MACQUARIE
University
SYDNEY·AUSTRALIA

EMPIRICAL STUDIES IN DEFAULT AND INSURANCE RISK

By

Feng Liu

Principal Supervisor: Prof. David Pitt

Associate Supervisor: Prof. Stefan Trück

A thesis submitted in fulfilment of the requirements for the degree of

PhD

in the

Faculty of Business and Economics

Department of Actuarial Studies and Business Analytics

September 13, 2018

Contents

Declaration

I hereby declare that this work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Feng Liu, Sydney, 13 September 2018

Acknowledgements

Foremost, I express my sincere gratitude to my thesis supervisors Professor David Pitt and Professor Stefan Trück for the continuous support of my PhD study and research, and for their patience, engagement, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. In particular, I would like to thank Professor David Pitt for his priceless advice both on research as well as on my career, and all the encouragements and comforting words during low times. I could not have imagined having a better mentor and I will always cherish everything I have learnt from him. I also want to thank Professor Stefan Trück for his valuable inputs, especially his generous help with the publication of my second paper. He has also put me forward for all kinds of academic activities, and has written too many reference letters for me. My two wonderful supervisors have made a great difference in my life. I will always be grateful for everything they have done for me.

I am also thankful to my beloved friends, for putting up with my worries and concerns. In particular, I want to thank my fellow PhD candidates, for all the conversations on the corridor and in the kitchen. I wouldn't be able to make this far if not for your friendship and support.

I would also like to thank my loving family: mother Qiaoyun, father Shixiang, and my big sister Liu, for their continuous and unconditional love, and never asking about my progress; my darling daughter who has been the sweetest angel, for managing to sleep through the night ever since she was one month and her lovely smiles that can always cheer me up.

A special thank you goes to my husband, for his constant love and faith in me. I am truly thankful for having you in my life.

MACQUARIE UNIVERSITY

Abstract

Faculty of Business and Economics
Department of Actuarial Studies and Business Analytics

PhD

EMPIRICAL STUDIES IN DEFAULT AND INSURANCE RISK

by Feng Liu

This PhD thesis evaluates three types of financial risks with analytical methods that improve the accuracy when making predictions for future risk events. It consists of three key chapters based on three research papers.

The research paper titled '*Application of the bivariate negative binomial regression model in analysing insurance count data*' analyses insurance claim frequency data using the bivariate negative binomial regression (BNBR) model, and claims data from third-party liability and comprehensive motor insurance. It is found that bivariate regression, with its capacity for modelling correlation between the two observed claim counts, provides a superior fit and improved out-of-sample prediction compared to the more common practice of fitting univariate negative binomial regression (UNBR) models separately to each claim type. Noting the complexity of BNBR models and their potential for a large number of parameters, this study explores the use of model shrinkage with the Lasso and ridge regression. Results show that our models estimated using shrinkage methods outperform the ordinary likelihood-based models when being used to make predictions out-of-sample. It can also be shown that the Lasso performs better than ridge regression as a method of shrinkage in the context considered here.

The research paper titled '*Assessing Sovereign Risk: A Bottom-Up Approach*' assesses sovereign default risk of individual states in the U.S. using information about default risk at the company level. The integrated risk factors of the private sector are linked to the overall sovereign risk of state governments in conjunction with additional financial variables. Using data from Moody's KMV expected default frequencies (EDFs) on corporate default risk, credit risk indicators for different industries are estimated. Building on these measures, state level credit risk indicators are developed encompassing industry compositions to explain the behaviour of credit default swap (CDS) spreads for individual states. It is found that market-based measures of private sector credit risk are strongly associated with subsequent shifts in sovereign credit risk premiums measured by CDS spreads. The credit risk indicators developed here are demonstrated to be highly significant in forecasting sovereign CDS spreads at weekly and monthly sampling frequencies.

The research paper titled '*A joint model for longitudinal and time-to-event data in corporate default risk modelling*' applies a joint model for longitudinal and time-to-event data to assess corporate default risk. A linear mixed-effects model is used to describe the trajectory of the predictor variable for company default. The output from this analysis is then used in a survived model where time to corporate default in the response variable of interest. The joint model does not assume constant values for the independent variable between observations, and can take advantage of the fully-specified subject-specific longitudinal trajectories. Data collected on U.S. listed companies from 1997 to 2016 is used to test the ability of the joint model to predict corporate default events. Two independent variables, the distance-to-default and the age of a company, are used to assess the company's probability of default over various time horizons. It is found that the joint models outperform the Cox model and the Weibull model in making predictions of default events. The results show that the joint model is more suitable in assessing corporate default risk than the selected standard survival models.

List of Figures

List of Tables

Chapter 1

Introduction

This thesis analyses three types of financial risks in a quantitative framework. Applied in empirical studies, the proposed analytical methods evaluate the financial risks in novel ways and help improve the accuracy when making predictions for future risk events. The results presented in this thesis contribute to the risk analysis discipline by modifying and enhancing established approaches for better risk management.

Financial risk arises due to instabilities in the financial markets and uncertainty relating to future financial events. It is one of the major concerns of businesses, individuals, government and regulators. In order to anticipate the potential impact of such risk, it is of great importance to evaluate the risk and make reasonable predictions for better decision-making. Output from this analysis can inform business strategies, investment options or the implementation of new regulations to protect the stability of the financial system.

As a result, the topic of financial risk modelling and management has been intensively studied in the literature. An accumulation of models have been developed to identify different financial risks and their sources, and forecast the probability of risk events using selected variables (Campbell *et al.*, 1997). Financial institutions, including banks, credit rating companies and insurance companies, also adopt and implement internal risk management systems, to evaluate their key risk exposure in preparation for a plan to react to adverse events (Saunders & Cornett, 2003).

Among the many different types of financial risks, this thesis focuses on the assessment of the following three risks using quantitative methods, each of which is addressed in one research paper:

- Insurance claim count risk
- Sovereign credit risk
- Corporate default risk

This chapter gives an introduction and an overview of the thesis. Section 1.1 discusses these financial risks and relevant analytical models in the current literature. Section 1.2 describes the structure of the thesis and the contributions made by each of the research papers.

1.1 Financial risk modelling and techniques

1.1.1 Insurance claim count risk

Insurance claim count risk refers to uncertainty in the number of claims made on an insurance company for a particular line of insurance business. Managing insurance claim count risk is crucial, as it is important for insurance companies to differentiate between low-risk and high-risk customers. Except for health insurance companies which operate under a community rating system, other insurance companies can discriminate between customers by charging different premiums, which reflect the perceived risk level of the customer based on the information collected about the customer. Over- or under-charging customers is not only detrimental to business profit, but also leads to problems such as under-reserving and adverse selection.

Often the predicted number of claims is analysed with the aid of policyholder's characteristics. For example, for motor insurance the claim frequency is commonly linked to the policyholder's age, driving history, living address and so on. Modelling insurance claim count data concerns the relationship between the observed counts and such explanatory variables.

Generalised linear models (GLMs) are commonly used as a mathematical formulation for this modelling purpose. A linear combination of explanatory variables, representing various characteristics of the policyholder, is used as the basis for estimating the expected claim count for this policyholder. The model can be calibrated based on the data of existing policyholders, in order to assess the risk level of new customers. Common distributions used in GLMs to analyse count data are the Poisson distribution and the negative binomial distribution (McCullagh & Nelder, 1989). The former assumes the mean and the variance of the response variable are equal, and the latter allows the variance to be higher than the mean. When the interest lies in the analysis of two types of policies held by the policyholder, a bivariate model with a properly designed correlation structure can be used as the claim counts of the two policies are likely to be correlated.

In empirical studies, it is important to choose the right distribution in the GLM and the association structure if two types of policies are studied. The other issue in analysing the claim count is the number of independent variables used in the model, as it is not always optimal to include as many variables. If the problem of over-fitting is present, this may lead to reduced out-of-sample prediction accuracy. Thus it may be necessary to shrink the model and only use the most important independent variables.

1.1.2 Sovereign credit risk

Sovereign credit risk is defined as the risk for a sovereign entity to default on its debt payments. Sovereign default events are infrequent but can lead to financial crisis or recession. Due to the serious consequences of such default events, managing sovereign risk is not only important for the investors who hold sovereign debts in their investment portfolio, but also for the sovereign government who issues sovereign debt instruments. The significance of this assessment has increased substantially since the global financial crisis when several European countries encountered periods of financial difficulties.

Sovereign risk is normally linked to the sovereign entity's fiscal condition, such as the budget deficit, trade balance and tax receipts. Economic factors can also help analyse sovereign risk, such as the GDP growth rate, inflation rate and the unemployment rate. The deterioration in these figures or ratios can help predict default events. However, most of these figures are only available at a low frequency,

released by the government's department of statistics with delays. So even if the model that links the figures to the sovereign risk is correctly specified, it is difficult for such model to give early warnings and provide time for reactions and responses. No matter how sophisticated the model is, the prediction result is not likely to be indicative.

As an attempt to mitigate the problem, a bottom-up approach is proposed in Altman & Rijken (2011*a*) that links the financial health of the sovereign entity to the health of its private sector, which is justified by the fact that the major source of income for the sovereign entity is the tax receipts collected in the private sector. This approach is further improved in this thesis, with more frequently updated and forward-looking market data as the inputs and focusing on state governments without existing financial problems.

1.1.3 Corporate default risk

Corporate default risk is defined as the risk for a corporation to default on its debt payments, which may be followed by the liquidation and bankruptcy of the company. Compared to sovereign default events, corporate default events are more frequent and thus provide ample data for model development. The literature in this discipline focuses on the modelling of the probability of default, which is linked to a range of independent variables that are believed to have explanatory power for the default event. Credit rating agencies are also contributors of the default analysis, assigning different ratings to companies based on probability of default computed using the internal models.

In general, there are three model-based approaches to the analysis of default risk. Structural models calculate the company's distance to default, which is defined as the number of standard deviations of the asset value between the mean asset value and the debt value. The larger the distance to default, the higher the chance to have adequate money for debt payments. Scoring models combine the company's accounting ratios in simple regression models to perform discriminant analysis. The dependent variable in the scoring model is the score for the company, which can be used for classification purposes. Because of their simplicity, scoring models are widely adopted in all types of credit risk analysis. Finally, survival models analyse the default intensity. The model assumes the default is a surprise event driven by exogenous random variables and the time to default is governed by the

hazard rate of default. The model automatically adjust for period at risk, which is one of the advantages over the structural model and the scoring model. These three types of models are discussed in detail in Chapter 4.

The thesis assesses corporate default risk using a joint model (Rizopoulos, 2012), which combines a linear mixed-effects (LME) model with a survival model, so that the longitudinal observations of the company is fitted with a smooth function, before it is linked to the hazard rate of default of the company. The model relaxes the assumption that the independent variables are constant between observations, and a well-modelled longitudinal trajectory helps increases the prediction accuracy as indicated by the results.

1.1.4 Statistical software

This thesis uses R to analyse data in empirical studies. It is used for data organisation, data processing and computation. The computation procedure involves fitting models to the data to estimate parameters, and generate predictions. Because the results of the proposed models are compared to benchmark models, R is also used to produce charts and tables for the presentation of the results.

One advantage of using R is the availability of a range of packages ready to be used for different computation purposes. The analysis of the data in this thesis is based on several main packages, including

- graphics (R Core Team, 2015*a*): For producing a variety of diagrams and plots.
- JMbayes (Rizopoulos, 2016*a*): The main package used in Chapter 4 for fitting the model and computing out-of-sample prediction results.
- MASS (Venables & Ripley, 2002*a*): A library of functions and datasets to support various calculations.
- nlme (Pinheiro *et al.*, 2017): A package for linear and nonlinear mixed-effects models, which is used together with JMbayes in Chapter 4 to fit a LME model.

- splines (R Core Team, 2015*b*): A package for regression spline functions and classes.
- stats (R Core Team, 2015*c*): A package containing functions for statistical calculations and random number generation. The function "lm" is used in Chapter 3 to fit regression models.
- survival (Therneau, 2015): A package for survival analysis, containing the basic routines such as the definitions of objectives in a survival model.

In addition to these packages, the author has also developed code as required in the analysis. As an example, part of the coding work involved in analysing the insurance claim data is presented in the appendices.

1.2 Thesis structure and contribution

The three main chapters of the thesis (Chapters 2-4) are based on three research papers, each of which focuses on one type of financial risk.

Chapter 2: *Application of the bivariate negative binomial regression model in analysing insurance count data*

This study analyses insurance claim frequency data using the bivariate negative binomial regression (BNBR) model, using general insurance data on claims from simple third-party liability insurance and comprehensive insurance. It is found that bivariate regression, with its capacity for modelling correlation between the two observed claim counts, provides a superior fit and improved out-of-sample prediction compared to the more common practice of fitting univariate negative binomial regression (UNBR) models separately to each claim type. Noting the complexity of BNBR models and their potential for a large number of parameters, this study explores the use of model shrinkage, namely the Lasso and ridge regression. Results show that models estimated using shrinkage methods outperform the ordinary likelihood-based models when being used to make predictions out-of-sample. It can also be shown that the Lasso performs better than ridge regression as a method of shrinkage in the context considered here.

This study contributes to the literature by demonstrating the importance of the BNBR model in analysing over-dispersed general insurance claims data, especially when interest lies in claim count data that are likely to be correlated. Moreover, the results obtained by shrinking the original models provide evidence that shrunken models provide higher out-of-sample prediction accuracy. This may due to the over-fitting problem of the original models where all explanatory variables are used.

Chapter 3: *Assessing Sovereign Risk: A Bottom-Up Approach*

This chapter assesses sovereign default risk of individual states in the U.S. using information about default risk at the company level. The integrated risk factors of the private sector is linked to the overall sovereign risk of state governments in conjunction with additional financial variables. Using data on Moody's KMV expected default frequencies (EDFs) on corporate default risk, credit risk indicators for different industries are derived. Building on these measures, state level credit risk indicators are developed encompassing industry compositions to explain the behaviour of credit default swap (CDS) spreads for individual states. It is found that market-based measures of private sector credit risk are strongly associated with subsequent shifts in sovereign credit risk premiums measured by CDS spreads. The developed credit risk indicators are highly significant in forecasting sovereign CDS spreads at weekly and monthly sampling frequencies.

The study contributes to the sovereign risk analysis literature by applying a novel approach that evaluates sovereign risk form a new perspective. The findings of the study suggest a strong predictive link between market expectations of private sector credit quality and expectations of sovereign credit quality - a connection that is not directly discernible from scoring models. Moreover, the study complements and extends earlier work on a bottom-up approach by using high-frequency forward-looking market data and analysing sovereign entities that are not selected with reference to their financial health to avoid survival bias.

Chapter 4: *A joint model for longitudinal and time-to-event data in corporate default risk modelling*

This chapter applies a joint model for longitudinal and time-to-event data to assess corporate default risk. The joint model analyses the independent variable in a linear mixed-effects model to assess the subject-specific time evolutions, before using it to evaluate event time and risk. The joint model does not assume

constant values for the independent variable between observations, and can take advantage of the fully-specified subject-specific longitudinal trajectories. Data collected on U.S. listed companies from 1997 to 2016 is used to test the ability of the joint model to predict corporate default events. Two independent variables, the distance-to-default and the age of a company, are used to assess the company's probability of default over various time horizons. It is found that the joint models outperform the Cox model and the Weibull model in making predictions of default events. The results show that the joint model is more suitable in assessing corporate default risk than selected standard survival models.

While the joint model is widely used in medical and biostatistical studies, it is to the best knowledge of the author that it is the first time that joint model is applied in credit risk analysis. It helps mitigate the problem of simply assuming the independent variable values are constant between observations. By better modelling the longitudinal trajectory of the independent variable, the joint model can better assess the relationship between the independent variable the probability of default, and thus making more accurate predictions.

To conclude the thesis, Chapter 5 summarises the main results from Chapters 2, 3 and 4 and the main contributions to the current literature. It also identifies several directions for future research in the area of financial risk analysis.

Chapter 2

Application of the bivariate negative binomial regression model in analysing insurance count data

Feng Liu (contribution 80%), David Pitt (contribution 20%)

A research paper based on this chapter has been published:

- Liu, F., & Pitt, D. (2017). Application of bivariate negative binomial regression model in analysing insurance count data. *Annals of Actuarial Science* 11(2), 390-411.

2.1 Introduction

We explore the use of a bivariate negative binomial regression (BNBR) model in the context of modelling bivariate insurance claim frequency data. Two types of insurance claims, the third party liability claim and the comprehensive cover claim, made by the same policyholder are assumed to be correlated and to be explained by a set of explanatory variables. By allowing a correlation between the two response variables, the performance of the BNBR is better than if two univariate negative binomial regression (UNBR) models are fitted separately, both in terms

of in-sample goodness-of-fit and out-of-sample prediction. We also find that the BNBR also outperforms the bivariate Poisson regression (BPR) model.

In addition, we apply two shrinkage techniques, the Lasso and ridge regression, to reduce the number of covariates used in the original unshrunken BNBR model. Although an increasing number of explanatory variables will increase in-sample goodness-of-fit, an overfitted model may result which performs less well in out-of-sample prediction. By selecting more relevant risk factors and removing unnecessary explanatory variables, we find that the shrunken models outperform the unshrunken model in out-of-sample prediction.

We use the model specification for BNBR in Famoye (2010*b*), where correlation structure allows for both a negative and a positive relationship between the two claim type frequencies.

The contributions of this chapter are threefold. First, we successfully demonstrate the importance of the BNBR model in analysing over-dispersed general insurance claim data, which outperforms the BPR model. Second, the correlation factor is found to be significant, with the implication that BNBR model is more suitable when the two claim counts are correlated. A similar conclusion is not evident in Famoye (2010*b*), where the correlation between the two variables considered is too low for useful dependence modelling, and thus univariate models seem to be adequate. Third, we shrink both BNBR models and UNBR models to reduce the size of coefficients of irrelevant explanatory variables, some of which are eliminated totally from the regression model. The shrinkage results are consistent with James *et al.* (2013, Chapter 6), in that the shrunken models provide much higher out-of-sample prediction accuracy, compared to the original full BNBR models.

The chapter is organised as follows: Section 2.2 gives a summary of existing methods to analyse claim counts, including univariate and bivariate generalised linear models. Section 2.3 describes the model used in this study as well as the shrinkage techniques. Section 2.4 introduces the claims data. Section 2.5 gives the modelling results and a discussion of findings. Section 2.6 concludes the chapter.

2.2 Literature review

Modelling of insurance claim count data has been an active area of research for some decades. The research interest often lies in modelling the relationship between the observed counts and a set of explanatory variables. Generalised linear models (GLMs) are very commonly used for this purpose as a mathematical formulation of the relationship. With a chosen link function, the mean of the distribution can be expressed as a linear function of the explanatory variables.

Under the GLM framework, the response variable is modelled using a member of the exponential dispersion family of distributions. Two common choices for this distribution in the case of insurance count data are the Poisson distribution and the negative binomial distribution (see McCullagh & Nelder, 1989). While the Poisson regression model assumes equality between the underlying mean and variance of the response variable, negative binomial regression relaxes the assumption and accounts for over-dispersion in the data (see Cameron & Trivedi, 2005). Both models have been widely adopted to analyse claim count data in general insurance. For a comprehensive review of the GLM, including different models and their specification, and applications and examples, refer to McCullagh & Nelder (1989).

An early example of the application of GLM in insurance modelling is Samson & Thomas (1987), where a GLM was applied to analyse claim costs for an automobile insurance account portfolio of a major British insurance company. They found that the categorical independent variables of policyholder age, area of residence, vehicle type, and no-claim discount (NCD) status were statistically significant predictors of claim costs. Dionne & Vanasse (1989) used both Poisson and negative binomial regression models for automobile insurance risk classification. Hürlimann (1990) studied the properties of maximum likelihood equations, based on the pseudo compound Poisson representation of any discrete distribution defined on the positive integers. Renshaw (1995) provides an overview of the potential of GLMs as a means of modelling the salient features of the claims process in the presence of rating factors. Haberman & Renshaw (1996) illustrated the use of the over-dispersed Poisson model in analysing life insurance claim counts, after presenting a summary of GLMs in actuarial science.

Various extensions to the basic GLM framework have been proposed in the statistics literature and explored in insurance contexts. For example, Generalised Additive Models (GAMs) are postulated by combining an original GLM with

additive models in the linear regression model, where smooth functions with semi-parametric or non-parametric forms are applied to explanatory variables. So with a chosen link function, the mean of the response variable is expressed as a linear function of unknown smooth functions of explanatory variables (see Hastie & Tibshirani, 1990). The GAM framework is adopted in Denuit & Lang (2004) to account for discrete, continuous, and spatial risk factors in a Bayesian framework for insurance ratemaking purposes. Mixtures of GLMs, such as Poisson mixtures, can be used to accommodate non-homogeneous populations (see Karlis & Xekalaki, 2005). More recently, increasing attention has been given to the application of extended GLMs in accounting for excess zeroes and over-dispersion in count data, especially for automobile insurance count numbers under no claim discount system. The proposed zero-inflated models are considered as a mixture of a zero point mass and a Poisson or negative binomial regression models under the original GLM framework. Yip & Yau (2005) provided a good summary of zero-inflated models with an application in general insurance count data. Heller *et al.* (2007) considered a group of candidate distribution to model claim counts, including Poisson, zero-inflated Poisson and negative binomial. Thorough reviews for count data regression can be found in Denuit *et al.* (2007) and Cameron & Trivedi (1998).

In addition to univariate models, bivariate regression models have been proposed to analyse two response variables that are possibly correlated. These models offer sufficient flexibility by allowing the two response variables to be affected by different predictive factors. Moreover, a bivariate model is more helpful for inference and prediction purposes because it allows us to properly specify the dependency between the two dependent variables (Shi & Valdez, 2014).

One way to introduce the correlation factor is to use copulas to analyse the correlation structure, by linking univariate marginals to the full multivariate distribution (see Frees & Valdez, 1998). The use of copulas is common in analysing correlation structure related to continuous variables such as claim losses. Denuit, Van Keilegom, Purcaru *et al.* (2006) used Archimedean copulas to analyse non-life insurance data, which was applied to an actual loss-ALAE (allocated loss adjustment expense) data set. Frees & Valdez (2008) adopted copula functions to specify the joint multivariate distribution of the claims arising from various claims types. They used two different copulas, the standard normal (Gaussian) copula and the t-copula. Czado *et al.* (2012) presented a mixed copula approach to allow

for dependency between the number of claims and its corresponding average claim size using a Gaussian copula.

In studying discrete variables such as the number of insurance claims, Cameron *et al.* (2004) used a bivariate copula in modelling the difference between self-reported and true doctor visits, but the application is limited to studying the distribution of the difference between two counts. Shi & Valdez (2014) considered three types of automobile claim counts using a mixture of copulas and the family of elliptical copulas. A review of using copulas to specify correlation structure can be found in a recent study by Chen & Hanson (2017).

Another group of studies analyse the correlation structure through the trivariate reduction method, where the pair of dependent variables are specified using three random variables. For example, by setting $Y_1 = X_1 + X_{12}$ and $Y_2 = X_2 + X_{12}$, where X_1, X_2 and X_{12} are independent Poisson random variables, Y_1 and Y_2 have a bivariate Poisson distribution with a covariance term derived from the use of the common Poisson variable X_{12} (see Kocherlakota & Kocherlakota, 1992; Johnson *et al.*, 1997).

The trivariate reduction method has been explored in many studies. For example, Jung & Winkelmann (1993) adopted a bivariate regression framework based on the trivariate reduction method for an analysis of data on two types of labour mobility. King (1989) proposed a joint Poisson regression estimator for the analysis of two contemporaneously correlated endogenous event count variables. Kocherlakota & Kocherlakota (2001) adopted the trivariate reduction method in specify a bivariate Poisson distribution and applied the method to simulated data.

In addition to the original trivariate reduction method, Karlis & Xekalaki (2005) proposed an extended model to allow for a combination of common random variables. Bermúdez & Karlis (2011) postulated a zero-inflated multivariate Poisson model to account for excess of zeros in automobiles insurance claim data. In another context of frequency modelling, a multivariate Poisson-lognormal regression model has been used for prediction of crash counts (Ma *et al.*, 2008). El-Basyouny & Sayed (2009) applied a similar multivariate Poisson-lognormal model to collision data to account for the correlation between two types of claims.

Although the trivariate reduction model can be extended to capture over-dispersion in the data, one drawback is that the correlation can only be positive (see Famoye, 2010*b*; Shi & Valdez, 2014). One way to address this

issue is to use an imposed parameter in the bivariate probability function to specify a covariance term to account for correlation. As the value of this correlation parameter can be negative, zero and positive, the limitation of positive correlation is removed. Thus the model is obviously more flexible with a more straightforward covariance structure. Lakshminarayana *et al.* (1999) defined a bivariate Poisson regression (BPR) model by including a multiplicative factor to capture the correlation between the two response variables. The probability function for the bivariate distribution is composed of two univariate Poisson probability functions, linked by the multiplicative correlation factor whose value depends on the embedded correlation parameter.

Based on a similar correlation structure, Famoye (2010*b*) applied a bivariate negative binomial regression (BNBR) model to analyse the bivariate distribution of two series of count data, while addressing over-dispersion in the sample. The study models marginal means of the two response variables with a set of explanatory covariates in a log-linear relationship. Data from the 1977-1978 Australian health survey is used to illustrate the model and the coefficients are estimated with maximum likelihood technique. The test results show that the BNBR model provides a better fit to the data than the BPR model, and supports the use of BNBR when the variance of the data is very different from the mean. However, the correlation parameter is not significant, thus two univariate negative binomial regression (UNBR) models may be able to provide similar results in his study.

2.2.1 Shrinkage methods

One drawback of the likelihood-based estimation of the regression models described above in the analysis of count data is that it commonly leads to a large number of variables being used. Although it is very tempting to incorporate as much information as possible to account for the heterogeneity in the population, this strategy is more time consuming in terms of model estimation. Too many explanatory variables in a regression model can also result in overfitting and consequently poor out-of-sample predictions.

The Lasso (least absolute shrinkage and selection operator) and ridge regression are two popular methods to shrink models (see Tibshirani, 1996; James *et al.*, 2013). Model shrinkage refers to the process of determining a smaller subset of variables that provide stronger explanatory power. Both techniques constrain the

coefficient estimates through a penalty term in the maximum likelihood estimation algorithm, comprised of the coefficient values and a shrinkage parameter ω. The higher the shrinkage parameter, the higher the impact of the shrinkage penalty. As a result, the coefficient values will approach zero as ω increases without bound. The optimal ω is commonly selected using cross-validation.

The two techniques differ in the way coefficient values are incorporated in the shrinkage penalty. The Lasso uses the sum of absolute values of coefficients, and ridge regression uses the sum of squared values. Ridge regression tends to shrink all coefficients towards zero, but will not generally set any of them to exactly zero. The Lasso is an alternative to ridge regression and can force some of the coefficient estimate to exactly zero if ω is sufficiently large. In other words, Lasso performs variable selection (see James *et al.*, 2013, Chapter 6).

The importance of model shrinkage has been recognised in the actuarial literature. First proposed by Tibshirani (1996), the Lasso has been extended to GLMs to handle count data (see Park & Hastie, 2007). Tang *et al.* (2014) applied adaptive Lasso to car insurance data. The risk factor selection improves the model goodness-of-fit both in the Poisson model as well as zero-inflated Poisson model. Wang *et al.* (2015) considered over-dispersed data and added a Lasso penalty to the maximum likelihood function of the negative binomial regression model. Their study concludes that a parsimonious model offers better prediction and interpretation. Both Tang *et al.* (2014) and Wang *et al.* (2015) used univariate regression models and applied the shrinkage technique to only one response variable. Ridge regression is shown to improve mean squared error in an early study by Hoerl & Kennard (1970). The technique is then applied to many areas of science. Some examples are Shen *et al.* (2013), Douak *et al.* (2013) and Meijer & Goeman (2013).

The two shrinkage methods can be applied to regression models to remove less significant variables. As a consequence, the unnecessary complexity in the model can be reduced and this leads to easier interpretation and potentially improved out-of-sample predicition (see James *et al.*, 2013, Chapter 6). It is these possibilities which we explore in the context of bivariate insurance claim data in this chapter.

2.3 Methodology

2.3.1 Bivariate negative binomial regression model

The bivariate Poisson distribution proposed in Lakshminarayana *et al.* (1999) has a probability function as the product of Poisson marginals with a multiplicative factor:

$$P(y_1, y_2) = \prod_{t=1}^{2} \frac{\theta_t^{y_t} e^{-\theta_t}}{y_t!} \times \left[1 + \lambda(e^{-y_1} - e^{-d\theta_1})(e^{-y_2} - e^{-d\theta_2})\right], y_1, y_2 = 0, 1, 2, \ldots \tag{2.1}$$

where $d = 1 - e^{-1}$. θ_t is the mean of Y_t $(t = 1, 2)$, and Y_1 and Y_2 are both Poisson distributed. The covariance between Y_1 and Y_2 is $\lambda\theta_1\theta_2 d^2 e^{-d(\theta_1+\theta_2)}$ and the correlation is $\rho = \lambda\sqrt{\theta_1\theta_2}d^2e^{-d(\theta_1+\theta_2)}$. Depending on the value of λ, the two response variables Y_1 and Y_2 can be positively or negatively correlated, or independent if λ is equal to zero.

By using a similar approach, Famoye (2010*b*) defined a bivariate negative binomial distribution. Following the same covariance specification as Lakshminarayana *et al.* (1999), a bivariate negative binomial distribution has the following probability function:

$$P(y_1, y_2) = \prod_{t=1}^{2} \binom{y_t + m_t^{-1} - 1}{y_t} \theta^{y_t}(1-\theta)^{m_t^{-1}} \times \left[1 + \lambda(e^{-y_1} - c_1)(e^{-y_2} - c_2)\right], \quad y_1, y_2 = 0, 1, 2, \ldots \tag{2.2}$$

Both Y_1 and Y_2 are random variables and follow a negative binomial distribution, with dispersion parameters m_1^{-1} and m_2^{-1} respectively. The mean of Y_t $(t = 1, 2)$ is $\mu_t = m_t^{-1}\theta_t/(1-\theta_t)$ and the variance is $\sigma_t^2 = m_t^{-1}\theta_t/(1-\theta_t)^2$. Also, $c_t = E(e^{-Y_t}) = [(1-\theta_t)/(1-\theta_t e^{-1})]^{m_t^{-1}}$.

Let n denote the sample size and Y_{it} $(t = 1, 2;\ i = 1, 2, \ldots, n)$ denote the count response variable, the corresponding vector of l explanatory variables is represented as $x_i = (x_{i0} = 1, x_{i1}, \ldots, x_{il})$. Assuming a log-linear model and the

same set of covariates as possible explanatory variables for both Y_{i1} and Y_{i2}, the means of the two response variables can be modelled as:

$$E(Y_{it}|x_i) = \mu_{it} = \exp(x_i\beta_t), t = 1, 2 \tag{2.3}$$

where $\beta_t^T = (\beta_{t0}, \beta_{t1}, \beta_{t2}, \ldots, \beta_{tl})$ and is the vector of the coefficients estimated using the maximum likelihood method. Given that $\theta_{it} = \mu_{it}/(m_t^{-1} + \mu_{it})$, equation (2.2) can be rewritten as:

$$\begin{aligned} P(y_{i1}, y_{i2}) = &\prod_{t=1}^{2} \begin{pmatrix} y_{it} + m_t^{-1} - 1 \\ y_{it} \end{pmatrix} \left(\frac{\mu_{it}}{m_t^{-1} + \mu_{it}}\right)^{y_{it}} \left(\frac{m_t^{-1}}{m_t^{-1} + \mu_{it}}\right)^{m_t^{-1}} \\ &\times \left[1 + \lambda(e^{-y_{i1}} - c_1)(e^{-y_{i2}} - c_2)\right]. \end{aligned} \tag{2.4}$$

The likelihood function, L is defined as:

$$L = \prod_{i=1}^{n} P(y_{i1}, y_{i2})$$

Accordingly, the log-likelihood function, which is set to a maximum to estimate the model parameters, for the unshrunken model is:

$$\begin{aligned} \log L = \sum_{i=1}^{n} \Bigg\{ \sum_{t=1}^{2} &\left[y_{it} \log \mu_{it} - m_t^{-1} \log m_t - (y_{it} + m_t^{-1}) \log(\mu_{it} + m_t^{-1}) - \log(y_{it}!) \right. \\ &\left. + \sum_{j=1}^{y_{it}-1} \log(m_t^{-1} + j)\right] + \log[1 + \lambda(e^{-y_{i1}} - c_1)(e^{-y_{i2}} - c_2)] \Bigg\}, \end{aligned} \tag{2.5}$$

where $c_t = (1 + d\mu_{it}m_t)^{-1/m_t}$ with $d = 1 - e^{-1}$. Equation (2.5) can be maximised with respect to β_t, m_t and λ. The asymptotic standard deviations of the estimated parameters are obtained in the usual way from Hessian matrix.

The deviance for a UNBR model, which is a measure of the goodness-of-fit for the model, is commonly defined as:

$$D_{UNBR} = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i(m^{-1}+\hat{\mu}_i)}{\hat{\mu}_i(m^{-1}+y_i)}\right) + m^{-1}\log\left(\frac{m^{-1}+\hat{\mu}_i}{m^{-1}+y_i}\right)\right]$$

Accordingly, the deviance for the BNBR model, is defined as:

$$\begin{aligned} D_{BNBR} =& 2\sum_{i=1}^{n}\left\{\sum_{t=1}^{2}\left[y_{it}\log\left(\frac{y_{it}(m_t^{-1}+\hat{\mu}_{it})}{\hat{\mu}_{it}(m_t^{-1}+y_{it})}\right) + m_t^{-1}\log\left(\frac{m_t^{-1}+\hat{\mu}_{it}}{m_t^{-1}+y_{it}}\right)\right]\right. \\ & \left.\log\left(\frac{1+\lambda\prod_{t=1}^{2}(e^{-y_{it}}-\bar{c}_t)}{1+\lambda\prod_{t=1}^{2}(e^{-y_{it}}-\hat{c}_t)}\right)\right\}, \end{aligned} \tag{2.6}$$

where $\bar{c}_t$ and $\hat{c}_t$ are the values of c_t evaluated at $\mu_{it} = y_{it}$ and $\mu_{it} = \hat{\mu}_{it}$ respectively, and $\hat{\mu}_{it}$ is the predicted value of μ_{it} found using equation (2.3) with estimated coefficients that maximise equation (2.5).

2.3.2 The Lasso and ridge regression

Given the BNBR model in equation (2.4), the coefficient vector β_t can be estimated by maximising equation (2.5). The resulting model will be called the full model in what follows. Here β_t (t=1,2) are vectors each having $k+1$ values. These relate to the model intercept and k explanatory variable coefficients. When k is large, the model may produce poor out-of-sample results because of an overfitting problem. It is therefore useful to shrink the estimated BNBR model using either the Lasso approach or ridge regression, by subtracting a shrinkage penalty from the log-likelihood function.

We define the log-likelihood function of the BNBR model in subsection 2.3.1, which is log L in equation (2.5). The new functions to be maximised under the two shrinkage approaches, with $2 \times l$ coefficients to be analysed are specified as:

$$\begin{aligned} &\text{The Lasso:} && \log L - \omega\sum_{t=1}^{2}\sum_{j=1}^{l}|\beta_{tj}| \\ &\text{Ridge regression:} && \log L - \omega\sum_{t=1}^{2}\sum_{j=1}^{l}\beta_{tj}^2 \end{aligned} \tag{2.7}$$

where ω is the shrinkage parameter. Adopting an increasing number of coefficients may increase the log-likelihood, however can also lead to an increase in both $\sum_{t=1}^{2}\sum_{j=1}^{l}|\beta_{tj}|$ and $\sum_{t=1}^{2}\sum_{j=1}^{l}\beta_{tj}^2$. These two shrinkage methodologies are used to find the optimal combination of coefficients that maximise equation (2.7), and some previously non-zero coefficients in the original full model may become zero in the shrunken models as a result.

Here t takes values 1 and 2, indicating that the shrinkage models consider regression coefficients for both y_1 and y_2. Thus the above equations specify the two shrinkage models in the context of a bivariate model, and y_1 and y_2 can represent, for example, claim numbers for two types of insurance policies bought by the same policyholder.

Note that we do not shrink the intercept coefficients (β_{t0}), as they simply constitute a measure of the mean value of the response variables when other explanatory variables are set to zero. Similarly, we also exclude the two over-dispersion parameters (m_1,m_2) and the correlation parameter (λ) from shrinkage, as we are focusing on shrinking the estimated association of each explanatory variable with the response. As a result, for each response variable, l regression coefficients are included in the shrinkage penalty.

When ω is equal to zero, both the Lasso and ridge regression will generate the same coefficients as the full model. A larger ω gives greater emphasis to model simplicity compared to in-sample goodness of fit. Consequently coefficient values will deviate from the maximum likelihood estimates, resulting in reduced in-sample goodness-of-fit. At the same time, the model is simplified with the potential for improved out-of-sample performance.

It is clear that different ω values will lead to different coefficients in the shrunken model and therefore differing out-of-sample prediction results. In order to perform the two shrinkage techniques as specified in equation (2.7) using the maximum likelihood method, the optimal value must be chosen for ω based on only the sample data to achieve the possibly best out-of-sample prediction accuracy. In this study we use k-fold cross-validation for this purpose, where commonly k is set to be 5 or 10 (Kohavi, 1995). In the cross-validation process, the sample data are randomly divided into k groups. One group is chosen as the validation set, while the model is fitted on the remaining $k-1$ groups. The fitted model is applied to the validation set to calculate the out-of-sample deviance, as the validation set is

held out in the model fitting process. As there are k groups, the procedure can be repeated k times resulting in k deviances when each of the k groups is held out as the validation set. The average of the k deviance values, each denoted deviance$_i$ (i=1,2,...,k), is taken as the cross-validation result, or k-fold CV, at a particular ω value (James *et al.*, 2013),

$$CV_{(\omega)} = \frac{1}{k}\sum_{i=1}^{k} \text{deviance}_i.$$

For each of the ω values, we perform the procedure as described previously. Among the grid of ω values, the most appropriate ω is the one that generates the lowest k-fold CV. As the CVs are calculated on the validation set, separated from the data to fit the model, when ω increases the CV is expected to decrease initially and later increase again when the impact from the penalty term is too strong. The ω that gives the minimum CV should be chosen.

For example, in a 5-fold cross-validation with a grid of 10 values chosen for ω $(\omega_1, \omega_2, \ldots, \omega_{10})$, the sample data will be divided into 5 groups. The first training set is composed of group 1 to group 4, and a BNBR model will be fitted to the training set, where the shrinkage parameter ω takes the value of ω_1. The fitted model is then used to calculate the out-of-sample deviance using group 5 which is the validation set. This process is repeated another 4 times at the same value of ω where each of the other 4 groups (group 1 to group 4) is held out as the validation set. So 5 deviances are generated at ω_1, and the average of the 5 deviances is the $CV_{(\omega_1)}$. The purpose of the cross-validation is find $\omega_i (i = 1, 2, \ldots, 10)$ among the 10 chosen values that returns the lowest $CV_{(\omega_i)}$.

We note here that although we develop different log-likelihood functions and shrinkage functions for the bivariate model, the validation process is standard. This is because the validation process only takes into consideration the deviances generate by a model, whether it is univariate or bivariate. Given the specified shrinkage models in equation (2.5), the validation process mentioned previously is proper for the BNBR model.

The shrinkage parameter, ω, is not assumed to be the same for the two shrinkage methods. A separate cross-validation is performed for each of the methods to locate the best ω value. Once this is achieved, the model is fitted again to the full set of data, disregarding the previously k group classifications. The shrunken

models can then be compared to the full model, which is estimated using maximum likelihood without any penalty term.

2.4 Data

The study is based on data from 14,000 automobile policies from a major insurance company in Spain, randomly selected from a pool of 80,994 policies. A subset of the data is also used in Brouhns *et al.* (2003), Bolancé *et al.* (2008), Bolancé *et al.* (2003), Boucher & Denuit (2008), Boucher *et al.* (2007), Bermúdez & Karlis (2011) and Boucher *et al.* (2009). We use 10,000 policies to estimate the model parameters, and the remaining 4,000 policies are used to test the model's out-of-sample prediction accuracy.

We model two types of claims, and their associated claim counts are recorded as Y_1 and Y_2. Y_1 represents the simple third-party liability with basic guarantees, and Y_2 stands for comprehensive cover. The same set of explanatory variables are assumed to affect both Y_1 and Y_2. The explanatory variables are summarised in Table 2.1. A similar table can also be found in Boucher *et al.* (2009).

Table 2.1. Explanatory variables in the regression model.

Variable	Description
v1	equals 1 for women and 0 for men
v2	equals 1 when driving in urban area, 0 otherwise
v3	equals 1 when zone is medium risk (Madrid and Catalonia)
v4	equals 1 when zone is high risk (Northern Spain)
v5	equals 1 if the driving license is between 4 and 14 years old
v6	equals 1 if the driving license is 15 or mode years old
v7	equals 1 if the client is in the company for more than 5 years
v8	equals 1 if the insured is 30 years old or younger
v9	equals 1 if includes comprehensive coverage (except fire)
v10	equals 1 if includes comprehensive and collision coverage
v11	equals 1 if horsepower is greater than or equal to 5500cc

We present in Table 2.2 a summary of the effects of the covariates on claim count based on all 80,994 policies[1]. The covariates are classified into eight groups. In the first column, we present the total number of policies that fall into each subgroup, followed by the percentage of policies with claim counts equal to 0, 1 or 2 (including higher than 2) for Y_1 and Y_2 respectively.

[1] Similar distribution figures can be generated for the sample chosen in this paper, which are not presented here.

For example, in the case of gender, we see here 12,957 of the policyholders are female. 93% of these female policyholders do not make a third party liability claim and 91.64% do not make a claim on the comprehensive cover. This is to be compared to the male policyholders, where 93.80% of them do not make a third party liability claim and 92.59% make no claim on the comprehensive cover. Ignoring other covariates and factors, female policyholders tend to have a slightly riskier profile compared to male policyholders.

Table 2.2. Summary statistics of claim frequencies as classified by the explanatory variables.

	Total	Y_1 (Third-party liability claim)			Y_2 (Comprehensive cover claim)		
		Count=0	Count=1	Count≥2	Count=0	Count=1	Count≥2
Gender							
Female (v1=1)	12957	93.29%	5.38%	1.33%	91.64%	6.14%	2.22%
Male (v1=0)	68037	93.80%	4.86%	1.34%	92.59%	5.60%	1.81%
Area							
Urban (v2=1)	54183	93.81%	4.86%	1.33%	92.21%	5.84%	1.95%
Other (v2=0)	26811	93.53%	5.10%	1.37%	92.89%	5.37%	1.74%
Zone risk level							
low (v3=0, v4=0)	45958	94.03%	4.65%	1.33%	93.78%	4.83%	1.39%
medium (v3=1, v4=0)	19320	93.78%	5.01%	1.22%	88.65%	8.14%	3.21%
high (v3=0, v4=1)	15716	92.73%	5.73%	1.55%	93.17%	5.17%	1.66%
Driver license							
below 4 years (v5=0, v6=0)	1894	90.87%	7.18%	1.95%	93.19%	5.33%	1.48%
between 4 and 14 years (v5=1, v6=0)	20854	92.93%	5.57%	1.51%	90.46%	7.19%	2.35%
above 14 years (v5=0, v6=1)	58246	94.09%	4.65%	1.26%	93.12%	5.16%	1.72%
Years with the company							
less than 5 years (v7=0)	11670	92.60%	5.79%	1.61%	90.26%	7.22%	2.53%
longer than 5 years (v7=1)	69324	93.90%	4.80%	1.30%	92.80%	5.43%	1.77%
Age							
30 years old or younger (v8=1)	7484	91.98%	6.27%	1.75%	90.62%	7.16%	2.22%
older than 30 years (v8=0)	73510	93.89%	4.81%	1.30%	92.62%	5.54%	1.84%
Insuarance cover							
no extra cover (v9=0, v10=0)	39791	93.97%	4.75%	1.29%	98.62%	1.17%	0.21%
only comprehensive (except fire) cover (v9=1, v10=0)	12613	93.61%	5.05%	0.90%	78.36%	14.39%	7.25%
both comprehensive and collision cover (v9=0, v10=1)	28590	93.41%	5.17%	1.42%	90.04%	8.13%	1.83%
Horsepower							
< 5500cc (v11=0)	15725	94.07%	4.67%	1.27%	96.09%	2.93%	0.98%
≥ 5500cc (v11=1)	65269	93.63%	5.01%	1.36%	91.56%	6.35%	2.08%
Mean			0.081			0.102	
Variance			0.123			0.168	

Similar observations can be made for the other groups of covariates. A lower claim count tends to be associated with driving in a low risk zone, a longer driving experience, a longer time with the insurance company, an older age and a smaller car horsepower. The effects of driving area (v2) and insurance cover (v9, v10) seem to be minimal based on this one-way analysis.

The estimated mean and variance of Y_1 and Y_2 are given at the end of Table 2.2. Y_1 has a lower mean and smaller variance compared to Y_2. Moreover, the variance is much higher than the mean for both claim types. This feature implies that

a model capable of handling over-dispersed data, such as the negative binomial regression model, is more appropriate compared to a Poisson regression model.

The correlation coefficient between Y_1 and Y_2 is 0.187, taking into account all 80,944 observations. A scatter plot is presented in Figure 2.1, including a trend line. The two variables can only take integer values. The number of observations at each of the dots is relatively indicated by the size of the dot, which is a rough reflection of the exact count summary shown in Table 2.3.

Figure 2.1. Scatter plot of two insurance claim counts. The size of the dot at each point gives a relative indication of the number of observations. The trend line is also presented.

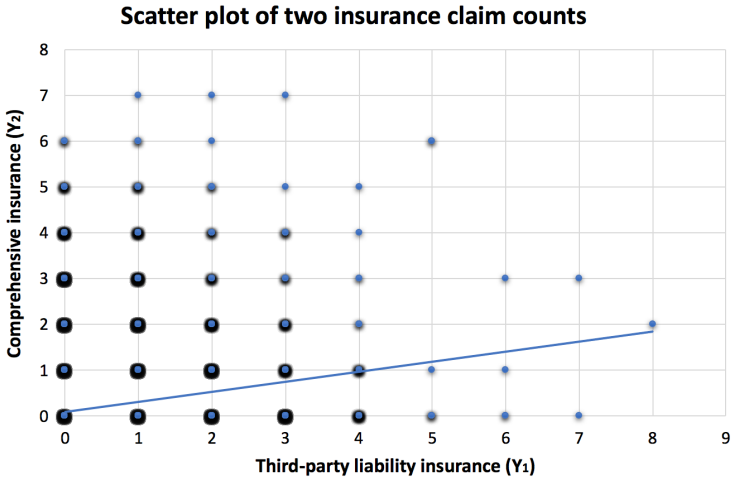


Table 2.3. Summary table of two types of insurance counts .

		Y_1								
		0	1	2	3	4	5	6	7	8
Y_2	0	71087	3022	574	149	29	4	2	1	0
	1	3722	686	138	42	15	1	1	0	0
	2	807	184	55	21	3	0	0	0	1
	3	219	71	15	6	2	0	1	1	0
	4	51	26	8	6	1	0	0	0	0
	5	14	10	4	1	1	0	0	0	0
	6	4	3	1	0	0	2	0	0	0
	7	0	1	1	1	0	0	0	0	0

We also analyse the correlation structure in the tail, when at least one of the claim counts is not zero. The correlation coefficient is computed at 0.126, which is lower than if all observations are considered. This is consistent with the two types of insurance counts presented in Table 2.3. If a higher right-tailed correlation is found, modelling tools such as copulas can be used to more accurately model the correlation structure (see Denuit, Dhaene, Goovaerts & Kaas, 2006, Chapter 4.4.4).

As tail dependency is not presented in this study, the model specified in equation (2.2) will suffice.

In addition to the variables listed in previous tables, we also consider two-way interaction effects among the variables. Adding interaction terms between independent covariates helps to relax the assumption that each of those independent variables only has additive effect in the regression model (see Fahrmeir *et al.*, 2013). Interaction effects are frequently analysed in regression models and have been considered in claim count models (see Yip & Yau, 2005; Shi & Valdez, 2014). We initially considered 14 potential two-way interactions. These terms cover the interaction effects between different groups of covariates, for example gender and driving experience, and are summarised in Table 2.4. We note that after model shrinkage many of the interaction terms were removed from the model.

Table 2.4. Interaction terms used in the regression model.

with v1	with v2	with v6	with v7	with v8
v1v2	v2v6	v6v7	v7v8	v8v11
v1v6	v2v7	v6v11	v7v11	
v1v7	v2v8			
v1v8	v2v11			
v1v11				

The total number of variables we use in the regression model is 25, excluding the intercept. Although we use the same set of variables for both response variables, we don't expect all explanatory variables to be significant in evaluating the claim counts, nor that the coefficients are the same for Y_1 and Y_2.

2.5 Results

2.5.1 Bivariate negative binomial regression model

We present in Table 2.5 the results of fitting four models: the BNBR model, UNBR model for Y_1, UNBR model for Y_2, and the BPR model. The four models are classified as full models as opposed to shrunken models, since at this stage we use all available variables including the chosen interaction terms. The BNBR model is specified in equation (2.4). The two UNBR models are fitted separately for each of the two response variables. The BPR model specification is the same

as in Lakshminarayana *et al.* (1999) and is given in equation (2.1). The function *glm.nb* from R package *MASS* is used to fit the univariate models (Venables & Ripley, 2002*b*). The R code developed to fit the two bivariate models, BNBR and BPR, can be found in Appendix A and Appendix B.

The results from the BNBR model are compared to the UNBR models. Coefficients from the BNBR model are consistent with those in UNBR models, both in terms of sign and statistical significance. By introducing a correlation factor λ, which is significant at the 1% level in the BNBR model, it is observed that the deviance of BNBR model is much lower than the sum of the deviances of the two UNBR models. This is true both in sample and out of sample, implying that the BNBR model provides a better in-sample goodness-of-fit, as well as more accurate out-of-sample prediction. It adds value to analyse the two correlated variables in a bivariate model, to properly account for the dependence between the two types of claim counts.

Consistent with expectation, the BNBR model also outperforms the BPR model. Although the BPR model recognises the correlation between the two response variables, the BNBR is more appropriate here when the data are over-dispersed and the variance of the claim counts is much higher than the mean for both types of claims as shown in Table 2.2. For this reason the BNBR generates both lower in-sample and out-of-sample deviances as expected.

2.5.2 The Lasso and ridge regression

The first step when applying the two shrinkage techniques is to choose the most optimal shrinkage parameter ω through cross-validation. We choose $k = 10$ and use 10-fold cross-validation which is widely used and effective, see for example Kohavi (1995)[2]. The two intercept coefficients (β_{10} and β_{20}), the dispersion parameters (m_1 and m_2) and the correlation parameter (λ) are excluded from the shrinkage process. For each of the two dependent variables, Y_1 and Y_2, 25 coefficients are estimated by maximising the penalised log-likelihood in equation (2.7).

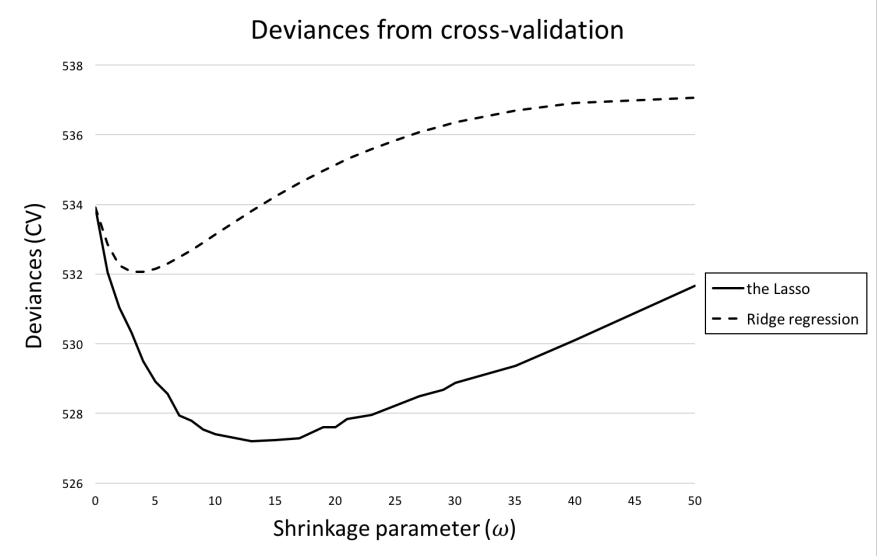
We select a grid of values for ω ranging from 0 to 50, and perform the procedure as described in Section 2.3.2. The sample data containing 10,000 policyholders is

[2] In addition to 10-fold cross-validation, we also conduct 5-fold cross-validation and the results are robust to the number of k. Here we present the results for $k = 10$.

randomly divided into 10 groups. One group is held out as the validation group while the model is fitted on the other nine groups at various ω values. This results in a number of different shrunken models and accordingly different deviance values based on equation (2.6) calculated on the validation set. Repeated 10 times for 10 different validation sets, we reach a series of $CV_{(a)}$ computed as the average of deviances from the 10 validation sets at a, where a denotes different ω values from 0 to 50.

We present in Figure 2.2 the $CV_{(a)}$ values from the cross-validation process. As expected, $CV_{(a)}$ decreases initially to a minimum before increasing again. When ω is zero, the shrunken models are equivalent to the full model. When ω increases, the deviances calculated using the held-out group firstly decrease, indicating better out-of-sample prediction results. Both of the curves increase again after reaching a minimum, where the shrinkage penalty is too strong and affects the models' prediction power.

Figure 2.2. Deviances from cross-validation at different ω values. Each deviance in the graph is calculated as the average of the ten deviances at the same ω generated in the 10-fold cross-validation process.



The shrinkage parameter, ω, in the Lasso and ridge regression are chosen using the cross-validation procedure. We get distinct optimal ω values that minimise deviance under the two different methods. As can be seen in Figure 2.2, the optimal ω chosen for the Lasso was found to be around 13, and the optimal ω for ridge regression was found to be around 4. We refit the BNBR model under two shrinkage approaches at given ω using the penalised log-likelihood as specified in equation (2.7). The estimated coefficients as well as the chosen ω are all presented

in Table 2.6. The full BNBR model fitted previously in Section 2.5.1 is also included.

Two observations from Table 2.6 can be made. First, the full model provides the best in-sample goodness-of-fit among the three, indicated by its lowest in-sample deviance. This is as expected as the full model is estimated to fit the sample data as closely as possible. Second, both shrunken models outperform the full model in out-of-sample prediction accuracy. The Lasso-shrunken model is the best among the three, with an out-of-sample deviance of 2586.82, lower than 2626.77 of the shrunken model obtained using ridge regression.

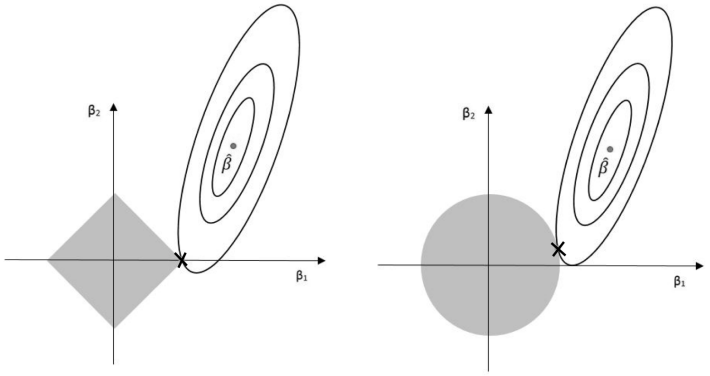
The shrinkage effect is more obvious in the Lasso-shrunken model. Many of the coefficients are forced to zero, including the insignificant ones identified in the full model. This indicates that those variables are not important in assessing the claim counts, and once removed, the out-of-sample prediction of the model is greatly improved. One possible explanation is that the full model overfits the sample data and thus underperforms shrunken models in making predictions. With fewer explanatory variables, the shrunken model is also much easier to interpret.

The shrinkage effect is not as obvious in the model regularised by ridge regression and none of the coefficients is zero after the shrinkage process. However, many coefficient values are more close to zero than in the full model, while the more significant variables, such the intercept, have a higher absolute coefficient and are still significant. This may explain why the shrunken model also outperforms the full model even when it uses a similar set of variables. Some coefficients of the regression may be reduced as ridge regression can be applied to treat the problem of collinearity between independent variables (see García *et al.*, 2015). In this study, we use categorical variables with values of 0 or 1, which may still lead to some potential for collinearity, for example between the policyholder's age and driving experience measured in years. As a result, treating the problem of collinearity may further improve the out-of-sample prediction accuracy.

The different results from the Lasso and ridge regression can also be explained with reference to Figure 2.3, which is similar to that in James *et al.* (2013, Chapter 6, page 222). The graph on the left refers to a two-dimensional coefficient scope of the Lasso, and the graph on the right represents the ridge regression. In both graphs, the dot inside the ellipses indicates the maximum likelihood estimate $\hat{\beta}$ without any shrinkage penalty. Assuming the same constraint amount s is used in

both methods, this means $|\beta_1|+|\beta_2| \leq s$ and $\beta_1^2+\beta_2^2 \leq s$, which can be represented by the grey area. If s is large enough to reach $\hat{\beta}$, the Lasso and ridge regression estimates will be the same as the maximum likelihood estimates (for example when $\omega = 0$).

Figure 2.3. Comparison of the Lasso (left) and ridge regression (right).



The ellipses around $\hat{\beta}$ represent regions of constant log-likelihood. The ellipses will expand away from $\hat{\beta}$ and touch the grey constraint area to satisfy the imposed shrinkage penalty. During this process, the Lasso is very likely to end up on one axis while ridge regression will land on the sphere, both shown in the graph as the cross. As a result, in the Lasso selection coefficients are commonly set to zero, while the same cannot be said for ridge regression. This simple graphical example can be extended to the higher dimensional case, when many Lasso estimated coefficients are equal to zero simultaneously.

To support the discussion and to show how the coefficient values react under the two shrinkage techniques, we present the shrunken coefficients at different ω values from the cross-validation procedure, computed as the average value across the 10 different models, each fitted when one group is held as the validation set. Note that we only plot the coefficients of explanatory variables, which are directly reduced in the shrinkage process. Figure 2.4 and Figure 2.5 show the results from the Lasso and ridge regression respectively, and present how the 25 coefficients change when the shrinkage parameter increases from 0 to 50 for Y_1 and Y_2 in separate graphes. As expected, all coefficients decrease with an increasing shrinkage parameter. They behave differently for Y_1 and Y_2, with some persistent coefficients significantly different from zero even at large ω values. These are specifically labelled on the figures.

However, it is quite noticeable that when ω is very large (i.e. set to 50), the coefficients in Figure 2.4 for the Lasso are much closer to zero, compared to those found in ridge regression in Figure 2.5. In particular, it can be observed that although the coefficients in Figure 2.5 approach zero initially and a few of them eventually become very close to zero in the end, most coefficients keep a constant distance away from zero which lasts to the end. The findings confirm the discussion made previously, that the two shrinkage techniques affect the coefficients in much distinctive ways.

Figure 2.4. Shrunken coefficients: the Lasso.

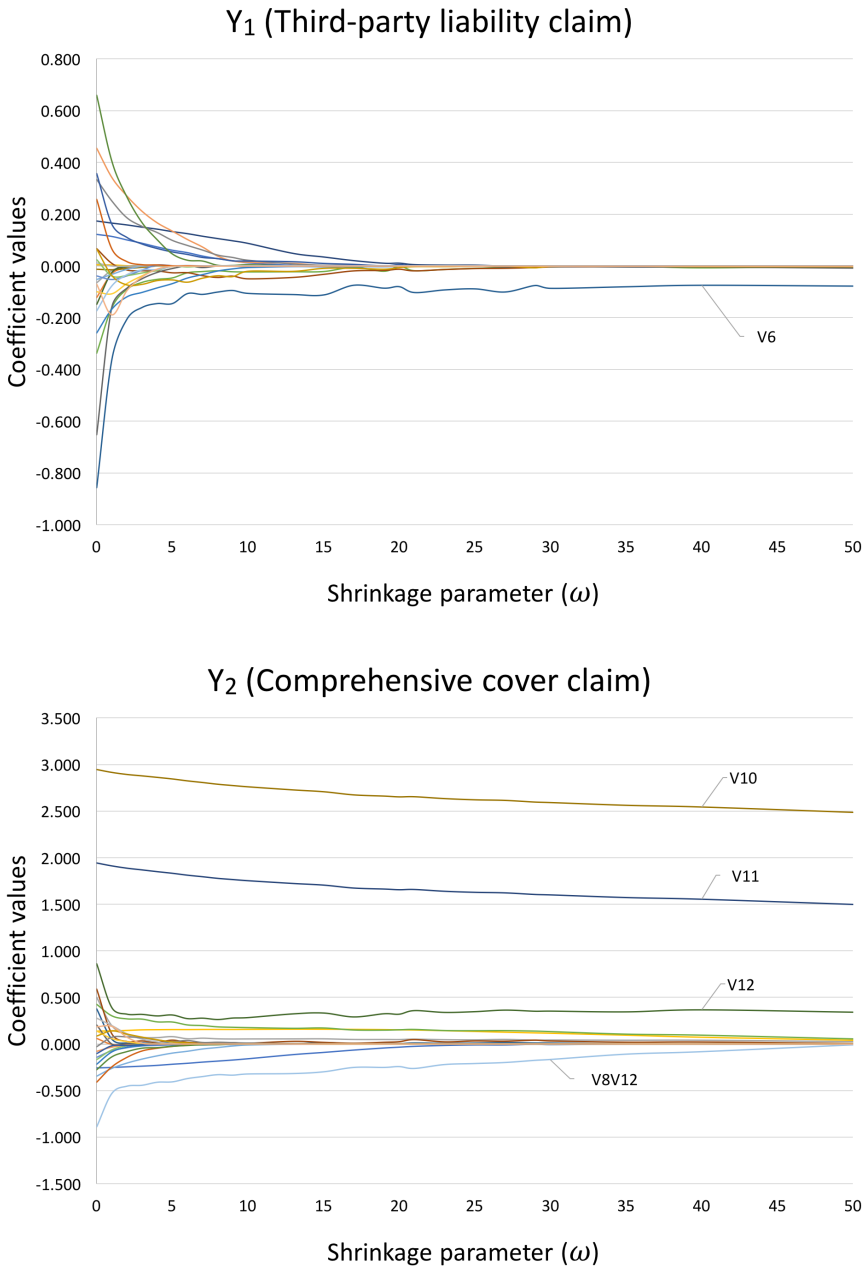
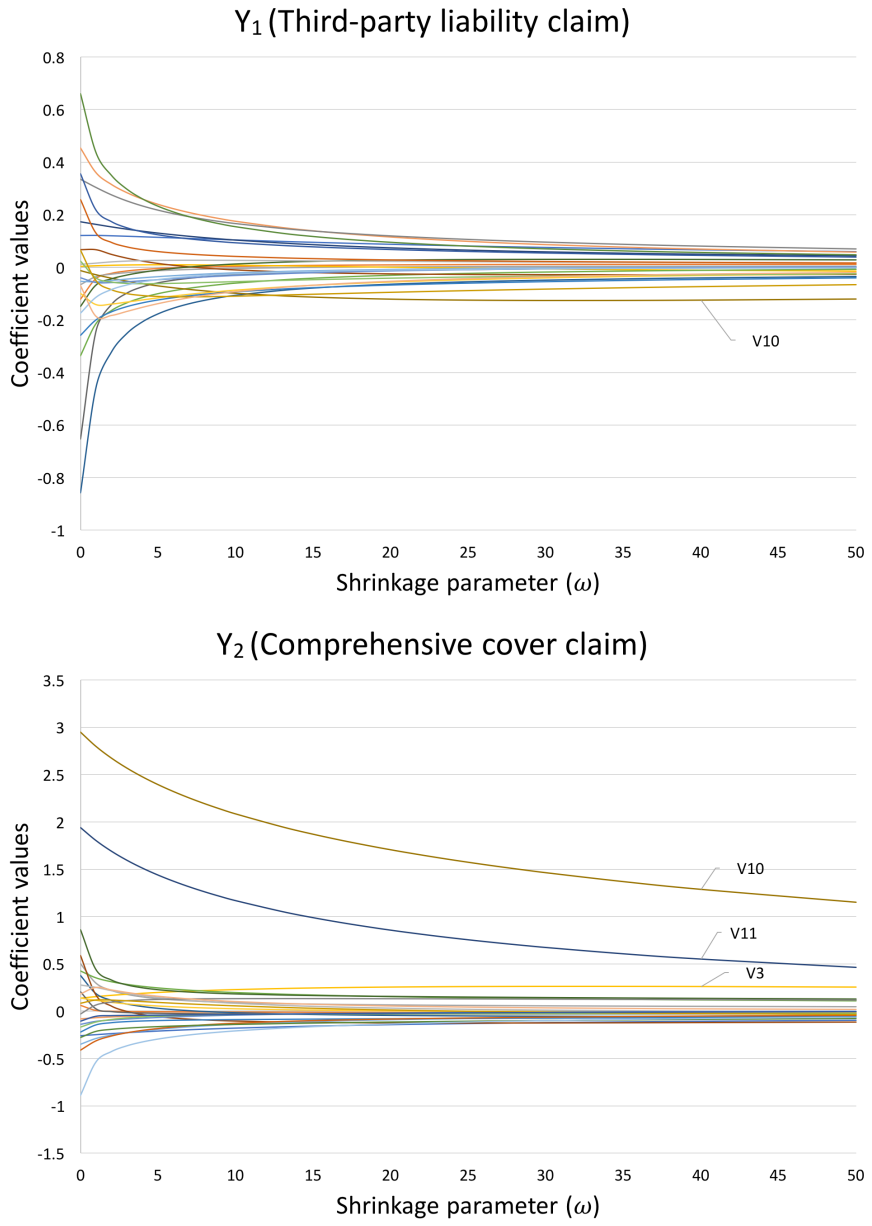


Figure 2.5. Shrunken coefficients: ridge regression .



The two shrinkage techniques are also applied to UNBR models in a similar way. For each response variable, two shrunken models are generated at given ω values selected by cross-validations. The results are presented in Table 2.7. Two full UNBR models estimated previously are also presented here.

Similar conclusions can be drawn from the shrunken UNBR models. For Y_1, both of the two shrunken models outperform the full model in out-of-sample prediction, implied by lower deviances. For Y_2, although the full model provides the best in-sample goodness-of-fit, it underperforms the shrunken models out-of-sample, with a lower log-likelihood and higher deviance.

By comparing the results for the Lasso-shrunken BNBR model in Table 2.6 and the two Lasso-shrunken UNBR models in Table 2.7, we see that the in-sample deviance of the BNBR model is much lower than the deviances from the two UNBR models combined, implying a better in-sample goodness-of-fit. The out-of-sample deviances are similar for the BNBR model and UNBR models. Obtained using ridge regression, the shrunken BNBR model outperforms the two shrunken UNBR models, providing both lower in-sample and out-of-sample deviances. This is in consistent with the conclusion we draw from the full models. It is beneficial to analyse the two response variables together in a bivariate model and properly account for the correlation structure between them.

Table 2.5. Modelling results of the BNBR model, two UNBR models and the BPR model, which are all classified as the full models. The coefficients of each variable are shown, followed by their standard deviation in parentheses. *,** and *** represent respectively statistical significance at the 10%, 5% and 1% level, calculated based on the t-statistics of coefficients of each variable.

Variable	**BNBR**	**UNBR(Y_1)**	**UNBR(Y_2)**	**BPR**
Y_1 **(Third-party liability claim)**				
Intercept	−1.984(0.570)***	−1.896(0.573)***		−1.990(0.452)***
v1	−0.114(0.431)	−0.186(0.434)		−0.123(0.341)
v2	−0.070(0.376)	−0.112(0.377)		−0.091(0.301)
v3	0.003(0.108)	0.036(0.109)		0.013(0.089)
v4	0.122(0.115)	0.113(0.115)		0.115(0.091)
v5	−0.341(0.295)	−0.374(0.300)		−0.316(0.224)
v6	−0.865(0.501)*	−0.952(0.507)*		−0.833(0.387)**
v7	0.053(0.437)*	0.056(0.438)		0.064(0.349)
v8	−0.655(0.592)	−0.064(0.594)		−0.710(0.476)
v9	−0.012(0.132)	0(0.133)		0.074(0.109)
v10	0.173(0.099)*	0.179(0.099)*		0.195(0.079)
v11	−0.157(0.433)*	−0.198(0.434)		−0.155(0.347)
v1v2	−0.056(0.253)	−0.017(0.252)		−0.061(0.200)
v1v6	0.452(0.277)	0.481(0.277)*		0.443(0.221)**
v1v7	0.011(0.322)	0.019(0.322)		0.035(0.253)
v1v8	−0.100(0.380)	−0.086(0.382)		−0.116(0.308)
v1v11	−0.042(0.282)	−0.027(0.282)		−0.052(0.226)
v2v6	0.023(0.241)	0.057(0.242)		0.010(0.193)
v2v7	−0.255(0.281)	−0.253(0.282)		−0.254(0.224)
v2v8	0.255(0.366)	0.244(0.369)		0.266(0.294)
v2v11	0.335(0.241)	0.360(0.242)		0.368(0.197)
v6v7	0.068(0.277)	0.059(0.279)		0.057(0.216)
v6v11	0.356(0.295)	0.381(0.296)		0.345(0.236)
v7v8	0.666(0.397)*	0.634(0.400)		0.688(0.322)**
v7v11	−0.170(0.344)	−0.179(0.346)		−0.197(0.282)
v8v11	−0.070(0.430)	−0.044(0.434)		−0.051(0.339)
m_1	6.454(0.649)***	6.440(0.648)***		
Y_2 **(Comprehensive cover claim)**				
Intercept	−5.104(0.629)***		−5.041(0.640)***	−4.732(0.525)***
v1	0.068(0.400)		0.039(0.412)	−0.001(0.324)
v2	0.489(0.346)		0.486(0.355)	0.355(0.279)
v3	0.136(0.084)		0.151(0.087)*	0.184(0.065)***
v4	−0.257(0.108)		−0.293(0.108)***	−0.253(0.089)***
v5	0.421(0.314)		0.431(0.323)	0.328(0.267)
v6	0.373(0.489)		0.252(0.506)	0.143(0.405)
v7	0.578(0.477)		0.572(0.476)	0.403(0.406)
v8	0.209(0.605)		0.194(0.609)	−0.064(0.510)
v9	2.946(0.130)***		2.943(0.131)***	2.942(0.120)***
v10	1.941(0.126)***		1.948(0.127)***	1.955(0.120)***
v11	0.848(0.497)		0.843(0.495)*	0.659(0.422)
v1v2	−0.351(0.225)		−0.331(0.222)	−0.278(0.181)
v1v6	−0.080(0.236)		−0.039(0.239)	−0.129(0.192)
v1v7	0.274(0.275)		0.246(0.275)	0.317(0.223)
v1v8	0.142(0.327)		0.132(0.327)	0.168(0.261)
v1v11	−0.142(0.283)		−0.140(0.283)	−0.183(0.231)
v2v6	−0.165(0.203)		−0.143(0.207)	−0.190(0.163)
v2v7	−0.213(0.228)		−0.209(0.231)	−0.171(0.177)
v2v8	−0.406(0.310)		0.506(0.314)	−0.404(0.247)
v2v11	−0.023(0.244)		−0.024(0.250)	0.119(0.201)
v6v7	0.080(0.216)		0.098(0.231)	0.234(0.172)
v6v11	−0.102(0.293)		−0.034(0.297)	−0.058(0.240)
v7v8	−0.276(0.305)		−0.199(0.312)	−0.075(0.240)
v7v11	−0.884(0.411)		−0.924(0.414)**	−0.865(0.361)***
v8v11	0.177(0.505)		0.177(0.513)	0.252(0.441)
m_2	2.532(0.254)***		2.504(0.254)***	
λ	5.663(0.396)***			5.748(0.371)***
In-sample log-likelihood	−5556.90	−2384.60	−2945.76	−5880.94
Out-of-sample log-likelihood	−2605.69	−1136.68	−1519.55	−2845.35
In-sample deviance	5215.18	2384.60	2866.00	8764.67
Out-of-sample deviance	2854.57	1025.36	1851.15	4494.50

.

Table 2.6. Modelling result for the original full BNBR model and shrunken models. The coefficients of each variable are shown, followed by their standard deviation in parentheses. *,** and *** represent respectively statistical significance at the 10%, 5% and 1% level, calculated based on the t-statistics of coefficients of each variable.

Variable	**Full model**	**the Lasso**	**Ridge regression**
ω	0	13	4
Y_1**(Third-party liability claim)**			
Intercept	−1.984(0.570)***	−2.371(0.283)***	−2.418(0.297)***
v1	−0.114(0.431)	0(0.009)	−0.008(0.241)
v2	−0.070(0.376)	0(0.009)	−0.018(0.225)
v3	0.003(0.108)	0(0.009)	0.009(0.103)
v4	0.122(0.115)	0(0.009)	0.118(0.108)
v5	−0.341(0.295)	−0.012(0.274)	−0.124(0.203)
v6	−0.865(0.501)*	−0.131(0.268)	−0.224(0.254)
v7	0.053(0.437)*	0.082(0.121)	0.026(0.230)
v8	−0.655(0.592)	0(0.009)	−0.081(0.258)
v9	−0.012(0.132)	0(0.009)	−0.059(0.122)
v10	0.173(0.099)*	0.058(0.091)	0.141(0.094)
v11	−0.157(0.433)*	0(0.009)	−0.023(0.230)
v1v2	−0.056(0.253)	0(0.009)	−0.042(0.194)
v1v6	0.452(0.277)	0(0.009)	0.268(0.201)
v1v7	0.011(0.322)	0(0.009)	0.026(0.213)
v1v8	−0.100(0.380)	0(0.009)	−0.127(0.246)
v1v11	−0.042(0.282)	0(0.009)	−0.053(0.203)
v2v6	0.023(0.241)	0(0.009)	−0.062(0.175)
v2v7	−0.255(0.281)	0(0.009)	−0.141(0.189)
v2v8	0.255(0.366)	0(0.009)	0.071(0.229)
v2v11	0.335(0.241)	0(0.009)	0.241(0.177)
v6v7	0.068(0.277)	−0.002(0.025)	−0.105(0.186)
v6v11	0.356(0.295)	0(0.009)	0.140(0.189)
v7v8	0.666(0.397)*	0(0.009)	0.275(0.233)
v7v11	−0.170(0.344)	0(0.009)	−0.061(0.200)
v8v11	−0.070(0.430)	0(0.009)	−0.159(0.238)
m_1	6.454(0.649)***	6.459(0.643)***	6.501(0.653)***
Y_2 **(Comprehensive cover claim)**			
Intercept	−5.104(0.629)***	−4.073(0.328)***	−3.895(0.294)***
v1	0.068(0.400)	0(0.009)	−0.001(0.229)
v2	0.489(0.346)	0.067(0.081)	0.158(0.211)
v3	0.136(0.084)	0.159(0.085)*	0.184(0.080)***
v4	−0.257(0.108)	−0.133(0.105)	−0.219(0.101)**
v5	0.421(0.314)	0.163(0.084)*	0.270(0.198)
v6	0.373(0.489)	0(0.009)	0.061(0.245)
v7	0.578(0.477)	0.010(0.331)	−0.022(0.221)
v8	0.209(0.605)	0(0.009)	−0.018(0.249)
v9	2.946(0.130)***	2.756(0.123)***	2.509(0.107)***
v10	1.941(0.126)***	1.748(0.120)***	1.545(0.104)***
v11	0.848(0.497)	0.375(0.323)	0.260(0.224)
v1v2	−0.351(0.225)	0(0.009)	−0.215(0.178)
v1v6	−0.080(0.236)	0(0.009)	−0.049(0.183)
v1v7	0.274(0.275)	0(0.009)	0.174(0.194)
v1v8	0.142(0.327)	0(0.009)	0.073(0.229)
v1v11	−0.142(0.283)	0(0.009)	−0.079(0.195)
v2v6	−0.165(0.203)	0(0.009)	−0.065(0.155)
v2v7	−0.213(0.228)	0(0.009)	−0.104(0.166)
v2v8	−0.406(0.310)	0(0.009)	−0.212(0.210)
v2v11	−0.023(0.244)	0(0.009)	0.117(0.170)
v6v7	0.080(0.216)	0(0.009)	0.104(0.162)
v6v11	−0.102(0.293)	0(0.009)	−0.043(0.184)
v7v8	−0.276(0.305)	0(0.009)	−0.174(0.210)
v7v11	−0.884(0.411)	−0.361(0.345)	−0.340(0.194)*
v8v11	0.177(0.505)	0(0.009)	0.186(0.235)
m_2	2.532(0.254)***	2.511(0.253)***	2.545(0.255)***
λ	5.663(0.396)***	3.797(0.429)***	5.774(0.413)***
In-sample log-likelihood	−5556.90	−5587.42	−5567.22
Out-of-sample log-likelihood	−2605.69	−2491.68	−2493.17
In-sample deviance	5215.18	5228.46	5228.87
Out-of-sample deviance	2854.57	2586.82	2626.77

Table 2.7. Modelling results of the original full UNBR model and UNBR models shrunken by the two methods. The coefficients of each variable are shown, followed by their standard deviation in parentheses. *,** and *** represent respectively statistical significance at the 10%, 5% and 1% level, calculated based on the t-statistics of coefficients of each variable .

Variable	Y_1 **(Third-party liability claim)**			Y_2 **(Comprehensive cover claim)**		
	UNBR	**the Lasso**	**Ridge regression**	**UNBR**	**the Lasso**	**Ridge regression**
ω	0	7	29	0	23	2
Intercept	−1.896(0.573)***	−2.469(0.206)***	−2.542(0.163)***	−5.041(0.64)***	−4.124(0.349)***	−4.111(0.373)***
v1	−0.186(0.434)	0(0.014)	0.008(0.110)	0.039(0.412)	0(0.007)	−0.014(0.283)
v2	−0.112(0.377)	0(0.012)	0.006(0.106)	0.486(0.355)	0.070(0.085)	0.221(0.256)
v3	0.036(0.109)	0(0.018)	0.021(0.082)	0.151(0.087)	0.151(0.088)*	0.178(0.084)**
v4	0.113(0.115)	0.020(0.110)	0.064(0.085)	−0.293(0.108)***	−0.024(0.106)	−0.269(0.105)***
v5	−0.374(0.300)	−0.001(0.012)	−0.013(0.107)	0.431(0.323)	0.134(0.091)	0.303(0.240)
v6	−0.952(0.507)*	−0.101(0.273)	−0.058(0.112)	0.252(0.506)	0(0.008)	0.015(0.312)
v7	0.056(0.438)	−0.054(0.243)	−0.032(0.107)	0.572(0.476)	0(0.352)	0.060(0.279)
v8	−0.064(0.594)	0(0.013)	0.003(0.114)	0.194(0.609)	0(0.161)	0.008(0.326)
v9	0(0.133)	0(0.013)	−0.011(0.091)	2.943(0.131)***	2.656(0.122)***	2.692(0.117)***
v10	0.179(0.099)*	0.120(0.092)	0.122(0.077)	1.948(0.127)***	1.690(0.119)***	1.719(0.113)***
v11	−0.198(0.434)	0(0.017)	0.032(0.107)	0.843(0.495)*	0.416(0.344)	0.335(0.283)
v1v2	−0.017(0.252)	0(0.013)	−0.011(0.110)	−0.331(0.222)	0(0.007)	−0.249(0.198)
v1v6	0.481(0.277)*	0.070(0.164)	0.097(0.111)	−0.039(0.239)	0(0.007)	−0.020(0.205)
v1v7	0.019(0.322)	0(0.015)	0.026(0.110)	0.246(0.275)	0(0.007)	0.197(0.224)
v1v8	−0.086(0.382)	0(0.013)	−0.040(0.121)	0.132(0.327)	0(0.270)	0.076(0.266)
v1v11	−0.027(0.282)	0(0.013)	−0.010(0.109)	−0.140(0.283)	0(0.008)	−0.096(0.224)
v2v6	0.057(0.242)	0(0.013)	−0.029(0.100)	−0.143(0.207)	0(0.013)	−0.065(0.175)
v2v7	−0.253(0.282)	−0.041(0.166)	−0.064(0.099)	−0.209(0.231)	0(0.023)	−0.125(0.191)
v2v8	0.244(0.369)	0(0.254)	0.022(0.115)	−0.506(0.314)	0(0.007)	−0.316(0.248)
v2v11	0.360(0.242)	0.077(0.161)	0.109(0.099)	−0.024(0.250)	0(0.011)	0.093(0.196)
v6v7	0.059(0.279)	−0.032(0.259)	−0.085(0.102)	0.098(0.231)	(0.009)	0.123(0.188)
v6v11	0.381(0.296)	0.010(0.174)	0.052(0.102)	−0.034(0.297)	0(0.007)	−0.004(0.215)
v7v8	0.634(0.400)	0(0.257)	0.071(0.115)	−0.199(0.312)	0(0.007)	−0.134(0.245)
v7v11	−0.179(0.346)	0(0.013)	−0.024(0.099)	−0.924(0.414)**	−0.289(0.365)	−0.458(0.243)*
v8v11	−0.044(0.434)	0(0.012)	−0.028(0.115)	0.177(0.513)	0(0.008)	0.184(0.299)
m	6.440(0.648)***	6.47(0.666)***	6.571(0.658)***	2.504(0.254)***	2.547(0.269)***	2.511(0.254)***
In-sample log-likelihood	−2725.21	−2732.89	−2731.50	−2945.76	−2961.10	−2949.04
Out-of-sample log-likelihood	−1136.68	−1112.983	−1111.99	−1519.55	−1400.59	−1453.35
In-sample deviance	2384.60	2394.80	2377.77	2865.60	2882.55	2869.90
Out-of-sample deviance	1025.36	976.22	968.92	1851.15	1608.78	1717.86

2.6 Conclusion

In this chapter we used the BNBR model to analyse general insurance claim data. We show that with a more flexibly specified correlation structure, the BNBR model adequately captures the relationship between the two claim counts and the set of explanatory variables. The correlation, which is totally ignored if two UNBR models are fitted separately, proves to be essential in analysing the two types of claim counts from the same policyholder. Note that the correlation coefficient between the two claim count is only 0.187 in this study which is considered as a weak correlation. When a higher correlation coefficient is present, it is likely that a bivariate model with a proper specification of the correlation structure is more suitable than a univariate model.

In addition, we apply two shrinkage techniques to choose core independent variables in modelling claim counts. The results from the Lasso and ridge regression are different, but both shrunken models outperform original full regression models which are likely to suffer from the overfitting problem. The shrunken models provide much better out-of-sample prediction accuracy in both UNBR and BNBR models. This automatic approach to model selection has considerable potential for application in actuarial modelling where very large numbers of variables and data points are often available. Moreover, the shrunken BNBR models also outperforms the two separately fitted shrunken UNBR models, which again emphasises the importance of properly accounting for the correlation structure between response variables.

In addition to BNBR model in this study, some extended Poisson model can also incorporate over-dispersion. For example the zero-inflated versions of multivariate Poisson models used in Bermúdez & Karlis (2011), where the correlation structure in equation (2.2) can be implemented instead of the full covariance specification. The bivariate generalised Poisson regression model in Famoye (2010*a*) follows a similar correlation structure as in this study, which also allows for over-dispersion. These potential alternative models can be considered in future research.

Chapter 3

Assessing Sovereign Risk: A Bottom-Up Approach

Feng Liu (contribution 80%), Egon Kalotay (contribution 10%), Stefan Trück (contribution 10%)

A research paper based on this chapter has been published:

- Liu, F., Kalotay, E. & Trueck, S. (2017). Assessing Sovereign Default Risk: A Bottom-Up Approach. *Economic Modelling* (70), 525-542.

3.1 Introduction

In recent years there has been an increased interest in sovereign credit risk, see, e.g., Pan & Singleton (2008); Caceres *et al.* (2010); Ang & Longstaff (2013); Longstaff *et al.* (2011); Aizenman *et al.* (2013); Janus *et al.* (2013). Sovereign risk is typically measured by credit spreads associated with the probability of default (PD) on sovereign debt securities, as there is uncertainty about receiving scheduled payments on time. Since the onset of the global financial crisis (GFC), Europe in particular has been the focus of much of this concern. While research on sovereign risk and advanced risk management tools had also accumulated before the European debt crisis, the crisis was still unseen by many market participants. Despite being preceded by the GFC, in early 2009 neither observed CDS spreads nor ratings for European sovereign entities provided an indication

of the magnitude of the soon-to-occur sovereign debt crisis. This may indicate a need to assess and predict sovereign credit risk using more responsive measures based on additional risk sources. Further, despite much effort from governments and global financial institutions, sovereign debt sustainability remains a major concern, which motivates us to develop a new framework for predicting sovereign default risk.

This study provides a new bottom-up approach to assess sovereign default risk at the state-level for 18 state governments in the U.S. As argued by Ang & Longstaff (2013), each U.S. state government retains the authority to establish its independent legal system and the ability to issue state bonds. As a result, state bonds are similar to federal bonds and the economic behaviour of a state government can be considered as being similar to a sovereign entity. Given the recent financial distress of large municipalities such as Detroit or the U.S. territory of Puerto Rico, we also believe that a more in depth analysis of sovereign debt at the state level is an important exercise more generally. We start with Moody's KMV expected default frequencies (EDFs) to assess credit risk at the corporate level for industries of economic importance. We then aggregate information extracted from EDFs at the company level to develop industry credit risk indicators (ICRIs). In a second step, the constructed ICRIs are then used to derive state credit risk indicators (SCRIs), based on the industry composition of each state. In this way we calculate real-time bottom-up credit risk indicators at the state-level. Clearly, our motivation for constructing the SCRIs is to better understand whether variation in default risk in the private sector can presage prediction for market views on a sovereign's ability to service its debt obligations. Our study follows the motivation of Altman & Rijken (2011*b*) in investigating the influence of the private sector on a sovereign entity's default risk. We assume that publicly listed companies contribute to a sovereign entity's wealth and, thus, also to its risk of default. The derived SCRIs are then investigated with regards to their predictive power for changes in credit default swap (CDS) spreads for the individual states. We find that the derived market-based measures of private sector credit risk are strongly associated with subsequent shifts in sovereign credit risk premiums, as measured by CDS spreads. Overall, the developed SCRIs are highly significant in forecasting sovereign CDS spreads at weekly and monthly sampling frequencies.

Traditionally, the assessment of sovereign risk has relied heavily on macroeconomic

variables containing information on economic conditions and aggregated national accounts. A variety of econometric frameworks using macroeconomic variables have been applied to explain the behavior of sovereign risk over time. Grinols (1976) applies both discriminant and discrete analysis to a sample of 64 nations to identify five significant national account variables in his assessment of debt service capability. Morgan (1986) studies debt rescheduling based on new short-term debt data and variables representing economic shocks, using logit and discriminant models. A more recent example is Haugh *et al.* (2009), where a range of macroeconomic explanatory variables are incorporated in a panel model to study sovereign spread differentials among European countries. Others studies such as Fuertes & Kalotychou (2004) and Hilscher & Nosbusch (2004) also examine the predictive power of similar variables.

A common approach across all these studies is the reliance on macroeconomic data, such as annual GDP growth rates, the balance of trade, tax receipts or debt servicing ratios, or similar. Although there is a significant body of research supporting the explanatory power of macroeconomic variables, the forecasting ability of these variables for crises or changes in credit quality of sovereigns has been questioned. In a comprehensive overview paper, Babbel (1996) argues that macroeconomic forecasting approaches generally fail to perform satisfactorily, and that the claimed predictive power of macroeconomic models is only illusory. The author argues that, upon closer inspection the studies are mostly unsuccessful. Bertozzi (1995) also questions the ability of macroeconomic models to provide a signal for early warning. One possible reason for the inadequate response times of macroeconomic models are the infrequent updates of input data, which are also subject to delayed release by government statistical offices. Thus for timely projections of changes in sovereign risk, it might be more beneficial to identify early warning signals in order to harness the limited time that policy makers and financial managers typically have to change strategies (Bertozzi, 1995; Neziri, 2009). Models that only use one set of observations per year will undoubtedly have difficulties in capturing changes in sovereign risk in a timely manner (Oshiro & Saruwatari, 2005). Aizenman *et al.* (2013) also argue that while macroeconomic factors are statistically and economically important determinants of sovereign risk, the pricing of this risk for Eurozone periphery countries is not predicted accurately either in-sample or out-of-sample with these factors.

Therefore, over the last decade sovereign risk has typically been measured by more

timely and frequently available data from financial variables such as sovereign bond prices or CDS spreads. Examples include Pan & Singleton (2008), Beber *et al.* (2009), Hui & Chung (2011), Fender *et al.* (2012), Aizenman *et al.* (2013), Ang & Longstaff (2013), Arce *et al.* (2013), Calice *et al.* (2013), Groba *et al.* (2013), Janus *et al.* (2013), Dewachter *et al.* (2015) and Chen *et al.* (2016). Recent studies have focussed on CDS spreads in particular, since they provide a more direct measure of sovereign risk. Pan & Singleton (2008) analyze default risk and recovery rates implicit in the term structure of sovereign CDS spreads. Ang & Longstaff (2013) adopt CDS spreads for the U.S. Treasury, individual U.S. states, and major Eurozone countries, to study the nature of systemic sovereign credit risk. Aizenman *et al.* (2013) examine CDS as a measure of sovereign default risk and argue that CDS spreads provide a good proxy for market-based pricing of default risk. The authors also provide a market-based real-time indicator of sovereign credit quality and default risk. Beber *et al.* (2009), Arce *et al.* (2013) and Calice *et al.* (2013) focus on price discovery, liquidity spill-over and flight-to-quality effects in the sovereign CDS market. Groba *et al.* (2013) focus on financially distressed economies inside the European Union and their impact on the CDS market.

One limitation of these studies in assessing sovereign risk is that so far little attention is given to the private sector, which can yield a more direct measure of economic activities within a sovereign entity. Generally, the productivity, profit and economic performance of companies in a state can be expected to have a direct impact on tax receipts and the wealth of a sovereign government. As a result, financial health of a sovereign entity will be sensitive to financial crises, the poor performance of major industries in a state or a slowdown of the economy. Incorporating forward-looking company level information into the risk assessment process therefore has the potential to provide important fundamental information that may help to predict financial distress at the state or government level. Due to the importance of measuring default risk at the firm level in financial markets, credit rating agencies such as Standard & Poor's, Fitch or Moody's KMV provide timely information on default risks at the company level (Trück & Rachev, 2009).

To take advantage of the abundant company level data for assessing sovereign risk, Altman & Rijken (2011*b*) were among the first to propose a bottom-up approach to incorporate private sector information in the assessment process, considering this information as a crucial determinant of sovereign risk. They test the predictive

power of factors generated from listed companies at country level, assuming that sovereign financial health relies on the economic performance of the private sector. Altman & Rijken (2011*b*) focus on major European countries during the debt crisis and assess the probability of sovereign default based on the credit risk of the private sector. Their prediction model demonstrates greater effectiveness in providing advance warnings compared to those of credit rating agencies. Incorporating listed company information also enlarges the available data points and gives greater opportunity for investigating sovereign default risk.

A potential disadvantage of the approach developed by Altman & Rijken (2011*b*) is its reliance on corporate credit scores, based on infrequently updated variables such as company leverage, profitability, and liquidity. Thus, corporate credit scores may provide a picture of retrospective rather than prospective company performance. In addition, macroeconomic variables such as GDP growth and inflation, that are available at a low frequency only, are also included in the model (Altman & Rijken, 2011*b*).

To overcome these shortcomings, this study assesses sovereign risk at the state level, using market variables encompassing industries of economic importance to each state. However state government defaults are different from corporate defaults because of different legal enforcements. Unlike the bankruptcy procedure following the default of a company, a state government's assets cannot be credibly liquidated or transferred to the debtor (Ang & Longstaff, 2013). Therefore, we argue that state governments can be considered as independent sovereign entities. Our motivation is to better understand whether variation in default risk in the private sector can improve prediction for a sovereign's ability to service its debt obligations. Our study follows the motivation of Altman & Rijken (2011*b*) by investigating the influence of the private sector on a sovereign entity's credit risk. We assume that publicly listed companies contribute to a sovereign entity's wealth and also its risk of default, and use Moody's KMV EDFs for individual companies to create industry-level and state-level credit risk indicators. EDFs are forward-looking measures of default risk, based on the structural model developed by Merton (1974) combined with information on historical defaults. The accuracy of EDFs for predicting defaults has been documented in a number of studies, see, e.g. Kealhofer (2003), Dwyer & Korablev (2007), Bharath & Shumway (2008). Next to the developed EDF-based industry- and state-level credit risk indicators, our model also incorporates additional financial variables that have been suggested

to have predictive power for default risk. In contrast to previous studies, we use a bottom-up approach to predict sovereign CDS spreads that may be particularly useful for capturing and forecasting short-term changes in sovereign risk. Due to the inclusion of the derived SCRIs, the proposed models may be able to better predict state CDS spreads at the weekly and monthly frequency. Thus, they could prove to be helpful to participants in financial markets, in particular those who trade credit-default swaps or other instruments related to default risk at the state level. Besides, the developed ICRIs and SCRIs may also provide policy makers with a monitoring tool or early warning indicators at longer time horizons. An upward-shift or permanent increase in the derived indicators may well be a sign of a possible increase in sovereign credit risk at the state level in future periods.

We examine the default risk of 19 state governments in the U.S., covering the time period from June 2006 to April 2013. In our analysis we treat each of the states as an independent sovereign entity. CDS spreads on state government debt are used to measure the default risk for each of these states. We first develop ICRIs that are then used to calculate the SCRIs based on the industry composition of each state. Each industry has its own default index that is built on the credit risk of listed companies in the sector.

Our results indicate that the developed SCRIs, using information from the private sector, are highly significant in predicting CDS spreads for the vast majority of the states considered in this study. Regression analysis strongly suggests the benefits of incorporating company level information on default risk to augment macro-financial variables for the assessment of sovereign credit risk. Our findings are also confirmed by robustness checks, using a variety of forecasting frequencies as well as alternative state credit risk indicators based on the major industries in each state only. We also apply quantile regression to estimate the coefficients for the independent variables at different quantiles of the distribution, and we test the predictive relationship using both through-the-cycle and point in time measures of company credit risk. The robustness checks confirm our findings on the usefulness of the developed credit risk indicators.

Overall, our results emphasize the importance of information from the private sector for predicting sovereign default risk. Our findings complement those of Altman & Rijken (2011*b*), by using a distinctively different and more timely assessment method. First, instead of using company scores, our study adopts Moody's KMV EDFs to assess corporate level credit risk. Second, our analysis

focuses on a significantly longer time horizon, examining the sovereign risk of state governments in the U.S. over seven years, also covering the pre- and post-financial crisis period. Moreover, constructing and incorporating ICRIs addresses the influence of variation in industry compositions on overall sovereign risk, an issue which was not examined by Altman & Rijken (2011*b*). Further, financial companies are not excluded from our study, addressing one of the caveats of Altman & Rijken (2011*b*). Finally, in contrast to previous studies that have applied bottom-up approaches to sovereign default risk, our model examines CDS spreads that may be particularly useful for capturing and forecasting short-term changes in sovereign risk.

The remainder of this chapter is organised as follows. Section 3.2 provides a brief review of the EDF measure and then describes our approach to derive bottom-up industry and state credit risk indicators. Section 3.3 describes the data and the applied models, while Section 3.4 provides results for the empirical analysis as well as various robustness checks. Section 3.5 concludes and provides suggestions for future work.

3.2 Bottom-Up Credit Risk Indicators

3.2.1 Industry Credit Risk Indicators (ICRIs)

In the following, we aim to derive industry and state specific credit risk indicators that will reflect information on default risk available at the company level. As a first step, for each industry a sector-specific indicator of default risk is developed that can then be used to derive state-specific indicators of default risk.

We use Moody's KMV one-year EDFs to measure corporate credit risk in the private sector as a predictive measure of credit risk at the firm level. We include all U.S. companies available in the Moody's KMV EDF universe and use one-year EDF estimates as measures for a company's credit risk. The timely availability of EDFs allows for almost immediate incorporation of new information relevant to measurement of sovereign credit risk. One-year EDFs provide an estimate of the probability of default for a particular company within a time horizon of twelve months. Unlike credit ratings that typically involve a relative rank-order scale, EDFs are measured on a quantitative scale (Moody's, 2012). Note that a number of

previous empirical studies has also confirmed the usefulness of EDFs or the Merton distance to default for predicting bankruptcies at the corporate level. Kealhofer (2003) shows that EDFs contain additional information for default prediction that is not captured in ratings. Results by Bharath & Shumway (2008) suggest that almost two thirds of defaulting firms had probabilities of default in the highest decile based on the Merton distance to default during the quarter they defaulted. Dwyer & Korablev (2007) also emphasize the usefulness of EDFs for predicting defaults, when computing accuracy ratios for North America, Europe and Asia.

The EDF model belongs to a class of structural credit risk models pioneered by Merton (1974), incorporating more realistic assumptions and enriching the originally suggested model with empirical data on defaults to reflect real-world measures of credit risk (Moody's, 2012). Starting from Merton's framework, the model assumes that a firm's value follows a stochastic process with an expected growth rate and volatility. The model assesses the probability of asset values falling below liabilities payable, the so-called default point. The distance to default (DD) is then calculated as the difference between the expected outcome for the firm's value based on the underlying stochastic process and the default point, measured by the number of standard deviations of the annual percentage change in the market value of the firm's assets. To derive real-world default frequencies, Moody's KMV then conduct a mapping process. The relationship between DD and default frequency is developed based on empirical observations to account for the actual number of default for different DDs. A mapping procedure is then applied to create EDFs for companies based on their DDs. Moody's KMV frequently update their EDF estimates and ratings to reflect a company's credit risk based on new information that affect price or volatility. The real-world probability of default at time t (EDF_t) can then be denoted by Equation 3.1 (De Servigny *et al.*, 2004):

$$EDF_t = F(-\frac{(log(V_t) - log(X) + (\mu - \sigma_V^2/2)(T-t)}{\sigma_V\sqrt{T-t}}), \tag{3.1}$$

where F is the function mapping the distance to default calculated by a Merton-type model to the actual EDF. V_t denotes the value of the firm at time t, X is the default threshold, μ is the expected return on assets, and σ_V is the asset volatility of the firm. $T-t$ is set equal to 1 in the calculation, according to the calibration of EDFs to a one-year horizon.

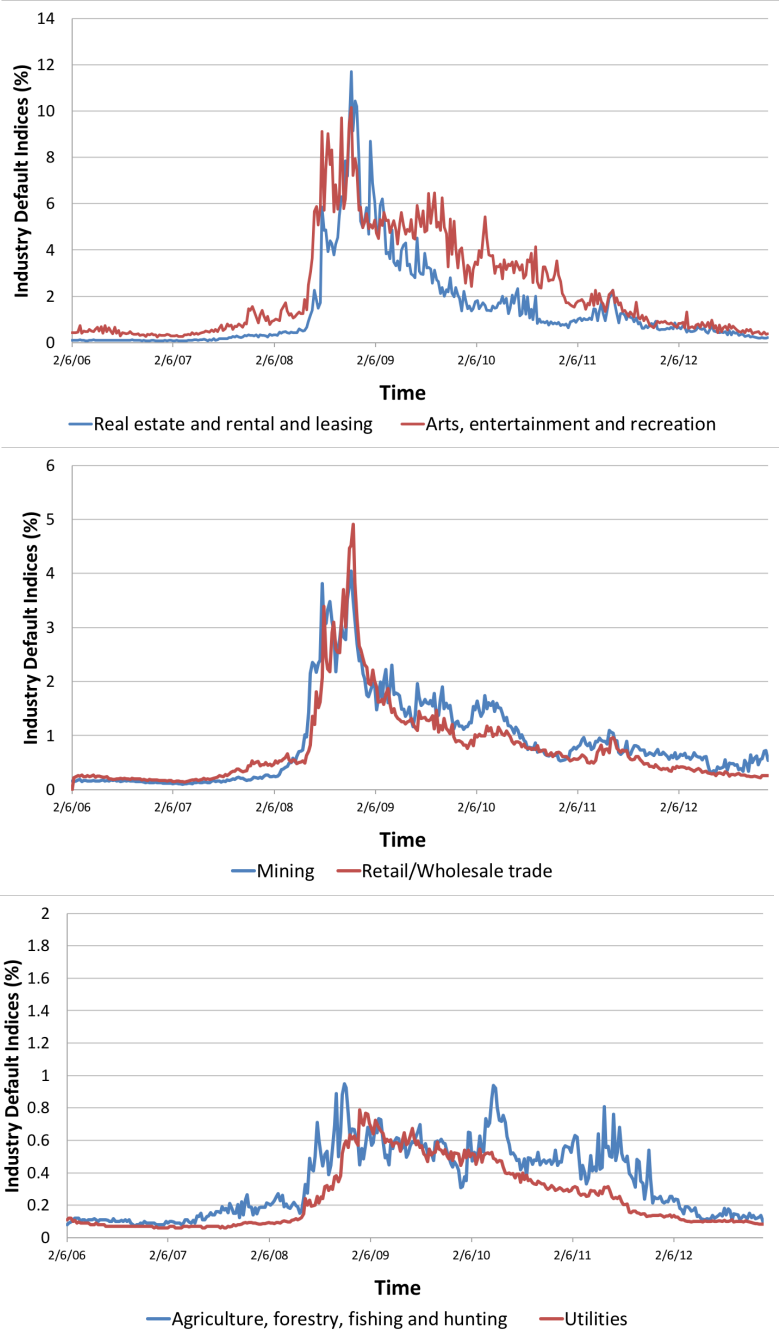
We use Moody's one-year EDFs from June 2006 to April 2013 for all U.S. listed companies at weekly and monthly frequency. Over the sample, the EDFs range from 1 basis point to a maximum of 35%. Over time, some companies have been delisted and newly listed companies have been added. A total of 8105 companies are included in our study over the sample period. Note that one-year EDFs are actually updated on a daily basis, while the average number of observations on a typical trading day is around 4300. We use end of week observations. When the Friday of a week coincides with a public holiday and no observation is available, the observed EDF on the Thursday of the same week is used. During the whole study period, we only observe 12 public holidays falling on a Friday.

In order to assign companies to industries, a set of industry classifications is constructed. The definition of each industry is based on Moody's EDF industry definitions as well as industry categories defined by the U.S. Department of Commerce, Bureau of Economic Analysis (BEA). Note that Moody's KMV assigns companies to a total of 61 detailed industries, including one unassigned category that contains companies without a clearly matching industry. The BEA on the other hand adopts 20 industry categories only. We define industry classifications by combining the two schemes, such that ICRIs can be constructed using Moody's EDF, and the indices can later be combined according to BEA's statistics on GDP compositions for each state. The industry mapping scheme is shown in Table 3.1 alongside the corresponding industries defined by Moody's KMV and the BEA.

According to the BEA industry classification, U.S. listed companies that were originally assigned to 61 industries by Moody's KMV, are then regrouped into 15 industries. Note that the Moody's KMV industry classification of a company may change multiple times throughout the study period, such that every time a change occurs the company will be assigned accordingly to the corresponding industry category. At each point in time, EDF values for all companies in the same industry are collected, and the ICRI is defined as the median of the observed EDFs for the industry cohort.

Figure 3.1 presents the derived ICRIs for six of the 15 industries. We find that the credit risk indicators behave quite differently across industries during the study period as a result of different sensitivities to movement in economic conditions. We also observe that all ICRIs increase significantly during the GFC period between 2008 and 2009, and tend to move back to their pre-crisis level afterwards. ICRIs for *real estate, rental and leasing* as well as for *arts, entertainment and recreation*

Figure 3.1. Time series of ICRIs for *real estate, rental and leasing* and *arts, entertainment and recreation (upper panel)*, *mining* and *retail/wholesale trade (middle panel)*, and *agriculture, forestry, fishing and hunting* and *utilities (lower panel)* for the sample period June 2006 to April 2013.



New category	BEA	Moody's
1.Agriculture, forestry, fishing and hunting	1.Agriculture, forestry, fishing and hunting	N02 Agriculture N33 Lumber & Forestry
2.Mining	2.Mining	N38 Mining N39 Oil Refining
3.Utilities	3.Utilities	N58 Utilities Nec N59 Utilities, Electric N60 Utilities, Gas
4.Construction	4.Construction	N13 Construction N14 Construction Materials
5.Durable goods	5.Durable goods	N05 Automotive N11 Computer Hardware N15 Consumer Durables N20 Electrical Equipment N21 Electronic Equipment N27 Furniture & Appliances N34 Machinery & Equipment N35 Measure & Test Equipment N36 Medical Equipment N49 Semiconductors N50 Steel & Metal Products N54 Transportation Equipment
6.Nondurable goods	6.Nondurable goods	N04 Apparel & Shoes N10 Chemicals N17 Consumer Products N40 Oil, Gas & Coal Expl/Prod N41 Paper N42 Pharmaceuticals N43 Plastic & Rubber N44 Printing N52 Textiles N53 Tobacco
7.Retail/Wholesale trade	7.Wholesale trade 8.Retail trade	N08 Business Products Whsl N16 Consumer Durable Retl/Whsl N18 Consumer Products Retl/Whsl N26 Food & Beverage Retl/Whsl
8.Transportation and warehousing	9.Transportation and warehousing	N03 Air Transportation N55 Transportation N56 Trucking
9.Information	10.Information	N07 Broadcast Media N12 Computer Software N45 Publishing N51 Telephone N61 Cable TV
10.Finance and insurance	11.Finance and insurance	N06 Banks and S&Ls N23 Finance Companies N24 Finance Nec N29 Insurance - Life N30 Insurance - Prop/Cas/Health N31 Investment Management N47 Real Estate Investment Trusts N48 Security Brokers & Dealers
11.Real estate, rental and leasing	12.Real estate and rental and leasing	N32 Lessors N46 Real Estate
12.Professional, scientific and technical services	13.Professional, scientific and technical services	N09 Business Services N19 Consumer Services N37 Medical Services
13.Arts, entertainment and recreation	18.Arts, entertainment and recreation	N22 Entertainment & Leisure
14.Accommodation and food services	19.Accommodation and food services	N25 Food & Beverage N28 Hotels & Restaurants
15.Other	14.Management of companies and enterprises 15.Administrative and waste management services 16.Educational services 17.Health care and social assistance 20.Other services, except government	N01 Aerospace & Defense N57 Unassigned

Table 3.1. Assigned industry classifications based on allocated industry definitions from BEA and Moody's KMV. The first column provides the classification of industries applied in this study, columns 2 and 3 present corresponding industries from BEA and Moody's KMV that were assigned to each category. The classification typically follows BEA.

are the two most volatile amongst the 15 indices and are presented in the upper panel of Figure 3.1. Median default probabilities for these categories reach as high as 12% during the GFC and exhibit a declining trend afterwards. It is no surprise that the real estate sector was affected greatly by the subprime credit crunch, while spillover effects are likely to have influenced industries such as entertainment.

The middle panel of Figure 3.1 illustrates that the industry categories *mining* and *retail/wholesale trade* are less volatile in comparison to the two industry categories

described above. Their ICRIs peak at approximately 5% in the middle of 2009 which is less than half of the median EDF for the real estate category. Finally, in the lower panel of Figure 3.1 we present the derived ICRIs for *agriculture, forestry, fishing and hunting* and *utilities*. These industries appear least affected by the GFC, indicated by the small change in the overall level of default probabilities. Throughout the entire sample period the median EDF does not exceed a value of 1% for these two industries.

3.2.2 State Credit Risk Indicators (SCRIs)

We combine the 15 ICRIs to generate predictive credit risk indicators at the state level. Henceforth, for each state we define an SCRI as the weighted average of the ICRIs. The weight of each industry as a contributor to the SCRI is based on an industry's GDP percentage contribution to the entire state GDP, following the decomposition for the prior 10 years, provided by the BEA. The BEA releases the total GDP for each state and for each industry in the regional economy each year. The contribution of each industry to a state's total GDP can then be calculated as the average percentage contribution over the ten year period from 2003-2013.[1] A summary of the compositions for industries across states can be found in Table 3.2, where the average composition, the maximum and minimum percentage, and difference between the two are shown for each industry. As expected, the GDP contribution for each industry varies significantly across the 19 states, yielding cross-sectional variation in risk profiles over the sample period.

Based on the 15 ICRIs and corresponding contributions for each state, the SCRIs are created as the weighted average of the 15 indices in each of the 19 states, calculated according to Equation (3.2):

$$\begin{bmatrix} ICRI_{1,1} & \cdots & ICRI_{1,15} \\ \vdots & \ddots & \vdots \\ ICRI_{1,360} & \cdots & ICRI_{15,360} \end{bmatrix} \times \begin{bmatrix} w_{1,1} & \cdots & w_{1,19} \\ \vdots & \ddots & \vdots \\ w_{15,1} & \cdots & w_{15,19} \end{bmatrix} = \begin{bmatrix} SCRI_{1,1} & \cdots & SCRI_{1,19} \\ \vdots & \ddots & \vdots \\ SCRI_{360,1} & \cdots & SCRI_{360,19} \end{bmatrix} \tag{3.2}$$

[1] Note that the contributions of the different industries to the GDP for a state are typically relatively stable through time. To examine whether the choice of the period for calculating the average contribution of an industry has an impact on the results in Section 4, we conducted various robustness checks, using the average industry composition also for shorter periods. The results were qualitatively the same and almost identical to those when the average composition for 2003-2013 was used.

	Average	Max	Min	Max-Min
Agriculture, forestry, fishing and hunting	0.70%	2.15%	0.18%	1.97%
Mining	0.91%	9.37%	0.01%	9.36%
Utilities	1.91%	2.57%	1.22%	1.35%
Construction	5.06%	8.75%	3.51%	5.24%
Durable goods	6.92%	15.39%	2.47%	12.92%
Nondurable goods	5.53%	14.82%	1.28%	13.54%
Retail/Wholesale trade	13.31%	16.26%	8.60%	7.65%
Transportation and warehousing	2.81%	3.96%	1.40%	2.56%
Information	4.85%	9.62%	2.16%	7.46%
Finance and insurance	11.85%	39.33%	6.09%	33.25%
Real estate, rental and leasing	15.01%	19.48%	10.33%	9.15%
Professional, scientific and technical services	8.72%	14.30%	5.50%	8.80%
Arts, entertainment and recreation	1.09%	2.60%	0.67%	1.94%
Accommodation and food services	3.74%	16.41%	2.06%	14.35%
Other	17.60%	21.98%	13.15%	8.83%

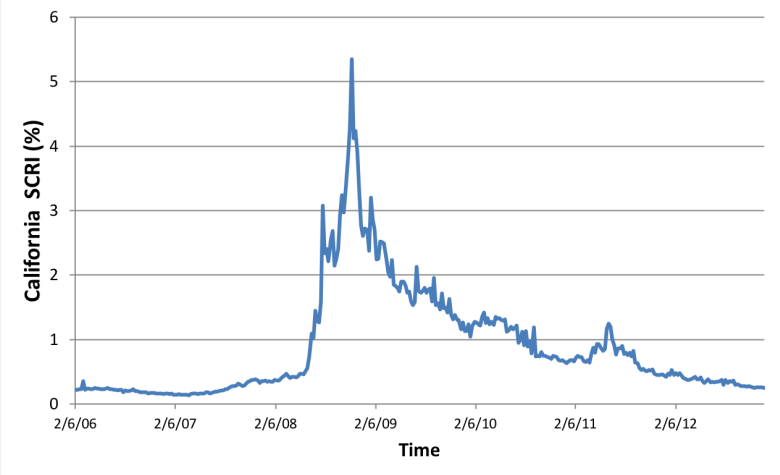
Table 3.2. Summary statistics of contribution to the GDP of the 19 states for different industries. We report average contributions for the time period 2003-2013 provided by the U.S. Department of Commerce, Bureau of Economic Analysis (BEA).

Recall that we have 360 weekly observations for the EDFs and ICRI. $ICRI_{t,j}$ then denotes the ICRI value at time t for industry j, with $t = 1, ..., 360$ and $j = 1, ..., 15$, while $w_{j,k}$ represents the weight for industry j in state k with $k = 1, ..., 19$. The product of the two matrices then yields 19 time series of SCRIs, one for each of the states considered in this study. Thus, $SCRI_{t,k}$ denotes the state risk indicator at time t for state k. We present the time series for the derived SCRI for California in Figure 3.2. Based on the different industry contributions for each state, we also observe a quite different behaviour in the developed state credit risk indicators throughout the sample period. Overall, we believe that the derived SCRIs will provide an appropriate measure of private sector credit risk for the considered states.

3.3 Data and Models

We use market data on CDS spreads of 19 US state governments for the time period June 2, 2006 to April 26, 2013 in order to examine the relationship between the constructed bottom-up SCRIs and sovereign default risk at the state level. Using weekly observations, we analyze CDS spreads for the following states: California, New York, Texas, Florida, Illinois, Pennsylvania, Ohio, New

Figure 3.2. Time series of California's constructed SCRI for the sample period June 2, 2006 to April 26, 2013.



Jersey, Michigan, Massachusetts, North Carolina, Virginia, Wisconsin, Maryland, Minnesota, Connecticut, Delaware, Nevada and Rhode Island. CDS spreads for all state governments are obtained from Datastream and Bloomberg and are quoted in basis points.[2] We decided to use five-year CDS spreads, because they have the greatest liquidity. Note that due to limited availability of CDS spreads for some of the states during the sample period, not all time series are of equal length. Table 3.3 provides summary statistics of the CDS spreads for the 18 state governments, including the number of observations, the mean, maximum and minimum spread and the standard deviation of the observed spreads. Table 3.3 also presents the initial point in time when information on CDS spreads was available for a particular state.[3]

In addition to SCRIs, we relate changes in CDS spreads to five additional financial variables with a view to capturing changes in general financial and economic conditions. Note that due to the forward-looking nature of this study and the use of data at a relatively high (weekly) frequency, we decided not to include

[2] The included states were selected as a result of data availability on CDS spreads. The beginning of the sample period (June 2006) was determined based on the fact that from this point onwards data on CDS spreads was available for at least six states, namely Florida, Maryland, New Jersey, New York, North Carolina, and Pennsylvania. We decided to only include states, where observations for CDS spreads were available from October 2010 at the latest, to guarantee at least 130 weekly observations for each state. Based on this criteria we have a cross-section of 19 states in our sample, see Table 3.3 for more details.

[3] Note that four states have missing values after the first available observation, namely North Carolina (from 24/09/2010 to 29/10/2010), Virginia (from 25/03/2011 to 29/06/2012), Delaware (from 09/12/2011 to 24/02/2012), Rhode Island (from 24/09/2010 to 22/10/2010).

any macroeconomic explanatory variables in our analysis. We argue that despite the impact of macroeconomic conditions on the ability of a state's government to service its debt, variables that are updated only infrequently throughout the year and often on a delayed basis are not appropriate for the purpose of this study.

We use the Chicago Board Options Exchange Market Volatility Index (VIX) as a forward-looking measure of volatility in the equity market, a proxy for the uncertainty faced by investors. We also include the term spread, i.e. the difference between short-term government bills (1 month) and long-term government bonds (20 years). Changes in the term spread are often used as an indicator of economic conditions and credit risk, since the spread is expected to contain information on future economic growth and has been successfully applied to forecasting the probability of a recession, see, e.g., Stock & Watson (1989), Dotsey (1998). Both the VIX and term spread are expected to be positively related to state CDS spreads, as both variables can be considered proxies for increasing risks in equity and debt markets. These variables have also been widely suggested as determinants of sovereign risk in previous studies, see for example, Hilscher & Nosbusch (2004), Giesecke & Kim (2011), Longstaff *et al.* (2011), Dieckmann & Plank (2012), just to name a few. Welch & Goyal (2008) also identify a link between equity premiums required by investors and the market volatility and the term spread. We also include returns of the S&P 500 index as a measure of stock market performance, as well as 5-year U.S. Treasury CDS spreads. Finally, we consider returns of the S&P U.S. issued investment grade corporate bond index (the CDX IG index). While the coefficients of the Treasury CDS spreads are expected to be positive for all states, market returns and the returns of CDX IG index are expected to be negatively correlated with state sovereign CDS spreads.

We include these additional market variables to complement the information contained in the derived bottom-up SCRIs. Note that a similar set of explanatory variables has also been applied by Ang & Longstaff (2013), when assessing systemic sovereign credit risk for several European countries and states in the U.S. We collect weekly observations for all predictive variables for the June 2006 to April 2013 period. Data for the VIX is available from Bloomberg, while yields on federal securities are available from the U.S. Treasury database. Data on the S&P 500 index is sourced from CRSP, while U.S. Treasury CDS spreads and the CDX IG index are available through Datastream.

We estimate the following model to measure the impact of the predictive variables on CDS spreads in the 19 states:

$$\begin{aligned} CDS_{i,t} =& \beta_{0,i} + \beta_{1,i} * SCRI_{i,t-1} + \beta_{2,i} * VIX_{t-1} + \beta_{3,i} * TS_{t-1} + \beta_{4,i} * SP500_{t-1} \\ &+ \beta_{5,i} * TCDS_{t-1} + \beta_{6,i} * CDX_{t-1} + \epsilon_i \end{aligned} \tag{3.3}$$

In the above equation, $CDS_{i,t}$ denotes the observed CDS spread for state i in period t, while $SCRI_{t-1}$ denotes the constructed state credit risk indicator at $t-1$. VIX denotes the Chicago Board Options Exchange Market Volatility Index, TS the term spread, $SP500$ the return on the S&P500 index, $TCDS$ the Treasury CDS spread, and CDX refers to the return on the CDX IG index. Since we are particularly interested in the predictive power of these factors, all explanatory variables are measured in period $t-1$. In contrast to some previous studies that focused on forecasting infrequent default events, such as Oshiro & Saruwatari (2005) and Giesecke & Kim (2011), we focus on CDS spreads as observable proxy measures of default risk for these states. Note that while the observed CDS spreads could also be transformed into PD estimates when additional assumptions on recovery rates for the states are applied, we don't make such adjustments in the current study.

Figure 3.3 provides a plot of the CDS spreads for the states of California, New York, Texas, Florida, Illinois and Ohio from December 2007 to April 2013. For all states we find that CDS spreads were at a low level at the beginning of 2007 and increase significantly during the GFC. The highest spreads could be observed in 2008, typically followed by several smaller peaks and troughs, with the CDS spreads exhibiting a declining trend in later periods of the sample.

3.4 Empirical Analysis

3.4.1 Baseline Model

As a first step, we estimate the coefficients for model (3.3), where all six explanatory variables are included to assess sovereign default risk. The model

	Obs	Mean	Max	Min	σ	Available from
California	281	198.3	455.0	46.0	83.9	Dec 2007
New York	360	121.1	356.8	29.0	69.4	Jun 2006
Texas	284	74.5	205.0	20.0	35.4	Nov 2007
Florida	360	104.4	273.0	35.0	46.8	Jun 2006
Illinois	284	184.0	360.0	24.3	81.6	Nov 2007
Pennsylvania	360	72.9	157.0	45.0	37.4	Jun 2006
Ohio	257	121.0	280.0	34.5	42.0	May 2008
New Jersey	360	125.8	370.0	29.0	75.0	Jun 2006
Michigan	272	162.7	404.8	39.0	79.4	Feb 2008
Massachusetts	280	107.1	246.0	20.6	49.1	Dec 2007
North Carolina	246	93.0	179.4	20.7	39.0	Jul 2008
Virginia	296	69.0	146.5	36.0	24.9	Jun 2006
Wisconsin	173	97.4	145.6	29.0	23.7	Jan 2010
Maryland	360	73.9	109.0	12.5	38.9	Jun 2006
Minnesota	130	70.2	175.2	12.5	19.4	Oct 2010
Connecticut	206	118.8	166.0	63.4	24.4	May 2009
Delaware	192	58.2	105.0	27.7	14.3	Jun 2009
Nevada	271	145.3	373.2	40.0	63.6	Feb 2008
Rhode Island	142	123.6	168.7	60.0	26.0	Jul 2010

Table 3.3. Summary statistics for weekly 5-year CDS spreads for the 19 states considered. Descriptive statistics are based on CDS spreads denoted in basis points. We report mean, maximum, minimum, standard deviation (σ) as well as the beginning of the sample period for each state. For all states, the last observation of the sample period is April 2013.

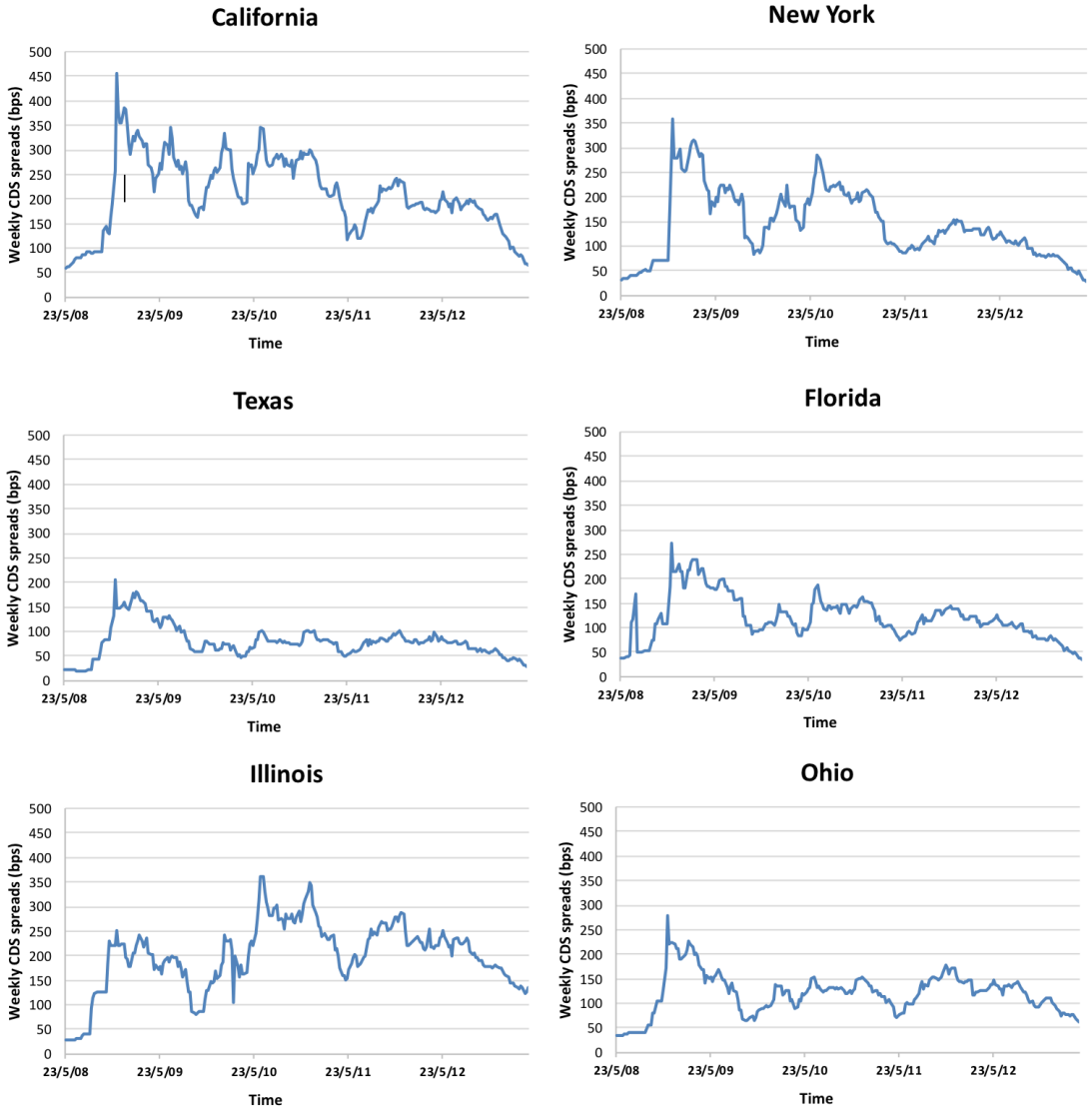
is estimated for each state separately, using the derived SCRI as well as the five additional forecasting variables.

Table 3.4 presents the results for the estimated regression model for each state. We find that the average explanatory power of the estimated models, measured by R^2, across all states is around 0.64. The coefficient of determination ranges from 0.28 for Connecticut up to 0.83 for New Jersey, and exceeds 0.5 for more than half the sample.

We are particularly interested in the predictive power of the SCRIs, presented in the first column of Table 3.4. The coefficient of the variable is positive and significant at the 1% level for 16 out of 19 states. The estimated coefficients in 16 states show the expected sign, suggesting that an increase in a state's credit risk at the firm level at time t will typically lead to an increase in sovereign risk for the state at time $t+1$. The results are consistent with our hypothesis that firm level information forecasts overall sovereign risk.

We further observe that while the coefficients for the SCRIs are mostly positive and significant for the 19 states, estimated coefficients for the other explanatory

Figure 3.3. Time series of weekly observations for CDS spreads for the states of California *(upper left panel)*, New York *(upper right panel)*, Texas *(middle left panel)*, Florida *(middle right panel)*, Illinois *(lower left panel)* and Ohio *(lower right panel)* from May 2008 to April 2013.



variables, except for the Treasury CDS spreads, are generally less significant with inconsistent signs. We argue that this is not necessarily inconsistent with the expected univariate relationship between these variables and default risk at the state level, since changes in less significant variables may well be captured by other variables, in particular the developed bottom-up SCRIs. For example, the estimation of EDFs for a company will take into account the volatility of the individual stock, that will also be related to the overall volatility of the equity market. Therefore, it is likely that information similar to that provided by the VIX will also be incorporated in the derived SCRIs. Changes in the VIX will most likely be accompanied by changes in EDFs, and consequently in SCRIs.

	R^2	Obs	SCRI	VIX	TS	S&P500	T-CDS	CDX IG
California	0.66	281	40.27***	-0.91***	17.80***	1.13	2.37***	6.59**
			(5.55)	(0.35)	(4.52)	(1.01)	(0.22)	(2.98)
New York	0.74	360	58.72***	-0.70***	-6.19***	1.52**	1.47***	1.97
			(3.74)	(0.24)	(1.81)	(0.70)	(0.14)	(2.10)
Texas	0.81	284	24.68***	0.22**	-4.29***	1.02***	1.01***	2.15**
			(2.03)	(0.11)	(1.37)	(0.32)	(0.07)	(0.94)
Florida	0.77	360	32.88***	0.23	-4.18***	1.35***	1.03***	1.83
			(2.21)	(0.15)	(1.11)	(0.44)	(0.09)	(1.31)
Illinois	0.56	284	-26.31***	-0.24	17.16***	1.27	4.06***	4.52
			(6.62)	(0.38)	(4.79)	(1.11)	(0.24)	(3.29)
Pennsylvania	0.55	360	-35.10***	-0.26	2.56**	0.25	1.81***	1.43
			(2.80)	(0.17)	(1.26)	(0.49)	(0.10)	(1.48)
Ohio	0.70	257	18.74***	0.33**	-12.59***	1.78***	1.73***	3.14**
			(3.05)	(0.17)	(2.39)	(0.49)	(0.12)	(1.42)
New Jersey	0.83	360	28.69***	0.10	3.82**	1.64***	2.36***	4.05**
			(3.11)	(0.21)	(1.57)	(0.62)	(0.13)	(1.84)
Michigan	0.76	272	61.26***	-0.14	4.29	2.12***	1.84***	6.47***
			(5.09)	(0.28)	(3.90)	(0.81)	(0.19)	(2.39)
Massachusetts	0.75	280	25.83***	-0.10	-3.18	1.31***	1.76***	2.59*
			(3.08)	(0.18)	(2.34)	(0.51)	(0.11)	(1.51)
North Carolina	0.62	246	38.04***	-1.23***	8.94***	0.46	0.26*	3.97***
			(3.17)	(0.17)	(2.57)	(0.52)	(0.14)	(1.51)
Virginia	0.77	295	32.58***	-0.82***	2.88***	-0.47*	-0.35***	-0.46
			(1.43)	(0.09)	(0.62)	(0.26)	(0.06)	(0.73)
Wisconsin	0.40	173	29.01***	0.75**	-1.71	0.99	0.58**	-0.41
			(6.78)	(0.32)	(2.96)	(0.69)	(0.25)	(1.56)
Maryland	0.60	360	26.39***	-1.06***	-0.82	0.41	0.90***	2.95**
			(2.47)	(0.17)	(1.24)	(0.49)	(0.10)	(1.46)
Minnesota	0.67	130	56.83***	-0.75**	37.49***	0.03	0.17	7.42***
			(12.97)	(0.31)	(4.46)	(0.03)	(0.20)	(0.88)
Connecticut	0.28	207	4.63	0.54	-5.19*	1.02	1.46***	0.89
			(5.18)	(0.33)	(3.06)	(0.68)	(0.21)	(1.65)
Delaware	0.41	192	11.65***	0.65***	-2.97	1.07***	0.47***	1.30
			(3.06)	(0.19)	(1.88)	(0.39)	(0.11)	(0.89)
Nevada	0.77	271	39.65***	-0.13	5.23	1.41	1.92***	4.79***
			(3.79)	(0.22)	(3.13)	(0.63)	(0.15)	(1.86)
Rhodes Island	0.53	142	57.90***	-0.16	-4.01	0.85	0.74***	-0.22
			(8.39)	(0.40)	(3.25)	(0.76)	(0.29)	(1.57)

Table 3.4. Results for regressing state CDS spreads on SCRI, VIX, TS, SP500, Treasury CDS, and CDX IG, using weekly observations. For each state, the coefficient of determination (R^2) is provided in the first column, followed by the number of observations in the second column. Estimated coefficients are reported in the subsequent columns, with heteroskedasticity and autocovariance consistent (HAC) standard errors (Newey & West, 1987) in brackets. *,** and *** indicate significance of the coefficients at the 10%, 5% and 1% level, respectively.

Therefore, estimated coefficients for the VIX may have unexpected signs since the linkage between the VIX and observed state CDS spreads is already partially explained by the estimated coefficient for the SCRI. Similar arguments can be made regarding the other explanatory variables.

In order to further investigate the predictive performance of the SCRIs, we

compare the full model regression results to results for restricted models without SCRI as explanatory variable.[4] We carry out model comparison tests for each state to examine the fit of the full model relative to the nested models and present the results in Table 3.5. The first two columns provide the R^2 values for the models, with the column indicating the goodness-of-fit for the nested model, and the second column referring to the full model (3.3) including SCRI as predictive variable. The third and fourth columns report the F-statistics and the corresponding p-values for significance of a superior fit of the full model.

	Restricted Model	Full Model	F-stat	p-value
California	0.59	0.66	52.70	0.00
New York	0.55	0.74	246.60	0.00
Texas	0.71	0.81	147.86	0.00
Florida	0.63	0.77	220.51	0.00
Illinois	0.53	0.56	15.82	0.00
Pennsylvania	0.35	0.55	157.17	0.00
Ohio	0.66	0.70	37.78	0.00
New Jersey	0.78	0.83	85.16	0.00
Michigan	0.63	0.76	144.82	0.00
Massachusetts	0.68	0.75	70.35	0.00
North Carolina	0.39	0.62	144.16	0.00
Virginia	0.36	0.77	518.67	0.00
Wisconsin	0.33	0.40	18.32	0.00
Maryland	0.47	0.60	114.32	0.00
Minnesota	0.62	0.67	19.21	0.00
Connecticut	0.27	0.28	0.80	0.37
Delaware	0.36	0.41	14.54	0.00
Nevada	0.68	0.77	109.41	0.00
Rhodes Island	0.36	0.53	47.66	0.00

Table 3.5. Results for model comparison tests to examine the superior fit of the full model in comparison to a restricted model that excludes the SCRIs. The first column provides the coefficient of determination for the nested model, the second column the coefficient of determination for the full model. The third and fourth columns present the F-statistic and the corresponding p-values for significance of a superior fit of the full model.

Our findings suggest that the coefficient of determination is typically much higher for the full model that includes the SCRIs. The average R^2 also increases from 0.53 for the restricted model in comparison to 0.64 for the full model, while the R^2 even doubles for the state of Virginia. Model comparison tests also support the full model as being superior for 18 out of 19 states. These results suggest that SCRIs are not only statistically significant but their inclusion substantially

[4] Model estimates and coefficients for the restricted models are not reported here but are available upon request to the authors.

increases predictive fit. SCRIs appear to offer substantial predictive gains with reference to CDS spreads.

Overall, our results strongly support the predictive power of granular information from the private sector in assessing sovereign default risk at the state level. Information on default risk at the company level appears to be an important determinant of the market's view on the ability of a state government to service its debt securities. The estimated positive coefficients for the SCRIs imply that changes in the median corporate credit risk within a state at time t can help predict sovereign risk at *t+1*. In the following sections we conduct a number of tests to examine the robustness of our results.

3.4.2 Robustness Checks

3.4.2.1 Results for Monthly Frequency

We first test the predictive power of SCRIs for state CDS spreads, by also looking at monthly observations.[5] Overall, financial variables such as CDS spreads are expected to react rather quickly to changes in market perceptions on credit risk conditions at the company or state level. However, information contained in SCRIs may also influence perceived risks for the credit quality of a state over a longer time horizon. Therefore, the SCRIs are expected to still have significant predictive power for state CDS spreads when the relationship is examined using monthly frequencies. We re-estimate model (3.3) using monthly observations, applying a one-month lag to the explanatory variables. If the suggested relationship between the derived bottom-up risk indicators and state CDS spreads is robust, the estimated models should remain significant.

Monthly regression results for monthly observations are shown in Table 3.6. Our findings illustrate that also for monthly frequencies, 15 out of 18 states yield positive and significant coefficients for SCRIs. Also Treasury CDS spreads and the CDX IG index remain highly significant in most of the estimated models. The average R^2 across all states is 0.70, and for most states is slightly higher in

[5] We exclude Minnesota in this robustness check, due to its low number of monthly CDS spread observations.

	R^2	Obs	SCRI	VIX	TS	S&P500	T-CDS	CDX IG
California	0.68	65	39.57***	-1.02	17.19*	-0.49	2.09***	13.54**
			(11.94)	(0.87)	(9.74)	(2.26)	(0.45)	(5.61)
New York	0.78	83	57.76***	-0.50	-5.58	0.29	1.32***	8.17**
			(7.52)	(0.55)	(3.59)	(1.47)	(0.28)	(3.87)
Texas	0.84	66	23.64***	0.14	-4.93*	1.25*	0.98***	3.46**
			(4.09)	(0.27)	(2.69)	(0.67)	(0.13)	(1.67)
Florida	0.83	83	31.45***	0.15	-5.04**	0.69	1.08***	7.72***
			(4.20)	(0.33)	(2.08)	(0.86)	(0.17)	(2.27)
Illinois	0.55	66	-31.90**	-0.13	18.65*	1.96	4.00***	10.14
			(15.05)	(1.02)	(10.59)	(2.63)	(0.52)	(6.59)
Pennsylvania	0.57	83	-39.98***	0.11	2.46	-0.01	1.86***	7.01**
			(5.97)	(0.42)	(2.65)	(1.09)	(0.21)	(2.88)
Ohio	0.76	60	18.96***	0.36	-14.98***	1.07	1.62***	5.55**
			(6.18)	(0.40)	(4.72)	(1.02)	(0.24)	(2.52)
New Jersey	0.84	83	28.71***	0.02	4.48***	1.68	2.22***	8.34**
			(6.46)	(0.51)	(3.22)	(1.33)	(0.26)	(3.50)
Michigan	0.80	63	64.46***	-0.31	-0.94	2.07	1.61***	7.13**
			(10.52)	(0.68)	(7.86)	(1.74)	(0.37)	(4.31)
Massachusetts	0.77	65	25.58***	-0.24	-5.43	1.89*	1.70***	5.99**
			(6.56)	(0.43)	(4.94)	(1.13)	(0.23)	(2.81)
North Carolina	0.68	57	39.17***	-1.23***	6.76	-0.71	0.07*	5.28**
			(6.44)	(0.43)	(5.06)	(1.10)	(0.28)	(2.68)
Virginia	0.81	68	33.29***	-0.85***	2.27*	-0.70	-0.37***	1.26
			(3.06)	(0.22)	(1.28)	(0.56)	(0.12)	(1.39)
Wisconsin	0.57	40	27.72*	1.11	-0.58	1.21	0.54	5.56**
			(14.42)	(0.69)	(6.54)	(1.57)	(0.49)	(2.62)
Maryland	0.62	83	26.95***	-1.35***	-0.66*	-0.86	0.87***	5.32**
			(5.22)	(0.41)	(2.59)	(1.07)	(0.21)	(2.82)
Connecticut	0.45	48	-2.44	0.98	-3.49	3.36**	1.21***	7.26***
			(10.03)	(0.65)	(6.14)	(1.48)	(0.37)	(2.61)
Delaware	0.59	45	13.72**	0.58*	-5.28	2.37***	0.44**	3.90***
			(5.84)	(0.35)	(3.72)	(0.79)	(0.20)	(1.36)
Nevada	0.82	63	43.94***	-0.44	-0.10	1.41	1.71***	6.49**
			(7.63)	(0.51)	(6.10)	(1.31)	(0.28)	(3.26)
Rhodes Island	0.64	33	60.25***	-0.38	-9.00	1.74	1.04*	4.47*
			(17.35)	(0.81)	(7.25)	(1.68)	(0.55)	(2.60)

Table 3.6. Results for regressing state CDS spreads on SCRI, VIX, TS, SP500, Treasury CDS, and CDX IG using monthly observations. For each state, the coefficient of determination (R^2) is provided in the first column, followed by the number of observations in the second column. Estimated coefficients are reported in the subsequent columns, with heteroskedasticity and autocovariance consistent (HAC) standard errors (Newey & West, 1987) in brackets. *,** and *** indicate significance of the coefficients at the 10%, 5% and 1% level, respectively.

comparison to the results for weekly data.[6] While the number of observations in each state is much smaller because of the change from weekly to monthly frequency,

[6] Note, however, that as pointed out by Boudoukh *et al.* (2008) higher levels of predictability with widening horizons are to be expected in longer term horizon regressions. As the sampling error that is almost surely present in small samples shows up in each regression, both the estimator and R^2 are proportional to the forecast horizon. Therefore, better results for long horizons in the form of higher increasing R^2s generally provide little if any evidence for a better forecasting performance over and above the weekly results. From this perspective, the increasing explanatory power of the applied models for monthly horizons should be interpreted with care.

our main results on the predictive power of the constructed SCRIs for sovereign default risk at the state level are confirmed.

3.4.2.2 Using ICRIs based on the mean of corporate risk measures

We also test the predictive relationship between the SCRIs and state CDS spreads, using a slightly different approach for the construction of the SCRIs. That is, instead of using the median of the observed EDFs for all companies in a specific industry, we use the mean of the EDFs to construct the ICRI. The mean is expected to be more sensitive to changes in EDFs of companies with a higher default risk. An SCRI based on such ICRIs is also likely to exhibit more volatility through time. Results for this alternative specification of the SCRIs are reported in Table 3.7.

Our findings suggest that the results are also quite robust with respect to the construction of the ICRIs and SCRIs. Again we find that for 16 out of 19 states the model yields positive and significant coefficients on SCRIs. The average R^2 for the estimated models is 0.64 and thus very similar to the baseline specification of the model. The significance of the coefficients and the explanatory power of the estimated models for the individual states are qualitatively unchanged. However, we observe that in comparison to the baseline model, the estimated SCRI coefficients are much smaller in magnitude. This is a result of the skewed distribution of company EDFs for each industry that leads to the mean typically being significantly higher than the median. At the same time, neither the sign nor the magnitude of the estimated coefficients for the other explanatory variables change significantly. Overall, these findings strongly support the robustness of the baseline results.

3.4.2.3 A bottom-up Credit Risk Indicator using Top Industries only

So far the derived SCRIs were based on all industries contributing to a state's GDP output. However, one could argue that it is typically the major industries with a large contribution to a state's GDP, that will have the largest influence on market perceptions of sovereign risk. Therefore, as an additional robustness check, we examine the sensitivity of our results with regards to the construction of the bottom-up indicators. To do this, we use an alternative approach and include only

	R^2	Obs	SCRI	VIX	TS	S&P500	T-CDS	CDX IG
California	0.70	281	25.57***	-1.28***	13.46***	0.41	2.11***	4.78*
			(2.62)	(0.33)	(4.26)	(0.95)	(0.22)	(2.98)
New York	0.72	360	27.47***	-0.56**	-11.08***	1.27*	1.31***	1.14
			(1.91)	(0.24)	(2.01)	(0.73)	(0.15)	(2.19)
Texas	0.83	284	12.59***	0.16	-5.98***	0.79***	0.97***	1.42
			(0.89)	(0.11)	(1.32)	(0.30)	(0.06)	(0.89)
Florida	0.77	360	16.26***	0.27*	-7.54***	1.14***	0.94***	1.20
			(1.11)	(0.15)	(1.20)	(0.44)	(0.09)	(1.32)
Illinois	0.54	284	-6.63**	-0.63	13.13***	1.04	3.87***	3.81
			(3.52)	(0.39)	(5.00)	(1.14)	(0.24)	(3.38)
Pennsylvania	0.48	360	-12.97***	-0.52***	4.04***	0.19	1.76***	1.43
			(1.36)	(0.18)	(1.43)	(0.53)	(0.11)	(1.59)
Ohio	0.73	257	10.78***	0.14	-14.27***	1.44***	1.73***	2.35*
			(1.30)	(0.16)	(2.24)	(0.47)	(0.11)	(1.36)
New Jersey	0.83	360	15.35***	0.08	0.12	1.37**	2.21***	3.25*
			(1.54)	(0.20)	(1.66)	(0.61)	(0.13)	(1.82)
Michigan	0.76	272	28.36***	-0.23	4.26	1.60**	1.93***	5.24**
			(2.27)	(0.28)	(3.82)	(0.81)	(0.18)	(2.37)
Massachusetts	0.76	280	13.32***	-0.18	-4.62**	1.03**	1.71***	1.87
			(1.41)	(0.17)	(2.31)	(0.50)	(0.11)	(1.48)
North Carolina	0.65	246	18.71***	-1.40***	9.31***	0.12	0.42***	3.22**
			(1.47)	(0.17)	(2.42)	(0.50)	(0.13)	(1.47)
Virginia	0.68	295	14.19***	-0.66***	-0.29	-0.46	-0.33***	-0.67
			(0.85)	(0.10)	(0.81)	(0.31)	(0.07)	(0.88)
Wisconsin	0.42	173	10.60***	0.96***	2.41	0.97	0.65***	-0.33
			(2.18)	(0.28)	(2.19)	(0.68)	(0.25)	(1.53)
Maryland	0.63	360	15.02***	-1.14***	-4.52***	0.10	0.74***	2.03
			(1.18)	(0.16)	(1.27)	(0.47)	(0.10)	(1.40)
Minnesota	0.62	130	27.37***	0.29	-3.76***	0.03	0.06	-6.06***
			(2.59)	(0.32)	(0.53)	(0.03)	(0.22)	(0.94)
Connecticut	0.28	207	2.58	0.51*	-5.38**	0.95	1.50***	0.82
			(1.97)	(0.30)	(2.62)	(0.68)	(0.21)	(1.64)
Delaware	0.42	192	4.97***	0.69***	-2.40	1.04***	0.50***	1.36
			(1.21)	(0.17)	(1.67)	(0.38)	(0.12)	(0.89)
Nevada	0.77	271	18.50***	-0.16	5.99*	1.14*	1.92***	3.97**
			(1.78)	(0.22)	(3.11)	(0.64)	(0.15)	(1.88)
Rhodes Island	0.62	142	21.06***	0.16	2.75	0.59	0.94***	-0.23
			(2.17)	(0.31)	(2.29)	(0.68)	(0.26)	(1.40)

Table 3.7. Results for regressing weekly state CDS spreads on the newly developed SCRI based on mean of the EDFs for each industry to construct the ICRIs. Additional explanatory variables are the same as in the baseline model, i.e., VIX, TS, SP500, Treasury CDS, and CDX IG. For each state, the coefficient of determination (R^2) is provided in the first column, followed by the number of observations in the second column. Estimated coefficients are reported in the subsequent columns, with heteroskedasticity and autocovariance consistent (HAC) standard errors (Newey & West, 1987) in brackets. *,** and *** indicate significance of the coefficients at the 10%, 5% and 1% level, respectively.

the five largest industries in a state (with regards to their contribution to GDP) when constructing the SCRIs. Typically, for our sample, the five major industries constitute more than half of the total GDP of a state and, thus, are expected to exert a large influence on a state's economy.

As a first step we identify the five largest industries in each state. We observe that

certain industries such as retail/wholesale trade and the real estate sector play a dominant role in most states, while the contribution of other industries varies significantly across the states. For example, the mining industry is the fourth largest industry in Texas, but it constitutes a small proportion of the GDP in most of the other states. We also find that companies in the information sector (media, computer software, publishing and telephone,) make a large contribution to state GDPs in California and Wisconsin, but not elsewhere.

The SCRIs are then computed as the weighted average of the corresponding ICRIs for the five major industries. We then test the predictive power, using the revised SCRIs with concentrated industry compositions together with the supplementary predictive variables. SCRIs based only on the most important industries for a state are still expected to provide important information for the assessment of sovereign default risk. Thus, we expect qualitatively similar results.

Table 3.8 shows regression results for the revised SCRIs. The average R^2 is 0.65 and 16 states have positive, significant coefficients also for the revised SCRIs. The higher overall R^2 suggests a slightly higher predictive power of the models when only major industries are used for construction of the SCRIs. Overall, the results affirm the predictive power of the developed bottom-up credit risk indicators.

3.4.2.4 Through-the-Cycle Credit Risk Measures

We next examine the predictive relationship when the effect of the credit cycle on company level default risk is filtered. Therefore, instead of using point-in-time EDFs at the company level to derive the ICRIs and SCRIs, we use Moody's KMV's through-the-cycle EDFs (TTCEDFs) to develop these indicators. These robustness tests will then help us to confirm whether credit risk information at the company level is significant in assessing sovereign risk independent of the credit cycle.

Like standard EDFs, TTCEDFs also provide a measure of credit quality for a firm over a one-year time horizon. However, standard EDFs are so-called point-in-time (PIT) measures and thus incorporate not only information about companies' individual credit risk profile, but also geographic, sectoral as well as cyclical macro-credit factors. Hence, Moody's KMV argue that standard PIT EDF measures as we have used them so far in this analysis typically provide early warning signals of rapid changes in default risk. On the other hand, TTCEDF

	R^2	Obs	SCRI	VIX	TS	S&P500	T-CDS	CDX IG
California	0.64	281	30.17***	-0.70*	21.99***	1.27	2.50***	7.24**
			(4.94)	(0.36)	(4.48)	(1.03)	(0.22)	(3.05)
New York	0.73	360	55.91***	-0.53**	-5.99***	1.57**	1.48***	2.17
			(3.67)	(0.23)	(1.83)	(0.71)	(0.15)	(2.13)
Texas	0.81	284	21.58***	0.20*	-3.07**	1.00***	1.01***	2.24**
			(1.76)	(0.11)	(1.32)	(0.32)	(0.07)	(0.94)
Florida	0.79	360	29.99***	0.32**	-3.59***	1.37***	1.04***	1.99
			(2.09)	(0.15)	(1.12)	(0.44)	(0.09)	(1.32)
Illinois	0.58	284	-26.35***	-0.25	17.56***	1.31	4.08***	4.58
			(6.14)	(0.37)	(4.73)	(1.10)	(0.23)	(3.27)
Pennsylvania	0.57	360	-33.39***	-0.35**	2.41*	0.23	1.80***	1.34
			(2.66)	(0.17)	(1.26)	(0.49)	(0.10)	(1.48)
Ohio	0.72	257	16.96***	0.38**	-12.00***	1.82***	1.74***	3.25**
			(2.90)	(0.16)	(2.38)	(0.49)	(0.13)	(1.43)
New Jersey	0.84	360	25.63***	0.20	4.23***	1.69***	2.38***	4.23**
			(2.97)	(0.21)	(1.58)	(0.62)	(0.13)	(1.86)
Michigan	0.77	272	55.34***	-0.08	9.04**	2.18**	1.91***	6.97***
			(4.98)	(0.29)	(3.85)	(0.83)	(0.19)	(2.45)
Massachusetts	0.77	280	24.46***	-0.03	-2.97	1.32***	1.77***	2.65*
			(2.91)	(0.17)	(2.33)	(0.51)	(0.11)	(1.51)
North Carolina	0.63	246	33.61***	-1.12***	9.59***	0.49	0.28*	4.11***
			(2.89)	(0.17)	(2.59)	(0.52)	(0.14)	(1.53)
Virginia	0.79	295	28.45***	-0.76***	2.92***	-0.47*	-0.33***	-0.41
			(1.24)	(0.08)	(0.62)	(0.26)	(0.06)	(0.73)
Wisconsin	0.40	173	23.84***	0.97***	0.53	1.14	0.55**	-0.26
			(6.96)	(0.32)	(2.91)	(0.70)	(0.25)	(1.58)
Maryland	0.62	360	24.93***	-1.02***	-0.57	0.41**	0.90***	3.00**
			(2.31)	(0.16)	(1.23)	(0.49)	(0.10)	(1.45)
Minnesota	0.67	130	50.74***	-0.66**	36.96***	0.03	0.13	7.21***
			(11.88)	(0.31)	(4.69)	(0.03)	(0.20)	(0.87)
Connecticut	0.28	207	5.13	0.51	-5.57*	1.00	1.47***	0.84
			(4.82)	(0.33)	(3.04)	(0.68)	(0.21)	(1.64)
Delaware	0.40	192	10.77***	0.68***	-2.79	1.09***	0.45***	1.33
			(3.03)	(0.19)	(1.94)	(0.39)	(0.11)	(0.90)
Nevada	0.76	271	36.57***	0.07	7.00**	1.50**	1.96***	5.14***
			(3.76)	(0.21)	(3.15)	(0.65)	(0.15)	(1.89)
Rhodes Island	0.52	142	56.95***	-0.12	-4.22	0.86	0.70**	-0.22
			(8.41)	(0.40)	(3.31)	(0.77)	(0.29)	(1.57)

Table 3.8. Results for regressing state CDS spreads on SCRI, VIX, TS, SP500, Treasury CDS, and CDX IG, using weekly observations and constructing the SCRIs based on the state's top five industries only. For each state, the coefficient of determination (R^2) is provided in the first column, followed by the number of observations in the second column. Estimated coefficients are reported in the subsequent columns, with heteroskedasticity and autocovariance consistent (HAC) standard errors (Newey & West, 1987) in brackets. *,** and *** indicate significance of the coefficients at the 10%, 5% and 1% level, respectively.

measures isolate a company's underlying credit trend from the macro-credit cyclical effect. TTCEDF are primarily driven by changes in a company's long-run credit quality, which tends to be more stable over time and exhibit less variation. Thus, while we expect EDFs and TTCEDFs to share a similar (long-run) trend, since they are both developed using a company's own credit risk profile, the values of the two measures may differ significantly during a major credit crisis, since

TTCEDFs will minimize the impact of the macro-credit cycle.

To derive the revised risk indicators, we implement the same approach as outlined earlier for the construction of the original SCRIs. Weekly TTCEDFs for all listed companies are collected based on the last trading day of each week. Using the same industry categories as described in Table 3.1, companies are then grouped into 15 industries and the median TTCEDF of all companies in the same industry is taken as the industry's TTCEDF. This results in 15 through-the-cycle industry credit risk indicators (TTCICRIs).

Figure 3.4 demonstrates the significant differences between the original ICRIs and the new TTCICRIs. As expected, the industry of *real estate, rental and leasing* was greatly affected during the subprime credit crisis, leading to a peak of default risk implied by the ICRI in late 2008 and during 2009. However, as shown in the upper panel, TTCICRIs for the industry are significantly less affected by the GFC period. Similar observations can be made for the industry of *retail/wholesale trade* and *utilities* in the middle and bottom panel of Figure 3.4. The effect of the GFC is diminished for the derived TTCICRIs, resulting in a more stable measure of credit risk at the company and industry level. Still we find that also for the constructed TTC industry measures average values as well as the dynamics of the risk indicators differ significantly for the 15 industries.

Based on an industry's percentage contribution to total GDP in each state, the TTCICRIs are then used to derive through-the-cycle state credit risk indicators (TTCSCRIs). The same set of industry weights that has been used for the construction of the SCRIs is adopted here, such that the TTCSCRI for each state is essentially a weighted average of the 15 TTCICRIs. We plot the original SCRI as well as the TTCSCRI for California in Figure 3.5. As expected, the two time series differ significantly, in particular, during the period from late 2008 to early 2010, when the TTCSCRI is significantly less affected by market conditions prevalent during the crisis period. Similar observations can be made for the other states in our sample.

We re-estimate model 3.3, replacing the SCRIs with the calculated TTCSCRIs, to examine the predictive power of the explanatory variables for sovereign default risks. We consider observations at weekly frequencies and present the results for the regression in Table 3.9. We find that the average explanatory power (measured by R^2) is 0.65, and therefore, quite similar to the results obtained for the original

Figure 3.4. Time series of constructed TTCICRI and ICRIs for selected industries: *real estate, rental and leasing (upper panel)*, *retail/wholesale trade (middle panel)*, and *utilities (lower panel)* for the sample period June 2006 to April 2013.

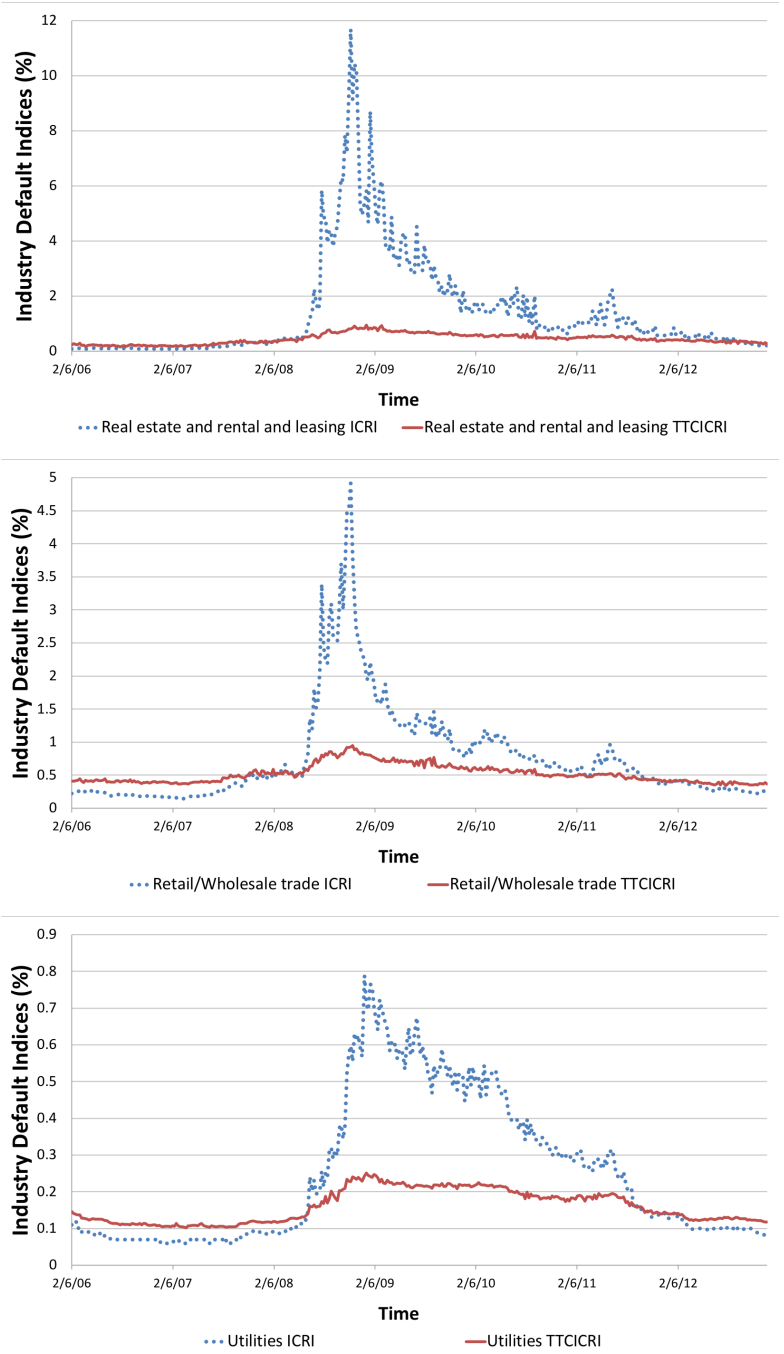
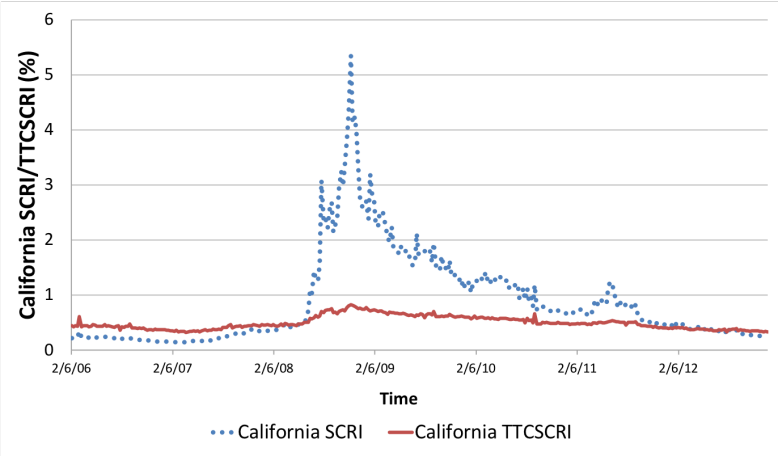


Figure 3.5. Time series of constructed TTCSCRI and SCRI for the state of California during the sample period June 2006 to April 2013.



SCRIs. Again we find that for 16 of the 19 states the estimated coefficients for the TTCSCRI are both positive and significant, which is consistent with earlier findings. Thus, our results on the predictive power of the constructed SCRIs also hold for measures being based on through-the-cycle EDFs. Overall, these results suggest the importance of adopting firm-level information in assessing sovereign risk independent of the credit cycle.

3.4.2.5 Predictive Model with Lagged CDS Changes

Up until now the CDS spreads have only been used as dependent variables to represent the level of sovereign risk. In this robustness test we include lagged changes in state CDS spreads to examine whether changes in CDS spreads help forecast future CDS spreads. We are interested in examining the new predictive relationship and in particular whether the SCRIs are still of incremental importance. If the coefficients for the SCRIs in this model are positive and significant for predicting CDS spreads, it can be concluded that the predictive power of the baseline model is not due to the correlation with the lagged changes in the dependent variable. If the relationships between the SCRIs and the dependent variables are robust, we expect SCRIs to retain their incremental importance.

Following the model in Aizenman *et al.* (2013), we test the predictive power of the regression model specified as follows, where we use the lagged changes in the dependent variable, $CDS_{i,t-1} - CDS_{i,t-2}$, to predict the value of the dependent variable at time t:

	R^2	Obs	TTCSCRI	VIX	TS	S&P500	T-CDS	CDX IG
California	0.72	281	461.97***	-1.75***	-5.08	-0.17	2.65***	3.86
			(41.96)	(0.33)	(5.03)	(0.93)	(0.18)	(2.73)
New York	0.76	360	437.08***	-0.76***	-12.35***	0.81	2.08***	1.20
			(24.93)	(0.22)	(1.84)	(0.68)	(0.13)	(2.01)
Texas	0.8	284	183.82***	0.15	-9.94***	0.69	1.22***	1.67*
			(16.41)	(0.12)	(1.74)	(0.33)	(0.06)	(0.98)
Florida	0.79	360	239.24***	0.29**	-8.70***	0.98**	1.39***	1.31
			(15.00)	(0.14)	(1.18)	(0.43)	(0.08)	(1.27)
Illinois	0.54	284	-66.67	-0.77*	12.83**	0.91	3.70***	3.33
			(55.67)	(0.41)	(6.01)	(1.16)	(0.22)	(3.41)
Pennsylvania	0.47	360	-176.98***	-0.60	4.26***	0.25**	1.45***	1.13
			(19.82)	(0.18)	(1.47)	(0.54)	(0.10)	(1.61)
Ohio	0.74	257	213.84***	0.01	-23.60	1.15*	1.77***	1.84*
			(23.32)	(0.16)	(2.78)	(0.46)	(0.11)	(1.33)
New Jersey	0.84	360	245.83***	-0.01**	-1.48***	1.11	2.64***	3.16*
			(21.20)	(0.20)	(1.62)	(0.59)	(0.11)	(1.75)
Michigan	0.79	272	553.22***	-0.65	-19.49***	0.83*	2.16***	4.00
			(38.1)	(0.27)	(4.52)	(0.77)	(0.16)	(2.24)
Massachusetts	0.75	280	214.57***	-0.25	-10.61***	0.93*	1.98***	1.87
			(24.61)	(0.18)	(2.83)	(0.51)	(0.10)	(1.51)
North Carolina	0.74	246	395.75***	-1.75***	-8.89***	-0.58	0.42***	1.89
			(22.48)	(0.15)	(2.67)	(0.44)	(0.11)	(1.28)
Virginia	0.66	295	174.91***	-0.59***	0.28	-0.54*	0.13**	-0.25
			(10.92)	(0.10)	(0.82)	(0.32)	(0.06)	(0.89)
Wisconsin	0.45	173	191.18***	0.36	-8.89***	0.35	0.54**	-0.90
			(31.63)	(0.32)	(3.33)	(0.68)	(0.24)	(1.49)
Maryland	0.54	360	139.57***	-0.79***	-2.60*	0.44	1.22***	3.21**
			(18.43)	(0.17)	(1.43)	(0.53)	(0.10)	(1.56)
Minnesota	0.72	130	207.28***	0.28	-38.03***	0.03	0.04	-6.07***
			(49.42)	(0.32)	(5.01)	(0.04)	(0.21)	(0.97)
Connecticut	0.28	207	53.33	0.38	-7.76**	0.83	1.47***	0.73
			(37.5)	(0.35)	(3.82)	(0.70)	(0.20)	(1.64)
Delaware	0.40	192	81.53***	0.67***	-3.55	0.97**	0.42***	1.36
			(25.34)	(0.20)	(2.29)	(0.40)	(0.11)	(0.90)
Nevada	0.8	271	378.34***	-0.45**	-11.11***	0.60	2.15***	3.30*
			(30.57)	(0.21)	(3.72)	(0.61)	(0.13)	(1.78)
Rhodes Island	0.62	142	338.28***	-0.45	-12.22***	0.03	0.56**	-0.81
			(35.93)	(0.35)	(3.29)	(0.70)	(0.26)	(1.42)

Table 3.9. Results for regressing state CDS spreads on TTCSCRI, VIX, TS, SP500, Treasury CDS, and CDX IG, using weekly observations. For each state, the coefficient of determination (R^2) is provided in the first column, followed by the number of observations in the second column. Estimated coefficients are reported in the subsequent columns, with heteroskedasticity and autocovariance consistent (HAC) standard errors (Newey & West, 1987) in brackets. *,** and *** indicate significance of the coefficients at the 10%, 5% and 1% level, respectively.

$$\begin{aligned} CDS_{i,t} =& \beta_{0,i} + \beta_{1,i} * SCRI_{i,t-1} + \beta_{2,i} * VIX_{t-1} + \beta_{3,i} * TS_{t-1} + \beta_{4,i} * SP500_{t-1} \\ &+ \beta_{5,i} * TCDS_{t-1} + \beta_{6,i} * CDX_{t-1} + \beta_{7,i} * (CDS_{i,t-1} - CDS_{i,t-2}) + \epsilon_i \end{aligned} \tag{3.4}$$

The regression results are quite robust and are presented in Table 3.10. The

	R^2	Obs	SCRI	VIX	TS	S&P500	T-CDS	CDX IG	$\Delta CDS_{i,t-1}$
California	0.67	279	42.12***	-1.23***	17.72***	1.72*	2.34***	6.39**	0.62***
			(5.44)	(0.35)	(4.47)	(0.99)	(0.21)	(2.92)	(0.15)
New York	0.75	358	59.42****	-0.85***	-5.48***	2.00***	1.41***	2.38	0.55***
			(3.65)	(0.23)	(1.78)	(0.69)	(0.14)	(2.05)	(0.12)
Texas	0.81	282	25.54***	0.16	-4.64***	1.08***	0.98***	2.40**	0.29**
			(2.05)	(0.12)	(1.41)	(0.32)	(0.07)	(0.94)	(0.12)
Florida	0.78	358	33.37***	0.14	-3.88***	1.59***	1.01***	2.50*	0.38***
			(2.18)	(0.15)	(1.10)	(0.43)	(0.09)	(1.29)	(0.09)
Illinois	0.55	282	-25.38***	-0.31	16.25***	1.54	4.03***	4.65	0.31
			(6.66)	(0.39)	(4.97)	(1.12)	(0.24)	(3.29)	(0.20)
Pennsylvania	0.56	358	-35.51***	-0.27	2.60**	0.49	1.82***	1.48	0.66***
			(2.77)	(0.17)	(1.25)	(0.49)	(0.10)	(1.46)	(0.20)
Ohio	0.70	255	19.93***	0.25	-13.20***	1.80***	1.67***	3.11**	0.22*
			(3.09)	(0.17)	(2.41)	(0.49)	(0.13)	(1.42)	(0.13)
New Jersey	0.83	358	29.81***	-0.06	4.16***	1.83***	2.33***	4.43**	0.42***
			(3.09)	(0.21)	(1.55)	(0.61)	(0.13)	(1.82)	(0.12)
Michigan	0.76	270	63.16***	-0.37	4.32	2.08**	1.81***	6.82***	0.44***
			(5.09)	(0.29)	(3.90)	(0.81)	(0.19)	(2.37)	(0.14)
Massachusetts	0.74	278	26.30***	-0.14	-3.42	1.40***	1.73***	2.84*	0.17
			(3.11)	(0.18)	(2.37)	(0.52)	(0.11)	(1.53)	(0.16)
North Carolina	0.64	244	38.86***	-1.32***	8.52***	0.43	0.17	3.71**	0.63***
			(3.07)	(0.17)	(2.49)	(0.50)	(0.14)	(1.47)	(0.16)
Virginia	0.79	293	32.47***	-0.85***	3.11***	-0.32	-0.36***	-0.42	0.52***
			(1.39)	(0.08)	(0.61)	(0.25)	(0.06)	(0.71)	0.11
Wisconsin	0.44	171	33.43***	0.57*	-2.08	1.30*	0.50**	-0.00	0.36
			(6.90)	(0.32)	(2.91)	(0.72)	(0.24)	(1.54)	(0.22)
Maryland	0.60	358	26.67***	-1.12***	-0.60	0.44	0.88***	3.00**	0.36***
			(2.45)	(0.17)	(1.24)	(0.49)	(0.10)	(1.45)	(0.13)
Minnesota	0.72	128	47.60***	-0.48	40.04***	0.02	0.18	7.54***	0.59***
			(12.59)	(0.30)	(4.35)	(0.03)	(0.19)	(0.83)	(0.17)
Connecticut	0.29	205	2.37	0.56	-4.27	1.17	1.45***	0.90	0.22
			(5.43)	(0.34)	(3.14)	(0.73)	(0.21)	(1.66)	(0.19)
Delaware	0.39	190	9.72***	0.68***	-2.47	1.16***	0.41***	1.24	0.17
			(3.01)	(0.18)	(1.83)	(0.37)	(0.11)	(0.87)	(0.16)
Nevada	0.78	269	40.85***	-0.26	5.17*	1.81***	1.90***	4.76***	0.40***
			(3.76)	(0.22)	(3.12)	(0.63)	(0.15)	(1.83)	(0.12)
Rhodes Island	0.54	140	57.95***	-0.14	-3.28	1.44*	0.60**	0.00	0.44*
			(8.59)	(0.40)	(3.27)	(0.82)	(0.30)	(1.57)	(0.24)

Table 3.10. Results for regressing state CDS spreads on SCRI, VIX, TS, SP500, Treasury CDS, CDX IG and CDS spreads changes from the previous period, using weekly observations. $\Delta CDS_{i,t-1}$ represents the new independent variable $CDS_{i,t-1} - CDS_{i,t-2}$. For each state, the coefficient of determination (R^2) is provided in the first column, followed by the number of observations in the second column. Estimated coefficients are reported in the subsequent columns, with heteroskedasticity and autocovariance consistent (HAC) standard errors (Newey & West, 1987) in brackets. *,** and *** indicate significance of the coefficients at the 10%, 5% and 1% level, respectively.

average R^2 across states is 0.66, which is 0.02 higher than that of the baseline model (3.3). We further observe that the coefficients of SCRI are positive and significant in 16 out of 19 states, indicating that the SCRIs remain important in predicting the changes in the dependent variables. This is consistent with the regression results from the baseline model. The coefficients of the new explanatory variable are all positive and some of them are significant at various levels, as

expected. The coefficients of the other variables are approximately the same, as are their standard deviations. Thus, the baseline results are largely unaffected by the inclusion of lagged changes in the dependent variable.

3.4.2.6 Contemporaneous Model

To this point, we have analyzed our baseline model (3.3) and assessed its performance by applying a range of robustness tests to it. We now examine the contemporaneous relationship between the independent variables used in model (3.3) and the sovereign risk measures. Weekly observations of SCRIs, which are the same as those used in the baseline model, are used in a contemporaneous model specified as follows:

$$\begin{aligned} CDS_{i,t} =& \beta_{0,i} + \beta_{1,i} * SCRI_{i,t} + \beta_{2,i} * VIX_t + \beta_{3,i} * TS_t + \beta_{4,i} * SP500_t \\ &+ \beta_{5,i} * TCDS_t + \beta_{6,i} * CDX_t + \epsilon_i \end{aligned} \tag{3.5}$$

The results for the contemporaneous model are presented in Table 3.11. The average R^2 across all states is about 0.63, which is slightly lower than that of the baseline model. Consistent with the baseline model, the SCRI coefficients are positive and significant at the 1% level in 16 out of the 19 states. The test results indicate that a higher value for SCRIs at time t is associated with higher state CDS spreads at that time. The results complement our previous findings, namely that the derived state credit risk indicators are not only important in forecasting CDS spreads, but also significant in a contemporaneous relationship.

3.4.2.7 Quantile Regression Models

As a last step, we conduct additional robustness tests in the form of quantile regressions. The method allows us to compute several different regression relations corresponding to various quantiles of the dependent variable and thus provides a more complete picture of the relationship between the variables (Mosteller & Tukey, 1977). In this way, results from quantile regression will provide additional information beyond the focus of least squares estimates only. As suggested by Koenker & Hallock (2001) and Cade & Noon (2003), quantile regression has

	R^2	Obs	SCRI	VIX	TS	S&P500	T-CDS	CDX IG
California	0.66	280	43.07***	-1.39***	17.05***	1.06	2.39***	6.27**
			(5.54)	(0.35)	(4.55)	(1.00)	(0.22)	(2.98)
New York	0.73	359	60.83***	-1.07***	-5.61***	0.53	1.43***	2.50
			(3.80)	(0.24)	(1.84)	(0.71)	(0.15)	(2.14)
Texas	0.80	283	27.35***	-0.00	-5.14***	1.05***	0.96***	2.28**
			(2.10)	(0.12)	(1.44)	(0.33)	(0.07)	(0.97)
Florida	0.76	359	34.09***	-0.01	-3.70***	1.16**	1.01***	2.45*
			(2.27)	(0.15)	(1.14)	(0.45)	(0.09)	(1.34)
Illinois	0.55	283	-23.93***	-0.47	15.38***	1.40	4.07***	5.25
			(6.68)	(0.39)	(4.90)	(1.12)	(0.24)	(3.32)
Pennsylvania	0.56	359	-35.59***	-0.27	2.47**	0.44	1.83***	1.52
			(2.78)	(0.17)	(1.25)	(0.49)	(0.10)	(1.47)
Ohio	0.68	256	22.20***	0.02	-13.89***	1.91***	1.65***	4.07***
			(3.16)	(0.17)	(2.49)	(0.50)	(0.13)	(1.47)
New Jersey	0.82	359	30.91***	-0.25	4.15***	1.60**	2.34***	5.43***
			(3.16)	(0.21)	(1.60)	(0.62)	(0.13)	(1.87)
Michigan	0.75	271	66.68***	-0.71**	3.22	2.04**	1.77***	6.40***
			(5.20)	(0.29)	(4.00)	(0.83)	(0.19)	(2.43)
Massachusetts	0.72	279	28.20***	-0.35*	-3.66	1.36**	1.69***	3.04*
			(3.22)	(0.18)	(2.46)	(0.53)	(0.12)	(1.58)
North Carolina	0.60	245	38.11***	-1.34***	8.80***	-0.04	0.22	3.52**
			(3.26)	(0.17)	(2.66)	(0.53)	(0.15)	(1.56)
Virginia	0.71	294	33.54***	-0.68***	3.57***	-0.22	-0.53***	-0.81
			(1.82)	(0.10)	(0.71)	(0.29)	(0.07)	(0.83)
Wisconsin	0.39	172	30.20***	0.61*	-2.21	1.33*	0.60**	-0.21
			(6.89)	(0.32)	(2.92)	(0.68)	(0.25)	(1.53)
Maryland	0.59	359	27.31***	-1.24***	-0.22	0.36	0.85***	2.84*
			(2.48)	(0.17)	(1.25)	(0.49)	(0.10)	(1.46)
Minnesota	0.65	129	53.44***	-0.79***	34.49***	0.03	0.18	7.45***
			(12.85)	(0.28)	(4.26)	(0.04)	(0.21)	(0.86)
Connecticut	0.27	206	2.57	0.62*	-4.84	1.40**	1.38***	1.06
			(5.18)	(0.34)	(3.06)	(0.70)	(0.21)	(1.66)
Delaware	0.38	191	13.16***	0.57***	-3.53*	1.43***	0.42***	1.07
			(3.10)	(0.18)	(1.83)	(0.39)	(0.12)	(0.91)
Nevada	0.76	270	43.65***	-0.48**	3.50	0.91	1.87***	4.14**
			(3.87)	(0.22)	(3.22)	(0.65)	(0.15)	(1.90)
Rhodes Island	0.54	141	57.65***	-0.01	-4.60	0.44	0.68**	0.26
			(7.58)	(0.36)	(3.11)	(0.77)	(0.28)	(1.52)

Table 3.11. Results for the contemporaneous model, using weekly observations. For each state, the coefficient of determination (R^2) is provided in the first column, followed by the number of observations in the second column. Estimated coefficients are reported in the subsequent columns, with heteroskedasticity and autocovariance consistent (HAC) standard errors (Newey & West, 1987) in brackets. *,** and *** indicate significance of the coefficients at the 10%, 5% and 1% level, respectively.

the potential to reveal the relationship between explanatory variables and the dependent variable that have been overlooked by standard regression models. Further, a conditional median regression is more robust than the conditional mean regression in terms of outliers in the observations (Yu & Moyeed, 2001).

In the following, we test the predictive relationship between the predictive variables and state CDS spreads at different quantiles. Naturally, we are specifically

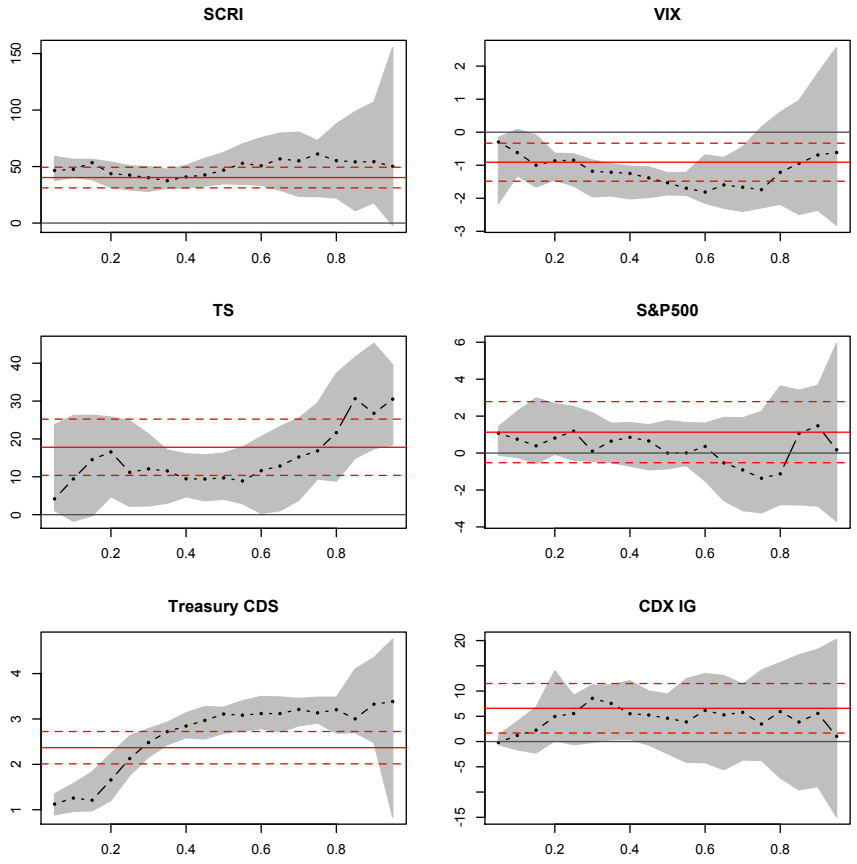
interested in the results for the SCRIs for various states. Results from quantile regressions can reveal possible changes in the predictive relationship for different ranges of the distribution for CDS spreads. Thus, they allow us to draw inferences at different levels of sovereign risk, for example, during crisis periods or periods where market participants had a rather low perception of sovereign risk. Results for the upper quantiles of the dependent variable are of particular interest, since higher values of market CDS spreads are indicative of a higher sovereign risk at the state level.

Figure 3.6 provides an overview of the results for the state of California and for the six predictive variables, namely the state-specific SCRI, the VIX, the term spread (TS), S&P500 returns, Treasury CDS spreads, and CDX IG. Estimated values of the coefficient for each quantile are represented by the black dotted line. The horizontal solid line represents the OLS estimate, while the two dashed lines represent the 90% confidence intervals for the least squares estimate from the previous regression model (3.3). The shaded grey area depicts a 90% confidence band for the quantile regression estimates.

As illustrated by the figure, the value of the estimated coefficient for SCRI in California is very steady around 50 in all quantiles, as indicated by the dotted line. The estimated coefficients for the conducted quantile regressions are also very close to the 90% confidence interval of the mean estimation of the coefficients, as marked by the two dashed lines. Thus, for California, our results suggest that the estimated coefficient for the SCRI is not only positive and significant at its mean value, but also at various quantiles of the distribution. Therefore, these results confirm our previous conclusions about the strong and significant predictive relationship between the derived SCRI and state CDS spreads, with little evidence of quantile effects.

However, quantile regression results for the other five explanatory variables show different variation patterns. For example, results for TS and TCDS show a clear upward trend from the lower quantiles to the higher quantiles, indicating a higher influence of these variables on sovereign CDS spreads in high distress risk states, while such behavior could not be observed for the estimated coefficients for the derived SCRI. This could imply that the high levels of sovereign risk are more likely to be able to be inferred from market conditions, while the influence of predictive signals from the private sector do not vary across different quantiles of the distribution. This is a possibility, considering that the performance in the

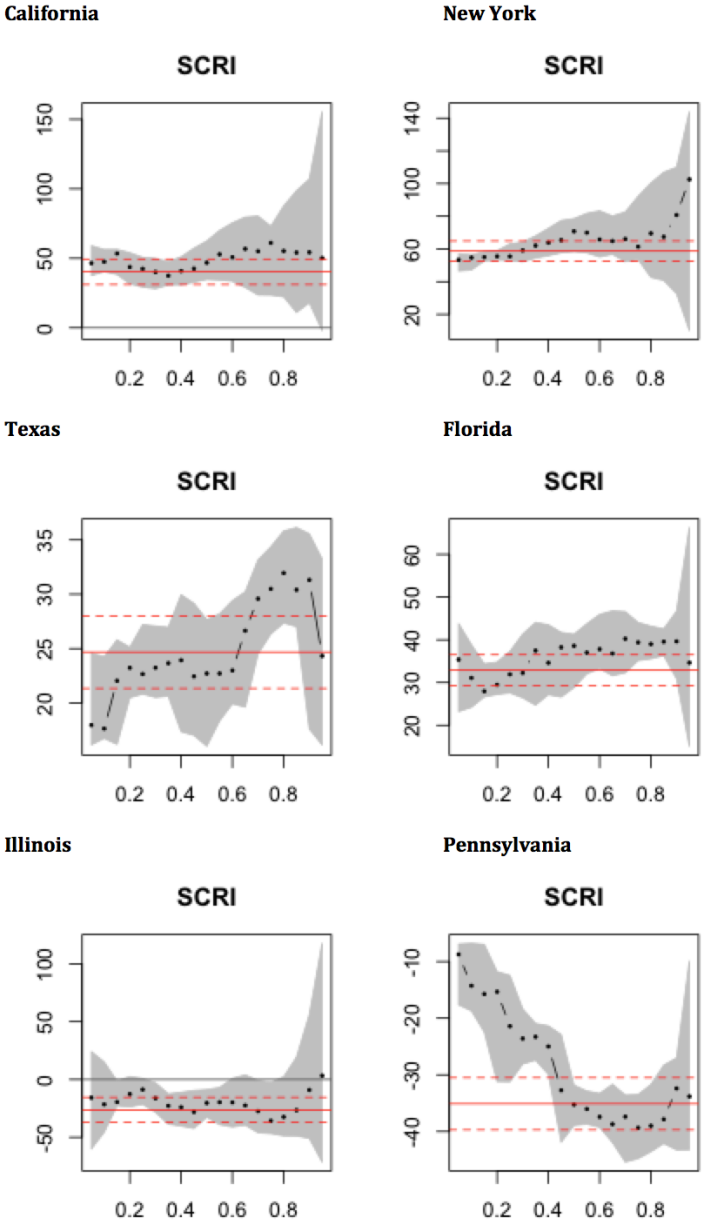
Figure 3.6. Results for conducted quantile regression based on data of California. Estimated coefficients for different quantiles are represented by the black dotted line. Each plot provides results for quantiles ranging from 0 to 1, while the vertical axis indicates the value of the estimated coefficient. The solid line in each graph shows the ordinary least squares estimate and the two dashed lines represent the 90% confidence intervals for the estimated coefficient using OLS regression. The shaded grey area depicts a 90% confidence band for the quantile regression estimates.



private sector aids in predicting the state government's intrinsic ability to service its debt payments, while the CDS spread is a market-based variable that is more likely to be influenced by overall market conditions.

Quantile regression results for the estimated coefficients of SCRI of six chosen states are presented in Figure 3.7. Clearly, the behaviour of estimated coefficients for SCRI at different quantiles of the CDS spreads varies from state to state. Typically the quantile regression estimates for the SCRI coefficients are consistent with OLS estimates, in particular for quantiles ranging from 0.4 to 0.6. Thus, while OLS and quantile regression coefficients are not identical, the difference between the estimates is often not significant for many of the considered quantiles, even when the confidence interval for the OLS coefficient is relatively narrow. However, for some states, including, e.g., New York and Pennsylvania, estimated coefficients

Figure 3.7. Quantile regression results based on data of California *(upper left panel)*, New York *(upper right panel)*, Texas *(middle left panel)*, Florida *(middle right panel)*, Illinois *(lower left panel)* and Pennsylvania *(lower right panel)*. We present quantile regression estimates (the black dotted line) for quantiles ranging from 0 to 1, while the vertical axis indicates the value of the estimated SCRI coefficient at different quantiles. The solid line in each graph shows the ordinary least squares estimate and the two dashed lines represent 90% confidence intervals of the OLS estimate. The shaded grey area depicts a 90% confidence band for the quantile regression estimates.



based on quantile regression significantly deviate from the OLS estimate of the coefficient. As illustrated in Figure 3.7 this is true in particular for very low and high quantiles.

The significant deviation from the mean estimate in the coefficients of SCRIs for higher quantiles indicates a possible change in the predictive relationship between SCRI and state CDS spreads. The coefficients are still significant in forecasting sovereign risk, however the expected change in a state's CDS spread caused by one unit variation in the SCRI depends on the level of CDS spreads. In particular, the size and the direction of the forecast changes differ when the CDS spread is very high, such as during a credit crisis. When the quantile regression estimate is outside the confidence band of the OLS estimate, the predictive relationship needs to be estimated with reference to the quantitle of the CDS spreads. In these cases, the influence of market-based variables such as VIX and CDX IG should be given special attention.

Overall, our results suggest that the coefficients for the SCRIs based on quantile regression are typically consistent with values from the OLS regression for most states. However, some states exhibit substantial quantile effects, in particular at very high or low quantiles of the distribution. Nevertheless, the coefficient for the SCRI is typically still positive and significant even in these quantiles. Overall the results do suggest further investigation of the predictive relationship between the derived SCRIs and state CDS spreads under extreme economic scenarios such as economic crises, using also non-linear models. We leave this task to future research.

3.5 Conclusion

In this paper we develop a new approach for assessing sovereign risk at the state level, using bottom-up credit risk indicators based on information about default risk at the company level. Our study is motivated by the simple rationale that the ability of state governments to service debt is affected by tax revenues from the private sector, the latter being dependent on the attendant economic activity and the performance of major industries in a state. The recent defaults of large municipalities such as Detroit and the U.S. territory of Puerto Rico also encourage

us to examine more thoroughly the dynamics and prediction of sovereign debt at the state level.

Using Moody's KMV EDF data to measure corporate default risk, we construct industry credit risk measures that are then used to derive state-specific indicators for default risk based on the industrial composition of a state. In combination with additional predictive variables, namely the VIX, the spread between short-term government bills and long-term government bonds, returns from the S&P500, U.S. Treasury CDS spreads, and investment grade corporate bond index returns, the derived credit risk indicators are then examined with respect to their ability to forecast U.S. state CDS spreads.

Our study complements and extends earlier work by Altman & Rijken (2011*b*) in several ways. First, in contrast to the reliance on scoring models in Altman & Rijken (2011*b*), our approach uses EDFs that are based on a structural model for quantifying credit risk at the company level. Market-based EDF measures are available at a daily frequency for a large universe of private companies. Thus, our approach overcomes many of the shortcomings of scoring models that are primarily reliant on accounting information. Second, this study examines a sample of sovereign U.S. state governments that are selected without reference to their financial health, thus extending the findings of Altman & Rijken (2011*b*) who focus solely on distressed European sovereigns.

Our results show that market-based measures of private sector credit risk are strongly associated with subsequent shifts in sovereign credit risk premiums, measured by CDS spreads. Our findings also suggest that SCRIs have higher predictive value than previously considered financial variables for forecasting sovereign CDS spreads at weekly and monthly sampling frequencies. These findings suggest a strong predictive link between market expectations of private sector credit quality and those of sovereign credit quality - a connection that is not directly discernible from scoring models. Moreover, we find that the link between private and public sector credit risk generalizes beyond the sample of distressed European sovereign entities studied by Altman & Rijken (2011*b*).

Our findings suggest that, at the very least, private sector based metrics complement market-based measures of macroeconomic expectations in forecasting sovereign risk. A closer look at company level information is also helpful for investors in making informed decisions. As our study suggests, fluctuations in

credit quality of resident corporations appear to be strongly linked to subsequent variation in sovereign credit quality. Therefore, based on our findings we strongly recommend additional research on the relationship between credit risk at the corporate level and sovereign default risk. For example, the credit risk on the company levels can be measured using different methods and focusing on particular aspects that are important to the assessment of sovereign default risk, such as the company's ability to pay tax. These credit risk measures are be aggregated with assigned weights which reflect the relative impact on the sovereign default risk. The results from these further studies can further improve the understanding and prediction of the sovereign default risk.

Chapter 4

A joint model for longitudinal and time-to-event data in corporate default risk modelling

Feng Liu (contribution 80%), David Pitt (contribution 10%), Stefan Trück (contribution 10%)

4.1 Introduction

Corporate default risk refers to the probability that a company fails to fulfil its debt obligations. Inability to pay debt is very often followed by insolvency events such as liquidation or bankruptcy. The need to manage corporate default risk arises due to the concern that such default events will lead to a loss for investors and creditors, and also for policyholders or depositors in the case of a financial company that declares bankruptcy. Default risk analysis is also useful because it can be used to calculate theoretical prices for corporate bonds, such as in Longstaff & Schwartz (1995).

There is considerable literature on assessing corporate default risk. The ability to service debt payments can be influenced by a variety of factors, such as general economic conditions (macroeconomic variables) and the financial status of the company (company-specific variables). For example, a high asset value or market value relative to interest payments on debt may indicate that the company is

capable of meeting its debt obligations. However, during an economic downturn with a low rate of GDP growth, a company's default probability may increase due to a fall in free cash flows or a drop in asset values. Other variables that are sometimes associated with a company's default probability include the Treasury Bill rate, the company's stock return, and accounting ratios such as working capital/total assets (for more on predictors of default, see for example Altman (1968); Zmijewski (1984); Shumway (2001); Duffie *et al.* (2007)).

As the corporate default risk varies with a number of factors, it is of interest to explore the relationship between this risk and independent variables. The results can then be used for forecasting future default events. Some models have been developed to calculate the exact default probability, while others aim to develop a classification system using discriminant analysis. An example of the latter is the Z-score model, which computes company credit scores as a linear combination of accounting ratios (Altman, 1968). The scores are then used to provide a basis for classifying companies as likely or unlikely to default over a given future specified time horizon. Structural models for default, first put forward by Merton (1974), calculate the probability that a company's asset value falls below the debt value as a measure of default risk, assuming that the asset value follows a stochastic diffusion process. An application of a structural model to measure default risk is the company expected default frequencies (EDF) publised by Moody's KMV. This is based on the distance to default (DD) defined as the number of standard deviations between the asset value and the default threshold.

More recently, hazard models have been used to estimate the default probability by modelling the hazard rate for default. Let T be a continuous random variable that denotes the company's future survival time, and $S(t) = P(T > t)$ be its survival function. The cumulative distribution function for survival time T can be denoted as $F(t) = P(T \leq t) = 1 - S(t)$. The probability density function for T is denoted $f(t)$. The hazard rate of default, $h(t)$, is the instantaneous default rate in the time interval $[t, t + dt]$, and $h(t) = f(t)/S(t)$ (Allison, 2010). A hazard model is used in Shumway (2001) to estimate corporate default events with the accounting ratios used in Altman (1968), and the model produces one-year default probability predictions that are more accurate than the existing scoring models discussed in the chapter. The one-period prediction model is extended by Duffie *et al.* (2007) and Duan *et al.* (2012) to allow for multi-period predictions.

The main problem with existing hazard models is the low frequency of the

measurement of the independent variables. The hazard function is continuous and smooth, while the independent variable values are only available periodically. For example, accounting reports are only available quarterly or sometimes half-yearly or even only once per year. There is a mismatch issue if we use a step function based on available covariate data to predict a continuous function of the hazard rate. Other than interpolation, the popular way to get around this issue is the so-called "Last Value Carried Forward (LVCF)" approach. The missing value is substituted by the most recent observation, and the covariate is assumed to be constant between observations, such as in Shumway (2001). Since the true values of an observed variable are never constant between observations, the assumption brings bias to the modelling results, and the corresponding relationship between the covariate and the probability of default will potentially be inaccurate and misleading.

In order to mitigate this problem, the joint model for longitudinal and time-to-event data (Rizopoulos, 2012) is investigated here as an extended hazard model that can also recognise the behaviour of the covariate between observations in a mixed-effects submodel. While reducing potential bias, the model can also be used for forecasting based on the well-modelled trajectory of the covariate. The joint model is a combination of two submodels: a mixed-effects model for analysing the longitudinal observations of independent variables, and a Cox model (Cox & Oakes, 1984) to assess the hazard rate of default and then the default probabilities. In a joint model, before inputting the observed values of the covariate, the covariate is first modelled using a mixed-effects model. This gives a model for the value of the covariate as a continuous function of time. The mixed-effects model describes the average longitudinal evolution in the population over all subjects, using fixed effects parameters, and each individual subject can have its own trajectory that deviates from the population mean as described by subject-specific random effects parameters (Rizopoulos, 2012). An example is presented in Figure 4.1. The dots represent the observations of a certain covariate for two companies, and the solid lines denote the modelling results from a linear mixed-effects (LME) model assuming a simple regression specification. Step functions are denoted by the dashed lines, which represent the values to be input in a hazard model with the LVCF assumption. The LME model allows for subject-specific intercept and slope coefficients, thus the two distinctive trajectories for the two companies. For more complicated trajectories, a spline function can be used in place of the simple linear function shown here.

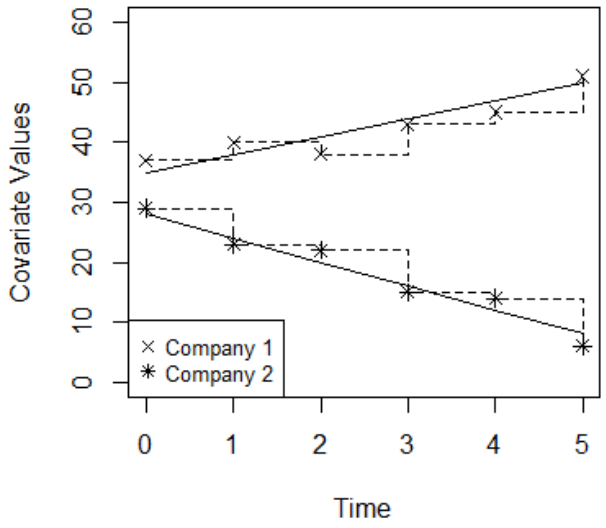


Figure 4.1. A representation of LME model results for two companies.

In this way, the trajectories of the covariates can be modelled and predicted, and the values of the covariates are no longer assumed to be constant between observations. The fitted values for our covariates from the mixed-effects model become the inputs to a Cox model (Cox *et al.*, 1972) to assess the association between the covariates and the corporate default event and predict rates of corporate default.

The joint model has been applied in many areas such as medical research and biomedical studies. It is particularly relevant to many cancer clinical trials where a patient's biomarkers, such as the blood pressure and cell counts, are observed and recorded repeatedly (Ibrahim *et al.*, 2010). Henderson *et al.* (2000) illustrate the use of the joint model with a clinical study to treat schizophrenia. Xu & Zeger (2001) uses the joint model to analyse clinical trial data comparing risperidone with a placebo for the treatment of schizophrenia. Elashoff *et al.* (2008) applies the joint model in a scleroderma lung study to evaluate the effect of oral CYC on the risk of treatment failure or death. The joint model has also been extended in various ways, such as allowing for multiple failure types, using an accelerated failure time (AFT) model to replace the Cox model (Tseng *et al.*, 2005; Elashoff *et al.*, 2008; Rizopoulos & Ghosh, 2011).

We use the joint model to analyse corporate default risk. The study focuses on

the default probabilities of U.S. listed companies over a 20-year time period. We calculate the distance to default of these companies using a structural model. The joint model is calibrated to the dataset to assess the association between the corporate default risk and the company's distance to default and time since being listed (age of the company). The relationship is used to predict default events in an out-of-sample study.

The results of the study show that the joint model can produce more accurate out-of-sample predictions of the default risk, compared to selected traditional survival models such as the Cox and Weibull model (Hosmer *et al.*, 2011). The results are consistent when predicting over different time horizons, and the superior performance of the joint model is more pronounced for the longer duration predictions.

The chapter is organised as follows. Section 4.2 gives a high level summary of existing methods to analyse default risk, including scoring models, structural models and hazard models. Section 4.3 describes the specifications of the joint model. Section 4.4 introduces the default risk data and predictor variables used in the joint model analysis. Section 4.5 presents the modelling results and a discussion of findings. Section 4.6 concludes the chapter.

4.2 Literature Review

Scoring models use a linear combination of independent covariates to compute a credit score, which is linked to the default probability. Structural models compare the asset value of a company to a predetermined default threshold, and calculate the probability that the asset value drops below the threshold, triggering a default event. Scoring models and structural models both focus on the state of the company at a particular time point to evaluate the company's default risk, and are classified as static models.

Hazard models explicitly recognise that a company's default risk changes through time. These models can use time-dependent covariates to evaluate the hazard rate, and can incorporate macroeconomic variables that are the same for all firms in addition to company-specific variables (Shumway, 2001). In addition to corporate default analysis, hazard models, such as the Cox model proposed in Cox *et al.*

(1972), have a wide application in survival analysis in many other academic areas as well and a review of these applications can be found in Nikulin & Wu (2016).

4.2.1 Scoring Models

A credit score for a particular company is calculated as a linear combination of the company's accounting ratios in a regression model. The score is designed to discriminate between default and non-default companies (Trueck & Rachev, 2009). Altman (1968) is generally recognised as the first to apply a scoring model in corporate default risk analysis to classify corporate borrowers. To calculate the so-called "Z-score" for a particular company, accounting ratios are used, specifically working capital/total assets, retained earnings/total assets, earnings before interest and taxes/total assets, market value of equity/book value of total liabilities, and sales/total assets. The Z-score is then found using

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_5 X_5, \tag{4.1}$$

where X_k $(k = 1, 2 \ldots 5)$ are the five variables used in the model (Altman, 1968). Based on empirical data of two groups of companies, which are the failed and solvent companies, the coefficients β_k $(k = 1, 2 \ldots 5)$ are estimated by maximising the between-group variance relative to within group variance in discriminant analysis.

A company's credit score can be used to flag high risk companies. This is also very useful for debt pricing or loan pricing. The credit scores can also be transformed to probabilities of default. The two most popular approaches are logit and probit transformation, both having the same general form (Trueck & Rachev, 2009)

$$Y = f(\beta_0 + \sum_{k=1}^{n} \beta_k X_k), \tag{4.2}$$

where Y indicates the probabilities of default, and f denotes the logit/probit transformation function applied to the credit score. Maximum likelihood estimation has been widely used to determine parameter estimates in these models. The goodness of fit and inferential statistics associated with the model are commonly based on log likelihood and chi-square test statistics.

The logit model is based on the assumption that the default probability is logistically distributed. The technique is discussed in detail with a number of possible applications illustrated by Green (1993) and Hosmer Jr *et al.* (2013). Other logit models include the O-scores developed in Ohlson (1980), and the technique is also applied in Zavgren (1985) and Engelmann *et al.* (2003).The probit model assumes that the probability of default is normally distributed, and some examples of its applications in default analyses are Zmijewski (1984), Platt & Platt (1991) and Amato & Furfine (2004).

Based on different performance measures, various scoring models are likely to rank differently. For the users of these scoring models, it is important to be clear about which aspect of the company performance is more important and then to set up criteria to select the most suitable model.

Although easy to understand and use, the scoring models have two major drawbacks. As pointed out by Ohlson (1980), the linear regression model used to calculate the credit score is restricted by the specific statistical requirements imposed on the distributional properties of the predictors. The scoring models also have the disadvantage of relying exclusively on accounting ratios, and the inadequacy of the simple statistical model in explaining potentially complicated associations with the default risk. Recent studies have been conducted to compare the performance of the simpler scoring models with more sophisticated statistical models such as the hazard models. These studies often conclude that the scoring models are insufficient and underperforming (Shumway, 2001). This may be due to the limitations of relying solely on accounting information and the fact that the model is static.

4.2.2 Structural Models

The original structural model to assess default risk was developed by Merton (1974). Structural models focus on the capital structure of the company and treat the default event as an outcome of the deterioration of the firm's value. The approach is sometimes referred to as the firm value approach, when the default event is triggered when the value of the firm falls below some threshold.

In a structural model, the value of the firm is assumed to follow a stochastic diffusion process with constant volatility. The only possible time for a default

event is at the maturity of the company debt. The distance to default at maturity is computed based on the value of the firm and the value of its debt, and defined as the number of standard deviations by which the asset value exceeds the company debt. A high distance to default is associated with a high asset value relative to debt obligations, indicating a high chance for the company to be able to repay its debt. A low or even negative distance to default raises concern and is likely to be associated with high probability of default, as the asset value may prove insufficient to cover debt payments.

The model has many assumptions, including a simple capital structure with only one debt, a constant risk free rate of interest and normally distributed asset returns. The probability of default, PD_t, is estimated as follows

$$PD_t = N(-\frac{log(V_t) - log(X) + (\mu - \sigma_V^2/2)(T-t)}{\sigma_V\sqrt{T-t}}), \tag{4.3}$$

where

$N(.)$ = the cumulative Gaussian distribution function,
V_t = the value of the firm at time t,
X = the default threshold,
σ_V = the asset volatility of the firm,
μ = the expected return on the assets of the firm
$T-t$ = prediction time horizon.

Later studies improved and modified the original structural model in various ways. Additional factors have been included in the model such as taxes, bankruptcy costs and protective covenants, and the original factors are modelled and computed in more sophisticated ways. The model has also been extended to predict the probability of default over more than one period. Some examples are Black & Cox (1976), Leland (1994), Longstaff & Schwartz (1995) and Zhou (1997).

One of the important applications of the original structural model is the development of the corporate Expected Default Frequency (EDF) published by Moody's KMV as a measure of default probabilities (Crosby, 1998). Based on empirical default frequency, EDF is calibrated to match historical default frequency based on the distance to default computed from the structural model, instead of directly obtaining the probability assuming the asset returns follow a

normal distribution. Moody's KMV has also introduced more advanced methods to compute the values of the factors in the structural model, such as an empirical procedure to estimate the value of the asset and its volatility, instead of relying on the balance sheet or historical records (Nazeran & Dwyer, 2015).

Compared to the accounting ratios used in scoring models, structural models take advantage of the forward-looking feature of market data as the market value of the company is used in the model. As market data is expected to reflect a wide range of relevant information relating to the firm, it should provide better predictor variables than those available from the balance sheet (Jovan & Ahčan, 2017).

One major drawback of structural models results from the many assumptions made by the original model as mentioned previously. Unfortunately many of these assumptions are likely violated in reality, such as the simplified assumptions about the capital structure of the firm and the normality of the asset return distribution (Crouhy *et al.*, 2000). Even if some or all of the assumptions can be relaxed in subsequent models, such as allowing for a more complicated capital structure with more than one debt or a less restrictive assumption of the distribution of the asset return, it is still difficult to assign values to some of the variables in the equation, such as the true value of the asset and the assumed constant volatility parameter.

Although the structural model was later shown to be inappropriate for predictions, Merton (1974) was successful in finding a way to estimate the probability of default implied by equity prices. The default probability is computed based on relevant market variables, which are generally considered to be forward-looking. In theory, this model is able to generate a marked-to-market assessment of the creditworthiness of the company, especially listed companies. As a result, the distance to default calculated from the model has become a popular covariate in the area of assessing default risk, to account for the impact of the company's capital structure on the probability of default.

4.2.3 Hazard Models

Hazard models differ from structural models in that the default is a surprise event, instead of an output from a process when the asset value approaches or crosses the default threshold. The key variable in these models is the time to default, governed by the hazard rate of default (Sundaresan, 2000; Duffie & Singleton,

2003). Commonly a Poisson process is assumed for the occurrence of the defaults in the literature as well as in industry practice, such as in the CreditRisk developed by Credit Suisse (Suisse, 1997).

Let T_i denote the observed event time for subject i. We define $T_i = \min(T_i^*, C_i)$ so that T_i is the minimum of the default time T_i^* and the censoring time C_i. The event indicator $\delta_i = I(T_i^* \leq C_i)$ takes the value 1 if the company defaults and 0 otherwise, where $I(\cdot)$ denotes the indicator function. If $p(\cdot)$ denotes the event time probability density function, the survival function is defined as

$$\mathcal{S}(t) = Pr(T^* > t) = \int_0^\infty p(s)ds = \exp\{-\int_0^t h(s)ds\}. \tag{4.4}$$

The hazard function $h(\cdot)$ can be used to describe the instantaneous risk for an event in the time interval $[t, t + dt]$ conditional on survival up to t. Let $y_i(t)$ denote the longitudinal covariate for subject i $(i = 1, 2, ..., n)$ observed at time t. If we assume that the covariates have a multiplicative effect on the hazard for a default event, a proportional hazards model can be postulated in the form of (Rizopoulos, 2012, Chapter 3, Chapter 5)

$$\begin{aligned} h_i(t|\mathcal{Y}_i(t), \omega_i) &= \lim_{dt \to 0} \frac{Pr(t \leq T_i^* < t + dt | T_i^* \geq t, \mathcal{Y}_i(t), \omega_i)}{dt} \\ &= h_0(t) \exp\{\gamma^\top \omega_i + \alpha y_i(t)\}, t > 0. \end{aligned} \tag{4.5}$$

In equation (4.5), $\mathcal{Y}_i(t)$ denotes the history of the longitudinal process up to time point t, $h_0(\cdot)$ denotes the baseline risk function, ω_i is a vector of baseline hazard covariates with a corresponding vector of regression coefficients γ and $y_i(t)$ is the value of the longitudinal measure for subject i at time t. The parameter α quantifies the effect of the underlying longitudinal outcome of subject i to the risk for a default event.

As discussed in Shumway (2001), the advantages of the hazard model over static models such as scoring models and structural models, are the automatic adjustment for period at risk, incorporation of time-varying covariates and the ability to produce more efficient out-of-sample forecasts. The hazard model is also suitable for incorporating censored observations. Since the focus of the model is the time until default or censoring over multiple periods, the status of each

individual can be observed and can contribute to the inference whether it survives or defaults. On the other hand, the structural and scoring models are considered to be less efficient in that not all information is used in the development of the association, when the model is calibrated solely using time of default data. The survival information which is equally important over time is not accounted for in the static models.

Hazard models have been applied in default risk analysis, and many models have been developed over the last twenty years. They have become a standard approach in the literature to assess and estimate corporate default risk (Orth, 2012). Shumway (2001) proposed the idea of a discrete-time hazard model in forecasting default probabilities based on data from 1962 to 1992. In the proportional hazards model specified in (4.5), the hazard rate is set to depend on some chosen covariates (in particular the accounting ratios as used in previous scoring models such as in Altman (1968)) and a logit estimation program is used to calculate maximum likelihood estimates. The fitted model is then applied to predict the one-period default rate, where the covariates are modelled in a Gaussian first-order vector autoregressive time series model. The study concludes that half of the previously used accounting ratios prove to be poor predictors, while market-driven variables are strongly related to bankruptcy probabilities. Chava & Jarrow (2004) extends the study to include financial companies and analyse the industry effect. They reach similar conclusions to Shumway (2001), namely that the hazard model outperforms existing accounting based models.

One other influential study is Duffie *et al.* (2007). This paper aims to forecast the default risk over multiple periods. The research is based on a dataset with more than 390,000 firm-months spanning 1980 to 2004. A proportional hazards model is used with a constant baseline hazard rate, and generates maximum likelihood estimates of term structures of conditional probabilities of corporate default, incorporating the dynamics of firm-specific and macroeconomic covariates over multiple time periods. Four covariates are modelled with Gaussian panel vector autoregressions. The method allows for the combination of traditional duration analysis of the dependence of event intensities on time-varying covariates with conventional time-series analysis of covariates, in order to obtain the maximum likelihood estimation of multi-period survival probabilities. The term structure of default hazard rates of individual firms depends significantly on the current state of economy and especially on the current leverage of the firm.

One potential drawback of Duffie *et al.* (2007) lies in the modelling of the covariates. Although only four variables are measured using a time series econometric model, the estimation of the model coefficients as well as the prediction process can be time-consuming when a large number of companies are analysed (Duan *et al.*, 2012). To address the issue, Campbell *et al.* (2008) uses a dynamic logit model to measure directly the association between the lagged covariates and the default event. The relationship is then used to forecast over multiple periods based on current values of the covariates. Duan *et al.* (2012) follow the same principle. By replacing the logit model with a proportional hazards model, they propose a forward intensity model for the prediction of corporate defaults over different future periods, to avoid the difficulty to specify and account for the dynamic nature of the covariates.

Other relevant applications of hazard models include Hillegeist *et al.* (2004), Beaver *et al.* (2005), Hwang (2012) and Orth (2013). The modelling techniques are similar, however they differ in the covariates employed to analyse the effects of different factors. The superior performance of the hazard model over other conventional models is evident in their studies. Moreover, the distance to default calculated from a structural model is found to be influential and significant. Although the structural model has drawbacks as mentioned previously, the distance to default proves to be an important covariate in assessing default risk.

4.3 Joint Model Specification

The purpose of the joint model is to measure the association between longitudinal observations of a subject and the subject's risk for an event. It augments the simple hazard model, in that the longitudinal covariates are analysed firstly in an LME model to assess the subject-specific time evolutions. This enables the joint model to take advantage of the fully-specified subject-specific longitudinal trajectories to help evaluate event time and risk. The joint model is well explained by Rizopoulos (2012). The author of the book is also responsible for the development of the R packages "JM" and "JMbayes", which are used to fit the joint model and conduct analysis in this research (Rizopoulos, 2010, 2016*b*).

4.3.1 Linear Mixed-Effects Model

A Linear Mixed-Effects model (LME) is used to model the trajectories of response variables taken on different subjects. It incorporates both fixed effects and random effects. The fixed effects are parameters associated with the entire population, while the random effects are associated with the individual behaviours of different subjects in the population.

In order to model the covariate longitudinal trajectory, let previously defined $y_i(t)$ denote the response of subject i, $i = 1, \ldots, n$, at time t. $y_i(t)$ can be expressed as a function of time t, with subject-specific coefficients. If subject i has n_i observations, let y_i denote the $n_i \times 1$ vector of longitudinal measures. Define t_i as a design matrix of time, then y_i can be expressed as the product of the design matrix and subject-specific coefficient vector β_i

$$y_i = t_i * \beta_i + \varepsilon_i. \tag{4.6}$$

The error terms ε_i are assumed to be normally distributed with mean zero and variance σ^2. The coefficient vector β_i in the formula above can be written as the sum of two new coefficient vectors β and b_i, where β describes the mean longitudinal evolution in the population averaged over all subjects and b_i is subject-specific.

In other words, the average evolution over all subjects is described with β, and different subjects can deviate from the population mean trajectory in their own manner, captured in b_i. Accordingly, parameters in β are called fixed effects, while parameters in b_i are called random effects, with a multivariate distribution with mean zero and covariance matrix D. Equation (4.6) can be reformulated into its general form, where $X_i(t)$ and $Z_i(t)$ are the design matrices for the fixed and random effects coefficients

$$\begin{cases} y_i = X_i(t)\beta + Z_i(t)b_i + \varepsilon_i, \\ b_i \sim \mathcal{N}(0, \mathcal{D}), \\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_{n_i}). \end{cases} \tag{4.7}$$

The LME model overcomes difficulties in analysing longitudinal data with standard statistical models. Repeated measures taken on the same subject are

expected to be correlated, meaning common statistical tools assuming independent observations are not appropriate (Rizopoulos, 2010). More importantly, longitudinal measures are often collected intermittently and with error at a set of a few time points for each subject. The so-called "Last Value Carried Forward (LVCF)" approach as discussed in Section 4.1, which assumes the covariate stays constant between observations, can result in significantly biased estimates for parameters and standard errors (Prentice, 1982). Finally, under the LVCF assumption, the covariate values follow a step function over time. The step function does not represent the evolution of the covariate, and is not appropriate to explain the continuous and smooth function of the hazard rate.

An LME model is more suitable in that it offers flexibility in modelling the trajectories by elaborating the specification of the time structure in $X_i(t)$ and $Z_i(t)$, expressed in terms of polynomials or splines. Moreover, the use of subject-specific random effects allows us to reconstruct the complete path of the true, unobserved value of the longitudinal outcomes, represented here with the term $m_i(t)$. Unlike $y_i(t)$ with a step function, $m_i(t)$ can be denoted with a continuous smooth function. This also solves the problem of the imbalance in the data, where different measurement frequencies may occur for different subjects or the measurements may be taken at different sets of time points (Rizopoulos, 2010). An example of $y_i(t)$ and $m_i(t)$ is in Figure 4.1, where $y_i(t)$ refers to the observed values (dots) and $m_i(t)$ is denoted by the straight lines fitted through the dots.

With the introduction of $m_i(t)$, equation (4.7) can be written as (Rizopoulos, 2012, Chapter 4)

$$\begin{cases} y_i = m_i(t) + \varepsilon_i, \\ m_i(t) = X_i(t)\beta + Z_i(t)b_i \\ b_i \sim \mathcal{N}(0, \mathcal{D}), \\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_{n_i}). \end{cases} \tag{4.8}$$

4.3.1.1 Parameter Estimation

The LME model makes it possible to estimate not only the fixed effects parameters that describe the mean response changes in the population, but also allows to measure how individual response trajectories change over time. The parameter

estimation is often based on maximum likelihood, where the marginal density of the observed outcome for the ith subject is

$$p(y_i) = \int p(y_i|b_i)p(b_i)db_i.$$

Given that the distribution of the random effects b_i is assumed to be normal, the above integral has a closed-form solution. This results in an n_i-dimensional normal distribution with mean $X_i\beta$ and variance-covariance matrix $V_i = Z_i\mathcal{D}Z_i^\top + \sigma^2 I_{n_i}$. Assuming independence across subjects, the log-likelihood of the LME model is (Rizopoulos, 2012, Chapter 2)

$$l(\theta) = \sum_{i=1}^{n} \log \int p(y_i|b_i;\beta,\sigma^2)p(b_i;\theta_b)db_i. \tag{4.9}$$

θ denotes the full parameter vector and can be decomposed into the subvectors $\theta^\top = (\beta^\top, \sigma^2, \theta_b^\top)$, and θ_b is the vectorisation of the matrix D written as $\theta_b = vech(\mathcal{D})$. If V_i is assumed to be known, it can be shown that the fixed effects vector β has the following form that corresponds to the generalised least squares estimator

$$\hat{\beta} = (\sum_{i=1}^{n} X_i^\top V_i^{-1} X_i)^{-1} \sum_{i=1}^{n} X_i^\top V_i^{-1} y_i. \tag{4.10}$$

If V_i is not known, the theory of restricted maximum likelihood (REML) estimation can be used, with the intuitive idea being to separate the part of the data used in the estimation of V_i from the part used for the estimation of β. The log-likelihood function to be maximised is slightly modified. Unlike the estimation of β, V_i does not have a closed form estimator and requires a numerical optimisation routine, such as the Expectation-Maximisation or the Newton-Raphson algorithm. For more details on the estimation of V_i, refer to Dempster *et al.* (1977), Laird & Ware (1982) and Lindstrom & Bates (1988).

Once both β and V_i are estimated, it can be shown that the variance of the generalised least squares estimator (4.10) is

$$\hat{var}(\hat{\beta}) = (\sum_{i=1}^{n} X_i^\top \hat{Q}_i X_i)^{-1}, \tag{4.11}$$

where $\hat{Q}_i = \hat{V}_i^{-1}$. The standard errors for the estimates of the unique parameters in V_i can be obtained from the inverse of the corresponding block of the Fisher information matrix

$$\hat{var}(\hat{\theta}_{b,\sigma}) = \left\{ E\left(-\sum_{i=1}^{n} \frac{\partial^2 l_i(\theta)}{\partial \theta_{b,\sigma}^{\top} \partial \theta_{b,\sigma}} |_{\theta_{b,\sigma}} = \hat{\theta}_{b,\sigma} \right) \right\}^{-1}.$$

4.3.2 Joint Model for Longitudinal and Time-to-Event Data

The joint model combines a standard proportional hazards model with an LME model. It aims to assess the risk for an event by evaluating the hazard rate, expressed as a function of the longitudinal observations. With the introduction of true values of the time-dependent covariate from the LME model, $m_i(t)$, the hazard rate function (4.5) can be updated to (Rizopoulos, 2012, Chapter 4)

$$\begin{aligned} h_i(t|\mathcal{M}_i(t), \omega_i) &= \lim_{dt \to 0} \frac{Pr(t \leq T^* < t + dt | T^* \geq t, \mathcal{M}_i(t), \omega_i)}{dt} \\ &= h_0(t) \exp\{\gamma^T w_i + \alpha m_i(t)\}, t > 0. \end{aligned} \tag{4.12}$$

As discussed before, the joint model is different from a standard survival model in that the time-dependent covariate $m_i(t)$ in the joint model is fully specified with a continuous smooth function. This solves the problem of using a standard proportional hazards model with the LVCF assumption, which postulates that the hazard for an event, at any time t, is associated with the extrapolated value of the covariate at the same time point.

The joint model solves this problem by associating the true longitudinal measures $m_i(t)$ with the risk of an event, when the covariate function is smooth and continuous. The modelling result is expected to be more accurate. We explore this model further in Section 4.4 and 4.5 in the context of the estimation of default probabilities.

4.3.2.1 Parameter Estimation

The main estimation method proposed for the joint model is semi-parametric maximum likelihood, and the estimates are derived as the modes of the log-likelihood function corresponding to the joint distribution of the observed outcome $\{T_i, \delta_i, y_i\}$. In this framework, the random effects account for both the association between the longitudinal and event outcomes, and the correlation between the repeated measurements in the longitudinal process. Under these assumptions, it can be shown that the log-likelihood contribution for the ith subject is given as (Rizopoulos, 2012)

$$\log p(T_i, \delta_i, y_i; \theta) = \log \int p(T_i, \delta_i, |b_i; \theta_t, \beta)[\prod_j p\{y_i(t_{ij})|b_i; \theta_y\}]p(b_i; \theta_b)db_i. \quad (4.13)$$

$\theta = (\theta_t^\top, \theta_y^\top, \theta_b^\top)^\top$ denotes the full parameter vector, with θ_t representing the parameters for the event outcome, θ_y the parameters for the longitudinal outcomes, θ_b the unique parameters of the random-effects covariance matrix, and y_i the vector of longitudinal responses of the ith subject. The joint density for the longitudinal responses together with the random effects is

$$\begin{aligned} p(y_i|b_i; \theta)p(b_i; \theta) &= \prod_j p\{y_i(t_{ij})|b_i; \theta_y\}p(b_i; \theta_b) \\ &= (2\pi\sigma^2)^{-n^2/2} \exp\{-||y_i - X_i\beta - Z_i b_i||^2/2\sigma^2\} \\ &\times (2\pi)^{-q_b/2} det(D)^{-1/2} \exp(-b_i^\top D^{-1} b_i/2), \end{aligned} \quad (4.14)$$

where q_b denotes the dimensionality of the random-effects vector, and $||x|| = \{\sum_i x_i^2\}^{1/2}$ denotes the *Edulidean* vector norm. The conditional density for the survival component, namely $p(T_i, \delta_i|b_i; \theta_t, \beta)$, takes the form

$$\begin{aligned} p(T_i, \delta_i|b_i; \theta_t, \beta) &= h_i(T_i|\mathcal{M}_i(T_i); \theta_t, \beta)^{\delta_i} \mathcal{S}_i(T_i|\mathcal{M}_i(T_i); \theta_t, \beta) \\ &= [h_0(T_i) \exp\{\gamma^\top \omega_i + \alpha m_i(T_i)\}]^{\delta_i} \\ &\times \exp(-\int_0^{T_i} h_0(s) \exp\{\gamma^\top \omega_i + \alpha m_i(s)\} ds). \end{aligned} \quad (4.15)$$

In a standard proportional hazards model, the distributional assumptions for T_i^* are hidden in the specification of the baseline hazard function. Cox *et al.* (1972) showed that estimation of γ and α based on the partial log-likelihood function does not require the specification of T_i^*. This is one of the advantages of the relative risk model, which means the standard errors and inference for the regression coefficients can be based on standard asymptotic distribution theory for maximum likelihood estimation. The maximum likelihood estimators do not depend on $h_0(t)$, so the baseline risk function can be specified as non-parametric.

Unlike a standard proportional hazards model, the baseline risk function $h_0(\cdot)$ is explicitly defined in the joint model, with a parametric yet flexible specification. Because of the use of random effects, the maximum likelihood estimator can not be derived using the asymptotic features and no closed-form solution can be found.

A feasible solution is to postulate a flexible parametric model for $h_0(t)$. Two commonly proposed options are to use cubic splines or a piecewise-constant model. In this way, various shapes $h_0(t)$ can be well captured by increasing the number of internal knots, and the estimation of standard errors directly follows from asymptotic maximum likelihood theory.

Following equation (4.14) and equation (4.15), the maximum of the log-likelihood function $l(\theta) = \sum_i \log p(T_i, \delta_i, y_i; \theta)$ with respect to θ can be found using standard algorithms such as the Expectation-Maximisation (EM). This studies follows the steps taken in Rizopoulos (2012) to apply the EM algorithm to generate the maximum likelihood estimates to previous equations. The process is composed of the Expectation (E) steps and Maximisation (M) steps. In the E-steps, the random effects are treated as "missing data". In order to find parameter values $\hat{\theta}$ that maximise the observed data log-likelihood $l(\theta) = \sum_i logp(T_i, \delta_i, y_i; \theta)$, the expected value of the complete data log-likelihood is maximised in the M-steps. For more details regarding the M-steps including the simulation techniques, refer to Rizopoulos (2012). Appendix B in Rizopoulos (2012) gives the specific formulation of the score vector and Hessian matrix.

4.4 Data

This chapter explores the use of the joint model to assess corporate default risk, by analysing the association structure between the default event and company-specific

covariates. A 20-year history of observations between 1997 and 2016 for all U.S. listed companies has been collected from from CRSP/Compustat database. Consistent with prior literature, companies analysed in this chapter are common firms (share code 10 and 11 in CRSP) and traded on NYSE, AMEX and Nasdaq (exchange codes 1 to 3).

A total of 12,698 companies were identified during this period, of which 788 defaulted and 3,592 were still listed at the end of the sample period in 2016. The average number of companies under observation per year is 7,110. A default event for a listed company is defined in this study as being delisted from the three stock exchanges as a result of declaring insolvency, bankruptcy or being liquidated. The corresponding delisting codes are those between 400 and 500 for liquidation, or equal to 572 or 574 for being insolvent consistent with the CRSP delisting code system. When a company is delisted for reasons other than default, such as merging with another company, being acquired and becoming a privately owned company, or moving to trade in a foreign exchange market, the company is considered to be censored.

The histogram of default frequencies is presented in Figure 4.2. Following the crisis in the late 1990s, the number of defaults stayed at a lower level from 2004 to 2007. The spike in the number of defaults during the global financial crisis in 2009 is quite obvious. The number then decreases following the crisis and reverts to the pre-crisis level in recent years.

4.4.1 Computing Distance-to-Default

A structural model assumes that a company defaults when its assets drop below the default threshold implied by its level of debt. The asset value is modelled with a geometric Brownian motion, and a firm's distance-to-default is defined as the number of standard deviations by which its asset value exceeds the default threshold (Merton, 1974; Leland, 1994; Duffie *et al.*, 2007). This chapter follows the approach in Hillegeist *et al.* (2004) to calculate the distance to default

$$DD_t = \frac{\ln(V_A/X) + (\mu - (\sigma_A^2/2))T}{\sigma_A\sqrt{T}}, \tag{4.16}$$

where V_A is the current market value of assets, X is the default threshold, T is the prediction time horizon, μ is the annual return on assets and σ_A is

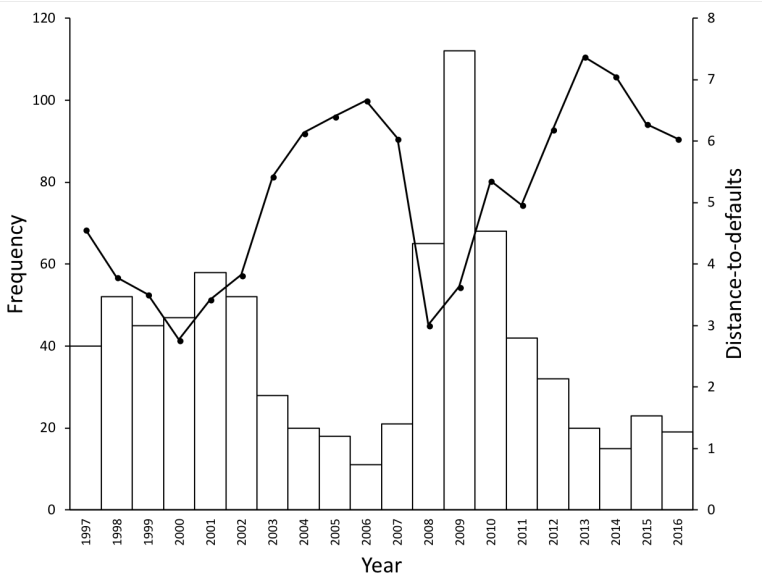


Figure 4.2. Histogram of default event and the average distance-to-defaults between 1997 and 2016.

the standard deviation of assets returns. V_A and σ_A are computed using the Black-Scholes-Merton Probability of Bankruptcy (BSM-Prob) framework (Merton, 1974; Black & Scholes, 1973)

$$\begin{aligned} V_E &= V_A e^{-\delta T} N(d_1) - X e^{-rT} N(d_2) + (1 - e^{-\delta T}) V_A, \\ \sigma_E &= (V_A e^{-\delta T} N(d_1) \sigma_A)/V_E, \end{aligned} \tag{4.17}$$

where V_E and σ_E are the value and volatility of equities, r is the annual risk-free rate, and $N(d_1)$ and $N(d_2)$ are the standard cumulative normal distribution function evaluated at d_1 and d_2, where

$$\begin{aligned} d_1 &= \frac{\ln[V_A/X] + (r + (\sigma_A^2/2))T}{\sigma_A \sqrt{T}}, \\ d_2 &= d_1 - \sigma_A \sqrt{T}. \end{aligned}$$

The one-year distance to default is computed empirically on a quarterly basis. The necessary accounting ratios are collected from quarterly accounting balance sheets in the Compustat database, and market information is from the CRSP database. Similar to Duffie *et al.* (2007) and Duan *et al.* (2012), X is measured as

the firm's short-term debt (larger of "debt in current liabilities" and "total current liabilities"), plus one half of the its long-term debt. We calculate σ_A from daily stock price using the rolling window approach as described in Duffie & Lando (2001). The one-year treasury bill yield is denoted by r and T is set to be 1.

A total of 380,719 quarterly observations are drawn for the 12,698 companies over the period of twenty years. The observations are truncated from the top at 20 (99th percentile) to avoid outliers due to errors in input data. The annual average DD during the sample period is presented in Figure 4.2. As shown in the figure, the number of defaults is negatively correlated with the average DD, indicating that a higher DD is likely to be associated with a lower default probability. A histogram of the distance to default for the remaining companies is presented in Figure 4.3. As shown, some DDs are negative, indicating that according to the applied model, for these companies the assets are lower than the debt values and thus are not sufficient to cover the corporate debt. These DDs are expected to be associated with high default probabilities.

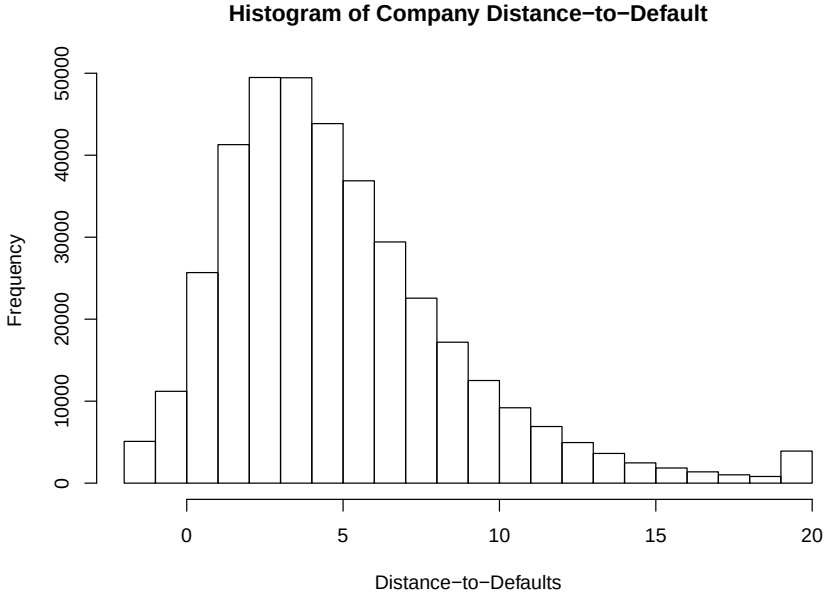


Figure 4.3. Histogram of distance-to-defaults for U.S. listed companies from 1996 to 2016.

4.4.2 Design of the Joint Model

In the LME model, the distance to default is the response variable that is described as a spline function of time,

$$\begin{aligned} y_i(t) =& m_i(t) + \epsilon_i(t) \\ =& (\beta_0 + b_{i0}) + \sum_{k=1}^{3} (\beta_k + b_{ik}) B_n(t, \lambda_k) + \epsilon_i(t). \end{aligned} \tag{4.18}$$

In the specification above, $y_i(t)$ represents the observed values of the distance to default and $m_i(t)$ denotes the estimated mean value of the distance to default, free of errors and in a form of spline curve. $\{B_n(t, \lambda_k); k = 1, 2, 3\}$ denotes a B-spline matrix for a natural cubic spline of time. We fit the LME model in two ways: one with two internal knots ($df = 3$) and the other with three internal knots in between the two boundary knots ($df = 4$). The spline coefficients have both fixed-effects and random-effects components, meaning the coefficients can vary for different companies.

We also incorporate the age of the company in the joint model as an additional explanatory variable, which is defined as the time since the company's Initial Public Offering (IPO). The *Age* variable is defined as 1, when the company is younger than five years (20 quarters), and 0 if the company has existed for more than five years.[1] As an example, for a company with an IPO at 1 Jan 1997, its age variable is 1 until 1 Jan 2002 when it becomes 0 subsequently.

Based on the two variables, the hazard rate of default is defined as follows

$$h_i(t|\mathcal{Y}_i(t), Age) = h_0(t) \exp\{\gamma Age + \alpha m_i(t)\}, t > 0. \tag{4.19}$$

The hazard rate $h_i(t)$ is expected to be positively related to the *Age* factor as companies may be more exposed to default risk in their earlier years, compared to more mature companies with established markets and business models. The hazard rate $h_t(t)$ is expected to be negatively related to the distance to default as discussed in Section 4.2.2. As a result, γ is expected to be positive and α is expected to be negative.

[1] We also tried different specifications for the distinction between relatively new and established companies, for example, a threshold of three years was chosen. However, this choice did not have a significant impact on the forecasting performance of the estimated models.

In addition to the standard joint model as specified in equation (4.19), we also consider another two extensions of the model with different parameterisations for the relationship between the longitudinal observations and the time to default. These extensions and the corresponding estimation process are explained in Rizopoulos (2012, Chapter 3).

In the first extended model, we consider the effect of the trend of DDs when analysing the default intensity. We assume that the risk at time t depends on the current true value of the longitudinal process ($m_i(t)$) as well as the slope of the true trajectory at time t, with the following specification

$$h_i(t|\mathcal{Y}_i(t), Age) = h_0(t)\exp\{\gamma Age + \alpha m_i(t) + \alpha' m'_i(t)\}, t > 0, \tag{4.20}$$

where $m'_i(t) = \frac{d}{dt}m_i(t)$.

Second, we consider the cumulative effect of DDs, as it may be beneficial to allow the risk to depend on the longitudinal marker history, not just the current value (Sylvestre & Abrahamowicz, 2009). We assume that the risk at time t depends on the whole trajectory history of the longitudinal process with different weights assigned to these past observations. As more recent observations are more important than those from further in the past, we assign a higher weight to the more recent observations

$$h_i(t|\mathcal{Y}_i(t), Age) = h_0(t)\exp\{\gamma Age + \alpha^* \int_0^t \varpi(t-s)m_i(s)ds\}, t > 0, \tag{4.21}$$

where $\varpi(\cdot)$ denotes the weight function which places smaller weights in points further in the past with a normal probability density function ($\varpi(x) = \exp(-x^2/2)/\sqrt{2\pi}$, with variance set to 1). More details on the methodology used to incorporate this weight function into modelling can be found in Rizopoulos (2012, Chapter 5).

4.5 Results

We use a randomly sample of 5,000 companies as the training set, and another 2,000 companies for out-of-sample prediction. We fit six joint models with different

specifications to the training set. First for the LME model, as mentioned in Section 4.4.2, we set the degrees of freedom to be either three (with four knots) or four (with five knots). Second, we combine each of the two LME models in a standard joint model (4.19), as well as the two extended joint models (4.20) and (4.21) described in Section 4.4.2. The six models are listed in Table 4.1. For example, Models 1, 2 and 3 use the same number of knots in the LME model and they differ in the joint model specification. Models 1 and 4 have the same joint model specification, but Model 4 has one more knot in the LME model.

Table 4.1. Summary of the six joint models fitted.

Model	LME	Joint Model
Model 1	$df = 3$	Standard joint model
Model 2	$df = 3$	Joint model with a parameter for slope
Model 3	$df = 3$	Joint model with a parameter for past observations with assigned weights
Model 4	$df = 4$	Standard joint model
Model 5	$df = 4$	Joint model with a parameter for slope
Model 6	$df = 4$	Joint model with a parameter for past observations with assigned weights

The six joint models are fitted using the R package "JMBayes" as described in Rizopoulos (2016*b*). The estimation process has the following three steps:

- Step 1: Fit a LME model to the time series data of DDs grouped by companies, using the "lme" function of the "nlme" package (Pinheiro *et al.*, 2017). The number of nodes in the spline function is specified using the *ns* function.
- Step 2: Fit a proportional hazards model using the "coxph" function of the "survival" package (Therneau, 2015). The function takes as inputs the start time and stop of the observation, and if a default event is observed during this time period. The inputs are fitted with the independent variable of company age, as described previously.
- Step 3: The fitted LME model and proportional hazards model are inputs in this step to fit a joint model to the data, using the "jointModelBayes" function of the "JMbayes" package. Additional specifications are added to the inputs for the extended models. For details on defining the additional specifications, refer to Rizopoulos (2016*b*).

4.5.1 Estimation results

The results for parameter estimation are presented in the Table 4.2. The estimation process is based on the entire 20-year history of the 5,000 in-sample companies.

Table 4.2. Modelling results of the 6 models. The coefficients of each variable are shown with their standard deviation in parentheses. *,** and *** represent respectively statistical significance at the 10%, 5% and 1% level, calculated based on the t-statistics of coefficients of each variable .

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Linear mixed-effects submodel						
Intercept	4.057***	4.061***	4.060***	5.115***	5.111***	5.119***
	(0.051)	(0.047)	(0.048)	(0.049)	(0.049)	(0.048)
β_1	2.248***	2.225***	2.252***	1.794***	1.775***	1.791***
	(0.101)	(0.102)	(0.105)	(0.076)	(0.081)	(0.078)
β_2	0.781***	0.725***	0.813***	-1.215***	-1.215***	-1.219***
	(0.128)	(0.121)	(0.126)	(0.096)	(0.094)	(0.096)
β_3	1.400***	1.388***	1.403***	-2.767***	-2.780***	-2.776
	(0.094)	(0.096)	(0.093)	(0.097)	(0.101)	(0.101)
β_4				4.066***	4.044***	4.059***
				(0.069)	(0.073)	(0.073)
Survival submodel						
Age	0.531***	0.342*	0.528**	0.382**	0.297	0.459**
	(0.199)	(0.212)	(0.210)	(0.194)	(0.208)	(0.190)
α	-0.527***	-0.558***	-0.656***	-0.526***	-0.570***	-0.526***
	(0.044)	(0.040)	(0.048)	(0.042)	(0.041)	(0.041)
α'		-4.556***			-3.033***	
		(0.489)			(0.477)	
Summary statistics						
DIC	830364	830267	830133	839282	839191	839204

For all six models, the parameter α is significant at the 1% level, indicating an important relationship between the distance to default and the company's default probability. In particular, as implied by the negative signs for the coefficients in the joint models, a higher distance to default is associated with a lower hazard rate. For example, in Model 1, a unit increase in the DD corresponds to a $\exp(-\alpha) = 0.59$ decrease in the hazard rate. This result is as expected, since a high distance to default is generated by a high asset value relative to the liability level, meaning the company is more likely to service its debt given the amount of assets available.

The coefficient of the *Age* variable is positive in all six models. As the variable is set to be 1 for a young company, the modelling results indicate that a young company is more likely to default, in comparison to a company that has survived for more than five years. A more mature company may have built a well known brand, acquired a good share of the market and possibly established good business

relationships with suppliers and customers. Thus, given that a company survives the first five years after being listed, it is less likely to default compared to a newly-listed company over the same period of time.

We also present summary statistics for the joint models in Table 4.2. The DIC represents the deviance information criterion and a lower DIC indicates a better goodness-of-fit (Rizopoulos, 2016*a*). The DICs for the six joint models are generally similar, and the extended joint models tend to generate a higher goodness-of-fit compared to a standard joint model. First, adding a slope parameter to the standard joint model helps to improve the modelling results what is illustrated by the lower DICs of Model 2 and Model 5 in comparison to Model 1 and Model 4. This implies that the default risk at time t depends not only on the DDs at time t, but also on the trend of the DDs measured as the slope of the longitudinal trajectory. Second, of the six joint models, Model 3 provides the best fit, indicating the significance of incorporating the history of DDs in assessing the hazard rate.

We note that Models 1, 2 and 3 have lower DICs than Models 4, 5 and 6, which have one more knot in the LME model. It may be expected that a more flexible specification of the LME model generates a more accurate in-sample goodness-of-fit. However, we argue that the DIC results are not necessarily a contradiction to this expectation. While adopting a spline curve with a higher degree of freedom still leads to a more accurate fit, these benefits are outweighed by the need to estimate additional parameters, resulting in higher DIC values. Due to this result, for the rest of the study, we focus on the first three joint models as they outperform the other three joint models in terms of in-sample goodness-of-fit, as well as out-of-sample prediction accuracy as shown in the next subsection.

In addition to the six joint models, for comparison purposes, we also fit a standard Cox model and an accelerated failure time (AFT) model with a Weibull response distribution (Cox & Oakes, 1984). These models are fitted using the "survival" package in R. It is expected that the signs of the coefficients of the two variables in the Cox model will be the same as the joint models, and will be the opposite to those in the Weibull model with an AFT specification. Consistent with the expectation, the parameter α is -0.315 in the Cox model and 0.064 in the Weibull model, and is significant at the 1% level in both models. The coefficient for the *Age* variable is 0.447 in the Cox model, significant at the 5% level. The *Age* variable is not significant in the Weibull model with a coefficient of -0.027.

4.5.2 Out-of-sample prediction

To test the predictive ability of the eight models (six joint models, the Cox model and the Weibull model) for company default, we apply these fitted models to 2,000 companies not used in the model fitting process. To start with, we first apply the eight models to make a five-year prediction given the observations in the previous 15 years. In other words, we use the first 60 quarters' observations to forecast the probability that the company will default within the next five years.

As an example, Figure 4.4 presents the fitted longitudinal process, $m_i(t)$, of a random company, generated by Model 1 and Model 4 as specified in Table 4.1. In the LME submodel, Model 1 has three internal knots ($df = 3$) and Model 4 has four internal knot ($df = 4$). In Figure 4.4, the dots represent the observed values of distance to default, $y_i(t)$, for this company over the 15 years (60 quarters). The step function connecting these dots is shown as the solid line. The fitted values generated by the joint models, which are used to make predictions of the continuous hazard rate, are denoted by the two smooth lines through the dots. The solid line representing Model 4 is more flexible than the dotted line. This is as expected since Model 4 allows for one more knot in the spline function.

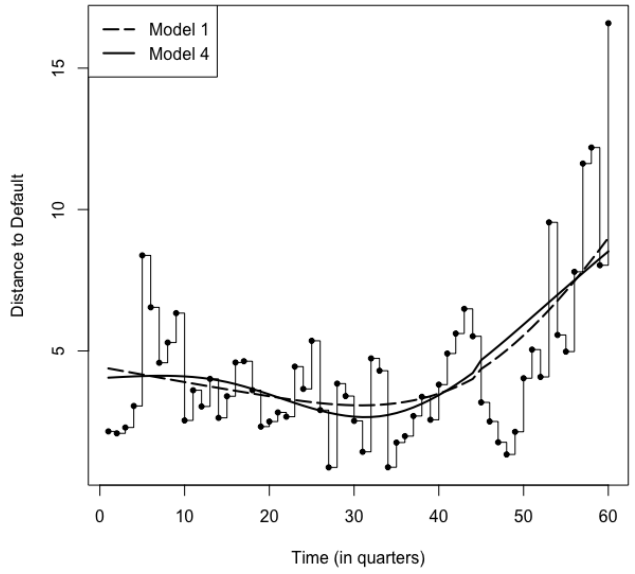


Figure 4.4. Fitted longitudinal process of distance to default for a randomly chosen company over 15 years.

The prediction results from the models are then compared to the actual default experience of these 2,000 companies. Receiver Operating Characteristic (ROC)

(Hanley & McNeil, 1982) curves are produced for the six joint models, the Cox model and the Weibull model, and they are presented in Figure 4.5 . We include the first three joint models to be compared to the Cox and Weibull models. The horizontal and vertical axes are the False Positive Rate (FPR) and the True Positive Rate (TPR) respectively. A curve that is closer to the vertical axis and the line TPR = 1 indicates good prediction power, as the model is able to discriminate an actual default companies with a high predicted default rate while not falsely indicating non-default companies.

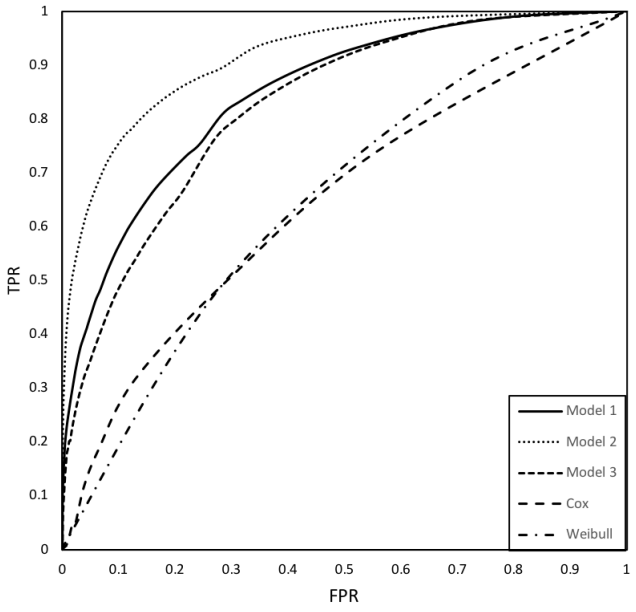


Figure 4.5. ROC curves generated by the 5 models when making out-of-sample 5-year predictions.

The prediction accuracy is measured using the Area Under the Curve (AUC) value, which is defined here as the total area under the ROC curve. The higher the AUC, the better the discriminatory ability. The highest AUC value is produced by Model 2 at 0.914, followed by Model 1 at 0.846. Model 5 generates an AUC at 0.717, lower than the other joint models but still higher than the Cox and Weibull models (0.661 and 0.672). Based on the AUC values, the joint models all outperform the Cox model and Weibull model.

4.5.3 Walk-forward prediction

In addition to the above results, we also test the prediction accuracy of the eight models over different time periods and for different prediction horizons. Note that

for this exercise we use the same 2,000 out-of-sample companies as in subsection 4.5.2. We make predictions for a horizon of two, three, four and five years using a rolling-window approach. For each of the models, we start by making predictions based on the first ten years (from 1997 to 2006 inclusive) for the probability of default in the next two years until 2008. Next we use observations from 1998 to 2007 to predict the probability for the company to default before 2009, given it is still listed and active at the end of 2009. This process is carried out until the prediction end point reaches the last quarter of 2016. So for a prediction horizon of two years, for each of the models we are able to make nine predictions. We then repeat the procedure for prediction horizon of three, four and five years, with the same one year rolling window.

Summary statistics are presented in Table 4.3, including the means and medians of the AUCs generated by all eight models. For example, 0.781 is the mean of the nine AUCs generated by Model 1 when predicting default events over a forecast horizon of two years. A higher mean indicates a better prediction accuracy on average. We note that the AUC values are higher for all joint models than for the Cox/Weibull models for all prediction horizons. The results for the joint models are similar to each other, varying by different horizons.

The average of the four means are also computed for all the models and given in the bottom row of Table 4.3. Using the average of means, which is an indication of the average prediction performance over all horizons, all joint models outperform the Cox and Weibull models as expected. This is consistent with the results for a five-year prediction as presented in subsection 4.5.2, indicating that joint models are able to produce more accurate predictions over different time periods and for various forecast horizons. Among the six joint models, Model 2 appears to be the best, confirming the results presented in Figure 4.5.

Similar to the results of the in-sample goodness-of-fit, we also note that a more flexible specification of the LME model does not necessarily lead to better out-of-sample prediction results. As shown in Table 4.3, Model 1 generates a higher average AUC than Model 4, which has the same structure as Model 1 except for the additional knot placed in the spline curve in the LME model function. The same conclusions can be drawn when comparing Model 2 and Model 5. Although opposite results are presented for Model 3 and Model 6, the two AUCs are approximately the same. These results may be due to the problem of overfitting, when the priority of estimating the models is to fit the training

data as closely as possible, which in this case reduces the models' ability to make out-of-sample predictions for companies that are different from the ones included in the training data.

Table 4.3. Summary statistics of AUCs for different prediction horizons for the 8 models .

Horizon	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Cox	Weibull
2 Year								
Mean	0.771	0.783	0.776	0.757	0.805	0.784	0.744	0.758
Median	0.761	0.773	0.775	0.757	0.804	0.789	0.763	0.779
3 Year								
Mean	0.782	0.821	0.790	0.767	0.767	0.799	0.687	0.717
Median	0.777	0.811	0.801	0.764	0.762	0.792	0.670	0.728
4 Year								
Mean	0.784	0.819	0.793	0.765	0.776	0.797	0.660	0.682
Median	0.785	0.785	0.808	0.763	0.776	0.799	0.640	0.674
5 Year								
Mean	0.781	0.812	0.794	0.754	0.751	0.789	0.658	0.653
Median	0.784	0.799	0.803	0.753	0.756	0.787	0.652	0.655
Average of Means								
Average	0.780	0.809	0.788	0.761	0.775	0.792	0.687	0.702

We plot the AUC values for different prediction horizons over time, presented in Figure 4.6, Figure 4.7, Figure 4.8 and Figure 4.9 respectively. We only include Models 1, 2 and 3 to be compared to the Cox/Weibull models, as they are better than or equivalent to their peer models as discussed previously. The minimum value on the vertical axis is set at 0.5 to better compare the models. The horizontal axis represents the starting point for prediction in quarters. For example, 40 refers to the 40th quarter, where the first 40 quarters are used to predict the default events in the next certain number of years.

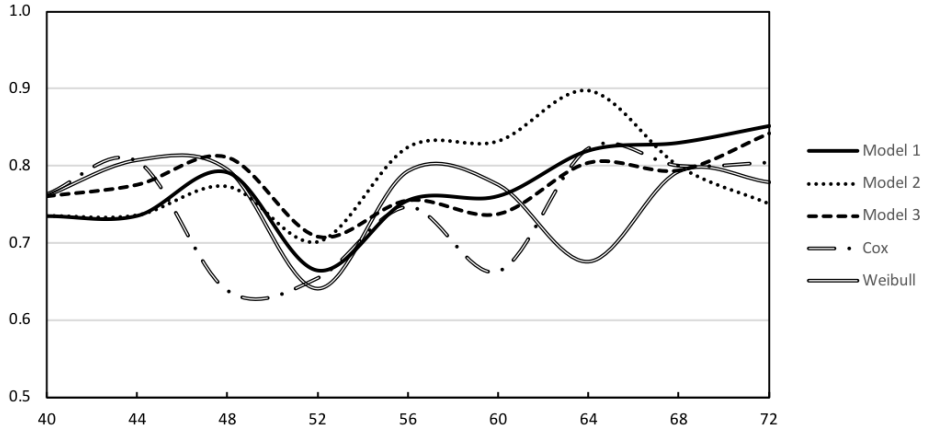


Figure 4.6. AUC values for 2-year prediction.

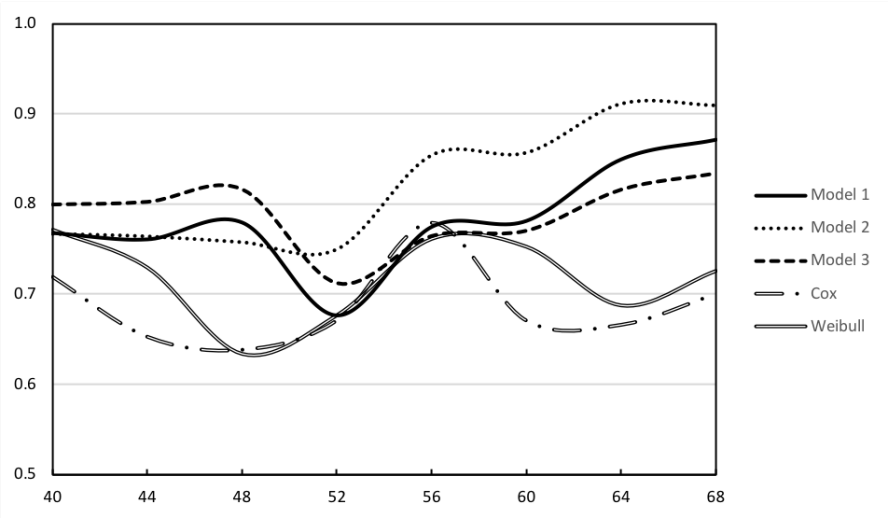


Figure 4.7. AUC values for 3-year prediction.

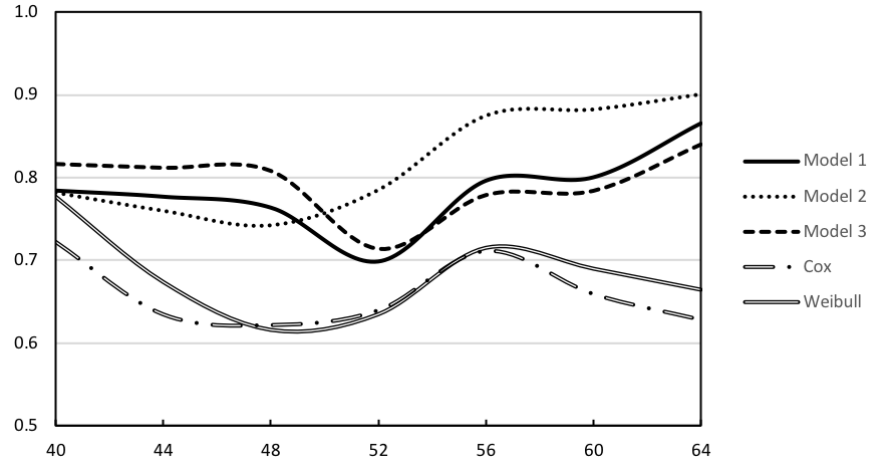


Figure 4.8. AUC values for 4-year prediction.

Some observations can be made by comparing the joint models to the Cox/Weibull models. First, it is confirmed again that the joint models outperform the Cox model and the Weibull model for all prediction horizons and over time, implied by higher AUC values. Second, the difference between the Cox or Weilbull models and the joint model widens with increasing prediction horizon. The eight models are relatively similar when projecting default events in the near future. However, the Cox model and Weibull model only use the most recent observations of the distance to default and ignore the longitudinal trajectory of the variable. When interest lies in forecasting long-term default rates, the joint models clearly perform much better. This is due to the joint model's ability to capture the dynamic nature of the distance to default in order to assess the hazard rate.

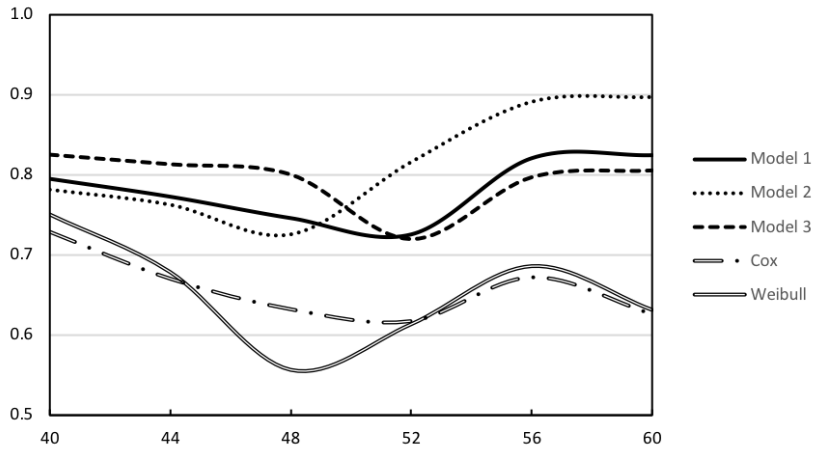


Figure 4.9. AUC values for 5-year prediction.

Overall, based on Figures 4.6-4.9, Model 2 appears to perform best also in a dynamic setting, confirming the static prediction results illustrated in Figure 4.5. Model 2 produces similar results as Model 1 and 3 in the early years, and it starts to outperform the two models significantly after quarter 52, corresponding to year 2011, i.e. the post-crisis period. It may imply that after the global financial crisis, also the trend in the DD of a company becomes more important in assessing the corporate default risk.

Moreover, Model 2 outperforms Model 3, which indicates that applying an additional weight function in the model estimation does not help to improve the results. In fact, Model 1 and Model 3 yield a similar performance during the post-crisis period, which suggests that including information on past observations of DDs does not improve the accuracy of predicting future default probabilities.

4.6 Conclusion and Future Plans

We apply the joint model for longitudinal and time-to-event data to assess corporate default risk. Our findings suggest that the joint models outperform selected traditional parametric survival models such as a Weibull and Cox model. We suggest that this outperformance is related to the fact that joint models initially apply a LME model for the DDs, before evaluating the association between DD and default risk. The superior performance is also evidenced by higher AUCs observed for the applied joint models at different prediction horizons.

There is scope for future research in the application of the joint model in this area. Currently the JMbayes package only allows for one longitudinal process when fitting a joint model: in our application this was the longitudinal process for the DD. As indicated by Rizopoulos (2016*b*), the package is to be expanded in the future to consider multiple longitudinal outcomes. Once this is achieved, more company-specific variables can be included in the model, for example the balance sheet ratios used in Duffie *et al.* (2007). This may result in more accurate prediction results.

Another potential modification relates to the censoring mechanism. Currently we only identify one exit risk which is by default. As the joint model can be extended to allow for competing risks, in future research, we can better distinguish and classify companies based on the actual reason for exiting the sample. For example, companies may exit through mergers and acquisitions, which are currently recognised as being censored. A clearly defined system of exit reasons might help us to find variables that are more closely linked to company default risk.

Chapter 5

Conclusion

5.1 Summary of main results

This thesis focuses on the evaluation of three types of financial risks, which are the insurance claim count risk, sovereign credit risk and corporate default risk. The aim of the three research papers is to improve the accuracy of the modelling results obtained previously in these three areas of applications. This is achieved by implementing novel analytical methods that haven't been applied before in these contexts, and/or using variables that can better explain the changes in the underlying risk.

Chapter 2 which is based on the research paper 'Application of the bivariate negative binomial regression model in analysing insurance count data' uses a bivariate negative binomial model to evaluate the claim count made on general insurance policies. The use of the BNBR model is justified for two reasons. First, when over-dispersion presents in the data, a negative binomial model is more suitable than a Poisson model. Second, when the aim is to jointly assess the claim counts made on two types of policies by the same policyholder, it makes sense to use a bivariate model to properly account for the correlation between the two count numbers. As expected, the BNBR model generates a higher in-sample goodness-of-fit as well as out-of-sample prediction accuracy. It outperforms the bivariate Poisson model as well as the two univariate negative binomial models combined, each assessing one policy independently.

In addition, two shrinkage methods are applied to the full model and the results show that both shrunken models generate better out-of-sample forecast figures than the full model. This implies that the shrunken models mitigate the problem of over-fitting, which arises as a result of incorporating too many policyholder features that limit the full model's ability to make predictions given different values of inputs. Moreover, the model shrunken by the Lasso performs better than the one shrunken by the ridge regression. As discussed previously in Chapter 2, the Lasso method tends to reduce variables to exact zero, which is more effective when performing variable selection.

Chapter **??** which is based on the research paper 'Assessing Sovereign Risk: A Bottom-Up Approach' shows the importance of the private sector's financial health in assessing the sovereign entity's credit risk. Compared to the traditional top-down approaches, where the independent variables are values aggregated on the sovereign level and are infrequently updated, the bottom-up approach takes advantages of the credit risk information on company level that are available at a much higher frequency. Moreover, the paper uses a market-based measure of credit risk which is also forward looking. This helps overcome the drawbacks of using just fiscal accounting ratios.

The empirical study performed on 18 U.S. state governments shows that the state credit risk factor, which is developed using company level credit risk information combined with each state's unique industry mix, is very important in forecasting the state CDS spreads. The inclusion of the state credit risk factor significantly increases the prediction accuracy of the nested model where only other macroeconomic independent variables are used. Given that these macroeconomic variables incorporate a wide range of market information that can be used to assess sovereign credit risk, this paper shows that crucial information is missing if the performance of the private sector is ignored. To conclude, the developed credit risk indicators are highly significant in forecasting sovereign CDS spreads. This finding is also supported by the consistent results generated in a range of robustness checks.

Chapter 4 which is based on the research paper 'A joint model for longitudinal and time-to-event data in corporate default risk modelling' analyses the corporate default risk with a novel approach. The joint model for longitudinal and time-to-event data combines a LME model with a standard survival model, so that a smooth function will be fitted to the independent variable and associated with the

continuous hazard rate of default for prediction. This overcomes the drawback of assuming constant values for the independent variables between observations or the so-called LVCF (last-value-carried-forward) approach. Moreover, the prediction results of the default events are more accurate with a well-analysed longitudinal trajectory for the independent variables.

The corporate DDs and the age of the company are used to assess the corporate default risk. Based on the empirical study covering U.S. listed companies over two decades, it is shown that a range of chosen joint models outperform a standard Cox and Weibull survival model, both in in-sample and out-of-sample predictions. The AUCs of the joint models are much higher than those of the Cox/Weibull models, supporting the importance of incorporating a LME model to analyse the independent variables. The results in the walk-forward predictions also show that the joint models outperform the two benchmark models over a range of prediction horizons, in particular the longer horizon. This provides the evidence that an analysed longitudinal trajectory for covariates aids in the prediction of default events.

5.2 Contributions and future research

As discussed briefly in Chapter 1, the results of the three research papers contribute to the financial risk assessment discipline, in terms of the evaluation approaches and the independent variables used to predict risk events. This section summarises the major contributions of each of the research papers, and discusses some directions for future research.

For insurance claim count analysis, the results presented in Chapter 2 demonstrate the use of a flexible correlation structure, which serves as an alternative to the copula or trivariate models. This correlation specification does not put restrictions on the relationship between the two dependent variables. The corresponding parameter estimation process is also relatively straight-forward by maximising the log likelihood, as shown in the R code presented in appendices. Another contribution of this research paper is questioning the advantages of using as much information as possible, as it is shown that the shrunken models with fewer independent variables outperform the full model incorporating the maximum number of independent variables. To the author's best knowledge, it is the first

time this area of research has been considered and addressed in an empirical study on general insurance data. The inclusion of every available characteristic of the policyholder may help fit the model as closely as possible to the in-sample data, but it may pose problems when the fitted model is to be used for forecasting claim counts for different policyholders. The over-fitting problem is very obvious in this empirical study, as both shrunken models produce more accurate out-of-sample prediction results compared to the full model. This has important implications for future studies in the area of claim count modelling and insurance policy pricing.

Future research may further apply a similar correlation structure to other models that can incorporate the over-dispersion presented in data, such as the extended Poisson models. Currently these extended Poisson models are used to assess correlated count data with a full covariance structure or copula, which may be replaced by the more flexible correlation specification used in Chapter 2. This will likely reduce the computation difficulty.

Regarding the sovereign credit risk analysis, the thesis discovers that the market expectation of the private sector's credit quality helps forecast the sovereign credit quality, which currently has received little attention in the sovereign credit risk literature. Moreover, this study is based on the earlier work by Altman & Rijken (2011*b*), and extends it further by removing the reliance on scoring models and less frequently updated accounting information. The implications for the investors who have risk exposure to sovereign risk is that a closer look at company level information is helpful. The fluctuations in the credit quality of the private sector can be used to predict the future variations in the sovereign credit quality, as argued by the results in Chapter 3.

Future research can look for alternative variables that can be linked to sovereign credit risk but supported with stronger economic reasons. Although the study in this thesis shows that the information from the private sector is important for the sovereign credit risk analysis, it does not provide the exact reason or evidence explaining this association other than through tax payments/receipts. Future studies can address this issue by using variables that can directly reflect the relationship between the corporate level and sovereign level credit quality. For example, one may consider the tax receipt figures released by state governments, and use the corporate tax component as a measure of the financial health of the private sector, which can be linked to the sovereign credit risk. Studies on

such variables will provide further evidence that the company level information is important in assessing sovereign credit risk.

The corporate default risk analysis in this thesis contributes to the current literature by modifying the exisiting survival models. Survival models are generally recognised as the most suitable approach to evaluate the corporate default risk, but the assumption that the independent variable is constant between observations can bring bias to the modelling results. The joint model for longitudinal and time-to-event data is able to mitigate the problem by first analysing the independent variables. Without the LVCF assumption, the joint model produces better prediction results. The application of the joint model in this empirical analysis demonstrates the necessity to evaluate the trajectory of the independent variable, which is not considered previously in the literature when a standard survival model is used.

Future research may modify the current joint model to allow for a more sophisticated analysis of the corporate default risk. One possibility is to introduce competing risks, so that the joint model can correctly account for other exit reasons such as mergers and acquisitions. Currently these types of the companies are classified as being censored. A more detailed classification system in terms of exit reasons may help increase the prediction accuracy. The other possibility is to include more longitudinal processes, aiming to increase further the modelling accuracy. When this is allowed for in the R package, it is very likely that the parameter estimation process takes much longer. So the benefits and costs of having multiple longitudinal processes have to be considered together, to justify the necessity of incorporating additional independent variables.

Appendix A

R code to fit a BNBR model to sample data

This appendix presents the R codes to fit a BNBR model to the insurance data in this study. The two types of policy claims are named "NCRC" for Y_1 and "NSRC" for Y_2 respectively, and the independent variables are as described in Table 2.1. The data is stored in the data-table of *sample* in R, and each row of *sample* represents the observations made on a policyholder for both dependent and independent variables. For example, the data for the first 3 policyholders are as follows. For convenience in the later part of the computation regarding the log-likelihood function, the intercept variable (*int* in column 1) and all interaction terms are calculated and listed as well.

```
> head(sample,n=3)
  int V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V1V2 V1V6 V1V8 V1V9 V1V12
1   1  0  1  0  0  0  1  1  0  0   1   1    0    0    0    0     0
2   1  0  0  0  0  1  0  1  0  0   1   1    0    0    0    0     0
3   1  0  1  0  1  0  1  1  0  0   1   1    0    0    0    0     0
  V2V6 V2V8 V2V9 V2V12 V6V8 V6V12 V8V9 V8V12 V9V12 NCRC NSRC
1    1    1    0     1    1     1    0     1     0    0    0
2    0    0    0     0    0     0    0     1     0    0    0
3    1    1    0     1    1     1    0     1     0    0    0
```

Given the data *sample*, the R codes to fit the model are presented here:

```
###Fit UNBR models for both policy types and extract the coefficients.
nbr1<-glm.nb(NCRC~V1+V2+V3+V4+V5+V6+V7+V8+V9+V10+V11
             +V1*V2+V1*V6+V1*V7+V1*V8+V1*V11+V2*V6+V2*V7
             +V2*V8+V2*V11+V6*V7+V6*V11+V7*V8+V7*V11+V8*V11,data=sample)
nbr2<-glm.nb(NSRC~V1+V2+V3+V4+V5+V6+V7+V8+V9+V10+V11
             +V1*V2+V1*V6+V1*V7+V1*V8+V1*V11+V2*V6+V2*V7
             +V2*V8+V2*V11+V6*V7+V6*V11+V7*V8+V7*V11+V8*V11,data=sample)
```

```
b_1<-nbr1$coefficients
b_2<-nbr2$coefficients
m_1<-1/nbr1$theta
m_2<-1/nbr2$theta

###Calculate the initial values of required variables.
###The 27th and 28th columns in sample store the claim numbers for
#the two types of policies.

theta_1<-1-mean(sample_i[,27])/sd(sample[,27])^2
theta_2<-1-mean(sample_i[,28])/sd(sample[,28])^2
c_1<-((1-theta_1)/(1-theta_1*exp(-1)))^(m_1^(-1))
c_2<-((1-theta_2)/(1-theta_2*exp(-1)))^(m_2^(-1))
a_1<-m_1^(-1)*theta_1*exp(-1)/(1-theta_1*exp(-1))-m_1^(-1)*theta_1/(1-theta_1)
a_2<-m_2^(-1)*theta_2*exp(-1)/(1-theta_2*exp(-1))-m_2^(-1)*theta_2/(1-theta_2)
###initial value for lambda
lmd<-cov(sample[,27],sample[,28])/(c_1*c_2*a_1*a_2)

###Group initial values.
par<-NULL
par[1:26]<-b_1
par[27:52]<-b_2
par[53]<-m_1
par[54]<-m_2
par[55]<-lmd

###Define the log-likelihood function.

log.lklh.bnb<-function(par){
    mu_1<-exp(as.matrix(sample[,1:26]) %*% par[1:26]);
    mu_2<-exp(as.matrix(sample[,1:26]) %*% par[27:52]);
    y_1<-sample[,27];
    y_2<-sample[,28];
    m_1<-par[53];
    m_2<-par[54];
    c_1<-(1+(1-exp(-1))*mu_1*m_1)^(-1/m_1)
    c_2<-(1+(1-exp(-1))*mu_2*m_2)^(-1/m_2)
    lmd<-par[55]
    a<-vector(length=x); ###x represents the number of rows in the data
    for (k in 1:x){
      a[k]<-y_1[k]*log(mu_1[k])-m_1^(-1)*log(m_1)
           -(y_1[k]+m_1^(-1))*log(m_1^(-1)+mu_1[k])-log(factorial(y_1[k]))
           +sum.log(m_1,y_1[k])+y_2[k]*log(mu_2[k])-m_2^(-1)*log(m_2)
           -(y_2[k]+m_2^(-1))*log(m_2^(-1)+mu_2[k])-log(factorial(y_2[k]))
           +sum.log(m_2,y_2[k])
           +log(1+lmd*(exp(-y_1[k])-c_1[k])*(exp(-y_2[k])-c_2[k]))
    }
    return(sum(a))
  }

###Extract the fitted coefficient values by maximising the defined
#log-likelihood function (minimising with R function "nlminb").
coe_bnbr<-nlminb(start=par, objective=-log.lklh.bnb)$par
```

Appendix B

R code to fit a BRP model

The R codes to fit a BPR model only differ from the codes in Appendix A in the definition of the log-likelihood function, presented as follows.

```
###Group initial values from the UNBR results.
par<-NULL
par[1:26]<-b_1
par[27:52]<-b_2
par[53]<-lmd

###Define the log likelhood function.
log.lklh.bpr<-function(par){
  mu_1<-exp(as.matrix(sample_i[,1:26]) %*% par[1:26]);
  mu_2<-exp(as.matrix(sample_i[,1:26]) %*% par[27:52]);
  y_1<-sample_i[,27];
  y_2<-sample_i[,28];
  lmd<-par[53]
  a<-vector(length=x);##x represents the number of rows in the data
  for (i in 1:x){
    a[i]<-y_1[i]*log(mu_1[i])+y_2[i]*log(mu_2[i])-(mu_1[i]+mu_2[i])
         +log(1+lmd*(exp(-y_1[i])-exp(-(1-exp(-1))*mu_1[i]))*(exp(-y_2[i])
         -exp(-(1-exp(-1))*mu_2[i])))-log(factorial(y_1[i]))
         -log(factorial(y_2[i]))
  }
  return(sum(a))
}

###Extract the fitted coefficient values by maximising the defined
#log-likelihood function (minimising with R function "nlminb").
coe_bpr<-nlminb(start=par,object=-log.lklh.bpr)$par
```

Appendix C

R code for model shrinkage

Two shrinkage methods are used in this study, namely the LASSO and ridge regression. The key thing to shrink the full model using these two methods is to include the penalty term in the log-likelihood function. It is the same for univariate models and bivariate models, and here we use the BNBR model as an example. As noted before, we are using standard shrinkage methods, so the penalty term is not unique to the BNBR model or any mentioned model in this study.

For the Lasso as specified in equation (2.7), the newly-defined function is presented as follows.

```
###Log-likelihood function for the LASSO
log.lklh.bnb<-function(par){
    mu_1<-exp(as.matrix(sample[,1:26]) %*% par[1:26]);
    mu_2<-exp(as.matrix(sample[,1:26]) %*% par[27:52]);
    y_1<-sample[,27];
    y_2<-sample[,28];
    m_1<-par[53];
    m_2<-par[54];
    c_1<-(1+(1-exp(-1))*mu_1*m_1)^(-1/m_1)
    c_2<-(1+(1-exp(-1))*mu_2*m_2)^(-1/m_2)
    lmd<-par[55]
    a<-vector(length=x); ###x represents the number of rows in the data
    for (k in 1:x){
      a[k]<-y_1[k]*log(mu_1[k])-m_1^(-1)*log(m_1)
           -(y_1[k]+m_1^(-1))*log(m_1^(-1)+mu_1[k])-log(factorial(y_1[k]))
           +sum.log(m_1,y_1[k])+y_2[k]*log(mu_2[k])-m_2^(-1)*log(m_2)
           -(y_2[k]+m_2^(-1))*log(m_2^(-1)+mu_2[k])-log(factorial(y_2[k]))
           +sum.log(m_2,y_2[k])
           +log(1+lmd*(exp(-y_1[k])-c_1[k])*(exp(-y_2[k])-c_2[k]))
    }
    ###lambda_i represents the shrinkage parameter
    return(sum(a)-lambda_i*sum(abs(par[c(2:26,27:52)])))
  }
```

The ridge regression differs from the LASSO in the specification of the penalty term, shown as follows.

```
###Log-likelihood function for ridge regression
log.lklh.bnb<-function(par){
    mu_1<-exp(as.matrix(sample[,1:26]) %*% par[1:26]);
    mu_2<-exp(as.matrix(sample[,1:26]) %*% par[27:52]);
    y_1<-sample[,27];
    y_2<-sample[,28];
    m_1<-par[53];
    m_2<-par[54];
    c_1<-(1+(1-exp(-1))*mu_1*m_1)^(-1/m_1)
    c_2<-(1+(1-exp(-1))*mu_2*m_2)^(-1/m_2)
    lmd<-par[55]
    a<-vector(length=x); ###x represents the number of rows in the data
    for (k in 1:x){
      a[k]<-y_1[k]*log(mu_1[k])-m_1^(-1)*log(m_1)
           -(y_1[k]+m_1^(-1))*log(m_1^(-1)+mu_1[k])-log(factorial(y_1[k]))
           +sum.log(m_1,y_1[k])+y_2[k]*log(mu_2[k])-m_2^(-1)*log(m_2)
           -(y_2[k]+m_2^(-1))*log(m_2^(-1)+mu_2[k])-log(factorial(y_2[k]))
           +sum.log(m_2,y_2[k])
           +log(1+lmd*(exp(-y_1[k])-c_1[k])*(exp(-y_2[k])-c_2[k]))
    }
    ###lambda_i represents the shrinkage parameter
    return(sum(a)-lambda_i*sum(par[c(2:26,27:52)]^2))
  }
```

Bibliography

Aizenman, J., Hutchison, M. and Jinjarak, Y. (2013). What is the risk of European sovereign debt defaults? Fiscal space, CDS spreads and market pricing of risk. *Journal of International Money and Finance* **34**, 37–59.

Allison, P. D. (2010). *Survival analysis using SAS: a practical guide*. SAS Institute.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance* **23**(4), 589–609.

Altman, E. I. and Rijken, H. A. (2011*a*). Toward a Bottom-Up Approach to Assessing Sovereign Default Risk. *Journal of Applied Corporate Finance* **23**(1), 20–31.

Altman, E. I. and Rijken, H. A. (2011*b*). Toward a Bottom-Up Approach to Assessing Sovereign Default Risk. *Journal of Applied Corporate Finance* **23**(1), 20–31.

Amato, J. D. and Furfine, C. H. (2004). Are credit ratings procyclical?. *Journal of Banking & Finance* **28**(11), 2641–2677.

Ang, A. and Longstaff, F. A. (2013). Systemic sovereign credit risk: Lessons from the US and Europe. *Journal of Monetary Economics* **60**(5), 493–510.

Arce, O., Mayordomo, S. and Pena, J. I. (2013). Credit-risk valuation in the sovereign CDS and bonds markets: Evidence from the Euro area crisis. *Journal of International Money and Finance* **35**, 124 – 145.

Babbel, D. F. (1996). Insuring sovereign debt against default. *World Bank Discussion Paper* .

Beaver, W. H., McNichols, M. F. and Rhie, J.-W. (2005). Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies* **10**(1), 93–122.

Beber, A., Brandt, M. W. and Kavajecz, K. A. (2009). Flight-to-quality or flight-to-liquidity? Evidence from the Euro-area Bond Market. *Review of Financial Studies* **22**(3), 925–957.

Bermúdez, L. and Karlis, D. (2011). Bayesian Multivariate Poisson Models for Insurance Ratemaking. *Insurance: Mathematics and Economics* **48**(2), 226–236.

Bertozzi, S. (1995). External debt models: Several approaches, few solutions - An annotated bibliography. *World Bank Discussion Paper* .

Bharath, S. T. and Shumway, T. (2008). Forecasting default with the Merton distance to default model. *Review of Financial Studies* **21**(3), 1339–1369.

Black, F. and Cox, J. C. (1976). Valuing corporate securities: Some effects of bond indenture provisions. *The Journal of Finance* **31**(2), 351–367.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *The journal of political economy* pp. 637–654.

Bolancé, C., Guillén, M. and Pinquet, J. (2003). Time-Varying Credibility for Frequency Risk Models: Estimation and Tests for Autoregressive Specifications on the Random Effects. *Insurance: Mathematics and Economics* **33**(2), 273–282.

Bolancé, C., Guillén, M. and Pinquet, J. (2008). On the Link Between Credibility and Frequency Premium. *Insurance: Mathematics and Economics* **43**(2), 209–213.

Boucher, J.-P. and Denuit, M. (2008). Credibility Premiums for the Zero-Inflated Poisson Model and New Hunger for Bonus Interpretation. *Insurance: Mathematics and Economics* **42**(2), 727–735.

Boucher, J.-P., Denuit, M. and Guillén, M. (2007). Risk Classification for Claim Counts: A Comparative Analysis of Various Zero-inflated Mixed Poisson and Hurdle Models. *North American Actuarial Journal* **11**(4), 110–131.

Boucher, J.-P., Denuit, M. and Guillen, M. (2009). Number of Accidents or Number of Claims? An Approach with Zero-Inflated Poisson Models for Panel Data. *Journal of Risk and Insurance* **76**(4), 821–846.

Boudoukh, J., Richardson, M. and Whitelaw, R. F. (2008). The Myth of Long-Horizon Predictability. *Review of Financial Studies* **21**(4), 1577–1605.

Brouhns, N., Guillén, M., Denuit, M. and Pinquet, J. (2003). Bonus-Malus Scales in Segmented Tariffs With Stochastic Migration Between Segments. *Journal of Risk and Insurance* **70**(4), 577–599.

Caceres, C., Guzzo, V. and Segoviano Basurto, M. (2010). Sovereign spreads: Global risk aversion, contagion or fundamentals?. *IMF working papers* pp. 1–29.

Cade, B. S. and Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment* **1**(8), 412–420.

Calice, G., Chen, J. and Williams, J. (2013). Liquidity spillovers in sovereign bond and CDS markets: An analysis of the Eurozone sovereign debt crisis. *Journal of Economic Behavior & Organization* **85**(C), 122–143.

Cameron, A. C., Li, T., Trivedi, P. K. and Zimmer, D. M. (2004). Modelling the Differences in Counted Outcomes Using Bivariate Copula Models with Application to Mismeasured Counts. *The Econometrics Journal* **7**(2), 566–584.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge University Press.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Campbell, J. Y., Hilscher, J. and Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance* **63**(6), 2899–2939.

Campbell, J. Y., Lo, A. W.-C., MacKinlay, A. C. et al. (1997). *The econometrics of financial markets*. Vol. 2. princeton University press Princeton, NJ.

Chava, S. and Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. *Review of Finance* **8**(4), 537–569.

Chen, S.-S., Chen, H.-Y., Chang, C.-C. and Yang, S.-L. (2016). The relation between sovereign credit rating revisions and economic growth. *Journal of Banking & Finance* **64**, 90–100.

Chen, Y. and Hanson, T. (2017). Copula regression models for discrete and mixed bivariate responses. *Journal of Statistical Theory and Practice* pp. 1–16.

Cox, D. R. and Oakes, D. (1984). *Analysis of survival data*. Vol. 21. CRC Press.

Cox, D. R. et al. (1972). Regression models and life tables. *JR stat soc B* **34**(2), 187–220.

Crosby, P. (1998). Modelling default risk. *Credit Derivatives: Trading & Management of Credit & Default Risk* .

Crouhy, M., Galai, D. and Mark, R. (2000). A comparative analysis of current credit risk models. *Journal of Banking & Finance* **24**(1), 59–117.

Czado, C., Kastenmeier, R., Brechmann, E. C. and Min, A. (2012). A Mixed Copula Model for Insurance Claims and Claim Sizes. *Scandinavian Actuarial Journal* **2012**(4), 278–305.

De Servigny, A., Renault, O. and De Servigny, A. (2004). *Measuring and managing credit risk*. McGraw-Hill New York, NY.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)* pp. 1–38.

Denuit, M., Dhaene, J., Goovaerts, M. and Kaas, R. (2006). *Actuarial theory for dependent risks: measures, orders and models*. John Wiley & Sons.

Denuit, M. and Lang, S. (2004). Non-Life Rate-Making with Bayesian GAMs. *Insurance: Mathematics and Economics* **35**(3), 627–647.

Denuit, M., Maréchal, X., Pitrebois, S. and Walhin, J.-F. (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. John Wiley & Sons.

Denuit, M., Van Keilegom, I., Purcaru, O. et al. (2006). Bivariate Archimedean Copula Models for Censored Data in Non-Life Insurance. *Journal of Actuarial Practice* **13**, 5–32.

Dewachter, H., Iania, L., Lyrio, M. and de Sola Perea, M. (2015). A macro-financial analysis of the Euro area sovereign bond market. *Journal of Banking & Finance* **50**, 308 – 325.

Dieckmann, S. and Plank, T. (2012). Default risk of advanced economies: An empirical analysis of credit default swaps during the financial crisis. *Review of Finance* **16**(4), 903–934.

Dionne, G. and Vanasse, C. (1989). A Generalization of Automobile Insurance Rating Models: the Negative Binomial Distribution with a Regression Component. *Astin Bulletin* **19**(2), 199–212.

Dotsey, M. (1998). The predictive content of the interest rate term spread for future economic growth. *FRB Richmond Economic Quarterly* **84**(3), 31–51.

Douak, F., Melgani, F. and Benoudjit, N. (2013). Kernel Ridge Regression with Active Learning for Wind Speed Prediction. *Applied Energy* **103**, 328–340.

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

Duffie, D. and Lando, D. (2001). Term structures of credit spreads with incomplete accounting information. *Econometrica* **69**(3), 633–664.

Duffie, D., Saita, L. and Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics* **83**(3), 635–665.

Duffie, D. and Singleton, K. J. (2003). *Credit risk: pricing, measurement, and management*. Princeton University Press.

Dwyer, D. and Korablev, I. (2007). Power and Level Validation of Moody's KMV EDFTM Credit Measures in North America. *Moody's KMV White Paper (Europe and Asia)* .

El-Basyouny, K. and Sayed, T. (2009). Collision prediction models using multivariate Poisson-lognormal regression. *Accident Analysis & Prevention* **41**(4), 820–828.

Elashoff, R. M., Li, G. and Li, N. (2008). A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics* **64**(3), 762–771.

Engelmann, B., Hayden, E. and Tasche, D. (2003). Testing rating accuracy. *Risk* **16**(1), 82–86.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer Science & Business Media.

Famoye, F. (2010*a*). A new bivariate generalized Poisson distribution. *Statistica Neerlandica* **64**(1), 112–124.

Famoye, F. (2010*b*). On the Bivariate Negative Binomial Regression Model. *Journal of Applied Statistics* **37**(6), 969–981.

Fender, I., Hayo, B. and Neuenkirch, M. (2012). Daily pricing of emerging market sovereign CDS before and during the global financial crisis. *Journal of Banking & Finance* **36**(10), 2786 – 2794.

Frees, E. W. and Valdez, E. A. (1998). Understanding Relationships Using Copulas. *North American actuarial journal* **2**(1), 1–25.

Frees, E. W. and Valdez, E. A. (2008). Hierarchical Insurance Claims Modeling. *Journal of the American Statistical Association* **103**(484), 1457–1469.

Fuertes, A.-M. and Kalotychou, E. (2004). Forecasting sovereign default using panel models: A comparative analysis. Technical report. Society for Computational Economics.

García, C., García, J., López Martín, M. and Salmerón, R. (2015). Collinearity: Revisiting the Variance Inflation Factor in Ridge Regression. *Journal of Applied Statistics* **42**(3), 648–661.

Giesecke, K. and Kim, B. (2011). Systemic risk: What defaults are telling us. *Management Science* **57**(8), 1387–1405.

Green, W. H. (1993). *Econometric Analysis, 1993*. Prentice Hall, New York.

Grinols, E. (1976). *International debt rescheduling and discrimination using financial variables*. U.S. Treasury Department, Washington, D.C.

Groba, J., Lafuente, J. A. and Serrano, P. (2013). The impact of distressed economies on the EU sovereign market. *Journal of Banking & Finance* **37**(7), 2520 – 2532.

Haberman, S. and Renshaw, A. E. (1996). Generalized Linear Models and Actuarial Science. *The Statistician* pp. 407–436.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve.. *Radiology* **143**(1), 29–36.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Vol. 43. CRC Press.

Haugh, D., Ollivaud, P. and Turner, D. (2009). What drives sovereign risk premiums?: An analysis of recent evidence from the Euro Area. Technical report. OECD Publishing.

Heller, G. Z., Mikis Stasinopoulos, D., Rigby, R. A. and De Jong, P. (2007). Mean and Dispersion Modelling for Policy Claims Costs. *Scandinavian Actuarial Journal* **2007**(4), 281–292.

Henderson, R., Diggle, P. and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**(4), 465–480.

Hillegeist, S. A., Keating, E. K., Cram, D. P. and Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies* **9**(1), 5–34.

Hilscher, J. and Nosbusch, Y. (2004). Determinants of sovereign risk. *Department of Economics, Harvard University, mimeo* .

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**(1), 55–67.

Hosmer, D. W., Lemeshow, S. and May, S. (2011). *Applied survival analysis*. Wiley Blackwell.

Hosmer Jr, D. W., Lemeshow, S. and Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Hui, C.-H. and Chung, T.-K. (2011). Crash risk of the Euro in the sovereign debt crisis of 2009-2010. *Journal of Banking and Finance* **35**(11), 2945 – 2955.

Hürlimann, W. (1990). On Maximum Likelihood Estimation for Count Data Models. *Insurance: Mathematics and Economics* **9**(1), 39–49.

Hwang, R.-C. (2012). A varying-coefficient default model. *International Journal of Forecasting* **28**(3), 675–688.

Ibrahim, J. G., Chu, H. and Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology* **28**(16), 2796–2801.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Vol. 112. Springer.

Janus, T., Jinjarak, Y. and Uruyos, M. (2013). Sovereign default risk, overconfident investors and diverse beliefs: Theory and evidence from a new dataset on outstanding credit default swaps. *Journal of Financial Stability* **9**(3), 330–336.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. Vol. 165. Wiley New York.

Jovan, M. and Ahčan, A. (2017). Default prediction with the Merton-type structural model based on the NIG Lévy process. *Journal of Computational and Applied Mathematics* **311**, 414–422.

Jung, R. C. and Winkelmann, R. (1993). Two Aspects of Labor Mobility: A Bivariate Poisson Regression Approach. *Empirical Economics* **18**(3), 543–556.

Karlis, D. and Xekalaki, E. (2005). Mixed Poisson Distributions. *International Statistical Review* **73**(1), 35–58.

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

King, G. (1989). A Seemingly Unrelated Poisson Regression Model. *Sociological Methods & Research* **17**(3), 235–255.

Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate Discrete Distributions*. New York: Marcel Dekker.

Kocherlakota, S. and Kocherlakota, K. (2001). Regression in the Bivariate Poisson Distribution. *Communications in statistics. Theory and methods* **30**(5), 815–825.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives* **15**(4), 143–156.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the 1995 International Joint Conference on Artificial Intelligence* pp. 1137–1143.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* pp. 963–974.

Lakshminarayana, J., Pandit, S. and Srinivasa Rao, K. (1999). On A Bivariate Poisson Distribution. *Communications in Statistics-Theory and Methods* **28**(2), 267–276.

Leland, H. E. (1994). Corporate debt value, bond covenants, and optimal capital structure. *The journal of finance* **49**(4), 1213–1252.

Lindstrom, M. J. and Bates, D. M. (1988). Newton?Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association* **83**(404), 1014–1022.

Longstaff, F. A., Pan, J., Pedersen, L. H. and Singleton, K. J. (2011). How sovereign is sovereign credit risk?. *American Economic Journal: Macroeconomics* **3**(2), 75–103.

Longstaff, F. A. and Schwartz, E. S. (1995). A simple approach to valuing risky fixed and floating rate debt. *The Journal of Finance* **50**(3), 789–819.

Ma, J., Kockelman, K. M. and Damien, P. (2008). A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis & Prevention* **40**(3), 964–975.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Vol. 37. CRC Press.

Meijer, R. J. and Goeman, J. J. (2013). Efficient Approximate k-Fold and Leave-One-Out Cross-Validation for Ridge Regression. *Biometrical Journal* **55**(2), 141–155.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates*. *The Journal of Finance* **29**(2), 449–470.

Moody's (2012). 'Public firm Expected Default Frequency (EDF) credit measures: Methodology, performance, and model extensions'.

Morgan, J. (1986). A new look at debt rescheduling indicators and models. *Journal of International Business Studies* **17**(2), 37–54.

Mosteller, F. and Tukey, J. (1977). *Data analysis and regression: A second course in statistics*. Addison-Wesley series in behavioral science. Addison-Wesley Pub. Co.

Nazeran, P. and Dwyer, D. (2015). 'Credit Risk Modeling of Public Firms: EDF9'.

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* **55**(3), 703–708.

Neziri, H. (2009). Can credit default swaps predict financial crises? Empirical study on emerging markets. *Journal of Applied Economic Sciences* **4**.

Nikulin, M. and Wu, H.-D. I. (2016). *The Cox model and its applications*. Springer.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research* pp. 109–131.

Orth, W. (2012). *Multi-Period Credit Default Prediction-A Survival Analysis Approach*. Shaker.

Orth, W. (2013). Multi-period credit default prediction with time-varying covariates. *Journal of Empirical Finance* **21**, 214–222.

Oshiro, N. and Saruwatari, Y. (2005). Quantification of sovereign risk: Using the information in equity market prices. *Emerging Markets Review* **6**(4), 346–362.

Pan, J. and Singleton, K. J. (2008). Default and Recovery Implicit in the Term Structure of Sovereign CDS Spreads. *Journal of Finance* **63**(5), 2345–2384.

Park, M. Y. and Hastie, T. (2007). L1-Regularization Path Algorithm for Generalized Linear Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(4), 659–677.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team (2017). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-131.
URL: *https://CRAN.R-project.org/package=nlme*

Platt, H. D. and Platt, M. B. (1991). A note on the use of industry-relative ratios in bankruptcy prediction. *Journal of Banking & Finance* **15**(6), 1183–1194.

Prentice, R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69**(2), 331–342.

R Core Team (2015*a*). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.
URL: *http://www.R-project.org/*

R Core Team (2015*b*). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.
URL: *http://www.R-project.org/*

R Core Team (2015*c*). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.
URL: *http://www.R-project.org/*

Renshaw, A. (1995). Modelling the Claims Process in the Presence of Covariates.. *Insurance Mathematics and Economics* **2**(16), 167.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software* **35**(9), 1–33.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.

Rizopoulos, D. (2016*a*). The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC. *Journal of Statistical Software* **72**(7), 1–45.

Rizopoulos, D. (2016*b*). The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC. *Journal of Statistical Software* **72**(1), 1–46.

Rizopoulos, D. and Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine* **30**(12), 1366–1380.

Samson, D. and Thomas, H. (1987). Linear Models As Aids in Insurance Decision Making: the Estimation of Automobile Insurance Claims. *Journal of Business Research* **15**(3), 247–256.

Saunders, A. and Cornett, M. M. (2003). *Financial institutions management: A risk management approach*. Irwin/McGraw-Hill.

Shen, X., Alam, M., Fikse, F. and Rönnegård, L. (2013). A Novel Generalized Ridge Regression Method for Quantitative Genetics. *Genetics* **193**(4), 1255–1268.

Shi, P. and Valdez, E. A. (2014). Multivariate Negative Binomial Models for Insurance Claim Counts. *Insurance: Mathematics and Economics* **55**, 18–29.

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

Stock, J. H. and Watson, M. W. (1989). New Indexes of Coincident and Leading Economic Indicators. *NBER Macroeconomics Annual* pp. 351–394.

Suisse, C. (1997). CreditRisk+: A Credit Risk Management Framework. *Credit Suisse Financial Products* .

Sundaresan, S. M. (2000). Continuous-Time Methods in Finance: A Review and an Assessment. *The Journal of Finance* **55**(4), 1569–1622.

Sylvestre, M. and Abrahamowicz, M. (2009). Flexible modelling of the cumulative effects of time-dependent exposures on the hazard. *Statistics in Medicine* **28**, 3437–3453.

Tang, Y., Xiang, L. and Zhu, Z. (2014). Risk Factor Selection in Rate Making: EM Adaptive LASSO for Zero-Inflated Poisson Regression Models. *Risk Analysis* **34**(6), 1112–1127.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. version 2.38.
URL: *https://CRAN.R-project.org/package=survival*

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Trück, S. and Rachev, S. (2009). *Rating Based Modeling of Credit Risk.*. Academic Press, Burlington, MA.

Trueck, S. and Rachev, S. T. (2009). *Rating Based Modeling of Credit Risk: Theory and Application of Migration Matrices*. Academic Press.

Tseng, Y.-K., Hsieh, F. and Wang, J.-L. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika* **92**(3), 587–603.

Venables, W. N. and Ripley, B. D. (2002*a*). *Modern Applied Statistics with S*. fourth edn. Springer. New York. ISBN 0-387-95457-0.
URL: *http://www.stats.ox.ac.uk/pub/MASS4*

Venables, W. N. and Ripley, B. D. (2002*b*). *Modern Applied Statistics with S*. fourth edn. Springer. New York. ISBN 0-387-95457-0.
URL: *http://www.stats.ox.ac.uk/pub/MASS4*

Wang, Z., Ma, S. and Wang, C.-Y. (2015). Variable Selection for Zero-Inflated and Overdispersed Data with Application to Health Care Demand in Germany. *Biometrical Journal* **57**(5), 867–884.

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies* **21**(4), 1455–1508.

Xu, J. and Zeger, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **50**(3), 375–387.

Yip, K. C. and Yau, K. K. (2005). On Modeling Claim Frequency Data in General Insurance with Extra Zeros. *Insurance: Mathematics and Economics* **36**(2), 153–163.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters* **54**(4), 437–447.

Zavgren, C. V. (1985). Assessing the vulnerability to failure of American industrial firms: a logistic analysis. *Journal of Business Finance & Accounting* **12**(1), 19–45.

Zhou, C. (1997). A jump-diffusion approach to modeling credit risk and valuing defaultable securities. *The Board of Governors of the Federal Reserve System* .

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research* pp. 59–82.