THE UNIVERSITY
of EDINBURGH

Modelling Loss Given Default of Corporate Bonds and Bank Loans

Xiao Yao

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Management Science and Business Economics

2015

Statement of Originality

This thesis has been composed by myself and contains no material that has been accepted for the award of any other degree at any university.

A part of this thesis has been published in the *European Journal of Operational Research:*

X. Yao, J. Crook and G. Andreeva (2014). “Support vector regression for loss given default”, forthcoming.

Permission to include text from that paper has been gained from the publisher and the authors.

To the best of my knowledge and belief this thesis contains no other material previously published by any other person except where due acknowledgment has been made.

Xiao Yao
Mar 2015

Table of Contents

List of Tables

List of Figures

Acknowledgement

When I complained to my supervisors how struggling I was during the writing of the Introduction and Conclusion chapters, I never believed this section could be the most difficult part in my dissertation. It is only because there are so many people I would like to thank for their help but I just can not list all the names in one page.

I would like to thank my supervisor Prof. Jonathan Crook for his patience on my endless questions in our fortnightly meetings and some of them are truly stupid when I look back. Without his valuable advice and guidance I would never be able to complete my PhD. I would also thank my supervisor Dr. Galina Andreeva who always gives me constructive opinions for improvement. I have also learnt a lot from her lectures and tutorials. I consider myself very fortunate to have them as my supervisors and it is a great honor and pleasure to work with them.

I would also like to express my gratitude to my friends and the staff in Credit Research Centre and Business School of Edinburgh University for the pleasant working atmosphere and all the discussions.

Last, but definitely the most important, is the support from my parents and all my families which motivates me to progress all the time. It is my parents' two hours nagging me every Saturday's afternoon that keeps me energetic on my research. I am also grateful for the support of my grandparents who are eager to hear every piece of good news from me.

Abstract

Loss given default (LGD) modelling has become increasingly important for banks as they are required to comply with the Basel Accords for their internal computations of economic capital. Banks and financial institutions are encouraged to develop separate models for different types of products. In this thesis we apply and improve several new algorithms including support vector machine (SVM) techniques and mixed effects models to predict LGD for both corporate bonds and retail loans.

SVM techniques are known to be powerful for classification problems and have been successfully applied to credit scoring and rating business. We improve the support vector regression models by modifying the SVR model to account for heterogeneity of bond seniorities to increase the predictive accuracy of LGD. We find the proposed improved versions of support vector regression techniques outperform other methods significantly at the aggregated level, and the support vector regression methods demonstrate significantly better predictive abilities compared with the other statistical models at the segmented level.

To further investigate the impacts of unobservable firm heterogeneity on modelling recovery rates of corporate bonds a mixed effects model is considered, and we find that an obligor-varying linear factor model presents significant improvements in explaining the variations of recovery rates with a remarkably high intra-class correlation being observed. Our study emphasizes that the inclusion of an obligor-varying random effect term has effectively explained the unobservable firm level information shared by instruments of the same issuer.

At last we incorporate the SVM techniques into a two-stage modelling framework to predict recovery rates of credit cards. The two-stage model with a support vector machine classifier is found to be advantageous on an out-of-time sample compared with other methods, suggesting that an SVM model is preferred to a logistic regression at the classification stage. We suggest that the choice of regression models is less influential in prediction of recovery rates than the choice of classification methods in the first step of two-stage models based on the empirical evidence.

The risk weighted assets of financial institutions are determined by the estimates of LGD together with PD and EAD. A robust and accurate LGD model impacts banks when making business decisions including setting credit risk strategies and pricing credit products. The regulatory capital determined by the expected and unexpected

losses is also important to the financial market stability which should be carefully examined by the regulators. In summary this research highlights the importance of LGD models and provides a new perspective for practitioners and regulators to manage credit risk quantitatively.

Chapter 1
Introduction

1.1. Introduction

The new Basel Accord issued in 2004 (Basel Committee, 2004) encouraged banks and financial institutions to manage financial risk quantitatively, where credit risk management is deemed to be an integrated component of risk management system. To better allocate capital, manage credit exposures and evaluate prices of financial instruments accurately, it is necessary to develop proper advanced internal quantitative models. The new Basel Accord suggested that credit risk should be evaluated based on three risk parameters including probability of default (PD), loss given default (LGD) and exposure at default (EAD). The expected and unexpected losses of credit portfolio are estimated based on the evaluation of the above key risk parameters. Previous research has been devoted to PD modelling largely. This thesis places the emphasis on modelling LGD for corporate bonds and bank loans by improving and developing innovative methodologies to increase the predictive accuracy of LGD. We make contributions to the literature by proposing several new algorithms. Firstly we develop new techniques based on support vector machine to improve the predictive performances of LGD for corporate bonds. Secondly we propose to model LGD of corporate bonds by accounting for the unobservable heterogeneity. Thirdly a new two-stage model that combines support vector machine technique and statistical regression models is developed to increase the predictive accuracy of LGD for retail credit cards. This chapter is organized as follows. Section 1.2 introduces the research background and motivations, and Section 1.3 lays out the research questions and contributions. Section 1.4 presents the structure of this thesis, and at last Section 1.5 concludes this chapter.

1.2. Research background and motivations

Loss given default (LGD) is the proportion of an exposure at the time of default that a lender does not recover. It is a key parameter in modelling both the regulatory capital that a bank is required to hold by regulators and the economic capital, the amount of capital that a bank believes it needs to hold to protect depositors. LGD equals one minus the recovery rate of the exposure. The Basel II Accord of 2004 stated that banks should keep adequate capital to buffer the credit losses based on the

evaluation of their credit exposures. Therefore it is necessary to estimate not only the probability that a financial instrument defaults, but also how much losses the bank will bear conditioning on default. This is true not only for banks estimating their expected and unexpected losses more accurately, but also for calculating the required regulatory capital properly to keep their competitiveness. The release of New Basel Accord (2004) encouraged financial institutions to shift from using the standardized approach where the estimates of risk parameters are provided by external agencies or regulators to adopting the internal rating based (IRB) approach including a foundation and an advanced approach. To be specific, banks that adopt a foundation IRB (FIRB) approach are only allowed to use the internal estimate of PD while estimates of other parameters are still given from outside. Under the FIRB approach banks have access to an array of pre-set LGD values with respect to the types of loans and collateral. In contrast, Banks that adopt the advanced IRB (AIRB) approach will be allowed to use the internal estimates of all parameters including LGD. In the case of retail portfolios FIRB is not available and thus the AIRB should be adopted. A bank may wish to predict LGD for defaulted loans at the level of the account for several reasons. First is to compute the amount of Regulatory and Economic capital it may wish to hold. Second is to allow a bank to segment its portfolios of accounts to apply different application scoring models to new borrowers or behavioural scoring models for existing customers.

While significant attention has been paid to the PD modelling, much less efforts have been made for LGD modelling at the level of the account. In recent years there have been an increasing number of studies dedicated to the subject of LGD estimation and forecast, but they are still relatively limited. During the credit crisis from 2007 to 2009 the default rates of corporate bonds and bank loans were observed to soar, accompanied by an increase in LGD. Altman et al (2005) examined the correlation between PD and LGD, and found that this correlation was likely to cause an increase in the pro-cyclicality effects of the amount of capital required under Basel II. This emphasized the pivotal role of LGD in credit risk modelling especially during the economic downturn. Failing to model LGD accurately may lead to an inaccurate estimate of portfolio losses under extreme risky conditions and trigger unexpected crisis because the bank may fail to cover the extreme losses and the contagion of default will spread across the financial market. Academic researchers and practitioners have recognized that an accurate estimate of LGD plays a key role in lending,

investing, trading or the pricing of loans, bonds and other credit derivatives. Gupton and Stein (2002) showed that errors in estimating LGD can be as damaging as an error in estimating the expected default frequency, and they argued that the precision of both regulatory and economic capital allocation can be improved with more accurate estimates of LGD. Compared with PD modelling, relevant studies have shown that LGD modelling has posed new challenges as follows:

• ***Lack of data***: Banks and other financial institutions often do not have sufficient data to develop and validate models. The new Basel Accord requires data used for developing internal LGD models should cover at least a complete economic cycle. In the case of corporate and sovereign loans it is recommended to have an observation period of seven years, whilst five years of data is considered to be adequate for retail loans. There has been relatively little published research on methods to predict LGD. This is partly because banks are reluctant to share their data with academia for the issue of confidentiality, and thus most research is based on the commercial databases related to corporate bonds. Studies related to modelling LGD of bank loans are limited resulting in the lack of empirical evidence to select proper determinants for estimation.

• ***Complexity of LGD distributions***: Previous research has shown that LGD distributions tend to present various types including bimodal, a U-shape, a reverse U-shape and an L-shape, etc. The irregular LGD distribution sometimes makes traditional statistical regression models ineffective to capture the characteristics and thus the model fit is relatively weak. Although Gupton and Stein (2002) proposed to fit the LGD model using a beta transformation in their internal model LossCalc, the latest version of this model in Dwyer and Korablev (2009) dropped this idea because the beta transformation does not improve the model fit significantly compared with an ordinary linear regression.

• ***Lack of methodologies***: Altman et al (2006) have given a detailed review of the methodologies applied to estimate credit risk. But most of them have only been used to estimate PD rather than LGD. Due to the limitation of data availability, a large amounts of research focuses on corporate bonds where asset pricing models for PD modelling can be adapted to LGD modelling. However, asset pricing models are not applicable for bank loans because generally bank loans are non-trading products without market values. Researchers are motivated to find other techniques beyond traditional statistical models. Apart from the more advanced statistical regression

models, machine learning techniques have also been investigated which are already widely applied to credit scoring and credit rating, although the relevant research is still very limited.

1.3. Research aims and questions

This thesis focuses on the LGD models for both corporate bonds and retail loans. It aims to develop and apply innovative methodologies to estimate and forecast LGD more accurately. In practice LGD models are built based on a portfolio of defaulted debts and applied on the live portfolio. However this research only considers the predictive models built on a dataset of closed defaulted debts and investigates their predictive power rather than monitor the model performance on the live portfolio in light of the given data. The research aims can be characterized into three research objectives as follows.

Question I: *How powerful are machine learning techniques as a way to model the LGD of corporate bonds*?

Machine learning techniques such as neural networks and support vector machines (SVM) have been successfully applied to credit classification and credit rating because of their outstanding discriminatory power but are still rarely used for LGD modelling. Machine learning techniques are well known for their flexibility and power to handle non-linear models. Theoretically SVM techniques could be a good alternative to established methods such as fractional logistic regression to model LGD for corporate bonds considering the irregular distributions of LGD. SVM techniques can not only be applied to predicting LGD directly, but can also be adapted to accounting for the unobservable seniority specific heterogeneity to improve predictive accuracy. We are interested in how much improvement the SVM techniques can make to predictive LGD of corporate bonds.

Question II: *Can we increase predictive accuracy by accounting for unobserved heterogeneity in modelling the LGD of corporate bonds*?

Unobservable heterogeneity of corporate bonds has never been studied in the context of LGD modelling. We consider heterogeneity to be embedded and reflected in the ultimate LGD, where an instrument of a higher seniority is expected to have a higher recovery rate. Similarly an instrument issued by a company with a relatively healthy financial condition should have a lower LGD. Related literature has explored various explanatory determinants including instrument characteristics and firm

accounting information. However, it is believed that the observable determinants are unable to explain the variations of LGD adequately given the low model fit reported in literature, and the incorporation of heterogeneity is expected to capture the embedded characteristics that can not be presented explicitly. The factor models which have been widely applied in PD modelling are a suitable instrument to investigate the impacts of unobservable heterogeneity on predictive accuracy for LGD modelling by incorporating both observable covariates and latent factors into a regression model.

Question III: *How powerful are machine learning techniques as a way to model the LGD of retail loans*?

Similar to Question I, we are interested in applying machine learning techniques to modelling the LGD of retail loans. Literature related to this topic is rather limited. For example, Loterman et al (2011) have studied six datasets of bank loans comparing a group of algorithms and found that SVM and neural network methods outperformed the other statistical models. Different from corporate bonds where the portfolio can be segmented based on seniority and firm characteristics to analyze the effects of unobservable heterogeneity, retail loans can hardly be modelled using the same methodologies as corporate bonds. It is also meaningful to explore how to improve the machine learning algorithms to be suitable for LGD modelling leading to our last research question.

Question IV: *Can we develop a new two-stage algorithm to predict LGD for retail loans which is more accurate than established methods in the literature.*

Modelling LGD of bank loans, especially credit cards has become a challenging topic, where a two-stage model has been proposed to handle the large proportion of cases concentrating at boundaries 0 and 1. The two-stage model proposed in literature combines a classification model to discriminate between the extreme cases at boundaries from the others and a regression method to estimate the cases in (0, 1). But not all empirical evidence finds that the two-stage model is significantly better than simple single-stage models (Bellotti and Crook, 2012). We assume that the key to improving the performance of a two-stage model to predict LGD is to choose the proper methodology for both stages, and often this differs from the commonly used methods such as logistic regression and OLS regression. To improve the classification accuracy of the first stage machine learning techniques will be applied which have been found to be more effective than the established statistical models. A collection of

regression models will also be compared to find out the best combination of classification and regression method for two-stage models.

1.4. Research findings and contributions

We have made several original contributions to the literature in this research as illustrated in three major studies. In our first substantive study we focus on the applications of support vector regression models to LGD prediction. SVR models are applied to predicting recovery rates of corporate bonds for the first time. To study the effects of unobservable heterogeneity of bond seniorities on recovery rates, we improve the original SVR model as described in Loterman et al (2011) by generalizing the fixed intercepts to different intercepts to adapt to different debt seniorities. Another improvement is to propose a semi-parametric SVR model where dummy variables for bond seniorities enter into the model linearly without applying kernel functions. The advantage of this setting is that the heterogeneity can be treated as fixed effects that influence the recovery rates linearly. Our findings show that the SVR models are substantially better than traditional regression models in terms of out-of-sample predictive accuracy, while the robustness of SVR models is as good as that of statistical models. Specifically, the proposed improved versions of SVR models significantly outperform the other methods at the aggregated level, and the original SVR model gives better performance compared with other statistical techniques where the improved versions are not applicable at the segmented level. Two LGD transformation methods including a logit and a beta transformation have also been explored but our empirical evidence shows that neither is able to improve the predictive accuracy of LGD compared with the original setting.

The proposed SVR models have never been presented in previous studies, and our empirical results find that new proposed SVR models can make improvements compared with the original SVR model setting of Suykens et al (1999, 2002). Although previous studies have shown that machine learning techniques are a competitive alternative to statistical regression models for LGD predictions, this study conducts a more comprehensive discussion related to SVR models and complements the literature by incorporating the unobservable heterogeneity into SVR models which has never been investigated in literature. We have shown that SVR techniques are promising in LGD modelling for corporate bonds and the empirical results support our improved SVR models. However, this study only considers the heterogeneity at

seniority level. Our next contribution is to explore the effects of heterogeneity at multiple levels.

In the second study we empirically examine the impacts of unobservable heterogeneity of corporate bonds on recovery rates modelling by making use of single factor models. Traditionally the latent systematic risk factor of single factor models is specified to be a time-varying random variable denoting unobservable economic trends (Frye, 2000a and 2000b). We apply the latent factor to multiple levels and find that the obligor-varying factor model presents a high model fit. This implies that by accounting for the firm level heterogeneity most variations of recovery rates can be explained. Furthermore when the observable firm specific characteristics are excluded from the regression model, we find that the model fit of this restricted model barely changes compared with a full model. The intra-class correlations of single factor models are also examined and we show that firm specific intra-class correlation is much higher than the other specifications when random effect is applied at seniority or time level, suggesting that the common unobservable firm characteristics shared by the instruments issued by the same obligor can be incorporated largely by the incorporation of an obligor-varying latent factor. We have also examined other distributional assumptions apart from the normal distribution for both fixed and random effect models. Among fixed effect regression models we find that fractional response regression model outperforms linear regression slightly, and that the inflated beta regression gives the worst performance. Similarly the inflated beta factor models are outperformed by other non-linear specifications significantly under the single factor modelling framework including the log-normal and logit-normal distributions. But none of those non-linear specifications shows better model fit than the linear single factor model. The impact of firm specific heterogeneity on portfolio loss distributions is also investigated at both aggregated and segmented levels. When the obligor-varying single factor is employed to be the AIRB approach for modelling LGD, we find that, at the aggregated level, methods used in FIRB may potentially underestimate the unexpected losses according to the calculated Value-at-Risk and Expected Shortfall compared with the AIRB approach. At a segmented level we find that the loss distributions generated by the AIRB approach are more right skewed for senior secured and unsecured bonds than that from the FIRB approach, but both of them present very a similar performance for subordinated bonds.

Literature related to analyzing LGD of corporate bonds has devoted to

discovering and examining the importance of observable determinants. This the first study that examines the effects of unobservable heterogeneity on modelling LGD for corporate bonds by employing the single factor models and applying the latent factors to multiple levels. Our study fills the gap that the firm specific heterogeneity is found to be especially crucial for corporate bonds when comparing with other models. This study also shows that a linear single factor model is most suitable for modelling LGD of corporate bonds. Implications to credit risk management have also been presented. The empirical evidence also suggests that LGD models of corporate bonds should be reviewed to calculate regulatory capital more properly, and financial institutions should be recommended to develop their internal LGD models. In summary Questions I and II developed in Section 1.3 are studied in the first two substantive chapters with respect to the effects of unobservable heterogeneity on modelling LGD of corporate bonds using both machine learning techniques and factor models.

The last substantive chapter pays attention to modelling LGD of retail credit cards to answer Questions III and IV. We propose to incorporate SVM techniques into the two-stage modelling framework for LGD modelling, and the SVR regression models are also applied to predicting LGD directly similar to the first study. Two-stage modelling is designed to deal with the large numbers of cases with recovery rates of 0 and 1. However, literature shows that two-stage modelling tends to give a close or less powerful predictive performances compared with established LGD models such as OLS and fractional response regression although it is convenient to implement (Wooldridge and Papke, 1996). To examine Question IV the SVM techniques are applied to the classification problem at the first stage to better discriminate the cases at the boundaries from the remaining ones. We do this because SVM models have been proved to be remarkably powerful in classification problems in literature. To investigate the performances of two-stage models we compare two classification methods at the first stage including a logistic regression and an SVM technique, and four regression methods at the second stage including OLS, fractional response regression, beta regression and SVR. According to our empirical results the two-stage models with an SVM technique at the first stage significantly outperform those with a logistic regression. The effects of choice of regression methods are examined by modelling in (0, 1) and [0, 1] separately. We find that the choice of regression methods does not influence the predictive accuracy as much as expected, except that beta regression is significantly less competitive than the other methods. Further tests

to compare SVM and logistic regression confirms that SVM is superior at the classification stage, and suggests that the predictive performance of two-stage models depends on the classification models more than on regression methods. The advantage of applying SVM techniques to two-stage models is that interpretability of the results can be preserved as much as possible while improving predictive accuracy. This idea has never been proposed in previous studies related to the modelling LGD of retail credit cards.

Our third study develops a new algorithm to predict LGD for bank retail loans. The two-stage models proposed in literature are expected to be a powerful alternative to estimate LGD. However, they are not shown to be more competitive than other simple regression models according to empirical results. Following our first study this study contributes to literature by proposing a hybrid two-stage model where SVM techniques are incorporated. The SVM techniques have been explored to predict LGD for retail loans directly in literature, but it has never been investigated in the two-stage models. This is the first study that applies SVM techniques to the classification stage in the two-stage framework and reveals that the performances of two-stage models are mainly dependent on the choice of classification algorithm. We have shown that the two-stage models are more effective when they are equipped with proper classification and regression algorithms. The empirical evidence which finds that the two-stage models with an SVM classifier significantly outperform those with a logistic regression classifier is also first reported in our study.

1.5. Structure of the thesis

This chapter has given an overview of the whole thesis introducing the research background and the contributions to the literature. Chapter 2 discusses the literature related to LGD modelling. It starts by explaining the background of the Basel Accord and relevant definitions, and then discusses the significant determinants of LGD with respect to both corporate bonds and bank loans that have been discovered in the literature. Finally it presents a comprehensive review of the methodologies that have been applied to modelling LGD from parametric regression models to non-parametric machine learning techniques, summarized by a comparative analysis to demonstrate the advantages and disadvantages of each category of methodologies.

Chapter 3 presents the data that will be used to model recovery rates of corporate bonds including details of the defaulted instruments of US companies from 1986 to

2012. It first explains how the sample is constructed from databases, and then the economic significance of all variables is introduced with summarized statistics presented as well. The empirical studies in Chapter 4 and 5 are both based on the data described in Chapter 3.

Chapter 4 explores the Question I in Section 1.3 by developing two improved versions of least squared support vector regression models to account for the seniority heterogeneity of corporate bonds in LGD modelling. To find out whether the new SVR techniques are effective or not all models are benchmarked at aggregated and segmented samples respectively. The results show that by accounting for the heterogeneity mentioned the new SVR models outperform the original setting of LS-SVR (Suykens et al, 1999 and 2002) and give much better out-of-sample predictions than other statistical regression methods. The results also show that the models tend to show better performances for bonds of higher seniorities.

To answer Question II Chapter 5 moves forward from Chapter 4 to investigate the effects of heterogeneity on recovery rates modelling at multiple levels including obligor, seniority and time levels. The major finding of this chapter is that by accounting for the firm specific heterogeneity, the single factor model shows an impressive model fit, and the firm specific intra-class correlation of instruments is found to be much higher than that at other levels. This chapter also discusses the implications of the results for risk management by comparing the simulated portfolio loss distributions of several methodologies, and suggests that it is beneficial for financial institutions to develop internal LGD models to better understand the portfolio risk of corporate bonds.

Chapter 6 tries to answer Question III and IV by proposing a hybrid two-stage modelling framework to predict recovery rates of credit cards, as introduced in Section 1.3. The data used in Chapter 6 is originally from a UK credit card lender covering recovery information of almost 300,000 defaulted customers. This sample has never been studied in literature. Out-of-sample predictions are compared across a collection of algorithms, where the classification performances of two-stage models are also examined. The major contribution of this chapter is that it illustrates how to combine SVM techniques with the statistical regression methods to improve the predictive accuracy of recovery rates. The empirical evidence finds that the hybrid two-stage models combining SVM and statistical models outperform the other kinds of two-stage and single-stage models significantly. Finally Chapter 7 draws some

conclusions from the research, notes some limitations and proposes ideas for future research.

Chapter 2
Literature Review

2.1 Introduction

Literature related to LGD/RR modelling can be generally categorized into two groups of products: corporate bonds and retail loans. The methodologies of modelling bond LGD models are more related to bond pricing models including both option pricing and jump diffusion models. For the bank loans for retail and company customers, econometric statistical models are commonly applied including linear and generalized linear regression models. This chapter aims to review the literature related to LGD for both corporate bonds and retail loans including unsecured and secured loans, and to discuss the commonly used methodologies of LGD with respect to different products. It is shown that Merton's structural models and factor models tend to give better predictive performances and parametric statistical regression are more advantageous to identify the important determinants of LGD. Survival regression models and other non-parametric machine learning techniques are regarded to be competitive alternatives. We also find PD/LGD correlation plays a vital role in modelling expected losses and demonstrate the relevant techniques. The transformation methods applied to LGD prior to modelling seems to be unnecessary, as empirical evidence suggests. This chapter is organized as follows. Section 2.2 gives an overview of Basel Accord and relevant concepts in credit risk modelling and capital requirements, and then the determinants of LGD for both retail loans and corporate bonds are discussed in Section 2.3. Section 2.4 presents the methodologies by categories in detail.

2.2 Overview

2.2.1 Basel accord and capital requirements

A Revised Framework on International Convergence of Capital Measurement and Capital Standards (Basel II) issued by Basel Committee in 2004 has been serving as the basis for regulators and global financial institutions to evaluate and manage financial risk in banks. Basel II is composed of three pillars: minimum capital requirements, a supervisory review process and market discipline. Under the first pillar financial institutions are required to calculate the total minimum capital requirements for credit, market and operational risk. Financial institutions can adopt

the Standardized Approach in which the capital requirements are specified by the Basel Committee with respect to each product. Basel II encourages financial institutions to move from the Standardized Approach to the Internal Rating Based (IRB) Approach to develop their own internal models to estimate the key drivers including Probability of Default (PD), Loss Given Default (LGD), Exposure at Default (EAD) and Maturity (M). Under the IRB Approach framework there are two broad approaches suggested by the Basel Committee: a foundation and an advanced termed as FIRB and AIRB respectively. Under FIRB Approach banks only need to provided estimates of PD and to apply the supervisory estimates for the other risk parameters while the other parameters including LGD and EAD are fixed values with respect to different products. In contrast, under the AIRB Approach banks provide their own estimates of PD, LGD, EAD and M.

There is no unique definition of default provided in Basel II, which requires banks to adopt a reference definition for their internal use. The Bank of International Settlement reference definition of default states that a default is considered to have occurred with regard to a particular obligor when one or more the following events has taken place when an obligor fails to pay its debt obligations including principal, interest or fees in full (Basel Committee, 2004). Schuermann (2004) has listed several scenarios that trigger a default according to Basel II. He indicates that for many instances of defaults under the definition no loss may result and thus it becomes arguable whether or not to include the relevant record into the bank's loss database. PD defines the probability that a default may happen and LGD measures the loss in the event of default. EAD denotes the outstanding dollar amount of an obligor or an account at default. It is common to define a credit conversion factor (CCF) which is the proportion of the facility's undrawn dollar amount at current time to be drawn down at default time. CCF is similar to LGD which are both bounded between 0 and 1. In practice many banks model the CCF rather than modelling the EAD dollar amount directly, since EAD varies significantly from a low amount to an extremely high amount.

The standardized approach is to be used by the banks that are not sufficiently sophisticated in the eyes of the regulators to use the IRB approaches. The rules for determining risk weights are specified with respect to the types of exposures with different credit ratings. For example, the risk weight for a sovereign exposure ranges from 0% to 150%, and the risk weight for a bank or corporate exposure ranges from

20% to 150%. The risk weight for retail lending is 75%, and a residential mortgage gives a risk weight of 35%. More details can be found in the First Pillar documentation of Basel II (2004). To understand IRB approaches we consider the losses for a given portfolio. There are composed of expected losses and unexpected losses, denoted as EL and UL respectively. We define the worst case default probability for the next year at 99.9% confidence level as conditional PD (CPD), which means the default rate will not be exceeded with a probability of 99.9%. Basel II aims to cover the UL which is given as the difference between the worst case loss rates (WCDR) and expected losses (EL) such that

$$EL = PD \cdot LGD$$
$$UL = WCDR - EL$$

where $WCDR = CPD \cdot LGD$. The reasoning of the IRB approaches is that banks are required to estimate EL and UL separately, where EL can be covered through provisioning by the bank, and UL is covered under the assumption of an extreme adverse scenario. Then the capital requirement formula is given such that

$$K = EAD \cdot UL \cdot MA$$

where MA denotes the maturity adjustment for the sovereign, bank and corporate exposures. Maturity adjustment is not needed for retail exposures. We can further calculate the risk weighted assets (RWA) such that

$$RWA = 12.5K = 12.5 \cdot EAD \cdot UL \cdot MA\,.$$

2.2.2 LGD measurements

LGD measures the ratio of losses to exposure at default. Schuermann (2004) suggested that there are three types of losses incurred during the recovery process:

- The loss of principal;
- The carrying cost of non-performing loans;
- Workout expenses.

There are three ways of measuring LGD for an instrument, which are Market LGD, Workout LGD and Implied Market LGD. Market LGD is given as the observed market price of the defaulted bonds and loans immediately after the actual default event, because the actual prices are based on the face value of 100 (par=100) and can be easily transformed into a recovery percentage. Market LGD reflects the market investors' expectation on the recovery which is suitably discounted including both discounted principal and missed interest payments. Workout LGD is calculated based on the cash flows during workout/collection process properly discounted. Schuermann

(2004) pointed out that a bank should be cautious when choosing a proper discount rate such as the bank's hurdle rate, and suggests that the coupon rate or risk-free rate could be inappropriate. Implied Market LGD is derived from non-default bonds using a bond pricing model. The credit spread that works as an indicator of the risk premium reflects the expected loss, or the product of PD and LGD. It has been a new topic in credit risk modelling on how to separate LGD from credit spread, and recent studies show that the derived recovery rates lie systematically below the actual recovery rates.

2.3 LGD Determinants

The determinants of LGD can be generally divided into four categories according to Resti and Sironi (2007): exposure (or debt) characteristics, borrower characteristics, bank's internal factors and other external factors. Previous studies of LGD/RR modelling only consider the exposure, borrower and external factors because a bank's internal factors are usually unavailable. The following sections review the determinants that are frequently examined in literature based on the credit product types: retail loans and corporate bonds.

2.3.1 Determinants of LGD for bank loans

For the studies related to LGD/RR modelling on bank loans, it is common for authors not to disclose the full information due to confidentiality requirements from the data providers. Typically bank loan recovery rates are modelled in terms of loan characteristics, borrower application variables and macroeconomic variables, and literature has found that loan characteristics influence the recovery rate more significantly than the other factors. For example, Grunert and Weber (2009) investigated the recovery rates of corporate loans in Germany. They considered more than 100 companies and included loan characteristics, company financial ratios and economic factors. They confirmed their hypotheses that a high quota of collateral leads to a higher recovery rate, and the creditworthiness has a positive correlation with the recovery rate. Another interesting finding was that the intensity of the client relationship was found to be positively correlated with recovery rate indicating that an intense relationship was able to improve the access to collateral and to increase the influences on the workout process of the company. They also found that all the macroeconomic factors were significant implying that banks may intensify their efforts to recover more debts during the financial distress periods. Considering the

small sample used in this study, they proposed that further analysis based on a larger sample was required in order to obtain further insight into the recovery rates. Similarly Khieu et al (2012) found that loan characteristics were more significant than borrower characteristics by analyzing the recovery rates of bank loans. They demonstrated that the loan contract features were strongly correlated with ultimate recovery rates. The macroeconomic variables included in the models were also shown to be significant. Zhang and Thomas (2009) modelled recovery rates of credit cards from a UK retail bank and found that the most significant determinant was the ratio between exposure at default to the total loan which was strongly negatively related to recovery rate. The other significant variables included some applicant characteristics such as employment status and residential status. However, they did not incorporate any macroeconomic variables but simply included year dummy variables which were shown to be significant.

For the recovery rates modelling of mortgage loans, it is more straightforward to observe that the characteristics related to the underlying asset purchased or the collateral are more important. For example, Qi and Yang (2009) studied LGD on a dataset of residential mortgage loans and found that the current loan-to-value (CLTV) ratio was the single most important determinant which affects LGD positively. They also found that the loss severity in distressed housing markets was significantly higher than that in the normal markets. Leow and Mues (2011) developed a probability of repossession model to estimate the probability of the collateral being repossessed, and they showed that the model with variables including CLTV, the type of security and the status on previous default was significantly better than a model with only CLTV variable. This probability of repossession model presented that both CLTV and previous default status were positively related to the repossession probability, and the type of security had a negative effect on LGD.

Effects of macroeconomic have also been investigated for both secured and unsecured loans. Caselli et al (2008) studied the relationship between macroeconomic conditions and LGD of bank loans in Italy including both household and corporate loans for SMEs. They conducted a comprehensive range of economic factors and developed separated models for households and SMEs with different set of macroeconomic variables. The LGD of household loans were found to be strongly correlated with default rate, unemployment rate and household consumption with the logarithmic form, showing that the factors influencing the household incomes affected

the recovery rate significantly. For the SMEs the GDP growth rate had an influential role for the recovery rate, indicating that SMEs asset value were more sensitive to the economic conditions. Bellotti and Crook (2012) and Leow et al (2013) also examined the effects of incorporating macroeconomic variables in retail loans recovery rates modelling. Bellotti and Crook (2012) found that the inclusion of macroeconomic variables did improve the forecasts across test quarters although the improvement in MSE was modest. However, the empirical results showed that such improvement at the portfolio level was more significant than at the segmented level with respect to separate products, and the inclusion of interaction terms between application and macroeconomic variables was not able to improve the predictive performances in general. Leow et al (2013) investigated the incorporation of macroeconomic variable in two data sets: a residential mortgage loans data set and a personal unsecured loans data set. They found that the predictive performances of mortgage loans LGD can be improved by the inclusion of macroeconomic variables where interest rate was the most beneficial one. But for the unsecured personal loans LGD little improvement was shown with only net lending growth being significant statistically, implying that the personal loans LGD was less sensitive to the economy than the mortgage loans LGD.

2.3.2 Determinants of LGD for corporate bonds

For corporate bonds most research was based on data from public resources such as Moody's Ultimate Recovery Database or Altman-NYU Salomon Centre Corporate Bond Default Master Database. The determinants of corporate bonds recovery rates can be generally grouped into four categories including instrument characteristics, firm specific effects, industry-wide factors and economic factors. Unlike recovery rates for bank loans, the recovery rates of corporate bonds are significantly correlated with all of these factors according to recent empirical studies. For example, Moody's special comment (2004) has investigated the determinants of recovery rates for bank loans and corporate bonds. It found that seniority and security played the most crucial role in explaining recovery rates. There were other significant factors including firm specific effects such as leverage, and industry and macroeconomic specific factors such as industry market-to-book and the health of the economy and of the stock market, which were both strongly correlated with recovery rates.

Acharya et al (2007) further examined industry-wide effects on recovery rates and found that when the default firm's industry is in distress, its instrument recovery

rate was expected to be between 10 to 15 percent lower than the case when the industry is healthy. The authors also found that the effects of industry return were always non-linear, suggesting that it was the distress effect that influenced the recovery rate. The industry-specific effects were also found to be economically significant and robust to the inclusion of controls for contract-specific effects and firm-specific effects. The industry-distress effect was presented to be robust to macroeconomic and bond-market conditions at the time of default by including aggregated default rate and aggregated supply of defaulted bonds in the regression model.

Jacobs and Karagozoglu (2011) demonstrated the statistical and economic significance of debt and equity market determinants of LGD such as the bond price at default, and they suggested that it is potentially beneficial to incorporate the market signals in order to improve forecasts of recovery rates. They also examined a wide range of determinants from instrument features, firm specific effects as well as industry and economic factors. Regarding macroeconomic and industry effects, they found both aggregated default rate and S&P 500 returns both significantly influenced LGD, and the industry profit margin and dummy variables for technology or utility industries also showed significant impacts that was consistent with Acharya et al (2007). In terms of the firm effects, they demonstrated the significance of firm size, leverage, asset tangibility, market valuation, cash flow and liquidity which all influence LGD negatively. For the debt contractual features, the collateral rank and relative seniority of creditor were also all found to be negatively correlated with LGD. The authors also showed that capital structure variables such as the number of creditor classes and the proportions of secured and bank debt were all inversely related to LGD.

Qi and Zhao (2011) examined the determinants of instrument recovery rate and found that firm specific variables were more critical determinants for the creditor's recovery than the industry and macroeconomic characteristics. They argued that the existing measures of debt seniority was not able to fully capture a firm's debt structure, and they proposed a new variable which incorporated the percentage of debt both more senior than, and at the same rank, to the instrument under consideration. They showed that the inclusion of this new debt structure variable increased the explanatory power of the model significantly as it turned out to be the most crucial determinant of recovery rate. In addition, the firm conditions measured by 12 months trailing stock

returns was found to be the second most important determinant but it was not available for private firms. They suggested that the PD/LGD correlation was more likely due to idiosyncratic risk than to systematic risk given the importance of firm characteristics shown in the empirical study.

Jankowitsch et al (2014) conducted a detailed analysis of US bond recovery rates by using a comprehensive set of determinants and found that all types of variables contributed to the explanatory power. They also examined the effects of some new features. For instrument characteristics they found that bonds that can be delivered into a credit default swap (CDS) contract influenced recovery rate positively. Bond covenants were also significant factors, indicating that the restrictions on investment and financing policy were an effective tool to increase the creditor's recovery rates. They found that illiquid bonds with higher transaction costs tended to recover less after default, showing a clear correlation between the defined liquidity measures for bonds and their recovery rates.

2.4 Methodologies

2.4.1 Merton's structural models

As introduced in 2.1.2, equity and bond pricing models have been applied to deriving the implied LGD from the observed equity prices. Studies related to modelling implied LGD are not as common as other topics. The implied LGD models are based on Merton's asset process model (1974). Jokivuolle and Peura (2000) presented a simple extension of modelling corporate debt and collateral values based on Merton's model. They proposed to model asset value and collateral value separately with correlated Wiener processes defined by a constant correlation parameter such that

$$dV_t = \mu_v V_t dt + \sigma_v V_t dW_t^V,\ 0 \le t \le T,$$

The collateral value C follows another stochastic process which is defined as

$$dC_t = \mu_c C_t dt + \sigma_c C_t dW_t^C,$$

where V_t and C_t denote asset value and collateral value processes of a given firm respectively. The parameters μ_v, σ_v, μ_c and σ_c are defined to be the mean and volatility of both processes. Two Wiener processes W_t^V and W_t^C are correlated with the parameter ρ. They investigated the collateral haircut of bank loans and suggested that for secured loans the recovery rate estimate should be a decreasing function of the

collateral volatility and the PD/RR correlation. However no empirical evidence was presented in their work.

Resti et al (2007) proposed to define the recovery rate as the ratio of the firm's asset value to the debt value such that

$$RR = E\left(\frac{V_T}{B} \mid V_T < B\right) = \frac{1}{B}E\left(V_T \mid V_T < B\right),$$

where V_T and B are the asset and debt value, and a closed form expression of expected recovery rate RR is given as

$$RR = E\left(\frac{V_T}{B}\right)\frac{N(-d_1)}{N(-d_2)}.$$

where $d_1 = \frac{\ln(V\ /\ B) + (r + 0.5\sigma_v^2)T}{\sigma_v\sqrt{T}}$ and $d_2 = d_1 - \sigma_v\sqrt{T}$. Seidler and Jakubik (2009) followed the approach in Resti et al (2007) and further extended it by taking dividend payouts and bankruptcy costs into consideration. They obtained a closed form solution of expected LGD (ELGD) in the physical measure such that

$$ELGD = 1 - \varphi\frac{V_T}{B}\exp[(\mu_V - \delta)T]\frac{N(-d_1)}{N(-d_2)},$$

where the bankruptcy costs is $1-\varphi$, and μ_V and δ denote market return and dividend rate respectively. They implemented this model on the data of companies listed on the Prague Stock Exchange to derive implied LGD, and found that the implied LGD was in the range between 20% and 45%. They further analyzed the sensitivity of implied LGD with respect to the company leverage, and found most of the companies had inelastic ELGD related to leverage. In summary the implied LGD modelling techniques have not been extensively studied, and most research focuses on modelling LGD explicitly instead of deriving LGD from equity prices.

2.4.2 Factor models

The single factor model was first proposed by Vasicek (1987) to estimate probability of default and portfolio losses based on Merton's model (1974). Similar to PD modelling, LGD can be also specified as a random variable under the single factor modelling framework. Frye (2000a) first modelled the collateral damage to be a random variable similar to Jokivuolle and Peura (2000). In this model, the collateral value of an obligor j is defined as follows

$$\begin{aligned} CV_i &= \mu(1 + \sigma C_i) \\ C_i &= \rho X + \sqrt{1-\rho^2}u_i \end{aligned},$$

where CV_i denotes the collateral value depending on a systematic risk factor C_i, and μ and σ represent the amount and volatility of collateral value. Specifically the collateral value was assumed to be driven by economic influences similar to an asset value process, which renders LGD depend on the economic state. This feature makes this paper different from pervious credit risk models where the LGD is specified as a determined value and independent with PD. The simulation in Frye's study showed that the conventional models that did not consider collateral damage, may underestimate expected portfolio loss compared with the proposed factor model when the economy slumps. In later work (Frye, 2000b) the same methodology was applied to modelling recovery rates of US corporate bonds. The model was estimated by a two-stage method. At the first stage the values of latent systematic risk factors across years were estimated. Next the unknown parameters relating to collateral value process were estimated given the implied systematic risk factors. Both of these two steps employed a maximum likelihood method, and the empirical results suggested similar implications that in an economic downturn the expected LGD tends to increase together with an increase in default probability.

Pykhtin (2003) modified the above model by allowing recovery rates to be related to the idiosyncratic factor. The model also assumed that the value of collateral followed a log-normal distribution and then a closed form of quantifying the expected LGD and economic capital was derived. In the following the recovery rate of obligor i is denoted as y_i. Pykhtin's model was given such that

$$\begin{aligned} CV_i &= \exp(\mu + \sigma y_i) \\ y_i &= \beta X + \gamma u_i + \sqrt{1-\beta^2-\gamma^2}\eta_i \end{aligned},$$

where η_i is also a standard normal distribution variable and independent of u_i. The simulation results also confirmed the conclusions of previous papers that the expected LGD would increase when PD increases.

In single factor models proposed by Frye (2000a, 2000b) and Pykhtin (2003) both systematic and idiosyncratic risk factors are unobservable, at least not directly observable directly. To increase the explanatory power observable covariates were included in later research such as macroeconomic and individual specific factors formulating mixed effects models. The mixed effects models were first applied in the context of PD modelling. Hamerle et al (2003a) incorporated lagged default rates into the single factor model to estimate the asset correlation parameter of the enterprises and bankruptcies of G7 countries. Hamerle et al (2003b) showed that with the

inclusion of observable economic specific effects the single factor model essentially becomes a non-linear mixed effects Probit model. They benchmarked a collection of models to estimate the PDs and asset correlations and found that by using the actual observable information the variance of estimates can be substantially reduced which is particularly useful to calculate portfolio Value-at-Risk.

Meanwhile, other research shed light on modelling downturn LGD as required in Basel II. In Basel II banks are required to use downturn LGD estimates in regulatory capital calculations considering the fact that PD/LGD correlation is not captured. Miu and Ozdemir (2006) showed that the PD/LGD correlation can be captured by including a certain degree of conservatism in cyclical LGD in a stylized modelling framework which can be estimated jointly. By using historical default data of a loan portfolio they estimated the PD/LGD correlation and found it substantially significant than LGD correlation. Finally they evaluated how much the expected LGD needs to be increased in order to compensate for the lack of consideration of PD/LGD correlation in the Basel capital formula. They found that given a moderate asset correlation the expected LGD needs to be increased by about 35% to 41% to achieve the correct regulatory capital. Li (2009) established a new modelling framework based on stochastic spot recovery for Gaussian copula. He discussed the large homogeneous pool limit and derived analytical formula for VaR and Expected Shortfall in the case of a single systematic factor. No empirical studies were conducted in this work although numerical examples were presented to compare the downturn LGD of a collection of methodologies.

Factor models based on Vasicek's single factor framework and Gaussian copula are widely applied to credit portfolio losses estimation and CDO pricing. Under the FIRB approach banks only need to estimate PD internally and rely on external agencies to determine LGD values. In this way single factor models simply treat LGD as a determined value and PD/LGD correlation is consequently ignored which may lead to underestimation of expected and unexpected losses. But it is highly inappropriate to recognize PD or LGD as an exogenous variable and plug it into a statistical regression model, because these two parameters are endogenously related. Therefore, an important feature of incorporating a separate LGD model into factor models has been a new trend in credit risk modelling where the dependence between PD and LGD can be explicitly incorporated and calibrated. Empirical studies have demonstrated strong evidence of negative correlation between PD and RR although

industry models including CreditRisk+ or Creditmetrics treat RR as a deterministic value or a stochastic variable but independent from PD. Based on Vasicek's single factor model, several approaches have been proposed to model this correlation explicitly, including Frye (2000a and 2000b), Pykhtin (2003), Dullman and Trapp (2004), and Rosch and Scheule (2005). Dullman and Trapp (2004) proposed to model PD and RR jointly where PD and RR are driven by a common systematic risk factor such that

$$\begin{aligned} A_i &= \sqrt{\rho} X_t + \sqrt{1-\rho} u_i \\ y_i &= \mu + \sigma\sqrt{\omega} X_t + \sigma\sqrt{1-\omega}\eta_i \end{aligned},$$

where A_i and y_i denote the asset return and recovery rate of obligor i. X_t is the time-varying systematic risk factor and ρ is the asset correlation parameter, and u_{it} and η_{it} are independent idiosyncratic risk factors both following standard normal distributions. This joint modelling specification shares common characteristics with Vasicek's model where the unconditional and conditional correlation between PD and RR can be derived

$$\begin{aligned} Cov(A_{it}, R_{it}) &= \sigma\sqrt{\omega\rho} \\ Cov(A_{it}, R_{it} \mid X_t) &= 0 \end{aligned},$$

Notice that the conditional correlation equals to zero which may underestimate the correlation between PD and LGD questioned by Peura and Jokivuolle (2005), who argue that this assumption was too restrictive even economically counterintuitive. Peura and Jokivuolle (2005) indicated that the correlation between LGD and PD is correlated with the sensitivities to the systematic factors, which may further underestimate their correlation.

Unlike Dullman and Trapp (2004) where no observable covariates were included, Rosch and Scheule (2005) incorporated macroeconomic variables into a multi-factor framework to model the aggregated annual default rates and recovery rates of corporate bonds jointly. The empirical findings suggested that default and recovery risk were negatively correlated and the incorporation of this correlation increased the simulated economic capital, indicating that the portfolio loss might be underestimated if banks failed to consider such correlation. The inclusion of macroeconomic variables tended to make the systematic risk factors less important and to decrease the influences of recovery rates correlation between different instruments. However, the firm and debt characteristics were not included in this paper because of the data limitations. A similar approach was adopted by Hamerle et al (2006) where the

observable explanatory variables such as macroeconomic conditions and individual specific characteristics were all incorporated to estimate the instrument level LGD. They found that the incorporation of macroeconomic variables reduced the variance of latent factors as well as the uncertainty of the predicted LGD. This finding was consistent with Rosch and Scheule (2005), but estimate of the variance of LGD was still rather high after including the macroeconomic variables and other factors implying that the variations of LGD should be further investigated through more potential predictors.

One pitfall in Dullman and Trapp (2004) and Rosch and Scheule (2005) is that the PD and LGD conditional correlation is zero conditional on a realization of the systematic risk factor, which may underestimate the PD and LGD correlation. Pykhtin (2003) proposed a two-factor model where the collateral value was modelled to be a combination of a systematic risk factor and two idiosyncratic risk factors, one of which also entered into the PD model. In this way the PD and LGD correlation comes from both systematic and idiosyncratic risk factors, and thus the conditional correlation between PD and LGD is not zero. In contrast Hillebrand (2005) developed a new two-factor model for LGD to estimate portfolio losses. Different from the two-factor model in Pykhtin (2003), the two-factor model in Hillebrand (2005) consists of two systematic risk factors that incorporates the dependence of PD and LGD and integrates the possibility of economic interpretation. Hillebrand showed that this two-factor framework was easy to implement and calibrate and it provides an excellent fit of corporate bond data.

Chava et al (2011) investigated the portfolio loss distribution by modelling the default and recovery risk separately. Based on the time-varying dynamic frailty model of Duffie et al (2009), they proposed a regime-dependent multiplicative frailty model to estimate default probability which placed the emphasis on the industry specific unobservable heterogeneity, and the recovery rates were fitted by a fractional response regression model. They showed that by accounting for the industry level heterogeneity the dynamic frailty model improved the out-of-sample predictions significantly which subsequently impacted the portfolio loss predictions. They found that the predicted default probabilities and recovery rates were negatively correlated but the magnitude of such correlation depended on the credit cycle and varies with industries and seniorities. However, Chava et al (2011) saw their main contribution in default risk modelling and suggested that the choice of recovery rate model had a

marginal impact on portfolio loss predictions. Bruche and Aguado (2010) proposed a new econometric model where default and recovery risk was driven by an unobservable factor denoting credit cycle which followed a regime-switching Markov chain process. The default risk was estimated by a discrete hazard model proposed in Shumway (2001) and the recovery rates by a beta regression model. The credit cycle was introduced as a frailty variable similar to Duffie et al (2009), and strong evidence was shown for the presence of the common latent factors led to more accurate predictions of both firm level default probabilities and portfolio losses. It also showed that the out-of-sample recovery rates predictions outperformed the methodology proposed in Chava et al (2011) in terms of RMSE. However, due to the limitations of their data, firm-specific variables were not included in their specifications. They suggested that further research was required to investigate not only the time-varying systematic risk, but also the default and recovery risk correlation at the firm level.

Frye and Jacobs (2012) developed a LGD model which assumed the LGD rate to be a function of the default rate. They considered three distributional assumptions including the Vasicek's distribution given in Bluhm et al (2003), a beta distribution and a log-normal distribution. Although the three distributions produce approximately the same PD/LGD relationship, they argued that the Vasicek's distribution was the easiest for a practitioner to apply since it has explicit formulas for its cumulative and inverse cumulative distribution functions, and the estimates of Vasicek correlation parameter also already existed within the current credit loss models. The model was developed based on a series of assumptions ended up with an explicit LGD function of three parameters including unconditional PD, expected loss and asset correlation.

Empirical studies related to PD and LGD correlation are largely based on corporate bonds at portfolio level. Altman et al (2005) examined the aggregated default rates and recovery rates of corporate bonds and found that the PD/RR correlation had significant impact on both credit VaR models and the procyclicality of capital requirements. According to simulation results they found that both expected and unexpected losses were highly underestimated if the PD/RR correlation was neglected, which underscored the importance of modelling such correlation especially under AIRB framework. Rosch and Scheule (2008) further investigated the PD/LGD correlation and focused on the link between bond recoveries with credit ratings and subordination levels. They extended the classical Tobit model to an econometric approach that accounted for the asset correlations, and found the regulatory capital

based on a constant recovery assumption might be underestimated by as much as 23%. Bade et al (2011) empirically investigated the default and recovery risk of corporate bonds following the Pykhtin's two-factor model (2003), because Pykhtin (2003) only provided a theoretical framework with a simulation study conducted. The empirical evidence showed that the rating grade and rating shift gave a highly remarkable explanation for default and recovery risk of US bonds.

Literature related to LGD factor models focuses on developing models to capture the influences of economic movements on portfolio losses as well as PD/LGD correlation, instead of aiming to improve the LGD predictive accuracies. In terms of improving LGD or RR predictions more attention has been paid to statistical regression models and other data mining techniques as introduced in the following sections.

2.4.3 Linear and generalized linear regression models

Statistical regression models including linear and generalized linear regression methods are most popular to explore the determinants of LGD/RR and to predict individual LGD/RR for either corporate bonds or retail loans. Given the large volume of literature related to the application of linear and generalized regression models to LGD modelling, this section is organized by the types of credit products including unsecured loans, mortgage loans and corporate bonds.

2.4.3.1 Introduction

Ordinary linear regression and fractional response regression dominate empirical studies related to LGD/RR at individual loan level. OLS is regarded as a benchmarking model and has shown consistently robust model fit with good interpretability. Fractional response regression was first proposed by Papke and Wooldridge (1996) and has been used because it is designed to model fractional response targets defined in the interval (0, 1) which is just suitable in the context of LGD modelling. However, it can be observed that a large number of cases concentrate at the boundaries 0 and 1 in the empirical LGD distributions which makes it inconvenient using these extreme values for fractional response regression model. Beta regression proposed by Ferrari and Neto (2004) has also drawn much attention because the beta distribution is considered to be a better candidate to approximate the LGD bi-modal distributions. However, one potential drawback in beta regression is that the response variable is defined in (0, 1) similarly to the fractional response regression. To overcome the issue of boundary cases, Ospina and Ferrari (2010)

improved beta regression proposing a mixture of distributions combing a continuous beta distribution in (0, 1) and a discrete Bernoulli distribution to capture the probability mass at 0 and 1. Similarly Tobit regression is also considered as an alternative to the OLS as it is designed to treat truncated cases. However, it is quite debatable to apply Tobit regression to modelling LGD because Tobit model assumes the cases are censored at a given point, but LGD is bounded between 0 and 1 for its definition instead of being censored. Another idea to address the issue of boundary points is to apply a two-stage modelling framework, which has been more widely accepted than inflated beta regression and Tobit models given its simplicity.

2.4.3.2 Unsecured loans

Studies on the LGD/RR of unsecured bank loans such as retail credit cards and SME loans have been attracting an increasing interest since more banks are willing to provide internal data for academic research to obtain more insights into the retail loans recovery characteristics. Dermine and Carvalho (2006) conducted a case study on the default loans of a European bank by applying mortality analysis and regression models. They first applied a univariate mortality-based approach to measuring the cumulative recovery rates and showed an average recovery estimate of 71%. Then the fractional response regression model with a complementary log-log link function was employed to identify the influences and significance of determinants. It was shown that the loan size, firm age, collateral characteristics, industry sector and year dummies were the significant determinants for loan recovery rates. Following this Dermine and Carvalho (2008) further investigated how to calculate a fair level of loan-loss provisions at default and after default. They proposed a dynamic provisioning scheme which was estimated on the non-performing loans given by a Portuguese bank by applying similar modelling approaches to the previous study. They found that the unsecured bad and doubtful loans exhibit better recoveries than the secured loans, and they suggested that the decision to issue unsecured loans had accounted for a higher expected recovery rate. By comparing with the Bank of Portugal mandatory provisioning rules it indicated that the provisioning in the long term after default tended to be conservative but in the near term more stringent provisions should be enforced.

Chalupka and Kopecsni (2009) investigated the determinants of LGD of small and medium sized enterprises (SMEs) loans from a Czech commercial bank. They applied three different models including a fractional response regression model, an

inflated beta regression model and an ordinal regression model for LGD grades estimation. The ordinary linear regression was used as the benchmarking model. They found the collateral types and time periods both had strong impacts on LGD. In terms of model fit linear regression performed similarly well compared with fractional response regression with either a logit or a complementary log-log link function. However, the inflated beta regression showed slightly worse results, unexpectedly.

Khieu et al (2012) further examined the determinants of bank loans with a wide range of characteristics with an OLS and a fractional response regression model. The main contribution of this study was that the loan characteristics are shown to be more significant than the borrowers' features prior to default and the industry and economic conditions were correlated for the pre-packaged bankruptcy arrangements. The fractional response regression showed better model fit indicating that loan recoveries varied significantly and nonlinearly with the length of time to emerge. They also investigated the efficiency of the 30-day post-default trading price which was commonly used proxy for recovery rates, and found that such a proxy was a biased and inefficient predictor of ultimate recovery rate although they were highly correlated.

Comparison studies on the performance of regression models are also based on the unsecured loan data. For example, Bastos et al (2010) evaluated both parametric and non-parametric methods to forecast the recovery rates of bank loans from a Portuguese private bank. A fractional response regression and a regression tree were compared in this study to obtain both out-of-sample and out-of-time predictive accuracies. The regression tree gives better results for shorter horizons of 12 and 24 months in terms of out-of-sample cross-validation performances. It also pointed out that the regression tree algorithm requires larger sample sizes compared with parametric regression models since the data must provide the model structure as well as the model estimates, but it gives best performances for out-of-time predictions. This study suggested that regression trees were an interesting alternative to parametric models because they are competitive and interpretable.

Calabrese (2014) applied inflated beta regression to modelling retail loans recovery rates from the Bank of Italy. This study showed the major advantage of inflated beta regression was that it is able to analyze the different influences of the same covariates on the extreme value of 0 or 1 and the recovery rates in the interval (0, 1). Compared with fractional response regression, inflated beta regression showed

consistently better out-of-sample predictive accuracies across different forecasting periods and different sample percentages of the extreme values.

Bellotti and Crook (2012) compared a collection of models including a Tobit regression, a two-stage model, a beta and fractional response regression model for LGD of retail credit cards of a UK retail bank. They investigated the influences of account and macroeconomic variables on forecasting LGD at the account and portfolio levels. The two-stage model proposed in this study is based on a decision tree framework where the extreme values of recovery rates of 0 or 1 are separated from the remaining values in the interval (0, 1) at the first stage, and a regression model is applied to fitting the interval cases at the second stage. However, this two-stage model did not outperform the other regression models. Instead they found that the benchmarking OLS regression model with macroeconomic factors showed the best out-of-sample predictive performances and that the inclusion of macroeconomic variables was crucial to model LGD during the downturn conditions. Although the improvement of predictive accuracies was modest compared with the regression model with account variables only, it was still worth mentioning that the improvements were seen across different forecast periods and more significant at the portfolio level.

Yang and Tkachenko (2012) proposed several practical approaches to estimate LGD and EAD factor or retail loans, both of which are values bounded in the unit interval. They considered using variable transformation techniques such as weight-of-evidence (WOE) and several methodologies including fractional response regression, mixture model and neural networks. They found mixture distribution model and neural networks were significantly better than the others, and the use of WOE transformation also provided decent improvements for all models. However, the empirical study was based on a sample of 500 cases which was quite small compared with other empirical studies making the evidence less reliable.

Different from above, the collection process was examined and incorporated into the modelling framework by Matuszyk et al (2010) which discussed modelling recovery rates of unsecured personal loans. They proposed to use a decision tree approach that was believed to be suitable to model both the decisions by lenders and the repayment risks of debtors. They suggested a two-stage model to obtain estimates where a logistic regression was applied to estimating which class the debtor belonged to and then a regression approach was adopted to estimate the LGD values. They

believed that this decision tree based approach allowed one to model downturn LGD because both lenders' collection policies and borrowers' repayment abilities might change, indicating that it is necessary to consider lenders and borrowers separately.

However, Bijak and Thomas (2014) pointed out two inherent problems in the two-stage modelling framework proposed in Matuszyk et al (2010): First the two models were estimated separately and the independent estimate can be incoherent, and thus a part of uncertainty of LGD estimates was lost; Second the selection of cut-off point at the first stage classification model could lead to a biased estimate of LGD. They proposed a Bayesian method which assumed that the LGD followed a mixture normal distribution with the weight probability of loss following a Bernoulli distribution. This approach was able to simulate the bimodal distribution of LGD and was free of the problems discussed above. The model was estimated by Markov Chain Monte Carlo (MCMC) procedure and applied to predicting LGD of retail unsecured loans of a UK bank. They found the estimates of the Bayesian model were very close to that estimated by the frequentist approach, and the predictive performances were also very close. The authors suggested that this Bayesian model generated similar estimates compared with the frequentist approach, and was free from the drawbacks of the frequentist approach which allowed for a much better description of uncertainty of the LGD estimates. To be exact, it obtained a predictive distribution for each single loan which can be further used for stress testing and estimating the downturn LGD.

2.4.3.3 Mortgage loans

Mortgage loans are secured by the collaterals where collateral value is more sensitive to the change of economic conditions, and thus it is necessary to consider new variables and methodologies. For example, Qi and Yang (2009) found that current loan-to-value ratio (CLTV) was the most crucial determinant that explained the majority of variation in the LGD regression model for residential mortgage loans. They reported that the adjusted R^2 decreased from 0.61 to 0.145 if CLTV was removed from the regression model. Apart from CLTV, other loan characteristics were also found to be significant such as loan size, loan purpose, property type and the age of loan, etc. They suggested that CLTV was a much better predictor than LTV and should be employed if available.

Leow et al (2013) also placed emphasis on the influences of economic conditions on the LGD of mortgage and unsecured personal loans. They found that the inclusion of House Price Index (HPI) and interest rate contributed to the model fit. But the

model with macroeconomic variables was only able to improve predictions for higher LGD bands, implying that the LGD predictions were skewed towards the downturn periods. They suggested that the economic factors might be non-linearly correlated with LGD for mortgage loans, or for different types of loans. In contrast the inclusion of macroeconomic variables brought little benefit for the predictions of unsecured loans LGD, suggesting that personal loan LGD was less affected by the economy after accounting for the loan characteristics.

Morone and Cornaglia (2010) presented a theoretical approach to estimate the downturn LGD for residential mortgages. They proposed a Bayesian approach to capture the effects of macroeconomic conditions on LGD and estimate the LGD together with the other economic drivers including default rates using vector regression model. The downturn LGD was then derived based on the estimates of default rates and real estate prices. They defined the downturn effect to be a ratio between two measures of loss: default rates times stochastic price dependent LGD and default rate times deterministic LGD, and they reported that such downturn effect was strongly negatively correlated with expected LGD, but it was far less sensitive to the change of other parameters, and suggested that this approach was sufficient to not only compute the LGD estimates, but also give benchmark values for the estimates from other models.

New methodologies have also been proposed to address the characteristics of mortgage loans. For example, Leow and Mues (2009) proposed to model the probability of repossession and the haircut, which is the reduction of in price of a repossessed property to its market valuation, separately and developed a two-stage model for mortgage loans LGD. They found that the incorporation of a probability of repossession model was significantly better than a model with only the commonly used CLTV. They further validated the two-stage model including a probability of repossession model and a haircut model and suggested that using the expected shortfall approach for haircut model was better than using a point estimate haircut model because the latter method tended to underestimate LGD. They reported the out-of-sample predictive results and showed that the two-stage models were significantly better than the single stage model, and the estimates from the expected shortfall haircut model were shown to be more robust and reliable in the low LGD bands according to the scatterplot.

Tong et al (2013) developed a zero-adjusted gamma regression model for LGD of

mortgage loans to distinguish the cases with zero loss from the others. They modelled the loss amount in GBP instead of LGD and assumed that the loss amount followed a mixed discrete-continuous distribution which was similar to the inflated beta regression model. In their study the loss amount was considered to be the response variable instead of LGD following a Bernoulli distribution and the non-zero cases were assumed to follow a gamma distribution which was considered to be suitable for the right-skewed LGD distribution. They further included observable covariates to reparameterize the three parameters of gamma distribution and incorporated non-parametric smoothing terms in order to identify the non-linear relationships between the predictors and the response variable. They compared the proposed model with other two methods including an OLS with beta transformation and a Tobit regression model, and reported that the zero-adjusted gamma model gave very close predictions to Tobit model and were consistently better than OLS with beta transformation over years. They suggested that the proposed model could be a powerful alternative to the existing LGD models which allowed one to model the loss amount without turning the resulting model into a 'black box' because the covariates were modelled using flexible non-parametric splines and easy to interpret. However, this model was not significantly more competitive compared with Tobit and other regression methods.

2.4.3.4 Corporate bonds

Different from bank loans, recovery information of corporate bonds is available from proprietary databases including Moody's Ultimate Recovery Database, S&P's Credit Pro Database and Altman-NYU Salomon Centre Corporate Bond Default Master Database, and thus most new methodologies have been first proposed for modelling recovery rates of corporate bonds. Altman and Kalotay (2010) proposed a mixture distribution model to model instruments ultimate recovery rates and presented an intuitive Bayesian approach for estimation. They found that it was flexible to accommodate important idiosyncratic features of recovery distributions using a mixture of three normal distributions, and the associated probability weights also had an intuitive economic interpretation. The use of information related to the debt cushion of defaulted exposures and industry level expectations of default enabled the model to give better performances for both in-sample and out-of-sample predictions.

Bruche and Aguado (2010) presented a new econometric model where both default and recovery rates were driven by a latent factor named as the "credit cycle".

The recovery rates were fitted with a beta regression model with the shape parameters reparameterized by a linear combination of observable covariates and latent credit cycle factor. The model was estimated by a state space filter method where the latent factor followed a Markov Chain process. The inclusion of a credit cycle showed strong evidence that default and recovery risk were both affected by the time-varying systematic risk, although such an effect on recovery risk was much smaller than the contribution of time variation in default probabilities to systematic risk. It found that the different phases of business and credit cycle were seen to be more evident in annual recovery rates.

Huang and Oosterlee (2011) extended the beta regression by including a random effect term into the linear predictors formulating a generalized beta regression model. They applied this model to estimating the recovery rates of corporate bonds and found that the inclusion of random effect significantly increased the model fit of beta regression, but it indicated that it was unnecessary to reparameterize both mean and dispersion parameters in the beta regression. However, they did not include any observable covariates in the empirical study which made the implications less convincing. They also demonstrated that the tail loss could be approximated efficiently based on the LGD estimates of generalized beta regression and PD estimates of single factor model using Normal approximation or Saddlepoint approximation.

Jacobs and Karagozoglu (2011) built a simultaneous equation system of beta-link generalized linear models for estimating LGD at both obligor and instrument levels. This beta-link generalized linear model can be regarded as a new fractional response model where the link function is denoted by a beta distribution instead of using a logit transformation. The in-sample and out-of-sample model performances showed that this new methodology improved LGD predictions effectively, and the instrument level LGD predictive accuracies were better than the obligor level as expected. They also demonstrated the economic and statistical significance of the explanatory variables in a unified framework including economic, industry, firm-wide effects as well as instrument characteristics, and documented a new finding that a larger firm tended to have a lower LGD but a larger loan would result in a significantly higher LGD.

In general there is no dominant methodology showing the best performance of LGD modelling until now. In terms of corporate bonds Qi and Zhao (2011) compared a collection of six algorithms for LGD modelling including parametric and

non-parametric regression models. They found non-parametric models such as regression trees and neural networks generally outperformed parametric regression methods including OLS and fractional response regression for both in-sample and out-of-sample performances. Among the parametric models, fractional response regression had a slight edge over OLS regression. They also carefully examined two transformation methods including inverse Gaussian and beta transformation and found the performances of transformation methods were sensitive to the choice of the selection of perturbation values which made them less attractive for LGD predictions. Here, a perturbation value ε is applied to transforming the boundary points 0 and 1 to ε and $1-\varepsilon$ as they are undefined for some distributions. For the non-parametric methods the regression tree overcame the black-box limitation of neural networks, and became more competitive when more splits were allowed. They suggested that the bi-modal pattern after transformation did not necessarily lead to a good model fit and accurate predictions, and the bi-modal distribution might be of only secondary concern when modelling LGD.

However, new evidence was presented in Yashkir and Yashkir (2013) who conducted a comprehensive analysis to compare the performances of a group of LGD models including censored least squares method, Tobit regression, three-tiered Tobit regression, inflated beta regression, beta regression and censored gamma regression models for corporate bonds. All above models were seemingly suitable to model LGD based on the their assumptions and characteristics, but the empirical study showed that model fit depended mainly on the choice of explanatory variables rather than the choice of the model used. They also reported that different types of debt depended on different sets of covariates according to the calibration results. They found the LGD models provided better model fit using the data during distressed business cycle periods and proposed to make use of the estimated parameters for stress testing. The findings of this study confirmed that the performance of LGD models relied on not only the methodologies but also on the choices of determinants, and no model outperformed the others for all types of debts.

2.4.4 Survival regression models

Literature on modelling LGD or RR by survival analysis is rather limited. Leow and Mues (2011) developed a competing risk survival analysis model to predict the time that a defaulted mortgage loan reached the occurrence of some event (repossession), and provided a more accurate prediction of the LGD on a UK

mortgage loan dataset than the traditional regression models. Zhang and Thomas (2009) compared OLS with survival regression models including both Cox proportional hazards (PH) regression model and accelerated failure time (AFT) models on retail credit cards recovery rates from a UK bank. They argued that survival models can not only treat the undergoing repayments as censored observations, but also provide a flexible choice of distributional assumption to model the irregular recovery rates distribution. Because the Cox PH model is free of distributional assumptions, they investigated three distributions for the AFT models including Weibull, Log-logistic and Gamma distributions. However, they found that the performances of those survival models were less competitive compared with ordinary linear regression in terms of R^2 and MSE. They also adopted the two-stage model in Bellotti and Crook (2012) and found that the OLS model gave better performances than Cox and AFT models in terms of R^2 and MSE. They suggested that one possible reason was that in AFT models the zero recovery rate cases had to be separated accurately first which was difficult to achieve. Their results also indicated that the recovery rates of cases in the testing set were unknown, and it is impossible to test the model predictions for those debtors who were still repaying, which weakened the competitiveness of survival models.

2.4.5 Non-parametric estimators

Non-parametric estimators have also been explored to estimate the mean or the probability density distribution of LGD, especially for workout LGD. Renault and Scaillet (2004) proposed to apply a beta kernel non-parametric estimator to analyze the recovery rates of corporate bonds. Compared with the most commonly used Gaussian kernel estimator, this method is free of boundary bias and is well suited for density estimation on the unit interval. They examined the properties of this method and compared its performances on both simulated and corporate bonds samples. They found that the non-parametric fit the density distribution of the recovery rates better than a parametric beta distribution by seniorities and industries, indicating that recovery rates were far from being beta distributed. They showed that the inappropriate use of a parametric beta distributional assumption of recovery rates distribution may lead to substantial underestimate of credit VaR, and argued that the beta distribution should be used carefully to calibrate the empirical mean and variance of recovery rates distribution at high loss probability levels such as 99.9%.

One major drawback of beta kernel estimator is that the estimate is not a density

function which may lead to a biased estimate of VaR. To overcome this drawback Calabrese and Zenga (2010) further extended the beta-kernel estimator and proposed a mixture of beta kernel estimator to estimate the density function of recovery rates. They showed that this new method is preferable to the beta kernel estimator based on Monte Carlo simulations. They also compared the performances of both estimators on the recovery rates of bank loans from Bank of Italy's database, and found that the new estimator is better to estimate the densities at the boundaries where total recovery and total loss cases are most likely to happen. Based on the empirical study they indicated that a suitable parametric model for recovery rates is a mixture of two random variables including a right-skewed and a symmetric one.

Although non-parametric estimators have been presented remarkable advantages of fitting density distributions of LGD, it is necessary to notice the limitations of this class of methodologies. The non-parametric estimators are restricted to fit the distribution, and they are not able to either make predictions or to analyze the determinants. Therefore the non-parametric estimators are not usually considered and applied in the empirical studies of LGD.

2.4.6 Support vector regression and other machine learning techniques

Machine learning techniques have been widely applied in credit scoring and credit rating, because most machine learning methods have presented more advantages to handle more complex non-linear relationship between independent and response variables. Regression trees have been explored to estimate recovery rates for both bank loans and corporate bonds. Bastos (2010) showed that the performances of regression trees outperformed the fractional response regression for shorter horizons but became less competitive for longer horizons, and he suggested that a regression tree could be an alternative to parametric models to forecast LGD. Qi and Zhao (2011) also examined regression trees to predict recovery rates of corporate bonds. They found that regression tree provided better fit than the neural network in the ten fold cross validation, and the over-fitting problem can be avoided by properly controlling the number of splits.

In the recent decade more complicated techniques such as neural networks (NNs) and support vector machines (SVMs) have shown impressive improvements on both classification and regression problems compared with the traditional statistical regression models. Compared with neural networks SVMs are particularly attractive in that it can effectively mitigate the overfitting issue because of the principle of

structural risk minimization. SVM was first proposed by Vapnik et al (1995, 1998) and the model with respect to classification and regression problems are formulated to be support vector classification (SVC) and regression (SVR) models. Large number studies have shown the advantage of SVCs and its variations in credit scoring and credit classification. Some popular variations of SVMs include the least-squared support vector machine (LS-SVM) proposed by Suykens et al (1999, 2002) and proximal support vector machine (PSVM) from Fung and Mangasarian (2001). For example, Gestel et al (2003) employed the LS-SVC to study the credit ratings of banks, and they found that the LS-SVC showed the best classification accuracy than other techniques such as ordinal logistic regression and neural networks. Baesens et al (2003) conducted a comparative study of various classification algorithms on eight real-life credit scoring datasets, and they found that LS-SVC with radial basis function (RBF) kernel and neural networks outperformed other techniques.

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. For example, Francis and Cao (2001) simply applied SVR model to predict the real life financial time series and compared it with a multi-layer neural network model. The empirical results showed that SVR was more advantageous than NN model. Another seminal paper is from Gestel et al (2001) who developed a Bayesian framework based LS-SVR model to predict financial time series and volatility, and they found that this new model gave significant improvement on the out-of-sample predictions of 90-day T-bill rate and daily DAX30 closing prices. More recently, research has focused on adapting SVR techniques to estimating survival regression models. Two main ideas have emerged in this topic. First is to model the time-to-event directly. Shivaswamy et al (2007) improved the SVR model to accommodate both left and right censored cases in the constraint conditions. Shim et al (2011) proposed a semi-parametric LS-SVM to model the accelerated failure time model and found that this model provided accurate estimates for the parametric and non-parametric components based on the small cell lung cancer data. The second idea is to find a utility function to map the instances with their failure times. For example, Van Belle et al (2010) proposed a ranking model under the SVR framework to classify the survival data to fit them into different categories which is similar to Cox's model. However, they found it difficult to incorporate time-dependent variables and competing risks into this model.

In terms of LGD modelling very limit research has been done so far to study the applications of NNs and SVR, although both NNs and SVRs have already demonstrated advantages in regression models and time series predictions. Qi and Zhao (2011) showed that neural networks delivered good model fit in terms of cross-validation predictions, but it had a potential risk of over-fitting which could be overcome effectively by SVM. Loterman et al (2011) have investigated the applications of SVR to LGD modelling. This study examined a total of 24 techniques on bank retail loan LGD from six datasets. They found the LGD variation in some cases remained largely unexplained as the average performances in terms of R^2 ranged from 0.04 to 0.43, and the non-linear techniques were shown to give consistently good predictive accuracies compared with linear models. Among the non-linear techniques neural networks and SVM outperformed the other benchmarking methods significantly across all the data sets in the empirical experiment, suggesting the strong presence of non-linear relationship between the independent covariates and LGD. They also proposed a hybrid model based on a combination of linear and non-linear techniques which showed a similarly good performance as the non-linear techniques with the added advantage of preserving a comprehensibility of the linear model component.

Tobback et al (2014) also studied LGD of bank retail loans from two US datasets and compared the performances of a collection of linear and non-linear models including linear regression, regression tree, SVR and a hybrid model similar to Loterman et al (2011). Different from the results in Loterman et al (2011), they found the best out-of-time performances were reported for the hybrid model combining a linear regression with a support vector regression on the error terms, and a regression tree showed the best out-of-time forecasting performance. They also documented the importance of incorporating macroeconomic variables, which improved the predictive performances and confirmed the impacts of business cycle on LGD that has been found in previous studies.

2.4.7 Transformations on LGD

Gupton and Stein (2002) proposed 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 conducted extensive validation studies showing that such beta transformed linear regression gave better predictions than historical average methods. Firstly we introduce the beta density function such that

$$f(y;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1},$$

where $\alpha > 0$ and $\beta > 0$ are two shape parameters and $\Gamma(\cdot)$ is Gamma function.

$$y_i^{Beta} = \Phi^{-1}(Beta(y_i,\alpha,\beta)),$$

where y_i^{Beta} denotes the transformed dependent variable, and $Beta(\cdot)$ defines a Beta distribution function while the inverse normal distribution function is given as $\Phi^{-1}(\cdot)$.

However, in the latest version LossCalc v3.0 (Dwyer and Korablev, 2009) the beta transformation was dropped out and an identity link function was adopted with more model transparency, and a transformation method was employed after fitting the model to ensure that the predicted LGDs were bounded between 0 and 1. Dwyer and Korablev (2009) justified the linear link function specification by using a bucket method, where all the observations were first grouped into 100 buckets based on their predicted recovery rates, and then the correlation of average actual and predicted recovery rates of each bucket was computed which showed a clearly linear relationship.

Another transformation method that has been frequently used in LGD modelling is a logistic transformation defined as

$$y_{logit} = \ln(\frac{y}{1-y}),$$

Dullman and Trapp (2004) presented a detailed discussion about the impact of different distributional assumptions on the expected recovery rates and on economic capital including normal distribution, log-normal distribution and logistic distribution. They demonstrated that the log-normal and normal distribution were shown to better explain the observed recovery rates. By analyzing the sensitivity of the recovery rate to the systematic risk factor, they found that the logit-normal distribution model was more stable and recommended in the sense that the sensitivity was less dependent on whether the recovery rate was calculated by the market prices at default or at emergence. Following the study of Dullman and Trapp (2004), Rosch and Scheule (2005) and Hamerle et al (2006) both adopted the logistic transformation method to model corporate bonds recovery rates.

In theory the model fit and predictive accuracies should be improved after the irregular LGD distributions are dealt with properly by some transformation method. However, empirical evidence does not support this assumption and previous studies have shown that the models with LGD transformations are outperformed in general.

For example in Bellotti and Crook (2012) the regression models with a logit, a Probit or a beta transformation had a lower out-of-sample MSE compared with the relevant models. Qi and Zhao (2011) also showed that the OLS regression models with a transformation method were very sensitive to the perturbation values presenting poor out-of-sample predictions. It is still arguable to apply a transformation to LGD as Qi and Zhao (2011) indicating that the bi-modal LGD distribution may not be the major concern for LGD modelling.

2.4.8 Comparative analysis

In previous sections we conduct a thorough survey of the literature on the LGD or recovery rate modelling for bonds and loans. Some studies focused on developing or seeking a better methodology to improve LGD predictions while others only aimed to identify the influences of the determinants of interest on LGD. In this section we restrict our interest to the methodologies presenting a comparative study to analyze their advantage and disadvantages. Table 2.1 and 2.2 present a list of literature regarding to the determinants and methodologies investigated in LGD/RR modelling of bank loans and corporate bonds.

In general machine learning techniques are more competitive for both corporate bonds (Qi and Zhao, 2009) and bank loans (Bastos, 2010; Loterman et al, 2011). The most advantage of the commonly used machine learning techniques such as regression trees and neural networks is that they are especially strong to fit non-linear relationships, but neural networks have long been blamed for its ‘black box’ quality which makes regression trees more attractive as an alternative method to the traditional parametric regression techniques. Another promising technique is support vector regression which has also been studied by Loterman et al (2011) and Tobback et al (2014), and it has been recognized to be as competitive as neural networks. But it also suffers from the same ‘black box’ problem similar with neural networks. To overcome this disadvantage Loterman et al (2011) proposed a hybrid model and they found this hybrid model was giving close performances to the SVR method while preserving the interpretability of a linear model. However, both Loterman et al (2011) and Tobback et al (2014) worked on the data of bank loans, and there is no further study related to applying SVR techniques to predicting recovery rates of corporate bonds yet. For the nature of ‘black box’ of neural networks and SVR models, it is recommended to apply a hybrid model that incorporate the non-linear techniques on the error terms with a linear model to increase the comprehensibility as well as to

improve the predictive accuracies.

Parametric models are more commonly applied to modelling LGD than non-parametric techniques. According to Table 2.2 it can be observed that OLS and fractional response regressions are most employed to identify the influences of predictors such as Acharya et al (2007), Dermine and Carvalho (2006, 2007) and Khieu et al (2012), and these studies all focused on corporate bonds or corporate loans. In terms of predictive accuracies Qi and Zhao (2009) and Loterman et al (2011) both demonstrated that non-parametric techniques consistently outperformed parametric regression models. Table 2.1 Panel A shows that most studies related to parametric models focus on recovery rates modelling of bank loans. Among the parametric models, however, there is no dominant methodology shown in the literature. In theory it is unreasonable to fit the irregular distributions of LGD using a normal distribution which defines the dependent variable in an infinite interval while LGD is bounded in a unit interval. But empirical studies found that OLS was likely to produce as good predictive performances as other complicated models. Apart from the commonly used models including OLS, fractional response regression and beta regression, other generalized linear models and survival regression models have been explored to improve the model fit. Empirical evidence supported the competitiveness of some new methodologies such as inflated beta regression (Calabrese, 2014) and zero-inflated gamma regression (Tong et al, 2013), but other evidence also found that OLS presented very robust predictions compared with other non-linear parametric models according to Zhang and Thomas (2010) and Bellotti and Crook (2012). Yashkir and Yashkir (2013) also pointed out that model performances relied on the choice of determinants more than that of methodologies. In overall the most commonly used parametric techniques such as OLS and fractional response regressions are no less competitive compared with other proposed complicated models.

For corporate bonds, Table 2.1 Panel B demonstrates that structural and factor models are most applied to estimating the default and recovery risk for the nature that they are derived from asset pricing models. The structural and factor models are capable of incorporating risk contagion effects conveniently but they are rarely used to make an out-of-sample forecast. Literature related to factor models is mainly interested in investigating the specific effects of interest on the recovery rates, benchmarking estimates of asset correlation, exploring the PD/LGD correlation and

estimating portfolio loss distributions. Single factor models were first applied to estimating PD and benchmarking asset correlation of corporate bonds and personal loans including Hamerle et al (2003a, 2003b) Crook and Bellotti (2012). The empirical evidence of these studies revealed that the asset correlation values specified in the Basel guidelines were too conservative, where the empirical asset correlation estimated from historical data was significantly lower than the Basel specifications. PD and LGD correlation has been explored by Frye (2000a, 2000b), Dullmann and Trapp (2004), Rosch and Scheule (2005, 2008) using the single factor joint modelling framework. Two factor models were also proposed to explain PD/LGD correlation from the perspective of both systematic and idiosyncratic risk factors instead of considering systematic risk factor alone such as Pythkin (2003), Hillebrand (2005), and Bade et al (2011). Two factor models also provide a closed form expression of expected losses although they are more difficult to calibrate. Because the factor models have latent time-varying random effect terms that can only be calibrated based on historical data, the factor models are rarely used to make out-of-sample predictions according to Table 2.1 Panel B. In summary factor models are not regarded to be a good choice for generating predictive performances, but they are convenient to estimate the correlation effects of default and recovery risk and to simulate loss distributions. We conclude this section claiming that non-parametric techniques show more advantages in predicting instrument level recovery rates and parametric are better to identify significant determinants, and we suggest that structural and factor models are convenient to estimate loss distributions at portfolio level.

2.5 Conclusions

We conclude this chapter summarizing the discussions presented above. In section 2.1 and 2.2 we introduce the definitions of terms related to Basel Accord and the key risk parameters including PD, LGD and EAD, and demonstrate three different measurements of LGD including market LGD, workout LGD and implied market LGD. In section 2.3 we discuss the determinants of LGD for bank loans and corporate bonds respectively. Empirical studies on retail loans show that the loan characteristics are more significant than the borrower's effects, and the effects of incorporating economic conditions depend on the products: mortgage loans are more sensitive to the macroeconomic movements than the retail credit cards. However, for corporate bonds it exhibits that all types of characteristics are important to explain LGD. More

specifically, characteristics of instrument, borrower, industry and economy are all shown to be significantly related to LGD.

In section 2.4 we focus on the methodologies used and proposed in empirical studies related to LGD or recovery rates modelling. Parametric statistical regression models have been widely considered for modelling both retail and corporate credit products. Evidence in literature has shown that the ordinary linear regression models have been presented to be consistently competitive compared with other generalized linear models. Other non-linear models including fractional response regression, inflated beta regression and zero-adjusted gamma regression also show their advantages to handle the unique characteristics of LGD distribution. Non-parametric techniques such as regression trees, neural networks and support vector regression models are shown to be more powerful to predict LGD than the parametric models at the expense of model transparency. But literature only analyzed the existed methodologies, none of which has further improved machine learning techniques to adapt to LGD predictions. We propose to improve support vector regression techniques to account for the unobservable heterogeneities of seniorities for corporate bonds. We report that the improved SVR techniques outperform the original SVR model, and all SVR modes show substantial advantages over the other statistical regression techniques. Regarding to the transformation techniques of LGD prior to modelling, we find that methods such as a beta or a logistic transformation can not make any significant improvement, which is consistent with the evidence discussed in section 2.4.7.

Merton's structural models and the evolved factor models are more suitable to estimate PD and LGD of corporate bonds. One advantage of structural based models is that it is easily to incorporate the PD/LGD correlation in model setting. However, the calibration of model raises difficulties because of the complexity of likelihood function, and thus the estimates of parameters based on historical data tend to be unstable. Empirical studies have investigated applying factor models to estimate LGD of corporate bonds. However, the systematic risk factor is specified to reflect the time-varying credit cycle, but it has not been applied to other levels. To examine the effects of unobservable heterogeneities we explore to specify random effect terms to multiple levels including time, obligor and seniority, and find when random effect is specified at the obligor level, the model fit is improved significantly for modelling instrumental level recovery rates of corporate bonds. The finding shows that the firm

specific heterogeneity can be effectively explained when obligor-level random effect has been incorporated. We also study the effects of distributional assumptions on the performances of factor models, and find that a linear model of normal distribution outperforms the other assumptions such as log-normal, logit-normal and beta distributions.

The decision tree based two-stage modelling framework adopted in literature has been demonstrated to be an effective way to account for the concentrated cases at the boundaries 0 and 1, but the performances of two-stage model are not satisfactory. It is noticed that the two-stage models only consider to applying the logistics regression at the first stage. Here we propose to apply SVM at the first stage to better discriminate the cases of zero and full recovery from the remaining others, and then we investigate other regression techniques at the second stage to predict recovery rates of personal credit cards. We find that the predictive accuracies of two-stage models are strongly dependent on the performances of the classification stage, but less related to the choice of the regression techniques. As SVM models outperform other traditional classification methods such as logistic regression, the two-stage model with SVM shows better predictive accuracies than the other settings.

In this chapter we have discussed the relevant empirical studies related to LGD modelling including the basic concepts, determinants and methodologies. We focus on the discussions of methodologies and present a comparative analysis to elaborate the advantages and disadvantages of techniques applied to different products. The gap has also been addressed in the discussions. In the following chapters we will discuss more details of the contributions developed in this thesis.

Table 2.1. Summary of literature of LGD modelling

Panel A. Bank loans

Author(s)	Year	Sample period	Sample size	Product type	Country	Methodologies
Dermine and Carvalho	2006	1995-2000	374	Corporate loans (SME)	Portugal	Complementary log-log regression
Dermine and Carvalho	2008	1995-2000	374	Corporate loans (SME)	Portugal	Complementary log-log regression
Caselli et al	2008	1990-2004	11649	Bank loans	Italy	OLS
Chalupka and Kopecsni	2009	1989-2007	N/A	Bank loans to SME	Czech	Fractional response and inflated beta regression
Grunert and Weber	2009	1992-2003	120	Corporate loans	Germany	OLS
Seidler and Jakubik	2009	2000-2008	37	Corporate loans	Czech	Structural joint model
Zhang and Thomas	2010	1987-1999	27278	Retail loans	UK	OLS and survival regression models
Bastos	2010	1995-2000	374	Bank loans to SME	Portugal	Fractional response regression and regression trees
Matuszyk et al	2010	1989-2004	50000	Retail loans	UK	Two-stage model
Calabrese	2014	2000-2001	144996	Retail loans	Italy	Fractional response and inflated beta regression.
Calabrese and Zenga	2010	2000-2001	149378	Retail loans	Italy	Mixture beta kernel estimator

Loterman et al	2011	NA	From 3351 to 119211	Bank loans and revolving credit	N/A	24 techniques including machine learning and statistical techniques
Khieu et al	2012	1987-2007	793	Corporate loans	US	OLS and fractional response regression.
Bellotti and Crook	2012	1999-2005	55000	Retail loans	UK	OLS, two-stage model and Tobit regression
Tobback et al	2014	2002-2008, 1984-2011	17,346,986	Revolving credit and corporate loans	N/A	OLS,OLS with Box-Cox transformation, regression tree, SVR, OLS+SVR
Bijak and Thomas	2014	1987-1998	50000	Retail loans	UK	Bayesian estimation for two-stage model
Qi and Yang	2009	1990-2003	241293	Mortgage loans	US	OLS
Leow and Mues	2011	1983-2001	140000	Mortgage loans	UK	OLS, two-stage model
Tong et al	2013	1988-2000	113000	Mortgage loans	UK	zero-adjusted gamma regression, OLS with beta transformation, Tobit regression
Leow et al	2013	Mortgage: 1990-2002 Retail: 1989-1999	Mortgage: 120000 Retail: 48000	Mortgage and retail loans	UK	OLS, two-stage model

Panel B. Corporate bonds

Author(s)	Year	Sample period	Sample size	Product type	Country	Methodologies
Frye	2000	1982-1999	N/A	Corporate bonds	US	A single factor joint model
Renault and Scaillet	2004	1981-1999	623	Corporate bonds	US	Beta kernel estimator
Dullmann and Trapp	2004	1982-1999	1511	Corporate bonds and loans	US	A single factor joint model
Rosch and Scheule	2005	1985-2004	N/A	Corporate bonds	US	A single factor joint model for PD and LGD
Hamerle et al	2006	1983-2003	1286	Corporate bonds	US	Single factor model
Acharya et al	2007	1987-1999	1511	Corporate bonds and loans	US	OLS
Bruche and Aguado	2010	1981-2005	898	Corporate bonds	US	generalized beta mixed model
Altman and Kalotay	2010	1987-2006	3492	Corporate bonds and loans	US	mixture normal distributed regression
Qi and Zhao	2011	1985-2008	3751	Corporate bonds and loans	US	OLS, fractional response regression, regression with transformations, regression tree, neural networks
Jacobs and Karagozoglu	2011	1986-2008	3902	Corporate bonds	US	Beta-link generalized linear regression
Chava et al	2011	1980-2008	46605	Coporate bonds	US	A dynamic frailty model for PD and OLS regression for LGD
Bade et al	2011	1982-2009	187638	Corporate bonds	US	A two-factor joint model
Frye and Jacobs	2012	1996-2009	6120	Corporate bonds	US	A single factor model

Yashkir and Yashkir	2013	1981-2011	4275	Corporate bonds and loans	US	censored least squared method, Tobit regression, Censored linear regression three-tiered Tobit model, inflated beta regression, Beta regression, censored gamma regression
Jankowitsch et al	2014	2002-2010	1270	Corporate bonds	US	Pooled linear regression model

Table 2.2. Determinants

Panel A. Bank loans

Author(s)	Year	Product type	Significant variables of interest
Dermine and Carvalho	2006	Corporate loans (SME)	Loan size
Dermine and Carvalho	2008	Corporate loans (SME)	Loan size, Age of firm
Caselli et al	2008	Bank loans	Customer segments, business investments, household consumption and industry distress dummies.
Chalupka and Kopecsni	2009	Bank loans to SME	Collateral types, Year of loan origination dummies
Grunert and Weber	2009	Corporate loans	Risk premium, Multiple loan contracts, Continuation indicator of company
Seidler and Jakubik	2009	Corporate loans	N/A
Zhang and Thomas	2010	Retail loans	Loan purpose dummies, Time at address dummies, Time in occupation dummies, Default rates
Bastos	2010	Bank loans to SME	Loan size, Personal guarantee, Ratings, Age of firm
Matuszyk et al	2010	Retail loans	Application score, Loan amount, Time of loan until default, Number of months in arrears in the last 12 months
Calabrese	2014	Retail loans	Geographic dummies

Calabrese and Zenga	2010	Retail loans	N/A
Loterman et al	2011	Bank loans and revolving credit	N/A
Khieu et al	2012	Corporate loans	Loan size, Loan type dummies
Bellotti and Crook	2012	Retail loans	N/A
Tobback et al	2014	Revolving credit and corporate loans	LTV, Utilization rate, Seniority and Rating dummies
Bijak and Thomas	2014	Retail loans	Age of exposure, Amount of loan, Number of months with arrears and Worst arrears within the life of the loan
Qi and Yang	2009	Mortgage loans	Loan-to-value, Current loan-to-value
Leow and Mues	2011	Mortgage loans	Loan-to-value, Loan-to-value at default, Security type dummies, Region dummies
Tong et al	2013	Mortgage loans	HPI, Time on books, Debt-to-value, Security type dummies
Leow et al	2013	Mortgage and retail loans	Unemployment rate, HPI growth rate, Interest rate

Panel B. Corporate bonds

Author(s)	Year	Product type	Significant variables of interest
Frye	2000	Corporate bonds	N/A
Renault and Scaillet	2004	Corporate bonds	N/A
Dullmann and Trapp	2004	Corporate bonds and loans	N/A
Rosch and Scheule	2005	Corporate bonds	Index of four coincident indicators and other ten leading indicators
Hamerle et al	2006	Corporate bonds	Debt ratings, Seniority dummies, Annual default rates
Acharya et al	2007	Corporate bonds and loans	Industry distress indicator
Bruche and Aguado	2010	Corporate bonds	Contractual, industry, and economic characteristics
Altman and Kalotay	2010	Corporate bonds and loans	Instrument characteristics
Bade et al	2011	Corporate bonds and loans	Ratings, Rating shifts, Gross private domestic investment
Qi and Zhao	2011	Corporate bonds	Collateral type dummies
Jacobs and Karagozoglu	2011	Coporate bonds	Contractual, financial, industry, and economic characteristics

Chava et al	2011	Corporate bonds	Total assets, industry dummies,T-bill rate
Frye and Jacobs	2012	Corporate bonds	N/A
Yashkir and Yashkir	2013	Corporate bonds and loans	Instrument characteristics, Capital structure, Industry characteristics
Jankowitsch et al	2014	Corporate bonds	Instrument characteristics, bond transaction information, economic factors and liquidity proxies

Chapter 3
Data Description

3.1 Introduction

This chapter gives an overview of data related to modelling recovery rates of corporate bonds from multiple sources including Moody's Ultimate Recovery Database (MURD), Compustat and other public available online databases. The data will be used in the empirical studies of the following two chapters (Chapter 4 and 5). Section 3.2 provides details of definitions of all variables including recovery rates, instrument and firm characteristics as well as macroeconomic variables, and Section 3.3 concludes this chapter.

3.2. Samples

The recovery information is extracted from MURD. This database contains a list of tables that covers a comprehensive set of recovery data, which is also the data that Moody's used to develop its internal LGD rating model. It contains more than 1000 US companies with over 3000 instruments including bank loans, revolving credit products and corporate bonds. MURD has been widely used in the empirical studies of LGD modelling for both academic researchers and practitioners. For example, studies such as Frye (2000b), Rosch and Scheule (2005), Hamerle et al (2006), Bade et al (2011), Jacobs and Karagozoglu (2011) and Frye and Jacobs Jr. (2012) have investigated using factor models to fit recovery rates and to estimate PD/LGD correlation based on the MURD. Furthermore, Qi and Zhao (2011) conducted a benchmarking study comparing a group of LGD models for both corporate bonds and bank loans. Khieu et al (2012) restricted their interests in bank loans and explored the determinants of recovery rates. We are only interested in the recovery rates of the corporate bonds, the selection criterion limits the sample to entities domiciled in the US because it is necessary to incorporate and economic factors and the economic conditions may differ across different countries. The final sample has 1413 observations of defaulted corporate bonds observed from 1986 to 2012.

Bankruptcy law in the US has defined the absolute priority rule (APR) which states how a bankrupt firm's value is to be distributed to suppliers of capital. In theory the most senior creditors should be fully satisfied before any distributions are made to other junior creditors, and junior creditors should be fully paid before common

shareholders. Although APR is usually violated in practice according to previous studies, we believe seniority plays a strong role to influence the recovery rate of corporate bonds. There are five seniorities in our sample: junior subordinated, subordinated, senior subordinated, senior secured and senior unsecured. Each obligor may issue more than one instrument with varied seniorities. The MURD provides three calculation measurements for ultimate recovery rates: settlement method, liquidity method and trading price method. We use the discounted prices where the recovery rates are discounted back from the instrument's trading date using a proper interest rate. Table 3.1 gives the definitions of three methods are given as follows according to the technical document of MURD (Moody's Analytics, 2011). To supplement the definitions above, we provide the definitions of the nominal recovery values for the three methods in Table 3.2.

An indicator of each recovery rates calculation method is given in the database to show if the method is recommended. We adopt the method recommended as the ultimate recovery rate of the instrument. Figure 3.1 exhibits the distribution of recovery rates in our sample. It is obvious that the observations are clustered at recovery rates equal to 0 and 1. Panels A and B in Table 3.3 present the summarized statistics of recovery rates by year and seniority respectively. Panel B presents the recovery rates breakdown by seniority. On average bonds with higher seniorities have higher recovery rates than the lower seniorities bonds, although the subordinated and senior subordinated bonds have very similar average recovery rates.

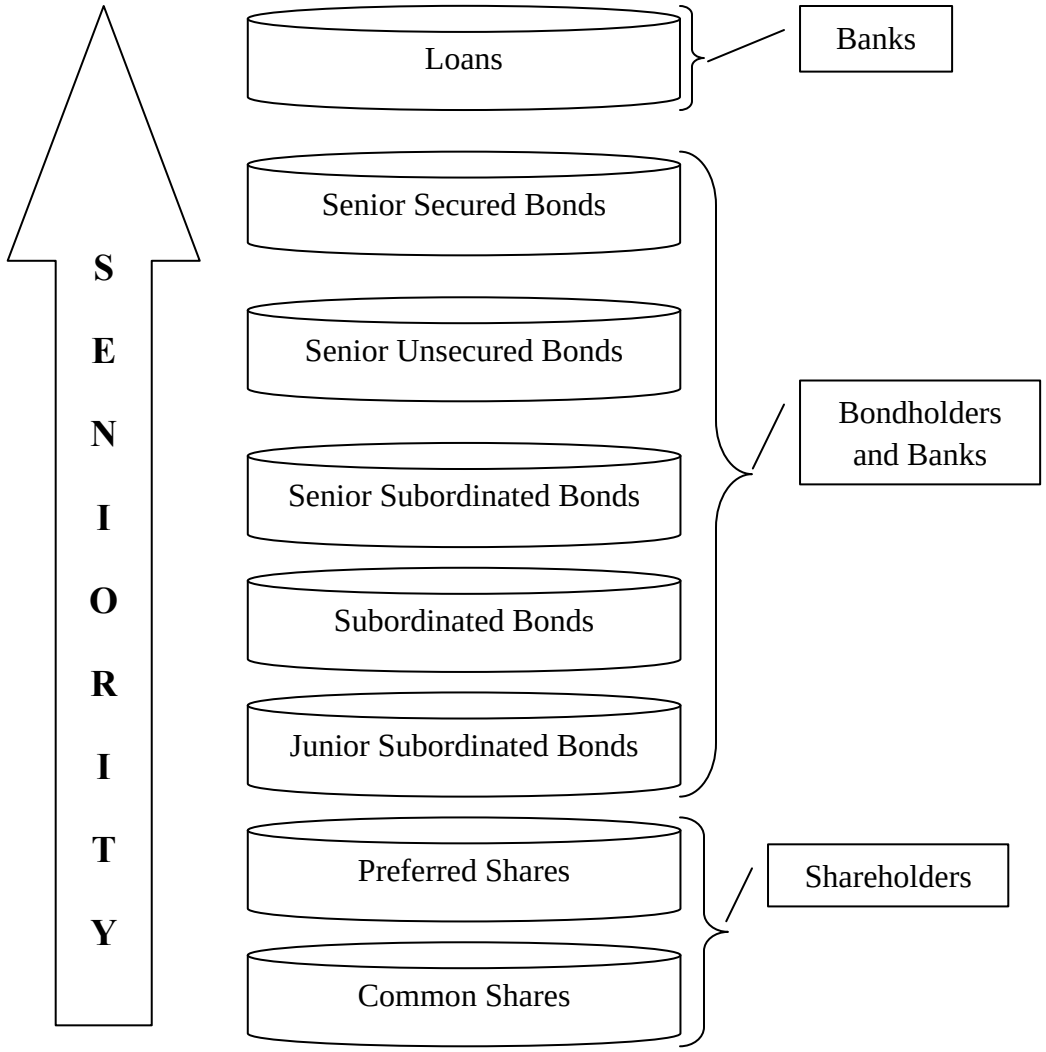
Table 3.1. Definitions of discounted methods of recovery rates

Discount Settlement Price	*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*
Discount Liquidity Price	*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*
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*

Table 3.2. Definitions of nominal methods of recovery rates

Nominal Settlement Total	*The sum value of the settlement instruments received for each defaulted instrument, taken at or close to emergence, divided by the total principal defaulted amount of the class, reflected as a percentage of the principal amount of at default*
Nominal Liquidity Total	*The sum value of the settlement instruments received for each defaulted instrument, taken at or close to emergence, divided by the total principal defaulted amount of the class, reflected as a percentage of the principal amount of at default*
Nominal Trading Price	*The average trading price at emergence of all instruments in the class, expressed as a percentage of par*

Figure 3.1. Flow chart of capital structure[1]



[1] This figure is from Schuermann (2004).

Table 3.3. Descriptive statistics of recovery rates

Table 3.3 presents the descriptive statistics of recovery rates by year and seniority respectively. Recovery rates are calculated based on the preferred methods provided in Moody's Ultimate Recovery Database (MURD).

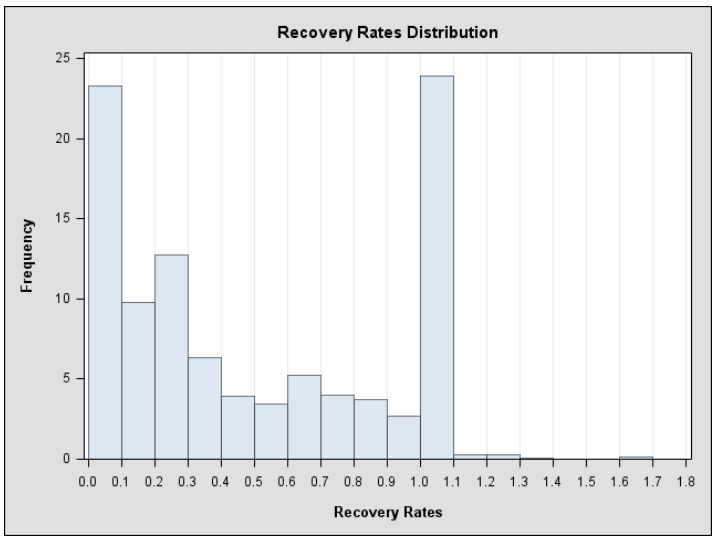
Panel A. Recovery rates breakdown by year

	No.	Mean	Std	Min	Max
1986	3	0.1436	0.1030	0.0272	0.223
1987	12	0.6195	0.2585	0.3185	1
1988	17	0.7850	0.3998	0.048	1
1989	14	0.1438	0.1695	0	0.4731
1990	65	0.2723	0.2849	0	1
1991	78	0.5629	0.3864	0	1
1992	47	0.6267	0.4287	0	1
1993	19	0.4104	0.4035	0	1
1994	13	0.6510	0.4177	0.0843	1
1995	25	0.6128	0.3806	0	1
1996	5	0.2989	0.4113	0.0024	1
1997	25	0.2687	0.1728	0	0.5577
1998	39	0.2977	0.3362	0	1
1999	54	0.3522	0.3558	0	1
2000	139	0.5449	0.4504	0	1
2001	192	0.3019	0.3360	0	1.012
2002	221	0.3840	0.3091	0	1.3691
2003	88	0.6578	0.3511	0	1.1298
2004	38	0.7244	0.4134	0	1.2766
2005	96	0.8026	0.2790	0	1.6978
2006	16	0.6299	0.4310	0	1.1567
2007	15	0.5704	0.4103	0.0024	1.013
2008	100	0.4910	0.3983	0	1.0029
2009	75	0.4903	0.3851	0	1.0373
2010	12	0.6228	0.4678	0.0039	1
2011	4	0.2870	0.2823	0.0455	0.5857
2012	1	0.2705	.	0.2705	0.2705
Total	1413	0.4806	0.3915	0	1.6978

Panel B. Recovery rates breakdown by seniority

	No.	Mean	Std	Min	Max
Junior Subordinate	28	0.1628	0.2634	0	1
Senior Secured	332	0.6292	0.3688	0	1.1298
Senior Subordinate	198	0.3150	0.3617	0	1.6978
Senior Unsecured	681	0.5100	0.3813	0	1.0499
Subordinate	174	0.3217	0.3743	0	1.3691
Total	1413	0.4806	0.3915	0	1.6978

Figure 3.2. Distribution of recovery rates



3.3. Variable selections

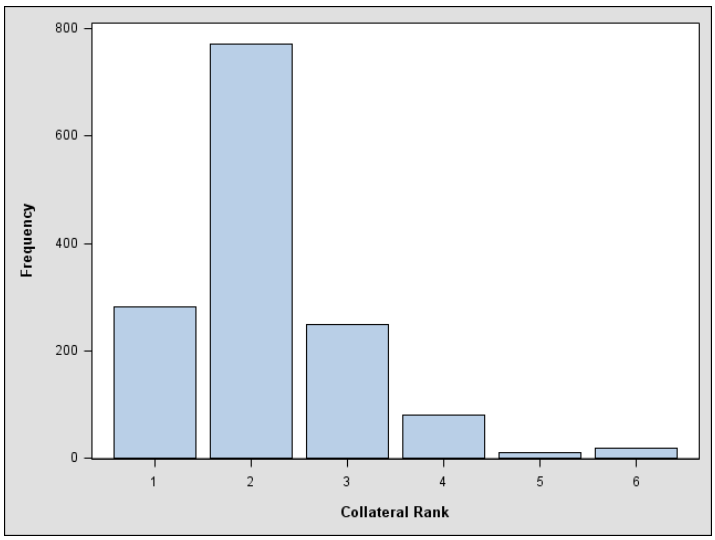
This section briefly introduces the descriptive statistics and distributions of independent variables. The descriptive statistics are reported in Table 3.4 with distributions presented individually for each variable. All the variables are measured at the ratio level except *Collateral Rank*. Based on its definition *Collateral Rank* is an ordinal variable, but it is regarded as a numeric variable in this study because its value is negatively correlated economically with recovery rate as discussed above. The two variables *Issue Size* and *Total Asset* are both subjected to a log transformation for scaling. Notice that three variables: *EBITDA*, *Debt Ratio* and *Book Value per Share* have abnormally large standard deviations. Note that the mean and median values of these three variables differ from each other significantly, and all of them show extremely skewed distributions according to the figures of their distributions. All these strongly suggest that there are outlier values of these variables. In the following we discuss the variables with more details.

3.3.1. Instrument characteristics

In MURD, characteristics of instrument are provided including the information of the instrument and recovery process. Instrument or contract characteristics have been proven to play an important role in explaining variations of recovery rates according to literature such as Acharya et al (2007), Jacobs and Karagozoglu (2011) and Khieu et al (2012). We select three variables to be the determinants in the empirical study including *Collateral Rank*, *Percent Above* and *Issue Size*.

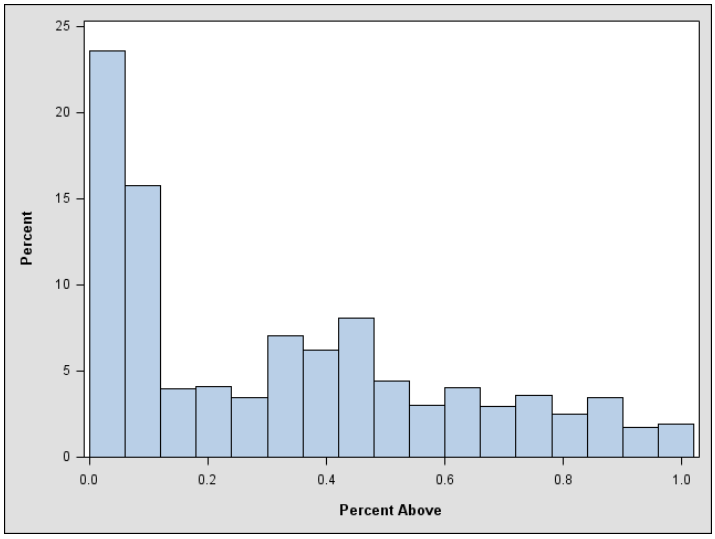
• ***Collateral Rank****: "The instruments in each event are ranked in relation to each other based on the structure prior to default, taking into consideration collateral and instrument type, based on analyst assessment". Collateral Rank* denotes the relative rank of the instrument for a given obligor ranged from 1 to 6. Figure 3.3 shows that the distribution of *Collateral Rank* is left skewed with most cases concentrated on the values of 2 and 3, and there are very few observations with a value higher than 4.

Figure 3.3. Frequency of Collateral Rank



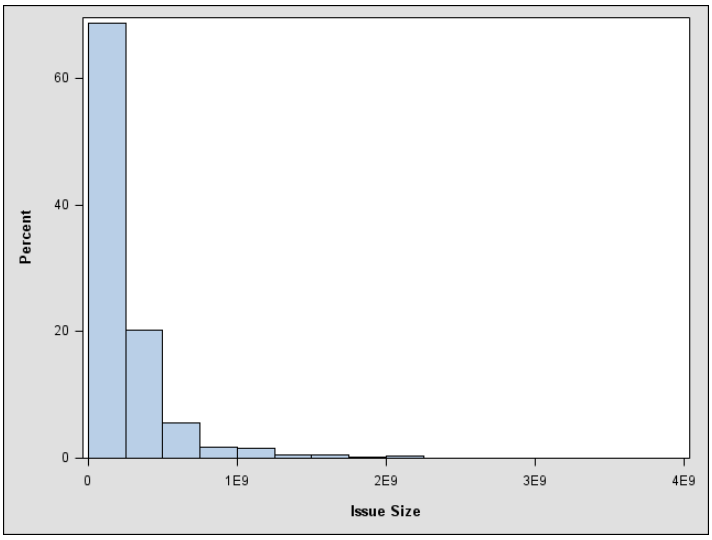
• ***Percent Above****: "The percentage of debt which is contractually senior to the current instrument. Percent of debt above is derived by taking the principal above in dollars and dividing it by the total issuer debt". Figure 3.4 shows that a large proportion of instruments have a Percent Above* between 0 and 0.2, and the observations are almost evenly distributed between 0.2 and 1. This variable is also included among the explanatory variables in Moody's internal LGD model LossCalc (Moody's special comment, 2004).

Figure 3.4. Distribution of Percent Above



• ***Issue Size***: *"Total original or face amount of this instrument in dollars"*. Figure 3.5 shows that the distribution is highly left skewed having a wide range from 1e6 to 4e9, and thus a log transformation is applied to scaling it into a small range.

Figure 3.5. Distribution of Issue Size



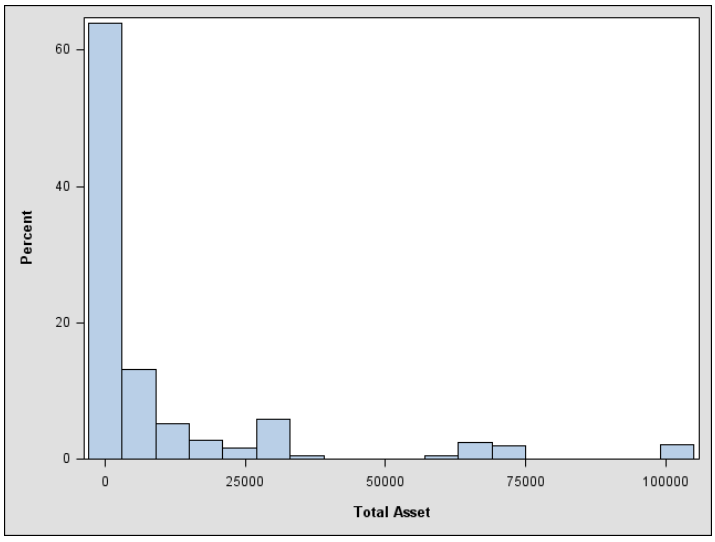
3.3.2. Firm characteristics

To study the effects of the firm level characteristics, we incorporate the accounting ratios from Compustat which is provided in Wharton Research Data Services (WRDS) into our sample, and merge them into the sample one year prior to the default dates of bonds based on the common identifier TICKER. In total, there are seven accounting ratio variables included for the empirical study. In the following the definitions and economic arguments will be given with details for all variables.

• ***Total Asset***: *The total asset value of the obligor.* Similar to *Issue Size*, the distribution is also greatly left skewed with most cases below 25000, and a log

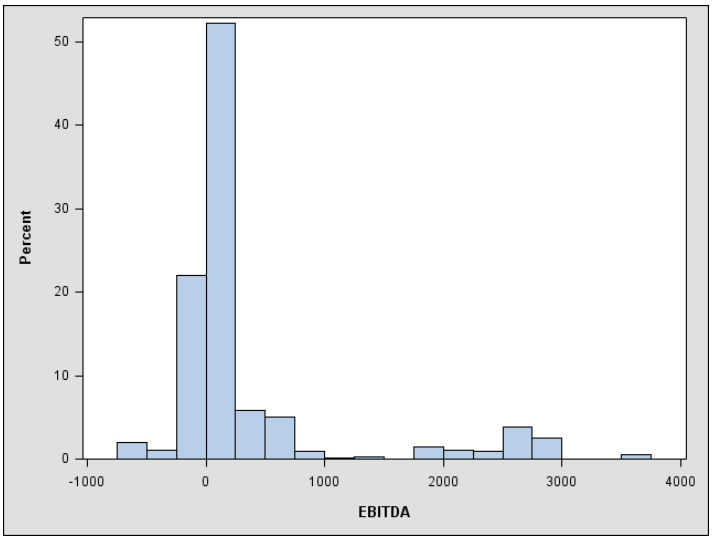
transformation is applied to *Total Asset* for scaling.

Figure 3.6. Distribution of Total Asset



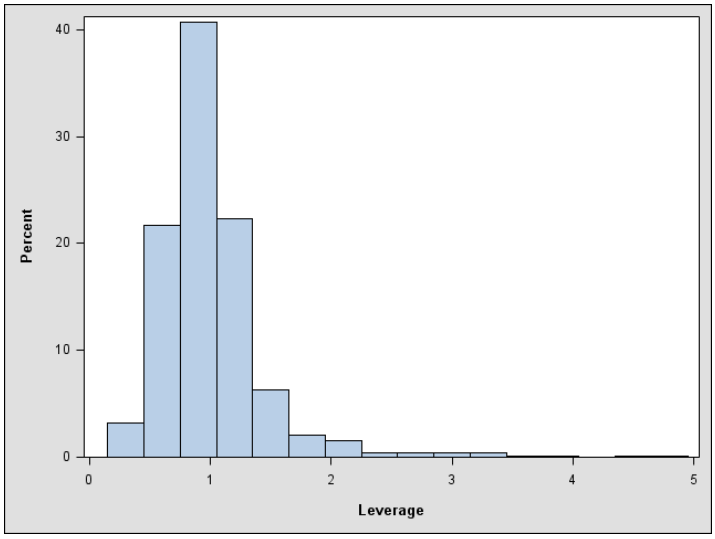
• ***EBITDA****: Earnings before interest, taxes, depreciation and amortization.* According to Table 3.4 it can be noted that there are outliers for this variable because the values of mean and median are far from normal and the standard deviation is also much greater than the mean. After deleting the cases with outliers it is clear to notice that most cases centred in the interval between -1000 and 1000, and a few of observations lie in the interval between 2000 and 3000.

Figure 3.7. Distribution of EBITDA



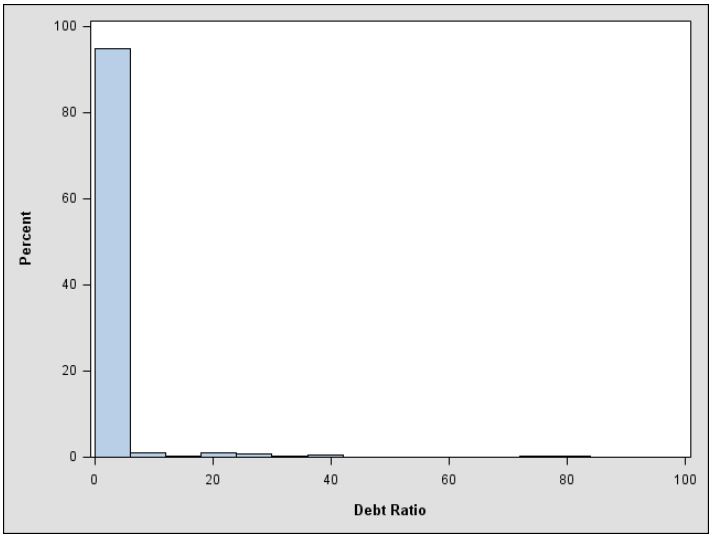
• ***Leverage***: *The ratio of total debts to total assets.* The distribution presented in Figure 3.8 finds that most cases are bounded between 0 and 2 and symmetrically distributed while the mean value is close to 1.

Figure 3.8. Distribution of Leverage



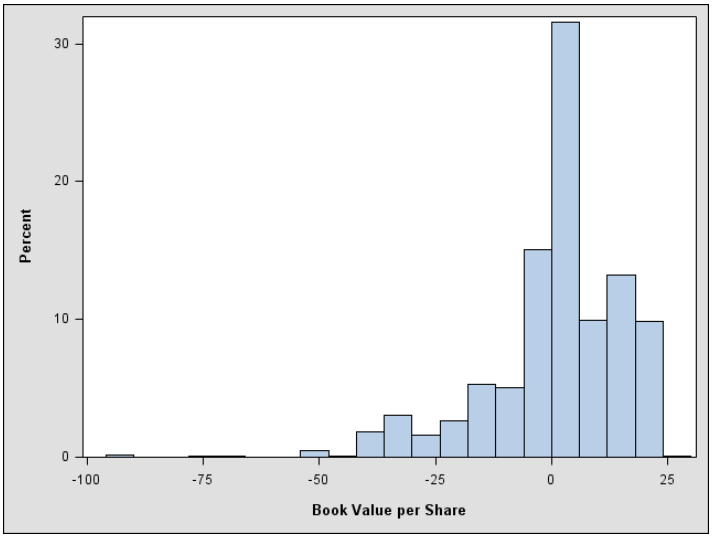
• ***Debt Ratio***: *The ratio of current liabilities to long term debt.* According to Table 3.4 we find there are outliers for this variable. Figure 3.9 shows the distribution after trimming the outliers, and it can be observed that the majority of observations are centred in the interval [0, 10] while others are sparsely distributed from 10 to 100.

Figure 3.9. Distribution of Debt Ratio



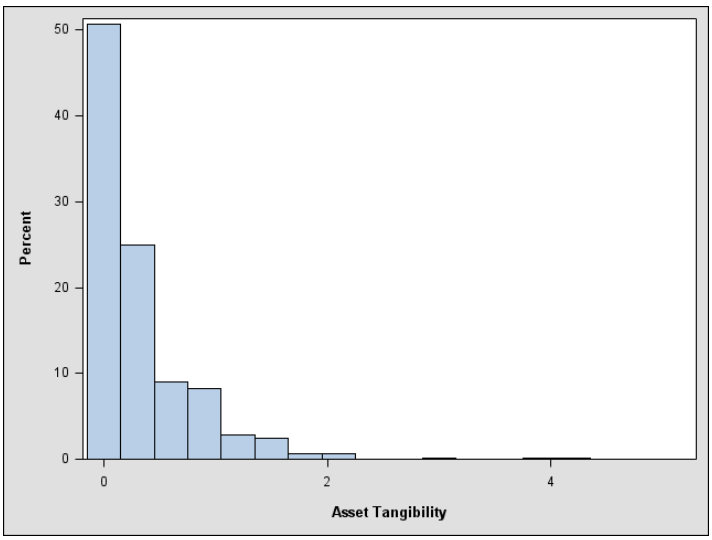
• ***Book Value per Share****: The book value of assets scaled by the total outstanding shares.* Similarly it can be found that there are outliers according to the huge standard deviation from Table 3.4. The right skewed distribution in Figure 3.10 shows that observations are concentrated in the interval between -50 and 25 with the mode value between 0 and 5.

Figure 3.10. Distribution of Book Value per Share



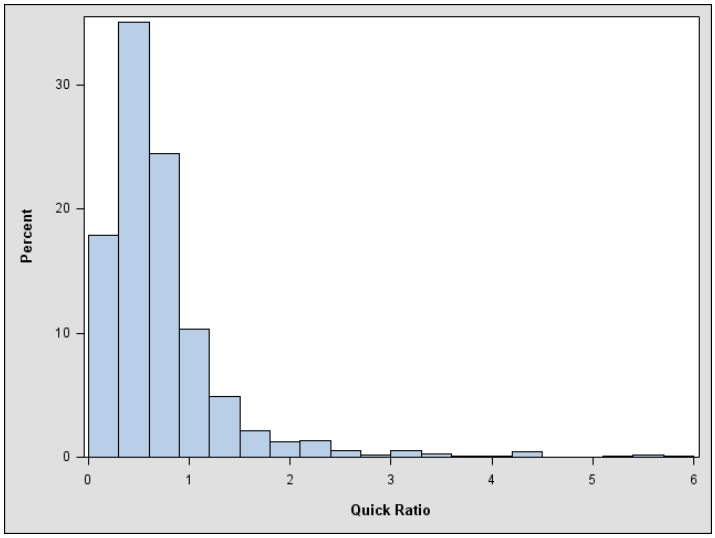
• ***Asset Tangibility***: *The ratio of tangible assets to intangible assets.* Table 3.4 shows that *Asset Tangibility* is bounded between 0 and 6 with a mean of 0.33. The left skewed distribution shown in Figure 3.11 indicates that for most companies in our sample the tangible assets are almost equal or twice as much as intangible assets.

Figure 3.11. Distribution of Asset Tangibility



• ***Quick Ratio***: *The sum of cash and short term investment and total receivables divided by the current liabilities.* Quick Ratio is also bounded in the interval [0, 6]. Table 3.4 shows that its mean and standard deviation values are 0.71 and 0.65 respectively, which implies the observations are centred between 0 and 2 and left skewed distributed based on Figure 3.12.

Figure 3.12. Distribution of Quick Ratio



3.3.3. Macroeconomic factors

There are four macroeconomic factors incorporated which are available from the open sources online to characterize the US economic conditions including *Growth Rate, Unemployment Rate, T-Bill Rate* and *Default Rate*. *Growth Rate* is defined as the US annual GDP growth rate. *T-Bill Rate* is the US three months Treasury bill rate as a proxy for the risk free interest rate. *Default Rate* is the annual issuer-weighted corporate default rate (Ou, 2013), which is defined by averaging the default rates for all ratings and is commonly used as a proxy of the default risk. Both accounting and macroeconomic covariates are incorporated one year prior to default.

Figure 3.13 shows the plot of annual default rates and recovery rates. Here we take the average of recovery rates with equal weights for each issue. The annual mean recovery rates present strong cyclicality, which we will further investigate by including the macroeconomic variables in the regression models. The aggregated recovery rate has a strong negative correlation with *Default Rate*. We also observe that *Growth Rate* presents a strong positive relationship with annual recovery rates. In contrast, the *T-Bill Rate* tends to move against with recovery rates. However, the relationship between *Unemployment Rate* and recovery rates is ambiguous according to Figure 3.13.

Figure 3.13. Plot of macroeconomic variables against recovery rates

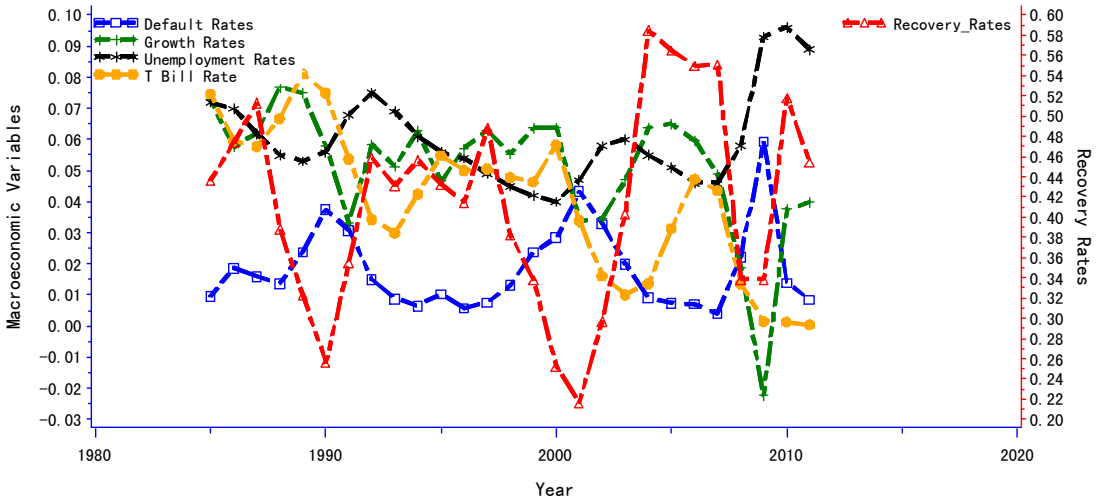


Table 3.4. Descriptive statistics of covariates

Table 3.4 presents the descriptive statistics of the covariates in the empirical study. All the variables are classified into three categories including instrument characteristics, firm characteristics and macroeconomic variables.

	Mean	Median	Std	Min	Max
			Instrument characteristics		
Collateral Rank	2.1677	2	0.9277	1	6
Percent Above	0.3230	0.2812	0.2894	0	1
Issue Size	2.299E8	1.427E8	3.081E8	100,000	3.987E9
			Firm characteristics		
Total Asset	10140.30	1570.91	20886.31	16.8320	103914
EBITDA	680.8151	55.6090	1920.55	-2439.94	10489
Leverage	0.9945	0.8843	0.4575	0.2893	4.8787
Debt Ratio	19.0926	0.4639	518.8207	0.0436	19455.2
Book Value per Share	1687.89	1.6962	46726.03	-875083	255000
Asset Tangibility	0.3310	0.1344	0.5004	0.0000	5.1922
Quick Ratio	0.7196	0.5565	0.6549	0.0124	5.9174
			***Macroeconomic variables* (%)**		
Growth Rate	5.1017	5.8080	1.6635	-2.2237	7.6852
T-Bill Rate	4.2160	4.36	1.9791	0.05	8.11
Default Rate	2.4307	2.3770	1.2968	0.3980	5.9340
Unemployment Rate	5.1183	4.7	0.9521	4	9.6

3.4. Expected signs

In this section we discuss the expected signs of independent variables individually based on their economic backgrounds and previous findings in literature.

• ***Collateral Rank***: The greater value of *Collateral Rank* means the instrument has a

relatively lower rank, indicating that it will be recovered after the instruments with a smaller value of *Collateral Rank* have been recovered. Therefore *Collateral Rank* is expected to have negative effects on recovery rate. Acharya et al (2007) and Qi and Zhao (2011) both found the effects of collateral types were significant in recovery rates modelling. However, they only used dummy variables to be proxies of collateral types. Jacobs and Karagozoglu (2011) found that *Collateral Rank* was statistically and economically significant showing that more significant and better secured bonds had better recoveries. It is believed that *Collateral Rank* contains more information than the collateral types' dummies because it shows the relative importance of the instrument of its issuer. Given that *Collateral Rank* enters into the models in above literature without being transformed to dummy variables, in the empirical studies in Chapters 4 and 5 we also consider it to be a numeric variable rather than make any further transformations. Limitations of such use will be discussed in Chapters 4 and 5.

• ***Percent Above***: A higher *Percent Above* indicates that there is more debt needed to be repaid prior to the repayment of current instrument, and thus it is expected to lead to a lower recovery rate. Qi and Zhao (2011) and Jacobs and Karagozoglu (2011) both found that *Percent Above* was significant statistically in the regression model for corporate bonds.

• ***Issue Size***: Previous findings in the literature on the effect of debt size are mixed. Dermine and Carvalho (2006) showed a significant negative sign of loan size on the recovery rate, and they suggested that the foreclosure of larger loans tended to be delayed by the bank and thus affected the recovery rates negatively. But Acharya et al (2007) found that the *Issue Size* was positively related to the recovery rate. They argued that a larger *Issue Size* indicated greater bargaining power for the obligor in the recovery process, and therefore a higher recovery rate was expected.

• ***Total Asset*****:** *Total Asset* can be interpreted as a proxy for firm size. It is expected that large companies will have lower probabilities of default and higher recovery rates because they have more resources to liquidate their assets to repay the debts during the bankruptcy process. But Acharya et al (2007) also pointed out that a large firm might have more difficulties in the process of debt reorganization with higher bankruptcy costs incurred which could lead to a lower recovery rate.

• ***EBITDA***: *EBITDA* is a measure of the profitability of a firm. The profitability of an obligor's asset is expected to influence recovery rate positively. A higher *EBITDA* of an obligor indicates that the firm has more cash to cover its debt and it is likely to

result in a higher recovery rate.

• ***Leverage***: *Leverage* is widely studied in the literature with controversial findings. Dwyer and Korablev (2009) argued that a higher leverage increased the PD, hence led to a lower recovery rate because of the negative correlation between PD and recovery rate. Also Acharya et al (2007) argued it was rather difficult to anticipate its effect on recovery rates ex ante. They noted that higher *Leverage* could be related to a more dispersed ownership structure that made the recovery process more complicated. On the other hand, Khieu et al (2012) suggested that a higher value of *Leverage* might imply that it was easier to restructure the debt after bankruptcy suggesting a higher recovery rate. Therefore we do not make any ex ante assumption on this variable.

• ***Debt Ratio***: It is reasonable to assume that for the short term obligations the funds would be withdrawn immediately and debt extension would be problematic, so the obligor with a high *Debt Ratio* would find it more difficult to repay debts. In summary a higher *Debt Ratio* is expected to affect the recovery rates negatively. Amiram (2011) has found that *Debt Ratio* was significantly positively related to LGD using the same data from MURD, which was consistent with our assumptions.

• ***Book Value per Share****:* A higher *Book Value per Share* implies that the company has a better position in the stock market and it should have more assets to repay debts and yield a higher recovery rate. Jacobs and Karagozoglu (2011) showed that book value of a firm was significantly negatively correlated with LGD in an empirical study. They suggested that a larger firm may have more power to negotiate with the lenders which was associated with superior recoveries.

• ***Asset Tangibility***: Since most intangible assets are difficult to liquidate after default, a higher value of *Asset Tangibility* should be associated with a higher recovery rate. Previous research (Schuermann, 2004) has revealed that defaulted companies whose assets are of a more tangible quality such as utilities or heavy manufacturing industries have significantly higher recoveries. Empirical studies such as Acharya et al (2007), Jacobs and Karagozoglu (2011) and Jankowitsch et al (2013) have all confirmed similar evidence that *Asset Tangibility* is positively related to recovery rate.

• ***Quick Ratio***: Similar to *Book Value per Share* and *Debt Ratio, Quick Ratio* is another proxy for the potential solvency capability of a firm. A higher *Quick Ratio* indicates an obligor has more cash and short term funds to cover its short term liabilities, which influences the recovery rates positively.

For the macroeconomic variables, *Growth Rate* and *Unemployment Rate* are used

to be the proxy indicators for the US economic activities. When the economic activity increases recovery rate should go up accordingly. In other words, recovery rate is expected to be positively correlated with annual GDP *Growth Rate* and negatively with the *Unemployment Rate*. A higher *T-Bill Rate* implies a higher cost during the debt recovery process, which subsequently affects the recovery rate negatively (Qi and Zhao, 2011). *Default Rate* has been found to be a powerful predictor of recovery rate that shows strong positive influences in previous studies (Altman et al, 2005).

3.5. Conclusion

This chapter provides an overview of the data for recovery rates modelling of corporate bonds. The data is extracted from commercial databases including MURD and Compustat as well as public online sources. The sample is constructed by matching the companies in MURD and Compustat according to the common identifiers and the macroeconomic variables are incorporated one year prior to default. The defaulted instruments are restricted to corporate bonds issued by US companies dating back to 1986. Summarized statistics of both dependent and independent variables are given together with the illustrations of the economic arguments. Macroeconomic variables presented to show the long term trend and correlation with recovery rates. Empirical studies in the following two chapters will develop and explore various new methodologies for predicting recovery rates of corporate bonds using the data sample constructed in this chapter.

Chapter 4
Support Vector Regression for Loss Given Default Modelling

4.1. Introduction

In this chapter the applications of machine learning techniques to LGD modelling will be introduced. It aims to investigate whether data mining technique are able to improve the predictive accuracy of LGD for corporate bonds, and focuses on support vector regression (SVR) models and a series of model variations to account for the unobservable heterogeneity of corporate bonds. The methodologies proposed in this chapter have been rarely applied to LGD modelling and are compared with other commonly used LGD models on the loss data introduced in Chapter 3.

LGD models including statistical regression models and data mining techniques have been explored in literature. The most popular methods of LGD modelling are OLS and fractional response regression for both corporate bonds and bank loans. For example, Acharya et al (2007) concluded from including the industry distress dummies into a linear regression model that industry distress conditions had strong negative effects on the RR of defaulted firms' debts. Qi and Yang (2009) in a study of LGD of residential mortgages demonstrated that LGD could be explained by linear regression that included debt characteristics, with loan-to-value playing the single most important role. These results were confirmed by Khieu et al (2012) which estimated RR of bank loans with loan characteristics, borrower characteristics and macroeconomic conditions using both OLS and fractional response regression models.

Empirical LGD distributions are often bi-modal and usually bounded between [0, 1], suggesting that more complicated regression models are needed. To improve the model fit and predictive accuracy of the model, various transformations of LGD have been tried prior to the modelling stage. For example Gupton and Stein (2002) proposed 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. Based on this a beta-link generalized linear model was proposed by Jacobs and Karagozoglu (2011) to estimate LGD at firm and instrument levels jointly which showed remarkable predictive power. Calabrese (2014) applied an inflated beta regression model to predict recovery rates of loans from The Bank of Italy where the dependent variable was 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. Bellotti and Crook (2012) benchmarked a number of different transformations and algorithms to predict the LGD for a credit cards data set. Surprisingly, they found that OLS with no variable transformations gave greater predictive accuracy.

Although parametric models are simple to implement and easy to explain, previous research reported rather poor predictions of LGD, and generalized linear regression models could not achieve significant improvements compared with linear regression. Zhang and Thomas (2010) compared both linear regression and survival regression for modelling RR of personal loans from a UK bank, and reported the out-of-sample R^2 as low as 0.0904 for linear regression, and the parametric survival models exhibited even poorer predictions. It is also 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) showed the model fit of simple linear regression to be rather weak with R^2 of 0.1428, and still the predictions of this model outperformed the other ones including logit and probit models.

In contrast, non-parametric methods provide 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) compared parametric fractional response regression and a non-parametric regression tree model to forecast bank loans RR and found that the latter was superior. More strong evidence came from Qi and Zhao (2011) who compared six modelling methods for a mixed portfolio of bonds and loans, and they found data mining techniques such as decision trees and neural networks performed significantly better than other parametric methods in terms of both model fit and prediction accuracy. Tong et al (2013) developed a zero adjusted gamma model to predict LGD of a UK bank and showed that such a semi-parametric formulation gave favourable 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 their application to LGD modelling. SVM was

first studied by Vapnik et al (1995, 1998) and has been 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 et al (2011), has investigated the application of SVR to LGD modelling. A comprehensive benchmarking study was given on six retail loan data sets with 24 techniques showing that neural networks and SVR models consistently outperformed other traditional linear methods. But it did not make any further improvements on SVR models.

This chapter makes three original 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 this chapter is organized as follows. Section 4.2 presents the models, and Section 4.3 discusses the results and conclusions are drawn in Section 4.4.

4.2. 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.

4.2.1. Linear Regression

Previous empirical research shows that linear regression models appear to be of comparable predictive accuracy as other more complicated statistical models (Qi and Zhao, 2011; Bellotti and Crook, 2012) 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

$$\begin{aligned} y_i &= \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma^2) \end{aligned}, \tag{4.1}$$

Maximum likelihood methods can be applied to estimate the parameters.

4.2.2. Fractional Response Regression

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

$$E(y_i \mid \mathbf{x}_i) = G(\boldsymbol{\beta}^T \mathbf{x}_i), \tag{4.2}$$

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

$$\begin{aligned} G(\boldsymbol{\beta}^T \mathbf{x}_i) &= \exp(\boldsymbol{\beta}^T \mathbf{x}_i) \,/\, (1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i)) \\ G(\boldsymbol{\beta}^T \mathbf{x}_i) &= \exp(-\exp(-\boldsymbol{\beta}^T \mathbf{x}_i) \end{aligned}, \tag{4.3}$$

and the quasi maximum likelihood function can be written as follows

$$\log L = \sum_i y_i \log(G(\boldsymbol{\beta}^T \mathbf{x}_i)) + (1 - y_i) \log(1 - G(\boldsymbol{\beta}^T \mathbf{x}_i)). \tag{4.4}$$

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 et al (1999, 2002). Two improved models are proposed based on LS-SVR.

4.2.3.1. Least Squares Support Vector Regression

Still consider the dataset described above, and the LS-SVR is defined based on the quadratic loss function such as

$$\begin{aligned} &\min\ J(\mathbf{w},b;u_i) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{N} u_i^2 \\ &s.t.\ \ y_i = \mathbf{w}^T\varphi(\mathbf{x}_i) + b + u_i, i = 1,...,N \end{aligned}, \tag{4.5}$$

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 regularization 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 α_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

$$\begin{pmatrix} 0 & \mathbf{e}^T \\ \mathbf{e} & \bar{\mathbf{K}} \end{pmatrix}\begin{pmatrix} b \\ \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{y} \end{pmatrix}, \tag{4.6}$$

where $\mathbf{e} = (\underbrace{1,...,1}_{1\times N})^T$, $\mathbf{y} = (y_1,...,y_N)^T$, $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_N)^T$, $\bar{\mathbf{K}} = \mathbf{K} + \frac{1}{C}\mathbf{I}$, 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

$$\begin{cases} \boldsymbol{\alpha}^* = \bar{\mathbf{K}}^{-1}(\mathbf{y} - b^*\mathbf{e}) \\ b^* = \dfrac{\mathbf{e}^T\bar{\mathbf{K}}^{-1}\mathbf{y}}{\mathbf{e}^T\bar{\mathbf{K}}^{-1}\mathbf{e}} \end{cases}, \tag{4.7}$$

Finally the estimated regression model can be written as

$$g(\mathbf{x}) = \sum_i \alpha_i^*\mathbf{K}(\mathbf{x}_i,\mathbf{x}) + b^*\,. \tag{4.8}$$

Note that in this work we only consider the LS-SVR model instead of the original SVR model based on the epsilon-insensitive loss function proposed by Vapnik (1996) for two reasons: First, LS-SVR model has only two parameters (the regularization parameter *C* and the kernel function parameter) to tune while the original SVR model has three (the loss function parameter epsilon as well as the above two). Second, LS-SVR has been proved to be more predictive than the original SVR model (Suykens et al; 1999, 2002). The following improved models are also based on the LS-SVR.

4.2.3.2. 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 $D = \{(\mathbf{x}_{kj}, y_{kj})\},\ j = 1,...,p_k,\ k = 1,...,M$ 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 $p_1 + p_2 + ... + p_M = N$, where M indicates the total number of groups in this dataset. The least squares SVR model with different intercepts can be constructed as follows

$$\begin{aligned} &\min\ J(\mathbf{w}, b_k; u_{kj}) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2}\sum_{k=1}^{M} b_k{}^2 + \frac{C}{2}\sum_{k=1}^{M}\sum_{j=1}^{p_k} u_{kj}^2 \\ &s.t.\ \ y_{kj} = \mathbf{w}^T\varphi(\mathbf{x}_{kj}) + b_k + u_{kj} \\ &\qquad k = 1,...,M \quad j = 1,...,p_k \end{aligned} \quad . \tag{4.9}$$

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 (4.9) can be written as

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

The KKT conditions are

$$\begin{cases} \dfrac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_k\sum_j \alpha_{kj}\varphi(\mathbf{x}_{kj}) = 0 \Rightarrow \mathbf{w} = \sum_k\sum_j \alpha_{kj}\varphi(\mathbf{x}_{kj}) \\ \dfrac{\partial L}{\partial b_k} = b_k - \sum_j \alpha_{kj} = 0 \Rightarrow b_k = \sum_j \alpha_{kj} \\ \dfrac{\partial L}{\partial u_{kj}} = Cu_{kj} - \alpha_{kj} = 0 \Rightarrow u_{kj} = \dfrac{\alpha_{kj}}{C} \end{cases} \quad . \tag{4.10}$$

Then the dual form problem is given as

$$\min\ \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{W}\boldsymbol{\alpha} + \frac{1}{2C}\boldsymbol{\alpha}^T\boldsymbol{\alpha} - \mathbf{y}^T\boldsymbol{\alpha}\,. \tag{4.11}$$

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 (4.11) with respect to $\boldsymbol{\alpha}$

$$(\mathbf{K} + \mathbf{W} + \frac{1}{C}\mathbf{I})\boldsymbol{\alpha} = \mathbf{y}\,, \tag{4.12}$$

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 equation (4.9) is obtained as

$$\begin{aligned} \mathbf{w}^* &= \sum_k \sum_j \alpha_{kj}^* \varphi(\mathbf{x}_{kj}) \\ b_k^* &= \sum_j \alpha_{kj}^* \end{aligned}\,. \tag{4.13}$$

4.2.3.3. 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

$$\begin{aligned} &\min\ J(\mathbf{w}, b_k; u_{kj}) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} + \frac{1}{2}b^2 + \frac{C}{2}\sum_{k=1}^{M}\sum_{j=1}^{p_k} u_{kj}^2 \\ &s.t.\ \ y_{kj} = \mathbf{w}^T\varphi(\mathbf{x}_{kj}) + \boldsymbol{\beta}^T\mathbf{z}_{kj} + b + u_{kj} \\ &\qquad k = 1,...,M \quad j = 1,...,p_k \end{aligned}\,, \tag{4.14}$$

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 (4.5). 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

$$\begin{cases} \dfrac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_k \sum_j \alpha_{kj}\varphi(\mathbf{x}_{kj}) = 0 \Rightarrow \mathbf{w} = \sum_k \sum_j \alpha_{kj}\varphi(\mathbf{x}_{kj}) \\ \dfrac{\partial L}{\partial \boldsymbol{\beta}} = \boldsymbol{\beta} - \sum_k \sum_j \alpha_{kj}\mathbf{z}_{kj} = 0 \Rightarrow \boldsymbol{\beta} = \sum_k \sum_j \alpha_{kj}\mathbf{z}_{kj} \\ \dfrac{\partial L}{\partial b} = b - \sum_k \sum_j \alpha_{kj} = 0 \Rightarrow b = \sum_k \sum_j \alpha_{kj} \\ \dfrac{\partial L}{\partial u_{kj}} = Cu_{kj} - \alpha_{kj} = 0 \Rightarrow u_{kj} = \dfrac{\alpha_{kj}}{C} \end{cases}\,. \tag{4.15}$$

The dual form is

$$\min\ \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Z}\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{V}\boldsymbol{\alpha} + \frac{1}{2C}\boldsymbol{\alpha}^T\boldsymbol{\alpha} - \mathbf{y}^T\boldsymbol{\alpha}\,, \tag{4.16}$$

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 (4.5). Model (4.16) can be solved with the same procedure as above and the linear equation systems can be obtained as follows

$$(\mathbf{K} + \mathbf{Z} + \mathbf{V} + \frac{1}{C}\mathbf{I})\boldsymbol{\alpha} = \mathbf{y}\,. \tag{4.17}$$

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

$$\begin{aligned} \mathbf{w}^* &= \sum_k \sum_j \alpha^*_{kj} \varphi(\mathbf{x}_{kj}) \\ \boldsymbol{\beta}^* &= \sum_k \sum_j \alpha^*_{kj} \mathbf{z}_{kj} \\ b^* &= \sum_k \sum_j \alpha^*_{kj} \end{aligned}\,. \tag{4.18}$$

4.2.4. Two-stage model

The two-stage modelling framework was 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<LGD<1 by a decision tree, and then estimated the LGD belongs to the interval (0, 1) by an ordinary linear regression. More details can be referred to Bellotti and Crook (2012).

4.2.5. Transformations

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

$$y_{logit} = \ln(\frac{y}{1-y})\,, \tag{4.19}$$

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

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1},\ p > 0, q > 0\,, \tag{4.20}$$

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

$$y_{beta} = N^{-1}(Beta(y; p, q))\,, \tag{4.21}$$

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. Similar to Qi and Zhao (2011) the zeros and ones are adjusted upward and downward by a small amount respectively to facilitate the beta

transformation.

4.3. Empirical analysis

4.3.1 Model specification

The empirical analysis is based on the data introduced in Chapter 3. The summarized statistics and details of recovery rate and all the other independent variables are provided and explained in Chapter 3. The empirical study includes two parts to investigate the effects of heterogeneity of bonds. Firstly all the models are fitted and examined on the whole sample where they are referred to be aggregated models. Next the whole sample will be segmented based on the seniority of bonds to explore the heterogeneity effects on the predictive performances. It is observed that subordinated bonds have relatively 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". Because the firm characteristics and macroeconomic variables are matched with one year latter, the regression model can be constructed and presented as follows

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

where subscript i denotes the *i*-th instrument and t is the related default year. The empirical experiment in this section is presented in two subsections: firstly for all seniorities pooled together and secondly, models for individual seniority are presented separately.

4.3.1.1. 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 split procedure is repeated 100 times where the samples are drawn at the instrument level with different random samples drawn each time to ensure the robustness of the results. The regularization 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. Alternative metrics are applied in Loterman et al (2011) which applied an overall average rank to compare the performance of a collection of models based on multiple

metrics. Here the pair-wise t-tests are used to examine the differences of each individual metric and the corresponding statistics are presented.

The models presented in Section 4.2 have been fitted to the aggregated training samples. For parametric regression techniques these include linear regression, fractional response regression where the 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, model (4.5)), least squares support vector regression with different intercepts (LS_SVR_DI, model (4.9)) and semi-parametric least squares support vector regression (Semi_LS_SVR, model (4.14)). 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.

4.3.1.2. 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 4.3.1. 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-*th* 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

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_i (r_i - \hat{r}_i)^2}$$
$$\mathrm{MAE} = \frac{1}{n}\sum_i \mid r_i - \hat{r}_i \mid \quad . \tag{4.22}$$
$$\mathrm{R}^2 = 1 - \frac{\sum_i (r_i - \hat{r}_i)^2}{\sum_i (r_i - \overline{r})^2},\ \overline{r} = \frac{1}{N}\sum_i r_i$$

All support vector models adopt the RBF kernels defined as

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

where σ is the scale parameter of the kernel and is tuned in cross-validation. The parameters tuning is carried out by a grid search method. The initial values of the regularization and kernel parameters are set to be 1 and 0.1 respectively, and the parameter values migrate in an interval of a collection of specified values such that

- *C*: 1, 2, 5, 10, 50
- σ : 0.1, 0.5, 1, 2, 5

As when the parameter value is out of the given interval the model performance significantly deteriorates. The combination of the two parameter values with the best performance on the training set is recorded and applied. It should be noted that this method is far from exhaustive because of the limitation of computation time, and the parameter tuning is one of the major difficulties in building support vector models which is out of the scope of this research. In that the parameters are tuned based on the performance of the training set, the overfitting risk is unavoidable which will be further discussed in the next section.

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.

4.3.2. Experimental Results

4.3.2.1. Results of Aggregated Models

Table 4.1 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 4.2 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.

From Tables 4.1 and 4.2 it is clear to see that all the SVR models (models

M5-M13) outperform the statistical models (models M1-M4)[2]. 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 4.2 shows that such an improvement is significantly better than the other methods except for LS_SVR_DI. Panel B in Table 4.2 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 4.2 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. It should be noted that the variable *Collateral Rank* has already contains partial information of the heterogeneity of seniorities given its definition. That said, the improvement made by the two proposed models may not be fully reflected because of the use of this variable. Even so the empirical results show that the two proposed models still outperform the original SVR, indicating that *Collateral Rank* does not fully explain the seniority heterogeneity. However, the model prediction will not be affect if *Collateral Rank* is coded into dummy variables, but the computation cost is increased.

Notice also from Table 4.1 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

[2] All the models are labelled. See the descriptions in Table 4.1.

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

Table 4.1. Cross validation results of aggregated models

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.

Models	C	σ	**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

Table 4.2. Paired t-test for comparisons of RMSE, MAE and R^2 for aggregated models

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. Note that * and ** means 5% and 1% significance level respectively. 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.

Panel A. RMSE

Models	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
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

Models	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
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

Models	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
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	–

4.3.2.2. 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 4.3 separately. The pair wise t-test results are presented in Table 4.4. 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 to A.3 in Table 4.4 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 4.4 Panels B.1 to B.3 and C.1 to 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 to 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.

In general it can be found that the SVR model prediction on the training set is much better than that on the testing set, where the concern of overfitting risk arises. However, given the evidence that the SVR models generally outperform the other statistical regression models it shows the overfitting issue does not impact the model performance seriously. In fact, the SVR models are less vulnerable to the overfitting risk because of the structural risk minimization principle of the support vector models. Also the out-of-time validation could be an alternative to avoid the overfitting issue.

Table 4.3. Cross validation results of segmented models

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.

Panel A. Senior secured bonds

Models	C	σ	RMSE	RMSE_sd	MAE	MAE_sd	R^2	R^2_sd
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

Models	C	σ	RMSE	RMSE_sd	MAE	MAE_sd	R^2	R^2_sd
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

Models	C	σ	RMSE	RMSE_sd	MAE	MAE_sd	R^2	R^2_sd
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

Table 4.4. Paired t-test for comparisons of RMSE, MAE and R^2 for segmented models

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. Note that * and ** means 5% and 1% significance level respectively. 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. Note that * and ** means 5% and 1% significance level respectively.

Panel A.1. RMSE on senior secured bonds

Models	M1	M2	M3	M4	M5	M8	M11
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

Models	M1	M2	M3	M4	M5	M8	M11
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

Models	M1	M2	M3	M4	M5	M8	M11
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

Models	M1	M2	M3	M4	M5	M8	M11
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

Models	M1	M2	M3	M4	M5	M8	M11
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

Models	M1	M2	M3	M4	M5	M8	M11
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

Models	M1	M2	M3	M4	M5	M8	M11
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

Models	M1	M2	M3	M4	M5	M8	M11
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	–

Panel C.3. R^2 on subordinated bonds

Models	M1	M2	M3	M4	M5	M8	M11
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	–

4.3.3.3. 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 equation (4.23). 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 RMSEi and MAEi. The combined RMSE, MAE and R2 are given as follows

$$\begin{aligned} \mathrm{RMSE}_{combined} &= \sqrt{\frac{1}{(n_1+n_2+n_3)}\sum_{i=1}^{3} n_i \times \mathrm{RMSE}_i^2} \\ \mathrm{MAE}_{combined} &= \frac{1}{(n_1+n_2+n_3)}\sum_{i=1}^{3} n_i \times \mathrm{MAE}_i \\ \mathrm{R}^2_{combined} &= \frac{1}{3}\sum_{i=1}^{3} \mathrm{R}_i^2 \end{aligned} \quad . \tag{4.23}$$

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 4.5. From Table 4.5 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 4.1 and Table 4.5 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 4.5. 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 et al, 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.

Table 4.5. Comparison of combined results of segmented models and aggregated models

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.

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

Table 4.6. 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

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 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 4.6. One of the most comparable studies was Qi and Zhao (2011) which compared six different techniques on the debt RR of MURD from 1985 to 2009 and achieved 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) proposed a beta-link generalized linear model to predict corporate bond instrument level RR from MURD from 1985 to 2008 and reported an out-of-sample R^2 of 0.6119. This is still outperformed by our proposed SVR models. Khieu et al (2012) considered the bank loans RR from MURD, and the in-sample R^2 of the models used in their study were reported to be between 0.2 and 0.3. Leow et al

(2013) obtained 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 was 0.1428 for personal loan LGD. Bellotti and Crook (2012) showed 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 was 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 was by Loterman et al (2011), which benchmarked 24 different techniques on six bank retail RR data sets and reported that the LS-SVR and neural network models consistently outperformed the other methods. They applied eight performance metrics to evaluate the model performances, and R^2 reported in this study was 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.

4.4. 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 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 parameterised on an aggregate sample, surprisingly, gives more accurate predictions than one parameterized on sub-samples. 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.

Chapter 5

Analyzing Corporate Bond Recovery Rates: An Empirical Study on the Impacts of Unobservable Firm Heterogeneity

5.1. Introduction

This Chapter studies the impacts of unobservable firm heterogeneity on recovery rates modelling for corporate bonds based on single factor models. As introduced in Chapter 2, Vasicek (1987, 2002) proposed a single latent factor framework based on Merton's model (Merton, 1974) to estimate default probabilities. Vasicek's single factor model assumes that the default contagion effect exists across debt instruments that are dependent on a common single latent factor, which is equivalent to a Probit model with the inclusion of a random effect term.

Single factor model accounts for the time-varying unobservable heterogeneity with the incorporation of the random effect term, which represents the latent economic trend as the single source of systematic risk. The single factor framework was first adapted to LGD modelling by Frye (2000a, 2000b), where the collateral value was modelled to be a function of a time-varying systematic risk factor. Dullman and Trapp (2004) estimated PD and LGD jointly under the single factor framework and found that incorporating systematic risk into recovery rates might lead to a dramatic increase in the regulatory capital but the distributional assumption of recovery rates played an insignificant role.

We build our study on previous research, in particular, on the methodology in Hamerle et al (2006) and improve the model fit under the single factor framework. We do not attempt to identify new determinants of recovery rates. Instead our primary purpose is to investigate the influences of unobservable heterogeneities on corporate bonds recovery rates by applying the latent factor at obligor, seniority and time levels.

The major contribution of this study consists in an empirical analysis of the impact of the unobservable heterogeneities on predictive accuracy of the recovery rates models and in extending the random effect, i.e., the systematic risk factor to multiple levels. Unlike the literature on LGD/RR modelling where the latent factor is assumed to be a time-varying variable representing the general economic condition, we find that by accounting for the firm specific unobservable heterogeneity the single factor model presents a dramatic improvement on both model fit and out-of-sample predictions. The empirical evidence presented in our study strongly supports the

necessity of including the obligor-varying latent factors with R^2 greater than 0.85 for both in-sample and out-of-sample predictions. This finding provides a different insight compared with literature indicating that there is significant amount of valuable accounting information about obligors that remains unobserved and can be explained by the inclusion of the random effect term. We further investigate the firm level heterogeneity by examining the intra-class correlations and the predicted latent factors, and find that the firm specific intra-class correlation is much higher than for the other random factor terms. The high intra-class correlation also suggests that the common accounting information shared by the instruments of the same issuer is largely explained by the obligor level random effect while the fixed effect regression models are not able to account for that. We also check the predicted latent obligor-varying factors by taking the average of them with respect to each year and find that the aggregated annual obligor-varying latent factors demonstrate similar patterns as the predicted time-varying latent factors. We suggest that the obligor-varying factor specification gives consistently remarkable performances at both instrument and yearly aggregated levels for recovery rates modelling.

Second, we consider a variety of distributional assumptions on recovery rates for both fixed and random effect models. We examine three different fixed effect regression models including linear regression, fractional response regression and inflated beta regression. We find that fractional response regression slightly outperforms linear regression, but in contrast to previous research, the inflated beta regression does not present any advantage compared with ordinary linear regression, which implies it is reasonable to use a linear relationship to estimate the recovery rates. Furthermore, we explore the factor models and find strong evidence in favour of a linear model with a normal distributional assumption of the recovery rates rather than the other non-linear specifications that are widely investigated in literature on LGD/RR modelling. We also examine three different non-normal specifications under the single factor framework including log-normal, logit-normal and inflated beta distributions. Besides beta distribution specification in Bruche and Aguado (2010) which considered a latent credit cycle factor, Huang and Oosterlee (2011) proposed a similar generalized beta regression model which inserted a time-varying random effect term into the linear predictor, and they found that this new setting provided a significant improvement according to the log-likelihood ratio. They also suggested that it was not necessary to reparameterize both mean and dispersion parameters like

Bruche and Aguado (2010). However, the generalized beta regression in Huang and Oosterlee (2011) was still not capable of modelling the observations at the boundaries 0 and 1, and no out-of-sample predictions were provided to show if there was any improvement on predictive accuracies. Here we follow Huang and Oosterlee (2011) and generalize the inflated beta regression by including the random effect term into the linear predictor showing that our generalized inflated beta model is able to improve the predictive accuracies compared with inflated beta regression. But it is outperformed by the other non-linear specifications. However, the empirical evidence suggests that both log-normal and logit-normal factor models suffer the same problem that the model fit is highly sensitive to the choice of the perturbation value at the boundaries making them less attractive for the lack of robustness.

We finally investigate the impact of firm level unobservable heterogeneities on portfolio loss distributions at both aggregated and segmented levels. As suggested in Basel II Accord (Basel Committee, 2005a, 2005b) we examine both AIRB and FIRB approaches to generate portfolio losses where the linear regression and single factor models are employed as AIRB models and the FIRB approach is implemented by specifying a determined value for the bonds with respect to their seniorities provided in Basel II. We find that the aggregated portfolio loss distribution generated by an obligor-varying factor model presents more skewed at the right tail than the time-varying factor and linear regression models. However, the evidence at the segmented levels is mixed. We notice that time-varying factor model presents slightly right skewed loss distributions than the obligor-varying on the senior secured and subordinated bonds but a significant left skewed distribution on the senior subordinated bonds. Another important finding is that the portfolio losses calculated by FIRB approach are significantly lower than the VaR and ES generated by the AIRB approaches. We believe that the LGD specifications of FIRB approach may underestimate the unexpected losses according to the calculated VaR and ES of AIRB approaches for the senior secured and unsecured bonds. However, for the subordinated bonds the FIRB specification provides very close performances to the AIRB models in terms of VaR and ES. The significant discrepancies of portfolio loss distributions between AIRB and FIRB approaches imply that the international regulators should review the LGD requirements of the defaulted corporate bonds to compute the regulatory capital more appropriately, and the financial institutions should be encouraged to develop their internal LGD/RR models to better manage the

credit portfolio risk.

The remainder of this chapter is structured as follows. The next section introduces the methodologies applied in the empirical study, and Section 5.3 presents the empirical evidence and analysis, and Section 5.4 analyzes the implications of this study to credit risk management. Finally Section 5.5 concludes the study.

5.2. Methodology

We first introduce and investigate the model specifications under the single factor framework and then we briefly introduce the benchmarking models used in the following empirical study.

Single factor model

The single factor model in Hamerle et al (2006) is defined as

$$\begin{array}{ll} \tilde{y}_i = \beta_0 + \mathbf{\beta}^T \mathbf{x}_i + \gamma_1 Z_t + \gamma_2 \varepsilon_i & t = 1,...,T \\ Z_t \sim N(0,1)\,, & \varepsilon_i \sim N(0,1) \end{array}, \tag{5.1}$$

where Z_t is the random effect term denoting the time-varying systematic recovery risk factor representing the unobservable heterogeneity of macroeconomic conditions and ε_i is the residual term denoting the idiosyncratic recovery risk factor. Here $\tilde{y}_i$ is the transformed recovery rate for instrument i by a logit transformation such that

$$y_i = \exp(\tilde{y}_i) \,/\, (1 + \exp(\tilde{y}_i))\,,$$

where y_i denotes the actual recovery rate. This model can be linked with the specification in Dullmann and Trapp (2004) such that

$$\tilde{y}_i = \beta_0 + \mathbf{\beta}^T \mathbf{x}_i + \sigma\sqrt{\rho} Z_t + \sigma\sqrt{1-\rho}\varepsilon_i\,, \tag{5.2}$$

where σ is the standard deviation of the recovery rates and ρ is the loading factor denoting the recovery rates correlation between two instruments i and j that default in the same year.

Model (5.2) is based on the assumption that the instruments are dependent on a common systematic economic state with respect to the year they default. In this study, we generalize this assumption by specifying that recovery rates of instruments depend on the systematic risk factor with respect to their corresponding obligor or seniority as well as default year. Hamerle et al (2006) found that the inclusion of observable macroeconomic variables rendered systematic risk factor less important statistically. In this study we simply assume recovery rate to follow a normal distribution instead of a logit-normal proposed by Dullmann and Trapp (2004) and Hamerle et al (2006),

and we model the actual recovery rate y_i directly. We show that the normal distributional assumption is more suitable than the other non-normal assumptions in the following empirical analysis. We define the obligor and seniority-varying random effect models by substituting the time-varying factor Z_t with Z_k, the *k*-th obligor of the total of K obligors and Z_s, the *s*-th type of the total of S seniorities as follows.

$$\tilde{y}_i = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i + \sigma\sqrt{\rho} Z_k + \sigma\sqrt{1-\rho}\varepsilon_i$$

$$\tilde{y}_i = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i + \sigma\sqrt{\rho} Z_s + \sigma\sqrt{1-\rho}\varepsilon_i$$

Note that different from the Merton-like single factor models, the above models also include the observable covariates formulated to be mixed effect models. The cluster intra-class recovery rate correlation of any two instruments is given as

$$\begin{aligned} Cov(y_i, y_j) &= \sigma^2\rho = \gamma_1^2 \\ Corr(y_i, y_j) &= \rho = \frac{\gamma_1^2}{\gamma_1^2 + \gamma_2^2} \end{aligned} . \tag{5.3}$$

The estimation procedure of this model starts by deriving the conditional probability density function, and then the unconditional density function can be obtained by integrating out the random effects. Conditioning on the realization of Z, y_i follows a normal distribution such as

$$f(y_i \mid Z) = \frac{1}{\sqrt{2\pi}\sigma_Z} \exp\left(\frac{(y_i - \mu_Z)^2}{2\sigma_Z^2} \right),$$

where

$$\begin{aligned} \mu_Z &= E(y_i \mid Z) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i + \gamma_1 Z \\ \sigma_Z &= \gamma_2 \end{aligned} .$$

The unconditional probability density function of y_i is then given as the product of the conditional probability density function and the marginal density function of Z such that

$$f(y_i) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_Z} \exp\left(\frac{(y_i - \mu_Z)^2}{2\sigma_Z^2} \right) d\Phi(Z) ,$$

where $\Phi(Z)$ denotes the cumulative standard normal distribution function. At the final stage the log likelihood function is given by

$$\log L = \sum_i \log f(y_i) .$$

The estimates of the parameters are generated by solving the log likelihood function with standard optimization algorithms. Because of the integral involved in the

likelihood function, Gaussian quadrature approximation is adopted before optimizing it.

Apart from the normal and logit-normal distributions introduced above, other specifications are also investigated in LGD/RR modelling. For example, Pykhtin (2003) proposed to estimate the loss distribution of corporate bonds with a log-normal distributional single factor model, where the recovery rate was modelled to be dependent on a single systematic risk factor and two idiosyncratic risk factors in order to strengthen the correlation between PD and LGD. Another two factor specification was Hillebrand (2006) which aimed to incorporate the PD and LGD correlation to estimate the loss distribution more accurately by assuming LGD to depend on two different systematic risk factors with a Probit link function applied. Our study focuses on modelling recovery rates instead of exploring the PD and LGD relationship, and we suggest the single factor specification is suitable to predict the recovery rates alone.

Benchmarking models

Three linear and generalized linear models are selected as the benchmarking models including ordinary linear regression, fractional response regression and inflated beta regression. The first two methods have been extensively investigated in LGD/RR modelling. Linear and generalized linear models are commonly used to identify the effects of the determinants of recovery rates. For example, Acharya et al (2007) investigated the main drivers of the recovery risk from the standpoint of the industry-equilibrium theory with a linear regression model. Khieu et al (2012) studied the ultimate recovery rates of bank loans and found that debt characteristics had more significant influences on recovery rates than the borrower characteristics. Fractional response regression is also widely considered in LGD modelling because it defines the dependent variable to be bounded in the open interval between 0 and 1 by imposing a link function such that:

$$E(y \mid \mathbf{x}) = G(\boldsymbol{\beta}^T \mathbf{x}), \tag{5.4}$$

where $G(\bullet)$ denotes some link function. We use a logit transformation link function such as

$$G(\boldsymbol{\beta}^T \mathbf{x}) = \exp(\boldsymbol{\beta}^T \mathbf{x}) \,/\, (1 + \exp(\boldsymbol{\beta}^T \mathbf{x})). \tag{5.5}$$

The inflated beta regression was proposed by Ospina and Ferrari (2010) to fit the fractional response variables, where dependent variable is defined over the interval [0, 1] that can be regarded as a mixture distribution of a beta distribution on (0, 1) and a

Bernoulli distribution on bounds 0 and 1. The probability density function is given as

$$bi_{01}(y;\pi,\psi,\mu,\phi) = \begin{cases} \pi(1-\psi) & if\ y=0 \\ \pi\psi & if\ y=1 \\ (1-\pi)f(y;\mu,\phi) & if\ y\in(0,1) \end{cases}. \tag{5.6}$$

The beta density function $f(y;\mu,\phi)$ is defined as

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},$$

where μ and ϕ are the mean and precision parameters. To formulate the inflated beta regression model, μ and ϕ can both be reparameterized by link functions such that $\mu = G(\mathbf{x}_\mu)$ and $\phi = H(\mathbf{x}_\phi)$. In this study we only reparameterize μ for comparison purpose. Because of $\mu\in(0,1)$ the link function $G(\mathbf{x}_\mu)$ is chosen to be a logit transformation such as specification (5.5). This model can also be written in an integrated form as

$$bi_{01}(y;\pi,\psi,\mu,\phi) = (\pi(1-\psi))^{\delta}(\pi\psi)^{c}((1-\pi)f(y;\mu,\phi))^{(1-\delta)(1-c)}, \tag{5.7}$$

where $\delta=1$ if $y=0$ and $\delta=0$ if $y\in(0,1]$, $c=1$ if $y=1$ and $c=0$ if $y\in[0,1)$. The expectation of the dependent variable is derived immediately such that

$$E(y) = \pi\psi + (1-\pi)\mu .$$

The estimates are then derived by maximizing the likelihood function based on (5.7). More details on fractional response and inflated beta regression models can be referred to Papke and Wooldrige (1996) and Ospina and Ferrari (2010).

The empirical study in Calabrese (2014) showed that the inflated beta regression model improved LGD predictive accuracies compared with linear regression models, and effectively captured the bi-modal distribution of recovery rates between 0 and 1 as well as accommodates the concentration of cases at boundaries 0 and 1. In the following empirical analysis, we show that the inflated beta regression just provides a marginal advantage over the linear regression, but the linear factor models outperform the inflated beta mixed model significantly.

5.3. Empirical results and analysis

The empirical study in this chapter is also based on the loss data of corporate bonds described in Chapter 3. We estimate four models including a single factor model and three other benchmarking fixed effect models. We report the parameter estimates as well as the goodness-of-fit for all models on the whole sample. For the

single factor model we examine the random effect specified at three levels: obligor, seniority and time. Table 5.1 exhibits the estimates of parameters for all covariates as well as the measures of model fit of each model discussed above, and in addition, the estimates of the intra-correlation for the factor model are also included.

A. Covariates interpretation

Instrument characteristics: First, it is noticed that all of the models give negative signs on the parameters related to the variables *Collateral Rank* and *Percent Above*. This is expected and is consistent with our previous analysis. The estimate of parameter *Collateral Rank* suggests that if the collateral of an instrument is downgraded by one grade, the related recovery rate is expected to decrease by 0.1 on average, implying that *Collateral Rank* plays a remarkably influential role in the model. *Percent Above* interprets the relative seniority in the recovery process of an instrument in a similar way with *Collateral Rank*, where a greater *Percent Above* indicates more debt should be recovered prior to recovering the current instrument. All the models demonstrate a significant negative sign of *Percent Above*, which is consistent with our ex ante hypothesis. Note the magnitude of *Percent Above* is nearly -0.25 for the linear regression, indicating a very strong negative influence on recovery rate. In other word, an upward change of 0.1 for *Percent Above* is expected to lead a decrease of more than two percentage points' change for the recovery rate on average. For the last instrument characteristic *Issue Size*, we notice that all the linear models show a negative sign except for the inflated beta regression. Note that both fractional response and inflated beta regression models give mixed but insignificant signs, and a clear-cut negative linear relationship with bond recovery rate should be recognized which conflicts to the findings in Acharya et al (2007). Although the empirical study in Khieu et al (2012) showed that *Issue Size* had a negative effect on recovery rate. However, the estimate of parameter is not significant statistically. We suggest that the difficulty for banks to foreclose the large size debts places higher influences than the bargaining power of the issuer and subsequently results in a negative effect on recovery rate.

Firm characteristics: All the models give a positive sign for the parameter *Total Asset* which conforms to the argument that restructuring follows default is processed more quickly by large companies leading to a higher recovery rate than by small companies. According to Khieu et al (2012) creditors tended to trust and to accept a restructuring plan from stockholders with more transparent information indicating

more advantages for the large firms. For the *EBITDA* only inflated beta regression presents a negative sign although it is not statistically significant. This coincides our expectations that a firm with better earning ability should be able to yield higher a recovery rate for its instruments. Table 5.1 shows that linear regression gives a significant positive sign for the parameter on *Leverage* and all the other models also confirm a positive effect on the recovery rate although not significant. Such evidence requires further investigations to explain the influences of firm debt structure. Next variable *Debt ratio* presents very controversial results in our study. A higher *Debt ratio* indicates the short term creditors dominate in the obligees, and the short term obligees would prefer to withdraw their funds immediately rather than an extension. Therefore, it is reasonable to observe a negative relationship between *Debt ratio* and recovery rate. According to Table 5.1 all models except linear regression present significant negative parameter estimate, which implies that a linear regression may misinterpret the effect of *Debt ratio*. For *Book Value per Share* only obligor-varying factor model gives a significant negative sign while the others all demonstrate insignificant estimates. This conflicts to our anticipation that a company with a higher *Book Value per Share* may result in a higher recovery rate suggesting that further examination is needed. As expected *Asset Tangibility* exhibits positive signs for all models and is highly significant statistically according to the obligor-varying factor model. *Quick Ratio* appears to be insignificant in all models and shows controversial signs. The unexpected mixed results may indicate that *Quick Ratio* is not an important determinant of recovery rates.

Macroeconomic variables: We find *Growth Rate* is positively correlated with recovery rate while *Default Rate* has a strongly negative influence as observed in Figure 3.13. *T-Bill Rate* also exhibits an expected significant negative sign. It conforms to economic intuition that a lower interest rate reduces the cost for an obligor to refinance and restructure its debt leading to a higher recovery rate. The only unexpected result is the significant positive sign of the parameter on *Unemployment Rate*. Economic intuition suggests that a higher *Unemployment Rate* means a more depressed economy where the recovery rate tends to be lower. Considering this variable is rarely used in previous research, we believe it might be better explained in further study.

Table 1 shows that the inclusion of random effect may change the sign of parameter estimates (e.g., Debt Ratio). The correlation of the above selected variables

may potentially affect the significance of the parameter estimates rather than the sign, and thus it is not a major concern in this research.

Table 5.1. Estimates of parameters

Table 5.1 demonstrates the estimated results of regression models applied to the whole sample data. Both Issue Size and Total Asset are subjected to a log transformation. Here *, ** and *** represent significance at 10%, 5% and 1% level respectively, and t-values of the estimated parameters are reported in parentheses. Note that t values and their corresponding significance levels are not consistent across models due to the change of degree of freedom. σ and ρ are the estimates of recovery rates volatility and intra-class correlation.

	Single factor models			**Linear regression**	**Fractional response regression**	**Inflated beta regression**
	Obligor	**Seniority**	**Time**			
Intercept	0.2768 ***	0.3842 *	0.7867 ***	0.4116 ***	-0.5989	-1.4478
	(2.64)	(2.41)	(3.24)	(2.72)	(-0.56)	(-1.54)
Collateral Rank	-0.1698 ***	-0.1199 ***	-0.1173 ***	-0.1204 ***	-0.6320 ***	-0.2199 ***
	(-19.09)	(-9.12)	(-8.94)	(-9.13)	(-6.37)	(-4.46)
Percent Above	-0.1113 ***	-0.2442 **	-0.2525 ***	-0.2505 ***	-1.2228 ***	-0.9410 ***
	(-3.88)	(-5.63)	(-6.17)	(-6.11)	(-4.18)	(-5.84)
Log(Issue Size)	-0.0168 ***	-0.0137 *	-0.0154 **	-0.0151 **	-0.0645	0.0219
	(-5.04)	(-2.22)	(-2.61)	(-2.53)	(-1.50)	(0.71)
Log(Total Asset)	0.0696 ***	0.0599 ***	0.0564 ***	0.0597 ***	0.2916 ***	0.1208 ***
	(14.35)	(9.24)	(8.63)	(9.24)	(6.15)	(4.92)
EBITDA	0.00001	0.00005 **	0.0001 ***	0.00004 ***	0.0003 **	-0.0001
	(1.56)	(3.67)	(3.61)	(3.61)	(2.54)	(-0.90)
Leverage	0.0322	0.0488	0.0214	0.0488 **	0.2403	0.0470
	(1.27)	(1.99)	(0.86)	(1.98)	(1.35)	(0.17)
Debt Ratio	-0.0008 *	-0.0033 **	-0.0031 ***	0.0009 ***	-0.0177 **	-0.0120 ***
	(-2.38)	(-2.99)	(-2.85)	(-3.02)	(-2.30)	(-2.52)
Book Value per Share	-0.0025 ***	0.0095	0.0006	-0.0033	0.0036	0.0057
	(-4.24)	(1.27)	(0.83)	(1.20)	(0.65)	(1.46)
Asset Tangibility	0.0670 ***	0.0007	0.0020	0.0004	0.0010	0.0274
	(4.15)	(0.04)	(0.11)	(0.02)	(0.01)	(0.15)
Quick Ratio	0.0173	-0.0103	-0.0007	-0.0105	-0.0556	-0.0337
	(1.23)	(-0.74)	(-0.05)	(-0.76)	(-0.57)	(-0.62)
Growth Rate	1.7795 **	0.0207 **	0.0124	2.0750 **	0.0961	0.0622
	(2.55)	(2.28)	(0.03)	(2.28)	(1.48)	(0.18)
T-Bill Rate	-0.0160 **	-0.0466 ***	-0.0406 *	-0.0467 ***	-0.2381 ***	-0.1813 ***
	(-2.63)	(-7.50)	(-1.98)	(-7.49)	(-5.25)	(-5.85)
Default Rate	-0.0115 **	-0.0416 **	-0.0339	-0.0413 ***	-0.1953 ***	-0.1228 ***
	(-2.12)	(-4.50)	(-1.69)	(-4.46)	(-3.04)	(-4.30)
Unemployment Rate	0.0538 ***	0.0731 ***	0.0117	0.0736 ***	0.3866 ***	0.2085 ***
	(5.68)	(6.80)	(0.32)	(6.87)	(4.77)	(3.94)
σ	0.3422	0.3186	0.3249	0.3177		
ρ	0.8208	0.0009	0.1214			
-2loglikelihood	-497.8	769.2	698.4	769.5	1625.7	1210.4
AIC	-463.8	803.2	732.4	801.5	1655.7	1302.4
BIC	-396	796.6	754.5	885.5	1734.5	1544.1
R^2	0.8964	0.3383	0.4029	0.3409	0.3628	0.3558

B. Goodness of fit

Table 5.1 demonstrates that the obligor-varying single factor model fits the data much better than the other methods. We notice that the obligor-varying factor model yields an outstanding model fit with the R^2 of 0.8964. Meanwhile the time-varying factor model also presents better model fit than the other benchmarking regression models. However, the seniority-varying factor model does not demonstrate any improvement in terms of the measure of fit.

We suggest that the improvements of model fit for the single factor model are caused by the inclusion of a random effect which effectively explains unobservable heterogeneity. Notice that the obligor specific intra-class correlation ρ is 0.8208, which is significantly higher than the seniority and time specific levels. It emphasizes that the instruments issued by the same company share a large amount of unobservable common characteristics represented by the high intra-class correlation, and the inclusion of an obligor-varying random effect explains such variations with a significant improvement of model fit. In contrast, the seniority specific intra-class correlation is rather small. One reason might be that the seniority specific unobservable heterogeneity has already been partially explained by the observable instrument level characteristics such as the *Collateral Rank*.

Among the fixed effect regression models, fractional response regression shows slightly better model fit compared with the ordinary linear regression, which is consistent with the findings in Qi and Zhao (2011) and Khieu et al (2012). It is also noticed that the inflated beta regression model only shows marginal advantage over the linear regression model, which conflicts with the findings in Calabrese (2014). One possible explanation is that although the inflated beta regression model accommodates modelling the clustered samples on the boundaries 0 and 1, it is not able to represent values between extremes accurately. Another possible reason is that the beta distribution can not fit our sample well. Also the model performance might be further improved if the dispersion parameter is reparamerised. In fact it is not unexpected to observe the relatively disappointing performances according to Qi and Zhao (2011), suggesting that the bi-modal distribution should be of secondary concern in LGD modelling. They find that using a beta transformation does not necessarily render a better model fit for linear regression model.

C. Unobservable heterogeneities

Our study shows strong evidence for the presence of instrument and

macroeconomic characteristics of all specifications. Khieu et al (2012) also reported that loan characteristics were more significant than borrower characteristics in general for the bank loans recovery rates. Another finding is that with the presence of a time-varying random effect, the macroeconomic variables become less significant indicating that the inclusion of time-varying latent factors weakens the importance of the observable macroeconomic covariates. However, it exhibits mixed results for the change of significance of accounting ratios by comparing the estimates of obligor-varying factor model and linear regression model according to Table 5.1. In the following we further investigate the influences of unobservable heterogeneities by examining the recovery rates intra-class correlation of single factor models.

The estimates of intra-class correlation and volatility of single factor models are exhibited in Table 5.1. First, it is clear that the firm specific intra-class correlation is significantly higher than that of the seniority and time specific levels. We suggest the instruments of the same issuer are highly correlated and the unobservable firm level information is effectively explained by the obligor-varying random effect. It is also straightforward to interpret the low correlation at the seniority level because the instruments with the same seniority do not necessarily demonstrate many common debt characteristics. Furthermore, instruments that defaulted in the same year can be considered to experience the same economic conditions, where the recovery rate correlation is higher than that at the seniority level but still much lower than that at the obligor level. Notice that the estimated volatility of the obligor-varying factor model is 0.3422, which underestimates the historical volatility of 0.3915 but it is the best estimate compared with the other random effect models.

To examine if the observable covariates can be completely replaced by the random effect defined at the same level, we perform an additional test on three restricted single factor models: the obligor-varying random effect is included with the firm accounting ratios excluded; the seniority-varying random effect is included with the instrument characteristics excluded; and the time-varying random effect is included with the macroeconomic variables excluded. The single factor models with all covariates included are referred to unrestricted models. The estimates of the restricted models are reported in Table 5.2. We notice that both the magnitudes and signs of the estimates of restricted models are not significantly changed compared with the unrestricted models, indicating that the estimates of factor models are robust enough. It is shown that the obligor-varying random effect can almost replace the

effects of observable firm characteristics, because the inclusion of observable firm accounting ratios does not provide any significant improvements in the measure of fit including AIC, BIC and R^2. Our test suggests that it is sufficient to explain the unobservable firm characteristics by accounting for the firm specific heterogeneity which makes observable accounting ratios almost replaceable. We also find similar evidence for the restricted time-varying factor model, which exhibits very close model fit to the related unrestricted model. The restricted time-varying factor model also shows that the economic cyclical effects can be sufficiently explained by the latent factors. In contrast, the inclusion of a seniority-varying random effect can not replace the observable debt characteristics where the model fit measurements deteriorate significantly when the instrument characteristics are excluded.

Table 5.2. Estimates of the restricted single factor models

Table 5.2 presents the estimated results of three restricted single factor models. Obligor: the obligor-varying model with firm characteristics excluded; Seniority: the seniority-varying model with instrument characteristics excluded; Time: the time-varying model with macroeconomic variables excluded. Here *, ** and *** represent significance at 10%, 5% and 1% level respectively, and t-values of the estimated parameters are reported in parentheses.

	Restricted single factor models		
	Obligor	**Seniority**	**Time**
Intercept	0.9485 ***	-0.6315***	0.6854***
	(10.12)	(-4.81)	(6.07)
Collateral Rank	-0.1868 ***		-0.1162 ***
	(-21.64)		(-8.98)
Percent Above	-0.0576 **		-0.2544 ***
	(-2.09)		(-6.33)
Log(Issue Size)	-0.0165 ***		-0.0154 **
	(-5.43)		(-2.62)
Log(Total Asset)		0.0475 ***	0.0556 ***
		(7.12)	(8.72)
EBITDA		0.00003	0.00005 ***
		(2.10)	(3.79)
Leverage		0.0308	0.0199
		(1.17)	(0.83)
Debt Ratio		-0.0034 **	-0.0030 **
		(-2.94)	(-2.77)
Book Value per Share		0.0027 **	0.0006
		(3.24)	(0.76)
Asset Tangibility		-0.0324	0.0045
		(-1.67)	(0.25)
Quick Ratio		0.0125	0.0007
		(0.85)	(0.02)
Growth Rate	0.8974	3.7624 **	
	(1.30)	(3.86)	
T-Bill Rate	-0.0169 ***	-0.0359 ***	
	(-3.75)	(-5.42)	
Default Rate	-0.0151 **	-0.0168	
	(-2.10)	(-1.73)	
Unemployment Rate	0.0346 ***	0.1036 ***	
	(3.86)	(9.04)	
σ	0.3408	0.3724	0.3433
ρ	0.8166	0.1628	0.2143
-2loglikelihood	-458.4	987.4	706.9
AIC	-438.4	1015.4	732.9
BIC	-398.5	1009.9	749.8
R^2	0.8948	0.2413	0.4043

Furthermore, to check statistical evidence on the inclusions of latent factors we use the Bayes factor by following the method adopted in Duffie et al (2009), which is represented as the twice differences of the log likelihood between the model with random effect (single factor model) and the null model (ordinary linear regression model). According to the literature cited in Duffie et al (2009), a value of Bayes factor

greater than 2 indicates positive strong evidence to include the random effect, and a value greater than 10 indicates very strong evidence. We find that the Bayes factor is 1267.3 for the obligor-varying single factor model and 71.1 for the time-varying model, which provides very strong evidence in favour of including the obligor and time specific latent factors. However, the seniority-varying factor model shows a Bayes factor of 0.3, which indicates it unnecessary to include the random effect at the seniority level. Such evidence further explains why the model fit can be improved when the obligor and time-varying random effects are included.

Finally we examine the patterns of the predicted latent factor values. The predicted time-varying latent factors are generally considered to represent the latent economic cycles. To explain the obligor-varying latent factors more intuitively, we aggregate the predicted values by taking the average of the obligor-varying latent factors with respect to each year in default and present them with the predicted time-varying latent factors together. We observe very similar movement tendencies of both in Figure 5.1. It implies that the unobservable firm level heterogeneity already contains the latent time-varying economic conditions. In other words, the latent economic cycle has been effectively accounted for by the obligor-varying latent factors. We also consider the yearly aggregated recovery rates estimated by the factor models as further evidence. We take the yearly average of the recovery rates of the instruments obtained from both obligor-varying and time-varying single factor models and plot them against the historical annual aggregated recovery rates exhibited in Figure 5.2. We observe an excellent fit of the aggregated annual recovery rates for both the obligor-varying and the time-varying factor models. The pattern presented in Figure 5.2 suggests that at the aggregated level both obligor-varying and time-varying factor models provide equally good model fit. But obligor-varying factor model performs considerable advantages over the others for estimating the instrument level recovery rates.

Figure 5.1. Plot of predicted time-varying and year aggregated obligor-varying latent factors

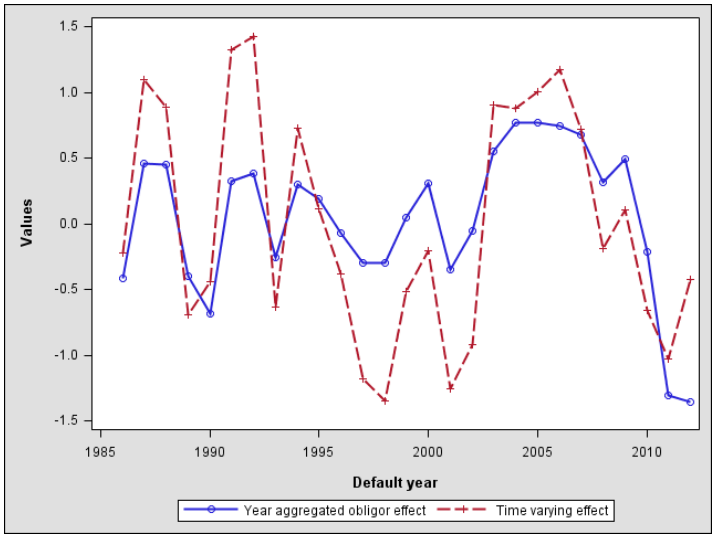
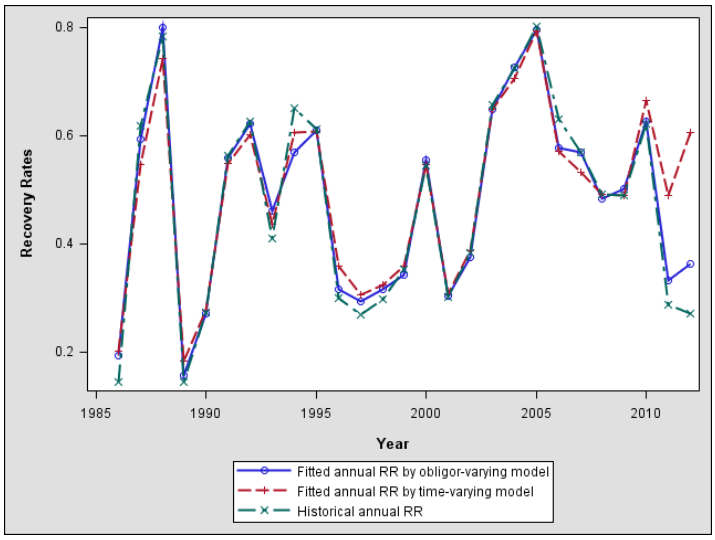


Figure 5.2. Plot of estimated recovery rates of time-varying and obligor-varying factor models



D. Out-of-sample prediction

We check the out-of-sample predictions of the above models in terms of R^2, MAE and RMSE. First we randomly generate a hold-out sample from the whole sample data. However, given that the latent factors are unobservable in the testing set for the factor model, it is necessary to design the experiment more carefully. Note that the individuals of the same group sharing the same latent factors. For example, if there are two instruments i and j issued by the same obligor k, we select the samples such that instrument i enters into the training set while instrument j is included in the testing set. So suppose we have fitted a linear mixed effects model on the training set, then the systematic risk factor u_k that has been estimated corresponding with

instrument i can be applied to predict the recovery rate of instrument j in the hold out sample. In other words, we make sure that any instrument in the testing set should have an instrument issued by the same obligor selected in the training set. We apply this rule also to seniority and time strata.

To summarize, we randomly divide all the samples into training and testing sets by a stratified sampling method. The strata are defined at obligor, seniority and time levels to be consistent with the random effect definitions. At each stratum approximately 70 percent of the observations are selected into the training set and the remaining observations are placed in the testing set. The summary statistics for the training and testing sets for different strata are given in Table 5.3, and the in-sample and out-of-sample predictions are reported in the Table 5.4.

For the benchmarking regression models, we notice that at obligor level stratum fractional response regression gives the highest R^2 of 0.4636 and the lowest RMSE and MAE of 0.2956 and 0.2469 respectively. It also outperforms the other two models at the time level stratum although the R^2 is down to 0.3249. The inflated beta regression model presents marginal advantages at seniority level, and obtains the highest R^2 of 0.3701 and the lowest RMSE of 0.3093. Another interesting finding is that the out-of-sample predictive performances at the obligor level stratum are significantly better than the other strata. This may suggest that when the data are sampled with respect to obligor stratum, the instruments in the holdout sample may share the same accounting information with the instruments of the same obligor in the training set. Therefore the model fitted on the training set should give more accurate predictions because the fitted model can be regarded to have obtained more prior knowledge of the testing set samples. Such evidence is demonstrated more clearly in the single factor models. Among the single factor models it is clear to see that the obligor-varying model outperforms the others substantially. It can be noticed that the time-varying factor model also gives better predictive accuracies than the benchmarking regression models on the holdout sample. This is consistent with the evidence of model fit, which further confirms that the inclusion of an obligor-varying random effect gives advantages in modelling instruments recovery rates. It also shows that single factor models are robust with similar performances presented on both the training and testing sets.

Table 5.3. Settings of training and testing sets

Table 5.3 shows the settings of training and testing sets for out-of-sample prediction. The training and testing sets are divided based on stratified sampling method with strata defined at obligor, seniority and time levels. At the obligor stratum, any instrument in the testing set should have an instrument issued by the same obligor selected in the training set. This rule is also applied to seniority and time strata.

	Training Set		Testing Set	
Strata	**Obligors**	**Instruments**	**Obligors**	**Instruments**
Obligor	398	1037	144	376
Seniority	352	991	196	422
Time	356	1002	197	411

Table 5.4. Out-of-sample prediction performances

Table 5.4 presents both in-sample and out-of-sample prediction performances including three performance metrics: R^2, RMSE and MAE. For single factor models, the sample strata are consistent with the random effect levels for out-of-sample predictions.

	Sampling Strata	In-sample			Out-of-sample		
		R^2	RMSE	MAE	R^2	RMSE	MAE
Single factor model	Obligor	0.8942	0.1256	0.0866	0.8667	0.1473	0.0981
	Seniority	0.3617	0.3119	0.2548	0.2855	0.3320	0.2744
	Time	0.4130	0.3010	0.2452	0.3444	0.3134	0.2568
Linear regression model	Obligor	0.2918	0.3249	0.2725	0.4456	0.3004	0.2524
	Seniority	0.3553	0.3141	0.2600	0.2884	0.3293	0.2714
	Time	0.3525	0.3162	0.2613	0.2992	0.3240	0.2670
Fractional response regression model	Obligor	0.3083	0.3211	0.2669	0.4636	0.2956	0.2469
	Seniority	0.3789	0.3083	0.2490	0.3184	0.3223	0.2620
	Time	0.3749	0.3106	0.2521	0.3249	0.3181	0.2578
Inflated beta regression model	Obligor	0.3014	0.3225	0.2743	0.4390	0.3030	0.2620
	Seniority	0.3312	0.3205	0.2678	0.3701	0.3093	0.2650
	Time	0.3629	0.3136	0.2614	0.3053	0.3227	0.2694

E. Non-normal distributional factor models

We have discussed the linear factor models based on the normal distributional assumption of recovery rates. To investigate the robustness of this assumption we compare goodness of fit of models based on alternative distributional assumptions about the single factor recovery rates models presented above. We consider three distributional assumptions here: log-normal, logit-normal and inflated beta distribution. The log-normal and logit-normal distributional recovery rates factor model were proposed by Pykhtin (2002) and Dullmann and Trapp (2004), and the specifications are given as follows.

Log-normal:

$$\ln(y_i) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i + \gamma_1 Z + \gamma_2 \varepsilon_i . \tag{5.8}$$

Logit-normal:

$$\ln(\frac{y_i}{1-y_i}) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i + \gamma_1 Z + \gamma_2 \varepsilon_i . \tag{5.9}$$

The estimation procedure is similar with the linear single factor model, where the conditional marginal distribution of recovery rate can be obtained by using the change-of-variable technique.

Additionally we formulate the generalized inflated beta model by inserting the random effect term into the linear predictors of the mean parameter as

$$\mu_Z = G(\mathbf{x}_\mu + \lambda Z), \tag{5.10}$$

where Z is the random effect term defined as a standard normal variable with the corresponding scale parameter λ. Specification (5.10) follows the study in Huang and Oosterlee (2011), where beta regression was generalized by including the random effect term into the predictor of expectation μ_Z. They suggested that the inclusion of a random effect improved the beta regression model significantly according to the log-likelihood ratio test. However, they did not include any observable covariates in their empirical study. In this research we estimate the generalized inflated beta regression model by adopting a similar method to Huang and Oosterlee (2011), where the marginal likelihood can be derived by using the conditional probability density function and integrating out the random effect.

Note that under the log-normal specification RR=0 is undefined and for the logit-normal specification, both RR=0 and RR=1 are undefined. Therefore, a small positive perturbation value τ is applied to transform the 0 and 1 to τ and $1-\tau$ in the implementation of these two models. However, it is rather tricky to select the optimal τ for the transformation. Qi and Zhao (2011) has conducted a detailed experiment to investigate sensitivities of τ to both in-sample and out-of-sample performances from 1e-11 to 0.5, and they find that the inverse Gaussian regression presents the best in-sample and out-of-sample predictive accuracies when $\tau = 0.05$. We argue that the selection of τ should not influence the distribution of recovery rates. The 10-th percentile in our data set is 0.0055. We choose 0.001 as the optimal value because when τ becomes smaller, the fitted recovery rates deviate from the actual values dramatically. The R^2 of the non-linear single factor models are reported in Table 5.5. We find that all the models demonstrate the best performances when an

obligor-varying latent factor is specified, and the logit-normal factor model gives a comparable R^2 with linear factor models. But with the random effect specified at seniority or time level, none of these models presents advantages over the benchmarking fixed effect models. In fact we find that both log-normal and logit-normal factor models are highly sensitive to the choice of τ implying their unreliable performances. For the generalized inflated beta models, the obligor-varying factor model gives the best model fit while the seniority-varying and time-varying models do not present any improvements. The empirical evidence suggests that none of the non-linear factor models outperforms the linear factor model. This is strongly in favour of a linear relationship under the normal distributional assumption for the recovery rate model.

Table 5.5. Model fit of non-linear single factor models

Table 5.5 shows the model fit of three non-linear single factor models. Three different non-normal distributional assumptions are considered to including log-normal, logit-normal and inflated beta distributions. The random effect is specified at obligor, seniority and time levels as above. Note that a small positive perturbation value τ is applied to transform the 0 and 1 to τ and $1-\tau$ in the implementation of log-normal and logit-normal factor models because the boundary points are not defined. We find that the model fit of these two models are highly sensitive to the choice of τ, and we choose 0.001 as the optimal value. R^2 is reported as the measure of model fit.

	Random effect	R^2
Log-normal	obligor	0.4955
	seniority	0.0122
	time	0.1160
Logit-normal	obligor	0.8151
	seniority	0.0524
	time	0.1132
Inflated beta	obligor	0.5994
	seniority	0.3309
	time	0.3726

5.4. Implications for credit risk management

We investigate the impacts of recovery rate models on the portfolio risk and consider three linear models including obligor-varying and time-varying single factor models as well as ordinary linear regression that may be used as Advanced Internal Rating Based Approach (AIRB) models to estimate the recovery rates. We simulate the loss distributions based on the AIRB models and compare the characteristics of the loss distributions generated by AIRB approaches by examining the loss distribution characteristics with the Foundation Internal Rating Based approach (FIRB). The implementation procedure is defined as follows.

1) Fit the models on the whole dataset, and collect the parameter estimates for the covariates. For the single factor model, sample the single systematic risk factor Z and the residual term ε_i from independent standard normal distributions. For the linear regression, sample the residual term ε_i from a normal distribution $\varepsilon_i \sim N(0, \sigma_{ols}^2)$ where σ_{ols} is the OLS estimate of volatility. Use the parameters estimated in step 1) to calculate the simulated recovery rates for instrument i. For the instrument i the simulated recovery rate $\hat{y}_i$, the related simulated LGD is given by $\hat{\text{LGD}}_i = 1 - \hat{y}_i$.

2) Set the default indicator d_i as $d_i = 1$ for the instrument that defaulted, and assume the exposure at default (EAD) of all instruments equals 1 for simplicity. Calculate the loss rates at the m-th iteration as $L_m = \frac{1}{N}\sum_{i=1}^{N} d_i \hat{\text{LGD}}_i$ since all the instruments in our sample have defaulted.

3) Repeat the above procedures M times and formulate a simulated loss rates distribution.

Here we consider three characteristics including Value-at-Risk (VaR), Expected Shortfall (ES) and Expected Loss (EL) where the last one is defined as the average of loss rates. The definitions of VaR and ES are given in Appendix A. According to the Basel II Accord (Basel Committee, 2005a, 2005b), under the FIRB approach the LGD of senior unsecured bond is assigned as 0.45 and the subordinated bond is assigned a value of 0.75. Considering there are five different seniorities in our sample data, we merge the three types of bonds “Junior subordinated bond”, “Subordinated bond” and “Senior subordinated bond” as a general type “Subordinated bond” for simplicity, and we keep the other two types “Senior secured bond” and “Senior unsecured bond”. The descriptive statistics of LGD with respect to the new categories are given in Table 5.6. For the LGD of senior secured bonds, banks need to calculate the exposure value after risk mitigation which is not available in MURD. Therefore, we use the historical average LGD of the senior secured bonds in our sample which is 0.3708 based on Table 5.6. We examine the portfolio loss distributions at both aggregated and segmented levels, where the aggregated portfolio refers to the whole sample and the segmented portfolio is given by segmenting the whole sample with respect to the seniorities defined above.

Figure 5.3 shows the comparisons of loss distributions at both aggregated and

segmented levels, and the related estimates of VaR, ES and EL are reported in Table 5.7. Note that the LGD is a determined value for each instrument under the FIRB approach. The loss rates calculated by the FIRB approach for each segment are just the same as the regulatory values and the aggregated portfolio loss rate is 0.5163. Meanwhile notice that VaR and ES of loss distributions given by FIRB are generally lower than that of the other AIRB approaches except for the subordinated bonds. We suggest that the FIRB approach may underestimate the extreme losses under a serious economic downturn. The loss distributions generated by linear regression are more concentrated than those of the linear factor models implying that the linear regression model is unable to capture the tail losses, which is clearly undesirable. For the aggregated portfolio the obligor-varying factor model obtains a more right skewed distribution than that of the time-varying factor model according to Panel A of Figure 5.3. Panel A of Table 5.7 also shows that at both 0.05 and 0.01 levels the obligor-varying factor model yields a higher VaR and ES, suggesting that there are more extreme losses discovered by the obligor-varying model under a severe economic downturn. Similar evidence is found for the senior unsecured bonds, where obligor-varying model generate a significant higher frequency of tail losses. However, for both senior secured and subordinated bonds, the time-varying model gives greater values of VaR and ES than the obligor-varying model although the differences are not very significant. We also find that the LGD specification of the subordinated bonds under FIRB approach is 0.75, which is quite close to the VaR and ES calculated by the AIRB models. We suggest that the LGD specification of subordinated bonds under the FIRB approach is reasonable to buffer the potential unexpected losses, but for the bonds of higher seniorities including senior secured and unsecured bonds FIRB approach may underestimate the tail risk according to the VaR and ES given by AIRB approaches.

Table 5.6. Summarized statistics of LGD for aggregated and segmented portfolios

Table 5.6 presents the summarized statistics of LGD for aggregated and segmented portfolios. The aggregated portfolio is represented by the whole sample and segmented by seniority. The three types of bonds "Junior subordinated bond", "Subordinated bond" and "Senior subordinated bond" are merged as a general type "Subordinated bond" for simplicity, and the other two types "Senior secured bond" and "Senior unsecured bond" are kept.

	No.	Mean	Std
Senior secured bonds	332	0.3708	0.3688
Senior unsecured bonds	681	0.4900	0.3813
Subordinated bonds	400	0.6927	0.3628
Aggregated portfolio	1413	0.5194	0.3915

Table 5.7. Descriptions of portfolio loss distributions

Table 5.7 shows the characteristics of loss distributions of both aggregated and segmented portfolios. Three measurements including Value-at-Risk (VaR), expected shortfall (ES), and expected loss (EL) are reported. VaR and ES are reported at both 0.05 and 0.01 levels. Both AIRB and FIRB approaches are examined. Under the FIRB approach the LGD of senior unsecured bond is assigned as 0.45 and the subordinated bond is assigned a value of 0.75. For the senior secured bond we use the historical average LGD of the senior secured bonds in our sample which is 0.3708 based on Table 5.6. Three models including obligor-varying, time-varying factor and linear regression model are considered to be AIRB approaches.

Panel A. Aggregated portfolio

	VaR		ES		EL
q	0.05	0.01	0.05	0.01	−
FIRB			0. 5163		
Obligor-varying factor model	0.6171	0.6405	0.6313	0.6515	0.5518
Time-varying factor model	0.5583	0.5823	0.5729	0.5940	0.5022
Linear regression	0.5333	0.5390	0.5368	0.5419	0.5192

Panel B. Senior secured bonds

	VaR		ES		EL
q	0.05	0.01	0.05	0.01	−
FIRB			0.3707		
Obligor-varying factor model	0.4317	0.4883	0.4660	0.5158	0.2960
Time-varying factor model	0.4615	0.5128	0.4931	0.5375	0.3387
Linear regression	0.3995	0.4116	0.4068	0.4172	0.3708

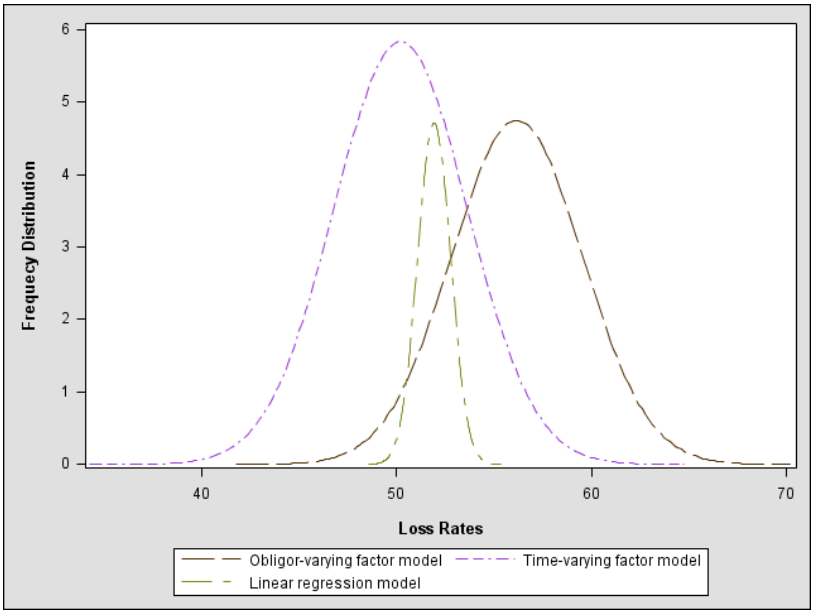
Panel C. Senior unsecured bonds

	VaR		ES		EL
q	0.05	0.01	0.05	0.01	−
FIRB			0.4500		
Obligor-varying factor model	0.6743	0.7080	0.6950	0.7249	0.5942
Time-varying factor model	0.5458	0.5762	0.5645	0.5918	0.4712
Linear regression	0.5099	0.5183	0.5150	0.5225	0.4900

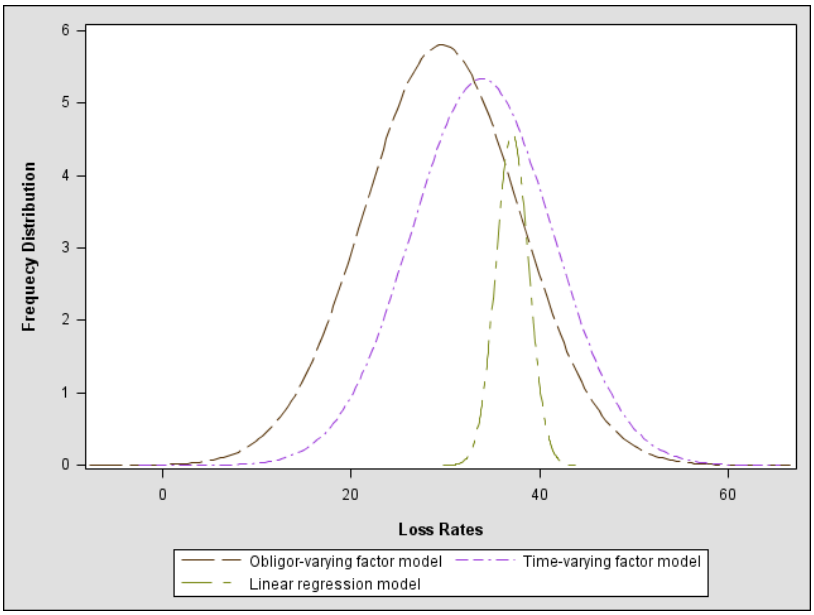
Panel D. Subordinated bonds

	VaR		ES		EL
q	0.05	0.01	0.05	0.01	–
FIRB			0.7500		
Obligor-varying factor model	0.7317	0.7487	0.7421	0.7573	0.6903
Time-varying factor model	0.7370	0.7619	0.7525	0.7752	0.6764
Linear regression	0.7189	0.7300	0.7256	0.7353	0.6927

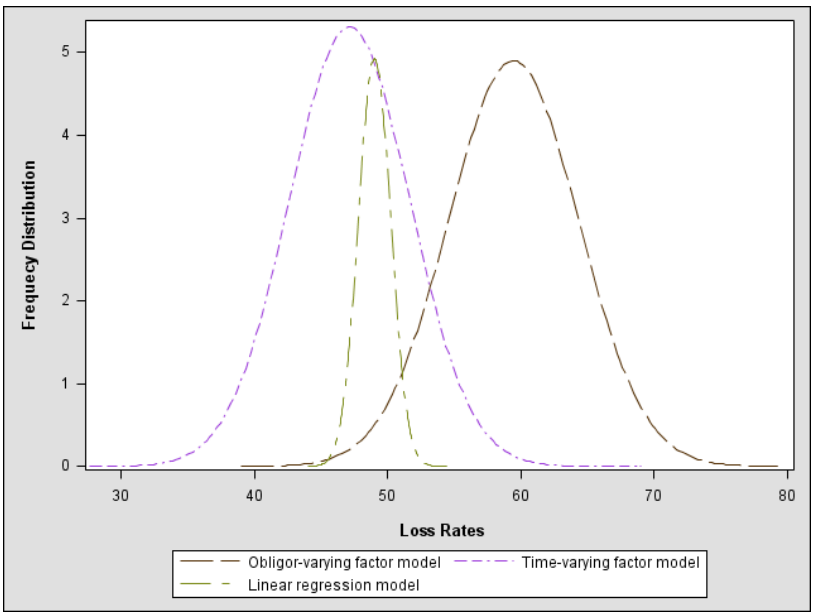
Figure 5.3. Plot of simulated loss distributions



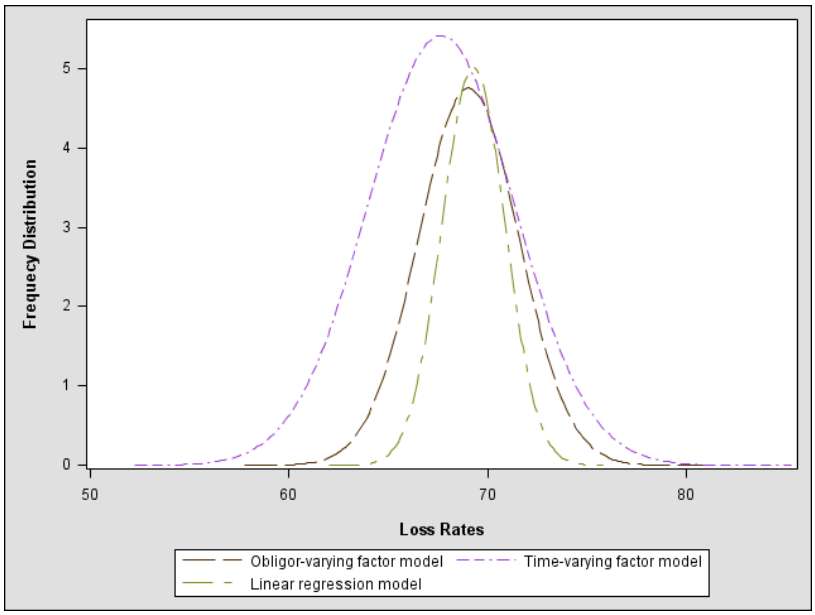
Panel A. Aggregated portfolio



Panel B. Senior secured bonds



Panel C. Senior unsecured bonds



Panel D. Subordinate bonds

5.5. Concluding remarks

Unobservable heterogeneity has been well investigated in PD modelling where default risk is assumed to be correlated across different instruments and firms are dependent on a common risk factor. In recovery rates models a latent time-varying systematic risk factor is commonly incorporated to explain the economic cyclical effects on recovery risk. In this paper we investigate the impact of firm specific

heterogeneity on modelling US corporate bond recovery rates and make the following contributions to the literature.

First we place the emphasis on the inclusion of firm heterogeneity in modelling instrument level recovery rates. By specifying the random effect at the obligor level, the single factor model shows a substantial improvement in model fit compared with the time-varying factor model and other traditional regression models. We suggest that the main reason is the unobservable obligor information is well explained by accounting for the firm specific heterogeneity. Unlike Hamerle et al (2006) which argued that more important observable variables were needed to explain the variations of LGD, our findings suggest that the reason why the variations of recovery rates can not be explained adequately is caused by the unobservable heterogeneity instead of the absence of other relevant observable determinants. Therefore the inclusion of obligor-varying random effect term improves the model fit significantly. The R^2 of 0.8964 obtained by the obligor-varying factor model presented in our study has never been reported in literature. One study that used a similar data set to ours is Jacobs and Karagozoglu (2011), which proposed a beta-linked generalized linear model to estimate recovery rates at firm and instrument levels jointly and reported an in-sample R^2 of 0.6997 and out-of-sample R^2 of 0.6119 at the instrument level. Qi and Zhao (2011) compared six different techniques on modelling recovery rates from MURD and reported that neural networks gave the highest R^2 of 0.529.

Next we carefully examine the predicted latent factors and their impact on aggregated recovery rates. We aggregate the predicted obligor-varying latent factors by year and find that their movement tendency is rather close to the predicted time-varying latent factors. We argue that the time-varying heterogeneity is well represented by the inclusion of firm level heterogeneity. Another interesting finding is that the predicted aggregated annual recovery rates of an obligor-varying factor model demonstrate an equally good fit of the historical annual recovery rates, implying its advantages at both yearly aggregated and instrument levels.

Furthermore, we show that the specification of normal distributional assumption is more appropriate than the other non-normal distributional assumptions. Our finding is consistent with Dwyer and Korablev (2009) which has shown that it was reasonable to assume a linear relationship for the recovery rate and its determinants. For the fixed effect regression models it is noticed that fractional response regression gives marginal advantages to linear regression and inflated beta regression in terms of

model fit and predictive accuracies. However, the linear single factor model is more robust with better model fit than the other non-linear specifications. We compare three other distributional assumptions on the factor models, and find that only the logit-normal specification gives a comparable model fit. Both log-normal and logit-normal factor models are extremely sensitive to the choice of the perturbation value at boundaries 0 and 1. The inflated beta mixed effects model proves to be more advantageous than the inflated beta regressions, but they are not comparable with the linear models. We believe a linear specification is the optimal choice for bonds recovery rates modelling.

Finally we investigate the impact of our models on credit risk management by comparing the simulated loss rates distributions generated by the AIRB approaches represented by single factor models and linear regression and the FIRB approach. We find that under the FIRB approach the aggregated portfolio loss is seriously underestimated. For the segmented portfolios the LGD specification of subordinated bonds under FIRB is quite close to the estimates of VaR and ES of the AIRB models and is appropriate for the loss predictions. But for the bonds with higher seniorities the FIRB approach may underestimate the unexpected losses based on our simulation results. We also find that both obligor-varying and time-varying models provide more frequent extreme losses than the linear regression method for both aggregated and segmented portfolios. We suggest the LGD specifications under FIRB approach may underestimate the potential unexpected losses, especially for the bonds with high seniorities.

One caveat in our study is that the default and recovery risk correlation is not taken into consideration in our modelling framework because of the limit of the data. We believe that with the incorporation of a default risk model, we expect to observe more interesting evidence of the impact of the firm heterogeneity on the credit portfolio losses. Another one is the limitation of the obligor-varying single factor model, which is that it is unable to predict the recovery rates of the instruments issued by new obligors. Such limitation also affects the experiment design which has been illustrated in Part D of Section 5.4. Similarly it is difficult for the time-varying and seniority-varying single factor models to make prediction for the instruments whose default dates or seniorities are out of the range defined in the training set. The time-varying random effect terms can be predicted by a time series model, and further investigation is required to improve the obligor-varying random effect model.

Chapter 6
Two-Stage Modelling for Recovery Rates: A Case Study of UK Credit Cards

6.1. Introduction

This chapter presents an empirical study using a real world dataset of recovery rates of credit cards from a UK credit card lender. The purpose of this study is to show a new methodology that is capable of improving the predictive accuracy of recovery rates modelling for credit cards. Unlike Chapters 4 and 5 where the research target is corporate bonds, there are more observations with the recovery rates of 0 and 1 in the portfolio of credit cards. In this chapter we focus on dealing with the extreme cases combining parametric and non-parametric techniques.

As introduced in Chapters 2 and 4, parametric models have been widely applied to predicting LGD of bank loans and identifying potential significantly useful predictors. For example, Qi and Yang (2009) identified the updated loan-to-value (CLTV) to be the single most important determinant of LGD for residential mortgage loans. Regarding credit cards, Bellotti and Crook (2012) discussed the influences of application and macroeconomic variables on recovery rates modelling. Similarly Khieu et al (2013) examined the determinants of bank loan recovery rates by applying both OLS and fractional response regression models. Leow et al (2013) found that the incorporation of macroeconomic variables was more effective on residential mortgage loans than unsecured personal loans.

Semi-parametric and non-parametric models have emerged to be an alternative to parametric statistical models, and recent research has found that they can effectively improve the predictive accuracies for recovery rates of bank loans compared with parametric models. Non-parametric statistical models that have been proposed for modelling recovery rates include the mixture beta-kernel estimator in Calabrese and Zenga (2010) and the zero-adjusted gamma regression model in Tong et al (2013). The mixture beta-kernel estimator was only used to fit the recovery rates distribution without any observable covariates included. Tong et al (2013) showed that the zero-adjusted gamma regression was flexible and competitive by reparameterizing the mean and dispersion parameters with additive non-parametric terms.

Machine learning techniques have also been investigated in a very limited number of studies including regression trees in Bastos (2010) and neural networks in

Qi and Zhao (2011), both of which found that machine learning methods outperformed parametric statistical regression models. A more comprehensive study was conducted by Loterman et al (2011) which benchmarked a total of 24 existed methods including both statistical regression models and machine learning techniques on six bank loans loss datasets. They found that machine learning techniques such as neural networks and SVMs tended to give the best performances across the datasets. However, they did not make any further improvement on the SVR models.

Chapter 4 has shown that SVM techniques are able to improve LGD predictive accuracy effectively for corporate bonds. In this chapter we conduct a further investigation by introducing SVM techniques into a two-stage modelling framework to predict recovery rates for credit cards. Two-stage methods developed in literature address a serious problem in recovery rates modelling, which is how to model the extreme cases concentrating on the boundaries at 0 and 1. Single-stage models assume that all cases are generated from the same distribution while two-stage models consider that the cases with recovery rates of 0 and 1 are intrinsically distinct from the cases between 0 and 1 which should be identified first. We develop the hypothesis that the performances of two-stage models are disappointing because the probabilities generated from a logistic regression model are not accurate enough to separate the cases at boundaries from those values in the interval (0, 1). In this study we seek to apply a least squares support vector classifier (LS-SVC) technique as an alternative method for the classification problem under the two-stage framework, and then the LS-SVC classification scores are transformed into probabilistic outputs by fitting a sigmoid form function using a maximum likelihood method proposed in Platt (1999). We find that the two-stage model equipped with a LS-SVC method gives significantly improved predictive accuracy of recovery rates compared with the other single-stage models, which suggests that the predictive performances of two-stage models rely heavily on the choice of classification model. To further examine our hypothesis we compare the classification accuracies between LS-SVC and logistic regression methods and find that LS-SVC consistently outperforms logistic regression for both of the two substages. Finally we study how the regression method influences the two-stage framework by modelling on cases with recovery rates in [0, 1] and (0, 1) separately. We find that when modelling on the cases in (0, 1), the least squares support vector regression model (LS-SVR) gives relatively close performances to an OLS model. But when LS-SVR is applied in the two-stage model, it is shown that the

combination of LS-SVC and LS-SVR is significantly outperformed by the combination of LS-SVC and OLS, although the margin is not remarkable. We conclude that the choice of regression methods plays a less crucial role than that of the classification methods.

The rest of this chapter is organized as follows. Section 6.2 introduces the methodologies applied in the empirical study where the kernel based support vector machine techniques will be presented with more details. Empirical evidence will be demonstrated in Section 6.3 including the interpretations of parameters and discussions of the model performances, and Section 6.4 concludes this chapter.

6.2. Models

6.2.1. Parametric models

We first introduce three parametric models that are commonly applied in LGD modelling including ordinary linear regression (OLS), fractional response regression and inflated beta regression methods. Both OLS and fractional response regression (Papke and Wooldridge, 1996) have been investigated extensively in LGD/recovery rates modelling for both corporate bonds and bank loans. Beta regression was proposed by Ferrari and Neto (2004) to fit the fractional response data with a beta distribution defined in (0, 1). The model is given as

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \tag{6.1}$$

where μ and ϕ are the mean and precision parameters that can be reparameterized with respect to the predictors. However, the beta regression model defines the dependent variable y in (0, 1) and thus neglects the boundary values 0 and 1 which are especially crucial to recovery rates modelling. To overcome this drawback Ospina and Ferrari (2010) proposed an inflated beta regression model to take the boundary values into consideration. It defines a mixture distribution for the dependent variable as a combination of a Bernoulli distribution and a beta distribution such that

$$bi_{01}(y;\pi,\psi,\mu,\phi) = \begin{cases} \pi(1-\psi) & if\ y=0 \\ \pi\psi & if\ y=1 \\ (1-\pi)f(y;\mu,\phi) & if\ y\in(0,1) \end{cases}. \tag{6.2}$$

The beta distribution assumption for recovery rates was first introduced by Gupton and Stein (2002) in Moody's internal LGD modelling framework LossCalc™. They suggested using a beta distribution to transform the recovery rates into a normally

distributed space and then to employ OLS to fit the transformed dependent variable, and finally the fitted dependent variables were transformed back to the fitted recovery rates. This idea has been widely accepted and adopted in the research on LGD/recovery rates modelling (Loterman et al, 2011; Bellotti and Crook, 2012). In contrast, Calabrese (2014) empirically studied the recovery rates of bank loans of the Bank of Italy showing that the inflated beta regression model demonstrated better out-of-time predictive accuracies compared with fractional response regression models, and that it was preferable for different forecasting periods of time and for different sample percentages of the extreme values of recovery rates. In our following study we adopt the same methodology from Calabrese (2014) to examine whether the inflated beta regression remains the most accurate for our data.

6.2.2. Support vector machine

The support vector machine was proposed by Vapnik (1995, 1998) and it has been an increasingly attractive technique in multiple areas. Compared with other parametric regression models, support vector regression has been found to be advantageous in recovery rates modelling (Loterman et al, 2011). In this section we introduce least squares support vector methods for both classification and regression problems respectively.

Classification

Suykens et al (1999) developed a least squares support vector classifier (LS-SVC) where the cost function was defined as the sum of squared error terms. One of the advantages of a LS-SVC is that it only needs to solve a linear system of equations instead of a quadratic programming problem as in the standard SVM models. Given a dataset $D = \{(\mathbf{x}_i, s_i)\}_{i=1}^{N}$, $\mathbf{x}_i \in R^m$ denote the covariates of *i*-th observation with the related labels s_i defined as $s_i \in \{-1, 1\}$. The LS-SVC is given as

$$\begin{aligned} &\min J(\mathbf{w}, b; \xi_i) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{N}\xi_i^2 \\ &s.t. \;\; s_i(\mathbf{w}^T\varphi(\mathbf{x}_i) + b) = 1 - \xi_i, \quad i = 1,...,N \end{aligned}, \tag{6.3}$$

where $\mathbf{w}$ denotes the parameter vector of the associated covariates and b is the intercept term. Here error terms, ξ_i^2, are scaled by a regularization parameter *C*, and $\varphi(\mathbf{x}_i)$ represents the kernel function that maps the data from original data space to a higher dimensional space. This model is then solved by its dual form problem derived

from a Lagrangian function

$$L(\alpha_i; \mathbf{w}, b, \xi_i) = J(\mathbf{w}, \xi_i) - \sum_i \alpha_i (s_i(\mathbf{w}^T \varphi(\mathbf{x}_i) + b) - 1 + \xi_i),$$

where α_i is the Lagrangian multiplier. Based on KKT conditions we have

$$\begin{cases} \nabla_w L = \mathbf{w} - \sum_i \alpha_i s_i \varphi(\mathbf{x}_i) = 0 \Rightarrow \mathbf{w} = \sum_i \alpha_i s_i \varphi(\mathbf{x}_i) \\ \nabla_b L = \sum_i \alpha_i s_i = 0 \\ \nabla_{\xi_i} L = \alpha_i - C\xi_i = 0 \Rightarrow \xi_i = \dfrac{\alpha_i}{C} \end{cases}. \tag{6.4}$$

After inserting the optimal conditions (6.4) back into the Lagrangian function, a linear system of equations is formulated as follows

$$\begin{pmatrix} 0 & \mathbf{s}^T \\ \mathbf{s} & \bar{\mathbf{H}} \end{pmatrix} \begin{pmatrix} b \\ \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{e} \end{pmatrix}, \tag{6.5}$$

where $\mathbf{e} = (\underbrace{1,...,1}_{N\times 1})^T$, $\mathbf{s} = (s_1,...,s_N)^T$, $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_N)^T$, $\bar{\mathbf{H}} = \mathbf{H} + \frac{1}{C}\mathbf{I}$, $\mathbf{H}_{ij} = s_i s_j \mathbf{K}(x_i, x_j)$, and $\mathbf{K}(x_i, x_j)$ defines the inner product of a pair of kernel functions as $\mathbf{K}(x_i, x_j) = \varphi(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j)$.

Denote the fitted classifier as $\hat{f}$ and its predicted output as $\hat{f}(\mathbf{x}_i)$. In the following we use $\hat{f}_i$ for short. To map SVM outputs to probabilistic outputs Platt (1999) proposed a parametric model to fit $\hat{f}_i$ using a sigmoid distribution, and the posterior probabilistic output $P(s_i = 1 \mid \hat{f}_i)$ is given such that

$$P(s_i = 1 \mid \hat{f}_i) = \frac{1}{1 + \exp(A\hat{f}_i + B)}, \tag{6.6}$$

where A and B are the unknown parameters to be estimated. The underlying assumption of this method is inspired by the conditional densities $P(\hat{f}_i \mid s_i = \pm 1)$, and a sigmoid form function is applied to fit such distributions. To estimate the parameters we first redefine the target variables as

$$t_i = \frac{s_i + 1}{2},$$

and then the estimates can be obtained by minimizing the negative log likelihood of the training data iteratively, which is defined as a cross-entropy error function such that

$$\min\ L(t_i, p_i; A, B) = -\sum_i t_i \log(p_i) + (1 - t_i)\log(1 - p_i), \tag{6.7}$$

where $p_i = P(s_i = 1 \mid \hat{f}_i)$.

Regression

The least squares support vector regression (LS-SVR) is formulated in a similar form. For more details see Section 4.2.3.1 of Chapter 4.

6.2.3. Two-stage model

We briefly introduce the two-stage modelling framework proposed by Bellotti and Crook (2012). First define the following notations such that

$$\begin{aligned} P_i^0 &= P(RR > 0) \\ P_i^{02} &= P(0 < RR < 1 \mid RR > 0), \\ P_i^{12} &= P(RR = 1 \mid RR > 0) \end{aligned} \tag{6.8}$$

and then the predicted recovery rate given by a two-stage model is defined such as

$$RR_i^{two-stage} = P_i^0 \times (P_i^{12} + P_i^{02} \times RR_i^{reg}), \tag{6.9}$$

where RR_i^{reg} denotes the predicted value by a regression model in the interval (0, 1). Bellotti and Crook (2012) suggested that it was normal to see that a customer in default either paid back all of the outstanding debt or paid back nothing. We believe the predictive performance of two-stage models depends on the choice of the classification methods at the final stage and thus propose to apply LS-SVC as an alternative classification method into the two-stage framework. For the regression methods we also investigate several different techniques besides OLS including fractional response regression, beta regression and LS-SVR techniques, all of which have been introduced in Chapter 4. Note that the inflated beta regression can be also regarded as a hybrid model that incorporates a logistic regression and a beta regression which is analogous to a two-stage model. The difference between the two methods lies in the estimation procedure: an inflated beta regression can be estimated by solving the likelihood function in a single step and the two-stage model has to be implemented step by step.

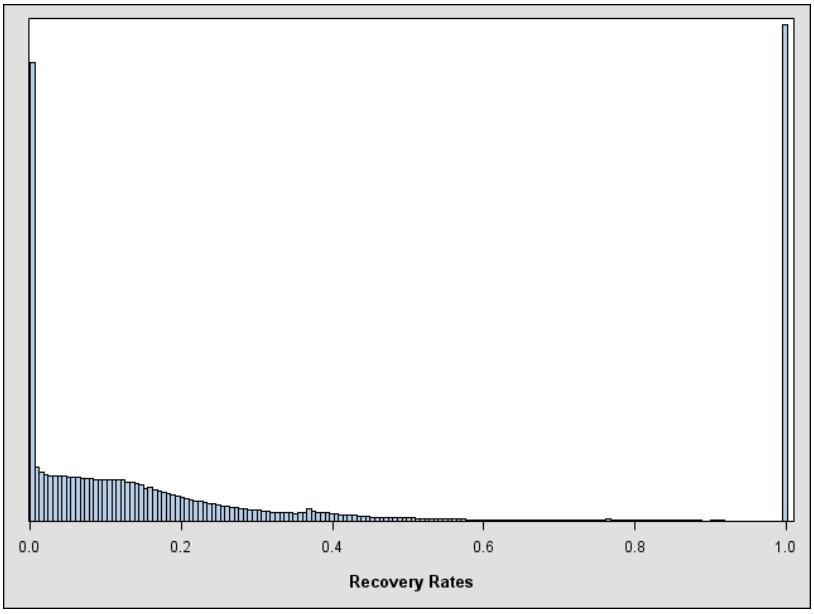
6.3. Empirical results

6.3.1. Data and setup

A data set of credit cards used in the analysis contains recovery rates information provided by a UK credit card lender. The data set consists of nearly 300,000 customers with more than 1,600,000 observations over periods from March 2009 to

February 2010. All customers are in default at the given month and for each customer there are records of the recovery rates during each month for a maximum of 12 months, and each individual observation is related to a final recovery rate. The dependent variable is termed as 24 months recovery rate after default which is provided by the lender in the data where overdue fees and accrued interest rate are included in the calculation. Therefore it is possible for the observed recovery rate to be greater than 1 for some observations. Without losing generalization we drop the cases with the recovery rates greater than 1 or less than 0. Figure 6.1 presents a histogram of recovery rates of the whole sample. It is very clear to observe that large numbers of cases concentrate at the boundaries.

Figure 6.1. Distribution of Recovery rates



We have nearly 40 potential predictors available for modelling. However, some of them have similar definitions and are highly correlated. We first generate the correlation matrix for all of the continuous variables based on Spearman's correlation coefficient and drop out the redundant ones if a correlation value is higher than 0.5 among several variables. The selected candidate variables include the account balance sheet and behavioural information. The outliers are defined as the values outside the interval between the 5 and 95 percentile of each variable and the observations with outliers are deleted. In total less than 5% of the total observations are deleted which does not affect the model estimates and the predictions significantly. Finally we have 13 account level variables for recovery rates models as are listed in Table 6.1 Panel A. Some candidate variables have been demonstrated to be important for LGD/RR modelling in the literature. For example Bellotti and Crook (2012) showed that both

Time on Book and *Time with Bank* had significant positive effects on recovery rate, and that *Balance at Observation* was negatively related to recovery rate. Our dataset also contains some new variables that have never been investigated in the literature. For example, the binary variable *Return on Order* identifies if a customer returned to order at any point in the last 12 months. It is expected that more outstanding debt can be recovered if the customer is shown to return. Another potentially useful variable is whether a customer is on a repayment plan or not. It can be inferred that a customer that is on a repayment plan should have a stronger will to repay their debt than a customer that is not. However, the data does not have any personal information relating to the customer such as marital status, educational background or family income, etc. After performing the correlation analysis eight account level variables are selected[3].

Macroeconomic variables are also incorporated to study the impacts on retail lending recovery risk. In the literature the influence of including macroeconomic variables on modelling recovery rates of unsecured retail loans is less evident than that of mortgage loans. Bellotti and Crook (2012) incorporated three variables including UK retail bank base interest rates, UK unemployment rate and UK earnings index, and they found that the inclusion macroeconomic variables increased model fit and improved out-of-time forecasts of recovery rates, although the authors mentioned that the data in this study spanned from 1999 to 2005 which did not cover an entire business cycle. Leow et al (2013) investigated a collection of variables from annual to monthly frequency indices to study the macroeconomic effects on LGD, and found that it was beneficial to incorporate macroeconomic variables for modelling the LGD of mortgage loans, but the parameter estimates of them were almost all statistically insignificant when modelling the LGD of personal retail loans. Khieu et al (2012) explored the determinants of bank loans from Moody's database and included both economic and industry indicators, where both annual GDP growth rate and the industry distress indicator were found to affect the recovery rate significantly.

Given that our data consists of monthly observations, only monthly varying macroeconomic variables including UK unemployment rate, Consumer Price Index (CPI) and Housing Price Index (HPI) are included. All of them are monthly data and are incorporated with one month lagged for each observation at default. As the time period covered in our data is quite short it would not be sensible to incorporate any

[3] The summarized statistics is omitted for confidentiality reason.

quarterly or annual data. The Bank of England interest rate is not included because there is little change since 2008.

To measure the forecast accuracy of recovery rates models we include the following performance metrics including Root Mean Squared Errors (RMSE) and Mean Absolute Errors (MAE) according to the literature. R Square (R^2) is also reported as an alternative measure of model fit. To test the robustness of each algorithm a bootstrapping method is applied which repeatedly draws a random 0.1 percent sample of the total observations to create a sub-sample. The procedure is repeated 1000 times to validate the robustness of the algorithms sufficiently. To assess the out-of-time predictions each sub-sample is divided into a training set and a testing set based on the observation date. The training set is defined to be from March 2009 to November 2009 and the testing set is from December 2009 to February 2010. We then report the mean and standard deviations of the performance metrics for each model.

Parameter tuning in SVC and SVR model is critical to the model performance. Similar to Chapter 4, the two parameters including regularization and kernel function parameter are tuned in the above bootstrapping procedure based on the performance of training set. The overfitting issue may potentially arise from the parameter tuning procedure because the predictive performance on the testing set is likely to be unsatisfied if the performance on the training set is maximised. However, the empirical evidence presented in Section 6.3.3 shows that the overfitting issue is not a major concern as SVR models still outperform the other statistical regression methods.

6.3.2. Model Interpretation

An explanatory analysis is performed on the whole sample with a robust linear regression model to adjust the estimated standard errors to account for the repeated observations over periods. OLS regression assumes that the residual terms are independent between observations, but in our sample the observations of each customer over periods are likely to be correlated. The estimates of coefficients of a robust regression are the same as the OLS estimates, but the standard errors take into account the correlation within a cluster to imply the correct statistical significance. The regression model is estimated with account level variables only and with both account level and macroeconomic variables, and the outputs of parameters estimates

and model fit are reported in Panels A and B of Table 6.1 respectively. To show the degree of multi-collinearity the VIF values for each parameter are also reported. It should be noticed that no variable has a VIF value greater than 5, which indicates that the model estimates are not significantly affected by multi-collinearity. It should be also noted that Table 6.1 shows that the incorporation of macroeconomic variables improves R^2 modestly from 0.1508 to 0.1515, although all three macroeconomic variables are statistically significant. It is observed that all the account level variables remain significant at the 0.01 confidence level with the inclusion of macroeconomic variables, indicating all account level variables are conditionally correlated with recovery rates.

Some straightforward conclusions on estimates of parameters can be taken from Table 6.1. For example, the number of months the account was with the bank (*Time with bank*) and the number of months that the customer has held the credit card (*Time on book*) both positively influence the recovery rate, showing that the longer a customer stays with the bank, a higher proportion of its debt will be recovered after default. According to Bellotti and Crook (2012), these two variables are the indicators of customer stability which are expected to lead to a lower recovery risk. *Balance at Observation* is shown to be negatively correlated with recovery rate, which indicates that the more outstanding debt a customer has, the more difficult it is to recover. It can be observed that the longer the customer is in arrears, the more will be repaid to the bank according to Table 6.1. One would expect that a bank would take more actions to urge the customer to pay back its debt if it finds the customer has been in default for a long time.

For the repayment behaviours the results show that the number of payments made in the last 12 months positively affects recovery rate, and as expected the average payment as a percentage of balance also positively influences the recovery rate. Next we suggest that a higher recovery rate is expected if a customer makes a higher payment most recently. There are three binary variables relating to the status of recovery process. Specifically, a higher recovery rate is observed if a customer returned to order in the last 12 months according to Table 6.1. However, contrary to the expectation that a customer is on a repayment plan would have a lower recovery rate. The negative effect may be explained that the customer who is assumed not to be able to repay its debt may be forced to join the repayment plan and is less capable of repaying debt.

Turning to the macroeconomic variables we notice that both CPI and HPI are negatively and significantly related to recovery rate. This implies that when price inflation increases customers are less capable of paying back their outstanding debts. The puzzling sign of the estimate of unemployment rate conflicts with the finding in Bellotti and Crook (2012), where the unemployment rate was shown to be negatively correlated to recovery rate. They found that the inclusion of macroeconomic variables generally improves the recovery rates predictions across test quarters modestly. One possible reason for our result is that our sample only covers one year which does not capture the long term correlation between recovery rates and economic variables very well, and we suggest that a data set with a longer time window is needed to investigate the impacts of macroeconomic conditions on modelling unsecured loans recovery rates.

Table 6.1. Explanatory models

Panel A. Modelling with account variables

	Estimate	p value	VIF
Intercept	0.4586 *** (0.0030)	<.0001	0
Time in months	0.0054 *** (0.0008)	<.0001	1.1881
Time on book	0.0002 *** (0.0000)	<.0001	1.2708
Sum of transactions across all current accounts	0.0064 *** (0.0001)	<.0001	1.0342
Number of months in arrears six months ago	0.0327 *** (0.0003)	<.0001	1.5260
Balance at observation	-0.0136 *** (0.0002)	<.0001	1.1486
Worst delinquency status in days across all products	-0.0001 *** (0.0000)	<.0001	1.0397
Number of payments made last 12 months	0.0025 *** (0.0003)	<.0001	2.2370
Average payment as percentage of balance in default summed over last 6 months	8.5094 *** (0.0852)	<.0001	2.0001
Status on if a customer returned to order	0.0029 * (0.0016)	0.0719	2.1807
Status on if a customer has spent 1-5 months in arrears	0.0189 *** (0.0012)	<.0001	2.0147
Status on if a customer is on a repayment plan	-0.1664 *** (0.0019)	<.0001	1.9355
Most recent payment received	0.0052 *** (0.0011)	<. 0001	1.5109
F value	24653.0	<.0001	
R^2	0.1500		
Adj R^2	0.1500		
RMSE	0.3297		

Panel B. Modelling with account and macroeconomic variables

	Estimate	p value	VIF
Intercept	0.6255 *** (0.0595)	<.0001	0
Time in months	0.0054 *** (0.0003)	<.0001	1.1894
Time on book	0.0002 *** (0.0000)	<.0001	1.2717
Sum of transactions across all current accounts	0.0064 *** (0.0000)	<.0001	1.0355
Number of months in arrears six months ago	0.0326 *** (0.0002)	<.0001	1.5391
Balance at observation	-0.0136 *** (0.0000)	<.0001	1.1518
Worst delinquency status in days across all products	-0.0001 *** (0.0000)	<.0001	1.0474
Number of payments made last 12 months	0.0025 *** (0.0001)	<.0001	2.2429
Average payment as percentage of balance in default summed over last 6 months	8.5050 *** (0.0387)	<.0001	2.0101
Status on if a customer returned to order	0.0021 *** (0.0008)	0.0051	2.1852
Status on if a customer has spent 1-5 months in arrears	0.0192 *** (0.0008)	<.0001	2.0151
Status on if a customer is on a repayment plan	-0.1650 *** (0.0009)	<.0001	1.9469
Most recent payment received	0.0050 *** (0.0006)	0.0018	1.5120
Monthly unemployment rate	0.0899 *** (0.0033)	<.0001	1.0229
Monthly CPI	-0.0035 *** (0.0003)	<.0001	1.8726
Monthly HPI	-0.0009 *** (0.0000)	<.0001	1.8577
F value	19822.1	<.0001	
R^2	0.1507		
Adj R^2	0.1507		
RMSE	0.3296		

6.3.3. Out-of-sample predictions

To investigate the effects of classification and regression in two-stage models we propose to model the cases with RR in [0, 1] and (0, 1) separately. Single-stage and two-stage models are all compared in [0, 1] and only single-stage models are

benchmarked in (0, 1). There are four single-stage methods investigated including OLS, fractional response regression, inflated beta regression and LS-SVR. For the two-stage models two classification methods are applied including logistic regression and LS-SVM, and there are four regression methods employed for the second stage that are the same as the single-stage models except that the inflated beta regression is replaced by a beta regression model because it is unnecessary to consider the cases at boundaries when modelling in (0, 1). In the following the abbreviations of two-stage models names are used for convenience. Terms of abbreviations and relevant full names are provided in Table 6.2. In total we have eight combinations for the two-stage models.

Table 6.2 Term of references

Table 6.2 presents the abbreviations and full names of the models in this study

	Abbreviations	Full names
Model1	OLS	Ordinary Linear Regression
Model2	Frac	Fractional Response Regression
Model3	Inflated Beta	Inflated Beta Regression
Model4	SVR	Least Squared Support Vector Regression
Model5	Logistic+OLS	Logistic Regression and Ordinary Linear Regression
Model6	Logistic+Frac	Logistic Regression and Fractional Response Regression
Model7	Logistic+Beta	Logistic Regression and Beta Regression
Model8	Logistic +SVR	Logistic Regression and Least Squared Support Vector Regression
Model9	SVC+OLS	Least Squared Support Vector Classification and Ordinary Linear Regression
Model10	SVC+Frac	Least Squared Support Vector Classification and Fractional Response Regression
Model11	SVC+Beta	Least Squared Support Vector Classification and Beta Regression
Model12	SVC+SVR	Least Squared Support Vector Classification and Least Squared Support Vector Regression

We first analyze the predictive performances of the cases with RR in [0, 1] and

report the outputs in Table 6.3. To compare model performances a paired t-test is applied to each performance metric and both the differences between each pair of models and the p values are reported in Table 6.4. First notice that OLS outperforms the other generalized linear regression models such as fractional response regression and inflated beta regression models in terms of out-of-sample predictive performances. From previous research such evidence is expected although the empirical recovery rates distribution is far from a Gaussian distribution. Both Zhang and Thomas (2010) and Bellotti and Crook (2012) have reported that the OLS regression model gave better predictions than other generalized linear models. Empirical evidence in Zhang and Thomas (2010) suggested that the flexibility of survival regression did not necessarily give better predictions than the OLS regression model because it was difficult to identify zero recovery rates cases when applying accelerated failure time models. In our study inflated beta regression, which is designed to accommodate the cases at the boundaries 0 and 1, does not show any advantages compared with OLS and fractional response regression. Notice that SVR model yields a consistently better model fit and higher predictive accuracy for both in-sample and out-of-sample tests. This result is also consistent with the findings in Loterman et al (2011) which showed SVR and neural networks significantly outperformed the other linear models for LGD prediction implying a strong non-linear relationship between LGD and its predictors.

The performances of the two-stage models are more straightforward. The two-stage logistic+OLS method proposed in Bellotti and Crook (2012) gave slightly better out-of-sample predictions than the single-stage OLS model. We replace the OLS with other techniques and find no noticeable improvement for either logistic+Frac or logistic+Beta. Instead logistic+OLS gives significantly better out-of-sample predictive accuracy than those. Furthermore it is noticed that logistic+SVR has a significantly lower R^2 and MAE and an insignificant improvement in terms of RMSE compared with logistic+OLS according to Table 6.4 Panel B. This indicates that the non-linear methods are not shown to improve the performances of two-stage models.

To check the hypothesis developed above, the logistic regression model is replaced by a LS-SVC technique under the two-stage modelling framework. We find that the two-stage models SVC+OLS and SVC+Frac significantly outperform all the other models. As can be seen from Table 6.4 there are insignificant differences between SVC+OLS and SVC+Frac in terms of R^2 and RMSE although SVC+Frac

shows a slightly significant better MAE. Notice that neither SVC+Beta nor SVC+SVR show better predictive accuracies than SVC+OLS or SVC+Frac. Also it is observed that the SVC+Beta model is much less competitive than any other two-stage method with a SVC technique. But SVC+Beta significantly outperforms the other single-stage statistical models, which implies that the cases with RR in (0, 1) may have a linear relationship between recovery rate and its predictors. Combined with the consistently poor performances of the inflated beta regression model, the results indicate that a beta distribution is not proving to be a superior model for recovery rates as expected. Yet when the SVC technique is applied as the classification method, all two-stage models present noticeable improvements compared with those using a logistic regression. This suggests that the probabilities of recovery rates being 0 or 1 generated in equation (6.6) from SVC techniques are more accurate than that those from logistic regression models.

Table 6.3. Model performances on cases with RR in [0, 1]

Table 6.3 presents the out-of-sample predictive accuracy for single-stage and two-stage models respectively.

Panel A. Single-stage models

	In sample			Out of sample		
	R^2	RMSE	MAE	R^2	RMSE	MAE
OLS	0.2014	0.2449	0.3176	0.0882	0.3424	0.2634
	(0.0393)	(0.0326)	(0.0187)	(0.0694)	(0.0213)	(0.0266)
Frac	0.2030	0.3173	0.2413	0.0778	0.3443	0.2678
	(0.0458)	(0.0329)	(0.0192)	(0.0707)	(0.0217)	(0.0273)
Inflated Beta	0.0690	0.3431	0.2721	0.0179	0.3556	0.2864
	(0.0318)	(0.0333)	(0.0252)	(0.0146)	(0.0221)	(0.0284)
SVR	0.6471	0.2112	0.1541	0.1214	0.3363	0.2538
	(0.0310)	(0.0283)	(0.0126)	(0.0570)	(0.0213)	(0.0219)

Panel B. Two-stage models

	In sample			Out of sample		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Logistic+OLS	0.2302	0.3118	0.2316	0.1018	0.3398	0.2549
	(0.0481)	(0.0335)	(0.0194)	(0.0723)	(0.0221)	(0.0266)
Logistic+Frac	0.2194	0.3203	0.2355	0.0894	0.3417	0.2509
	(0.0542)	(0.0682)	(0.0191)	(0.0698)	(0.0220)	(0.0246)
Logistic+Beta	0.2207	0.3131	0.2351	0.0880	0.3423	0.2513
	(0.0452)	(0.0164)	(0.0197)	(0.0736)	(0.0220)	(0.0251)
Logistic +SVR	0.2871	0.2938	0.2099	0.0825	0.3389	0.2616
	(0.0483)	(0.0191)	(0.0219)	(0.0709)	(0.0222)	(0.0240)
SVC+OLS	0.4794	0.2553	0.1707	0.1710	0.3256	0.2534
	(0.0430)	(0.0142)	(0.0220)	(0.0607)	(0.0215)	(0.0359)
SVC+Frac	0.4771	0.2550	0.1726	0.1744	0.3263	0.2509
	(0.0493)	(0.0151)	(0.0141)	(0.0654)	(0.0182)	(0.0202)
SVC+Beta	0.4534	0.2612	0.1874	0.1329	0.3349	0.2605
	(0.0475)	(0.0137)	(0.0226)	(0.0787)	(0.0217)	(0.0375)
SVC+SVR	0.5476	0.2378	0.1483	0.1628	0.3278	0.2444
	(0.0465)	(0.0156)	(0.0209)	(0.0767)	(0.0219)	(0.0245)

Table 6.4. Out-of-sample comparisons on [0, 1]

Table 6.4 presents the absolute differences of each performance metric between the model of related row and the model of related column. Paired t-test is conducted with p values reported in parenthesis. ***,** and * indicate the significance at 0.01, 0.05 and 0.1 confidence level respectively.

Panel A. R^2 table

	OLS	Frac	Inflated Beta	SVR	Logistic+OLS	Logistic+Frac	Logistic+Beta	Logistic +SVR	SVC+OLS	SVC+Frac	SVC+Beta	SVC+Beta
OLS	-											
Frac	-0.0104 *** (0.0009)	-										
Inflated Beta	-0.0703 *** (0.0000)	-0.0599 *** (0.0000)	-									
SVR	0.0332 *** (0.0000)	0.0436 *** (0.0000)	0.1035 *** (0.0000)	-								
Logistic+OLS	0.0136 *** (0.0000)	0.0240 *** (0.0000)	0.0839 *** (0.0000)	-0.0196 *** (0.0000)	-							
Logistic+Frac	0.0012 (0.6999)	0.0116 *** (0.0002)	0.0715 *** (0.0000)	-0.0320 *** (0.0000)	-0.0124 *** (0.0001)	-						
Logistic+Beta	-0.0002 (0.9502)	0.0102 *** (0.0016)	0.0701 *** (0.0000)	-0.0334 *** (0.0000)	-0.0138 *** (0.0000)	-0.0014 (0.6626)	-					
Logistic +SVR	-0.0057 * (0.0694)	0.0047 (0.1379)	0.0646 *** (0.0000)	-0.0389 *** (0.0000)	-0.0193 *** (0.0000)	-0.0069 ** (0.0284)	-0.0055 * (0.0889)	-				
SVC+OLS	0.0828 *** (0.0000)	0.0932 *** (0.0000)	0.1531 *** (0.0000)	0.0496 *** (0.0000)	0.0692 *** (0.0000)	0.0816 *** (0.0000)	0.0830 *** (0.0000)	0.0885 *** (0.0000)	-			
SVC+Frac	0.0862 *** (0.0000)	0.0966 *** (0.0000)	0.1565 *** (0.0000)	0.0530 *** (0.0000)	0.0726 *** (0.0000)	0.0850 *** (0.0000)	0.0864 *** (0.0000)	0.0919 *** (0.0000)	0.0034 (0.2284)	-		
SVC+Beta	0.0447 *** (0.0000)	0.0551 *** (0.0000)	0.1150 *** (0.0000)	0.0115 *** (0.0002)	0.0311 *** (0.0000)	0.0435 *** (0.0000)	0.0449 *** (0.0000)	0.0504 *** (0.0000)	-0.0381 *** (0.0000)	-0.0415 *** (0.0000)	-	
SVC+SVR	0.0746 *** (0.0000)	0.0850 *** (0.0000)	0.1449 *** (0.0000)	0.0414 *** (0.0000)	0.0610 *** (0.0000)	0.0734 *** (0.0000)	0.0748 *** (0.0000)	0.0803 *** (0.0000)	-0.0082 *** (0.0081)	-0.0116 *** (0.0003)	0.0299 *** (0.0000)	-

Panel B. RMSE table

	OLS	Frac	Inflated Beta	SVR	Logistic+OLS	Logistic+Frac	Logistic+Beta	Logistic +SVR	SVC+OLS	SVC+Frac	SVC+Beta	SVC+Beta
OLS	-											
Frac	0.0019 **	-										
	(0.0483)											
Inflated Beta	0.0132 ***	0.0113 ***	-									
	(0.0000)	(0.0000)										
SVR	-0.0061 ***	-0.0080 ***	-0.0193 ***	-								
	(0.0000)	(0.0000)	(0.0000)									
Logistic+OLS	-0.0026 ***	-0.0045 ***	-0.0158 ***	0.0035 ***	-							
	(0.0075)	(0.0000)	(0.0000)	(0.0003)								
Logistic+Frac	-0.0007	-0.0026 ***	-0.0139 ***	0.0054 ***	0.0019 *	-						
	(0.4698)	(0.0079)	(0.0000)	(0.0000)	(0.0542)							
Logistic+Beta	-0.0001	-0.0020 **	-0.0133 ***	0.0060 ***	0.0025 **	0.0006	-					
	(0.9178)	(0.0408)	(0.0000)	(0.0000)	(0.0113)	(0.5420)						
Logistic +SVR	-0.0035 ***	-0.0054 ***	-0.0167 ***	0.0026 ***	-0.0009	-0.0028 ***	-0.0034 ***	-			-	
	(0.0003)	(0.0000)	(0.0000)	(0.0076)	(0.3637)	(0.0047)	(0.0006)					
SVC+OLS	-0.0168 ***	-0.0187 ***	-0.0300 ***	-0.0107 ***	-0.0142 ***	-0.0161 ***	-0.0167 ***	-0.0133 ***	-			
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)				
SVC+Frac	-0.0161 ***	-0.0180 ***	-0.0293 ***	-0.0100 ***	-0.0135 ***	-0.0154 ***	-0.0160 ***	-0.0126 ***	0.0007	-		
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.4321)			
SVC+Beta	-0.0075 ***	-0.0094 ***	-0.0207 ***	-0.0014	-0.0049 ***	-0.0068 ***	-0.0074 ***	-0.0040 ***	0.0093 ***	0.0086 ***	-	
	(0.0000)	(0.0000)	(0.0000)	(0.1456)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)		
OLS	-0.0146 ***	-0.0165 ***	-0.0278 ***	-0.0085 ***	-0.0120 ***	-0.0139 ***	-0.0145 ***	-0.0111 ***	0.0022 **	0.0015 *	-0.0071 ***	
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0235)	(0.0959)	(0.0000)	

Panel C. MAE table

	OLS	Frac	Inflated Beta	SVR	Logistic+OLS	Logistic+Frac	Logistic+Beta	Logistic +SVR	SVC+OLS	SVC+Frac	SVC+Beta	SVC+Beta
OLS	-											
Frac	0.0044 *** (0.0003)	-										
Inflated Beta	0.0230 *** (0.0000)	0.0186 *** (0.0000)	-									
SVR	-0.0096 *** (0.0000)	-0.0140 *** (0.0000)	-0.0326 *** (0.0000)	-								
Logistic+OLS	-0.0085 *** (0.0000)	-0.0129 *** (0.0000)	-0.0315 *** (0.0000)	0.0011 (0.3128)	-							
Logistic+Frac	-0.0125 *** (0.0000)	-0.0169 *** (0.0000)	-0.0355 *** (0.0000)	-0.0029 *** (0.0054)	-0.0040 *** (0.0005)	-						
Logistic+Beta	-0.0121 *** (0.0000)	-0.0165 *** (0.0000)	-0.0351 *** (0.0000)	-0.0025 ** (0.0177)	-0.0036 *** (0.0019)	0.0004 (0.7190)	-					
Logistic +SVR	-0.0018 (0.1123)	-0.0062 *** (0.0000)	-0.0248 *** (0.0000)	0.0078 *** (0.0000)	0.0067 *** (0.0000)	0.0107 *** (0.0000)	0.0103 *** (0.0000)	-				
SVC+OLS	-0.0100 *** (0.0000)	-0.0144 *** (0.0000)	-0.0330 *** (0.0000)	-0.0004 (0.7636)	-0.0015 (0.2885)	0.0025 * (0.0694)	0.0021 (0.1297)	-0.0082 *** (0.0000)	-			
SVC+Frac	-0.0125 *** (0.0000)	-0.0169 *** (0.0000)	-0.0355 *** (0.0000)	-0.0029 *** (0.0021)	-0.0040 *** (0.0002)	0.0000 (1.0000)	-0.0004 (0.6947)	-0.0107 *** (0.0000)	-0.0025 * (0.0551)	-		
SVC+Beta	-0.0029 ** (0.0462)	-0.0073 *** (0.0000)	-0.0259 *** (0.0000)	0.0067 *** (0.0000)	0.0056 *** (0.0001)	0.0096 *** (0.0000)	0.0092 *** (0.0000)	-0.0011 (0.4347)	0.0071 *** (0.0000)	0.0096 *** (0.0000)	-	
OLS	-0.0190 *** (0.0000)	-0.0234 *** (0.0000)	-0.0420 *** (0.0000)	-0.0094 *** (0.0000)	-0.0105 *** (0.0000)	-0.0065 *** (0.0000)	-0.0069 *** (0.0000)	-0.0172 *** (0.0000)	-0.0090 *** (0.0000)	-0.0065 *** (0.0000)	-0.0161 *** (0.0000)	-

We further explore the advantages of SVC techniques by comparing the classification accuracies of logistic regression and SVC models in terms of AUC (Area under curve). AUC is a statistics related to a ROC (Receiver Operating characteristic) curve to measure the overall performance of the classifier scores. A simple method of AUC calculation of a classifier G was presented in Hand and Till (2001) as equation (6.10).

$$\hat{AUC} = \frac{\sum_i r_i - n_0(n_0 + 1)/2}{n_0 n_1}, \tag{6.10}$$

where n_0 and n_1 are the numbers of positive and negative cases respectively, and r_i denotes the rank of *i*-th positive case in the ranked list of the predictive values from the logistic regression. In two stage models there are two classification events involved. Event 1: RR=0 vs. RR>0; Event 2: RR=1 vs. 0<RR<1. The same scheme is applied as for recovery rates prediction to generate the classification predictions repeatedly for 1000 times and we report both the means and the standard errors of both in-sample and out-of-sample performances in Table 6.5. A paired t-test is employed for out-of-sample predictions comparisons. This shows that SVM models excel in general for both events in terms of in-sample and out-of-sample AUC. It is also noticed that both logistic regression and SVM perform fairly well for Event 1 with an AUC higher than 0.85 with mild advantage shown by SVM. SVM gives a better performance on Event 2 with significant improvement on AUC. Table 6.6 confirms the expectations that the SVM technique is able to generate relatively better probabilistic outputs than logistic regression. It should be noted that it is more difficult to separate the cases with RR in (0, 1) from those with RR=1, suggesting that customers who are willing to repay all debts are more difficult to separate from those who are unable to repay any debt. Our results suggest that non-linear models, including both statistical and machine learning techniques, do not exhibit advantages over OLS in the two-stage frameworks no matter whether a logistic regression or a SVC technique is applied as the classification method

Table 6.5. AUC comparisons of classification

Table 6.5 presents the AUC values and paired t test results of the two classifications methods in two-stage models

	In sample		Out of sample	
	Event 1	Event 2	Event 1	Event 2
Logistic Regression	0.9089	0.8152	0.8671	0.7648
	(0.0248)	(0.0311)	(0.0406)	(0.0544)
LS-SVC	0.9245	0.9549	0.8725	0.8013
	(0.0300)	(0.0116)	(0.0443)	(0.0572)
t value			-2.84 ***	-4.61 ***
			(0.0045)	(0.0000)

To examine the effects of regression models for modelling RR in (0, 1) four methods are applied including OLS, fractional response regression, beta regression and a SVR technique. The performance metrics and model comparison results are reported in Tables 6.6 and 6.7 respectively. According to Table 6.7 the SVR technique outperforms the other methods significantly in terms of out-of-sample MAE, but it presents an insignificant advantage compared with OLS in terms of R^2 and RMSE. This suggests that the SVR is considered to be as accurate as OLS when modelling RR in (0, 1), and both of them are significantly better than the fractional response and beta regression models. This is consistent with the evidence presented above and it can be concluded that SVC+OLS and SVC+SVR give similarly accurate out-of-sample predictions. Similarly Logistic+OLS shows marginal advantage over Logistic+SVR in terms of R^2 and MAE.

Table 6.6. Performances of single-stage models in 0<RR<1

Table 6.6 presents the in-sample and out-of-sample performances of single-stage models on the sample with recovery rate between (0, 1).

	In sample			Out of sample		
	R^2	RMSE	MAE	R^2	RMSE	MAE
OLS	0.1951	0.1564	0.1072	0.0792	0.2037	0.1480
	(0.0567)	(0.0165)	(0.0114)	(0.0944)	(0.0270)	(0.0215)
Frac	0.1705	0.1589	0.1097	0.0513	0.2069	0.1476
	(0.0547)	(0.0166)	(0.0117)	(0.0906)	(0.0272)	(0.0205)
Inflated Beta	0.0633	0.1686	0.1180	0.0277	0.2179	0.1538
	(0.1073)	(0.0192)	(0.0161)	(0.0726)	(0.0306)	(0.0282)
SVR	0.6579	0.1015	0.0691	0.0828	0.2035	0.1363
	(0.0370)	(0.0113)	(0.0076)	(0.0827)	(0.0265)	(0.0160)

Table 6.7 Comparisons of single-stage models

Table 6.7 presents the absolute differences of each performance metric between the model of related row and the model of related column. Paired t-test is conducted with p values reported in parenthesis. ***,** and * indicate the significance at 0.01, 0.05 and 0.1 significance levels respectively.

Panel A. R^2

	OLS	Frac	Inflated Beta	SVR
OLS	-			
Frac	-0.0279 ***	-		
	(0.0000)			
Inflated Beta	-0.0515 ***	-0.0236 ***	-	
	(0.0000)	(0.0000)		
SVR	0.0036	0.0315 ***	0.0551 ***	-
	(0.3645)	(0.0000)	(0.0000)	

Panel B. RMSE

	OLS	Frac	Inflated Beta	SVR
OLS	-			
Frac	0.0032 ***	-		
	(0.0083)			
Inflated Beta	0.0142 ***	0.0110 ***	-	
	(0.0000)	(0.0000)		
SVR	-0.0002	-0.0034 ***	-0.0144 ***	
	(0.8672)	(0.0047)	(0.0000)	

Panel C. MAE

	OLS	Frac	Inflated Beta	SVR
OLS	-			
Frac	-0.0004 (0.6703)	-		
Inflated Beta	0.0058 *** (0.0000)	0.0062*** (0.0000)	-	
SVR	-0.0117 *** (0.0000)	-0.0113 *** (0.0000)	-0.0175 *** (0.0000)	-

6.4. Conclusions

This chapter evaluates the performances of a group of statistical and machine learning techniques to predict the recovery rates of a large sample of UK bank credit cards, and shows that machine learning techniques are an effective supplement to statistical regression models to improve predictions of recovery rates. Kernel based least squares support vector machine techniques are applied in two ways. First the recovery rates were modelled with support vector regression directly and SVR demonstrated better predictions than the other linear or generalized linear models in terms of both in-sample and out-of-sample predictive metrics on average, although the improvements of RMSE and MAE are not as remarkable as R^2. Second the support vector machine was incorporated into a two-stage modelling framework where the cases with zero and one recovery rates were separated by a least squares support vector classifier, and then the cases in the interval (0, 1) were modelled with other regression models. It can be concluded that the combination of LS-SVC and OLS gives the best out-of-sample predictive accuracies in terms of out-of sample RMSE and MAE. It is also noticed that this two-stage model outperforms the single-stage support vector regression model significantly in terms of the out-of-sample R^2. For the other combinations of two-stage models, where the OLS is replaced by other statistical or machine learning methods, the predictive performances are not as good as the SVC+OLS model. We suggest that choice of algorithm at the separation stage of the two-stage model plays an evidently crucial role in the predictive accuracy of recovery rates modelling while the choice of algorithms at the regression stage in (0, 1) is less important.

Chapter 7
Conclusions

7.1. Summary

Loss given default modelling has been a new challenge in industry applications and academic research since the issue of the new Basel Accord. The new capital guideline encourages banks to develop internal models to estimate risk parameters with respect to various products including retail and wholesale loans. In the literature statistical regression models such as ordinary linear regression and fractional response regression are most widely applied to loans and bonds to seek the significant determinants of LGD. Limited efforts have been made to develop new LGD models improve the predictive accuracy, and this thesis makes an attempt to apply innovative methodologies to estimating LGD for corporate bonds and retail loans. Chapter 1 provides a high level overview of credit risk modelling under the guideline of the Basel Accords, and then introduces the research questions and contributions of this thesis. Chapter 2 reviews the literature related to LGD modelling for bank loans and corporate bonds exhaustively. Chapters 4 and 5 discuss the impacts of heterogeneity on modelling LGD of corporate bonds based on the loss data of corporate bonds described in Chapter 3. Chapter 6 introduces a new modelling algorithm to predict LGD for retail credit cards. This chapter concludes the whole thesis by summarizing the contributions and limitations of this research, and lays out the topics for further study.

7.2. Contributions

This thesis consists of three substantive empirical studies to answer the research questions addressed in Chapter 1 by making contributions related to building up innovative algorithms to improve the LGD predictions for corporate bonds and retail credit cards. In Chapter 4 we develop a new support vector regression model to predict LGD to account for the seniority heterogeneity of corporate bonds for the first time. Here two improved versions are proposed to incorporate the seniority heterogeneity into SVR models aiming to predict LGD accurately. The proposed models have never been reported in the literature. Our first study exhibits the power of machine learning techniques especially the support vector regression models in modelling LGD which is consistent with the findings in Loterman et al (2011). But it

is the first detailed study related to applying SVR models to predicting LGD for defaulted corporate bonds. We conduct the empirical study at both aggregated and segmented levels by comparing the predictive performances of SVR models and statistical methods. At aggregated level we find that the two proposed models in our study both outperform the original SVR model significantly and other statistical models that have been widely investigated in the literature. At segmented seniority level we show that SVR models still give the best predictions of LGD and the superiority is even more evident for the bonds of lower seniority. We find SVR models consistently give more accurate predictions of LGD for corporate bonds than the other statistical models in the literature, and demonstrate that the SVR models are easy to be generalized to account for the heterogeneity of bond seniorities and the proposed improved SVR models are shown to present improvements compared with the original SVR models. We suggest that SVR models are a promising alternative to the traditional regression methods for modelling LGD of corporate bonds.

The second major contribution is based on an empirical study related to the impacts of firm specific heterogeneity illustrated in Chapter 5. It presents that an obligor-varying single latent factor model is significantly better than other models in terms of model fit. This is the first study showing that the single latent factor model, which has been accepted to be a benchmarking methodology for modelling PD of credit portfolios, can effectively explore the effects of firm specific heterogeneity on corporate bonds. It shows that the unobservable heterogeneity has a pivotal role in the LGD modelling for corporate bonds which has never been reported in the literature. Different from the first study which models LGD from a purely methodological perspective, the second study seeks to improve the model fit for corporate bonds by investigating heterogeneities at multiple levels. It is found that the obligor-varying single factor model shows the best model fit but the seniority level heterogeneity single factor model is not able to achieve significant improvements. We suggest it is due to that a large proportion of variations of LGD have been explained with the inclusion of firm specific heterogeneity. This study also shows that the time-varying information has been embedded in the predicted obligor-varying factors. The empirical results of out-of-sample predictive performances have also been studied and are found to be consistent with the evidence related to model fit. Our study fills the gap that firm specific heterogeneity should be accounted for modelling LGD of corporate bonds. We also contribute to literature by simulating the portfolio losses

based on the estimated LGD models and discuss the implications on regulatory capital. We find that the credit losses are significantly underestimated if an FIRB approach is adopted.

The third contribution is made to LGD modelling for retail credit cards. This study proposes a new two-stage model that incorporates the SVM techniques to the classification stage in order to separate the extreme cases at the boundaries of 0 and 1 from the others more accurately. This new model differs from the previous ones in that it is powerful to classify the extreme cases leading to more satisfactory predictive performances based on a loss data set of UK retail credit card. The study in Chapter 4 has proved the advantage of SVM techniques as a regression model. The study in this chapter contributes to the literature by applying SVM techniques to a hybrid modelling framework as a classification model for the first time, and it suggests that this new two-stage model with SVM techniques presents significant improvements in LGD predictions. In contrast, the two-stage models with a logistic regression do not exhibit any advantage compared with the single-stage methods. We also find that both account level and macroeconomic variables are significantly related to LGD but the incorporation of macroeconomic variables barely contributes to the model fit. The major implication in this study is that the classification accuracy at the first stage in the two-stage models is strongly correlated with the overall predictive accuracy, which is confirmed by examining the AUC of the classification methods. We show that SVM is substantially better than logistic regression and thus generate more accurate probabilistic output which has never been reported in the literature. This study suggests that the choice of algorithms at the classification stage of two-stage modelling framework influences the model performances more than the choice of regression models.

7.3. Implications

It is important to understand the implications arising from Basel II outputs due to their impact on risk weighted assets (RWAs) and capital calculations. The calculations of risk parameters affect setting credit risk strategies such as scorecard cut-offs and pricing credit products. In addition, it is also essential for regulators to monitor and review the impacts of risk models submitted by financial institutions. In the following the implications of our research are presented from the perspectives of both practitioners and regulators.

7.3.1. Implications for practitioners

First this research gives a new angel to develop a better LGD model. It suggests that machine learning techniques are a promising and competitive alternative to statistical regression models for estimating credit risk of both commercial banking business such as corporate bonds and retail business such as credit cards. The SVM techniques can either be applied to fitting LGD directly or be incorporated into a two-stage modelling framework combined with other regression methods which both demonstrates significant improvements according to the empirical results. It is also beneficial for banks to consider developing new algorithms based on SVM techniques.

Secondly this research shows that it is necessary to account for heterogeneity when modelling the LGD for an aggregated portfolio. Usually segmentation is applied to avoiding the issue of heterogeneity in practical business. Our studies have found that when the heterogeneity is considered, the predictive accuracy of LGD can be improved significantly for an aggregated portfolio. In terms of SVM techniques two different versions of improved models are proposed which both outperform the original SVM model. A more detailed study is conducted to explore the heterogeneity at multiple levels using a mixed effects regression model and it is found that the firm specific unobservable heterogeneity can explain large amounts of variations of recovery rates. Our studies strongly support the inclusion of firm specific heterogeneity for LGD modelling of corporate bonds.

Finally it is a key requirement for banks to understand implications when setting risk appetite. For the same exposures, Basel II methodologies deliver different RWA numbers compared with Basel I. The RWAs are dependent on the PD, LGD and EAD, and these will vary at business unit level. It is noted that the estimate of LGD is as important as the estimate of other risk parameters for banks to calculate the RWAs and hence for regulatory capital calculations for Pillar 1 credit risk, although the actual capital is unlikely to be reduced at least in the short term due to capital requirements for other risks. According to the empirical evidence in Chapter 5, the FIRB approach tends to underestimate the unexpected losses under extreme conditions compared with the LGD model developed in our study. Therefore, it is important to develop a sound LGD model to calculate RWAs properly so that the banks can allocate proper capital to buffer the unexpected losses.

7.3.2. Implications for regulators

Similar to practitioners in the financial institutions, regulators also need to understand the impacts of LGD models. First it is necessary for the regulators to urge banks to better manage the collection process to guarantee the data quality. Since more sophisticated approaches have been proposed to be the internal models in banks, the reliability of model performances is highly dependent on data quality. Our research finds that the inclusion of unobservable heterogeneity contributes to the LGD model fit significantly, but it would be more beneficial to have more information related to the instrument, contractual, and repayment characteristics.

Second it is necessary to review the estimation of downturn LGD. The new Basel Accord suggested that a mapping function could be proposed and applied to extrapolating downturn LGD based on long-term average LGD, or the downturn LGD could be estimated internally during the downturn conditions subject to supervisory standards (Basel Committee, 2005a). Our research suggests that the downturn LGD should be simulated based on a group of correlated stressed scenarios of economic factors, indicating the importance of incorporating macroeconomic variables into the LGD models. To give a conservative estimate of LGD, our research also suggests that the values of LGD specified in the FIRB approach may underestimate the LGD for lower seniority corporate bonds. It is recommended that the regulators should review and set a more proper value of LGD of portfolios for banks that adopt the FIRB approach.

Finally we suggest that regulators should review the approaches of treating zero and full recovery cases. This research shows that the performances of two-stage models are dependent on how well those cases at boundaries 0 and 1 can be separated from the remaining ones. The more accurately the zero and full recovery cases can be identified, the better predictive accuracy of the two-stage models can be expected. Our research raises the awareness that those zero and full recovery cases are somewhat special, and suggest that the regulators should consider the issues such as if it is appropriate to include them into the sample for modelling, and if any special attention should be paid conditional they are included.

7.4. Limitations

In spite of the efforts have been made in this research the limitations should be addressed in this section. First it is noted that our studies only focus on methodological development instead of discovering significant variables. However,

the selection of covariates plays a crucial role in modelling LGD for both corporate bonds and bank loans. In our empirical studies the variables are chosen based on literature and statistical significance, but we have not attempted to explore other potential significant observable variables for LGD modelling. Meanwhile it should be noted that banks are required to estimate LGD on a historical data set with a minimum of five years according to the new Basel Accord. The data set related to corporate bonds in Chapters 4 and 5 satisfies that requirement, but the time period of the retail credit cards data in Chapter 6 only covers one year although the sample size is large enough. Hence it is difficult to find a significant contribution of the inclusion of the economic variables for modelling LGD of credit cards.

Next we address the issue of model transparency in this research. The SVM models investigated in our studies are a non-parametric techniques which apply a kernel trick to mapping the original sample points into a high-dimensional space. The advantage of the kernel trick is that it enables the samples that are not linearly separable in a low-dimensional space to be linearly separable in a high-dimensional space. The principle of structural risk minimization also guarantees that SVM can effectively avoid the over-fitting problem. However, the kernel trick is similar to a multiple non-linear transformation that makes the model hard to explain. The black box characteristic of SVM is similar to neural networks which may make them less attractive to the risk analysts and modellers in industry. It is not surprising that the traditional statistical regression models such as ordinary linear regression and logistic regression still dominate the risk models of most banks because they are easy to implement and explain, and compliant with regulations. However, the two-stage modelling that incorporates SVM techniques shows its advantage over other methods while preserving relatively good explanatory power, indicating that the hybrid modelling framework that combines machine learning and statistical models together could be more promising.

Modelling LGD for non-defaults is another limitation in this research. The LGD models built are all based on the defaulted accounts which actually take up a very small percentage in the whole sample for a given product. Theoretically it is not appropriate to apply the LGD models that we have estimated to predicting LGD for non-defaulted accounts because it may lead to biased estimates of LGD. This research has not pursued to find out how to predict LGD for non-defaulted accounts as we only discuss modelling LGD of defaulted accounts. Until now very limited literature from

academia or industry that has made efforts to solve this problem although it seems to be of high importance.

A further limitation is that we have not explained modelling PD and LGD correlation, which is beyond the scope of this thesis. The LGD models discussed in our studies are built alone without considering the impacts of PD. A limitation of our data is that we are unable to build models to estimate PD and LGD point-in-time correlation although the correlation for long term default rates and loss rates can be calculated based on the historical data.

7.5. Further study

Based on the above discussions we address several topics for further study. First, it is necessary to further improve the SVM techniques to enhance the explanatory power. Data mining techniques such as neural networks and SVMs suffer the problem of being a black box making them less favoured by practitioners. Therefore it has been increasingly crucial to develop a more transparent machine learning model while preserving its superior predictive power. It is also interesting to explore more methodologies for LGD modelling. We have investigated a collection of algorithms including statistical regression models and machine learning techniques in our research whereas this is far from exhaustive. There are still some other potentially suitable and powerful choices for LGD modelling which are worth exploring. For example, Zhang and Thomas (2009) discussed the applications of survival regression models to estimating LGD for credit cards. Loterman et al (2011) gave a detailed benchmarking study for some 24 techniques including machine learning techniques such as neural networks and statistical methods. Tong et al (2013) developed a new zero inflated gamma regression model which is similar to the inflated beta regression but formulated in a semi-parametric way. The new proposed ideas in recent literature are shown to be promising for LGD modelling of mortgage loans and credit cards. We believe it worth exploring the applications of these methodologies to corporate bonds or other commercial loans.

Second, it is of great importance to discover useful determinants of LGD for different products. Although plenty of studies have been conducted to identify the significant determinants of LGD or recovery rate for various products, it is still worth more efforts to find out more potential useful variables due to the low model fit presented in the empirical studies of LGD modelling. This research covers two

products including corporate bonds and credit cards. For corporate bonds, it is believed that the industry and other contractual characteristics are also significantly correlated with the LGD (Archarya et al, 2007). Loan characteristics have been found to be most significant to LGD for bank loans portfolio according to the study in Khieu et al (2012). We suggest it could be useful to include repayment behaviours to predict the ultimate recovery rate of retail loans such as credit cards if possible. Other factors such as the actions that a bank has taken during recovery process are also considered to be helpful but they are more difficult to obtain.

Last but not least, modelling PD and LGD correlation for retail loans is another important but challenging topic in credit risk modelling. It is well accepted that PD and LGD are positively correlated and this correlation influences the estimate of portfolio losses significantly. It is necessary to include PD/LGD correlation to capture the credit cycle when estimating portfolio losses so that banks can allocate the regulatory capital accurately. Altman et al (2005) has highlighted the issue of pro-cyclicality of capital requirements, and argued that this effect tends to be exacerbated by the correlation between PD and LGD. In other words, when the economy goes down, the increase of PD combined with the rise of LGD will lift the capital charges and thus limit the credit supply. However, literature related to this topic has all been devoted to corporate bonds due to the data availability, and it is still unclear what the impact of PD and LGD correlation is for retail loans. Some new challenges are expected when modelling PD and LGD correlation on retail loans. For example the asset pricing models that have been applied to corporate bonds are no longer appropriate for retail loans since there is no market price of a retail loan. New methodologies should be proposed to model the PD and LGD correlation while remains to be applicable to retail loans. Also the sample size of a typical retail portfolio is much larger than commercial loans or corporate bonds where special efforts might be needed to calibrate the model.

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.

Altman, E., Brady, B., Resti, A. & Sironi, A. (2005). The link between default and recovery rates: Theory, empirical evidence, and implications, *Journal of Business*, 78(6), 2203-2227.

Altman, E., (2006). Default recovery rates and LGD in credit risk modeling and practice, working paper, Stern School of Business, New York University.

Altman, E. & Kalotay, E., (2010). A flexible approach to modeling ultimate recoveries on defaulted loans and bonds, working paper, Stern School of Business, New York University.

Bade, B., Rosch, D. & Scheule, H. (2011). Default and recovery risk dependencies in a simple credit risk model. *European Financial Management*, 17(1), 120-144.

Baesens, B., Gestel, van., Viaene, S., Stepanova, M., Suykens, J & Vanthienen, J.. (2003). Benchmarking state-of-art classification algorithms for credit scoring, *Journal of the Operational Research Society*, 54, 627-635.

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.

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, (2006). Basel II: International convergence of capital measurement and capital standards: A revised framework.

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

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.

Bijak, K. & Thomas, L. (2014). Modelling LGD for unsecured retail loans using Bayesian methods. *Journal of Operational Research Society*, 1-11.

Bluhm, C., Overbeck, L. & Wagner, C., (2003). An introduction to credit risk modeling. Chapman & Hall/CRC.

Bruche, M. & Gonzalez-Aguado. Carlos., (2010). Recovery rates, default probabilities, and the credit cycle. *Journal of Banking and finance*, 34, 754-764.

Calabrese, R. (2014). Predicting bank loan recovery rates in a mixed continuous-discrete model, *Applied Stochastic Models in Business and Industry*, 30(2), 99-114.

Calabrese, R. & Zenga, M. (2010). Bank loan recovery rates: Measuring and nonparametric density estimation. *Journal of Banking and Finance*, 34, 903-911.

Caselli, S., Gatti, S., & Querci, F., (2008). The sensitivity of the loss given default rate to systematic risk: New empirical evidence on bank loans. *Journal of Financial Service Research*, 34, 1-34.

Chalupka, R. & Kopecsni, J. (2009). Modeling bank loan LGD of corporate and SME segments: A case study. *Czech Journal of Economics and Finance*, 59(4), 360-382.

Chava, S., Stefanescu, C. & Turnbull, S., (2011). Modeling the loss distribution. *Management Science* 57(7), 1267-1287.

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

Dermine, J. & Neto De Carvalho, C. (2008). Bank loan-loss provisioning, central bank rules vs. estimation: The case of Portugal. *Journal of Financial Stability*, 4, 1-22.

Dullmann, K. & Trapp, M., (2004). Systematic risk in recovery rates-an empirical analysis of US corporate credit exposures. Deutsche Bundesbank Discussion Paper.

Duffie, D., Eckner, A., Horel, G. & Saita, L., (2009). Frailty correlated default. *Journal of Finance*, 64(6), 2089-2132.

Dwyer, D. & Korablev, I., (2009). Moody's KMV LossCalc V3.0. Moody's Analytics.

Ferrari, S. & Neto, F. (2004). Beta regression for modeling rates and proportions, *Journal of Applied Statistics*, 31(7), 799-815.

Francis, E. H. T. & CAO, L., (2001). Application of support vector machines in financial time series forecasting. *Omega: The International Journal of*

Management Science, 29(4), 309-317.

Frye, J., (2000a). Collateral damage: A source of systematic credit risk. *Risk*, 91-94.

Frye, J., (2000b). Depressing recoveries. *Risk*, 108-111.

Frye, J. & Jacobs Jr., M. (2012). Credit loss and systematic loss given default. *Journal of Credit Risk*, 8(1), 109-140.

Fung, G. & Mangasarian, O. L. (2001). Proximal support vector machine classifiers. In F. Provost & R. Srikant (Eds.), Proceedings KDD-2001: Knowledge discovery and data mining, 77–86. San Francisco, CA, New York: Asscociation for Computing Machinery. ftp://ftp.cs.wisc.edu/pub/ dmi/techreports/01-02.ps.

Gordy, M., 2003. A risk-factor model foundation for rating-based bank capital rules. *Journal of Financial Intermediation*, 12, 199-232.

Grunert, J. & Weber, M. (2009). Recovery rates of commercial lending: Empirical evidence for German companies. *Journal of Banking and Finance*, 33, 505-513.

Gupton, G. M. & Stein, R. M. (2002). LossCalc^{TM}: Model for Predicting Loss Given Default (LGD). Moody's KMV.

Hamerle, A., Knapp, M. & Wildenauer, N. (2006). Modeling loss given default: A "point in time"-approach, in The Basel II risk parameters: Estimation, validation, stress testing – with applications to loan risk management, edited by Engelmann, B. & Rauhmeier, R..

Hamerle, A., Liebig, T. & Rosch, D. (2003a). Credit risk factor modeling and the Basel II IRB approach, discussion paper, Series 2: Banking and Financial Superivision.

Hamerle, A., Liebig, T. & Rosch, D. (2003b). Benchmarking asset correlations, *Risk*, 77-81.

Hillebrand, M., (2005). Modeling and estimating dependent loss given default. *Risk*, 120-125.

Huang, X. & Oosterlee, C. W., (2011). Generalized beta regression models for random loss given default. *Journal of Credit Risk*, 7(4), 1-26.

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

Jankowitsch, R., Nagler, F. & Subrahmanyam, M. (2014). The determinants of recovery rates in the US corporate bond market. *Journal of Financial Economics*, 114(1), 155-177.

Jokivuolle, E. & Peura, S. (2000). A model for estimating recovery rates and collateral

haircuts for banks loans, Bank of Finland Discussion Papers.

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.

Li, H. (2009). On models of stochastic recovery for base correlation, working paper. Online at http://mpra.ub.uni-muenchen.de/15750/, MPRA Paper No. 15750.

Loterman, G., Brown, I., Martens, D., Mues, C. & Baesens, B. (2011). Benchmarking Regression Algorithms for Loss Given Default Modeling. *International Journal of Forecasting*, 28(1), 161-170.

Matuszyk, A., Mues, C. & Thomas, L. (2010). Modelling LGD for unsecured personal loans: Decision tree approach. *Journal of Operational Research Society*, 61, 393-398.

Merton, R. C., (1974). The pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, 29, 449-470.

Miu, P. & Ozdemir, B. (2006). Basel requirements of downturn loss given default: Modeling and estimating probability of default and loss given default correlations. *Journal of Credit Risk*, 2(2), 43-68.

Moody's Analytics (2012). Default & Recovery Database DRD Technical Specifications.

Morone, M. & Cornaglia, A. (2010). An econometric model to quantify benchmark downturn loss given default on residential mortgages. *Journal of Risk Model Validation*, 4(3), 27-51.

Ospina, R & Ferrari, S., (2010). Inflated beta distributions. *Statistics Papers*, 51, 111-126.

Ou, S., (2013). Annual default study: Corporate default and recovery rates, 1920-2012. Moody's Investors Service.

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.

Peura, S. & Jokivuolle, E. (2005). LGD in a structural model of default, in Recovery Risk: The next challenge in credit risk management, edited by Altman, E., Resti, A. & Sironi, A.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, in Advances in Large Margin Classifiers, Smola, A., Bartlett, P., Scholkopf, B. & Schuurmans, D. MIT Press.

Pykhtin, M., (2003). Unexpected recovery risk. *Risk*, 74-48.

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 Modeling 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.

Renault, O. & Scaillet, O. (2004). On the way to recovery: A nonparametric bias free estimation of recovery rate densities. *Journal of Banking and Finance*, 28, 2915-2931.

Rosch, D. & Scheule, H., (2004). Forecasting retail portfolio credit risk. *Journal of Risk Finance*, 5, 16-32.

Rosch, D. & Scheule, H., (2005). A multi-factor approach for systematic default and recovery risk. *Journal of Fixed Income*, 15(2), 63-75.

Rosch, D. & Scheule, H., (2008). The empirical relation between credit quality, recoveries, and correlation in a simple credit risk model, working paper, http://www.efmaefm.org/0EFMAMEETINGS/EFMA%20ANNUAL%20MEETINGS/2009-Milan/papers/19.pdf.

SAS Institute Inc (2009). SAS 9.2 User's Guide, Cary, NC.

Schonbucher, P. J., (2001). Factor models for portfolio credit risk. *Journal of Risk Finance*, 3, 45-56.

Schuermann, T. (2004). What do we know about loss given default, working paper.

Seidler, J. & Jakubik, P. (2009). Implied market loss given default in the Czech Republic. *Czech Journal of Economics and Finance*, 59(1), 20-40.

Shim, J., Kim, C. & Hwang, C. (2011). Semiparametric least squares support vector machine for accelerated failure time model. *Journal of the Korean Statistical Society*, 40, 75-83.

Shivaswamy, P. K., Chu, W & Jansche, M., (2007). A support vector approach to censored targets, in *Proceedings of IEEE International Conference on Data*

Mining (ICDM-2007).

Shumway, T., (2001). Forecasting bankruptcy more accurately: A simple hazard model. *Journal of Business*, 74(1), 101-124.

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

Suykens, J. A. K. & Vandewalle, J. (1999). Least Squares Support Vector Machine Classifiers. *Neural Processing Letters*, 9, 293-300.

Suykens, J. A. K., Gestel, T. V., De Brabanter, J., De Moor, B. & Vandewalle, J. (2002). *Least Squares Support Vector Machine*, Singapore: World Scientific.

Tasche, D., (2004). The single risk factor approach to capital charges in case of correlated loss given default rates. Deutsche Bundesbank Working Paper.

Tobback, E., Martens, David., Gestel, T & Baesens, B. (2014). Forecasting loss given default models: Impact of account characteristics and the macroeconomic state. *Journal of Operational Research Society*, 65, 376-392.

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.

Yashkir, O. & Yashkir, Y. (2013). Loss given default modeling: A comparative analysis. *Journal of Risk Model Validation*, 7(1), 25-59.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. New York: Springer-Verlag.

Vapnik, V. (1998). Statistical Learning Theory, New York: John Wiley.

Varma, P. & Cantor, R. (2004). Determinants of recovery rates on defaulted bonds and loans for North American corporate issuers: 1983-2003, Moody's Investors Service Special Comment.

Van Belle, V., Pelckmans, K., Suykens, J., & Van Huffel, S. (2010). Additive survival least-squares support vector machines, *Statistics in Medicine*, 29, 296-308.

Van Gestel, T., Baesens, B., Garcia, J. & Dijcke, P van, (2003). A support vector machine approach to credit scoring, Bank en Financiewezen, 2, 73-82.

Van Gestel, T.. Van Suykens, J. K, Baestaens, D. E, Lambrechts, A., Lanckriet, G. Vandaele, B., De Moor, B. & Vandewalle, J. (2001). Financial time series prediciton using least squares support vector machines within the evidence framework, *IEEE Transactions on Neural Networks*, 12(4), 809-821.

Vasicek, O. (1987). Probability of loss on loan portfolio. KMV Corporation.

Vasicek, O. (2002). Loan portfolio value. *Risk*, 160-162.

Yang, B. & Tkachenko, M. (2012). Modelling exposure at default and loss given default: Empirical approaches and technical implementation. *Journal of Credit Risk*, 8(2), 81-102.

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

Appendices

Appendix A. Definitions of Value-at-Risk (VaR) and Expected Shortfall (ES)

Given a confidence level $q \in (0,1)$, the VaR and ES of a loss distribution are defined as

$$\begin{aligned} \mathrm{VaR}_q(L) &= \min\{l \mid P(L > l) \le 1-q\} \\ \mathrm{ES}_q &= E(L \mid L > \mathrm{VaR}_q) \end{aligned},$$

where l is the smallest value such that the probability that the loss rate L exceeds l is $1-q$ at most. Here we use the empirical quantile as the estimate of VaR such that

$$\mathrm{VaR}_q(L) = \hat{L}_q,$$

where $\hat{L}_q$ satisfies that $P(L \ge \hat{L}_q) = 1-q$. Given the estimate of VaR, $\hat{L}_q$, we can derive the empirical estimate of the ES as follows:

$$\mathrm{ES}_q = \frac{1}{N_q}\sum_{j=1}^{M} L_j I\left[L_j > \hat{L}_q\right],$$

where L_j denotes the loss rate simulated at *j*-th iteration, and $I[\bullet]$ is an indicator function, and N_q is the count that the loss rate L exceeds $\hat{L}_q$.

Appendix B. Publications

ARTICLE IN PRESS

European Journal of Operational Research xxx (2014) xxx–xxx

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 xxxx

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/

Please cite this article in press as: Yao, X., et al. Support vector regression for loss given default modelling. *European Journal of Operational Research* (2014), http://dx.doi.org/10.1016/j.ejor.2014.06.043

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.

Please cite this article in press as: Yao, X., et al. Support vector regression for loss given default modelling. *European Journal of Operational Research* (2014), http://dx.doi.org/10.1016/j.ejor.2014.06.043

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 α_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 jth sample in the kth 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

Please cite this article in press as: Yao, X., et al. Support vector regression for loss given default modelling. *European Journal of Operational Research* (2014), http://dx.doi.org/10.1016/j.ejor.2014.06.043

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

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

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

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.

Please cite this article in press as: Yao, X., et al. Support vector regression for loss given default modelling. *European Journal of Operational Research* (2014), http://dx.doi.org/10.1016/j.ejor.2014.06.043

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

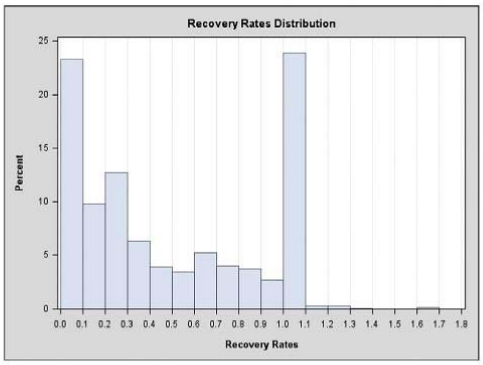


Fig. 1. Distribution of recovery rates.

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

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

where subscript i denotes the ith 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

Please cite this article in press as: Yao, X., et al. Support vector regression for loss given default modelling. *European Journal of Operational Research* (2014), http://dx.doi.org/10.1016/j.ejor.2014.06.043

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	σ	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 RMSE_i and 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.

Please cite this article in press as: Yao, X., et al. Support vector regression for loss given default modelling. *European Journal of Operational Research* (2014), http://dx.doi.org/10.1016/j.ejor.2014.06.043

(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	–

(continued on next page)

Please cite this article in press as: Yao, X., et al. Support vector regression for loss given default modelling. *European Journal of Operational Research* (2014), http://dx.doi.org/10.1016/j.ejor.2014.06.043

(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	–

(continued on next page)

Please cite this article in press as: Yao, X., et al. Support vector regression for loss given default modelling. *European Journal of Operational Research* (2014), http://dx.doi.org/10.1016/j.ejor.2014.06.043

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.

Please cite this article in press as: Yao, X., et al. Support vector regression for loss given default modelling. *European Journal of Operational Research* (2014), http://dx.doi.org/10.1016/j.ejor.2014.06.043