: Least Squared Support Vector Regression with a Logistic Transformation; M11: Least Squared Support Vector Regression with a Beta Transformation.

of predictive accuracy in terms of all metrics. However, transformations of RR appear to reduce predictive accuracy.

#### 4.3.2. Results of segmented models

We test seven methods on the three seniorities of bonds and the out-of-sample performances are reported in Panel A to Panel C in Table 5 separately. The pair wise $t$-test results are presented in Table 6. In general LS_SVR outperforms the other models on all three subsets. Both Log_LS_SVR (M8) and Beta_LS_SVR appear to be inferior in terms of predictive abilities compared with the LS_SVR model without any transformation. In comparison, linear regression, fractional response and two-stage models are comparable with each other and their performances improve considerably for bonds of higher seniority, while linear regression with a beta transformation always performs worst among all methods on all subsets.

Turning to the results by segments, Panels A.1–A.3 in Table 6 exhibit the results for senior secured bonds. Whilst LS_SVR and the fractional response model obtain similar results on RMSE without significant differences, LS_SVR has a significantly lower MAE than the fractional response model. Panel A.3 shows LS_SVR has a significant larger $R^2$ than Log_LS_SVR or Beta_LS_SVR. Beta_LS_SVR shows a comparable performance on MAE with LS_SVR although this does not hold in terms of RMSE. In contrast, Log_LS_SVR gives the second worst performance on RMSE and $R^2$ while its MAE is comparable with the fractional response model. Linear regression with a beta transformation still gives the poorest predictions.

Now considering senior secured and senior unsecured bonds (Table 6 Panels B.1–B.3 and C.1–C.3), it can be seen that LS_SVR model and the Beta_LS_SVR model have no significant differences in terms of RMSE, MAE and $R^2$. Log_LS_SVR gives the least accurate predictions among the three SVR models as seen in Panels C.1–C.3 although the differences between Log_LS_SVR and Beta_LS_SVR in terms of RMSE and MAE on subordinated bonds are not statistically significant.

#### 4.3.3. Comparison of aggregated and segmented models

In this section we combine the results of segmented models to examine if the support vector models can give better results through segmenting the dataset, as compared to models estimated on the aggregated dataset. The combined results are yielded by Eq. (23). Denote the number of observations of each segment testing set as $n_i$, $i = 1, 2, 3$ and the RMSE and MAE of each segment as $\mathrm{RMSE}_i$ and $\mathrm{MAE}_i$. The combined RMSE, MAE and $R^2$ are given as follows

$$\begin{aligned}
\mathrm{RMSE}_{combined} &= \sqrt{\frac{1}{(n_1+n_2+n_3)}\sum_{i=1}^{3} n_i \times \mathrm{RMSE}_i^2} \\
\mathrm{MAE}_{combined} &= \frac{1}{(n_1+n_2+n_3)}\sum_{i=1}^{3} n_i \times \mathrm{MAE}_i \\
\mathrm{R}^2_{combined} &= \frac{1}{3}\sum_{i=1}^{3} \mathrm{R}_i^2.
\end{aligned} \tag{23}$$

Because there is no explicit form to calculate $R^2$ across different groups, we simply use the arithmetic average as the combined value. The combined results of the segmented models are given in Table 7. From Table 7 we see that in general the combined results of segmented models are better than models built on aggregated samples for almost all methods, indicating that modelling each segment separately and then combining the results together can give better predictions than modelling without segmentation. But comparisons of Table 3 with Table 7 show that the two improved versions of SVR models proposed in this study, that is LS_SVR_DI (M9) and Semi_LS_SVR, outperform all of the combined results from Table 7. This is a surprising result because the segmented models allow for all parameters to be estimated for each segment separately, whereas the DI models only allow the intercept to be sector specific. This type of result has been observed before in the context of default prediction (Banasik, Crook, & Thomas, 1996) and may be due to the smaller sample size when segmented data used. This result suggests that given the sizes of available datasets our improved SVR models can capture the characteristics of each segment better than segmented models. In other words, the segmented models take longer to estimate and are less accurate compared with our proposed methods.

In summary several conclusions can be drawn as follows.

(i) LS_SVR models present superior in-sample model fitting and out-of-sample predictive abilities compared with statistical models when used to model RR of corporate bonds at both an aggregated and at a segmented level. For aggregated models, given the sizes of available data sets, the improved model LS_SVR_DI proposed in this paper is able to make better use of the bond seniority characteristics and give significantly lower RMSE and MAE values and higher $R^2$ than LS_SVR models. Another improved version, Semi_LS_SVR, which assumes that dummy variables have linear effects on the dependent variable, also suggests that such modifications can yield similar performances to LS_SVR_DI.

(ii) For the segmented models, fractional response regression and the LS_SVR give close predictions for senior secured bonds, but LS_SVR is more accurate when it comes to lower seniority bonds. Among the statistical models, fractional response models show the most accurate predictions for all seniorities of bonds, but their performances are inferior to SVR models. Linear regression with a beta transformation always gives the poorest performance throughout the study.

(iii) We explore the effects of the transformations of RR. For aggregated models no matter whether RR is transformed by a logistic or a beta distribution, the performances of all SVR models are noticeably worse than without the transformation. The MAE of Beta_LS_SVR_DI and Beta_Semi_LS_SVR are lower compared with LS_SVR, but there are no significant differences compared with LS_SVR_DI and Semi_LS_SVR. Little improvements can be seen in terms of

**Table 6**
Paired $t$-test for comparisons of RMSE, MAE and $R^2$ for segmented models.

| Models | M1 | M2 | M3 | M4 | M5 | M8 | M11 |
|---|---|---|---|---|---|---|---|
| *Panel A.1. RMSE on senior secured bonds* | | | | | | | |
| M1 | – | | | | | | |
| M2 | 4.2193** | – | | | | | |
| M3 | −6.3657** | −11.6768** | – | | | | |
| M4 | −1.1074 | −5.7171** | 6.3095** | – | | | |
| M5 | −5.1038** | −9.9516** | 0.9854 | −4.6113** | – | | |
| M8 | 0.5951 | −4.0130** | 8.6644** | 1.9188 | 6.6668** | – | |
| M11 | −2.5102* | −7.2554** | 4.8509** | −1.5707 | 3.1757* | −3.5687* | – |
| *Panel A.2. MAE on senior secured bonds* | | | | | | | |
| M1 | – | | | | | | |
| M2 | 0.0300 | – | | | | | |
| M3 | −11.2477** | −2.0470 | – | | | | |
| M4 | −3.2966* | −0.8894 | 5.5337** | – | | | |
| M5 | −11.1374** | −2.7858* | −4.1781** | −7.0920** | – | | |
| M8 | −6.1284** | −1.8077 | 0.7645 | −3.0209* | 3.2018* | – | |
| M11 | −11.5111** | −2.7062* | −4.1880** | −7.1467** | 0.3784 | −3.0328* | – |
| *Panel A.3. $R^2$ on senior secured bonds* | | | | | | | |
| M1 | – | | | | | | |
| M2 | −6.1199** | – | | | | | |
| M3 | 2.2557 | 17.0256** | – | | | | |
| M4 | 1.5053 | 20.3137** | −1.5199 | – | | | |
| M5 | 4.5856** | 27.1789** | 3.8762** | 6.3998** | – | | |
| M8 | −0.5423 | 6.8192** | −3.4423* | −2.6400* | −6.3744** | – | |
| M11 | 2.0040 | 13.6154** | −0.2106 | 1.0672 | −3.5703* | 3.0420* | – |
| *Panel B.1. RMSE on senior unsecured bonds* | | | | | | | |
| M1 | – | | | | | | |
| M2 | 8.1333** | – | | | | | |
| M3 | −4.0908** | −12.9118** | – | | | | |
| M4 | 1.0200 | −7.1014** | 5.2266** | – | | | |
| M5 | −11.9775** | −17.9103** | −11.9484** | −12.6102** | – | | |
| M8 | −2.0877 | −6.9875** | −0.4918 | −2.6382* | 4.8354** | – | |
| M11 | −11.2767** | −17.3489** | −11.0859 | −11.9320** | 0.7897 | −4.3748** | – |
| *Panel B.2. MAE on senior unsecured bonds* | | | | | | | |
| M1 | – | | | | | | |
| M2 | 2.2326 | – | | | | | |
| M3 | −3.8669** | −4.7025** | – | | | | |
| M4 | −1.5027 | −3.1841* | 1.4355 | – | | | |
| M5 | −21.2486** | −17.1638** | −22.3545** | −17.8611** | – | | |
| M8 | −10.6433** | −10.5417** | −9.4989** | −9.1976** | 3.2986* | – | |
| M11 | −20.0585** | −16.6321** | −20.6173** | −17.0064** | 0.0967 | −3.1583* | – |
| *Panel B.3. $R^2$ on senior unsecured bonds* | | | | | | | |
| M1 | – | | | | | | |
| M2 | −6.5535** | – | | | | | |
| M3 | 0.6542 | 7.2133** | – | | | | |
| M4 | −0.8846 | 5.7324** | −1.5545 | – | | | |
| M5 | 11.1183** | 14.0324** | 11.5590** | 11.2633** | – | | |
| M8 | 2.0096 | 6.8221** | 1.6582 | 2.5613* | −4.5843** | – | |
| M11 | 10.4546 | 13.6151** | 10.8193** | 10.6522** | −0.8095 | 4.1903** | – |
| *Panel C.1. RMSE on subordinated bonds* | | | | | | | |
| M1 | – | | | | | | |
| M2 | 3.6883* | – | | | | | |
| M3 | −3.2730* | −7.2476** | – | | | | |
| M4 | 0.1339 | −3.5638* | 3.4417* | – | | | |
| M5 | −5.8778** | −8.7823** | −4.7156** | −5.9934** | – | | |
| M8 | −0.6255 | −3.3397* | 1.2095 | −0.7191 | 3.6257* | – | |
| M11 | −2.6911* | −5.4022** | −1.0418 | −2.7879* | 1.8724 | −1.6165 | – |
| *Panel C.2. MAE on subordinated bonds* | | | | | | | |
| M1 | – | | | | | | |
| M2 | 0.2974 | – | | | | | |
| M3 | −2.5791* | −2.0565 | – | | | | |
| M4 | −0.0994 | −0.3676 | 2.3355 | – | | | |
| M5 | −9.6819** | −7.5865** | −9.5604** | −9.2977** | – | | |
| M8 | −6.3248** | −5.6337** | −5.4075** | −6.1621** | 0.4294 | – | |
| M11 | −8.6197** | −7.3007** | −7.9806** | −8.3707** | −0.7897 | −0.9933 | – |

**Table 6** (*continued*)

| Models | M1 | M2 | M3 | M4 | M5 | M8 | M11 |
|---|---|---|---|---|---|---|---|
| *Panel C.3. $R^2$ on subordinated bonds* | | | | | | | |
| M1 | – | | | | | | |
| M2 | 1.8956 | – | | | | | |
| M3 | 0.3250 | −1.7400 | – | | | | |
| M4 | −0.1815 | −1.9545 | −0.4900 | – | | | |
| M5 | 7.4640** | 3.1252* | 7.6934** | 7.0602** | – | | |
| M8 | 0.7905 | −0.9349 | 0.5940 | 0.8930 | −4.4203** | – | |
| M11 | 3.6127* | 1.0305 | 3.5179* | 3.5777* | −2.3473* | 2.1030 | – |

Values are paired t statistics where a positive value means the accuracy statistic for the model on the horizontal axis is better than that for the model on the vertical axis, and vice versa. M1: Linear Regression; M2: Linear Regression with a Beta Transformation; M3: Fractional Response Regression; M4: Two-stage Model; M5: Least Squared Support Vector Regression; M8: Least Squared Support Vector Regression with a Logistic Transformation; M11: Least Squared Support Vector Regression with a Beta Transformation.

* 5% Significance level.

** 1% Significance level.

**Table 7**
Comparison of combined results of segmented models and aggregated models.

| Models | Combined results | | | Aggregated models | | |
|---|---|---|---|---|---|---|
| | RMSE | MAE | $R^2$ | RMSE | MAE | $R^2$ |
| M1 | 0.3091 | 0.2392 | 0.2848 | 0.3258 | 0.2678 | 0.3044 |
| M2 | 0.3746 | 0.2483 | 0.1019 | 0.3931 | 0.2761 | 0.0137 |
| M3 | 0.2732 | 0.2118 | 0.3467 | 0.3193 | 0.2628 | 0.3263 |
| M4 | 0.3090 | 0.2281 | 0.3035 | 0.3343 | 0.2628 | 0.2673 |
| M5 | 0.2280 | 0.1398 | 0.6029 | 0.2357 | 0.1455 | 0.6353 |
| M8 | 0.2962 | 0.1623 | 0.3216 | 0.3021 | 0.1817 | 0.3999 |
| M11 | 0.2495 | 0.1386 | 0.5002 | 0.2442 | 0.1486 | 0.6100 |

M1: Linear Regression; M2: Linear Regression with a Beta Transformation; M3: Fractional Response Regression; M4: Two-stage Model; M5: Least Squared Support Vector Regression; M8: Least Squared Support Vector Regression with a Logistic Transformation; M11: Least Squared Support Vector Regression with a Beta Transformation.

$R^2$. For the segmented models it is noticed that the Beta_LS_SVR model shows superior performances compared with Log_LS_SVR, but it does not make significant improvements compared with LS_SVR. Therefore applying a transformation to RR before modelling is not necessarily a desirable thing to do.

(iv) In this study we focus on the predictive abilities of RR models and consider three performance metrics where RMSE and MAE are absolute measure of goodness of fit and $R^2$ is the relative measure. Most empirical research on LGD modelling is interested in identifying the determinant variables instead of the out-of-sample prediction accuracies. It is hard to compare our empirical results with other similar research directly because the data set and variables used in this study are not completely the same as in the others. However, it is still interesting to make some comparisons to show the superior performance of our proposed SVR models. We selectively summarize the most recent studies on LGD/RR modelling in Table 8. One of the most comparable studies is Qi and Zhao (2011) which compares six different techniques on the debt RR of MURD from 1985 to 2009 and achieves the best cross-validation results with $R^2$ of 0.529 for neural networks, (for comparison, Semi_LS_SVR model proposed in our study achieves a $R^2$ of 0.7002). Jacobs and Karagozoglu (2011) propose a beta-link generalized linear model to predict corporate bond instrument level RR from MURD from 1985 to 2008 and report an out-of-sample $R^2$ of 0.6119. This is still outperformed by our proposed SVR models. Khieu et al. (2012) consider the bank loans RR from MURD, and the in-sample $R^2$ of the models used in their study are reported to be between 0.2 and 0.3. Leow et al. (2013) obtain an out-of-sample $R^2$ of 0.3129 for mortgage loan LGD from a two-stage model, and the out-of-sample $R^2$ is 0.1428 for personal loan LGD. Bellotti and Crook (2012) show a very weak fit of linear regression with a $R^2$ of 0.11 for credit card recovery data from a UK bank. In Leow and Mues (2011) the proposed two-stage model is reported to obtain an out-of-sample $R^2$ as 0.268 compared with 0.233 from a single stage model on a UK residential mortgage loan recovery data set. The only paper that studies SVR for LGD modelling is by Loterman et al. (2011), which benchmarks 24 different techniques on six bank retail RR data sets and reports that the LS-SVR and neural network models consistently outperform the other methods. They apply eight performance metrics to evaluate the model performances, and $R^2$ reported in this study is still less than 0.5 for the best model for each data set. Whilst comparison across different studies should be treated with care because of differences in the data (as already noted above), we have shown that our SVR models make substantial increases in predictive accuracies compared with the literature.

## 5. Conclusions

As far as we know there is no paper that compares the predictive performances of SVR methods to predict the RR of defaulted corporate instruments. The aims of this research were first to

**Table 8**
Comparisons of LGD/RR predictive performances of selective literature.

| Authors | Data | Techniques | $R^2$ |
|---|---|---|---|
| Qi and Zhao (2011) | MURD, loans and bonds, from 1985 to 2009 | Neural networks | 0.529 |
| Jacobs and Karagozoglu (2011) | MURD, bonds, from 1985 to 2008 | Beta-link generalized linear model | 0.6119 |
| Khieu et al. (2012) | MURD, loans, from 1987 to 2007 | Linear and fractional response regression | 0.2–0.3 |
| Leow et al. (2013) | Mortgage loan | Two-stage model | 0.3129 |
| Leow et al. (2013) | Personal loan | Linear regression | 0.1428 |
| Bellotti and Crook (2012) | Personal loan | Linear regression | 0.11 |
| Leow and Mues (2011) | Mortgage loan | Two-stage model | 0.233 |
| Loterman et al. (2011) | Bank personal loan | Both parametric and non-parametric methods | 0–0.5 |

investigate whether SVR methods give more accurate predictions of RR for such instruments than other methods in the literature and, second, to devise novel SVR methods that are able to explain the unobservable heterogeneity of bond seniorities which would allow a financial institution to predict RR for these instruments more accurately than other currently available techniques. We have proposed two SVR models; one that specifies different intercepts for the seniorities of the instruments and a second includes dummy variables as a semi-parametric SVR.

By comparing the predictive accuracy of these two models with available techniques using a large sample of defaulted instruments that are observed between 1985 and 2012 we draw the following conclusions. First, when treating all of the instruments in aggregate, both SVR techniques allow more accurate predictions of RR to be made than linear regression, fractional response regression or a two-stage method that is commonly used in practice. Second, if we consider instruments segmented into seniority classes and model the RR within each class separately, SVR gives more accurate predictions than the other techniques for more senior categories of bonds and LS_SVR and fractional response models give predictions with similar accuracy at lower levels of seniority. Third, by incorporating unobservable heterogeneity the improved SVR methods parameterized on an aggregate sample, surprisingly, gives more accurate predictions than one parameterized on subsamples. Fourth, transformations of the RR do not improve the predictive accuracy of SVR models and may well make things worse. Fifth, although published work has used different datasets, over different time periods and for different credit segments compared with our work, the proposed SVR methods we present appear to give more accurate predictions than those quoted by other papers.

A limitation of using SVR techniques to predict RR is that they have the characteristics of a 'black box' in that the role of each variable is difficult to discern. Nevertheless in the context of predicting RR this is less important than predictive accuracy.

As we explained earlier LGD is an important component of the regulatory capital formula in the Basel Accords. By adopting a more accurate method to predict LGD than the method currently used, a bank can more accurately compute the regulatory capital that is required and so gain a more accurate estimate of the amount of Tier 1 capital that it needs to hold so as to fulfil the requirements of its national regulator.

## References


Acharya, V. V., Bharath, S. T., & Srinivasan, A. (2007). Does industry-wide distress affect defaulted firms? Evidences from creditor recoveries. *Journal of Financial Economics, 85*, 787–821.

Banasik, J., Crook, J., & Thomas, L. (1996). Does scoring a subpopulation make a difference. *International Review of Retail Distribution and Consumer Research, 6*(2), 180–195.

Bastos, J. A. (2010). Forecasting bank loans loss-given-default. *Journal of Banking and Finance, 34*(10), 2510–2517.

Bellotti, T., & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting, 28*(1), 171–182.

Basel Committee on Banking Supervision (2005a). *Guidance on paragraph 468 of the framework document*.

Basel Committee on Banking Supervision (2005b). *An explanatory note on the Basel II IRB risk weight functions*.

Basel Committee on Banking Supervision (2011). *Basel III counterparty credit risk frequently asked questions*.

Calabrese, R. (2010). Predicting bank loan recovery rates in a mixed continuous-discrete model. *Working paper*.

Dermine, J., & Neto De Carvalho, C. (2006). Bank loan losses-given-default: A case study. *Journal of Banking and Finance, 30*, 1219–1243.

Gupton, G. M., & Stein, R. M. (2002). LossCalc™: Model for predicting loss given default (LGD). *Moody's KMV*.

Jacobs, M., Jr., & Karagozoglu, A. K. (2011). Modelling ultimate loss-given-default on corporate debt. *Journal of Fixed Income, 21*(1), 6–20.

Khieu, H. D., Mullineaux, D. J., & Yi, H. C. (2012). The determinants of bank loan recovery rates. *Journal of Banking and Finance, 36*, 923–933.

Leow, M., & Mues, C. (2011). Predicting loss given default (LGD) for residential mortgage loans: A two-stage model and empirical evidence for UK bank data. *International Journal of Forecasting, 28*(1), 183–195.

Leow, M., Mues, C., & Thomas, L. (2013). The economy and loss given default: evidence form two UK retail lending data sets. *Journal of Operational Research Society*, 1–13.

Loterman, G., Brown, I., Martens, D., Mues, C., & Baesens, B. (2011). Benchmarking regression algorithms for loss given default modelling. *International Journal of Forecasting, 28*(1), 161–170.

Moody's Analytics (2012). *Default & recovery database DRD technical specifications*.

Papke, L., & Wooldridge, J. (1996). Econometric method for fractional response variables with an application to the 401(K) plan participation rates. *Journal of Applied Econometrics, 11*, 619–632.

Qi, M., & Yang, X. (2009). Loss given default of high loan-to value residential mortgages. *Journal of Banking and Finance, 33*, 788–799.

Qi, M., & Zhao, X. (2011). Comparison of modelling methods for loss given default. *Journal of Banking and Finance, 35*, 2842–2855.

Resti, A., & Sironi, A. (2007). *Risk management and shareholders' value in banking*. John Wiley & Sons, Ltd.

SAS Institute Inc. (2009). *SAS 9.2 user's guide*. Cary, NC.

Staelin, C. (2003). Parameter selection for support vector machines. *Technical Report HPL-2002-354*. HP Laboratories Israel.

Suykens, J. A. K., Gestel, T. V., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machine*. Singapore: World Scientific.

Suykens, J. A. K., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters, 9*, 293–300.

Tong, E. N. C., Mues, C., & Thomas, L. (2013). A zero-adjusted gamma model for mortgage loan loss given default. *International Journal of Forecasting, 29*, 548–562.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Vapnik, V. (1998). *Statistical learning theory*. New York: John Wiley.

Zhang, J., & Thomas, L. (2010). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting, 28*(1), 204–215.