A Comparative Analysis of the Frequentist and Bayesian Approaches to Stress Testing

Zheqi Wang

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

The University of Edinburgh

2020

Declaration

This thesis has been composed by myself and contains no material that has been accepted for the award of any other degree at any university. To the best of my knowledge and belief, this thesis contains no material previously published by any other person except where due acknowledgement has been made.

Zheqi Wang

2020

Table of Contents

List of Tables

List of Figures

Acknowledgement

First and foremost, I would like to express my deepest gratitude to my supervisor Prof. Jonathan Crook. Many thanks to him for giving me the privilege to be his PhD student. He is one of the smartest and most gifted people I know. I thank him for his academic guidance, for inspiring me, and for his unparalleled patience and generosity in helping me improve my thesis. I would also like to thank my second supervisor Dr. Galina Andreeva, for her guidance and support and her invaluable, always timely and helpful comments and advice.

I am really lucky to have Jonathan and Galina as my supervisors. They are professional and strict, but also the nicest, warmest and funniest people. I learn a lot from meetings and discussions with them. They are role models for me in so many ways, from doing research, to becoming a hardworking person. I am forever grateful for all the time and energy they spent in helping me and I deeply appreciate the opportunity to work with them.

Thanks to my dear friends and colleagues who make my time in Edinburgh fun and happy. My appreciation also goes to the UEBS staff for providing kind and swift help when it's needed.

Thanks to the China Scholarship Council for funding my PhD study. Their financial support helped me focus on my research.

Special thanks to my internal examiner, Prof. Jake Ansell, and my external examiner, Prof. Paolo Giudici for their very constructive and helpful comments and suggestions which help me improve my thesis and also my future academic career development.

Last but not least, I would like to thank my parents. Without their unconditional love, nothing I did would have been possible.

List of Abbreviations

AC	Autocorrelations	K-NN	K-NearestNeighbour
AR	Autoregressive	LCR	Liquidity Coverage Ratio
ARIMA	Autoregressive Integrated Moving Average	LDA	Linear Discriminant Analysis
ARMA	Autoregressive Moving Average	LGD	Loss Given Default
ASF	Available Stable Funding	LP	Linear Programming
AUC	Area Under the Curve	MA	Moving Average
Bag	Bootstrap Aggregating	MARS	Multivariate Adaptive Regression Splines
BagDT	Bag Decision Tree	MCMC	Markov Chain Monte Carlo
BagNN	Bag Neural Network	M-H	Metropolis-Hastings Algorithm
BCBS	Basel Committee on Banking Supervision	MLP	Multilayer Perceptron Neuron Networks
BMA	Bayesian Model Averaging	MP	Mathematical Programing
BMS	Bayesian Model Selection	NB	Naïve Bayes
BN	Bayesian Network	NN	Neural Network
BoE	Bank of England	NSFR	Net Stable Funding Ratio
Bst	Boosting	PA	Probit Analysis
CART	Classification and Regression Tree	PAC	Partial Autocorrelations
CCAR	Comprehensive Capital Analysis and Review	PCC	Percentage Correctly Classified
CDF	Cumulative Distribution Function	PD	Probability of Default
DFAST	Dodd-Frank Act Stress Test	RF	Random Forest
DMslct	Dynamic Ensemble/Classifier Selection	ROC	Receiver Operating Characteristic
DR	Default Rate	Rot F	Rotation Forest
DT	Decision Trees	RS	Random Subspace
EAD	Exposure at Default	RSF	Required Stable Funding
EBA	European Banking Authority	RWA	Risk Weighted Assets
EL	Expected Loss	SCAP	Supervisory Capital Assessment Program
EMC	Estimated Misclassification Cost	GB	Gradient Boosting
F	Fuzzy Classifiers	SIFI	Systematically Important Financial Institution
FRB	Federal Reserve Board	Stck	Stacking
FSAPs	Financial Sector Assessment Programmes	SVM	Support Vector Machine
Hbrd	Hybrid	UL	Unexpected Loss
HQLA	High Quality Liquid Asset Amount	UWA	Uncertainty Weighted Accuracy
IRB	Internal Ratings Based	V	Voting
K-mns	K-means Clustering	VaR	Values at Risk

Abstract

Stress testing is necessary for banks as it is required by the Basel Accords for loss predictions and regulatory and economic capital computations. It has become increasingly important especially after the 2008 global financial crisis. Credit models are essential in controlling credit risk. The search for new ways to more accurately predict credit risk continues. This thesis concentrates on stress testing the probability of default using the Bayesian posterior distribution to incorporate estimation uncertainty and parameter instability. It also explores modelling the probability of default using Bayesian informative priors to enhance the model predictive accuracy.

A new Bayesian informative prior selection method is proposed to include additional information to credit risk modelling and improve model performances. We employ cross-sectional logistic regressions to model the probability of default of mortgage loans using both the Bayesian approach with various priors and the frequentist approach. In the Bayesian informative prior selection method that we propose, we treat coefficients in the PD model as time series variables. We build ARIMA models to forecast the coefficient values in future time periods and use these ARIMA forecasts as Bayesian informative priors. We find that the Bayesian models using this prior selection method outperform both frequentist models and Bayesian models with other priors in terms of model predictive accuracy.

We propose a new stress testing method to model both macroeconomic stress and coefficient uncertainty. Based on U.S. mortgage loan data, we model the probability of default at the account level using discrete time hazard analysis. We employ both the frequentist and Bayesian methods in parameter estimation and default rate (DR) stress testing. By applying the parameter posterior distribution obtained in the Bayesian approach to simulating the Bayesian estimated DR distribution, we reduce the estimation risk coming from employing point estimates in stress testing. We find that the 99% value at risk (VaR) using the Bayesian posterior distribution approach is

around 6.5 times the VaR at the same probability level using the frequentist approach with parameter mean estimates.

We further simulate DR distributions based on models built on crisis and tranquil time periods to explore the impact changes in model parameters between different scenarios have on stress testing results. We apply the parameter posterior distribution obtained in a Bayesian approach to stress testing to reduce the estimation risk that results from using parameter point estimates. We compute the VaRs and required capital with both parameter instability between scenarios and with estimation risk considered. The results are compared with those obtained when coefficient changes in stress testing models or coefficient uncertainty are neglected. We find that the required capital is considerably underestimated when neither parameter instability nor estimation risk is addressed.

Chapter 1
Introduction

1.1 Introduction

Credit risk management has long been an essential point of attention for financial practitioners and academics alike. The 2008 financial crisis further emphasised the need to increase the accuracies of credit scoring models and the importance of credit risk stress testing for financial institutions and regulators to measure and prevent potential losses.

Researchers constantly look for ways to improve the predictive accuracy of credit scoring models either by including additional information or by using better models. For instance, some include additional covariates (Altman, Sabato, & Wilson, 2010; Duffie, Saita, & Wang, 2007; Hu & Ansell, 2007; Tinoco & Wilson, 2013; Wilson & Altanlar, 2014) or improve sample data quality (Forrest, 2011; Mujalli, López, & Garach, 2016; Siddiqi, 2012) to better explain failure events. Some opt for the generally best performing individual classifiers such as logistic regression, support vector machines and neural networks, etc. in the literature to predict default rates (Bellotti & Crook, 2009; Bravo, Thomas, & Weber, 2015; Brown & Mues, 2012; Leong, 2016; Marqués, García, & Sánchez, 2012a), or use ensemble and hybrid models to combine a set of base learners to increase classification accuracy (Abedini, Ahmadzadeh, & Noorossana, 2016; Abellán & Castellano, 2017; Fallahpour, Lakvan, & Zadeh, 2017; Lessmann, Baesens, Seow, & Thomas, 2015). As a way of both incorporating additional information and obtaining better models, we consider the use of an informative Bayesian approach and finding more appropriate Bayesian informative priors selection methods could be effective in improving model performance.

Stress testing is one of the most fundamental risk related topics as whether a risk exposure can survive stresses is a very intuitive and important question. When banks cannot sustain a stress event, the depositors, the banks' shareholders, financial stability, and the financial system all suffer. A bank's assets (A) are mainly what are owed to the bank and what the bank owns. They consist of liquid assets such as cash, loans and advances to borrowers and other banks, marketable securities it holds, and fixed assets such as real estates, etc. A bank's liabilities (L) are mainly what the bank owes. Deposits, securities and financial instruments that a bank has issued are examples of its liabilities. A bank's equity (E) is the difference between the market value of its assets and its liabilities' book value: $E = A - L$. Under extreme but plausible stresses, many loans turn bad as the borrowers become unable to repay. Therefore the value of the loans drops and subsequently the financial institution's asset value decreases. Then the equity value decreases given the liabilities stay unchanged. The institution becomes insolvent when the asset value drops to the point that it is less than its liability value and the equity value becomes negative. When a bank is insolvent it does not have sufficient funds to repay its liabilities. In this case, depositors cannot receive their deposits back if they all ask for their deposits back at the same time. The stock price plummets. The bank may enter bankruptcy which leads to unemployment. Besides, it will collectively hurt the financial stability and the macroeconomy if the same happens to groups of banks. Furthermore, taxpayers lose when they have to bail out banks. Therefore, it is of utmost importance for banks to maintain enough capital (such as equity capital) to absorb decreases in asset value. It is also of utmost importance to regularly stress test banks' resilience to ensure this is highly likely to be the case.

To increase safety against bankruptcy and to maintain its competitiveness against other banks, a bank evaluates and maintains the capital needed to absorb potential losses under stress using its own predictive models to maintain a level of failure probability based on its risk appetite. The capital calculated in this way is called economic capital. Banks can use any method they see fit to model economic capital.

To protect financial stability and protect the depositors, financial regulators also require banks to adhere to regulatory capital requirements. Different institutions are subject to different regulators and regulatory capital requirements. Among them, the Basel Accords are some of the most widely used standards for regulatory capital determination adopted by over 100 countries and regions (BCBS, 2001; Thomas, Crook, & Edelman, 2017). The Basel Accords require banks to set aside enough regulatory capital to absorb losses resulting from various types of risks so that they will not become insolvent when assets fall. Based on the Basel Accords, regulatory capital for credit risk is calculated as a percentage of the risk weighted average assets with the weights decided by the risk types. Basel II allows banks to calculate the capital needed using their internal ratings based approach (IRB) by developing their own models to estimate the risk parameters such as the probability of default. Banks are required to stress test these risk parameters under severe but plausible economic conditions. In addition to internal stress tests required of the banks, regulators also regularly conduct macro stress tests on banks of large size to evaluate the ability of the banking sector to withstand losses under stressed scenarios.

Although they seemed well capitalised based on regulatory capital requirements, during the financial crisis many banks still had insufficient capital which led to their failure or near-failure (Schuermann, 2014; Thomas et al., 2017). In response to the financial crisis, the U.S. government authorised $475 billion to purchase equity and toxic assets from banks to improve the solvency and liquidity of these banks in case of a total collapse of the U.S. financial system. For instance, around $200 billion was spent in purchasing preferred stock and equity warrants from hundreds of banks through the Capital Purchase Program. Purchasing illiquid mortgage-backed securities and assisting residential mortgage loan foreclosures cost more than $65 billion. Around $70 billion was spent in stock purchase of the American International Group, and $40 billion in that of Citigroup and Bank of America. In response to the financial crisis, the UK government announced a bank rescue package of £500 billion to restore confidence in and stabilise the British banking system. £50 billion was made available to recapitalise the banks through common and preferred stocks purchase. The government invested in the banks short of capital, making them partly nationalised through taxpayers' money. For instance, the Royal Bank of Scotland raised £20 billion capital, and Lloyds banking group £17 billion through the Bank Recapitalisation Fund.

Inadequate bank capital was attributed to many factors, such as insufficient minimum capital ratio requirements, the definition of capital being too wide, excessive leverage, procyclical amplification of financial shocks, insufficient liquidity requirements, etc. (BCBS, 2011). We consider another reason could be the neglect of uncertainty in the regulatory and bank internal risk models in use to assess the capital needed. For instance, the Basel accords require regulatory capital to cover credit risk, operational risk and market risk, etc., but did not include estimation risk which is the uncertainty of the coefficient estimates in the risk models. This type of risk may be present when modelling all types of risks that are required of the banks. We consider that incorporating the coefficient estimation risk in stress testing may increase perceived risk and provide more conservative predictions of losses and required capital, therefore helping financial institutions make better and safer capital planning decisions.

In summary, in the context of the financial crisis, stress testing has become a standard tool in the macroprudential framework (Borio, Drehmann, & Tsatsaronis, 2014). It is of vital importance for financial institutions to use appropriate stress testing methods to accurately evaluate portfolio riskiness given the worst case scenarios and to make suitable risk management decisions. At the same time, it is a constant point of interest for researchers to find new ways to improve the predictive power of the credit scoring models to better predict credit defaults and potential losses.

1.2 Credit risk stress testing in practice

The amount of regulatory capital required by the first Basel Accord, Basel I, was no less than 8% of the banks' risk weighted assets to cover their credit risk (BCBS, 1988). Stress testing was not required in the Basel I Accord. Credit risk is not the only risk that may reduce banks' asset values and cause insolvency. Market risk, such as fluctuations in interest rates, exchange rates, commodities prices, etc. may cause stock value and foreign currency to drop in value, and subsequently decrease the bank's total asset value. Since Basel I predominantly focused on credit risk, between Basel I and II Accords, the Basel Committee on Banking Supervision (BCBS) issued the 1996 amendment to Basel I to incorporate market risks. BCBS first introduced stress testing in the 1996 Amendment. According to the 1996 Amendment, banks using internal models to meet capital standards for market risk must perform stress tests which include scenarios that cover low probability events of all major risk types that could lead to extraordinary losses. The ability to have rigorous and comprehensive stress test programs is a requirement to gain regulatory approval for banks to use internal models (BCBS, 1996).

While introducing stress testing and allowing internal models for market risk, the 1996 Amendment made no mention of internal models or stress testing for credit risk. This was addressed in the Basel II Accord that followed. As in the Basel I Accord, in the Basel II Accord the minimum regulatory capital is 8% of the banks' risk weighted assets. Upon regulatory approval, banks can use their internal models to compute the risk weighted assets (RWA) of a portfolio by calculating its unexpected loss (UL):

$$RWA = 12.5UL \, .$$

The UL of a portfolio is calculated as a function of the probability of default (PD), loss given default (LGD) and exposure at default (EAD):

$$UL = EAD * LGD * (VaR - E(PD))$$

in which VaR denotes the expected PD when the stress factors are at the lowest confidence level, such as there is only a 0.1 percent chance of a higher default rate, and $EAD*LGD*E(PD)$ denotes the expected value of loss[1]. Therefore, based on the internal ratings based (IRB) approach, to calculate the required capital for credit risk we may first calculate the three risk components. Under Pillar 1 Minimum Capital Requirements of Basel II, banks using internal models are required to stress test these risk components. In Pillar 2, the Supervisory Review Process of Basel II, stress testing is also required of banks so that regulators can review the banks' risk control and capital assessment (BCBS, 2004). Table 1.1 shows a summary of the credit risk stress testing requirements in Basel II.

Table 1.1 Credit risk stress testing requirements of the Basel II Accord

	Stress test requirements (BCBS, 2004)
Pillar 1	1. Banks using the internal ratings based approach must have meaningful and reasonably conservative stress testing programs that assess their capital adequacy and ability to withstand stress. 2. Stress tests conducted should identify potential events or economic condition changes including (i) economic or industry downturns; (ii) market-risk events; (iii) liquidity conditions that could have adverse effects on credit exposures. 3. Consider at least the effect of mild recession scenarios. For instance, use two consecutive quarters of zero growth to assess the effect on the bank's PDs, LGDs, and EADs, taking into account the bank's international diversification on a conservative basis.

[1] In some cases, such as for corporate loans, for banks using an advanced IRB approach, RWA is required to be adjusted by loan maturity using a function of effective maturity (M) as the default probability could be related to length of loan term. Then $RWA = 12.5*EAD*LGD*(VaR - E(PD))*M$. For banks using the foundation IRB approach, M is a constant value given by the Accord. For retail IRB capital calculation M is not required.

	4. Consider ratings migration under mild deterioration or a stressed environment. 5. National supervisors may guide the stress testing approaches.
Pillar 2	1. Rigorous, forward-looking stress tests are included in the overall capital adequacy evaluation process where the board and senior management's oversight; sound and comprehensive capital and risks assessment; monitoring and reporting, and an internal control review are featured. 2. Stress testing should be included in the internal control review feature of the said process to evaluate and ensure its soundness. Supervisors should consider the stress test when assessing the overall process. Supervisors should also consider the extent of extremeness of the stressed scenarios. The sophistication of stress testing methods and the range of scenarios used should be appropriate for the bank's activities. 3. Under Pillar 2, supervisors may wish to review the credit risk stress tests a bank performs under Pillar 1. They may require risk reduction or additional capital holding if they decide a bank has insufficient capital. 4. Credit concentration risk is not addressed in the capital charge for credit risk in Pillar 1. Under Pillar 2, stress tests of major credit risk concentrations should be conducted regularly to identify and respond to possible changes in market conditions that could lead to unfavourable impacts on the bank's performance. Supervisors should review the concentration risk stress tests.

The Basel III Accord was published in 2009 in response to the financial crisis. It seeks to enhance banks' resilience against system-wide shocks. It made no major revision or addition to Basel II in terms of the above stress testing requirements. However, it revised the requirements for the minimum capital, introduced countercyclical measures, leverage measures, and liquidity measures, etc., which are all efforts to strengthen the banks' ability to tackle finance stress.

The Basel Committee is raising the resilience of the banking sector against stress firstly by strengthening the regulatory capital framework. In Basel II, the minimum capital requirements include 3 tiers, whereas there are 2 tiers and 2 capital buffers in Basel III. Table 1.2 and Table 1.3 present the contents of the tiers, buffers, their contents and requirements of Basel II and Basel III.

Table 1.2 Contents and specifications of minimum capital requirements of Basel II

Tiers	Contents and specifications
Tier 1 (core capital)	• Include common stock + non-cumulative perpetual preferred stock + disclosed reserves from post-tax retained earnings. • $\geq$ 4% of risk weighted assets (RWA).
Tier 2 (supplementary capital)	• Include undisclosed reserves + asset revaluation reserves + general provisions + hybrid debt capital instruments + subordinated term debt with at least 5 years of original time to maturity. • The subordinated term debt component of Tier 2 capital cannot be more than 50% of Tier 1 capital. • Tier 2 capital $\leq$ Tier 1 capital.
Tier 3	• Include fully paid up unsecured short-term subordinated debt that has an original maturity of at least two years. • Only to meet market risk. • Can only be included with the approval of the national regulator. • Must be no more than 250% of a bank's Tier 1 capital that is allocated to cover market risks. • Tier 1 + Tier2 + Tier3 $\geq$ 8% RWA.

Table 1.3 Contents and specifications of minimum capital requirements of Basel III

Tiers and Buffers	Contents and specifications
Tier 1 Common Equity	• Common shares + stock surplus + retailed earnings + other comprehensive income and other disclosed reserves + regulatory adjustments. • ≥ 4.5% of RWA.
Capital conservation buffer	• In the form of Tier 1 Common Equity. • 2.5% of RWA. • The purpose is to maintain capital above the required minimum at normal conditions. • If Tier 1 Common equity falls below 7% (including the conservation buffer), a bank is required to retain a proportion or all of its earnings.
Countercyclical buffer	• In the form of Tier 1 Common Equity. • 0% - 2.5% of RWA. • The purpose is to take into account the macro-financial environment and protect banks from macroeconomic recessions and high system-wide risk. • The national authorities decide when to raise the countercyclical buffer and the size of the buffer.
Additional Tier 1 capital	• Instruments qualified for inclusion in additional Tier 1+ stock surplus from additional Tier 1 instruments + instruments issued by subsidiaries qualified for inclusion in additional Tier 1 + regulatory adjustments. • Tier 1 Common Equity (not including conservation buffer) + Additional Tier 1 capital ≥ 6.0% of RWA.
Tier 2 capital	• Instruments that are issued and paid in and subordinated to depositors and is callable only after 5 years + stock surplus resulting from instruments included in Tier 2 capital +

	instruments issued by subsidiaries qualified for inclusion in Tier 2 + certain loan loss provision + regulatory adjustments. • Tier 1 + Tier2 $\geq$ 8% RWA.

Compared to Basel II, Basel III raised both the quality and quantity of the regulatory capital. The types of capital are aggregated into tiers based on their liquidity. Tier 1 capital items are the easiest to liquidate and the most reliable type of capital as it can turn into cash without too much value reduction. Because the banks had insufficient capital during the financial crisis, Basel III raised the capital requirements especially the percentage of the Tier 1 capital and equity capital. For instance, the minimum of Tier 1 capital in Basel III is 6% of RWA compared to 4.5% in Basel II. Equity capital, such as common shares and retained earnings, is the most liquid and it is taken first should a crisis event happen. Since Basel III introduced a capital conservation buffer in the form of common equity of the amount of at least 2.5% RWA, it in effect raised the requirement for common equity Tier 1 to 7% RWA. The total capital required in Basel III including the capital conservation buffer is also higher than that of Basel II (10.5% RWA compared to 8% RWA). The Tier 3 of Basel II, which consists of unsecured short-term subordinated debt and only used to cover market risks, is removed from Basel III.

The Basel committee considers that banks' procyclical behaviour was one of the most destabilising elements during the crisis. In times of economic recession, loan defaults and losses increase. Banks then reduce their lending due to reduced asset value and capital. Fewer lendings then cause a reduction in aggregate demand which will intensify the economic downturn. Aggravated economic conditions then further reduce banks' equity value which forces banks to further reduce their exposure (BCBS, 2011). The vicious circle continues. Therefore as further tools to tackle financial stress, Basel III introduced the countercyclical buffer of up to 2.5% RWA specifically to deal with capital cyclicality in times of stress and system-wide risk. It also required banks to maintain a leverage ratio of at least 3%.

During the crisis, many banks had adequate capital levels. However, they did not manage their liquidity prudently. Furthermore, during stress periods, funding that was readily available in the prosperous period became especially difficult to obtain. Therefore those banks still experienced difficulties (BCBS, 2011). Therefore, apart from revising the minimum capital requirements and introducing the capital buffers, the Basel Committee also introduced global liquidity standards including the Liquidity Coverage Ratio (LCR) and the Net Stable Funding Ratio (NSFR).

The LCR is defined as the ratio of the high quality liquid asset amount (HQLA) to the total net cash flow amount, over 30 days. It aims to ensure that banks have liquid assets sufficient to fund cash outflows for at least 30 days of acute, short-term stress and liquidity disruptions comparable to the 2007 financial crisis (BCBS, 2011). The HQLA consists only of those that can easily and quickly be converted to cash, such as government-issued securities, high quality non-financial sector common stocks and corporate debts, etc. For each systematically important financial institution (SIFI), this value must be at least 100%:

$$LCR = \frac{HQLA}{Total\ net\ cash\ flow} \geq 100\%$$

The LCR standard improves financial institutions' short-term resilience against stress. The NSFR enhances their long-term resilience and reduces the chance of bank distress caused by stress in funding as it requires banks to use more stable funding on an ongoing basis. The NSFR is the ratio of a bank's available stable funding (ASF) to its required stable funding (RSF). ASF includes capital and liabilities that are still fully available in over 1 year. RSF is the quantity of stable funding a bank needs based on the liquidity of its assets. That is, illiquid assets require stable funding. NSFR must be at least 100%:

$$NSFR = \frac{ASF}{RSF} \geq 100\%$$

While internally testing the resilience of a risk exposure to adverse conditions, either based on the Basel Accords or using a banks' own methodologies, is not new, system-wide regulatory stress tests came into existence mainly after the recent financial crisis. Since the first US macro-prudential stress test, the Supervisory Capital Assessment Program (SCAP) in 2009, national regulators such as the Federal Reserve Board (FRB), the Bank of England (BoE) and the European Banking Authority (EBA) regularly conduct macro stress tests on groups of financial institutions above certain asset values whose assets make up a great percentage of the total assets of U.S., UK, and European banks. Table 1.4 shows a summary of the most recent macro stress tests conducted by the FRB, EBA, and BoE: Comprehensive Capital Analysis and Review (CCAR), EU-wide stress test, and UK banking system stress test, respectively.

Table 1.4 Summary of the most recent macro stress tests

	The minimum Capital requirement under stress	Number of Banks and inclusion criteria or total coverage	Scenarios	Result	Risk types Included
CCAR (2018)	T1C: 6.3% T1: 8.0% Total: 10.5 leverage: 5.0 [a]	• 35 US banks • total consolidated assets ≥ \$100 billion • 80% of the total assets of all U.S. financial companies	See DFAST 2018: 1.Baseline 2.adverse 3.severely adverse [b]	1 out of the 35 banks did not pass	Credit Market

EBA (2018)	T1C: 4.5 T1: 6.0 Total:8.0	• 48 banks from 15 EU and EEA countries • broadly 70% of total EU banking sector assets	1.baseline 2.adverse [c]	Predicted 3-year Credit risk losses in the adverse scenario: €358bn	Credit Market Operational Liquidity Model
BoE (2018)	T1C: 6.6% T1 leverage: 3.25%	• 7 UK banks • Covering 80% UK lending	1. scenarios more extreme than the financial crisis. 2. including paths for economic and financial market variables [d]	1.Sound UK banking system [e] 2.No banks need to strengthen their capital positions	Credit Market Operational

a: T1C: Common equity tier 1 capital ratio; T1: Tier 1 capital ratio; Total: Total capital ratio; T1 leverage: Tier 1 leverage ratio.

b: Dodd-Frank Act Stress Test (DFAST) 2018: Supervisory scenarios of 28 variables. 16 of them capture economic activity, asset prices, and interest rates in the U.S. economy and financial markets and 3 of them are (GDP growth, inflation, and the U.S./foreign currency exchange rate) for each of four countries/country blocks (FRB, 2018).

c. scenarios of a 3-year horizon with the end of 2017 data as the starting point. The adverse scenario identifies a set of systemic risks that may pose a threat to the financial stability of the EU banking sector and trigger specific shocks, including a growth in gross domestic product (GDP) in the EU of -1.2%, -2.2% and 0.7% as of 2018, 2019 and 2020 respectively, with a deviation of -8.3% from its baseline level as of the end of 2020 (EBA, 2018).

d: macroeconomic scenarios include that world, China and UK GDP falls by 2.4%, 1.2%, and 4.7% respectively; UK unemployment and Bank Rate rises to 9.5% and 4% respectively; UK

residential property prices, commercial real estate prices and exchange rate index fall by 33%, 40% and 27% respectively.
e: stress testing shows that the UK banking sector is resilient to deep simultaneous UK and international economic recessions more severe overall than the global financial crisis combined with large market and operational risks (BoE, 2018).

The macro stress testing programs currently in use and the stress testing requirements in the Basel Accords have been developed over several years. However, addressing model risk in stress tests is rather recent. For instance, the Bank of England (BoE, 2017, 2018) introduced model risk into stress testing in 2017, and the EBA (EBA, 2018) in 2018. The description of model risk in the stress testing programs is either relatively preliminary (BoE, 2017; 2018) or vague in how it is implemented (CCAR 2017, 2018; EBA, 2018). However, it is an important risk as it could lead to a significant loss. For instance, the influence of a model uncertainty shock totalled a €21bn loss in the EBA stress test of 2018 (EBA, 2018). The Basel III Accord requires a leverage ratio of at least 3% to constrain excess leverage and provide extra protection against model risk and measurement error (BCBS, 2011). It also introduced a multiplier of a constant value (1.25) to compensate for model risk in credit valuation adjustment risk calculation (BCBS, 2017), while it made no mention of specifically and individually addressing model risk in modelling other types of risks. We consider addressing model coefficient estimation risk in the stress testing process using a Bayesian approach which we believe could be a useful addition to the existing stress testing methodologies.

The European Central Bank uses a satellite Bayesian model design which is an autoregressive distributed lag model structure to translate the macroeconomic scenarios into various forms of risks faced by banks and uses a Bayesian model averaging approach to address model uncertainty (Henry & Kok, 2013). As the Bayesian model averaging approach combines a selection of the best performing models, it can make use of large quantities of variables and reduce the model uncertainty of using each individual model. Our stress testing method is similar to the satellite Bayesian stress testing analysis that the European Central Bank employs mainly in two ways. Firstly, the two methods both use the Bayesian approach for its benefits in accounting for model or parameter uncertainty. Secondly, the satellite Bayesian stress testing approach and our stress testing procedure are both models that translate presumed stress or non stressed scenarios into a path for the risk indicator of interest. The main difference between our stress testing method and the Bayesian satellite stress testing model that ECB employs is the type of risk that the two methods intend to reduce. The Bayesian satellite model uses a Bayesian model averaging approach to address the model uncertainty coming from imperfect data quality and short historical time series that is rather common for the explanatory variables for the banks' risk indicators (Henry & Kok, 2013). Some papers use this approach in stress testing. For instance, see Petropoulos et al. (2018), Siemsen and Vilsmeier (2018), etc. Our stress testing model uses a Bayesian posterior distribution approach and concentrates on the estimation uncertainty of model parameters.

The frequentist and Bayesian approaches are the two major approaches in econometrics. One major way in which the Bayesian approach differs from the frequentist approach is that prior information can be included in the Bayesian estimation but not in the frequentist approach. By adding available information into the modelling process, Bayesian approaches allow for reducing the risk of neglecting useful information. Another difference between the two approaches is that in Bayesian econometrics, the model parameters are random variables as opposed to fixed scalars in the frequentist approach. Treating the model parameters as random variables and including all ranges of possible parameter estimates addresses the coefficient uncertainty and allows for the reduction of coefficient estimation error. The motivations, aims, and contributions of this thesis are based on these two major differences between the frequentist and Bayesian approaches.

1.3 Motivation

In the literature of credit risk modelling, most papers use the frequentist approach to model the probability of default (Baesens et al., 2003; Brown & Mues, 2012; Lessmann et al., 2015). Among the papers that use a Bayesian approach, some employ non-informative priors (Bijak & Thomas, 2015; Miguéis, Benoit, & Van den Poel, 2013; Park, Amarchinta, & Grandhi, 2010). Some use subjective informative priors to include expert opinions (Jacobs & Kiefer, 2010; Kiefer, 2009). Some use a more objective way to elicit priors such as semi-automatic priors (Maltritz & Molchanov, 2008) and data-based informative priors (Bijak & Matuszyk, 2017). Mira and Tenconi (2004) find that Bayesian models with informative priors outperform frequentist models in predictive accuracies for predicting default probability.

In conventional stress testing practice, the estimation uncertainty of model coefficients is neglected since only the coefficient point estimates are used in the credit loss distribution simulations or point forecasts even though estimation risk, commonly expressed using coefficient standard errors, is present (Bellotti & Crook, 2013; Bikker & Hu, 2002; Breeden, 2016; Busch, Koziol, & Mitrovic, 2018; Laeven & Majnoni, 2003; Sorge & Virolainen, 2006). One gap in the literature is that assuming coefficients are fixed and omitting estimation errors in stress testing may underestimate credit loss and subsequently the capital needed since a source of risk is not considered.

In the stress testing literature, as in the method proposed by Berkowitz (1999), most papers use the same parameter estimates between the tranquil and stress scenarios either in the frequentist framework (Jacobs, 2018; Jokivuolle & Viren, 2013; Kapinos & Mitnik, 2016; Sorge & Virolainen, 2006; Wong, Choi, & Fong, 2008) or in the Bayesian framework (Louzis, 2017; Papadopoulos, Papadopoulos, & Sager, 2016; Petropoulos et al., 2018; Siemsen & Vilsmeier, 2018). Few papers consider structural breaks between scenarios in the stress testing literature, and only the frequentist methods are employed when they do (Jacobs & Sensenbrenner, 2018; Petropoulos et al., 2018; Tsukahara, Kimura, Sobreiro, & Zambrano, 2016). Therefore, one gap in the literature is that no prior research incorporates model parameter instability between normal and stressed scenarios into a Bayesian stress testing methodology which allows for the inclusion of coefficient uncertainty.

1.4 Aims

Considering the gaps in the literature listed above, this thesis has three major aims as follows:

- To include useful available information using the Bayesian approach with informative priors and improve model performance.

In classic econometric methods, only information contained in a sample of data is taken into account in the estimation, while other information that may potentially contribute to the estimation and prediction accuracy is neglected. Therefore we aim to use a new Bayesian informative prior selection method which employs time series forecasts of PD model coefficients as informative priors to reduce model risk of neglected information.

- To reduce estimation risk in stress testing by using the Bayesian coefficient posterior distribution as opposed to frequentist coefficient point estimates.

In conventional stress testing methods, coefficient estimation uncertainty is neglected since estimation errors of the model coefficients are not included in the stress testing procedure. In the Bayesian framework, the coefficients are variables instead of fixed values as in the frequentist method. We aim to reduce estimation risk and subsequently credit risk by using the Bayesian coefficient posterior distribution which includes all possible coefficient estimates instead of only the mean estimates in stress testing.

- To explore the effect that macroeconomic shocks have on stress testing models and the influence that changes in model parameters between scenarios have on stress testing results.

Most stress testing methods in the literature employ the same model with the same model coefficients for different scenarios in the credit loss simulation. However, macroeconomic stress could cause parameter changes between models built on different scenarios. We aim to include parameter instability between normal and stress scenarios into stress testing by building individual models for different scenarios. We also aim to incorporate this stress testing method into a Bayesian framework so that coefficient estimation risk could be taken into account.

1.5 Contributions

This thesis makes several contributions to the literature. In chapter 3, we propose a new Bayesian informative prior selection method. That is, we treat model parameters in the credit scoring models built on consecutive time periods as time series variables and forecast their values in future time periods using ARIMA models. We then use these parameter forecasts as priors in the informative Bayesian credit scoring models. This method of prior selection has the benefit firstly because it is data-based therefore relatively objective. Secondly, the prior selection is systematic as every coefficient prior is selected based on the same method. We show that this method of prior selection enhances predictive accuracy compared to both frequentist models and various other Bayesian models.

Second, in chapter 4 we propose a new Bayesian stress testing method that uses the Bayesian coefficient posterior distribution in the stress testing procedure instead of frequentist coefficient point estimates. In the default rate distribution simulation in this stress testing method, we not only simulate the macroeconomic scenarios but also randomly take draws from the coefficient posterior distribution to include other possible coefficient estimates so we do not ignore the estimation errors surrounding the coefficient mean estimates. Since additional sources of risks are included, more extreme estimates of losses are simulated and higher required capital is calculated. Therefore compared to using conventional stress testing methods which neglect estimation risk, using the stress testing method that we propose, banks would be more conservative and cautious and maintain higher levels of capital thus are safer against stressed conditions. Moreover, we give plausible stress testing results that avoid overestimating credit loss since the less likely coefficient estimates in the posterior are given less weight proportionate with posterior probabilities.

In chapter 5, we contribute in two ways. Firstly, we incorporate parameter instability between models built on crisis and non-crisis time periods into a Bayesian stress testing methodology. In contrast, most existing stress testing papers only explore changes in macroeconomic variables between stress and normal scenarios while ignoring changes in model parameters between the two scenarios. Secondly, the use of Bayesian estimation in our work allows the inclusion of estimation risk in a stress testing approach that addresses parameter instability by using a Bayesian coefficient posterior distribution as the source of coefficient estimates. In contrast, the few stress testing papers that do address parameter changes are only in the frequentist framework and use coefficient point estimates which ignore coefficient estimation risk.

1.6 Methods and main findings

This thesis consists of three main chapters (Chapter 3, 4 and 5) that employ different samples of U.S. mortgage loan data from the Freddie Mac database. In chapter 3, we estimate the probability of default of mortgage loans using cross-sectional logistic regression models. For all the samples, we model the probability of default using frequentist maximum likelihood estimation, Bayesian estimation with both non-informative priors and various informative priors, a Bayesian model averaging method, and a Bayesian model selection method. We take random samples of mortgage loans originated before, during, and after the financial crisis, respectively, as training samples.

Our main finding in chapter 3 is that our method of Bayesian prior selection using ARIMA forecasts of coefficients outperform all other models employed in chapter 3, frequentist or Bayesian, on all samples in terms of model predictive accuracies regardless of the time periods on which the models are built or the economic environments.

In Chapter 4, we use a discrete time hazard model to estimate the probability of default based on a panel dataset. The stress testing model is the latent variable interpretation of the logistic regression. We stress test the probability of default using the distribution approach. Both the frequentist method and the Bayesian method are employed in estimation and stress testing. We use the Bayesian approach so that the posterior distribution of the model parameters can be obtained. The coefficient posterior draws obtained in the Bayesian approach are subsequently applied to simulate the Bayesian estimated default rate (DR) distribution. The Bayesian simulated DR distribution is then compared with the simulated DR distribution obtained using a frequentist approach with coefficient point estimates.

Our main finding in chapter 4 is that the 99% VaR of the simulated default rate distribution obtained using a Bayesian approach with a coefficient posterior distribution is around 6.5 times as large as the VaR at the same probability level using the frequentist approach with coefficient mean estimates. It shows that neglecting coefficient uncertainty in stress testing may significantly underestimate credit loss and the capital required.

In Chapter 5, using the stress testing method proposed in Chapter 4, we model and stress test the probability of default separately for the financial crisis and post-crisis tranquil time periods to study the influence that coefficient changes between scenarios have on stress testing results. We use both frequentist and Bayesian stress testing methods to study the influence that coefficient uncertainty has on stress testing results given the same scenario.

Our main finding in chapter 5 is that without estimation risk, the default rate distribution that we simulate based on a model built on crisis period data have higher VaRs and variance than that built for the tranquil period. For models built on the same scenario, the default rate distributions with estimation risk included, using Bayesian methods, have larger variances and VaRs than the distributions obtained using the frequentist point estimates approach without considering estimation risk. Both estimation risk and macro shocks on model parameters cause the simulated DR distributions to be more spread-out, with a combined influence increasing the required capital by around 170% compared to the baseline distribution which incorporates neither risk.

1.7 Outline structure

Chapter 1 has been an overview of the thesis and has outlined its motivations, aims, contributions, methods, and main findings. The structure of the rest of the thesis is outlined as follows.

Chapter 2 is a literature review for the thesis which includes two main sections: different classification techniques for PD modelling and the literature review on stress testing. In the first section, we summarise the individual classifiers and ensemble methods commonly used in the credit scoring literature. We then review the main performance measures for the classifiers and compare the classification algorithms based on the model performance results in the literature. In the second section, we review the literature on credit risk stress testing.

Chapter 3 introduces the methodology used in this chapter including the logistic regression model and ARIMA model, frequentist and Bayesian estimation methods, and different prior selection methods including using ARIMA forecasts of model coefficients as informative priors. We then present the data and variables description. The estimation results using the frequentist and Bayesian approaches as well as model averaging and selection results are presented. We also give the post estimation convergence diagnostics. We then show the model performance comparisons for the frequentist and Bayesian methods trained on all samples.

In chapter 4, we outline the methodology employed in this chapter including the discrete time hazard analysis for probability of default modelling. It also defines the coefficient estimation risk and proposes a Bayesian method to address it in stress testing. It then explains the frequentist and Bayesian stress testing models and procedures. After describing the data and variables used in chapter 4, the estimation, prediction, and performance results are presented. We then compare the stress testing results when including and omitting estimation risk.

In Chapter 5, we first outline our stress testing method that addresses both parameter changes between models built on stressed and normal scenarios and estimation risk. We then elaborate on the stress testing models and simulation methods with the frequentist coefficient point estimates and the Bayesian posterior distribution. Subsequently, we describe the data and variables used in chapter 5. We then present the estimation, performance, and stress testing results and discuss their implications.

Chapter 6 is the concluding chapter. We review the objectives of the thesis and summarise the main findings and contributions. Policy implications, limitations, and future work are also suggested.

Chapter 2

Literature Review

2.1 Introduction

In this chapter, we firstly review the state of the art classification algorithms for modelling the probability of default and various model performance measures in the literature of credit scoring. On comparing model performances of the most popular individual and ensemble classifiers, we find that support vector machines, neural networks, and logistic regression are the best performing models. Ensemble classifiers generally have good model performances; however there is no consensus regarding the best ensembles. We find that compared to the frequentist approach, the use of Bayesian estimation in PD modelling is relatively underdeveloped in the literature. We consider the benefit of Bayesian informative priors in model performance improvement and model risk reduction is an interesting point of future research. Secondly, we review the literature on credit risk stress testing. We summarise the stress testing literature according to the scale of the tests, stress testing methods used, and the risk indicators being tested. We find that papers differ vastly in terms of the stress testing models used, the scenarios built, the risk exposures being tested, etc., but they consistently employ point estimates of model coefficients, while little attention is paid to the role coefficient estimation uncertainty plays in stress testing. Since a source of risk is neglected, the predicted loss may be underestimated in the stress testing literature. As model coefficients are random variables in the Bayesian method as opposed to fixed values in its frequentist counterpart, we consider using the Bayesian approach may be an effective way to introduce estimation risk into stress testing.

2.2 Literature review on credit scoring

In this section, we first summarise the state of the art classification algorithms in credit scoring and compare their performances, and then we review the papers that use Bayesian estimation in modelling the probability of default.

2.2.1 Classification techniques for PD modelling

In this literature review, we focus on modelling PD using classification methods which usually estimate the probability that an account will default. Papers using classification analysis differ in the classification algorithms they employ and in the performance measures of the classifiers.

In our literature review, we follow Lessmann et al. (2015) in categorising different classification models into individual models and ensembles. They group all classifiers into three categories, namely individual classifiers, homogenous ensembles, and heterogeneous ensembles. The individual classifiers use a single model to assess default probability. The latter two combine multiple classifiers to increase predictive accuracy (Finlay, 2011). The main difference between the latter two is that homogenous ensemble classifiers use the same classification algorithm to develop base models and combine their predictions, whereas heterogeneous ensemble classifiers use different classification algorithms for the base models. Unlike homogeneous ensemble classifiers, heterogeneous ensemble classifiers can also search through base models and select an appropriate subset of models. Individual classifiers and homogenous ensemble classifiers can be seen as special cases of heterogeneous ensembles (Lessmann et al., 2015). They obtain mixed results in comparing different types of algorithms. They find that some advanced algorithms especially heterogeneous ensembles outperform simple ones. However, they also argue that advanced methods do not necessarily improve accuracy since they find that simple classifiers such as logistic regression outperform some of the more sophisticated ones such as dynamic ensemble selection.

Some of the most popular individual classifiers in credit scoring include linear/quadratic discriminant regression analysis, logistic regression analysis, decision trees, neural networks, support vector machines, Bayesian networks, k nearest neighbours, mathematical programming, multivariate adaptive regression splines, etc. (Abdou, Pointon, & El-Masry, 2008; Abdou, 2009; Akkoç, 2012; Baesens et al., 2003; Bellotti & Crook, 2009; Brown & Mues, 2012; Chen, Ma, & Ma, 2009; Crone, Lessmann, & Stahlbock, 2006; Finlay, 2011; Hens & Tiwari, 2012; Sinha & May, 2004; Tsai, 2014). Most widely employed ensembles include bagging, boosting, stacking, random forest, random rotation, dynamic ensemble selection, etc. (Abellán & Mantas, 2014; Finlay, 2011; Ko, Sabourin, & Britto Jr, 2008; Kruppa, Schwarz, Arminger, & Ziegler, 2013; Marqués et al., 2012a; Partalas, Tsoumakas, & Vlahavas, 2009, 2010; Tsai, 2014; Tsai & Wu, 2008; Wang, Hao, Ma, & Jiang, 2011; Wang & Ma, 2012; Zhang, Zhou, Leung, & Zheng, 2010). Below we briefly outline the most widely discussed classifiers in the literature on credit scoring.

2.2.1.1 Individual classifiers

1. Linear Discriminant analysis

Fisher (1936) proposed linear discriminant analysis as a technique for classification and discrimination. Linear discriminant analysis seeks to distinguish amongst different classes and to predict which class new cases belong to. The LDA model is: $z_i = \boldsymbol{x}_i^{'}\boldsymbol{b}$, where z_i is the discriminant score and $\boldsymbol{x}_i$ is a vector of explanatory variables, and $\boldsymbol{b}$ denotes a vector of coefficients. The coefficients vector $\boldsymbol{b}$ is estimated such that for discriminant scores, the difference of the means between classes is maximised in relation to the variances within classes. This method assumes that the independent variables conditional on all classes are multivariate normal and have the same covariance matrix. (Abdou et al., 2008; Brown & Mues, 2012; Sinha & May, 2004).

Linear discriminant analysis is one of the earliest and most commonly employed statistical algorithms in credit scoring. It can narrow down the list of attributes to an appropriate combination in a fast manner (Thomas et al., 2017). However, the assumptions of a linear relationship between dependent and explanatory variables, of equal covariance matrices, as well as the assumption of a multivariate normal distribution are often criticised. The covariance matrices of different classes are believed to be considerably different. In theory, quadratic discriminant analysis could be used when the equal covariance matrix assumption is violated; however this classifier is reported to be actually more sensitive to this assumption while LDA is more robust (Dillon & Goldstein, 1984; Lee & Chen, 2005; Thomas et al., 2017).

2. Logistic regression

Harrell and Lee (1985) found that logistic regression is as accurate as linear discriminant analysis. It is widely used in two-category classifications (Sinha & May, 2004). In logistic regression, the logarithm of the odds ratio of a dichotomous outcome is dependent upon a set of predictors: $\log[p_i/(1-p_i)] = \boldsymbol{x'}_i\boldsymbol{\beta}$, where p_i is the probability of the event of interest for the individual i (e.g. the probability of default in the case of PD modelling). $\boldsymbol{x}_i$ is the vector of explanatory variables while $\boldsymbol{\beta}$ is the vector of coefficients. The logistic regression can also be written as $p_i = \frac{1}{1+e^{-x_i'\beta}}$. The estimator computes coefficients $\boldsymbol{\beta}$ such that the probability of the observed outcome is maximised. The ordinary least squares method can be used to estimate a linear discriminant function while maximum likelihood estimation is commonly employed to estimate the coefficients in logistic models.

Like many other parametric models such as linear discriminative analysis and probit regression, logistic regression allows the use of statistical tools such as confidence intervals and hypothesis testing. Therefore different features' relative importance and the model's discriminating power are easily interpretable (Thomas et al., 2017). Another advantage is that unlike linear discriminant analysis, neither the multivariate normal assumption nor the equality of covariance matrices is required in logistic regression. However a linear relationship is still assumed between the logit and the independent variables. Unlike some classifiers such as multivariate adaptive regression splines (MARS) that model interactions between variables in an automatic fashion, variable interactions generally have to be modelled manually when using logistic regressions.

3. Probit regression

Probit analysis was first proposed in the 1930s (Abdou, 2009). Like logit analysis, probit analysis allows the probability of a dichotomous variable to be estimated. The linear combination of explanatory variables is transformed into the dichotomous dependent variable's cumulative probability using an inverse normal distribution as the link function. In the estimation of a probit model, the coefficients of the explanatory variables are estimated such that the probability of the binary dependent variable being one equals the cumulative normal density: $prob(y_i = 1 | \boldsymbol{x}_i) = \Phi(\boldsymbol{x'}_i \boldsymbol{\beta})$, in which y is the 0-1 dichotomous dependent variable; Φ is the cumulative distribution function of a normal distribution. $\boldsymbol{x'}_i \boldsymbol{\beta}$ is as above.

Probit analysis is similar to logistic regression in that they are both generalised linear models that model categorical outcome against predictors using a link function. It is an alternative to logistic regression and is often used in comparison with other classifiers (Abdou et al., 2008). As is logistic regression, probit analysis is also criticised for its linear setting and relatively low accuracy compared to some nonlinear models.

The main difference between the two is the error distribution assumed. In logistic regression, a logistic error distribution is assumed and the logistic function is used to model the default probability. In probit analysis, a standard normal residual distribution is assumed and the probit function is used. The coefficients in logistic regression are interpreted as a log odds ratio or odds ratio after exponentiation. The coefficients in probit models, on the other hand, are interpreted as the differences in Z scores for a standard normal distribution associated with differences in explanatory variables, which do not have a natural intuitive interpretation like odds ratio. Secondly, the logistic distribution has heavier tails than a normal distribution. Greene (2011) considers that logistic regression gives a higher (lower) probability for event happening when the linear combination of attributes are extremely small (large). Hahn and Soyer (2005) consider that the logistic model fits better than the probit model for cases of extreme independent variable values. They both also argue that in most cases both models fit data equally well and provide similar probability (Greene, 2011; Hahn & Soyer, 2005). Greene (2011) also considers that for practical grounds one may favour one model over the other, but theoretical justification is difficult and generally unresolved. Another difference lies in the probability estimation process. In logistic and probit regression, the probability of an event is a cumulative logistic distribution and cumulative normal distribution function respectively. The former cumulative distribution function (CDF) is an integral that has a simple closed form whereas the latter CDF does not have a closed form. Since the normal distribution function cannot be easily integrated and has no closed form probability expression, simulation is typically required in estimation when using a probit model (Train, 2009).

4. k-nearest neighbours algorithm (k-NN)

k-NN is a data mining algorithm that uses a distance measure to predict the outcome of a new data point. In this method, the new case is classified to the majority class among its k nearest cases within the training sample. Therefore we need to measure the distance between cases and find the ones that are closest to the case of interest. One commonly used distance measure in the literature is the Euclidean distance measure: $dist(i,j) = \| \boldsymbol{x}_i - \boldsymbol{x}_j \| = [(\boldsymbol{x}_i - \boldsymbol{x}_j)'(\boldsymbol{x}_i - \boldsymbol{x}_j)]^{\frac{1}{2}}$ where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ represents the vector of input variables for cases i and j. The probability of the new case belonging to a class is estimated as the ratio of the number of cases in the majority class over k (Baesens et al., 2003; Bellotti & Crook, 2009; Brown & Mues, 2012).

K-NN method has the advantage that the credit scoring system can be easily updated by adding new cases with known classification outcomes and deleting the oldest cases in the training sample. However, it is difficult to find an appropriate distance measure and to evaluate the performance of the classifier. Unlike the regression methods, this algorithm shares the disadvantage with decision tree methods that it does not provide credit scores for the loan applications (Thomas et al., 2017).

5. Neural networks

The neural network (NN) algorithm is inspired by biological nervous systems in terms of information processing. The most popular NN algorithm, the multilayer perceptron neural network, consists of inputs layers, hidden layers and output layers in its information processing structure. Each of the layers has several neurons. Information fed into the network is processed in the neurons of one layer, and the output is sent to the subsequent layer for further processing. The output a neuron transmits to the next layer is represented as: $h_i = f(b_i + \sum_{j=1}^{n} w_j x_j)$ where h_i is the output that neuron i sent to the next layer; b_i stands for bias input comparable to the intercept in a regression model; x_j is the j th input; w_j denotes the weight connecting the j th unit and the neuron. Unlike linear discriminant analysis and logistic regression, with the transfer function f , neural networks can handle both linear and non-linear relationships between the dependent and explanatory variables.

Neural networks are reported by some literature to have higher accuracy than some statistic methods such as linear discriminant analysis (Lee & Chen, 2005). However, it is often criticised that due to its black-box nature it is not as comprehensible as the linear models. For instance, the relative importance of different input variables is difficult to identify. Therefore it can be hard to explain to customers why they are rejected. The long training process is another weakness. For instance, Leong (2016) finds that although a neural network has the better overall predictive performance than a Bayesian network and a logistic regression, it is approximately 50-500 times and 20-150 times more time consuming to build the model than a Bayesian network and a logistic regression, respectively.

6. Bayesian Networks

A naive Bayes classifier learns the probability of each attribute x_k conditional on a given class y_i. Then using Bayes rule we can compute the probability of each class given all attributes. A new data point is assigned to the class that generates the highest posterior probability of class membership. In Naive Bayes, it is assumed that all attributes are independent conditional on the class: $p(\boldsymbol{x} \mid y_i) = \prod_{k=1}^{n} p(x_k \mid y_i)$. Therefore the posterior probability of a case being class y_i is: $p(y_i \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid y_i) p(y_i) = \prod_{k=1}^{n} p(x_k \mid y_i) p(y_i)$.

The assumption that all the inputs are conditionally independent of each other permits the use of the Bayesian rule of probability. A criticism of this method is that this assumption is not true in practice. However, this independence assumption works well for most cases, even if in actuality the variables are not really independent (Twala, 2010).

The naive Bayes method can be seen as a special case of the Bayesian network method (Thomas et al., 2017). Bayesian networks make few assumptions of independent attributes, and the dependence between attributes in a Bayesian Network has to be modelled. Methods such as tree augmented naive Bayes relaxes the assumption of independent attributes (Baesens et al., 2003; Friedman, Geiger, & Goldszmidt, 1997).

Bayesian networks and Bayesian inference are closely related concepts. Bayesian networks are classification models based on Bayes theorem. The modelling process of Bayesian networks consists of two parts: a network structure in which the relationship of the attributes can be represented by a diagram; and a learning algorithm to estimate the attributes' conditional probabilities (Thomas et al., 2017). For simple Bayesian networks such as Naive Bayes where attribute independence is assumed, the probability of a case's class membership is a product of the conditional probability of the attributes. For Bayesian networks in which the dependence of attributes is taken into account, learning algorithms based on maximum likelihood are commonly used.

Bayesian inference is also based on Bayes theorem. The fundamental interest is to learn the distributions or the characteristics of the distributions of the *parameters* in the model, conditional on the data, i.e. $p(\theta \mid y)$ and $E(\theta \mid y)$. It parallels frequentist econometrics methods such as ordinary least squares, maximum likelihood, etc., which can be used to estimate parameters in classifiers such as discriminant analysis and logistic regression, etc.

7. Support vector machines

A support vector machine (SVM) is a classifier that separates two groups of data points with a hyperplane such that the distance between the cases and the hyperplane is maximised (Bellotti & Crook, 2009; Crone et al., 2006).

The binary classes y_i take the value of either -1 or +1. To separate the two classes as much as possible on the separating surface, the two classes satisfy the following condition: $y_i[\boldsymbol{x'}_i \boldsymbol{w} + b] \geq 1, \quad i = 1,...n.$ The border between class -1 and class +1 is the hyperplane: $\boldsymbol{x'}_i \boldsymbol{w} + b = 0$

To maximise the margin width and minimise the misclassification, SVM can be written as an optimisation problem:

$$\min_{w,b,\xi} \frac{\boldsymbol{w}'\boldsymbol{w}}{2} + C\sum_{i=1}^{n}\xi_i$$
$$s.t.: y_i(\boldsymbol{x}'_i\boldsymbol{w} + b) \geq 1 - \xi_i, \quad i = 1,...,n.$$
$$\xi_i \geq 0, \quad i = 1,...,n.$$

where $y_i \in \{-1,1\}$, ξ_i is a slack variable to take into account classification errors; C is a tuning parameter to control the trade-off between margin maximisation and classification error minimisation (the first and second term of the objective respectively). The coefficients $\boldsymbol{w}^*$ and b^* obtained from the optimisation define the location of the hyperplane, thus deciding the classification of cases. The classifier takes the form: $y_i = \text{sgn}(\boldsymbol{x}'_i\boldsymbol{w}^* + b^*)$. The hyperplane is linear in this case. A non-linear hyperplane to separate data points can be obtained when the input vector $\boldsymbol{x}_i$ is mapped into a high dimensional feature space using a nonlinear function.

Support vector machine models are generally reported to have good classification performances. However, their performances are highly influenced by the kernel parameters setting, sample settings, the algorithm for the quadratic programming and different classes' importance (Yu, Yao, Wang, & Lai, 2011; Zhou, Lai, & Yu, 2010). The problem of optimal input features selection is also a common one (Yao & Lu, 2011). They also share with neural networks the disadvantage of being very difficult to interpret due to the complex and high dimensional relationship being estimated (Thomas et al., 2017).

8. Linear Programming

Linear programming (LP) in credit scoring is an optimisation problem to minimise the sum of classification errors over all cases (Baesens et al., 2003) under the conditions that the linear combination of attributes for good (bad) cases are above (below) a cutoff:

$$\min_{\beta,\varepsilon} \sum_{i=1}^{n} \varepsilon_i$$

$$s.t.$$

$$\begin{cases} y_i(\boldsymbol{x}_i^{'}\boldsymbol{\beta} - c) \geq -\varepsilon_i & y_i \in \{-1,1\}, \\ \varepsilon_i \geq 0, \quad i = 1,...,n \end{cases}$$

in which ε_i is the misclassification error for case i, and c is the cutoff pre-specified by the modeller (Baesens et al., 2003; Crook, Edelman, & Thomas, 2007; Thomas et al., 2017). Integer programming is a related classification approach in which the objective is to minimise the sum of misclassification cost instead of misclassification errors as in linear programming.

Baesens et al. (2003) consider linear programming has the advantage that priori knowledge can be included easily to the model by adding extra constraints. Thomas et al. (2017) consider the lack of statistical support as a disadvantage against the linear programming method. Since linear programming is not a statistical method, the results of coefficient estimates and hypothesis testing of variable significance cannot be obtained directly. Unlike regression methods that choose characteristics based on their discriminating power, using linear programming often relies on resampling techniques for parameter estimation and characteristics selection, and it involves solving optimisation problems numerous times.

9. Multivariate adaptive regression splines

Multivariate adaptive regression splines (Friedman, 1991) can be seen as an extension of linear regression analysis in that it models nonlinearity and variable interactions. MARS takes the form $\hat{y} = \sum_{i=1}^{k} b_i A(x_i)$ in which $A(x_i)$ is a basis functions that can be a constant, a hinge function (i.e. $\max(0, x_i - a)$ or $\max(0, a - x_i)$ where a is a value selected as a knot for variable x_i), or a product of multiple hinge functions. A MARS model selects variables and the knots in hinge functions, and the combination of hinge functions automatically to maximise the goodness of fit and model parsimony. It separates data into regions and treats them differently with different slopes to create piecewise linear functions. These piecewise functions are also further multiplied together to model nonlinearity and interactions between variables. The model building process of MARS starts by repeatedly adding basis functions that reduce residual sum of squares the most, until adding new terms do not substantially reduce residual errors. A subset of the added basis functions is then selected to give both high goodness of fit and model parsimony.

One advantage of MARS is that it is more flexible than linear models such as LDA and LR since it models non-linear relationships. Compared to the black-box models such as neural networks it is easy to interpret since it can identify potential explanatory variables' relative importance. In addition, training and prediction are not as time consuming as models such as neural networks and support vector machines therefore it can handle large datasets and make predictions rather quickly. However, since MARS is a nonparametric method, unlike in linear models, parameter checks such as confidence intervals cannot be directly calculated.

10. Decision Trees

A decision tree is a classifier that uses a flowchart-like diagram to split the observations of characteristics in order to separate customers into different classes, repeatedly, until the subset of a node is all of the same class or further splitting does not increase prediction accuracy. The data are of the form $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k, \boldsymbol{y})$ in which $\boldsymbol{y}$ is the target variable that we would like to classify; $\boldsymbol{x}_k$ are the explanatory variables whose values are split recursively to put cases into different target variable groups. Algorithms that measure homogeneity within subsets of a node are used in each branch of the decision tree to measure the classification performance of the target variable in order to decide the best split of the observations of characteristics.

There are two main types of decision trees: classification trees and regression trees. In classification trees the outcome is discrete, whereas in regression trees the target variable can be continuous.

Decision trees as well as neural networks, and support vector machines are considered as some of the best individual classifiers among the data mining methods in terms of model performances. Like statistical classifiers, decision trees have good interpretability (Thomas et al., 2017). Unlike statistical classifiers, they have the advantages of automatic attributes selection and insensitivity to monotone transformation of attributes. However, decision trees such as classification and regression tree (CART) are criticised for their high variance. In particular, minor changes in the data can yield completely different trees, model interpretations, and predictions (Kruppa et al., 2013; Thomas et al., 2017).

2.2.1.2 Ensemble and hybrid methods

Ensemble learning is a machine learning method that combines a set of individual classifiers to solve the same problem. Individual classifiers employed to form the ensemble are called base learners (Polikar, 2006; Wang et al., 2011). In classifier ensembles, multiple learning algorithms are employed. For instance, homogeneous models trained on different training datasets through modifying datasets or generating new training datasets (Nanni & Lumini, 2009), different base learners with different subsets of the features in each base learner (Brown & Mues, 2012) trained on the same datasets, etc. The outputs are then further combined to give a data point an overall classification. The errors made by each individual learner on different areas of the input space complement each other through model combination (Tsai, 2014). In this section we summarise some of the most popular ensembles.

One of the early and commonly used ensemble algorithms is bootstrap aggregating (Bagging). In a bagging algorithm, subsets of the training sample are drawn and each of them is used to train a base learner. The results of individual learns are then combined using voting or averaging. A voting strategy is commonly used in classification problems. In majority voting, classification of a case into good or bad is based on the majority classification results from the base learners. In weighted voting, a weight based on the base learner’s accuracy is attached to each base learner’s decision. Weighted averaging is commonly used for regression problems in which the ensemble decision is an average of base learners’ output with weights based on their performance. The Bagging method is particularly useful when the data size is limited (Wang et al., 2011). A Bag decision tree and bag MLP are bagging ensembles that have decision trees and multilayer perceptron neural networks as base learners, with some changes from the traditional bagging, such as including features sampling instead of using the full set of features in training each base learner (Zhang et al., 2010).

Boosting is a method that increases classification accuracy by reweighting cases in training samples iteratively. After each round of training using one base learner, the classification error is computed and the weight for the correctly (incorrectly) classified cases will be lowered (increased) in the sample for the next round of training. Another base learner is then used to train the reweighted training sample. After a prescribed number of iterations, the ensemble is formed as a weighted average of the base learners, the weight being the model performance of each base learner.

Stacking is a two-stage ensemble algorithm. Different base learners are used to train the first stage training sample. These learners are then tested on a test sample. The predictions the base learners give along with their true classes make a new sample with meta-instances, also called second-stage training sample. A second stage learner is then trained on this sample and gives the final output (Partalas et al., 2009).

Bagging, boosting, and stacking are some of the most popular ensemble methods in the literature. Numerous other ensembles have also been developed and employed. For instance, Nanni and Lumini (2009) utilise a random subspace ensemble, in which each base learner randomly selects a subset of all characteristics for training and testing. Then the classifiers are combined to build the ensemble. A random forest is a kind of random subspace ensemble with decision trees as base learners (Ho, 1995). It can also incorporate bagging in that the trees are trained on bootstrap samples of the training data (Breiman, 2001; Brown & Mues, 2012). In a rotation forest ensemble, every decision tree base learner is trained on a random subset of the input variables after principal component analysis is first applied (Nanni & Lumini, 2009). In each base learner of the stochastic gradient boosting ensemble, an error term from the previous base learner is minimised, so that the predictive accuracy is incrementally improved (Friedman, 2002).

Hybrid is a method that combines multiple machine learning techniques sequentially (Tsai, 2014). For some artificial intelligence models such as neural networks and support vector machines, it is difficult to choose the best features or interpret their relative importance. Lee and Chen (2005) argue that there is not a theoretical method to optimally choose input nodes in neural networks. They consider a hybrid model can be used to solve this input selection problem. Specifically, models without black box characteristics can be first utilised to select the optimal features thus making the effect of each predictor and their relative importance interpretable. Secondly, these significant variables obtained from the first step can then serve as the input nodes of a neural network (Akkoç, 2012; Lee & Chen, 2005).

2.2.2 Performance of classification techniques

2.2.2.1 Percentage correctly classified, misclassification cost, ROC, AUC, GINI, K-S statistic

Different evaluation criteria are used to measure classifier performance. One of the most commonly employed criteria is the percentage correctly classified (PCC) which is the proportion of cases in a dataset that is correctly classified (Abdou, 2009; Paliwal

& Kumar, 2009). A related measure are the type I and type II error rates which are the ratio of good cases predicted as bad over the number of actual good cases, and the ratio of bad cases considered as good over the number of actual bad cases, respectively. The above concepts can be explained in a classification matrix - 'confusion matrix' in table 2.1.

Table 2.1 an example of a confusion matrix

		Predicted		
		Good (g)	Bad (b)	Total
actual	Good (G)	Gg	Gb	TG
	Bad (B)	Bg	Bb	TB
	Total	Tg	Tb	T

(Note: Gg : actual good cases predicted as good; Gb : actual good cases predicted as bad; Bg : actual bad cases predicted as good; Bb : actual bad cases predicted as bad; TG : the total number of actual good cases; TB : the total number of actual bad cases; Tg : the total number of predicted good cases; Tb : the total number of predicted bad cases)

PCC= $\frac{Gg+Bb}{T}$

Type I error rate= $\frac{Gb}{TG}$

Type II error rate= $\frac{Bg}{TB}$

Estimated misclassification cost (EMC) incorporates the costs of making type I and type II errors into the model performance measure. EMC takes the form: $EMC=\frac{Gb}{TG}\pi_1 c_I+\frac{Bg}{TB}\pi_2 c_{II}$, in which c_I and c_{II} are the misclassification costs for type I and II errors respectively; $\frac{Gb}{TG}$ and $\frac{Bg}{TB}$ are type I and II errors; π_1 and π_2 are prior probabilities of good and bad loans respectively. The prior probabilities are taken into account because the prevalence of good and bad loans are rarely equal. This method also takes into account the belief that the costs related to type I and II errors are significantly different. However, this method is not as widely used since reliable and consistent estimates for misclassification costs of type I and II errors can be difficult to obtain (Lee & Chen, 2005; Abdou, 2009).

The concepts of sensitivity and specificity are also derived from the confusion matrix. Sensitivity is the true positive rate $\frac{Bb}{TB}$, while specificity is the true negative rate $\frac{Gg}{TG}$. The Receiver Operating Characteristic (ROC) curve is a plot of a classifier's sensitivity (true positive rate $\frac{Bb}{TB}$) against 1-specificity (false positive rate $\frac{Gb}{TG}$) at different cutoffs. Each point on the curve represents the model's true/false positive rate when using a given cutoff probability. For instance, the (0,0) point corresponds to a model that predicts every data point as good at a threshold, while the point (1,1) represents that it always predicts data points as bad. The upper left corner corresponds to classifying every good case as good and every bad case as bad. The closer a point in ROC is to the upper left of the ROC space, the better the classification is. The diagonal line corresponds to random guessing. Figure 2.1 shows a typical ROC curve.

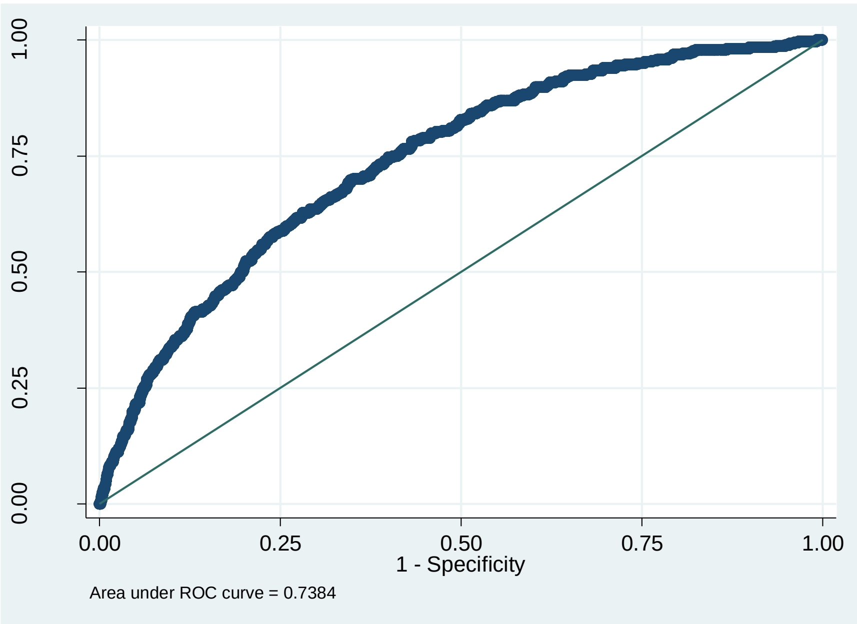

Fig. 2.1 An example of an ROC curve

The area under the curve (AUC) measures the overall performance of the classifier across the whole spectrum of decision cutoffs. As its name suggests, it is the integral of the ROC curve $f(x)$ between (0, 1): $\int_0^1 f(x)dx$. Its value is typically between 0.5 and 1 if the model performance is better than random guessing. A related measure is the Gini coefficient which is the area between the ROC curve and the diagonal line compared to the area of half a standardised square, and it also equals to 2AUC-1: $\frac{\int_0^1 f(x)dx - 0.5}{0.5} = \frac{2\int_0^1 f(x)dx - 1}{1} = 2AUC - 1$. AUCH measure, which is the area under the convex hull of a ROC curve, is another common performance measure based on the ROC curve. AUCH is equal to or larger than AUC as the ROC curve may have concave regions. Another popular performance measure related to the ROC curve is the Kolmogorov–Smirnov (K-S) statistic. Among the whole ROC curve, the point that has the maximum vertical distance from the diagonal line (i.e., $f(x) - x$) has the threshold that best separates good cases from bad ones. This distance is the K-S statistic.

The Percentage Correctly Classified measure gives accuracy at one specific cutoff. This method is criticised for strong bias regarding imbalanced data and proportions of prediction correctness of good and bad cases (Marqués, García, & Sánchez, 2012b). In contrast, the ROC related measures such as AUC and Gini are not restricted to a specific cutoff. They illustrate model performance across the full range of cutoffs and measure the aggregate performance of classification models (Sinha & May 2014). Therefore these measures are not influenced by class distribution in data (Marqués et al., 2012b). On the other hand, AUC also has its disadvantages. Kruppa et al. (2013); Kruppa, Ziegler, and König (2012) argue that a minimal increase in AUC could mean a considerable improvement in prediction at a certain cutoff. That is, it is based on the cases correctly classified in the data, but is not a function based on the observed and predicted probability at an individual level (Janes & Pepe, 2008; Kruppa et al., 2013). Hand (2009) also argues that with AUC performance measure the misclassification cost is different for different classifiers when misclassification cost should be independent of the classifier chosen. They proposed the H measure, an alternative performance measure free from this problem as it assumes the cost weight distribution to be the same between classifiers. The larger the values of H measure, the better the model performance, similar to other performance measures.

A Brier score is a performance measure to evaluate predicted probability. It is often utilised in medical applications and weather forecasting (Kruppa et al., 2013). A Brier score is an average of all individuals' squared difference between actual and predicted probabilities. A Brier score takes the form:

$BS = \frac{\sum_{i=1}^{N}(y_i - \hat{p}(y_i = 1 | \boldsymbol{x}_i))^2}{N}$, in which $\hat{p}(y_i = 1 | \boldsymbol{x}_i)$ is the predicted probability of default for data point $i, i = 1, ..., N$ given input features $\boldsymbol{x}_i$.

2.2.2.2 Comparison of algorithms based on model performances

To compare model performances of the most commonly used individual classifiers and ensembles in the literature, we record their model performances based on two of the most popular performance measures: PCC and AUC. Table 2.2 – Table 2.3 and Table 2.4 – Table 2.5 show the model performances of individual classifiers and ensembles, respectively. For each paper, the classifier with the greatest accuracy is highlighted in bold.

We calculate and record model performances in the literature in the following way:

1. Data: for each algorithm if a paper only uses one dataset and one sub-model of the model family, or homogeneous ensembles with only one set of base learners, we record it as the paper shows. If the classifiers are based on multiple datasets, we follow Lee and Chen (2005); Wang et al. (2011) to calculate the average accuracy for each model across datasets.

2. Cut-offs: When using PCC as a performance measure, some papers report results from multiple cut-offs. After averaging over datasets, we report the result from the cut-off that gives the best performance.

3. Sub-model: some papers train several sub-models in each type of classifier. Specifically, many papers use different decision trees, neural networks, and SVM models with different kernels, etc. In such cases, we report the best performance within a classifier family.

4. For papers using homogeneous ensembles with multiple sets of base learners, we report the best performance the ensemble obtains.

Table 2.2 Papers using *individual* classification algorithms for PD modelling (*PCC*)[2]

*Authors (*Date)	Algorithms[3]												
	LDA	LR	DT	NN	SVM	PA	BN	kNN	F	QDA	NB	MARS	MP
Baesens et al. (2003)	79.3^{3}	79.3^{3}	79.0^{4}	79.4^{2}	**79.7^{1}**		77.1^{6}	78.2^{5}		70.5^{8}	74.2^{7}		79.0^{4}
Malhotra and Malhotra (2003)	65.2^{2}			**69.4^{1}**									
Lee and Chen (2005)	75.5^{4}	76.1^{3}		**84.8^{1}**								81.0^{2}	
Li, Shiue, and Huang (2006)				73.2^{2}	**84.3^{1}**								
Huang, Chen, and Siew (2006)			82.8^{2}	**82.9^{1}**							82.9^{1}		
Lee, Chiu, Chou, and Lu (2006)	69.1^{5}	72.0^{4}	**79.2^{1}**	74.5^{3}								78.8^{2}	
Xiao, Zhao, and Fei (2006)	73.8^{5}	**79.0^{1}**	71.8	77.7^{4}	78.2^{2}			73.2^{6}				77.9^{3}	
Huang, Chen, and Wang (2007)			79.8^{2}	**82.3^{1}**									
Yu, Wang, and Lai (2008)		73.3^{3}		77.2^{2}	**78.9^{1}**								
Abdou et al. (2008)	86.9^{4}	88.3^{2}		**96.2^{1}**		87.8^{3}							
Yu, Wang, and Lai (2009)	70.5^{4}	73.0^{3}		75.6^{2}	**75.7^{1}**								
Šušteršič, Mramor, and Zupan (2009)		76.1^{2}		**79.3^{1}**									
Tsai, Lin, Cheng, and Lin (2009)	74.5^{3}	84.4^{2}		**93.6^{1}**									
Nanni and Lumini (2009)				**82.9^{1}**	80.7^{2}			69.2^{3}					

[2] Each classifier's rank is at the right corner of the performance result. Bold font shows the best classifier in the papers that use more than 2 classifiers.
[3] LDA: linear discriminant analysis; K-NN: K-nearest neighbour; SVM: support vector machine; DT: decision trees; NB: Naïve Bayes; NN: Neural Network; PA: probit analysis; BN: Bayesian network; MP: mathematical programing; MARS: multivariate adaptive regression splines; F: fuzzy classifiers. Akkoc (2012) and Sohn (2016) applied fuzzy rule on neural networks and logistic regression respectively.

Chen et al. (2009)			**81.8**^{1}		81.3^2							80.5^3	
Zhang et al. (2010)			79.4^3	80.7^2	**82.4**^{1}								
Yu, Yue, Wang, and Lai (2010)	60.8^4	64.7^2		58.9^5	**65.2**^{1}					64.3^3			
Zhou et al. (2010)	66.9^7	75^2	71.0^5	73.9^3	**75.5**^{1}		68.4^6	72.7^4		61.5^8			
Yu et al. (2011)	81.5^3	81^5	82.4^2	81.2^4	**84.5**^{1}		80.0^6	78.7^7		77.0^9	77.2^8		80.1^6
Wang et al. (2011)		**78.3**^{1}	78.1^2	75.3^4	76.5^3								
Yao and Lu (2011)	75.6^6	79.1^4	76.6^5	81.0^2	**82.1**^{1}			79.3^3					
Yap, Ong, and Husain (2011)		71.2^2	71.9^1										
Finlay (2011)	85.9^2	85.9^2	84.9^4	**86.0**^{1}				85.4^3					
Akkoç (2012)	57.2^4	57.8^3		58.6^2					**60**^{1}				
Hens and Tiwari (2012)				81.7^2	**81.8**^{1}								
Li, Tsang, and Chaudhari (2012)		80.6^2			**82.3**^{1}								
Marqués et al. (2012a)		**83.2**^{1}	81.4^3	81.4^3	82.8^2			80.0^4			66.2^5		
Tsai (2014)		82.2^2	82.1^3	**87.4**^{1}									
Florez-Lopez and Ramon-Jeronimo (2015)	75.0^5	74.6^6	**80.5**^{1}	77.6^3	75.2^4			79.4^2					
Liang, Tsai, and Wu (2015)			83.1^2	80.1^4	**84.0**^{1}			80.8^3			76.1^5		
Bahnsen, Aouada, and Ottersten (2015)		**93.8**^{1}	93.5^2										
Malekipirbazari and Aksakalli (2015)		54.5^3			63.3^2			**70.1**^{1}					
Bae and Kim (2015)		71.0^2	68.7^3	**72.8**^{1}									
Koutanaei, Sajedi, and Khanbabaei (2015)			85.9^2	**87.2**^{1}	85.9^2						82.1^3		
Leong (2016)		88.3^3		**94.0**^{1}			89.8^2						
Li, Niskanen, Kolehmainen, and Niskanen (2016)		71.6^2		**75.4**^{1}									
Abedini et al. (2016)		**75.8**^{1}	71.6^4	74.2^3	75.1^2								
Zhang, Jia, Diao, Hai, and Li (2016)		77.1^3	**81.2**^{1}	77.4^2									
Sohn, Kim, and Yoon (2016)		**71.5**^{1}							70.0^2				
Van Sang, Nam, and Nhan (2016)			79.8^2	78.0^3	**81.4**^{1}			77.7^4			69.5^5		
Ala'raj and Abbod (2016b)			80.6^3	82.5^2	**83.4**^{1}						77.0^4		
Ala'raj and Abbod (2016a)		82.9^4	**86.2**^{1}	82.0^5	84.9^2						56.8^6	83.3^3	
Zhang, Liu, Zhang, and Almpanidis (2017)		77.4^3	79.7^2		**84.2**^{1}			76.3^4			71.9^5		
Xia, Liu, Li, and Liu (2017)		74.8^2	70.7^4	71.3^3	**76.5**^{1}								
Fallahpour et al. (2017)		**74.0**^{1}	71.6^4	73.4^2	72.2^3								
Saritas and Yasar (2019)				**87.0**^{1}							83.5^2		

Table 2.3 Papers using *individual* classification algorithms for PD modelling (*AUC*)

*Authors (*Date)	Algorithms												
	LDA	LR	DT	NN	SVM	PA	BN	kNN	F	QDA	NB	MARS	MP
Baesens et al. (2003)	76.2^{4}	76.3^{3}	71.3^{9}	**77.0^{1}**	76.5^{2}		74.7^{5}	73.6^{7}		72.6^{8}	74.0^{6}		70.7^{10}
Sinha and Zhao (2008)		**85.5^{1}**	74.3^{6}	84.8^{2}	78.2^{4}			77.8^{5}			82.4^{3}		
Nanni and Lumini (2009)				**87.7^{1}**	85.7^{2}			73.3^{3}					
Yu et al. (2009)	70.8^{4}	73.3^{3}		75.8^{2}	**76.0^{1}**								
Bellotti and Crook (2009)	78.1^{2}	77.9^{3}			**78.3^{1}**			75.6^{4}					
Zhou et al. (2010)	**63.7^{1}**	**63.7^{1}**	62.4^{4}	61.1^{6}	63.5^{2}		56.5	63.4^{3}		61.4^{5}			
Paleologo, Elisseeff, and Antonini (2010)			53.0^{3}		**60.0^{1}**			54.0^{2}					
Brown and Mues (2012)	81.3^{2}	78.2^{4}	73.5^{7}	78.7^{3}	**84.2^{1}**			76.6^{5}		73.9^{6}			
Akkoç (2012)	62.1^{4}	62.2^{3}		62.5^{2}					**63.1^{1}**				
Marqués et al. (2012b)		**83.7^{1}**	75.3^{3}	82.2^{2}	72.8^{4}			72.7^{5}					
Kruppa et al. (2013)		**77.9^{1}**						68.5^{2}					
Koutanaei et al. (2015)			**89.3^{1}**	84.5^{2}	81.9^{3}						79.8^{4}		
Tomczak and Zięba (2015)		**63.4^{1}**	62.4^{3}	63.2^{2}	61.6^{4}								
Florez-Lopez and Ramon-Jeronimo (2015)	76.2^{6}	76.9^{4}	80.2^{2}	78.4^{3}	76.5^{5}			**89.9^{1}**					
Malekipirbazari and Aksakalli (2015)		**68.0^{1}**			62.0^{2}			53.0^{3}					
Bravo et al. (2015)		**66.8^{1}**		66.6^{2}									
Ala'raj and Abbod (2016b)			76.1^{3}	79.7^{2}	**80.8^{1}**						75.5^{4}		
Ala'raj and Abbod (2016a)		**83.6^{1}**	**80.0^{2}**	73.7^{4}	77.6^{3}						69.6^{5}	**83.6^{1}**	
Zhang et al. (2017)		81.0^{4}	84.6^{2}		**85.2^{1}**			82.8^{3}			80.1^{5}		
de Melo Junior, Nardini, Renso, and de Macêdo (2019)	**86.2^{1}**	85.8^{2}	80.8^{6}	85.6^{3}	85.5^{4}			83.5^{5}	80.8^{6}		78.2^{7}		

Among the numerous datasets employed in the literature, many of the studies use the same German dataset and Australian dataset. These are publicly available credit scoring datasets available from < http://kdd.ics.uci.edu/ >. The German dataset consists of 1000 cases of which 300 are bad. The Australian dataset consists of 690 cases of which 383 are bad. Both datasets have far fewer observations than in a dataset a financial institution would have and with much higher proportions of bad cases than are usually observed in practice.

In the literature that compares different individual classifiers, many papers find that more recent and complex machine learning methods outperform linear statistical methods. It can be seen from Table 2.2 and Table 2.3, SVM and NN are the best classifiers in data mining, whereas LR is the best statistical method. However, the performances of SVM is largely influenced by the kernel and parameter employed.

For instance, in Bellotti and Crook (2009), SVM with linear and RBF kernels give the best performances whereas SVM with polynomial kernel gives the worst among classifiers employed. Neural networks suffer from its black box nature as some argue that although they give better predictions, the training process can be slow and the predictors' effects are difficult to interpret.

Xiao et al. (2006) find that logistic regression, SVM and MARS, etc. give good classification ratios while linear discriminant analysis and decision trees yield better accuracy. Based on 5 datasets, Brown and Mues (2012) find that LDA and LR give competitive results compared to more recent ensemble techniques such as gradient boosting and random forests, whereas QDA and C4.5 decision tree significantly underperform. Baesens et al. (2003) find that radial basis function kernel least squares support vector machines and neural networks have the best PCC and AUC performances. LDA and LR also have good performances indicating weakly nonlinear data. Hens and Tiwari (2012) use the F score method to choose features in their SVM model to reduce computational time. They find their method is computationally efficient and at the same time does not induce large decreases in accuracy. Yu et al. (2011) use least squares SVM to solve the quadratic programming problem in the traditional SVM. They find their method performs better when applied in both the German and the Australian data in terms of overall accuracy and type I accuracy compared to other models. They also argue that LDA and LR are not significantly different from their method with respect to accuracy, yielding very good performance. Furthermore, NB and KNN give the best performance with respect to type II errors.

Table 2.4 Papers using *ensembles and hybrid* classification algorithms for PD modelling (*PCC*)

Authors (Date)	Algorithms[4]													
	RS	Bag	Hbrd	Bag NN	Bst	RF	Rot F	GB	V	K-mns	Stck	DM slct	UW A	Bag DT
Lee and Chen (2005)			84.7											
Rodriguez, Kuncheva, and Alonso (2006)		84.7[4]			85.03[3]	85.05[2]	**87.0[1]**							
Huang et al. (2007)			82.4											
Ko et al. (2008)	83.8[4]	85.7[3]			86.4[2]							**87.8[1]**		
Yu et al. (2008)		86.5[1]	81.5[2]											
Chen et al. (2009)			86.8											
Nanni and Lumini (2009)	**82.8[1]**	**82.8[1]**					82.6[2]							
Partalas et al. (2009)									80.8					
Yu et al. (2010)		71.2												
Partalas et al. (2010)													86.8	
Zhang et al. (2010)			86.1[2]	81.8[3]										**86.8[1]**
Zhou et al. (2010)					73.3[2]				**75.6[1]**					
Li, Wei, Li, and Xu (2011)					89.6									
Finlay (2011)		86.7[2]			**87.5[1]**				86.1[3]			85.5[4]		
Yao and Lu (2011)			82.1											
Wang et al. (2011)		**80.8[1]**			79.5[3]						80.4[2]			
Akkoç (2012)			60.0											
Marqués et al. (2012b)	83.6[3]	84.1[2]			82.9[4]		**84.3[1]**							
Tsai (2014)										87.0				
Florez-Lopez and Ramon-Jeronimo (2015)						79.3[3]		79.6[2]	**80.5[1]**					
Bahnsen et al. (2015)						**85.6**								
Koutanaei et al. (2015)		88.5[2]			**91[1]**	87.2[3]					85.9[4]			
Malekipirbazari and Aksakalli (2015)						78.0								
Van Sang et al. (2016)						80.6								
Li et al. (2016)			84.6											
Abedini et al. (2016)	73.7[4]	77.3[2]			75.8[3]		75.8[3]		**78.3[1]**					

[4] UWA: Uncertainty weighted accuracy; Bag: bootstrap aggregating ; RF: Random forest ; Rot F: Rotation forest; RS: random subspace; Hbrd: hybrid; Bst: boosting; GB : gradient boosting; K-mns: K-means clustering; Stck: stacking; DMslct: dynamic ensemble/classifier selection; BagDT: bag decision tree; BagNN: bag multilayer perceptron; V: voting

Ala'raj and Abbod (2016b)						83.9								
Ala'raj and Abbod (2016a)						**88.1**[1]			83.9[2]					
Abellán and Castellano (2017)	83.9[3]	84.1[2]			83.1[4]		**84.3**[1]							
Xia et al. (2017)				72.7[4]		76.1[2]		**77.2**[1]						73.8[3]
Zhu, Xie, Wang, and Yan (2017)	77.5[2]	77.2[3]			**85.4**[1]									
Zhang et al. (2017)						**86.4**[1]		86.0[2]						
Fallahpour et al. (2017)		76.2[2]			**77.4**[1]									

Table 2.5 Papers using *ensembles and hybrid* classification algorithms for PD modelling (*AUC*)

*Authors (*Date)	Algorithms													
	RS	Bag	Hbrd	Bag NN	Bst	RF	Rot F	GB	V	K-mns	Stck	DM slct	UW A	Bag DT
Nanni and Lumini (2009)	88.2[2]	88.2[2]					**88.3**[1]							
Paleologo et al. (2010)		**66.0**[1]			59.0[2]									
Zhou et al. (2010)					51.4[2]				**64.0**[1]					
Akkoç (2012)			63.1											
Marqués et al. (2012b)	83.7[3]	85.2[2]			83.5[4]		**85.8**[1]							
Brown and Mues (2012)						**81.5**[1]		80.7[2]						
Kruppa et al. (2013)				68.1[2]		**95.9**[1]								
Abellán and Mantas (2014)	88.2[2]	**88.6**[1]												
Florez-Lopez and Ramon-Jeronimo (2015)						**85.2**[1]		84.3[3]	85.1[2]					
Malekipirbazari and Aksakalli (2015)						71.0								
Tomczak and Zięba (2015)		**63.0**[1]			59.9[3]	61.4[2]								
Koutanaei et al. (2015)		90.3[2]			**91.2**[1]	89.1[3]					82.4[4]			
Ala'raj and Abbod (2016b)						87.0								
Ala'raj and Abbod (2016a)						**84.2**[1]			80.1[2]					
Fitzpatrick and Mues (2016)					**77.7**[1]	77.1[2]								
Abellán and Castellano (2017)	85.0[2]	84.9[3]			83.3[4]		**85.8**[1]							
Zhu et al. (2017)	85.8[3]	84.6[2]			**91.0**[1]									
Zhang et al. (2017)						87.0[2]		**87.4**[1]						
de Melo Junior et al. (2019)						**87.1**[1]		86.6[2]						86.4[3]

In the literature that uses ensemble methods, it can be seen from Table 2.4 and Table 2.5 that most ensembles have been considered the best by at least one of the papers based on one of the performance measures. No model significantly outperform others with all performance measures on all data sets. Therefore there is no consensus in terms of one single best ensemble. The best performance of each of the ensembles recorded is good, with PCC and AUC generally above 80%.

As to comparing ensembles with individual classifiers, many papers find that ensembles improve model performance while some find no significant performance improvement compared to individual classifiers. Based on two credit datasets from Germany and Australia, Zhang et al. (2010) find that their bagging decision tree ensemble method outperforms single classifiers such as C4.5, multilayer perceptron neuron networks (MLP), and support vector machines (SVM) concerning classification accuracy. They also find that the weighted vote strategy outperforms the majority vote. West, Dellana, and Qian (2005) use cross-validation, bagging and Adaboost ensembles with MLP as the base classifier. They find that ensembles consolidating cross-validation, bagging and boost outperform single ones with respect to generalisation errors. They also find that bagging and boosting have no significant difference with simple cross-validation in terms of accuracy. Their performances differ between datasets. Cross-validation works best for German data while bagging is most accurate for Australian data and their 'bankruptcy data'. Yu et al. (2010) use multistage ensembles based on SVM classifiers with majority voting and weight averaging as combining strategies. They find their approach outperforms individual classifiers and some comparable ensembles based on error rates and overall accuracy. Yu et al. (2008) find similar results with their multistage ensemble based on neural network base learners. Marqués et al. (2012a) study the performances of base learners in an ensemble. They find that concerning accuracy and error rates, decision tree classifier is the best. However, MLP, LR and SVM are not far behind. Paleologo et al. (2010) use subbagging ensemble method to cope with highly imbalanced data. Based on their Italian corporate loan data, they find that this ensemble works best with decision tree, linear SVM and RBF SVM as base learners in terms of AUC measure. Tsai and Wu (2008) consider that in theory their neural network ensemble should outperform single classifiers. But in their experiments, the best single neural network outperforms ensembles in terms of average accuracy. With respect to error rate measure, there is no significant winner between individual models and ensembles. Zhou et al. (2010) find mixed results from their ensembles based on least squares support vector machine models with voting and reliability based strategies. They find their reliability based ensembles outperform individual models when tested on German data. However single model kNN has the best performance when tested on British data.

Overall, we have the following summaries about literature in PD classification. Firstly, SVM and NN are some of the best classifiers amongst data mining methods, while logistic regression is one of the best statistical models. However, they all have their respective disadvantages. Secondly, there is no consensus in terms of the best ensemble methods. Most of them have good model performances. Thirdly, the majority of papers find that ensembles outperform individual classifiers, whereas there are also papers that find no significant difference in model performances between the two.

2.2.3 Modelling PD, LGD, and EAD using Bayesian estimation in statistical models

In the literature using statistical models for credit scoring, papers normally use the frequentist estimation methods such as ordinary least squares for discriminant analysis or maximum likelihood for logistic regression. There are also some papers that use a Bayesian approach in estimating the regression coefficients. In this section, we summarise papers that use the Bayesian approach in parameter estimation and inference for statistical parametric models in credit scoring.

To estimate corporate default probability, Mira and Tenconi (2004) use a Bayesian hierarchical logistic regression model and include sector dependence, therefore the expert prior opinion on PD as well as cross sectional correlations are taken into consideration. They use both informative and non-informative prior in estimation and find similar results. Since the data is fairly large, posterior is more influenced by the likelihood of the data than by the priors. With cross validation analysis, they find that the Bayesian regressions outperform the frequentist models in predictive performance. Park et al. (2010) employ a small area estimation approach for a hierarchical Bayesian model to estimate the probability of default for corporate loans. With their method, the missing values of needed variables are substituted by values from related areas of those variable, thereby increasing the estimation accuracy. Similar to Mira and Tenconi (2004), Park et al. (2010) take into account the cross sectional and serial correlations into PD estimation by introducing into the model latent variables, whose variances represent obligor correlations, and whose means over time represent serial correlations.

Many papers use non-informative priors on the basis that likelihood of the data normally overwhelms priors asymptotically with large data. However, Jacobs and Kiefer (2010); Kiefer (2009) emphasise the importance of expert information and use informative priors. Maltritz and Molchanov (2008) study the determinants of sovereign default risk. They consider that including only the significant explanatory variables in credit risk modelling could induce the model risk of omitted variables. Therefore they use Bayesian model averaging to average over all possible models with different explanatory variables. They find that the debt to GDP ratio, history of default, and credit ratings are the most important determinants. In analysing highly dimensional complex datasets, Giudici (2001) use graphical models and Bayesian model selection to increase model flexibility and computational efficiency. Since each alternative model in the model space is attached with a model score, the search for an appropriate model using their Bayesian graphical method is more interpretable and efficient than the classical backward variable selection method. They find that the model selected is more parsimonious using their Bayesian method than using the classical model selection. Miguéis et al. (2013) argue that models that give estimates of the relationship between explanatory variables and the expectation of the dependent variable provide only average effects, and overlook the potential coefficient change for the extreme values of the dependent variable. Therefore, to estimate the probability of default of credit cards they use quantile regression models so that coefficient estimations are provided for every quantile of the distribution of the dependent variable. Miguéis et al. (2013) use a Bayesian approach to estimate credit card default probability. They find that each explanatory variable has a different coefficient estimate for each PD quantile, and the signs are robust across all dependent variable values.

Few papers use Bayesian estimation for loss given default. Bijak and Thomas (2015) use a Bayesian hierarchical model apart from a two-step frequentist method to estimate loss given default. Unlike the classic approach which only gives point estimates for each loan, the Bayesian models produce LGD distributions. Secondly, with the classic estimation approach, predicting LGD is difficult because the two steps of estimation are done separately. With a Bayesian hierarchical model, on the other hand, we only need to build one single model instead of two regression models. Bijak and Thomas (2015) find the parameter estimation results from using a Bayesian method are similar to those from the frequentist approach. To the best of our knowledge, there are presently no papers that model exposure at default using the Bayesian estimation approach.

Using Bayesian methods to estimate credit risk has many advantages, such as providing full parameter distributions; can work on a limited amount of data; can incorporate prior information. Powerful computation techniques such as the Metropolis-Hastings algorithm could be employed with Bayesian estimation (Jacobs & Kiefer, 2010). Using the Bayesian method can also reduce the imbalanced data bias derived from rare default events (King & Zeng, 2001; Jacobs & Kiefer, 2010; Kiefer, 2009) and the missing value problem (Park et al., 2010). Some papers use Bayesian methods for PD, LGD models, and state the advantages of coefficient distributions obtained from the Bayesian approach over the point estimates from the frequentist approach (Bijak & Thomas, 2015).

We find a few gaps in the literature on PD modelling. Firstly, most papers use the frequentist approach to modelling PD; and many use the Bayesian approach. However, few papers compare the performance of the two approaches. The second gap is the relatively rare use of informative priors. The frequentist methods rely entirely on data information. In contrast, the Bayesian approaches allow the incorporation of expert knowledge or other non-data information as priors into the estimation. It may be difficult to elicit and represent this expert prior information in a probability distribution fashion (Kiefer, 2009). If prior information is not available, non-informative priors are also developed which gives Bayesian methods more flexibility. In the literature using Bayesian methods in credit risk modelling, a lot of papers use non-informative priors. We consider that although non-informative priors can be used, using them misses the opportunity of adding expert information to the model, and subsequently the potential of improving model performance through adding available useful information.

2.3 Literature review on credit risk stress testing

Stress testing is a term that originated in engineering. It is a technique to evaluate the stability of a material, building, or machine, etc. under adverse circumstances. In stress testing, an object is put under different levels of high stress to test its resilience (Borio et al., 2014). Stress testing in finance is an important tool to assess a bank's risk level and to provide a basis to assist the decision making of financial institutions and regulators (Schechtman & Gaglianone, 2012). The focus of bank stress testing is on credit risk which represents a banking system's most significant risk (Sorge & Virolainen, 2006). A stress test examines the performance of an entity, such as a bank portfolio, a balance sheet or the vulnerability of a banking sector, under severe but plausible economic conditions (Misina, Tessier, & Dey, 2006).

2.3.1 Different aggregations

In this review, we summarise and evaluate the literature according to whether the stress test is carried out at an aggregate financial system level or an individual bank/portfolio/customer level. We further categorise stress tests into whether the risk exposures subject to stress are elements of the lenders' balance sheets (such as loan loss provisions, non-performing loans) or the borrowers' credit loss parameters (such as the probability of default).

This way of categorising stress testing literature is similar to that of Sorge and Virolainen (2006), in which they perform macro stress tests and generally group the macro stress testing methods into the balance sheet approach and the distribution approach. The former is to make point forecasts for credit risk indicators in a bank's balance sheet under changes of macroeconomic variables. The latter is to simulate probability distributions of estimated PD/LGD/loss for portfolios of accounts, under tranquil and stress economic scenarios. However, it should be noted the two categories can overlap. For instance, some papers stress test the balance sheet risk indicators by forming probability distributions of such indicators (Schechtman & Gaglianone, 2012) instead of forming point predictions of these accounting variables under assumed macroeconomic stresses. Potentially, stress tests of customer credit risk can also be carried out without simulating a probability distribution of the estimated PDs or credit losses but by predicting point estimates of such indicators under point changes of macroeconomic variables which is one or several scenarios, as opposed to a group of simulated scenarios that form a scenario distribution.

Therefore, based on the balance sheet/VaR categorisation of Sorge and Virolainen (2006), we further grouped the stress testing literature into the following three subgroups: macro versus micro, lender versus borrower, and point predictions versus probability distribution. Most types of credit risk stress testing methods in the literature generally fall in either category in these three groups.

Macro/Micro

A major difference between the micro and macro stress testing is the exposure being tested. Micro stress testing is to measure the vulnerability of single portfolios or financial institutions and to support capital management and crisis resolution (Crook, Leow, & Bellotti, 2015). Macro stress testing is to evaluate the vulnerability of a group of financial institutions that is large enough to impact the whole economy (Borio et al., 2014).

Lender/borrower

In this framework, papers can be separated by whether they stress test the credit risk of the lender or of the borrower. Lenders are in the sense of financial intermediaries such as banks and building societies as opposed to the depositors. Borrowers are corporates, households, individuals, etc. In the modelling process, for the former, the credit risk of interest is commonly represented by banks' accounting measures of vulnerability (Schechtman & Gaglianone, 2012; Crook & Banasik, 2012; Delgado & Saurina, 2004; Salas & Saurina, 2002; van den End, 2006; Drehmann, Patton, & Sorensen, 2005; Hoggarth, Sorensen, & Zicchino, 2005; Delgado & Saurina, 2004; Bikker & Hu, 2002; Laeven & Majnoni, 2003; Sorge & Virolainen, 2006). Covariates included are macroeconomic variables and other variables specific to the financial institutions indicating their profitability, size, risk diversification, etc. For the latter, the credit risk of interest is commonly represented by the debtors' default probability or estimated credit loss with covariates being macroeconomic variables and information that reflects debtors' credit risk level (Bellotti & Crook, 2013, 2014; Jokivuolle & Viren, 2013; Sorge & Virolainen, 2006; Wong et al., 2008).

Point predictions/distribution

Some papers use a point prediction approach (BoE, 2018; FRB, 2018; Breeden, 2016; Busch et al., 2018; EBA, 2018). In this approach, for given baseline and stress scenarios, one or several potential values for credit risk are obtained. These scenarios can be either historical or hypothetical. A historical stress scenario is an event that happened in the past whereas a hypothetical scenario is one decided by experts. This approach is mainly in use for macro stress testing banks' aggregate balance sheet indicators.

The stress testing approach predominantly employed in the literature is the distribution approach (Bellotti & Crook, 2013, 2014; Jokivuolle & Viren, 2013; Kanas & Molyneux, 2018; Sorge & Virolainen, 2006; Wong et al., 2008). This approach can be further separated by whether the risk distribution is formed by simulations or simple samplings of the historical scenarios. It can also be separated by whether one single distribution or separate distributions are generated for baseline and stress scenarios

i) Simulation/simple random sampling.

Simulations are experiments conducted to imitate real situations (Greene, 2011). Monte Carlo simulation is a procedure to take draws from pseudo-random variables (Sawilowsky, 2003). Simple random sampling is a process where a designated number of observations are drawn from a population (Starnes, Moore, Yates, & Tabor, 2014). The sample should be taken randomly and represent the population. Simulations rely on a random sampling technique, but they are a fictitious representation of reality. In stress testing using the distribution approach, scenarios can be generated by simulation or simple random sampling of observed previous values.

For the simulation-based distribution approach, both hypothetical and historical scenarios are used in the literature (Bellotti & Crook, 2013, 2014; Jokivuolle & Viren, 2013; Sorge & Virolainen, 2006; Wong et al., 2008). It can be impractical and subjective to come up with hypothetical macroeconomic scenarios individually in a distribution approach. Therefore papers using the distribution approach with hypothetical scenarios commonly simulate the error terms of the stress testing models and add hypothetical shocks to these error terms. These hypothetical shocks are then transformed into the macroeconomic variables and subsequently the default probability through their VAR type system of equations (Jokivuolle & Viren, 2013; Sorge & Virolainen, 2006; Wong et al., 2008). Papers using a simulation-based distribution approach also employ historical scenarios that are based on historical values of the macroeconomic variables (Bellotti & Crook, 2013, 2014). In this case, both error terms and macro-level covariates are simulated. One reason that simulation is needed for the macroeconomic variables is to preserve the structure of dependency between these variables as Crook et al. (2015) argue that individually univariate sampling of the macroeconomic variables is incoherent. Bellotti and Crook (2013) and Bellotti and Crook (2014) use Cholesky decomposition and principal component analysis respectively to preserve the variance-covariance structure between the macroeconomic time series. They normalise the macro variables/factors since Cholesky decomposition requires the variables to be normally distributed so that random draws can be conveniently taken from a known distribution in the simulation process. As for simple random sampling, a few papers (Berkowitz, 1999) that use distribution approach employ this method.

Most papers using the distribution approach simulate the historical macroeconomic scenarios since individually sampling from each of the original variables ignores the structure between variables (Crook et al., 2015). However, a few papers (Berkowitz, 1999) treat the observations of macroeconomic variables in each time period as a vector and draw simple random samples of these vectors over time. There are several benefits to this method. Firstly, the dependence structure in the macroeconomic variables is preserved in this way since the values of macro variables are drawn simultaneously from each time period. Secondly, the simulation method is an approximation of reality and often has other requirements such as normalisation of the original variables beforehand. The simple random sampling is of the original variables thus better fits the plausibility requirement of stress testing.

ii) A single distribution for stress and normal scenarios/different distributions for different scenarios

The stress testing steps of the distribution approach are similar at both the macro and micro levels. First consider a time horizon, for instance, 1 year in the future, for the stress testing to be performed on. Second, take random draws from macroeconomic variables both in the non-stressed baseline and stressed circumstances to generate normal and stress scenarios to form one loss distribution for each of the two circumstances. This is based on the argument that stress will shift the loss distribution to the right and change the shape of the distribution. This is the method suggested by Berkowitz (1999). Some employ this method (Jokivuolle & Viren, 2013; Sorge & Virolainen, 2006; Wong et al., 2008).

Some, on the other hand, use simulated historical scenarios to form one single loss distribution and assume the high loss with low probability is under extreme circumstances while the mild loss with high probability is under baseline circumstances (Bellotti & Crook, 2013, 2014).

2.3.2 Macro level stress testing

With macro stress testing, time series data and models are often employed to test the credit risk of the financial system at an aggregate level.

2.3.2.1 Estimating banks' Balance sheet indicators

The commonly used balance sheet indicators to measure credit risk are non-performing loans, loan loss provisions, loan write-offs, etc. The macro variables used are similar in all papers with slight variations. The macroeconomic variables considered to impact the credit risk are GDP, industrial production index, interest rates, exchange rates, inflation, etc. (Delgado & Saurina, 2004; Hoggarth, Sorensen, & Zicchino, 2005).

Data is usually aggregate time series data of a country/industry/sector (Sorge & Virolainen, 2006), or panel data for several countries' banking sectors (Bikker & Hu, 2002; Laeven & Majnoni, 2003). The models used include linear regressions (Sorge & Virolainen, 2006); vector autoregressive (Hoggarth et al., 2005); vector error correction model (Crook & Banasik, 2012; Delgado & Saurina, 2004), etc., studying the relationship between financial system instability and macroeconomic indicators. As for the stress testing approach, some seek point predictions of risk indicators under assumed stress scenarios (Sorge & Virolainen, 2006), while some consider the distribution of loan loss provisions potentially arising under given shocks (Schechtman & Gaglianone, 2012).

2.3.2.2 Estimating borrowers' credit risk

The expected loss is computed as the product of the risk exposure at the time of default (Exposure at Default, or EAD), the percentage of the exposure that will be at lost if a default event occurs (loss given default, or LGD), and the probability of a default happening (probability of default, or PD). Assuming exposure and loss given default are 1, and assuming the three parameters are independent, the expected credit loss depends on its probability of default.

At a macro level, credit loss is usually measured by the probability of default with LGD and EAD considered constant. Since papers in this category study default probability at an aggregate level, this variable is commonly measured by the number of default companies divided by the number of all companies in a sample or the sum of debts of the failed companies over the total debt (Jokivuolle & Viren, 2013; Sorge &

Virolainen, 2006; van den End, 2006; Wong et al., 2008). Similarly, aggregate default rates data is used in stress testing consumer loan loss (Rösch & Scheule, 2004). In order to keep the dependent variable (probability) between 0 and 1, a logistic transformation is commonly applied.

Apart from the models discussed in 2.3.2.1 such as VAR (Misina et al., 2006; van den End, 2006), VECM (Crook & Banasik, 2012), Sorge and Virolainen (2006); Wong et al. (2008) and Jokivuolle and Viren (2013) use a system of regression equations for several industries/sectors using seemingly unrelated regression.

As for the stress testing approach, many papers (Jokivuolle & Viren, 2013; Rösch & Scheule, 2004; Sorge & Virolainen, 2006; Wong et al., 2008) use the distribution approach by generating shocks from macroeconomic variables and simulate future default rates under such shocks. With the distribution of default rate simulated, expected and unexpected loss can be obtained.

2.3.3 Micro level stress testing

With micro stress testing, panel data and models are commonly employed to test the credit risk of financial institutions, corporates, or consumers at an individual level.

2.3.3.1 Estimating banks' Balance sheet indicators

On estimating banks' Balance sheet indicators, the micro level stress testing papers are similar to aggregate macro stress testing papers in the use of dependent variables and macroeconomic variables. Different from the aggregate macro level analysis, papers that stress test individual banks' balance sheets at the micro level also add a cross-sectional dimension to times series analysis using panel data. Bank level characteristics are added as explanatory variables to model banks' loan loss provision, non-performing loan, etc. (Salas & Saurina, 2002; van den End, 2006).

Apart from the models that directly link credit risk with a macroeconomic shock through panel regressions (van den End, 2006), there are also papers employing multi-stage structural models which add intermediate variables between credit risk indicators and macro factors (Drehmann, Patton, & Sorensen, 2005).

2.3.3.2 Estimating borrowers' credit risk

The concept of bank stress testing originated at the micro level (Borio et al., 2014). A VaR approach to stress test banks' portfolios is a standard procedure adopted by the industry (Sorge & Virolainen, 2006). Our stress testing research falls into this category.

Logistic regression models are commonly used to find the variables impacting credit failure and to estimate the default probabilities of obligors. The standard estimation technique is maximum likelihood methods (Bellotti & Crook, 2013, 2014), and there are also papers employing Bayesian estimators for PD (Miguéis et al., 2013). Bangia, Diebold, Kronimus, Schagen, and Schuermann (2002) consider that asset return is determined by systematic and idiosyncratic risk, and it follows a normal distribution. They follow a Merton approach, in which default occurs when asset value drops under the value of outstanding debt. The probability of default for companies in the same credit rating (as ascribed by a rating agency) is considered the same. With the empirically estimated probability of one rating and the rating transition probability, the probability of default for companies of every rating can be computed. Instead of comparing credit value with assets, Pesaran, Schuermann, Treutler, and Weiner (2006) compare it with equity. The probability of default is determined by the probability that equity return falls below a threshold measured by the rating of the company. Therefore the problem of default rates becomes the problem of equity return. The return is regressed against lags of national and global macroeconomic variables using a global VAR (GVAR) model.

Papers find that macro variables such as output growth and inflation have negative impacts, while interest rate, unemployment, equity price, etc. have positive impacts on credit risk (Bellotti & Crook, 2013, 2014; Bikker & Hu, 2002; Hoggarth et al., 2005; Jokivuolle & Viren, 2013; Laeven & Majnoni, 2003; Pesaran et al., 2006; Sorge & Virolainen, 2006; van den End, 2006; Wong et al., 2008). They also find that credit losses under stress scenarios with shocks from macroeconomic variables are higher than those under baseline scenarios (Bellotti & Crook, 2013, 2014; Hoggarth et al., 2005; Jokivuolle & Viren, 2013; Pesaran et al., 2006; Sorge & Virolainen, 2006; van den End, 2006; Wong et al., 2008). For instance, in Schechtman and Gaglianone (2012) a GDP shock of 2 standard deviations causes the solvency probability to drop by nearly 8%. In Jokivuolle and Viren (2013), the expected probability of default under stressed scenario is nearly twice that under the baseline scenario, with shocks to GDP, interest rate, debt, profit rate, etc. considered. In Breeden (2016), with a longer time horizon the forecasted PD under the stressed scenario is found to be increasingly higher than that under the baseline scenario. At the end of the stress testing period, the forecasted PD under the severely adverse scenario is found to be around 10% higher than that under the baseline scenario.

One gap in the literature is that papers do not include coefficient estimation risk in their stress testing models and procedures. Although the scales, objects, models, estimation methods for credit risk stress testing are vastly different, the stress testing process itself is essentially the same, especially how the coefficient estimates are used. After getting the estimation results, most papers substitute the coefficient mean estimates with the macro scenarios in the model and obtain estimated values of the risk indicator. However, in the frequentist framework, the sample mean estimates of coefficients are only approximations of the population means. The coefficient estimation errors play little role in the stress testing literature. Various methods are employed in the literature to ensure that the macroeconomic scenarios represent reality well enough, but little attention has been paid to estimation risk and the misleading results that neglecting this source of risk may bring to credit loss prediction. We consider that the use of the Bayesian approach to simulate the Bayesian coefficient posterior distribution and to include the posterior distribution in the stress testing process may be a potential way to solve the problem of neglecting coefficient uncertainty. This method firstly has the benefit that it introduces estimation risk into stress testing as it takes into consideration all the possible coefficient values in the coefficient distribution, instead of point estimates, thus addressing the problem of ignoring estimation error and underestimating estimation risk. Second, since in the posterior distribution the areas of unlikely coefficient values have correspondingly lower probability, using a Bayesian posterior distribution takes into account the varying probabilities of different coefficient values, thus avoiding credit risk overestimation. Furthermore, since in the Bayesian framework coefficients are random variables, not a fixed value, it is conceptually appropriate to use different regions in the posterior distribution, as opposed to mean estimates.

2.4 Conclusion

In the first section of this chapter, we first reviewed different classification algorithms for modelling the probability of default in credit scoring literature. We then discussed the model performances of some of the most commonly used individual and ensemble models based on two of the most popular performance measures in the literature. We find that support vector machines, neural networks, and logistic regression are among the best performing models. We also find that although ensemble classifiers generally have good model performances, there is no consensus as to which are the best ensembles. Lastly, we reviewed the papers that use Bayesian econometrics in parameter estimation. We think that prior information is a major advantage in the Bayesian approach compared to the frequentist econometric methods. We consider the influence of prior information on coefficient estimates and model predictive accuracy have not been fully studied in the literature of PD modelling; therefore deserve the attention for future study.

We then reviewed the literature on credit risk stress testing in the second section of this chapter. We firstly categorised the stress testing methods based on whether the stress tests are at the macro or micro levels; whether the risk exposures being tested are the credit risk indicators of the lenders' balance sheet or the borrowers' credit loss parameters; and whether the stress testing approach used is the point prediction approach or the distribution approach. We then reviewed papers in each category. We found that in the literature, the credit losses, when unexpected stress happens, are higher than under tranquil economic conditions. Therefore banks should keep enough capital to cover any shortfalls due to unexpected credit losses. We found that papers tend to neglect model coefficient estimation errors with their use of the frequentist approach and model coefficient mean estimates. We consider the credit loss predictions provided in the literature may be underestimated as a source of risk - estimation risk - is neglected in the stress test modelling. Since the Bayesian coefficient posterior distribution includes the full ranges of possible coefficient estimates instead of a vector of mean estimates, we believe the use of the Bayesian approach may potentially reduce the estimation risk and credit risk underestimation problem.

Chapter 3

Improving Model Predictive Accuracy Using a Bayesian Approach: Application to PD Modelling of Mortgage Loans

3.1 Introduction

Frequentist and Bayesian approaches are the two fundamental approaches in econometrics. Bayesian approaches have the advantage that they can make use of prior information we have other than the data used for estimation. By adding additional information that is available but usually neglected into the modelling process, Bayesian approaches have the potential of improving model performance.

In the literature of credit risk modelling, the majority of papers use a frequentist approach in the PD modelling (Abdou et al., 2008; Abdou, 2009; Abellán & Mantas, 2014; Akkoç, 2012; Baesens et al., 2003; Bellotti & Crook, 2009; Chen et al., 2009; Crone & Finlay, 2012; Finlay, 2011; Hens & Tiwari, 2012; Kim & Sohn, 2010; Kruppa et al., 2013; Lessmann et al., 2015; Marqués et al., 2012b; Sinha & May, 2004; Tsai & Hung, 2014; Tsai & Wu, 2008; Wang & Ma, 2012; Zhang et al., 2010), while some use a Bayesian approach (Bijak & Matuszyk, 2017; Bijak & Thomas, 2015; Miguéis et al., 2013; Park et al., 2010). Among the papers using a Bayesian approach, many of them use non-informative priors on the basis that the likelihood normally overwhelms priors asymptotically with large data (Bijak & Thomas, 2015; Miguéis et al., 2013; Park et al., 2010). Some papers emphasise the importance of expert information and use that to gain informative priors (Jacobs & Kiefer, 2010; Kiefer, 2009). Mira and Tenconi (2004) use a Bayesian hierarchical logistic regression to model the probability of default. They take the expert prior opinion on PD as well as cross sectional correlations into consideration. They find that the Bayesian regressions outperform the frequentist models in predictive performance. Some papers use a more objective way to elicit priors. For instance, Maltritz and Molchanov (2008) use semi-automatic informative priors to study the determinants of sovereign default risk. Bijak and

Matuszyk (2017) use coefficient estimates of past data as priors for current data. However, we find that few do a comparative study between the two approaches. The second gap is the relatively rare use of informative priors. We consider that although non-informative priors can be used, using them misses the opportunity of including additional information available to the model, which is a major advantage of the Bayesian methodology and may improve the model's predictive accuracy.

This chapter makes the contribution that we propose a new method to gain Bayesian informative priors. In this method, we treat the values of a coefficient in the frequentist credit scoring model built on consecutive time periods as a time series and build an ARIMA model to forecast its value in future time periods. We use the ARIMA forecasts of the coefficients as priors for the Bayesian informative credit scoring models. This method of prior selection is relatively objective and systematic compared to the subjective ways of choosing Bayesian priors which have been a main point of criticism for Bayesian approaches (Koop, Poirier, & Tobias, 2007). Furthermore, we show that the informative Bayesian models using this method of the prior selection outperform both frequentist models and other types of Bayesian models based on various model performance measures, regardless of the time periods over which the models are built or the economic conditions. Secondly, we compare the frequentist and Bayesian approaches with non-informative and informative priors in modelling the probability of default. In contrast, the existing literature generally uses either one of the frequentist and Bayesian methods, while some papers (Bijak & Thomas, 2015) compare the frequentist approach and the Bayesian approach with only non-informative priors.

In this chapter, we model the probability of default of mortgage loans using both the frequentist and Bayesian approaches. We use cross-sectional logistic regression as the credit scoring model. The dataset for this research is the U.S mortgage loan data from the Freddie Mac database. We take random samples of mortgage loans originated before, during, and after the financial crisis of 2008 as three training samples. For all the samples we model the probability of default using the frequentist maximum likelihood estimation, Bayesian estimation with non-informative and informative priors, a Bayesian model averaging method, and a Bayesian model selection method. We then present comparative model performances for the frequentist and Bayesian models for three different samples. We find that the estimation results are similar between using the frequentist and Bayesian approach with non-informative priors. With informative priors, the Bayesian estimates tend to move towards the means of the prior distributions. In Bayesian model averaging, due to model combination, the coefficients for variables with low model probability are smaller compared to the corresponding frequentist estimates, representing a more modest effect of the explanatory variables on default. In terms of model performances, we find that our method of Bayesian prior selection using ARIMA forecasts of coefficients as priors outperform all other models, frequentist or Bayesian with different priors, on all samples.

The organisation of this chapter is as follows. It first introduces the methodology used in this paper such as the logistic regression model, the ARIMA model, the frequentist and Bayesian methods, the use of the ARIMA models in Bayesian prior selection, etc. It then presents the data and variables description, estimation results using frequentist and Bayesian approaches. Thirdly, it gives the post estimation diagnostics and performance measures for all models. This chapter finishes with a conclusion section.

3.2 Models and Methods

3.2.1 Logistic Regression Model

In the linear regression model, the dependent variable can range from $-\infty$ to $+\infty$. However, credit default probability cannot be smaller than 0 or larger than 1. Therefore we use a model for binary outcomes instead of the conditional mean to model the probability of events. Logistic regression is one of the possible binary dependent variable models and is commonly used in credit risk modelling. It limits the dependent variable to be within the (0, 1) range.

A standard logistic distribution function has the form:

$f(y) = \frac{1}{1+e^{-y}} \quad y \in (-\infty, +\infty)$, in which y is the expected value of the latent variable as a linear combination of explanatory variables.

Figure 3.1 shows the cumulative density distribution of $f(y)$, and as can be seen, the logistic function transforms the range of (-∞, +∞) into the range between 0 and 1.

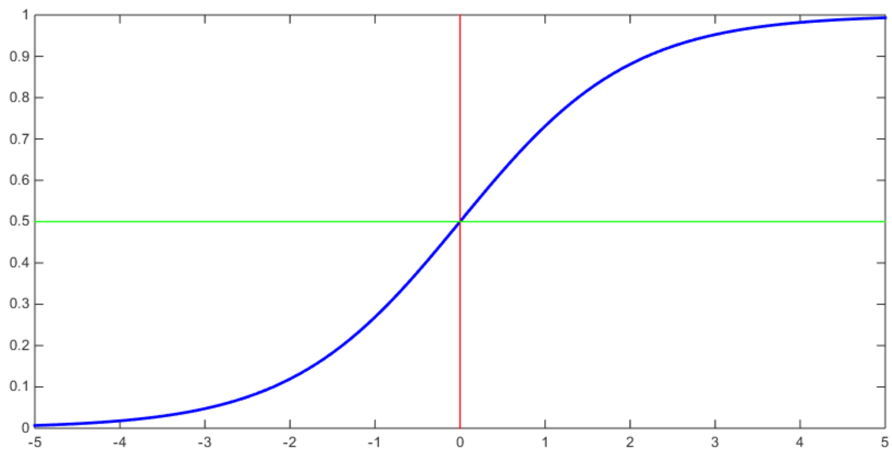

Fig. 3.1 CDF of a standard logistic function

Logistic regression applied in credit risk modelling takes the form:

$$p(d_i = 1 | x_i) = \frac{1}{1+e^{-x_i^{'}\beta}}, \ i = 1, 2, ..., n \qquad (3.1)$$

in which:

d_i denotes the credit default performance for account i. For each account, it either takes the value of 1 or 0, meaning default and non-default respectively.

$\boldsymbol{x}_i$ denotes a $k \times 1$ column vector of k explanatory variables for account i : $\begin{pmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,k} \end{pmatrix}$.

In our case these variables are the origination variables when the mortgage loan was originated. We expect them to have an influence on the probability of credit default.

$\boldsymbol{\beta}$ denotes a $k \times 1$ column vector of coefficients to be estimated and each element measures the extent of impact an explanatory variable has on credit default: $\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$.

$p(d_i = 1 | \boldsymbol{x}_i)$ denotes the probability of default conditional on the explanatory variables. It ranges from 0 to 1.

3.2.2 Autoregressive integrated moving average model

An autoregressive integrated moving average (ARIMA) model is commonly used in time series analysis to explore the pattern of data and to obtain forecasts of the series in future time periods. In an ARIMA model, the variable of interest is regressed against its own lagged terms and the lagged terms of the regression error, corresponding to the 'AR' and 'MA' part of the model, respectively. An initial step of differencing can be applied to the variable if it is originally non-stationary, hence the 'integrated' part of the model. An autoregressive moving average (ARMA) model, an autoregressive (AR) model, and a moving average (MA) model are all special cases of an ARIMA model.

An ARIMA(p,0,q) model can be written as a structural equation and a disturbance function:

$$y_t = \mathbf{x}_t^{'}\boldsymbol{\beta} + u_t$$
$$u_t = \sum_{l=1}^{p} \pi_l u_{t-l} + \sum_{m=1}^{q} \varpi_m \varepsilon_{t-m} + \varepsilon_t \tag{3.2}$$

In which

π_l denotes autocorrelation parameter for lag l

ϖ_m denotes moving average parameter for lag m

ε_t denotes regression disturbance that is assumed a white-noise process.

Substitute the second equation into the first and the two equations can be further written as:

$$y_t = \mathbf{x}_t^{'}\boldsymbol{\beta} + \sum_{l=1}^{p} \pi_l (y_{t-l} - \mathbf{x}_{t-l}^{'}\boldsymbol{\beta}) + \sum_{m=1}^{q} \varpi_m \varepsilon_{t-m} + \varepsilon_t \tag{3.3}$$

The above can be easily extended to the general ARIMA (p,d,q), when the original series is non-stationary, by replacing the original values of the series y_t with the differenced values Δy_t, and where d denotes the order of differencing.

3.2.3 Estimation Methods

3.2.3.1 Estimating logistic regression using a frequentist approach

To estimate logistic regression, a maximum likelihood method is the standard estimation method in the frequentist framework. The maximum likelihood estimator is an estimator that yields the parameter estimate that maximises the probability of observing the data, which is the likelihood, given specific assumptions about the distribution of y_i.

For independent observations, the joint probability density function of these observations is the product of their individual densities. This joint probability density function is the likelihood:

$$L(\boldsymbol{\beta}) = p(y_1, y_2, ..., y_n \mid \boldsymbol{\beta}) = \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\beta}) \tag{3.4}$$

The maximum likelihood method seeks to find the parameters that maximise this likelihood, and since the log function is monotone and simpler to work with, it is customary to maximise $\ln L(\boldsymbol{\beta})$.

For the logistic regression model, the likelihood function and log likelihood functions are derived as follows. Each observation of $y_i = 1$ or $y_i = 0$ is seen as a draw from a Bernoulli distribution, with the probability of $y_i = 1$ being p_i and the probability of $y_i = 0$ being $1 - p_i$, hence:

$$p(Y_i = y_i) = p_i^{y_i} (1 - p_i)^{(1-y_i)} \tag{3.5}$$

Since observations are independent of each other, the joint probability density function, or likelihood function, for a sample with *n* observations can be conveniently written as:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} (p_i)^{y_i} (1 - p_i)^{1-y_i} \tag{3.6}$$

Substituting $p(y_i = 1 | \boldsymbol{x}_i) = \dfrac{1}{1+e^{-x_i'\boldsymbol{\beta}}}$ into the likelihood function Eq. (3.6) we derive:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{x}'_i\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}'_i\boldsymbol{\beta}}}\right)^{y_i} \left(\frac{1}{1+e^{\boldsymbol{x}'_i\boldsymbol{\beta}}}\right)^{1-y_i} \tag{3.7}$$

Taking log of the likelihood function Eq. (3.6) and substituting in Eq. (3.7) we obtain the log likelihood function:

$$\begin{aligned} \ln L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} [y_i \ln p_i + (1-y_i)\ln(1-p_i)] \\ &= \sum_{i=1}^{n} [y_i \ln(\frac{1}{1+e^{-\boldsymbol{x}'_i\boldsymbol{\beta}}}) + (1-y_i)\ln(\frac{e^{-\boldsymbol{x}'_i\boldsymbol{\beta}}}{1+e^{-\boldsymbol{x}'_i\boldsymbol{\beta}}})] \end{aligned} \tag{3.8}$$

Taking the first derivative of $\ln L(\boldsymbol{\beta})$, the parameter estimate obtained achieves a local peak (either a maximum or a minimum) if:

$$\boldsymbol{g} = \frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{0}$$

For the logistic model the second derivative is always negative definite (Greene, 2011):

$$\boldsymbol{H} = \frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \quad \textit{is negative definite}$$

Therefore the maximum likelihood estimates achieve a global maximum of the likelihood.

3.2.3.2 Estimating logistic regression using a Bayesian Approach

Bayes theorem and posterior distribution

Bayesian econometrics is based on Bayes theorem:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \tag{3.9}$$

in which:

$\boldsymbol{y}$ denotes the observed data;

θ denotes a vector of parameters;

$p(\boldsymbol{\theta} \mid \boldsymbol{y})$ denotes the posterior density function (density function of the parameters given the data.);

$p(\boldsymbol{y} \mid \boldsymbol{\theta})$ denotes the likelihood function (density function of the data conditional on the parameters.);

$p(\boldsymbol{\theta})$ denotes a prior distribution of the parameters (non-data information we have prior to seeing the data.);

$p(\boldsymbol{y})$ denotes the marginal density function of the data.

Since $p(\boldsymbol{y})$ is the distribution of the data that does not include the parameters of interest, we treat it as constant. Therefore based on Bayes' rule in Eq. (3.9), the posterior density is proportional to the product of the prior density and the likelihood, which is:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \tag{3.10}$$

Therefore we update our prior understanding of parameters using the data information. The posterior that we obtain using the Bayesian approach is a combination of data and prior information.

Bayesian posterior function for logistic regression

In the Bayesian approach, for the regression parameters of the logistic regression we may use normal priors $N(\underline{\boldsymbol{\beta}},\ \underline{\boldsymbol{V}})$, in which we use parameters with underscores (i.e. $\underline{\boldsymbol{\beta}}$ and $\underline{\boldsymbol{V}}$) to denote hyper-parameters of the prior. $\underline{\boldsymbol{\beta}}$ is a k dimensional vector of prior means: $\underline{\boldsymbol{\beta}} = \begin{pmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \\ \vdots \\ \underline{\beta}_k \end{pmatrix}$

$\underline{\boldsymbol{V}}$ is the $k \times k$ prior covariance matrix: $\underline{\boldsymbol{V}} = \begin{pmatrix} \underline{\sigma}_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \underline{\sigma}_k^2 \end{pmatrix}$

The likelihood of the logistic model has a logistic distribution.

The posterior distribution is the product of the prior and the likelihood function Eq. (3.6). Therefore the posterior function takes the form:

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) \propto N(\underline{\boldsymbol{\beta}},\ \underline{\boldsymbol{V}}) L(\boldsymbol{\beta})$$
$$\Rightarrow p(\boldsymbol{\beta} \mid \boldsymbol{y}) \propto \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi}\underline{\sigma}_j} \exp[-\frac{(\beta_j - \underline{\beta}_j)^2}{2\underline{\sigma}_j^2}] \prod_{i=1}^{n} (p_i)^{y_i} (1-p_i)^{1-y_i} \tag{3.11}$$

The posterior distribution for logistic regression has no closed form, and we cannot obtain the moments of the posterior distribution in an analytical manner by integration. Therefore we use the simulation approach to approximate the characteristics of the posterior distribution.

Bayesian simulation

With data and prior distributions, we obtain the posterior distribution. For presentation purposes, and to have features comparable to the frequentist approach, we can derive point or interval estimates of the coefficients.

In some cases, the posterior has a common distribution form; the point estimates of the posterior distribution can be derived analytically. For instance, by taking the expectation of the posterior distribution, we get the first moment point estimate:

$E(\boldsymbol{\theta} \mid \boldsymbol{y}) = \int \boldsymbol{\theta} p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}$. However, for most posterior distributions such integrals rarely have a closed form. Therefore in most cases, we need to use simulation based estimation to obtain characteristics of the posterior distribution.

With an infinite number of sampling instances from the posterior distribution, the averaged sample mean approximates the mean of the parameter distribution based on the weak law of large numbers. That is:

$$\frac{1}{S}\sum_{s=1}^{S} g(\boldsymbol{\theta}^{(s)}) \rightarrow E(g(\boldsymbol{\theta}) \mid \boldsymbol{y})$$

in which:

$\boldsymbol{\theta}$ denotes parameters, for instance regression coefficients $\boldsymbol{\beta}$

$g(\boldsymbol{\theta})$ denotes a function of interest.

S denotes the infinite number of random draws

Therefore posterior distribution characteristics can be deduced based on sample characteristics. With the posterior simulation methods, distribution characteristics such as the posterior mean and variance from the distribution can be obtained. The Markov chain Monte Carlo (MCMC) algorithms are commonly used simulation methods. Metropolis-Hastings algorithm is one of the most popular MCMC simulation algorithms. In this research, we use the random-walk Metropolis-Hastings algorithm for Bayesian posterior simulation.

Markov Chain Monte Carlo algorithms

Markov chain Monte Carlo (MCMC) are stochastic simulation algorithms used for simulating from a desired probability density distribution through the construction of a Markov chain. The Markov chain converges to an equilibrium distribution that is the desired distribution. Applied to Bayesian estimation, the equilibrium distribution is the Bayesian posterior distribution. Therefore the summary statistics of the posterior distribution can be obtained by analysing samples from the Markov chain. In a Markov chain, the current draw $\boldsymbol{\theta}^{(s)}$ only depends on the last draw $\boldsymbol{\theta}^{(s-1)}$: $p(\boldsymbol{\theta}^{(s)} | \boldsymbol{\theta}^{(s-1)}) = p(\boldsymbol{\theta}^{(s)} | \boldsymbol{\theta}^{(s-1)}, ..., \boldsymbol{\theta}^{(0)})$. Given specific assumptions, the distribution of the Markov chain $\boldsymbol{\theta}$ converges to the posterior distribution $p(\boldsymbol{\theta} | \boldsymbol{y})$ regardless of the initial state of the chain $\boldsymbol{\theta}^{(0)}$, as $s \to \infty$.

Metropolis-Hastings (M-H) algorithm

The Metropolis-Hastings algorithm is a generic posterior simulator that can be used in any models (Koop et al., 2007). Some posterior density distributions do not have a known form and it is hard to take draws from them directly. M-H algorithm uses a candidate generating density to approximate the posterior distribution $p(\boldsymbol{\theta} | \boldsymbol{y})$. The M-H algorithm is a Markov Chain Monte Carlo (MCMC) algorithm. In other words, in the candidate density function, the candidate draw of the parameters $\boldsymbol{\theta}^*$ depends on the last draw $\boldsymbol{\theta}^{(s-1)}$ of the parameters: $q(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{(s-1)})$ and all the draws form a chain.

We correct the difference between posterior and candidate generating density by not accepting all the draws of the candidate density function. The acceptance probability $\alpha(\boldsymbol{\theta}^{(s-1)}, \boldsymbol{\theta}^*)$ is the probability that we accept the candidate draw $\boldsymbol{\theta}^*$ as our current draw $\boldsymbol{\theta}^{(s)}$. $1 - \alpha(\boldsymbol{\theta}^{(s-1)}, \boldsymbol{\theta}^*)$ is the probability we reject the candidate draw $\boldsymbol{\theta}^*$ and stay with the last draw $\boldsymbol{\theta}^{(s-1)}$. Therefore when the acceptance probability is low in some area, $\boldsymbol{\theta}^{(s-1)}$ will swiftly move away from it, whereas if the acceptance probability is high in some area, the Markov chain tends to stay there (take many draws from that area). Since the acceptance probability is derived in a way that it is high (low) in areas where posterior probability is high (low), the M-H algorithm in effect gives more (less) weight to the area of high (low) posterior probability.

The acceptance probability of the candidate draw in the M-H algorithm takes the form:

$$\alpha(\boldsymbol{\theta}^{(s-1)},\boldsymbol{\theta}^{*}) = \min[\frac{p(\boldsymbol{\theta}^{*} \mid \boldsymbol{y})q(\boldsymbol{\theta}^{s-1} \mid \boldsymbol{\theta}^{*})}{p(\boldsymbol{\theta}^{s-1} \mid \boldsymbol{y})q(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}^{s-1})},1] \quad (3.12)$$

Since it is a probability, it is always less than 1, thus the 'min' operator. The accepting and rejecting a candidate draw based on the acceptance probability is equivalent to taking a random draw u from a standard uniform distribution $U(0,1)$, and $\boldsymbol{\theta}^{s} = \begin{cases} \boldsymbol{\theta}^{*} & if\ \frac{p(\boldsymbol{\theta}^{*} \mid \boldsymbol{y})q(\boldsymbol{\theta}^{s-1} \mid \boldsymbol{\theta}^{*})}{p(\boldsymbol{\theta}^{s-1} \mid \boldsymbol{y})q(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}^{s-1})} > u \\ \boldsymbol{\theta}^{s-1} & otherwise \end{cases}$ (Lynch, 2007).

Random walk chain Metropolis-Hastings algorithm:

A random walk chain Metropolis-Hastings is one of the most common strategies in choosing a candidate generating density. We use it in this research. The random walk chain Metropolis-Hastings algorithm makes no attempt to choose a candidate generating density similar to the posterior, but uses one that wanders widely across various regions, and takes draws of the density distribution in proportion with the posterior.

The draws from the candidate density distribution is a random walk process:

$$\boldsymbol{\theta}^{*} = \boldsymbol{\theta}^{(s-1)} + \boldsymbol{z}. \quad (3.13)$$

in which $\boldsymbol{z}$ is a symmetric distribution such as a multivariate normal distribution, centred over $\boldsymbol{\theta}^{(s-1)}$.

That is the draws are taken from either direction of the last draw. The acceptance probability ensures draws are taken in the appropriate direction. As the candidate density is a symmetric distribution centred over the previous draw, $q(\boldsymbol{\theta}^{s-1} \mid \boldsymbol{\theta}^{*}) = q(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}^{s-1})$ is always the case (Koop et al., 2007; Lynch, 2007). Therefore, together with Eq. (3.12), for the random walk Metropolis-Hastings algorithm, the acceptance probability takes the form:

$$\alpha(\boldsymbol{\theta}^{(s-1)},\boldsymbol{\theta}^{*}) = \min[\frac{p(\boldsymbol{\theta}^{*} \mid \boldsymbol{y})}{p(\boldsymbol{\theta}^{(s-1)} \mid \boldsymbol{y})},1] \tag{3.14}$$

in which:

$p(\boldsymbol{\theta} \mid \boldsymbol{y})$ denotes the posterior density function.

As can be seen from Eq. (3.14), the acceptance probability of the candidate draw in the candidate density is the ratio of the posterior probability of the candidate draw and the posterior probability of the last draw. The acceptance rate for the candidate draw is high if the posterior probability of the candidate draw is high compared to that of the last draw. For areas that the posterior probability density distribution has a high probability, more draws from the candidate distribution are accepted, and for areas where the posterior distribution has a low probability, we accept less draws from the candidate distribution. Therefore based on the number of draws in different areas of the candidate distribution, we approximate the posterior distribution.

As described in the previous paragraphs, the steps of M-H sampling are:

- Set the starting value: $\boldsymbol{\theta}^{(0)}$
- Draw $\boldsymbol{\theta}^{*}$ from candidate generating distribution $q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(s-1)})$
- Calculate the acceptance probability for the draw $\boldsymbol{\theta}^{*}$. Based on the acceptance probability we obtain the probability that the candidate draw is accepted and rejected, respectively:
  $p(\boldsymbol{\theta}^{(s)} = \boldsymbol{\theta}^{*}) = \alpha(\boldsymbol{\theta}^{(s-1)},\boldsymbol{\theta}^{*});\quad p(\boldsymbol{\theta}^{(s)} = \boldsymbol{\theta}^{(s-1)}) = 1-\alpha(\boldsymbol{\theta}^{(s-1)},\boldsymbol{\theta}^{*})$
- Repeat sampling S times until convergence is reached.
- Discard the burn-in period of the initial s_1 draws and use the $\boldsymbol{\theta}^{s_1+1},...,\boldsymbol{\theta}^{S}$ draws as the MCMC output sample.
- Obtain the summary statistics of the posterior distribution

3.2.4 Priors used for our Bayesian models:

The prior distribution for coefficients is set as multivariate normal. We use both non-informative priors and different sets of informative priors.

3.2.4.1 Non-informative priors:

For non-informative priors we use small prior precisions (large prior variances) indicating low confidence in the priors and high uncertainty about what the likely values of the coefficients are. We set the prior mean for each coefficient to 0. The prior precision of each coefficient is set to 0.000001. Since the prior variance is the inverse of prior precision ($\underline{\boldsymbol{V}}^{-1}$), the prior variance for each coefficient is set to 1000000:

$$\underline{\boldsymbol{\beta}} = \boldsymbol{0},\ \underline{\boldsymbol{V}} = \begin{pmatrix} 1000000 & & 0 \\ & \ddots & \\ 0 & & 1000000 \end{pmatrix}.$$

Compared to the coefficient estimation results obtained from the frequentist approach, these prior hyper parameters are rather non-informative since in our frequentist estimation the standard errors estimated for the coefficients in each regression are rarely more than 1.

3.2.4.2 Informative priors:

For informative priors, we use the coefficient estimates from a frequentist approach based on earlier or later samples of our data, as well as ARIMA forecasts of the coefficients as priors.

Naive forecasts of coefficients as priors

The naive forecast method uses the observed value of the last period as a forecast for the subsequent period. The coefficient mean estimates and standard errors obtained from frequentist logistic models based on accounts originated 1 year prior to the origination of each training data are used as prior means and standard deviations for the Bayesian logistic models. We call these priors naive forecasts. More specifically, frequentist estimates of 2003 are used as priors for Bayesian models trained on the 2004 data. Frequentist estimates of 2006 are used as priors for Bayesian models trained on the 2007 data. Frequentist estimates of 2010 are used as priors for Bayesian models trained on the 2011 data.

Updating

By the time the test data is available, all information about accounts originated prior to the time period of the test data is available. Therefore we can use newer information as priors to update models built earlier. Based on this argument, the coefficient mean estimates and standard errors obtained from frequentist logistic models based on accounts originated 1 year prior to the origination of the *test data* cases are used as the prior means and standard deviations for the Bayesian logistic models. We call this prior information 'updating'. More specifically, frequentist estimates of 2013 are used as priors for Bayesian models trained on the 2004, 2007, and 2011 samples.

ARIMA(p,d,q) forecasts of coefficients as priors:

Firstly, we estimate separate cross-sectional models for the probability of default over 24 months outcome period using the frequentist method based on accounts originated in each quarter, up until the time period of the test data. We have 60 different quarters of origination in the dataset (1999q1 – 2013q4) and so estimated

60 cross section logistic regression models. For each coefficient, we collect the estimation results obtained based on these consecutive quarterly datasets and treat them as a time series variable.

Secondly, we fit the estimation results of all the coefficient variables obtained in the first step using ARIMA models. We then make forecasts of these coefficient variables into the time period of the test sample. We use these ARIMA forecasts as informative priors in the Bayesian logistic models.

1) ARIMA forecasts of coefficients as priors (static forecast): Bayesian models with informative priors obtained from static ARIMA forecasts of the coefficients.

Since by the time the test data is available, all information of accounts originated prior to that is available, the ARIMA estimation is done on the time series of coefficient mean estimates obtained from PD models based on accounts originated in each quarter between 1999q1 and 2013q4. One step ahead static forecasts for the first quarter of 2014 are used as prior means for coefficients in the informative Bayesian logistic model trained on pre-crisis, crisis, and post-crisis training samples.

That is, we model the estimated values of a coefficient in the frequentist PD model using the ARIMA model $\Delta y_t = \alpha + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \sum_{j=1}^{q} \gamma_j \varepsilon_{t-j} + \varepsilon_t$, where y_t denotes the time series of a coefficient's estimates obtained from repeatedly estimating a frequentist cross sectional PD model based on accounts originated in each quarter between 1999q1 and 2013q4.

We then make a static forecast into the first quarter of the test sample: $\Delta \hat{y}_f = \hat{\alpha} + \sum_{i=1}^{p} \hat{\beta}_i \Delta y_{f-i} + \sum_{j=1}^{q} \hat{\gamma}_j \varepsilon_{f-j}$. The prediction at period f is based on the actual values of lagged dependent variable and error terms up to the period $f-1$. These

ARIMA forecasts of the coefficients are used as prior means for the Bayesian PD model.

2) ARIMA forecasts of coefficients as priors (dynamic forecast): Bayesian models with informative priors obtained from dynamic ARIMA forecasts of the coefficients.

ARIMA models are estimated based on coefficient data from 1999q1 to the last quarter of the training samples instead of 2013q4, and a dynamic forecast instead of a static forecast is used. Dynamic forecasts of the first quarter of 2014 are used as prior means for the coefficients. The dynamic forecast is only applied to the post-crisis training sample.

That is, the ARIMA model is: $\Delta y_t = \alpha + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \sum_{j=1}^{q} \gamma_j \varepsilon_{t-j} + \varepsilon_t$, where y_t denotes a time series of coefficient estimates obtained from repeatedly estimating frequentist cross sectional PD models based on accounts originated in each quarter from 1999q1 to the end of the latest training sample 2011q4. We then make a dynamic forecast into the first quarter of the test sample. A dynamic ARIMA forecast at period $f+1$ is: $\Delta \hat{y}_{f+1} = \hat{\alpha} + \sum_{i=1}^{p-1} \hat{\beta}_i \Delta y_{f-i} + \hat{\beta}_0 \Delta \hat{y}_f + \sum_{j=1}^{q-1} \hat{\gamma}_j \varepsilon_{f-j} + \hat{\gamma}_0 \hat{\varepsilon}_f$. The forecast at period $f+1$ is based on the *forecasted* values of the dependent variable and error term at period f, and the *observed* values of lagged terms of the dependent variable and error term prior to f. That is, in the case of the dynamic forecasts for our coefficient variables, the forecast value of a coefficient variable of the first quarter after the training sample is based on the *observed* value up to the last quarter of the training sample. The forecasts for quarters afterwards are based on the *forecasted* values of that coefficient variable instead of observed values.

The above explains how the prior mean for a coefficient is obtained. As for prior standard deviation for a coefficient, we try two sets of prior standard deviations: 1.

the standard error of the static ARIMA forecast for that coefficient; 2. the static ARIMA forecast for the times series of the coefficient standard errors obtained from the frequentist models built on the quarterly samples. Same as the ARIMA forecast for the coefficient mean, the ARIMA forecast for the coefficient standard error is also for the first quarter of the test sample.

3.2.5 Posterior model probability and Bayesian model averaging

In this section, we introduce the Bayesian posterior model probability and the method of Bayesian model averaging.

The Bayes rule applied in a statistical model with data y and model M_a provides the posterior model probability:

$$p(M_a \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid M_a) p(M_a)}{p(\boldsymbol{y})} \tag{3.15}$$

In which:

$p(M_a \mid \boldsymbol{y})$ denotes the posterior model probability for model M_a conditional on the data;

$p(\boldsymbol{y} \mid M_a)$ denotes the likelihood distribution of data conditional on the model M_a;

$p(M_a)$ denotes prior belief of the probability of model M_a, and $\sum_{a=1}^{n} p(M_a) = 1$;

$p(\boldsymbol{y})$ denotes the marginal probability density of the data.

Since the marginal probability density of the data $\boldsymbol{y}$ is difficult to calculate in practice, it is common to obtain the posterior model probability by comparing models. Models are compared using the posterior odds ratio. The posterior odds ratio between two models (M_1 and M_2) is:

$$PO_{1,2} = \frac{p(M_1 \mid \boldsymbol{y})}{p(M_2 \mid \boldsymbol{y})} = \frac{p(\boldsymbol{y} \mid M_1) p(M_1)}{p(\boldsymbol{y} \mid M_2) p(M_2)} \tag{3.16}$$

The higher the posterior odds ratios are, the stronger evidence there is against the use of model M_2.

If we assume the model priors are equal for different models, then the posterior odds ratio is decided by the marginal likelihoods of models, and the ratio is called a Bayes factor:

$$BF_{1,2} = \frac{p(\boldsymbol{y} \mid M_1)}{p(\boldsymbol{y} \mid M_2)} \tag{3.17}$$

The posterior model probability for a model can be obtained by comparing two models through the posterior odds ratio or Bayes factor. For instance, if there are two potential models M_1 and M_2 :

$$\begin{aligned} p(M_1 \mid \boldsymbol{y}) &= \frac{PO_{1,2}}{1 + PO_{1,2}} \\ p(M_2 \mid \boldsymbol{y}) &= \frac{1}{1 + PO_{1,2}} \end{aligned} \tag{3.18}$$

In which:

$PO_{1,2}$ denotes the posterior odds ratio between models M_1 and M_2

In choosing the most appropriate set of predictors, we can use Bayesian model selection or Bayesian model averaging. In Bayesian model selection, the one model that has the highest posterior model probability is used. Bayesian model averaging uses a weighted average over all potential models, with the weight for each model being its posterior model probability.

The posterior distribution and posterior mean estimates in Bayesian model averaging are:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \sum_{a=1}^{n} p(\boldsymbol{\theta} \mid \boldsymbol{y}, M_a) p(M_a \mid \boldsymbol{y}) \tag{3.19}$$

and:

$$E(\boldsymbol{\theta} \mid \boldsymbol{y}) = \sum_{a=1}^{n} E(\boldsymbol{\theta} \mid \boldsymbol{y}, M_a) p(M_a \mid \boldsymbol{y}) \quad (3.20)$$

in which

M_a denotes the ath model

n denotes the total number of models

Eq. (3.19) shows that the posterior distribution in a Bayesian model averaging method is the weighted average of the posterior distributions of all the n models, and the weight is the posterior model probability of each of the n models. Eq. (3.20) shows that the posterior mean estimates in a Bayesian model averaging method is the weighted average of the posterior mean estimates in all the n models, and the weight is the posterior model probability of each of the n models. Koop (2007) argues that Bayesian model averaging is better than model selection in that it does not simply discard the models that have lower model probabilities, instead it still takes them into account, but with less weight.

3.3 Data and Variables

3.3.1 Data

The data for this research is the Freddie Mac single-family loan-level dataset. The loans are fully amortising long term mortgages. We use the mortgage accounts that originated in 2004, 2007, and 2011 as our training sample0, training sample1, and training sample2, respectively. We use accounts originated in 2014 as the test sample. For each sample we take a 24 months observation period. We sample this way so the length of the origination and observation periods are the same for all samples. These periods cover pre-crisis, financial crisis, and post-crisis circumstances. For each year we randomly select 50000 loans. We consider an account is in default if it has in its payment history a record of delinquency of at least 90 days. We delete the loans that have missing values of delinquency. Therefore we have a dataset of 198,906 loans altogether including training and test samples. Table 3.1 shows the number of loans, defaults and default rates in each sample.

Table 3.1 Training and test samples

	Train 0 (2004) Pre-crisis	Train 1 (2007) crisis	Train 2 (2011) Post-crisis	Test (2014)
Number of loans	49432	49975	49499	50000
Defaults	441	2344	106	171
Default rate	0.89%	4.7%	0.21%	0.34%
All samples have a 24 month observation period				

3.3.2 Variables

For our cross sectional binomial logistic regression, the dependent variable is either 1 (default) or 0 (non-default) for each account. There are 3062 defaults in total in our samples. The explanatory variables we use include continuous variables such as debt to income ratio, original interest rate, etc., categorical variables such as loan purpose. Table 3.2 gives a full list of the explanatory variables for this research.

Table 3.2 Full list of explanatory variables

Group	Variable name	Definition
Continuous variables	original interest rate	The original interest rate in mortgage note
	mortgage insurance percentage	Percentage of mortgage that is covered by insurance
	original loan term	The mortgage's number of scheduled monthly payments based on the first payment date and date of maturity in mortgage notes
	log original UPB	Log of the original unpaid balance of the mortgage
	original combined loan to value	(original mortgage loan amount+ secondary mortgage loan)/property appraised value or purchase price
	original debt-to-income ratio	the sum of monthly debt/sum of monthly income calculated at loan origination
	original loan-to-value	Original loan amount/appraised loan value or purchase price
	number of units	The number of units in the property
Categorical variables	loan purpose	Whether the mortgage is a purchase mortgage (P), cash-out refinance mortgage (C) or no cash-out refinance mortgage (N)
	number of borrowers	The number of borrowers who are obligated to repay the mortgage note secured by the mortgaged property. It is in a categorical form of either 1 borrower or more than 1 borrower.

Source: Freddie Mac database

We include these explanatory variables for the following reasons. We expect a higher interest rate to discourage borrowers from paying back loans, therefore to have a positive impact on default. For an insured loan, a mortgage insurer will cover the losses if the borrower defaults. Therefore, we expect the higher the mortgage insurance percentage, the higher the probability of default. We expect indicators of the loan size, e.g. the original unpaid balance to influence default. This impact could be positive since the larger the loan size, the more difficult the task of repayment is. We expect the loan to value ratios and debt to income ratio to have positive signs since the low level of these ratios shows a high capability of the borrowers to repay. We expect the number of borrowers to have a negative sign since multiple borrowers share the risk of default. We consider different loan purposes to have different impact on loan default. For instance, refinancing is sometimes undertaken by borrowers facing financial difficulty and seek to reduce monthly payment. Therefore we expect a refinance mortgage to have a higher default rate than a purchase mortgage.

Table 3.3 gives the descriptive statistics for the 'coefficient variables' obtained from the logistic models using frequentist estimation based on quarterly mortgage loan data originated between 1999q1 and 2013q4. We have 12 coefficient time series in total with each having 60 observations. ARIMA models are later used to fit these series and forecasts are made to serve as informative priors in Bayesian logistic models.

Table 3.3 Descriptive statistics for the 'coefficient variables' obtained from logistic models using frequentist estimation based on quarterly data originated between 1999q1 and 2013q4

Name	Descriptive statistics				
	Observations	Mean	SD	Min	Max
Intercept	60	-21.4181	15.1527	-79.4815	-79.4815
original_loan_to_value	60	0.0238	0.0247	-0.0177	0.1019
original_combined_loan_to_value	60	0.0029	0.0218	-0.0698	0.0304
original_debt_to_income_ratio	60	0.0277	0.0151	0.0065	0.0563
original_interest_rate	60	1.2536	0.2615	0.7401	1.9405
mortgage_insurance_percentage	60	0.0162	0.0053	-0.0036	0.0267
number_of_units	60	-0.1596	0.2637	-1.1387	0.3407
log_original_upb	60	-0.1657	0.3712	-0.7536	0.6994
original_loan_term	60	0.0226	0.0414	-0.0063	0.1823
as.factor(n_loan_purpose)2	60	0.7672	0.2051	0.3406	1.1304
as.factor(n_loan_purpose)3	60	0.5837	0.2835	-0.1567	1.0925
as.factor(number_of_borrowers)2	60	-0.8374	0.1444	-1.1963	-0.6286

3.4 Results

3.4.1 ARIMA models for coefficient forecasts to be used as informative priors in Bayesian logistic models

ARIMA models require stationary time series whose mean and autocovariance do not depend on time. If the original time series is nonstationary, integration i.e. differencing operations are needed to make it stationary before it can be fitted in an ARIMA model. The order of integration is the number of times a variable must be differenced to make a nonstationary series stationary. It equals the number of unit roots in the series. A differenced stationary series is denoted as I(d) in which d is the order of integration or the number of unit roots. For instance, a stationary variable is an I(0) process while a series with 1 unit root is an I(1) process.

Therefore before fitting the time series coefficient variables using ARIMA models, we first use different unit root tests to check the stationarity of these coefficient series to decide the order of integration needed for the series to become stationary if the original series is nonstationary. These unit root tests include Augmented Dickey-Fuller test, GLS-detrended Dickey-Fuller test, and KPSS test.

For instance, to test the unit roots of an AR(1) process, one models $y_t = \rho y_{t-1} + \boldsymbol{x}_t'\boldsymbol{\varsigma} + \varepsilon_t$, in which $\boldsymbol{x}_t$ are exogenous variables which may include a constant and a trend, ρ and $\boldsymbol{\varsigma}$ are coefficients to be estimated, and ε_t is the error term which is assumed to be a white noise process if the AR model fits the time series well. The variable y_t is a nonstationary series if $|\rho| \geq 1$ and stationary if $|\rho| < 1$. Therefore the stationarity of a series can be tested based on the relationship between $|\rho|$ and 1. Different unit root tests have different null hypotheses. For Augmented Dickey-Fuller test and GLS-detrended Dickey-Fuller test, the null hypothesis is nonstationarity against the alternative hypothesis of stationarity, i.e. $H_0 : \rho = 1$ against $H_1 : \rho < 1$. In other words, in these two tests the series is stationary if the null hypothesis is rejected. For KPSS test, the null hypothesis is stationarity against the alternative hypothesis of nonstationarity, i.e. $H_0 : \rho < 1$ against $H_1 : \rho = 1$. In other words, in this test the series is stationary if we cannot reject the null hypothesis.

To decide the ARIMA models used for coefficient forecasts, we first examine the coefficient series and their differenced series. If the series is non-stationary, we first difference the variables to remove the nonstationarity. Fig 3.2 – Fig 3.3 and Table 3.4 give an example of the original and differenced series. They show that the original series is non-stationary and is stationary after first differencing. Therefore for this coefficient variable, we choose an I(1) term in the ARIMA model. That is, we use an ARIMA(p,1,q) model.

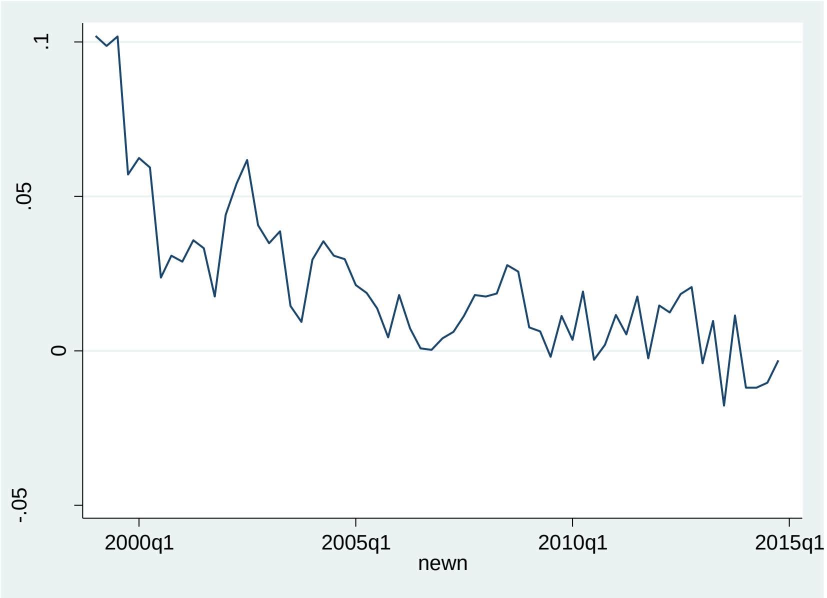

Fig. 3.2 The coefficient for 'original_loan_to_value'

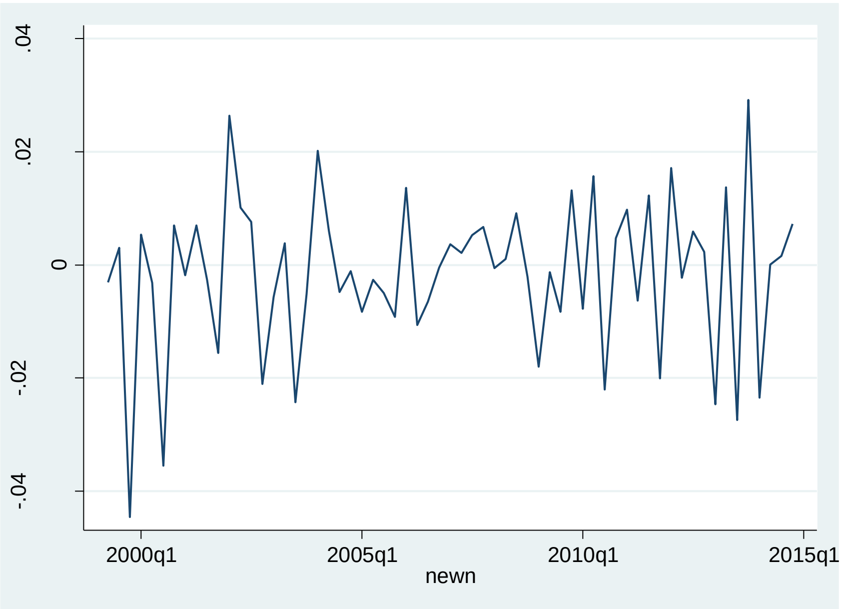

Fig. 3.3 The first differenced term of the coefficient for 'original_loan_to_value'

Table 3.4 Unit root tests of the coefficient for 'original_loan_to_value'

Unite Root Test / Variables	ADF test		DF-GLS test		KPSS test		Conclusion
	P-value	Stationarity	T-statistic	Stationarity	LM-stat	Stationarity	
Coefficient variable: 'original_loan_to_value'							
Level	0.1836	No	-1.2897	No	0.1645**	No	nonstationry
1st Dif	0.0055***	Yes	-5.133***	Yes	0.0803	Yes	stationry

We then check the differenced term of the coefficient variable's autocorrelations (AC) and partial autocorrelations (PAC) for a recommendation of autoregressive (AR) and moving averaging (MA) terms. Fig 3.4 – Fig 3.5 show the AC and PAC for the differenced coefficient variable. The AC figure recommends 1 MA term: MA(1). The PAC figure recommends 1 AR term: AR(1).

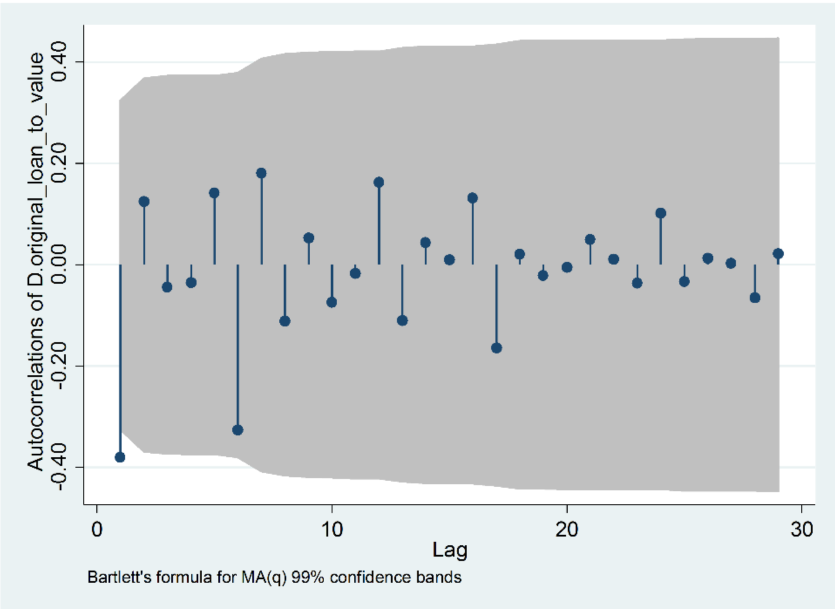

Fig. 3.4 Autocorrelation for the first differenced term of the coefficient for 'original_loan_to_value'

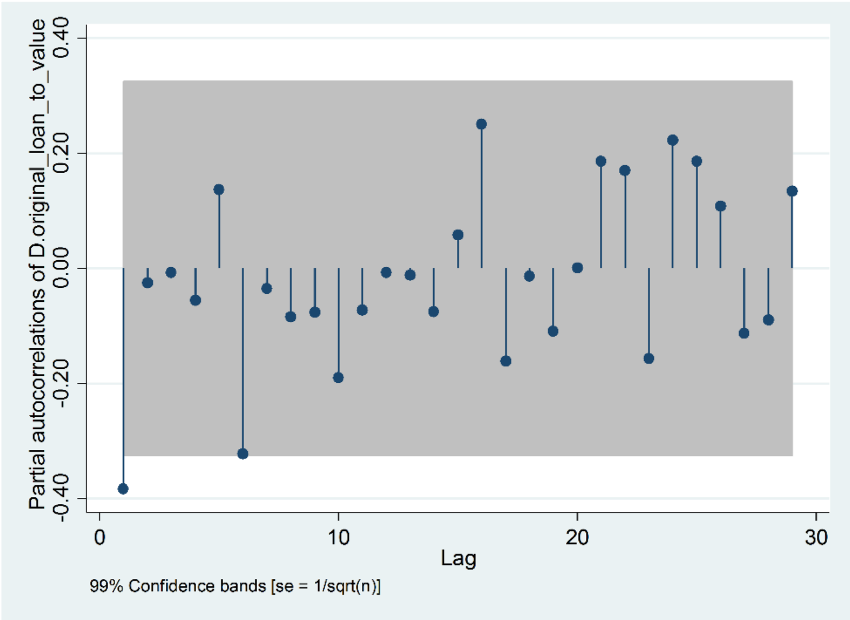

Fig. 3.5 Partial Autocorrelation for the first differenced term of the coefficient for 'original_loan_to_value'

Based on the recommendations from the autocorrelations (1 MA term), partial correlations (1 AR term), and the unit root test results (first order integration), we try three models: ARIMA(1,1,1), ARIMA(1,1,0) and ARIMA(0,1,1). We decide the final model based on the information criteria of the models, shown in Table 3.5.

Table 3.5 Information criteria for potential ARIMA models for the series "coefficient for 'original_loan_to_value'"

Model	Observations	Log likelihood	AIC	BIC
arima(1,1,1)	59	**172.3**	-336.584	**-328.2739**
arima(1,1,0)	59	**172.3**	**-338.5565**	-332.3238
arima(0,1,1)	59	172.0203	-338.0407	-331.8081

Model ARIMA(1,1,1) and ARIMA(1,1,0) are both good according to multiple information criteria. We choose ARIMA(1,1,0) model based on model parsimony and the significance of lagged AR and MA terms.

Fig 3.6 presents the autocorrelation plot for the residual in the ARIMA model. The correlation between the disturbance of the ARIMA regression and its own lags are relatively low which resembles a white noise process, showing that no pattern is remained in the disturbance term after fitting the data using the ARIMA model, and that our ARIMA model has taken into account most of the information in the series.

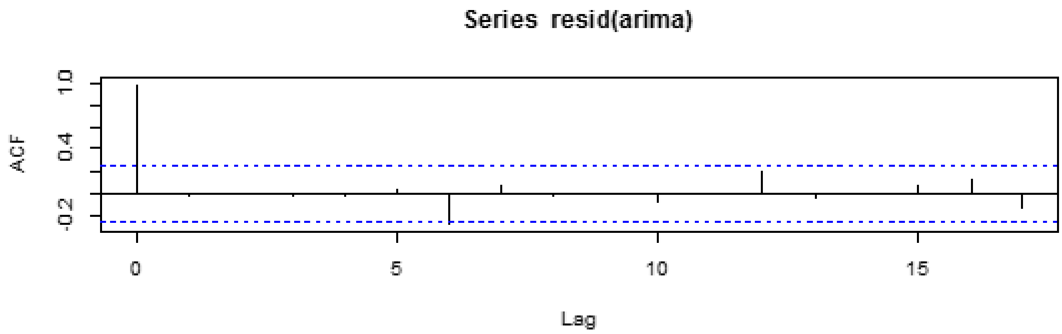

Fig. 3.6 Autocorrelation for the residual in the ARIMA model for the coefficient for 'original_loan_to_value'

We then make both static and dynamic forecasts of the coefficient using the ARIMA(1,1,0) model listed above, and use the forecasts of the coefficient in the first quarter of 2014 as coefficient prior means for that coefficient for the Bayesian logistic models. The forecasts for the coefficient of the variable 'original_loan_to_value' is shown in Fig 3.7. The blue shades show 95% and 80% confidence intervals of the static forecasts.

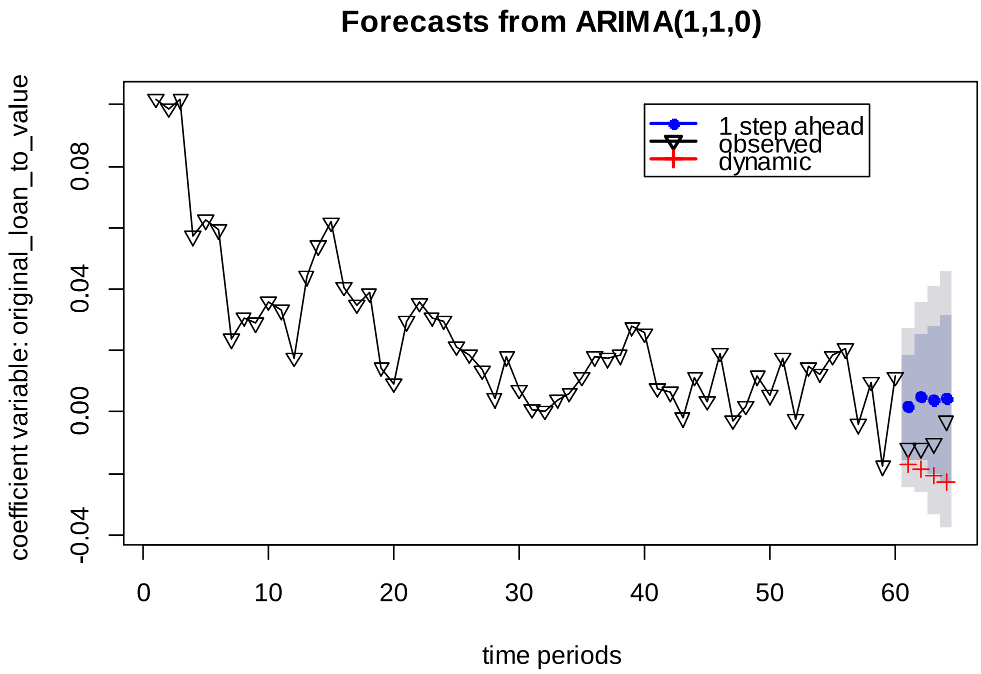

Fig. 3.7 static and dynamic forecasts and observed values of the coefficient for 'original_loan_to_value'

Based on the unit root tests, autocorrelations, partial correlations, preliminary estimations, and information criteria, we use the following ARIMA models for our data fitting and coefficient forecasting for all the coefficient time series, shown in Table 3.6.

Table 3.6 ARIMA models for the coefficients to make ARIMA forecasts as prior means in Bayesian logistic models

coefficient variable	ARIMA model used for forecasts
Intercept	ARIMA (4,1,0)
original_loan_to_value	ARIMA (1,1,0)
original_combined_loan_to_value	ARIMA (6,1,0)
original_debt_to_income_ratio	ARIMA (0,1,5)
original_interest_rate	ARIMA (0,1,1)
mortgage_insurance_percentage	ARIMA (6,1,1)
number_of_units	ARIMA (4,1,0)
log_original_upb	ARIMA (5,1,0)
original_loan_term	ARIMA (4,1,0)
as.factor(n_loan_purpose)2	ARIMA (0,1,1)
as.factor(n_loan_purpose)3	ARIMA (0,1,2)
as.factor(number_of_borrowers)2	ARIMA (2,1,0)

As for prior standard deviations, we use two sets of prior standard deviations. The first set is the standard errors of the ARIMA forecasts of the coefficient means, which are obtained simultaneously when the ARIMA forecasts of coefficient means are obtained. The second set is the ARIMA forecasts of the coefficient standard errors in the frequentist logistic regressions based on the consecutive quarterly data. In Bayesian models with ARIMA forecasts as informative priors, we report the estimation and model performance results with the set of prior standard deviations that gives the better performances.

3.4.2 Estimation results in logistic models

We use our samples in the following way. We estimate models on each of the training samples using the frequentist approach, the Bayesian approach with non-informative priors, the Bayesian approach with different sets of informative priors, Bayesian model averaging, and Bayesian model selection. Table 3.7 - Table 3.9 show the estimation results using the frequentist approach and Bayesian approach with non-informative and informative priors based on training samples 0, 1 and 2.

Table 3.7 Estimation Results using the frequentist approach and the Bayesian approach with non-informative and informative priors based on training sample 0

	Pre-crisis training sample originated in 2004 with 24 months observation period				
Variables	Estimate (std.error)	Posterior mean (std.dev)			
	Frequentist	Bayesian			
		Non-informative	Informative	Informative	Informative
		Non-informative	naive forecast	ARIMA forecast (static)	Bayesian updating
Intercept	-7.016 (1.601)	-6.987 (1.581)	-10.041 (0.053)	-10.170 (0.373)	-8.751 (0.091)
original_loan_to_value	0.011 (0.014)	0.013 (0.014)	0.007 (0.001)	-0.0004 (0.001)	-0.026 (0.001)
original_combined_loan_to_value	0.012 (0.013)	0.011 (0.013)	0.028 (0.001)	0.008 (0.001)	0.036 (0.001)
original_debt_to_income_ratio	0.013** (0.004)	0.013 (0.004)	0.020 (0.001)	0.043 (0.001)	0.047 (0.002)
original_interest_rate	0.863*** (0.127)	0.858 (0.129)	0.943 (0.009)	1.228 (0.015)	0.967 (0.019)
mortgage_insurance_percentage	0.027*** (0.006)	0.027 (0.006)	0.025 (0.002)	0.018 (0.00004)	0.029 (0.003)
number_of_units	0.145 (0.221)	0.107 (0.225)	0.159 (0.049)	-0.466 (0.013)	-0.855 (0.088)
log_original_upb	-0.486*** (0.105)	-0.490 (0.104)	-0.368 (0.005)	-0.308 (0.030)	-0.314 (0.008)
original_loan_term	0.000 (-)	0.000 (-)	0.000 (-)	0.000 (-)	0.000 (-)
loan_purpose:					
P(Excluded category)					
C	0.876*** (0.134)	0.879 (0.135)	0.823 (0.083)	0.617 (0.011)	0.643 (0.102)
N	0.848*** (0.126)	0.834 (0.126)	0.709 (0.070)	0.166 (0.003)	0.586 (0.102)
Number_of_borrowers:					
1(excluded category)					
> 1	-0.614*** (0.106)	-0.619 (0.107)	-0.689 (0.068)	-0.996 (0.006)	-0.733 (0.088)
Log likelihood = -2175.5589 Prob > chi2 = 0.0000		Number of draws in MCMC = 100000 Burn-in = 200000			

Table 3.8 Estimation Results using the frequentist approach and the Bayesian approach with non-informative and informative priors based on training sample 1

	Crisis training sample originated in 2007 with 24 months observation period				
Variables	Estimate (std.error)	Posterior mean (std.dev)			
	Frequentist	Bayesian			
		Non-informative	Informative	Informative	Informative
		Non-informative	naive forecast	ARIMA forecast (static)	Bayesian updating
Intercept	-23.007*** (0.710)	-23.016 (0.709)	-14.654 (0.034)	-7.260 (0.038)	-9.017 (0.092)
original_loan_to_value	0.011* (0.005)	0.011 (0.005)	0.009 (0.0004)	-0.003 (0.0005)	-0.022 (0.001)
original_combined_loan_to_value	0.020*** (0.004)	0.020 (0.004)	0.019 (0.0004)	0.007 (0.0005)	0.038 (0.001)
original_debt_to_income_ratio	0.023*** (0.002)	0.023 (0.002)	0.021 (0.001)	0.034 (0.0009)	0.040 (0.001)
original_interest_rate	1.176*** (0.051)	1.176 (0.051)	0.891 (0.005)	1.087 (0.009)	0.994 (0.018)
mortgage_insurance_percentage	0.013*** (0.003)	0.013 (0.003)	0.021 (0.001)	0.022 (0.001)	0.024 (0.002)
number_of_units	0.014 (0.084)	0.009 (0.084)	0.052 (0.030)	-0.679 (0.036)	-0.511 (0.077)
log_original_upb	0.606*** (0.045)	0.606 (0.045)	0.108 (0.003)	-0.189 (0.003)	-0.264 (0.007)
original_loan_term	0.004*** (0.001)	0.004 (0.001)	0.003 (0.0001)	-0.004 (0.0001)	-0.0003 (0.0002)
Loan_purpose:					
P(Excluded category)					
C	0.892*** (0.060)	0.894 (0.060)	1.014 (0.037)	0.337 (0.040)	0.798 (0.053)
N	0.879*** (0.061)	0.878 (0.061)	0.960 (0.045)	0.402 (0.047)	0.819 (0.056)
Number_of_borrowers:					
1(excluded category)					
> 1	-0.692*** (0.047)	-0.693 (0.047)	-0.645 (0.036)	-0.795 (0.037)	-0.584 (0.045)
Log likelihood =-8043.2411 Prob > chi2 = 0.0000		Number of draws in MCMC = 100000 Burn-in = 200000			

Table 3.9 Estimation Results using the frequentist approach and the Bayesian approach with non-informative and informative priors based on training sample 2

	Post-crisis training sample originated in 2011 with 24 months observation period					
Variables	Estimate (std.error)	Posterior mean (std.dev)				
	Frequentist	Bayesian				
		Non-informative	Informative	Informative	Informative	Informative
		Non-informative	naive forecast	ARIMA forecast (static)	ARIMA forecast (dynamic)	Bayesian updating
Intercept	-17.849***	-17.929	-13.783	-9.435	-8.722	-8.854
	(2.491)	(2.518)	(0.085)	(0.439)	(0.438)	(0.094)
original_loan_to_value	0.026	0.042	-0.010	-0.001	-0.016	-0.026
	(0.037)	(0.041)	(0.001)	(0.001)	(0.001)	(0.001)
original_combined_loan_to_value	-0.005	-0.019	0.038	0.008	0.025	0.037
	(0.037)	(0.041)	(0.001)	(0.001)	(0.001)	(0.001)
original_debt_to_income_ratio	0.064***	0.065	0.051	0.044	0.060	0.061
	(0.012)	(0.013)	(0.002)	(0.001)	(0.001)	(0.002)
original_interest_rate	1.928***	1.919	1.566	1.240	1.743	1.055
	(0.264)	(0.268)	(0.017)	(0.015)	(0.015)	(0.021)
mortgage_insurance_percentage	0.013	0.012	0.013	0.018	0.013	0.018
	(0.012)	(0.012)	(0.008)	(0.00004)	(0.00004)	(0.005)
number_of_units	0.050	-0.049	0.051	-0.466	-0.445	-0.857
	(0.307)	(0.337)	(0.080)	(0.013)	(0.013)	(0.093)
log_original_upb	-0.049	-0.049	-0.335	-0.251	-0.540	-0.285
	(0.173)	(0.173)	(0.007)	(0.032)	(0.033)	(0.008)
original_loan_term	-0.001	-0.001	0.001	-0.002	-0.004	-0.001
	(0.002)	(0.002)	(0.0003)	(0.001)	(0.001)	(0.0003)
Loan_purpose:						
P(Excluded category)						
C	0.577*	0.587	1.244	0.619	0.741	0.483
	(0.257)	(0.261)	(0.106)	(0.011)	(0.011)	(0.142)
N	0.149	0.142	0.400	0.165	0.367	-0.045
	(0.274)	(0.275)	(0.156)	(0.003)	(0.003)	(0.152)
Number_of_borrowers:						
1(excluded category)						
> 1	-1.106***	-1.121	-0.884	-0.998	-1.085	-1.110
	(0.223)	(0.223)	(0.124)	(0.006)	(0.006)	(0.142)
Log likelihood = -653.15584		Number of draws in MCMC = 100000				
Prob > chi2 = 0.0000		Burn-in = 200000				

In the frequentist approach, the variables loan to value ratio, debt to income ratio, and interest rate consistently have positive influences on default, while the number of borrowers has a significantly negative impact on default. These results are in accordance with economic theory and intuition. For instance, a large loan size compared to house value and a low income compared to the loan size make the borrowers more likely to default since it may be difficult for people with low income or a large debt to repay. As people repaying the loan together share the burden and risk together, they are less likely to default. Therefore the number of borrowers has a significantly negative sign. The variable mortgage insurance percentage has a significantly positive impact on default pre and during the crisis but insignificant post-crisis. We consider the positive impact is due to moral hazard of the borrowers. In all samples, the loan purpose of refinance has a larger positive impact on default than the purpose of property purchase. We consider this is because refinances are in many cases due to financial distress of the borrowers when they change payment schedules.

As expected, for the non-informative Bayesian approach, the coefficients and standard deviations estimated are very similar to the coefficients and standard errors obtained from the frequentist approach in each sample because both methods rely entirely on data information. In the Bayesian estimation with informative priors, the use of informative priors adds to the estimation results the information we have prior to seeing the data. The estimation results are a combination of data and prior information, which in our case is the frequentist estimation results based on samples of earlier or later time periods, and forecasts of the coefficient values. Generally, the estimation results such as the posterior means are between the frequentist estimation results and the prior means. For some of the coefficients, for instance the coefficients for debt to income ratio, interest rate, mortgage insurance percentage, etc., the difference of estimation results between using informative priors and non-informative priors is not large. The quantities of posterior means move moderately towards prior means while the coefficient signs generally stay unchanged. On the other hand, some coefficients changed signs compared to the non-informative Bayesian estimation because the highly informative priors have different signs from the non-informative Bayesian estimates. For instance, in the estimation results using informative Bayesian models based on training sample 2, the coefficients for the variable loan to value are opposite to that of the Bayesian model using non-informative priors.

In the Bayesian framework, unlike in the frequentist approach, there is no concept of significance. As a rule of thumb, if the posterior mean is more than two standard deviations away from 0, the corresponding variable is considered to have an impact on the dependent variables. Using this criterion, the variables that influence loan default in the Bayesian approach with non-informative priors are the same as the variables that are significant in the frequentist approach.

3.4.3 Bayesian model averaging and model selection

Table 3.10 shows the estimation results such as posterior means, standard deviations in Bayesian model averaging (BMA) based on loans initiated in 2004, 2007, and 2011. “P!=0” shows the probability that each explanatory variable has an impact on default probability.

Table 3.10 Bayesian model averaging results for accounts initiated in 2004, 2007 and 2011

	2004		2007		2011	
	P!=0	Posterior mean (std.dev)	P!=0	Posterior mean (std.dev)	P!=0	Posterior mean (std.dev)
Intercept	100%	-7.260 (1.888)	100%	-22.856 (0.706)	100%	-17.251 (1.274)
original_loan_to_value	47.7%	0.011 (0.013)	5.8%	0.0006 (0.003)	12.2%	0.002 (0.007)
original_combined_loan_to_value	52.3%	0.011 (0.011)	100%	0.027 (0.003)	8.0%	0.001 (0.005)
original_debt_to_income_ratio	33.4%	0.004 (0.007)	100.0%	0.023 (0.002)	100.0%	0.061 (0.012)
original_interest_rate	100.0%	0.878 (0.128)	100.0%	1.181 (0.051)	100.0%	1.963 (0.227)
mortgage_insurance_percentage	100.0%	0.028 (0.006)	100.0%	0.016 (0.002)	0.0%	0.000 (-)
number_of_units	0.0%	0.000 (-)	0.0%	0.000 (-)	0.0%	0.000 (-)
log_original_upb	96.4%	-0.427 (0.134)	100.0%	0.607 (0.045)	0.0%	0.000 (-)
original_loan_term	-	-	100.0%	0.004 (0.001)	0.0%	0.000 (-)
loan_purpose:						
P(Excluded category)	-	-	-	-	-	-
C	100.0%	0.873 (0.133)	100.0	0.913 (0.059)	0.0%	0.000 (-)
N	100.0%	0.809 (0.126)	100.0	0.889 (0.061)	0.0%	0.000 (-)
number_of_borrowers:						
1(excluded category)	-	-	-	-	-	-
> 1	100.0%	-0.645 (0.108)	100.0	-0.694 0.047	100.0%	-1.126 0.220

Fig 3.8 – Fig 3.10 show the potential models in Bayesian model averaging based on each of the three samples, and the variables included in each model. The horizontal axis shows the number of each model. The models are numbered in a descending order based on their posterior model probabilities. That is, the model with the largest posterior model probability is model #1. The vertical axis shows which variables are included in each model. A white square shows the variable is excluded in a model, while a red or blue square shows the variable is included in the model. Table 3.11 – Table 3.13 presents the posterior model probability and information criterion for each model.

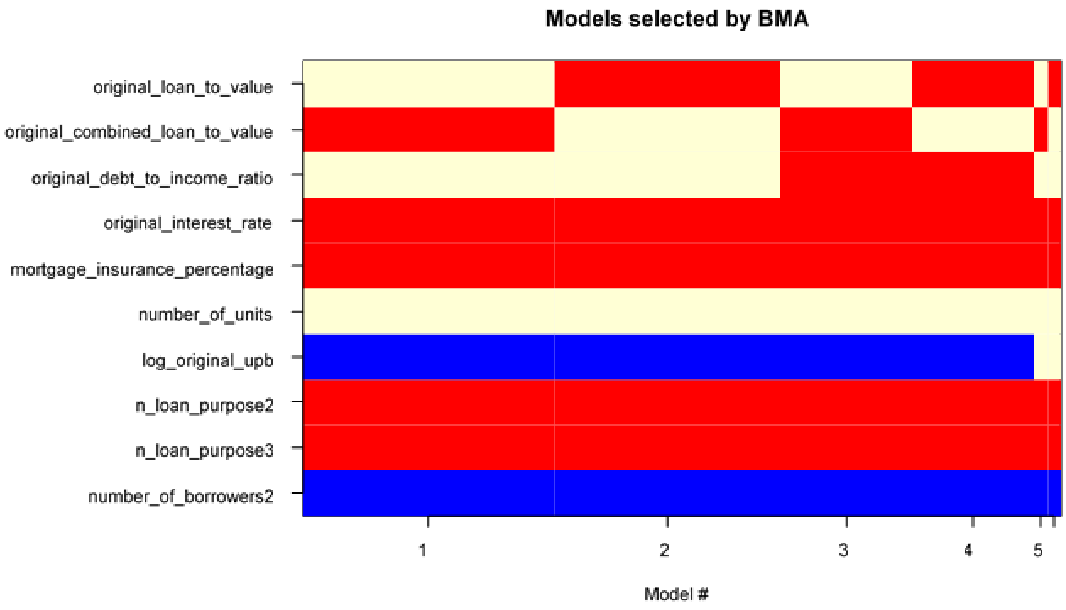

Fig. 3.8 models selected based on a sample of 2004

Table 3.11 Number of variables, information criterion and posterior model probability for each model selected on sample 2004

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Number of variables	7	7	8	8	6	6
BIC			-5.183e+05			
Posterior model probability	0.332	0.299	0.173	0.161	0.019	0.016

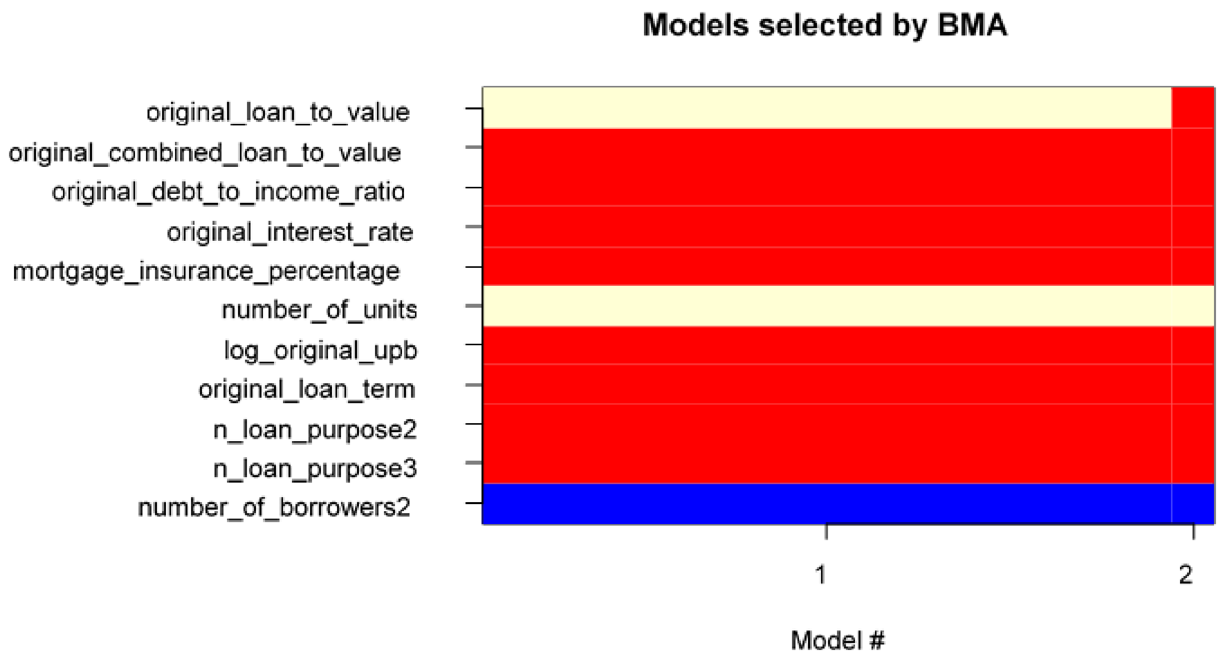

Fig. 3.9 models selected based on a sample of 2007

Table 3.12 Number of variables, information criterion and posterior model probability for each model selected on sample 2007

	Model 1	Model 2
Number of variables	9	10
BIC	-5.085e+05	
Posterior model probablity	0.942	0.058

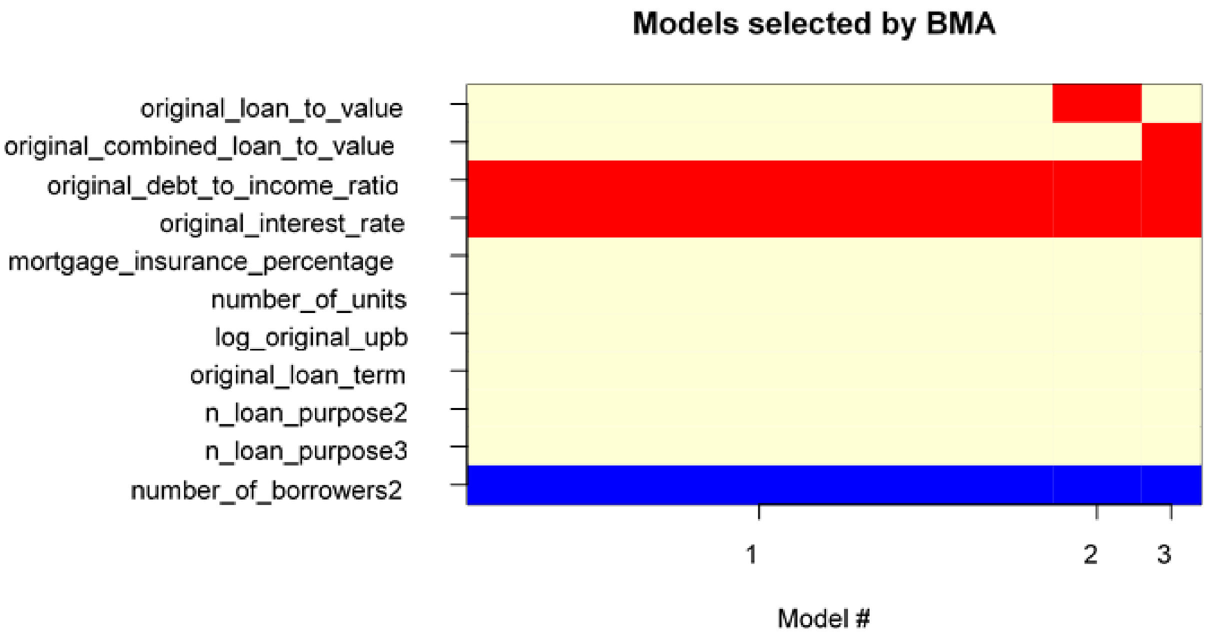

Fig. 3.10 models selected based on a sample of 2011

Table 3.13 Number of variables, information criterion and posterior model probability for each model selected on sample 2011

	Model 1	Model 2	Model 3
Number of variables	3	4	4
BIC	-5.336e+05		
Posterior model probability	0.798	0.122	0.080

The Bayesian model averaging method provides a probability that a variable could have an impact. For instance, as Fig 3.8 and Table 3.11 show, based on the 2004 sample, 6 models are selected in the Bayesian model averaging process. Within them, model 3 and model 4 select the variable 'original debt to income ratio' while others do not. Model 3 and Model 4 have posterior model probability 0.173 and 0.161 respectively. Therefore, in model averaging in the 2004 sample the variable 'original debt to income ratio' has a probability of 33.4% to have an impact which is the sum of the posterior probabilities of the models that select this variable. In BMA, the posterior mean of the coefficient is the weighted average of the posterior means of that coefficient in all models and the weights are the posterior model probability of each model: $0.173E(\beta_3 \mid y, M_3) + 0.161E(\beta_3 \mid y, M_4)$, based on Eq. (3.21). Since the coefficient estimate of a variable in Bayesian model averaging is a weighted average of the coefficient estimates of that variable over all models, for variables with low model probability the coefficient estimates tend to be smaller than that using a single model. For instance, for the variable 'original debt to income ratio' in sample 2004, the point estimate in the frequentist single model is about 0.013 (see Table 3.7), whereas it is 0.004 in the Bayesian model averaging approach (See Table 3.10). In the former approach this variable is considered to have a more important impact on default while in the latter approach the variable is of less importance.

In pre-crisis, during crisis, and post-crisis samples, there are 6, 2, and 3 potential models, respectively. Each model has a posterior model probability less than 1. The Bayesian model selection method uses the model that has the highest posterior model probability. Variables that have low influence on default probability are often excluded from the model. For instance, variables such as 'debt to income ratio', 'interest rate', 'number of borrowers', 'loan purpose' are selected by all models estimated on all samples. On the other hand, in model 1 based on sample 2011, variable 'original loan to value' is not included in the model although it still has 12.2% probability of having an impact on default. In model 2 based on sample 2004, the same variable is included in the model although it has only 47.7% probability of having an impact on default. By choosing a single model either with or without a variable, the model risk of variable selection is induced. Including the variable in the models means ignoring the chance that it does not impact on default, whereas excluding the variable means ignoring the chance that it does. Therefore compared to Bayesian model averaging, Bayesian model selection method brings in a larger variable selection risk.

3.5 Post-estimation convergence diagnostics and model performances

3.5.1 Bayesian diagnostics:

Bayesian diagnostics show that every parameter estimate converges to their corresponding posterior mean after 100000 simulations, with 200000 burn-ins discarded. As an example, in Fig 3.12 we show the summary of the convergence diagnostics for one coefficient in Bayesian estimation based on sample train2 with static ARIMA forecasts of coefficients as prior means and forecast standard errors as prior standard deviations.

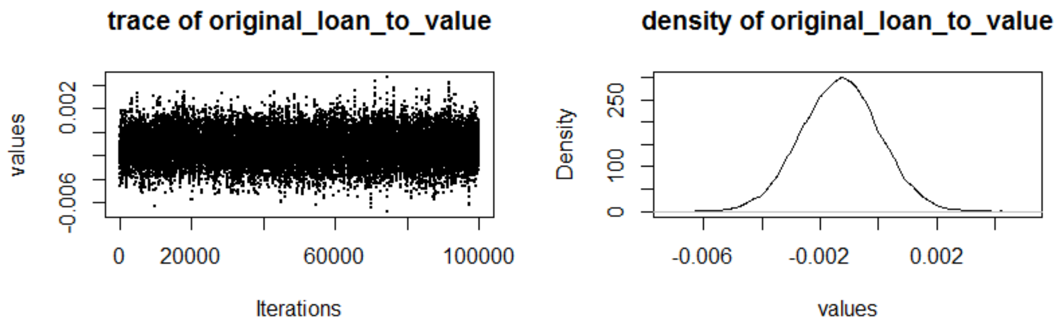

Fig. 3.11 Convergence diagnostics for the coefficient for 'orginal_loan_to_value' in the Bayesian regression based on sample train2 with static ARIMA forecasts as priors

The trace plot shows the parameter sampling against the number of iterations. The MCMC chain is mixing well and converges to the posterior: it rapidly traverses the marginal posterior domain. It visits areas with high posterior probability with high frequency and centres around the mean of the posterior distribution but also explores other areas proportionate with their corresponding posterior probability. The trace plots show no sign of trend or autocorrelation. The coefficient density plot shows the simulated marginal posterior distribution for the coefficient. The density is unimodal and symmetric which resembles the shape of a normal distribution that we choose as the prior. These demonstrate that the algorithm converges well. The diagnostic plots for other coefficients show similar characteristics.

Table 3.14 shows the Geweke diagnostic test of convergence for the Bayesian regression based on sample train2 with ARIMA forecasts of coefficients as priors. According to the z-test, $|z|<2$ for all parameters confirms that all variables in our models achieved good convergence.

Table 3.14 Geweke diagnostic for the Bayesian regression based on sample train2 with static ARIMA forecasts as priors

variables	z
Intercept	1.7660
original_loan_to_value	-1.0175
original_combined_loan_to_value	-0.8488
original_debt_to_income_ratio	-0.0558
original_interest_rate	1.4394
mortgage_insurance_percentage	-0.9417
number_of_units	-0.0567
log_original_upb	-1.2875
original_loan_term	-0.6874
as.factor(n_loan_purpose)2	0.1238
as.factor(n_loan_purpose)3	0.0579
as.factor(number_of_borrowers)2	-1.5850

3.5.2 Model performances

Table 3.15 - Table 3.17 show the model performances of each method based on the three training samples. The AUC (area under ROC) and AUCH (area under the Convex Hull of the ROC Curve) for the test samples range between 0.65 and 0.75, and the GINI coefficients fluctuate around 0.45. The H measure and K-S statistic range between 10%-20% and 25%-40%, respectively. The overall accuracy of our model differs between using different training samples and different performance measures.

Table 3.15 Performance measures for models trained on pre-crisis sample (2004) and tested on the test sample (2014)

Method	Sample	H-measure	GINI	AUC	AUCH	K-S
Frequentist	Test	0.1333932	0.4267079	0.7133539	0.7239204	0.348526
Bay non-inf	Test	0.1324768	0.4256598	0.7128299	0.723011	0.3463307
Bay inf naive forecast	Test	0.1341928	0.414865	0.7074325	0.7242675	0.355709
Bay inf ARIMA forecast (static)	Test	**0.1769625**	**0.4876301**	**0.743815**	**0.7558264**	**0.3965884**
Bay inf Bayesian updating	Test	0.1602184	0.4595142	0.7297571	0.740917	0.3782065
BMA	Test	0.1257415	0.4148326	0.7074163	0.7172335	0.3251104
BMS	Test	0.1209886	0.409416	0.704708	0.7153748	0.3176249

Table 3.16 Performance measures for models trained on crisis sample (2007) and tested on the test sample (2014)

Method	Sample	H-measure	GINI	AUC	AUCH	K-S
Frequentist	Test	0.1001499	0.3214536	0.6607268	0.674129	0.2627422
Bay non-inf	Test	0.1005414	0.321644	0.660822	0.6744331	0.2634848
Bay inf naive forecast	Test	0.1203579	0.3715313	0.6857656	0.699313	0.2633904
Bay inf ARIMA forecast (static)	Test	**0.1765579**	**0.4779653**	**0.7389827**	**0.7521907**	**0.3928992**
Bay inf Bayesian updating	Test	0.1554517	0.4487751	0.7243875	0.7380265	0.3516263
BMA	Test	0.1027469	0.3226562	0.6613281	0.6759636	0.2514019
BMS	Test	0.102888	0.3227015	0.6613507	0.6763563	0.2519265

Table 3.17 Performance measures for models trained on post-crisis sample (2011) and tested on test sample (2014)

Method	Sample	H-measure	GINI	AUC	AUCH	K-S
Frequentist	Test	0.1651203	0.4738612	0.7369306	0.7484631	0.3692304
Bay non-inf	Test	0.1650112	0.4700597	0.7350298	0.7474616	0.3653979
Bay inf naive forecast	Test	0.1781688	0.4770383	0.7385192	0.751733	0.3727387
Bay inf ARIMA forecast (static)	Test	0.1805296	0.4968523	0.7484262	0.7588825	0.390492
Bay inf ARIMA forecast (dynamic)	Test	**0.1949956**	**0.5007703**	**0.7503851**	**0.7630892**	**0.3914605**
Bay inf Bayesian updating	Test	0.1717078	0.4738927	0.7369463	0.7490461	0.3861991
BMA	Test	0.1588343	0.4603332	0.7301666	0.7407223	0.3414858
BMS	Test	0.1612778	0.4596614	0.7298307	0.7426359	0.3420132

In general, models trained on accounts that originated in 2011 generate the highest overall accuracy. Models trained on the crisis sample don't perform as well as those trained on non-crisis samples.

Bayesian models with informative priors consistently outperform the frequentist models and Bayesian models with non-informative priors in all samples regardless of sample periods and economic conditions. Among the informative Bayesian models, the Bayesian models with ARIMA forecasts as priors consistently perform the best compared to all other frequentist and Bayesian models used in this research regardless of sample periods and economic conditions on which the models are built, showing the advantage of the prior selection method we propose.

3.6 Conclusion

The frequentist approach is commonly used in credit scoring. However, it has the disadvantage that it relies entirely on the data. With the Bayesian approach, we can add information that we have about the model coefficients as Bayesian informative priors to the modelling process.

In this chapter, we model the probability of default of mortgage loans using the frequentist and different Bayesian approaches. The credit scoring model used in this chapter is the cross sectional logistic regression model. We train our model on samples of pre-crisis, during crisis, and post-crisis time periods. We use the maximum likelihood method and random walk chain Metropolis-Hastings algorithm for the frequentist and Bayesian estimations respectively.

We intend to reduce the risk of neglecting useful available information and enhance model performance by using multiple sets of informative priors. The informative prior selection methods that we use include a naive forecast method which uses the estimated coefficients obtained based on the last period as a forecast of the current period, and an ‘updating’ method which uses the coefficient estimates based on the latest time period as priors to update the model built earlier. We also use the Bayesian model averaging and selection methods. The main contribution of this chapter is that we propose an innovative informative prior selection method. With this method, we treat each coefficient in the credit scoring models built on consecutive periods as a time series variable. We fit the coefficient values estimated using ARIMA models. We then use the ARIMA forecasts of the coefficients to the time period of the test sample as informative priors in the Bayesian logistic model.

We show that the method of using the ARIMA forecasts of coefficients as informative priors gives the best model predictive accuracies among all the frequentist and Bayesian methods used in this research regardless of the economic conditions. Therefore, with the informative prior selection method using ARIMA forecasts of coefficients as priors that we propose, we reduced the risk of ignoring advantageous information in addition to that contained in the data, and thus improved the predictive performances of the credit scoring model.

Chapter 4

Reducing Estimation Risk Using a Bayesian Posterior Distribution Approach: Application to Stress Testing Mortgage Loan Default

4.1 Introduction

Stress testing is an important operational research tool to assess bank risk levels and to provide a basis to assist decision making by financial institutions and regulators (Breeden, 2016; Ju, Jeon, & Sohn, 2015; Schechtman & Gaglianone, 2012). Stress tests are designed to measure how sensitive risk exposures are to external or internal shocks to a financial system, an individual financial institution, a portfolio or an account (Misina et al., 2006). Based on stress testing results, regulators can assess if the financial system is stable enough to tolerate extreme but plausible economic conditions. Banks can decide how much capital they should keep to protect depositors in case such conditions occur. In practice, the financial sector assessment programmes (FSAPs) of the IMF and World Bank, as well as the financial authorities of various countries, regularly apply stress testing on financial institutions to assess the stability of the financial system (Schuermann, 2014; Sorge & Virolainen, 2006).

The focus of stress testing is largely on credit risk which is the most significant risk in banking systems (Sorge & Virolainen, 2006). The credit risk stress testing methods in the literature can be summarized as a three-step procedure (Borio et al., 2014; Kapinos, Martin, & Mitnik, 2018). First, empirical models exploring the relationship between an indicator of bank credit risk and macroeconomic variables are built and model coefficients estimated. Many papers (Kanas & Molyneux, 2018; Schechtman & Gaglianone, 2012; Vazquez, Tabak, & Souto, 2012) perform macro stress tests with aggregate data at the system level or with data of groups of financial institutions. Some (Bangia et al., 2002; Bellotti & Crook, 2013, 2014; Breeden, 2016; Ju et al., 2015; Pesaran et al., 2006) implement micro stress testing methods with granular data at an individual account, portfolio or institution level. Studies find that increases in output growth tend to reduce credit risk (Bikker & Hu, 2002; Laeven & Majnoni, 2003; Sorge & Virolainen, 2006), whereas rises in interest rate or unemployment tend to increase credit risk (Bellotti & Crook, 2013; Bikker & Hu, 2002; Pesaran et al., 2006). In the second step, stress scenarios for the macroeconomic variables are constructed on a historical (Bellotti & Crook, 2013, 2014; Sorge & Virolainen, 2006) or hypothetical (Jokivuolle & Viren, 2013; Tsukahara et al., 2016) basis using distribution simulation methods (Bellotti & Crook, 2013, 2014; Kanas & Molyneux, 2018) or point prediction methods (BoE, 2018; FRB, 2018; Breeden, 2016; Busch et al., 2018; EBA, 2018). The third step is to apply the stress scenarios of the macroeconomic variables to the empirical model to measure the extent of impact they have on the credit risk indicator of interest, such as the probability of default. Previous research has found that credit losses under stress scenarios with shocks from macroeconomic variables are higher than those under baseline scenarios (Bellotti & Crook, 2013, 2014; Bikker & Hu, 2002; Jokivuolle & Viren, 2013; Laeven & Majnoni, 2003; Sorge & Virolainen, 2006). For example, Bellotti and Crook (2013) found that in the worst 1% economic scenarios, the default rate is 1.73 times the median using account level credit card data. Rösch and Scheule (2004) found the 99% Value at Risk to be between 1.50 and 8.32 times the mean for real estate loans based on aggregate data.

Estimation risk and model risk have been recognised in the financial risk management literature. For instance, Escanciano and Olmo (2010) take into account estimation risk when backtesting market risk models so that how appropriate and conservative those models are can be better assessed. Gourieroux and Zakoïan (2013) argue that using an estimate based on a sample to approximate a true parameter is asymptotically biased and would cause VaR underestimation. Therefore they propose to substitute an adjusted estimate to the true parameter which would result in larger estimates of VaR. Some papers address estimation risk and model risk in credit risk stress testing. Philippon, Pessarossi, and Camara (2017) provide the first assessment method for the stress tests in the European Union which can be seen as an effort to detect model risks in existing stress testing models. In general, they find no evidence of biases in scenario building and loss forecasting. Jacobs, Karagozoglu, and Sensenbrenner (2015) use a Bayesian approach to address estimation risk in stress testing in the sense that they use informative priors to include expert knowledge. However, the importance of estimation risk has not been fully addressed in the credit risk stress testing literature. Although parameter estimation risk is gaining increasing attention, none of the papers introduces this type of uncertainty as a source of stress scenario input into the stress testing procedure as they do the risk of macroeconomic shocks. The majority of the literature on credit risk stress testing uses the frequentist estimation approach in the first step of the three-step procedure to obtain model parameter estimates. Such fixed scalars are used in the simulation of the loss distribution. However, there are estimation errors inherent in coefficient estimates. In stress testing practice in the literature, this estimation risk is ignored since only the mean estimates are substituted into the DR simulations in stress testing. For papers that do consider estimation risk in their modelling (Escanciano & Olmo, 2010; Gourieroux & Zakoïan, 2013), this is still the case. Some papers use a Bayesian method for stress testing (Jacobs et al., 2015; Louzis, 2017; Petropoulos et al., 2018). However, their stress test approaches still employ Bayesian coefficient point estimates without addressing all of the range of possible coefficient estimates.

The contribution this chapter makes to the literature is to propose that by using a Bayesian approach and a Bayesian coefficient posterior distribution in stress testing, we take into account parameter uncertainty and reduce risk underestimation. That is, a more prudential amount of capital that a bank would need in order to maintain a given risk level to protect depositors is estimated than with conventional methods. In Bayesian econometric theory, both parameters and explanatory variables are random variables instead of scalars and variables respectively, as in the frequentist approach. When doing Bayesian stress testing and simulating the estimated default rate distribution, we not only take random draws from the historical scenarios of the macroeconomic variables but also simulate from the coefficient posterior to take into account other possible coefficient estimates in the coefficient posterior distribution.

Therefore by employing Bayesian simulation of coefficients in stress testing, we model the uncertainty of coefficients thus reducing credit risk underestimation that arises from neglecting estimation risk. Moreover, since the number of draws taken from different regions of the posterior distribution is proportionate to the posterior probability of these regions, when we include the less likely coefficient estimates from the posterior distribution, we also take their corresponding low probability into account. Therefore this stress testing method also has the benefit that it avoids risk overestimation.

This chapter estimates and stress tests the probability of default at the micro/account level using a dataset of U.S. mortgage loans. The distribution approach is used to form the simulated default rate (DR) distribution and to obtain Value at Risk (VaR) at different percentiles. A discrete time hazard model is employed to analyse the relationship between default behaviour and macroeconomic as well as account level covariates, and to make forecasts for the probability of default. We use both the frequentist and Bayesian methods in estimation and stress testing. A Bayesian approach is used in order to simulate the posterior distribution of the model coefficients. We employ non-informative priors so that the differences between the simulated DR distributions using the frequentist and Bayesian approaches in the stress testing stage do not come from subjectivity introduced in the estimation stage. The coefficient posterior draws obtained in the Bayesian approach are subsequently applied to simulate the Bayesian estimated DR distribution. The Bayesian simulated DR distribution is then compared with the simulated DR distribution obtained using a frequentist approach with coefficient point estimates. In detail our method involves: 1) modelling the probability of default of mortgage loans and estimating the relationship between the probability of default and macro and micro predictors using both frequentist and Bayesian methods; 2) stress testing the impact of rare but plausible macro events as well as coefficient uncertainty on default rate by using historical scenarios of macroeconomic variables, the Bayesian coefficient posterior distribution, and simulated error terms, to simulate the Bayesian estimated DR distribution; 3) comparing the Bayesian DR distribution using a posterior distribution with the frequentist DR distribution using coefficient point estimates and computing the VaRs accordingly.

We find that the estimation and forecast results are similar using a frequentist method and a Bayesian method with non-informative priors. But in stress testing, the 95% and 99% VaRs of the simulated DR distribution obtained using a Bayesian approach with a coefficient posterior distribution are 3.7 and 6.5 times as large as the VaRs at the same probability levels respectively using the frequentist approach with coefficient mean estimates. The expected monetary values of losses estimated based on the 99% VaRs show that neglecting coefficient uncertainty in stress testing may considerably underestimate credit losses and the capital needed by a bank.

The structure of this chapter is as follows. Section 2 outlines the discrete time hazard model employed to estimate the probability of default and defines the estimation risk and a Bayesian approach that can be used to address it. Section 3 describes the stress testing models and procedures using the frequentist and Bayesian approaches. Section 4 describes the data and variables used in this research. Section 5 presents the estimation, prediction, performance, and stress testing results. Section 6 provides a summary of the main findings and concludes.

4.2 Methodology

4.2.1 Discrete time hazard model, default rate, and expected loss

Discrete time hazard model

To estimate the probability of default, we use a discrete time hazard model, which is a logistic regression model using panel data (Bellotti & Crook, 2013):

$$\text{logit}(p_{i,t}) = \alpha + \boldsymbol{z}_{t-3}'\boldsymbol{\beta}_1 + \boldsymbol{w}_i'\boldsymbol{\beta}_2 + \boldsymbol{u}_{i,t-3}'\boldsymbol{\beta}_3 + \boldsymbol{g}(\boldsymbol{t})'\boldsymbol{\beta}_4 \quad (4.1)$$

$p_{i,t}$ denotes the probability of default for account i at the duration time t; $\boldsymbol{z}_{t-3}$ denotes a vector of macroeconomic variables lagged three months; $\boldsymbol{w}_i$ denotes a vector of application variables for account i; $\boldsymbol{u}_{i,t-3}$ denotes a vector of behavioural variables lagged three months for account i; $\boldsymbol{g}(\boldsymbol{t})$ denotes functions of loan duration time.

We arrange the data such that the observations after the first default for any account are set to missing values. This ensures that the econometric model is parameterised using data up until the first default, which is what is required for the single event hazard distribution (See Singer & Willett, 1993). Functions of duration are included as explanatory variables. In this way, our model is a discrete time survival model with the event of interest being loan default (Bellotti & Crook, 2013). Using an appropriate estimation method, and with the accounts' defaults and covariates information in the logistic model, we can predict the probability of default for each account at duration time t , and how much impact each explanatory variable has on the logit. Maximum likelihood estimation is used in the frequentist approach. The random walk chain Metropolis-Hastings algorithm is used in the Bayesian approach. Non-informative priors are employed in the Bayesian method. For details, see Koop et al. (2007).

Survival probability

The survival probability to a duration time period can be computed from the probability of default at each time period. The predicted survival probability at duration time $t = q$ for account i is the product of the probability of account i not defaulting in each of the time periods until q:

$$\hat{S}_{i,q} = \prod_{t=1}^{q}(1 - p_{i,t}) \quad (4.2)$$

The cumulative probability of default is the complement of the survival function during each time period. It provides the probability of default at any time within the duration of q time periods:

$$\hat{H}_{i,q} = 1 - \hat{S}_{i,q} \tag{4.3}$$

The predicted and observed default rate at the aggregate level

Suppose t_c denotes calendar time. t_{a_i} denotes an account's opening time. $d_{i,t_c - t_{a_i}}$ denotes default of account i at calendar time t_c. $p_{i,t_c - t_{a_i}}$ denotes the probability of default for account i at calendar time t_c. n denotes the number of accounts. Then the default rate at calendar time t_c is computed as the ratio of the number of defaults at calendar time t_c divided by the total number of accounts at risk at that time:

$$R_{t_c} = \frac{1}{n} \sum_{i=1}^{n} d_{i,t_c - t_{a_i}} \tag{4.4}$$

The predicted probability of default at calendar time t_c at the aggregate level is:

$$P_{t_c} = \frac{1}{n} \sum_{i=1}^{n} p_{i,t_c - t_{a_i}} \tag{4.5}$$

The expected loss (EL) of an account is calculated as the product of the probability of default (PD), exposure at default (EAD) and loss given default (LGD):

$$EL = PD \times EAD \times LGD \tag{4.6}$$

4.2.2 Coefficient uncertainty and Bayesian stress testing

Frequentist methodology

In the frequentist approach, data is repeatable and random while parameters are fixed. The frequentist approach assumes the parameters from a population form a vector of scalars, $\boldsymbol{\theta}$. The estimator $\hat{\boldsymbol{\theta}}$ is a function of the data, which is a repeated sample from a population. As the sample size increases, an unbiased estimator converges to the true parameter $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$. Therefore the coefficient estimate, which is the expectation of the unbiased estimator, tends to the true parameter $E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$. In reality, the sample size is finite; and a difference between the estimate and the true parameter exists. Estimation risk arises but is ignored in stress testing because coefficient estimation standard errors are ignored.

Bayesian methodology

In the Bayesian approach, the data, which is the observed sample, is fixed, while parameters are random. The Bayesian approach treats the parameters as random variables that have their own distributions since in the Bayesian approach anything uncertain can be expressed using probability (Koop et al., 2007). Suppose $\boldsymbol{y}$ and $\boldsymbol{\theta}$ are the data and a vector of parameters respectively. Based on Bayes' rule, the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is proportional to the product of the prior distribution $p(\boldsymbol{\theta})$ and the likelihood distribution $p(\boldsymbol{y} \mid \boldsymbol{\theta})$:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$$

To obtain coefficient estimates in the Bayesian approach, instead of random sampling from data as in the frequentist approach, the Bayesian method involves random sampling from the parameter posterior distribution. Each region within the posterior distribution has a probability. Therefore unlike the frequentist approach which assumes there is a true parameter with certainty, in the Bayesian approach no coefficient estimate is the right or wrong one. The more likely parameter regions have higher probabilities while the unlikely ones have lower probabilities.

Bayesian stress testing

In our Bayesian stress testing application that addresses the estimation risk of coefficient uncertainty, we use the Bayesian parameter variables $\boldsymbol{\theta}$ as opposed to the frequentist point estimates $E(\hat{\boldsymbol{\theta}})$. We randomly sample from the posterior distribution of the parameters, and apply these random draws to stress testing to avoid the estimation risk that arises from omitting the differences between the draws, as well as the different probabilities of different regions in the posterior distribution. In other words, with our Bayesian approach in the stress testing application, the simulated value of a dependent variable $\boldsymbol{y}_j$ for each observation j using the kth draw is:

$y_{j,k} = \boldsymbol{x}_j '\boldsymbol{\theta}_k + \varepsilon_{j,k}$, in which $k = 1, 2, ..., K$. (4.7)

In contrast, with the frequentist stress testing method using coefficient mean estimates it is:

$$y_j = \boldsymbol{x}_j ' E(\hat{\boldsymbol{\theta}}) + \varepsilon_j \tag{4.8}$$

To compare with frequentist stress testing using coefficient mean estimates, we also carry out Bayesian stress testing using Bayesian coefficient posterior mean estimates:

$$y_j = \boldsymbol{x}_j ' E(\boldsymbol{\theta} \mid \boldsymbol{y}) + \varepsilon_j \tag{4.9}$$

4.3 Stress testing

4.3.1 Stress testing model

Our stress testing method is a combination of posterior simulation of coefficients, simple random sampling of historical macroeconomic scenarios, and Monte Carlo simulation of the error terms. Consider the latent variable interpretation of the logistic regression over the time period that stress testing is applied to:

$$d^*_{j,i,t_s} = I(y^*_{j,i,t_s} = \boldsymbol{x}'_{j,i,t_s}\boldsymbol{\beta} + \varepsilon_{j,i,t_s} > 0) \quad (4.10)$$

in which

i denotes the ith account. $i = 1, 2, ..., n$.

j denotes the jth macroeconomic scenario. $j = 1, 2, ..., l$.

$\boldsymbol{x}_{j,i,t_s}$ denotes a vector of covariates including macroeconomic covariates that take their historical values, account application variables $\boldsymbol{w}_i$, account behavioural variables $\boldsymbol{u}_{i,t_s - t_{a_i} - 3}$ and duration functions $\boldsymbol{g}(t_s - t_{a_i})$, in which t_s is the calendar time stress testing is applied to, and t_{a_i} is the calendar time of the opening of account i.

y^*_{j,i,t_s} denotes the simulated value of the latent variable in the logistic regression for account i in scenario j at calendar time t_s.

d^*_{j,i,t_s} denotes the simulated default behavior for account i in scenario j at calendar time t_s. d^*_{j,i,t_s} takes the value 1 when an event occurs and 0 when it does not occur:

$$d^*_{j,i,t_s} = \begin{cases} 1 & y^*_{j,i,t_s} > 0 \\ 0 & else \end{cases}.$$

Application variables $\boldsymbol{w}_i$, account behavioural variables $\boldsymbol{u}_{i,t_s-t_{a_i}-3}$ and duration functions $\boldsymbol{g}(t_s - t_{a_i})$ can all be considered account specific variables. Suppose we represent the vector $\boldsymbol{x}_{j,i,t_s}$ as follows. $\boldsymbol{x}_{j,i,t_s} = \begin{pmatrix} \boldsymbol{m}_j \\ \boldsymbol{a}_{i,t_s} \end{pmatrix}$, in which $\boldsymbol{m}_j$ denotes the values of macroeconomic variables for scenario j. $\boldsymbol{m}_j$ represents the observed values of the macroeconomic variables in a random month in history before t_s. $\boldsymbol{a}_{i,t_s}$ denotes the values of the account-specific variables for account i at the time period stress testing is applied to. $\boldsymbol{a}_{i,t_s}$ includes account application variables $\boldsymbol{w}_i$, account behavioural variables $\boldsymbol{u}_{i,t_s-t_{a_i}-3}$ and duration functions $\boldsymbol{g}(t_s - t_{a_i})$ at calendar time t_s. $\boldsymbol{\beta}^{(m)}$ denotes a $v_1 \times 1$ column vector of parameter mean estimates for the constant and the macroeconomic variables using frequentist estimation. In this paper, $v_1 = 10$. $\boldsymbol{\beta}^{(a)}$ denotes a $v_2 \times 1$ column vector of coefficient mean estimates for account specific variables using frequentist estimation. In this paper, $v_2 = 7$. $\boldsymbol{b}^{(m)}$ is a $v_1 \times 1$ column vector of Bayesian posterior mean estimates for the constant and the coefficients for the macroeconomic variables. $\boldsymbol{b}^{(a)}$ is a $v_2 \times 1$ column vector of Bayesian posterior mean estimates for the coefficients for the account-specific variables. k denotes the kth draw from the K number of random draws from the posterior distribution. $\boldsymbol{b}_k^{(m)}$ denotes a $v_1 \times 1$ column vector of the kth draw from the Bayesian posterior distribution for the constant and the coefficients for the macroeconomic variables. Then Eq. (4.10) is further written as follows.

(1) Stress testing using the frequentist and Bayesian coefficient mean estimates

Frequentist:

$$d^*_{j,i,t_s} = I(y^*_{j,i,t_s} = \boldsymbol{m}'_j \boldsymbol{\beta}^{(m)} + \boldsymbol{a}'_{i,t_s} \boldsymbol{\beta}^{(a)} + \varepsilon_{j,i,t_s} > 0) \tag{4.11}$$

Bayesian:

$$d^{*}_{j,i,t_s} = I(y^{*}_{j,i,t_s} = \boldsymbol{m}^{'}_{j}\boldsymbol{b}^{(m)} + \boldsymbol{a}^{'}_{i,t_s}\boldsymbol{b}^{(a)} + \varepsilon_{j,i,t_s} > 0) \quad (4.12)$$

(2) Bayesian stress testing using the Bayesian coefficient posterior distribution:

$$d^{*}_{j,k,i,t_s} = I(y^{*}_{j,k,i,t_s} = \boldsymbol{m}^{'}_{j}\boldsymbol{b}^{(m)}_{k} + \boldsymbol{a}^{'}_{i,t_s}\boldsymbol{b}^{(a)} + \varepsilon_{j,k,i,t_s} > 0), \quad k = 1, 2, ..., K \quad (4.13)$$

The simulated default rates using the mean estimates and posterior distribution approaches:

1) The simulated default rate at t_s in scenario j using the frequentist and Bayesian coefficient mean estimates is:

$$\hat{R}_{j,t_s} = \frac{1}{n}\sum_{i=1}^{n} d^{*}_{j,i,t_s} \quad (4.14)$$

There are l scenarios in total.

2) In the Bayesian posterior distribution approach, the simulated default rate at t_s in scenario j,k, which is in the jth macroeconomic scenario and using the kth Bayesian coefficient draw, is:

$$\hat{R}_{j,k,t_s} = \frac{1}{n}\sum_{i=1}^{n} d^{*}_{j,k,i,t_s} \quad (4.15)$$

There are $l * K$ scenarios in total.

4.3.2 Stress testing procedure

The stress testing procedure we propose is as follows:

1. Estimate the discrete hazard model Eq. (4.1) based on the training sample using the frequentist and Bayesian approaches and obtain the frequentist and Bayesian coefficient mean estimates as well as the Bayesian posterior draws.
2. Choose a time period in the test sample for stress testing to apply to, which in our case is the end of the test sample time period October 2017. Simulate the default/non-default events of all the accounts alive during the stress testing period in each scenario in both frequentist and Bayesian frameworks by sampling from historical macroeconomic scenarios, posterior distribution (in the Bayesian posterior distribution approach) and error terms. The simulated default event of account i in scenario j for the point estimate approach is based on Eq. (4.11) – Eq. (4.12). The simulated default event of account i in scenario j,k for the Bayesian posterior distribution approach is based on Eq. (4.13). Next, calculate the simulated default rate across the accounts in each scenario. The simulated default rate in scenario j using the coefficient mean estimate approach is calculated based on Eq. (4.14). The simulated default rate in scenario j,k using the coefficient distribution approach is calculated based on Eq. (4.15).
3. Build the frequentist and Bayesian distributions of simulated default rates and compute the VaRs at different probability levels.

The panel data of the macroeconomic variables is arranged as a matrix, $\boldsymbol{M}_{v_1 \times T}$, of v_1

macroeconomic variables and T time periods[5]. In order to keep the dependence structure between the macroeconomic variables, we draw the historical values of all the macroeconomic variables simultaneously as opposed to sampling historical values of each variable individually. To give more details, we draw l simple random samples[6] with replacement of the columns[7] of $\boldsymbol{M}_{v_1 \times T}$. Each draw represents a macroeconomic scenario $\boldsymbol{m}_j$. All the l scenarios form a matrix of macroeconomic scenarios $\boldsymbol{M}_{v_1 \times l}$. The values of the v_2 number of account level variables for the n accounts alive at stress testing time t_s is $\boldsymbol{A}_{v_2 \times n}$.

In the Bayesian posterior distribution approach, we take K number of draws for the constant and macroeconomic coefficients from the posterior distribution, thus forming a matrix of v_1 number of parameters and each having K draws: $\boldsymbol{B}^{(m)}_{v_1 \times K}$. For coefficients for the v_2 number of account level variables, we use their posterior mean estimates: $\boldsymbol{b}^{(a)}_{v_2 \times 1}$.

Since there are three components, which are the *macroeconomic component* $\boldsymbol{m}_j^{'}\boldsymbol{\beta}^{(m)}$ (or $\boldsymbol{m}_j^{'}\boldsymbol{b}^{(m)}$, $\boldsymbol{m}_j^{'}\boldsymbol{b}_k^{(m)}$), the *account level component* $\boldsymbol{a}_{i,t_s}^{'}\boldsymbol{\beta}^{(a)}$ (or $\boldsymbol{a}_{i,t_s}^{'}\boldsymbol{b}^{(a)}$), and the *error term component* ε_{j,i,t_s} (or ε_{j,k,i,t_s}) in the latent variable y^*_{j,i,t_s} (or y^*_{j,k,i,t_s}) for each account in each scenario in Eqs. (4.11) – (4.13), to build the simulated default behaviours d^*_{j,i,t_s} and d^*_{j,k,i,t_s} in Eqs. (4.11) – (4.13), we first build the three components in y^*_{j,i,t_s} and y^*_{j,k,i,t_s} for all accounts in all scenarios.

---

[5] T equals the number of months in our macroeconomic data before t_s. The first row of $\boldsymbol{M}_{v_1 \times T}$ is a vector of $\mathbf{1}$.

[6] l can be larger or smaller or equal to T , and in our case is larger.

[7] Each column represents the values of the macroeconomic variables in a same month.

1) Stress testing procedure for the mean estimate approach:

We use the frequentist mean estimate approach to illustrate in this section. The stress testing process is very similar using the Bayesian mean estimate approach.

The *macroeconomic component* is: $\mathbf{M}_{l\times 1} = (\boldsymbol{M}_{v_1 \times l})' \boldsymbol{\beta}^{(m)}_{v_1 \times 1}$ with each scalar in $\mathbf{M}_{l\times 1}$ being m_j.

The *account level component* is: $\mathbf{A}_{n\times 1} = (\boldsymbol{A}_{v_2 \times n})' \boldsymbol{\beta}^{(a)}_{v_2 \times 1}$ with each scalar in $\mathbf{A}_{n\times 1}$ being a_i.

The *error term component* is: $n \times l$ draws from a standard logistic distribution: $\boldsymbol{\varepsilon}_{(ln)\times 1}$.

Repeat ***each scalar*** in the macroeconomic component $\mathbf{M}_{l\times 1}$ by n times, and we obtain:

$$\mathbf{M}_{ln\times 1} = \begin{pmatrix} \mathrm{m}_{1,1} \\ \vdots \\ \mathrm{m}_{1,n} \\ \vdots \\ \mathrm{m}_{j,1} \\ \vdots \\ \mathrm{m}_{j,n} \\ \vdots \\ \mathrm{m}_{l,1} \\ \vdots \\ \mathrm{m}_{l,n} \end{pmatrix}_{(ln)\times 1}$$

, in which $\mathrm{m}_{j,1} = \cdots = \mathrm{m}_{j,n} = \mathrm{m}_j$.

Repeat the account level component $\mathbf{A}_{n\times 1}$ ***as a whole*** l times, and we obtain:

$$\mathbf{A}_{ln\times 1}=\begin{pmatrix} a_{1,1} \\ \vdots \\ a_{n,1} \\ \vdots \\ a_{1,j} \\ \vdots \\ a_{n,j} \\ \vdots \\ a_{1,l} \\ \vdots \\ a_{n,l} \end{pmatrix}_{(ln)\times 1} \text{, in which } \begin{pmatrix} a_{1,1} \\ \vdots \\ a_{n,1} \end{pmatrix}_{n\times 1}=\ldots=\begin{pmatrix} a_{1,j} \\ \vdots \\ a_{n,j} \end{pmatrix}_{n\times 1}=\ldots=\begin{pmatrix} a_{1,l} \\ \vdots \\ a_{n,l} \end{pmatrix}_{n\times 1}=\mathbf{A}_{n\times 1}.$$

Then add $\mathbf{M}_{ln\times 1}$, $\mathbf{A}_{ln\times 1}$, $\boldsymbol{\varepsilon}_{(ln)\times 1}$ together. The intuition is that each one of the n accounts in the stress testing time period t_s faces l number of potential parallel scenarios, and each one of the l scenarios should have all the n accounts. Specifically, consider n accounts and l macroeconomic scenarios. The full results of the right-hand side of the latent logistic function for all the accounts in all the scenarios are:

$$\begin{matrix} (a) & & (b) & & (c) & & (d) \\ \begin{pmatrix} y^*_{1,1,t_s} \\ \vdots \\ y^*_{1,n,t_s} \\ \vdots \\ y^*_{j,1,t_s} \\ \vdots \\ y^*_{j,n,t_s} \\ \vdots \\ y^*_{l,1,t_s} \\ \vdots \\ y^*_{l,n,t_s} \end{pmatrix}_{(ln)\times 1} & = & \begin{pmatrix} \mathrm{m}_{1,1} \\ \vdots \\ \mathrm{m}_{1,n} \\ \vdots \\ \mathrm{m}_{j,1} \\ \vdots \\ \mathrm{m}_{j,n} \\ \vdots \\ \mathrm{m}_{l,1} \\ \vdots \\ \mathrm{m}_{l,n} \end{pmatrix}_{(ln)\times 1} & + & \begin{pmatrix} \mathrm{a}_{1,1} \\ \vdots \\ \mathrm{a}_{n,1} \\ \vdots \\ \mathrm{a}_{1,j} \\ \vdots \\ \mathrm{a}_{n,j} \\ \vdots \\ \mathrm{a}_{1,l} \\ \vdots \\ \mathrm{a}_{n,l} \end{pmatrix}_{(ln)\times 1} & + & \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \\ \vdots \\ \varepsilon_{(j-1)n+1} \\ \vdots \\ \varepsilon_{jn} \\ \vdots \\ \varepsilon_{(l-1)n+1} \\ \vdots \\ \varepsilon_{ln} \end{pmatrix}_{(ln)\times 1} \end{matrix} \qquad (4.16)$$

Each scalar in vector (a) in Eq. (4.16) represents the simulated value of the latent variable for an account in a scenario. For instance, y^*_{j,n,t_s} represents the simulated latent variable for account n in scenario j at stress testing time t_s. Divide vector (a) equally into l sections, one for each scenario, with each section having n scalars. Rearrange vector (a) into a new matrix with the number of rows being the number of accounts n, and the number of columns being the number of scenarios l. That is, put the first n scalars from vector (a) into the first column, the next n scalars into the second column, the j th n scalars into the j th column, so on until the last n scalars from vector (a) are put into the last column of the new matrix. In this way, each column of the new matrix has the simulated values of the latent variable for all the accounts in the same scenario, and all columns represent all the l scenarios:

$$\begin{pmatrix} y^*_{1,1,t_s} & \cdots & y^*_{j,1,t_s} & \cdots & y^*_{l,1,t_s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y^*_{1,n,t_s} & \cdots & y^*_{j,n,t_s} & \cdots & y^*_{l,n,t_s} \end{pmatrix} \qquad (4.17)$$

Based on Eq. (4.11), compare the simulated scalars in Eq. (4.17) with 0 and decide whether each of the n accounts is predicted to default or not to default in each scenario:

$$\begin{pmatrix} d^{*}_{1,1,t_s} & \cdots & d^{*}_{j,1,t_s} & \cdots & d^{*}_{l,1,t_s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d^{*}_{1,n,t_s} & \cdots & d^{*}_{j,n,t_s} & \cdots & d^{*}_{l,n,t_s} \end{pmatrix} \tag{4.18}$$

Subsequently, based on Eq. (4.14), the simulated default rate in each of the l scenarios can be computed by averaging over the simulated default behaviours of all the n accounts in each column in Eq. (4.18), and obtain the simulated default rates in all l scenarios: $\begin{pmatrix} \hat{R}_{1,t_s} & \cdots & \hat{R}_{j,t_s} & \cdots & \hat{R}_{l,t_s} \end{pmatrix}$.

2) Stress testing procedure for the Bayesian posterior distribution approach:

The *macroeconomic component* is: $\mathbf{M}_{l\times K} = (\boldsymbol{M}_{v_1\times l})' \boldsymbol{B}^{(m)}_{v_1\times K}$ with each scalar in $\mathbf{M}_{l\times K}$ being $\mathrm{m}_{j,k}$.

Convert the *macroeconomic component matrix* $\mathbf{M}_{l\times K}$ into a *macroeconomic component vector* $\mathbf{M}_{lK\times 1}$ with each scalar being $\mathrm{m}_{j,k}$: $\mathbf{M}_{lK\times 1} = \begin{pmatrix} \mathrm{m}_{1,1} \\ \vdots \\ \mathrm{m}_{j,k} \\ \vdots \\ \mathrm{m}_{l,K} \end{pmatrix}_{(lK)\times 1}$. In Bayesian stress testing, the coefficient draws also contribute to scenario building. Therefore instead of l scenarios, we have $l \times K$ scenarios in total using the Bayesian posterior distribution method.

The *account level component* is: $\mathbf{A}_{n\times 1} = (\boldsymbol{A}_{v_2\times n})' \boldsymbol{b}^{(a)}_{v_2\times 1}$ with each scalar in $\mathbf{A}_{n\times 1}$ being a_i.

The *error term component* is: take $l \times K \times n$ draws from a standard logistic distribution: $\boldsymbol{\varepsilon}_{(lKn)\times 1}$.

Repeat ***each scalar*** of the macroeconomic component vector $\mathbf{M}_{lK\times 1}$ by n times, and we obtain: $\mathbf{M}_{lKn\times 1} = \begin{pmatrix} \mathrm{m}_{1,1,1} \\ \vdots \\ \mathrm{m}_{1,1,n} \\ \vdots \\ \mathrm{m}_{j,k,1} \\ \vdots \\ \mathrm{m}_{j,k,n} \\ \vdots \\ \mathrm{m}_{l,K,1} \\ \vdots \\ \mathrm{m}_{l,K,n} \end{pmatrix}_{(lKn)\times 1}$, in which $\mathrm{m}_{j,k,1} = \cdots = \mathrm{m}_{j,k,n} = \mathrm{m}_{j,k}$.

Repeat the account level component $\mathbf{A}_{n\times 1}$ ***as a whole*** lK times, and we obtain: $\mathbf{A}_{lKn\times 1} = \begin{pmatrix} \mathrm{a}_{1,1,1} \\ \vdots \\ \mathrm{a}_{n,1,1} \\ \vdots \\ \mathrm{a}_{1,j,k} \\ \vdots \\ \mathrm{a}_{n,j,k} \\ \vdots \\ \mathrm{a}_{1,l,K} \\ \vdots \\ \mathrm{a}_{n,l,K} \end{pmatrix}_{(lKn)\times 1}$, in which $\begin{pmatrix} \mathrm{a}_{1,1,1} \\ \vdots \\ \mathrm{a}_{n,1,1} \end{pmatrix}_{n\times 1} = \ldots = \begin{pmatrix} \mathrm{a}_{1,j,k} \\ \vdots \\ \mathrm{a}_{n,j,k} \end{pmatrix}_{n\times 1} = \ldots = \begin{pmatrix} \mathrm{a}_{1,l,K} \\ \vdots \\ \mathrm{a}_{n,l,K} \end{pmatrix}_{n\times 1} = \mathbf{A}_{n\times 1}$.

Then add $\mathbf{M}_{lKn\times 1}$, $\mathbf{A}_{lKn\times 1}$, and $\boldsymbol{\varepsilon}_{(lKn)\times 1}$ together. The intuition is that all the n accounts in the stress testing time period t_s face lK number of potential parallel scenarios, and each one of the lK scenarios should have all the n accounts. Specifically, consider n

accounts, l macroeconomic scenarios, and K draws from the posterior distribution. The full results of the right-hand side of the latent logistic function for all the accounts in all scenarios are:

$$\overset{(e)}{\begin{pmatrix} y^*_{1,1,1,t_s} \\ \vdots \\ y^*_{1,1,n,t_s} \\ \vdots \\ y^*_{j,k,1,t_s} \\ \vdots \\ y^*_{j,k,n,t_s} \\ \vdots \\ y^*_{l,K,1,t_s} \\ \vdots \\ y^*_{l,K,n,t_s} \end{pmatrix}_{(lKn)\times 1}} = \overset{(f)}{\begin{pmatrix} \mathrm{m}_{1,1,1} \\ \vdots \\ \mathrm{m}_{1,1,n} \\ \vdots \\ \mathrm{m}_{j,k,1} \\ \vdots \\ \mathrm{m}_{j,k,n} \\ \vdots \\ \mathrm{m}_{l,K,1} \\ \vdots \\ \mathrm{m}_{l,K,n} \end{pmatrix}_{(lKn)\times 1}} + \overset{(g)}{\begin{pmatrix} \mathrm{a}_{1,1,1} \\ \vdots \\ \mathrm{a}_{n,1,1} \\ \vdots \\ \mathrm{a}_{1,j,k} \\ \vdots \\ \mathrm{a}_{n,j,k} \\ \vdots \\ \mathrm{a}_{1,l,K} \\ \vdots \\ \mathrm{a}_{n,l,K} \end{pmatrix}_{(lKn)\times 1}} + \overset{(h)}{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \\ \vdots \\ \varepsilon_{(jk-1)n+1} \\ \vdots \\ \varepsilon_{jkn} \\ \vdots \\ \varepsilon_{(lK-1)n+1} \\ \vdots \\ \varepsilon_{lKn} \end{pmatrix}_{(lKn)\times 1}} \tag{4.19}$$

Each scalar in vector (e) in Eq. (4.19) represents the simulated value of the latent variable for an account in a scenario. For instance, y^*_{j,k,n,t_s} represents the simulated latent variable for account n in macroeconomic scenario j using the k th draw from the Bayesian posterior at stress testing time t_s. Divide vector (e) equally into lK sections, one for each scenario, with each section having n scalars. Rearrange vector (e) into a new matrix with the number of rows being the number of accounts n, and the number of columns being the number of scenarios lK. That is, put the first n scalars from vector (e) into the first column of the new matrix, the next n scalars into the second column, the jk th n scalars into the jk th column, so on until the last n scalars from vector (e) are put into the last column of the new matrix. In this way, each column of the new matrix has the simulated values of the latent variable for all the accounts in the same scenario, and all columns represent all the lK scenarios:

$$\begin{pmatrix} y^*_{1,1,1,t_s} & \cdots & y^*_{j,k,1,t_s} & \cdots & y^*_{l,K,1,t_s} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y^*_{1,1,n,t_s} & \cdots & y^*_{j,k,n,t_s} & \cdots & y^*_{l,K,n,t_s} \end{pmatrix} \tag{4.20}$$

Based on Eq. (4.13), compare the simulated scalars in Eq. (4.20) with 0 and decide whether each of the n accounts is predicted to default or not to default in each scenario:

$$\begin{pmatrix} d^*_{1,1,1,t_s} & \cdots & d^*_{j,k,1,t_s} & \cdots & d^*_{l,K,1,t_s} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d^*_{1,1,n,t_s} & \cdots & d^*_{j,k,n,t_s} & \cdots & d^*_{l,K,n,t_s} \end{pmatrix} \tag{4.21}$$

Subsequently, based on Eq. (4.15), the simulated default rate in each of the lK scenarios can be computed by averaging over the simulated default behaviours of all the n accounts in each column in Eq. (4.21), and obtain the simulated default rates in all lK scenarios: $\left(\hat{R}_{1,1,t_s} \quad \cdots \quad \hat{R}_{j,k,t_s} \quad \cdots \quad \hat{R}_{l,K,t_s}\right)$.

Once we have the simulated default rates $\hat{R}_{j,t_s}$ and $\hat{R}_{j,k,t_s}$ in all the l and lK scenarios using the frequentist and Bayesian approaches, we use these simulated default rates to form the empirical simulated default rate distributions and obtain the VaRs.

4.4 Data and Variables

4.4.1 Data

The data we use to illustrate our methods are from the Freddie Mac single-family loan level dataset[8]. The loans are fully amortizing long term mortgages. We use the mortgage accounts that originated during the 12 months in 2014 as a training sample. We use accounts originated during the 12 months in 2015 as a test sample. For the training sample, we take December 2016 as the observation date. For the test sample, we take October 2017 as the observation date. For each year we use a random sample of 50000 loans. We consider an account is in default if it has in its payment history record no less than 60 days delinquency. Table 1 shows the number of accounts and defaults in each sample cross-sectionally.

Table 4.1 Training and test samples

	Train (2014)	Test (2015)
Number of accounts	50000	50000
Defaults of accounts	415	295

4.4.2 Variables

For our discrete time hazard model, the event of interest is default with the event indicator being 1 (default) and the non-event being 0 (non-default).

Table 2 gives a full list of the explanatory variables for this research. We include macroeconomic variables and application as well as behavioural variables of the accounts. To avoid trends, the macroeconomic variables are first differenced. To enable prediction and to avoid endogeneity, macroeconomic variables and behavioural variables are lagged 3 months.

[8] Dataset url: http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.page

Table 4.2 Full list of explanatory variables

Group	Variable name	Definition
Macroeconomic	d_l_tbill_3m	Three months treasury bill interest rate
	d_l_unemployment_rate	Unemployment rate
	d_l_CPI	Consumer price index
	d_l_consumer_confidence	Consumer confidence
	d_l_retail_sales	Log of retail sales
	d_l_personal_earnings	Log of personal earnings
	d_l_IPI	Industrial production index
	d_l_dowjones_index	Dow Jones stock price index
	d_l_CS_houseprice_index	House price index
Application	original_debt_to_income_ratio	The sum of monthly debt/sum of monthly income calculated at loan origination
	original_loan_to_value	Original loan amount/appraised loan value or purchase price
Behavioural	l_current_actual_upb	Log of the current unpaid balance of the mortgage
	l_current_interest_rate	Current interest rate
	l_remaining_months	The remaining months from the loan term in the mortgage note
Duration	loan_age	The duration of the loan since its origination
	loan_age_sq	The squared term of loan age

Source: Freddie Mac database for account specific variables, Datastream for macroeconomic variables

4.5 Results

4.5.1 Estimation results for the discrete time hazard model

We estimate models on the training sample using a frequentist approach and a Bayesian approach. Table 3 illustrates the estimation results.

Table 4.3 Estimation Results using the frequentist approach and the Bayesian approach with non-informative priors

Variables (Macro and behavioural variables are lagged 3 months)	Training sample: accounts originated during 2014 with December 2016 as the observation date	
	Frequentist estimate (std.error)	Bayesian posterior mean (std.dev)
Intercept	-14.7275682 ***	-14.758733
	(1.2980562)	(1.2752167)
d_l_tbill_3m	0.2126658	0.195955
	(0.9666853)	(0.9538694)
d_l_unemployment_rate	-0.2944642	-0.319338
	(0.4911965)	(0.5026099)
d_l_CPI	-0.0446704	-0.047746
	(0.1540015)	(0.1548964)
d_l_consumer_confidence	0.0079945	0.008596
	(0.0144115)	(0.0145325)
d_l_retail_sales	23.0846659 *	23.398211
	(13.8015895)	(14.0626727)
d_l_personal_earnings	24.9354680	24.469273
	(38.1177126)	(37.9106425)
d_l_IPI	0.1759931	0.182078
	(0.1345957)	(0.1334646)
d_l_dowjones_index	1.7978390	1.888173
	(2.0682810)	(2.0835078)
d_l_CS_houseprice_index	0.6831190 ***	0.679918
	(0.2085248)	(0.2080482)
original_debt_to_income_ratio	0.0346552 ***	0.035027
	(0.0061243)	(0.0060389)
original_loan_to_value	0.0044147	0.004576
	(0.0034320)	(0.0034240)
l_current_actual_upb	-0.1502279 **	-0.151780
	(0.0876762)	(0.0859478)
l_current_interest_rate	1.5111287 ***	1.505791
	(0.1570621)	(0.1572061)
l_remaining_months	-0.0031977 ***	-0.003141
	(0.0011235)	(0.0011293)
loan_age	0.0837527 **	0.085288
	(0.0346547)	(0.0347888)
loan_age_sq	-0.0017150 **	-0.001753
	(0.0009199)	(0.0009265)
	Log likelihood = -3579.424 Prob > chi2 = 0.0000	Number of draws in MCMC = 100000 Burn-in = 200000

The estimation results of the frequentist and Bayesian noninformative methods are very similar since both are based on information contained in the data entirely. In the frequentist approach, the effects of accounts' individual interest rate and debt to income ratio are significantly positive which is consistent with the expectation that the higher the interest rate and the amount of debt compared to borrowers' income the more likely a borrower is to default. Among the macroeconomic covariates, the house price index and the retail sales have a significantly positive impact on default probability also as expected. Loan duration and its squared term have positive and negative signs respectively showing that the default probability is nonlinear over time. The ratios of posterior means to posterior standard deviations show the variables that have an important impact on default rates in the Bayesian approach are similar to those in the frequentist approach. The coefficient signs of these variables in the Bayesian framework are the same as those in the frequentist framework. Table 4.4 presents the Geweke diagnostic results for the Bayesian estimations. The z-scores for the coefficients are within the [-2, 2] range. Fig. 4.1 gives the trace and density plots of the marginal posterior distribution of one coefficient which demonstrate good mixing. The marginal posterior distributions of the other coefficients have similar characteristics. Based on the Bayesian coefficients convergence diagnostics, the Markov chain converges successfully, and the Bayesian estimation is reliable.

Table 4.4 Bayesian coefficients convergence diagnostics

Geweke diagnostic of convergence	z-scores
Intercept	1.08196
d_l_tbill_3m	-1.11816
d_l_unemployment_rate	-0.70955
d_l_CPI	0.87389
d_l_consumer_confidence	-0.01717
d_l_retail_sales	-0.80482
d_l_personal_earnings	-1.09172
d_l_IPI	-0.20158
d_l_dowjones_index	-0.82063
d_l_CS_houseprice_index	-1.91440
original_debt_to_income_ratio	0.94838
original_loan_to_value	0.96615
l_current_actual_upb	-0.63741
l_current_interest_rate	-1.12304
l_remaining_months	0.63890
loan_age	-0.51490
loan_age_sq	0.32803

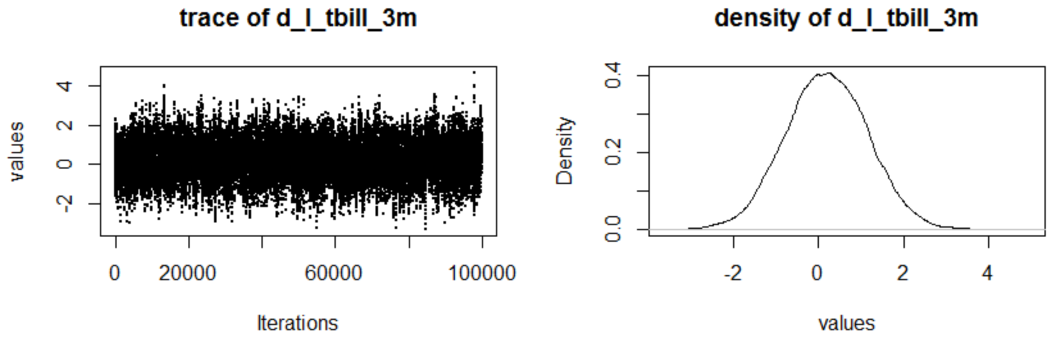

Fig. 4.1 Trace and density plots for the marginal posterior distribution of a coefficient

4.5.2 Prediction results using frequentist and Bayesian methods

The predicted and observed default rates are calculated based on Eqs. (4.4) and (4.5). Fig. 4.2 and Fig. 4.3 show that the default rates in the training and test samples are well predicted using both the frequentist and Bayesian methods. The default rate predictions follow the trend and fluctuation of the observed default rate along the time periods.

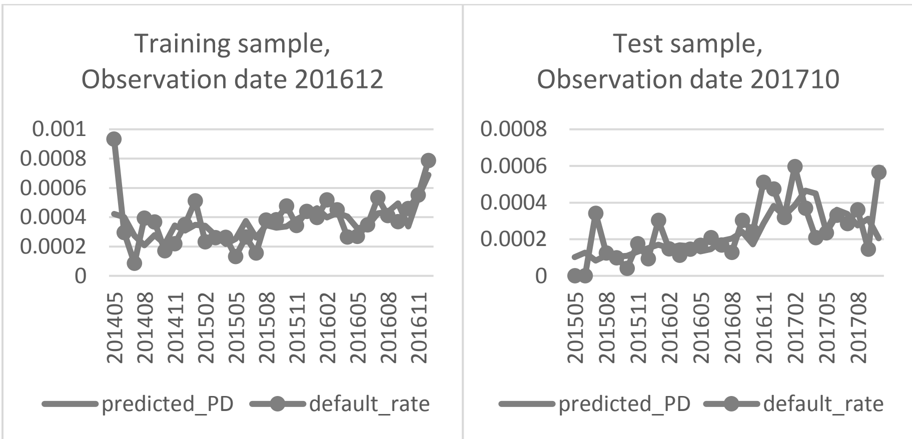

Fig. 4.2 Predicted and observed DR in the training and test samples using a frequentist approach

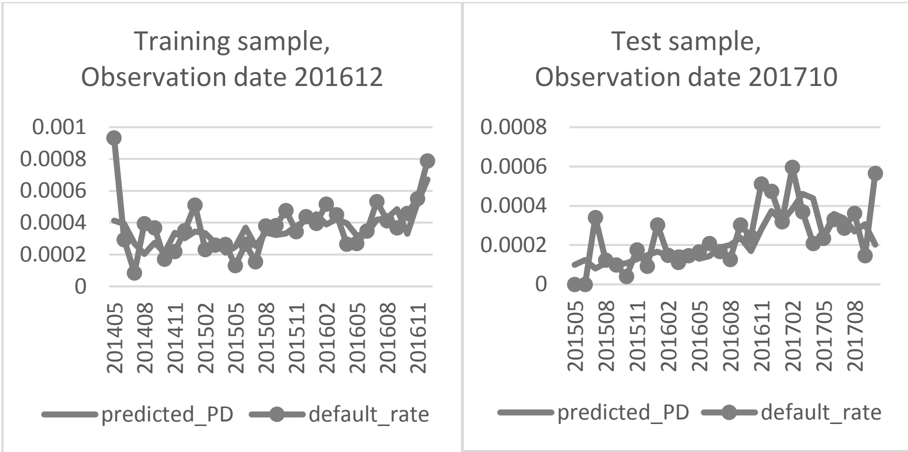

Fig. 4.3 Predicted and observed DR in the training and test samples using a Bayesian approach

Table 4.5 presents the mean absolute differences between the predicted and observed default rates in the training and test samples using the frequentist and Bayesian approaches. The default rate predictions are close to the observed default rates on average with the mean absolute differences in the two samples using the two approaches being approximately 0.000095.

Table 4.5 Mean absolute difference between the predicted and observed default rates

Measure	Sample	Frequentist	Bayesian
Mean absolute difference	Train	0.0000984	0.0000993
	Test	0.0000928	0.0000923

4.5.3 Performance results using the frequentist and Bayesian methods

Table 4.6 shows the performance results on the training and test samples using the frequentist and Bayesian methods in the duration of the first 12 months since each account's opening based on Eq. (4.3).

Table 4.6 Performance results in the duration of the first 12 months

Approach	Sample	Performance measures				
		H-measure	GINI	AUC	AUCH	K-S
Frequentist	Train	0.1122979	0.3965329	0.6982665	0.7068052	0.3230177
	Test	0.1047113	0.3407828	0.6703914	0.6856272	0.2827534
Bayesian	Train	0.1119204	0.3965452	0.6982726	0.7067148	0.3222444
	Test	0.1042898	0.3405327	0.6702664	0.6852344	0.2839313

The AUC and AUCH measures based on the training sample are around 70% whereas the AUC and AUCH measures based on the test sample are around 68%. The GINI coefficients based on the training sample are around 40% whereas those based on the test sample are about 34%. The H measure results based on the training sample are about 11% whereas those on the test sample are a little above 10%. The performance results show good predictive accuracy of the models.

4.5.4 Stress testing results using the frequentist and Bayesian methods

We have carried out stress testing in 3 ways using frequentist coefficient mean estimates, Bayesian coefficient mean estimates, and the Bayesian coefficient posterior distribution. We ensure that there are equal numbers of scenarios using all three methods. In the frequentist and Bayesian stress tests using coefficient mean estimates, we take 22500 random draws with replacement from past economic scenarios between Jan 1999 to Sept 2017. In the Bayesian stress test, using the posterior distribution and taking both macroeconomic risk and estimation risk into consideration, for computational efficiency, we take each economic scenario between Jan 1999 to Sept 2017 once (225 observations altogether). The values of the macroeconomic variables in each time period are drawn simultaneously.

In the Bayesian stress test using the posterior distribution approach, we take 100 random draws from the posterior distribution. That is, each of the 225 vectors of macroeconomic values is combined with 100 draws from the posterior distribution. Each draw forming the Bayesian posterior distribution includes all the coefficients.

We take each draw of the coefficients simultaneously as opposed to sampling from the marginal posterior distribution of each coefficient individually. For the coefficients for the macroeconomic variables and the constant, values in the posterior random draws are used. For the coefficients for the account specific variables, Bayesian coefficient mean estimates are used.

Stress testing is performed on the test sample. For computational efficiency, we take a random sample of 50% of the accounts in the test sample. We then use all the accounts in this sample that live to the time period upon which stress testing is performed. The time period that stress testing is performed upon is October 2017 which is the observation date of the test sample.

To avoid sampling bias, we apply bootstrapping for stress testing computations. For each one of three approaches (i.e. the frequentist mean estimate approach, the Bayesian mean estimate approach, and the Bayesian posterior distribution approach), the stress testing procedure is repeated 100 times, each with random simulations for the samplings of macroeconomic scenarios, the Bayesian coefficient posterior distribution (when using the posterior distribution approach), and the error terms. We then collect all the estimated default rates obtained in the 100 computations to build the empirical simulated default rate distribution for each of the three stress testing methods.

4.5.4.1 Stress testing results

Table 4.7 presents a comparison of different VaRs, means, and standard deviations of the simulated default rate distributions using the frequentist and Bayesian approaches as well as the observed default rate in October 2017.

Table 4.7 Statistics of the simulated default rate distributions using frequentist mean estimates, Bayesian mean estimates, and random samples from the Bayesian posterior distribution and the observed default rate in the test sample in October 2017

Approaches / Statistics of DR distributions	Frequentist mean estimates	Bayesian mean estimates	Bayesian posterior distribution
Mean of simulated DR distribution	0.000319	0.000314	0.000784
St.d of simulated DR distribution	0.000270	0.000267	0.002887
95% VaR of simulated DR distribution	0.000782	0.000782	0.002884
99% VaR of simulated DR distribution	0.001320	0.001271	0.008554
Observed DR in October 2017 in the test sample		0.000565	

Using the frequentist mean estimates approach, the mean of the simulated DR distribution is lower than the observed default rate. The standard deviation of the distribution is approximately 0.00027. The 95% and 99% VaRs of the DR distribution are about 0.0008 and 0.0013 respectively. Both are larger than the observed default rate. In other words, the stress testing succeeds in yielding VaRs above the observed default rate if we use the frequentist approach without considering estimation risk. The statistics for the simulated DR distribution using Bayesian mean estimates give the same conclusions since firstly both estimation methods rely fully on information contained in the data and secondly, both stress testing methods only consider macroeconomic scenarios without considering estimation risk.

For the simulated DR distribution using the Bayesian approach with random draws from the Bayesian posterior, the distribution mean is about 0.00078, much larger than the simulated DR distribution mean using the frequentist or Bayesian coefficient mean estimates. The 95% and 99% VaRs are about 0.0029 and 0.0086 respectively. The standard deviation of the distribution is approximately 0.0029, much higher than that of the simulated DR distribution using the frequentist or Bayesian coefficient mean estimates. In this approach, the mean, 95% and 99% VaRs of the simulated DR distribution all successfully exceed the observed default rate when we use random draws from the Bayesian posterior distribution with both macroeconomic scenarios and estimation risk taken into account. Notice, the 99% VaR obtained using this stress testing method is very close to the observed default rates during the 07/08 financial crisis. For instance, the observed default rate for accounts originated in 2007 with December 2009 as the observation date is 0.008485 at the observation time, based on a random sample of 50000 accounts from the same database.

In summary, the observed default rate in the stress testing period is within the 95% and 99% VaRs both using coefficient mean estimates methods and the Bayesian posterior distribution method. The stress testing results show that statistics such as the VaRs and the standard deviation of the simulated DR distribution increase as estimation risk is introduced.

We propose the following way to measure the relative sizes of macroeconomic stress and estimation risk. The distribution mean of the frequentist simulated DR distribution using coefficient mean estimates represents the expected default rate under normal macroeconomic circumstances with neither macroeconomic stress nor estimation risk considered. We use this value as a benchmark to measure macroeconomic stress and estimation risk. The 99% VaR of the frequentist simulated DR distribution represents the simulated default rate in stressed macroeconomic conditions but without considering coefficient uncertainty. The 99% VaR of the Bayesian simulated DR distribution using the coefficient posterior distribution approach is the simulated default rate both in stressed macroeconomic circumstances and with coefficient uncertainty addressed. Therefore the unexpected loss that comes from macroeconomic stress can be quantified in the traditional way by comparing the distribution mean (0.0003) and the 99% VaR (0.0013) of the frequentist DR distribution. Furthermore, a combination of the stress from macroeconomic stress and the coefficient uncertainty can be quantified by comparing the mean (0.0003) of the frequentist simulated DR distribution, which is the expected default rate in tranquil economic circumstances and without estimation risk, and the

99% VaR (0.0086) of the Bayesian simulated DR distribution that uses the coefficient posterior distribution approach, which both addresses macroeconomic stress and estimation risk. In other words, the size of macroeconomic stress is approximately 0.001 (=0.001320-0.000319) measured in simulated default rate. The size of a combination of macroeconomic stress and the coefficient uncertainty is around 0.0082 (=0.008554-0.000319) measured in simulated default rate. Therefore, estimation risk contributes much higher than macroeconomic stress to the simulated default rate.

4.5.4.2 Stress testing results comparison

Fig. 4.4 compares the simulated DR distributions between using the frequentist and Bayesian point estimate approaches. The simulated DR distributions using frequentist and Bayesian coefficient mean estimates are almost identical since the coefficient estimates are very similar between the frequentist and non-informative Bayesian approaches.

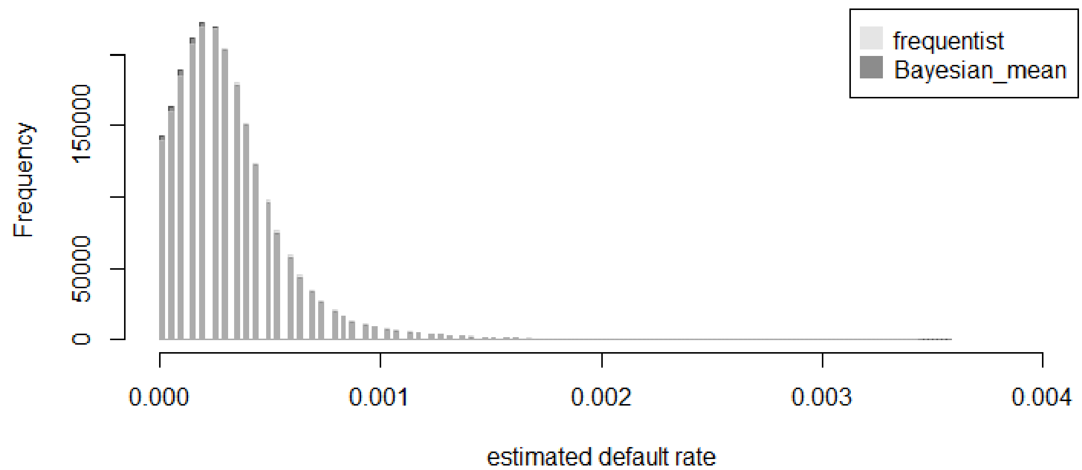

Fig. 4.4 Histograms of the simulated default rate distributions using frequentist and Bayesian coefficient mean estimates

Fig. 4.5 compares the simulated DR distributions between using the frequentist point estimate approach and the Bayesian posterior distribution approach. Fig. 4.6 Shows the tails of the two distributions. It can be seen from the two figures that the simulated DR distribution has a fatter and longer tail when using the Bayesian posterior distribution approach compared to when using the mean estimate approach. At midrange default rates, the simulated DR distribution has higher frequencies using the mean estimate approach than when using the Bayesian distribution approach. At other default rates, the reverse is true.

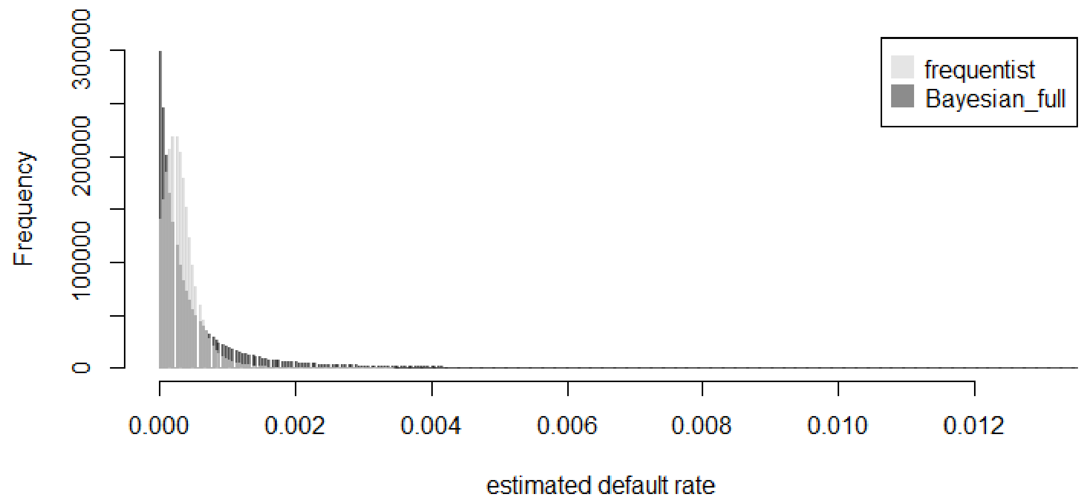

Fig. 4.5 Histograms of simulated default rate distributions using frequentist coefficient mean estimates and the Bayesian coefficient posterior distribution

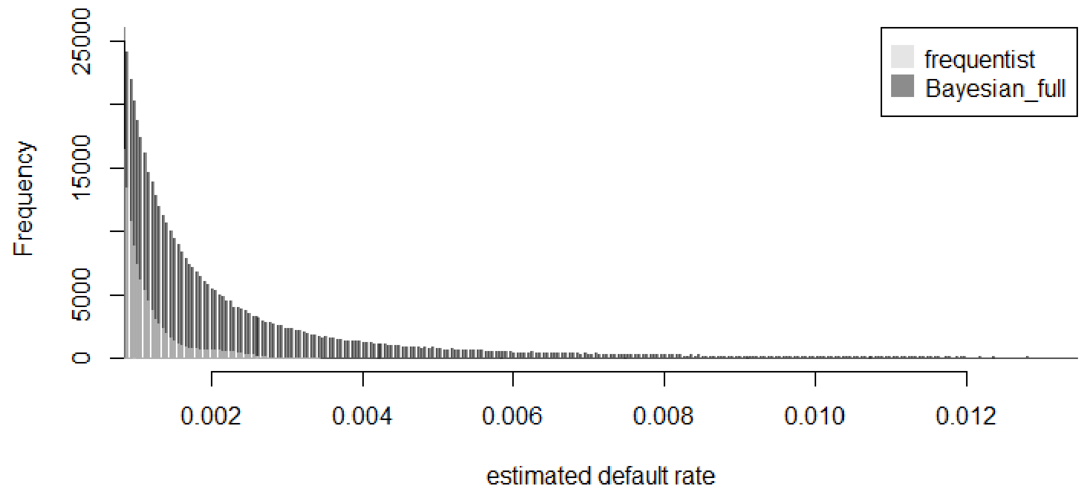

Fig. 4.6 Tails of the simulated default rate distributions using frequentist coefficient mean estimates and the Bayesian coefficient posterior distribution

Since for the Bayesian posterior distribution approach there are two sources of variation, that is from both the macroeconomic variables and coefficient variables instead of just from the macroeconomic variables alone, the measurements of variation such as the standard deviation and variance are larger.

In the literature, when macroeconomic shocks such as decreases in GDP and increases in interest rates are introduced, the simulated default rates increase (Jokivuolle & Viren, 2013; Sorge & Virolainen, 2006). Similarly, when introducing estimation risk into stress testing, we expect the simulated default rate of a scenario to increase further.

In the Bayesian posterior distribution approach, higher estimation risk is taken into account which results in higher simulated default rates in scenarios that use draws from areas of the posterior distribution that are far away from the coefficient mean estimates, such as the tails, causing the simulated DR distribution to have higher VaRs compared to using the coefficient mean estimates approach.

Table 4.8 shows the monetary values of credit loss based on the 99% VaRs for an average account in October 2017. For PD, we use the 99% VaRs of the simulated DR distributions. We assume the EAD of an account is the average current unpaid balance among accounts originated in 2015, which is the population data from which the test sample is taken, and alive in October 2017. We assume the fraction of EAD that is not recovered is 100%.

Table 4.8 Estimated monetary values for the 99% VaRs at the observation date based on stress tests using coefficient mean estimate approaches and the Bayesian posterior distribution approach for an average account

EAD (Average current unpaid balance in 201710)	$ 204115.5
The monetary value for 99% VaR (frequentist coefficient mean estimate approach)	$ 269.4
The monetary value for 99% VaR (Bayesian coefficient mean estimate approach)	$ 259.4
The monetary value for 99% VaR (Bayesian posterior distribution approach)	$ 1746.0

The estimated monetary value of the 99% VaR for an account on average is about $ 264 when ignoring estimation risk. The loss is around $ 1746 when estimation risk is included in stress testing. Considering there were 1.065 million accounts that were both originated in 2015 and alive in October 2017 in the Freddie Mac dataset population, it is clear that when the stress testing exercises only address macroeconomic shocks and ignore estimation uncertainty, they underestimate credit loss considerably.

4.6 Conclusions

Credit risk stress testing is a topic that attracts a growing research interest in the operational research literature. Our paper contributes to the literature in that we introduce estimation risk into stress testing to reduce credit risk underestimation. We demonstrate how a Bayesian approach and the Bayesian coefficient posterior distribution can be employed in stress testing to account for the potential credit risk underestimation induced by ignoring parameter uncertainty and estimation risk. In the stress testing application, we model both macroeconomic stress and coefficient uncertainty. We apply the Bayesian coefficient posterior distribution instead of coefficient point estimates to the stress test model to include various possible coefficient values and their corresponding probabilities.

In this paper, we use discrete time hazard analysis to model credit default risk over time based on U.S. mortgage loan data. We employ maximum likelihood estimation and the Metropolis-Hasting algorithm respectively for the frequentist and Bayesian approaches. In the Bayesian PD modelling and coefficient estimation, we use Bayesian non-informative priors to ensure the coefficient point estimate results are essentially the same between the frequentist and Bayesian methods, so that the differences in the stress testing results between using posterior distribution and point estimates are mainly due to the accommodation of estimation risk.

In the stress testing step, our Bayesian framework not only takes random draws from the historical scenarios of the macroeconomic variables but also considers estimation risk by simulating from the Bayesian coefficient posterior distribution. By employing Bayesian simulation of coefficients in stress testing, we model the uncertainty of coefficients thus providing more conservative estimates of credit risk by addressing the estimation variation. Furthermore, since the number of draws from different areas of the posterior is proportionate to the posterior probability of these areas, when we include the less likely coefficient estimates from the posterior distribution, we also take into consideration their corresponding low probability thus avoiding unnecessarily putting high weight on unlikely estimates.

Our main finding is that with the Bayesian stress testing approach using the posterior distribution, we obtain a broader simulated default rate distribution with higher VaRs and larger variance compared to stress testing approaches using coefficient mean estimates. The simulated DR distribution obtained using the Bayesian posterior distribution approach has a standard deviation 10.7 times as large as that using the parameter mean estimates approach. Moreover, the 95% and 99% VaRs of the estimated DR distribution using the Bayesian posterior distribution approach are around 3.7 and 6.5 times the 95% and 99% VaRs using the point estimate approach. The credit loss computed when estimation risk is included is much higher, around 6.5 times as much as the credit loss when estimation risk is ignored.

The results show that if the financial institutions use the traditional stress testing methods without addressing coefficient uncertainty, they could substantially underestimate default rates, and credit loss levels. Therefore it is essential for financial institutions and regulators to include estimation risk in their stress testing applications so that they do not underestimate credit risk and so that the amount of capital they keep accordingly does not fall short of the amount needed to give conventional levels of protection to depositors.

Chapter 5

Macroeconomic shocks on model parameters: Stress Testing Mortgage Loan Default Rate

5.1 Introduction

The 2008 financial crisis raised major concerns about assessments of borrowers' credit risk and the soundness of the financial system. While there is a growing interest in stress testing credit risk in response to macroeconomic shocks, research on the effect of macroeconomic shocks on stress testing models when the influence of changes in model parameters are considered is rare in comparison. Furthermore, little research has stress tested credit risk in response to model changes in a Bayesian framework. In this chapter, we consider the impact of the differences between parameters of the stress testing models between normal and stressed scenarios on default rate predictions using both frequentist and Bayesian methods. By using the Bayesian method, we also introduce estimation risk into the stress testing process.

We contribute to the literature in the following ways. Firstly, we incorporate parameter instability between crisis and non-crisis time periods into a stress testing methodology in a Bayesian framework. Prior work on stress testing mostly focuses on the changes in the values of macroeconomic variables while using the same model parameters between the two scenarios (Dua & Kapur, 2018; Jacobs, 2018; Jokivuolle & Viren, 2013; Kanno, 2015; Kapinos & Mitnik, 2016; Papadopoulos et al., 2016; Sorge & Virolainen, 2006). Some papers investigate parameter changes in credit risk models between different time periods, but do not apply it to stress testing (Jacobs & Sensenbrenner, 2018; Leow & Crook, 2016). The few stress testing papers that address parameter changes are primarily in the frequentist framework (Tsukahara et al., 2016). Siemsen and Vilsmeier (2018) consider the impact of variable selection risk on stress testing results using Bayesian methods. However, they consider model risk within the same scenario instead of model risk that arises from shifting between scenarios. Secondly, the use of Bayesian estimation in our work allows the introduction of estimation risk in a stress testing approach that addresses parameter instability by using a Bayesian coefficient posterior distribution as the source of coefficient estimates, instead of point estimates as in the frequentist approach.

A variety of modelling approaches have been used for stress testing in the frequentist framework but without considering parameter instability between models built on different scenarios. Kapinos and Mitnik (2016) employ a LASSO variables selection method and principal component analysis to identify macroeconomic and bank-specific drivers that influence banks' net revenue and net charge-offs and to preserve the variance-covariance structure between these variables. They estimate the model parameters based on 2013 and 2014 data relating to medium and large U.S. banks, each using several alternative models. In the forecasted values of the dependent variables, they employ the same parameter estimates for the baseline and stressed scenarios for each dataset. Based on their stress testing results, they argue that the capitalization of the banks in the dataset is not adequate under the stressed scenarios. Kanno (2015) and Dua and Kapur (2018) stress test companies' credit scores and banks' balance sheet indicators respectively with panel data models at micro level using distribution and point forecasting stress testing approaches, respectively. Both researchers employ the same parameter estimates between baseline and stressed scenarios in stress testing. Jacobs (2018) employs vector autoregressive and multivariate adaptive regression splines models to study the relationship between macroeconomic drivers and net charge-off rates. The parameter estimates are applied in both the stressed and baseline scenarios. Other stress testing papers (Jokivuolle & Viren, 2013; Sorge & Virolainen, 2006) also focus on the impact of macroeconomic stress without considering the instability of model parameters between the stressed and normal scenarios.

A few stress testing papers in the frequentist framework take into consideration the changes in model parameters between stressed and normal scenarios. Tsukahara et al. (2016) simulate both hypothetical default behaviour and values of explanatory variables in normal and adverse scenarios to stress test the influence that changes in the mean, variance, and correlation of the covariates have on validation measures, such as AUROC. They also employed an empirical dataset for the same task. They develop account level logistic models for the baseline and stressed scenarios for estimation and stress testing separately hence taking the changes in the model parameters into consideration. Jacobs and Sensenbrenner (2018) use vector autoregressive and Markov switching vector autoregressive models to study the relationship between credit loss and macroeconomic variables. They find that the Markov switching model gives more accurate forecasts due to its regime switching framework that can better accommodate normal and extreme events observed in history.

Some papers incorporate Bayesian methods into credit stress testing. But few take into account model parameter changes between scenarios, and none address coefficient estimation uncertainty. Louzis (2017) investigates the response of the consumer price index and loan size to the latest Greek crisis using a Bayesian vector autoregressive model at an aggregate level. Petropoulos et al. (2018) use Bayesian model averaging, regime switching, and a linear regression model to study the macro determinants of the size of financial institutions' non-performing loans at an aggregate level. They find that the Markov regime switching models which capture the parameter changes between recession and expansions have the best predictive accuracy among all models. Papadopoulos et al. (2016) use a Bayesian model averaging method to reduce variable selection risk in studying the relationship between non-performing loans and macroeconomic covariates. They apply their estimation results in a stress testing exercise that considers the adverse 1% probability levels of the macroeconomic variables as the stressed scenario. Siemsen and Vilsmeier (2018) also employ a Bayesian model averaging method. They take both macro shocks and model uncertainty into consideration by exploring the influence that different model combinations in Bayesian model averaging have on the stress testing results.

Our work is closely related to Berkowitz (1999) in that it considers shifts in a predicted loss distribution due to macro stress. Our work also closely relates to Bellotti and Crook (2013, 2014) in that it assumes a single unified distribution of macroeconomic variables. This research is also closely related to Leow and Crook (2016) in that it emphasizes the importance of structural breaks on models' parameters due to the financial crisis.

We model and stress test the probability of default with discrete time hazard analysis using both frequentist and Bayesian estimation methods based on U.S. mortgage loan data during the 2007 - 2009 financial crisis and a post-crisis tranquil time period. We find that the default rate distribution that we simulate, based on models built on crisis period data, is shifted to the right relative to that built for the tranquil period, with higher VaRs and variance. For instance, when using a frequentist approach without considering estimation risk, the VaR is approximately 60% higher using model parameters estimated on the crisis dataset than on the tranquil dataset. For models built on the same scenario, the default rate distribution with estimation risk included, using Bayesian methods, has a larger variance and larger VaRs than the distribution that ignores estimation risk using frequentist point estimates. For instance, if under crisis scenarios, the VaR obtained when estimation risk is included is approximately 90% higher than when estimation risk is ignored. Both estimation risk and macroeconomic shocks on model parameters cause higher VaRs in the simulated DR distributions, and subsequently higher required capital, with a combined influence causing the required capital to increase by over 1.7 times compared to the baseline that addresses neither risk. The influence of estimation risk on simulated default rate distributions and VaRs is more drastic than the influence of parameter instability between models built on the stressed and non-stressed scenarios.

The structure of this chapter is as follows. Section 2 presents the discrete time hazard model and stress testing models using the frequentist coefficient point estimates and the Bayesian posterior distribution approaches. It also outlines our stress testing method that incorporates both parameter changes between scenarios and the estimation risk. Section 3 describes the data and variables for this research. Section 4 presents the estimation, performance, and stress testing results and discusses their implications. Section 5 provides a summary of the main findings and concludes.

5.2 Methodology

5.2.1 Discrete time hazard model, frequentist and Bayesian stress testing models

Discrete time hazard model

To estimate PD, we use the same discrete time hazard model in Chapter 4. The probability of default $p_{i,t}$ for account i at duration time t is:

$$p_{i,t} = p(d_{i,t} = 1 \mid d_{i,t_0} = 0,\ for\ all\ t_0 < t) = \frac{1}{1+e^{-y_{i,t}}} \tag{5.1}$$

where $\boldsymbol{y}_{i,t} = \boldsymbol{x}_{t-3}'\boldsymbol{\beta} + \boldsymbol{w}_i'\boldsymbol{\delta} + \boldsymbol{u}_{i,t-3}'\boldsymbol{\gamma} + \boldsymbol{g}(\boldsymbol{t})'\boldsymbol{\rho}$, and where $d_{i,t} = 1$ if account i defaults at duration time t, and $d_{i,t} = 0$ otherwise, $\boldsymbol{x}_{t-3}$ denotes macroeconomic variables at duration time t lagged 3 months, $\boldsymbol{w}_i$ denotes application variables for individual i, $\boldsymbol{u}_{i,t-3}$ denotes behavioural variables for individual i at duration time t lagged 3 months, and $\boldsymbol{g(t)}$ denotes the account duration functions at duration time t for individual i. Together, $\boldsymbol{g(t)}$, $\boldsymbol{u}_{i,t-3}$, $\boldsymbol{x}_{t-3}$, and $\boldsymbol{w}_i$ are the covariates, and $\boldsymbol{\beta}$

[9], $\boldsymbol{\delta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\rho}$ are vectors of coefficients for macroeconomic, behavioural, application, and duration variables respectively. This model follows Bellotti & Crook (2009, 2013, 2014).

We use the maximum likelihood estimation method (Greene, 2011) to obtain the frequentist coefficient mean estimates and the random walk Metropolis-Hastings algorithm with non-informative priors (Koop et al., 2007) to obtain the simulated Bayesian posterior distribution.

Frequentist and Bayesian stress testing models

Our simulated default rates are computed using the same logistic models as in chapter 4. We briefly recap on them here. Our stress testing model is the latent variable model for logistic regression at t_s, which is the calendar time at which stress testing is applied. t_{a_i} denotes the calendar time of the opening of account i. $\hat{\boldsymbol{R}}_{j,t_s}$ denotes the simulated proportion of individuals in the risk set that are predicted to default in scenario j at the calendar time t_s, given that they did not default prior to t_s. Similarly, $\hat{R}_{j,d,t_s}$ denotes the corresponding simulated default rate in the Bayesian approach using the dth random draw from the posterior distribution. In the frequentist stress testing approach, we use coefficient mean estimates. In the Bayesian stress testing approach, to address estimation risk, for the constant and macroeconomic coefficients we use Bayesian posterior draws as coefficient estimates instead of the mean estimates, so that the ranges of possible coefficient estimates are taken into account. Bayesian mean estimates are used for the coefficients of the account level variables. $\boldsymbol{x}_j$ represents the observed values of the macroeconomic variables in a random month in history before t_s. $\boldsymbol{\beta}_d$ denotes the

[9] The constant is included in $\boldsymbol{\beta}$.

dth draw from the Bayesian posterior distribution for the constant and the macroeconomic coefficients.

1) Following Bellotti and Crook (2013, 2014), the simulated default rate $\hat{R}_{j,t_s}$ at stress testing time period t_s in the jth macroeconomic scenario using the frequentist coefficient mean estimates is assumed to follow the model:

$$\hat{R}_{j,t_s} = \frac{1}{n}\sum_{i=1}^{n} I(y^*_{j,i,t_s} = \boldsymbol{x}'_j\boldsymbol{\beta} + \boldsymbol{w}'_i\boldsymbol{\delta} + \boldsymbol{u}'_{i,t_s-t_{a_i}-3}\boldsymbol{\gamma} + \boldsymbol{g}'(t_s - t_{a_i})\boldsymbol{\rho} + \varepsilon_{j,i,t_s} > 0) \tag{5.2}$$

2) The simulated default rate $\hat{R}_{j,d,t_s}$ at the stress testing time period t_s in the jth macroeconomic scenario and using the dth Bayesian coefficient draw is assumed to follow the model:

$$\hat{R}_{j,d,t_s} = \frac{1}{n}\sum_{i=1}^{n} I(y^*_{j,d,i,t_s} = \boldsymbol{x}'_j\boldsymbol{\beta}_d + \boldsymbol{w}'_i\boldsymbol{\delta} + \boldsymbol{u}'_{i,t_s-t_{a_i}-3}\boldsymbol{\gamma} + \boldsymbol{g}'(t_s - t_{a_i})\boldsymbol{\rho} + \varepsilon_{j,d,i,t_s} > 0), \tag{5.3}$$
$$d = 1, 2, \ldots, D$$

5.2.2 A new stress testing method

5.2.2.1 Notation

Berkowitz (1999) expresses his model in terms of asset returns whereby if the asset return falls below a cutoff, default is implied. He argues that stressed economic conditions cause the underlying factors explaining asset returns to follow a different distribution from the factors under normal scenarios. The shift in the distribution for the explanatory variables is transferred to the asset returns distribution. Therefore, instead of following the same distribution of predicted returns under normal scenarios, predicted returns under stressed scenarios follow a separate distribution.

The valuation models used are assumed the same between the stressed and normal scenarios.

Briefly, in Berkowitz (1999), the simulated asset return distributions under stressed and normal scenarios are the distributions of values represented as:

$$\hat{r}_{j,stress} = P(\hat{\boldsymbol{x}}_j(\boldsymbol{h}_{stress}(\bullet))) \tag{5.4}$$

$$\hat{r}_{j,normal} = P(\hat{\boldsymbol{x}}_j(\boldsymbol{h}_{normal}(\bullet))) \tag{5.5}$$

In which:

$\hat{r}_{j,stress}$ denotes an estimated asset return from the simulated asset return distribution under stressed scenarios.

$\hat{r}_{j,normal}$ denotes an estimated asset return from the simulated asset return distribution under normal scenarios.

$\boldsymbol{h}_{stress}(\bullet)$ denotes a multivariate distribution describing the joint behaviour of the explanatory variables under stressed economic conditions.

$\boldsymbol{h}_{normal}(\bullet)$ denotes a multivariate distribution describing the joint behaviour of the explanatory variables under normal economic conditions.

$\hat{\boldsymbol{x}}_j$ denotes a vector of the predicted values for the explanatory variables for the time period stress testing is applied to. It is the $j\,th$ draw of repeated draws from the explanatory variables distribution $\boldsymbol{h}_{stress}(\bullet)$ or $\boldsymbol{h}_{normal}(\bullet)$.

$P(\bullet)$ denotes a valuation model to simulate the predicted return. In Berkowitz (1999), this model is assumed the same under stressed and normal scenarios.

Fig. 5.1 illustrates the relative positions of hypothetical credit loss distributions in stressed and normal macroeconomic scenarios in credit risk stress testing using the method proposed by Berkowitz (1999). The loss distribution is assumed to shift under the stressed scenario from the distribution under the normal scenario.

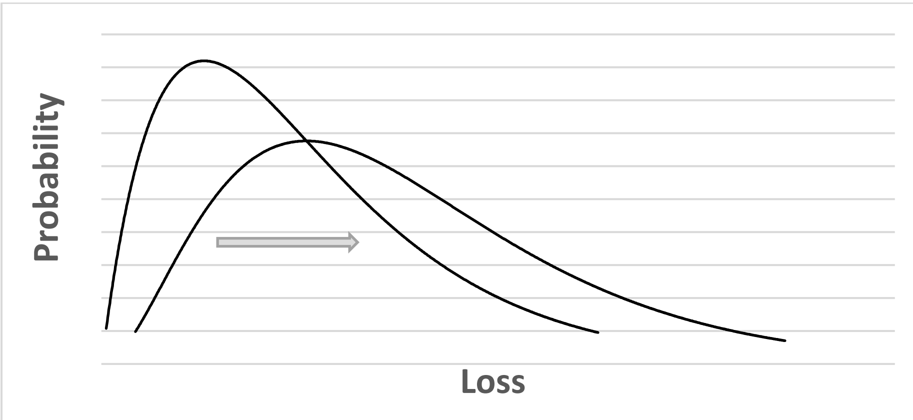

Fig. 5.1 Hypothetical relative positions of simulated default rate distributions under stressed and normal scenarios using the two distributions approach proposed by Berkowitz (1999). Adapted from Sorge and Virolainen (2006)

Whilst our model differs crucially from that of Berkowitz (1999), we also use a two distributions approach to stress testing. That is, we assume stress would shift the simulated default rate distribution from the simulated default rate distribution under normal scenarios. In addition, we <u>also</u> consider the <u>changes to parameters</u> in the credit risk models caused by stress. Therefore, in our method, we use different models for stressed and normal scenarios whereas Berkowitz (1999) uses the same model for the two scenarios. In other words, in Berkowitz (1999), stress is added to the input vector $\boldsymbol{X}$ whereas our stress is added to the parameters $\boldsymbol{\theta}$ [10] as well as the

[10] $\boldsymbol{\theta}$ includes all parameters β , δ , γ , ρ .

input vector $\boldsymbol{X}$. Furthermore, as a model contribution to the stress testing literature, we do not just have different coefficient means but also different coefficient dispersions for the stressed and non-stressed distributions with the use of a Bayesian coefficient posterior distribution.

To give more details, the simulated default rate distributions under stressed and normal scenarios, with and without estimation risk addressed are:

$$\hat{R}_{j,normal} = f_{normal}(\hat{\boldsymbol{x}}_j(\boldsymbol{h}(\bullet)), E(\hat{\boldsymbol{\beta}}_{normal})) \tag{5.6}$$

$$\hat{\boldsymbol{R}}_{j,normal,ER} = f_{normal}(\hat{\boldsymbol{x}}_j(\boldsymbol{h}(\bullet)), \boldsymbol{\beta}_{normal}) \tag{5.7}$$

$$\hat{R}_{j,stress} = f_{stress}(\hat{\boldsymbol{x}}_j(\boldsymbol{h}(\bullet)), E(\hat{\boldsymbol{\beta}}_{stress})) \tag{5.8}$$

$$\hat{\boldsymbol{R}}_{j,stress,ER} = f_{stress}(\hat{\boldsymbol{x}}_j(\boldsymbol{h}(\bullet)), \boldsymbol{\beta}_{stress}) \tag{5.9}$$

in which:

$\hat{R}_{j,normal}$ denotes an estimated default rate in the simulated default rate distribution using observed past values of the macroeconomic variables in month j, for normal scenarios and without estimation risk considered. All the $\hat{R}_{j,normal}$, $j = 1, 2, \cdots T$ values form the simulated default rate distribution $\hat{\boldsymbol{R}}_{normal}$.

$\hat{\boldsymbol{R}}_{j,normal,ER}$ denotes a vector of D estimated default rates in the simulated default rate distribution for month j using the observed past values of the macroeconomic variables in month j, for normal scenarios but with estimation risk considered using D random draws from the Bayesian posterior distribution. All the $\hat{\boldsymbol{R}}_{j,normal,ER}$, $j=1,2,\cdots T$ values form the simulated default rate distribution $\hat{\boldsymbol{R}}_{normal,ER}$.

$\hat{R}_{j,stress}$ denotes an estimated default rate in month j in the simulated default rate distribution using the observed past values of the macroeconomic variables, for stressed scenarios but without estimation risk considered. All the $\hat{R}_{j,stress}$, $j=1,2,\cdots T$ values form the simulated default rate distribution $\hat{\boldsymbol{R}}_{stress}$.

$\hat{\boldsymbol{R}}_{j,stress,ER}$ denotes a vector of D estimated default rates in the simulated default rate distribution in month j using the observed past values of the macroeconomic variables, for stressed scenarios and <u>with</u> estimation risk considered using D random draws from the Bayesian posterior distribution. All the $\hat{\boldsymbol{R}}_{j,stress,ER}$, $j=1,2,\cdots T$ values form the simulated default rate distribution $\hat{\boldsymbol{R}}_{stress,ER}$.

$\boldsymbol{h}(\bullet)$ denotes a multivariate distribution describing the joint behaviour of the macroeconomic explanatory variables. We can assume $\boldsymbol{h}(\bullet)$ follows a theoretical distribution, such as following a normal distribution. Alternatively, we can use an empirical distribution derived by historical simulation. Thus, $\boldsymbol{h}(\bullet)$ can be formed by randomly resampling from past observations. That is, $\boldsymbol{h}(\bullet)$ assigns a probability $\frac{1}{T}$ to each of the historical observations $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ... \boldsymbol{x}_T\}$ (Berkowitz, 1999). In this paper, we employ the empirical distribution method by sampling from the past observations of the macroeconomic variables.

$\hat{\boldsymbol{x}}_j$ denotes a vector of the predicted values of the explanatory variables for the stress testing time period t_s. It consists of repeated draws from the distribution $\boldsymbol{h}(\bullet)$, $j=1,2,\cdots,T$.

$f_{normal}(\bullet)$ denotes a credit risk model to simulate default rates under normal scenarios.

$f_{stress}(\bullet)$ denotes a credit risk model to simulate default rates under stressed scenarios.

$E(\hat{\boldsymbol{\beta}}_{normal})$ denotes a vector of frequentist coefficient mean estimates for the macroeconomic variables and the constant obtained from the model built on the normal scenario based on the tranquil time period dataset.

$E(\hat{\boldsymbol{\beta}}_{stress})$ denotes a vector of frequentist coefficient mean estimates for the macroeconomic variables and the constant obtained from the model built on the stressed scenario based on the crisis time period dataset.

$\boldsymbol{\beta}_{normal}$ denotes a matrix of D random draws from the Bayesian posterior distribution for the macroeconomic coefficients and the constant obtained from the model built on the normal scenario based on the tranquil time period dataset.

$\boldsymbol{\beta}_{stress}$ denotes a matrix of D random draws from the Bayesian posterior distribution for the macroeconomic coefficients and the constant obtained from the model built on the stressed scenario based on the crisis time period dataset.

In the rest of this section, we further explain the simulation method for the macroeconomic scenarios and the macroeconomic coefficient variables with the Bayesian approach when estimation risk is considered. We illustrate the simulation method and procedure using models built on the stressed scenario dataset with and without estimation risk Eq. (5.8) and Eq. (5.9) as examples.[11]

5.2.2.2 The simulation method

5.2.2.2.1 Simulated values of $\hat{R}_{j,stress}$

To show the simulation method of the macroeconomic variables, Eq. (5.8) can be further written as Eq. (5.10). Individual univariate sampling of the macroeconomic variables may be incoherent (Crook et al., 2015). To preserve the structure of dependency between the macroeconomic variables, we simultaneously draw the values of all the macroeconomic variables in one past month $\hat{\boldsymbol{x}}_j(\boldsymbol{h}(\bullet)) = \begin{pmatrix} x_{j,1} & x_{j,2} & \cdots & x_{j,k} \end{pmatrix}_{1\times k}$ from the multivariate distribution of the macroeconomic explanatory variables $\boldsymbol{h}(\bullet) = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ \vdots & \vdots & \cdots & \vdots \\ x_{j,1} & x_{j,2} & \cdots & x_{j,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T,1} & x_{T,2} & \cdots & x_{T,k} \end{pmatrix}_{T\times k}$ [12],

where k denotes the number of coefficients for the macroeconomic variables and

[11] The simulation method for models built on the normal scenario dataset (i.e. Eq. (5.6) and Eq. (5.7)) are similar.

[12] To accommodate the constant, the first column of $\boldsymbol{h}(\bullet)$ is a vector of 1: $\begin{pmatrix} x_{1,1} \\ \vdots \\ x_{j,1} \\ \vdots \\ x_{T,1} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix}$

the constant. We can write $E(\hat{\boldsymbol{\beta}}_{stress}) = \begin{pmatrix} E(\hat{\beta}_1) \\ E(\hat{\beta}_2) \\ \vdots \\ E(\hat{\beta}_k) \end{pmatrix}_{k\times 1}$ [13], which is the vector of frequentist mean estimates for the constant and the macroeconomic coefficients.

Then $\hat{R}_{j,stress}$, the simulated DR in month j using the frequentist mean estimate approach without estimation risk, but built on the stressed scenario based on the crisis time period dataset is:

$$\hat{R}_{j,stress} = f_{stress}(\hat{\boldsymbol{x}}_j(\boldsymbol{h}(\bullet)), E(\hat{\boldsymbol{\beta}}_{stress}))$$
$$\Rightarrow \hat{R}_{j,stress} = f_{stress}(\begin{pmatrix} x_{j,1} & x_{j,2} & \cdots & x_{j,k} \end{pmatrix}_{1\times k} \times \begin{pmatrix} E(\hat{\beta}_1) \\ E(\hat{\beta}_2) \\ \vdots \\ E(\hat{\beta}_k) \end{pmatrix}_{k\times 1}) \tag{5.10}$$

We can decompose the vector $\hat{\boldsymbol{R}}_{stress}$ using Eq. (5.11). $\hat{\boldsymbol{R}}_{stress}$ consists of all the simulated DRs that form the simulated DR distribution for all the T macroeconomic scenarios using the frequentist mean estimate approach without including estimation risk but built on the stressed scenario based on the crisis time period dataset. The values for the macroeconomic variables in the grey frame are those obtained in one draw from all the macroeconomic variables, taken simultaneously from $\boldsymbol{h}(\bullet)$ in the same month in history, j.

[13] $E(\hat{\beta}_1)$ is the constant.

$$\underset{[T\times 1]}{\hat{\boldsymbol{R}}_{stress}} = f_{stress}\left(\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ \vdots & \vdots & \cdots & \vdots \\ \boxed{x_{j,1} \quad x_{j,2} \quad \cdots \quad x_{j,k}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T,1} & x_{T,2} & \cdots & x_{T,k} \end{pmatrix}_{T\times k} \times \begin{pmatrix} E(\hat{\beta}_1) \\ E(\hat{\beta}_2) \\ \vdots \\ E(\hat{\beta}_k) \end{pmatrix}_{k\times 1}\right) \tag{5.11}$$

5.2.2.2.2 Simulated values of $\hat{\boldsymbol{R}}_{j,stress,ER}$

We next show the simulation method for the macroeconomic variables and the Bayesian coefficients for the macroeconomic variables, Eq. (5.9) can be further written as Eq. (5.12). Since each draw forming the Bayesian (joint) posterior distribution includes all the coefficients, we use values of the macroeconomic coefficients and constant from the same draw from the Bayesian posterior distribution as opposed to simulating each independently from the marginal posterior distribution of each coefficient.

Thus we can write $\boldsymbol{\beta}_{stress} = \begin{pmatrix} \beta_{1,1} & \cdots & \beta_{1,d} & \cdots & \beta_{1,D} \\ \beta_{2,1} & \cdots & \beta_{2,d} & \cdots & \beta_{2,D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_{k,1} & \cdots & \beta_{k,d} & \cdots & \beta_{k,D} \end{pmatrix}_{k\times D}$, which denotes the D random draws from the Bayesian posterior distribution for the macroeconomic coefficients and the constant, and $\beta_d = \begin{pmatrix} \beta_{1,d} \\ \beta_{2,d} \\ \vdots \\ \beta_{k,d} \end{pmatrix}_{k\times 1}$ denotes the dth random draw from the Bayesian posterior distribution for the macroeconomic coefficients and the constant, $d = 1, 2, \cdots, D$.

Then $\hat{\boldsymbol{R}}_{j,stress,ER}$, the vector of D simulated DRs in month j using the Bayesian posterior distribution approach with estimation risk included and built on the stressed scenario based on the crisis time period dataset is:

$$\hat{\boldsymbol{R}}_{j,stress,ER} = f_{stress}(\hat{\boldsymbol{x}}_j(\boldsymbol{h}(\bullet)), \boldsymbol{\beta}_{stress})$$

$$\Rightarrow \hat{\boldsymbol{R}}_{j,stress,ER} = f_{stress}(\begin{pmatrix} x_{j,1} & x_{j,2} & \cdots & x_{j,k} \end{pmatrix}_{1\times k} \times \begin{pmatrix} \beta_{1,1} & \cdots & \beta_{1,d} & \cdots & \beta_{1,D} \\ \beta_{2,1} & \cdots & \beta_{2,d} & \cdots & \beta_{2,D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_{k,1} & \cdots & \beta_{k,d} & \cdots & \beta_{k,D} \end{pmatrix}_{k\times D})$$

(5.12)

We can decompose the vector $\hat{\boldsymbol{R}}_{stress,ER}$ using Eq. (5.13). $\hat{\boldsymbol{R}}_{stress,ER}$ consists of all the simulated DRs that form the simulated DR distribution for all the T macroeconomic scenarios and using all the D random draws from the posterior distribution using the Bayesian approach with estimation risk included and built on the stressed scenario based on the crisis time period dataset. The values for the constant and the macroeconomic coefficients in the grey frame are from the same draw from the Bayesian posterior distribution.

$$\underset{[T\times D]}{\hat{\boldsymbol{R}}_{stress,ER}} = f_{stress}(\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ \vdots & \vdots & \cdots & \vdots \\ x_{j,1} & x_{j,2} & \cdots & x_{j,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T,1} & x_{T,2} & \cdots & x_{T,k} \end{pmatrix}_{T\times k} \times \begin{pmatrix} \beta_{1,1} & \cdots & \boxed{\beta_{1,d}} & \cdots & \beta_{1,D} \\ \beta_{2,1} & \cdots & \beta_{2,d} & \cdots & \beta_{2,D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_{k,1} & \cdots & \beta_{k,d} & \cdots & \beta_{k,D} \end{pmatrix}_{k\times D})$$

(5.13)

In Fig. 5.2 we illustrate the potential relative positions of the hypothetical simulated default rate distributions based on baseline and stressed scenarios with and without estimation risk included. We use three Chi-square distributions to denote the hypothetical simulated default rate distributions under baseline scenarios using the frequentist approach, stressed scenarios using the frequentist approach, and stressed scenarios using the Bayesian approach. We assume that both parameter instability due to macroeconomic shocks and estimation risk will induce shifts from the baseline probability distribution based on the tranquil time period data and without estimation risk.

We expect the loss distribution in a stressed scenario, using the frequentist approach, to have a higher expected loss and larger variance than that of the baseline scenario using the same approach because that by addressing changes in model parameters in the stressed scenario from those in the baseline scenario we include the additional impact that macroeconomic risk has on the stress testing model; that is the impact on model parameters. We expect the loss distribution of a stressed scenario using the Bayesian distribution approach to have a higher expected loss and larger variance than that of the same scenario using the frequentist point estimate approach because of the inclusion of parameter uncertainty which is an additional source of variation and risk. It is unclear, on a hypothetical basis, whether shocks on model parameters or parameters as random variables (estimation risk) have a stronger effect on how spread out the loss distributions are. Therefore we omit the hypothetical loss distribution for the tranquil scenario using the Bayesian approach.

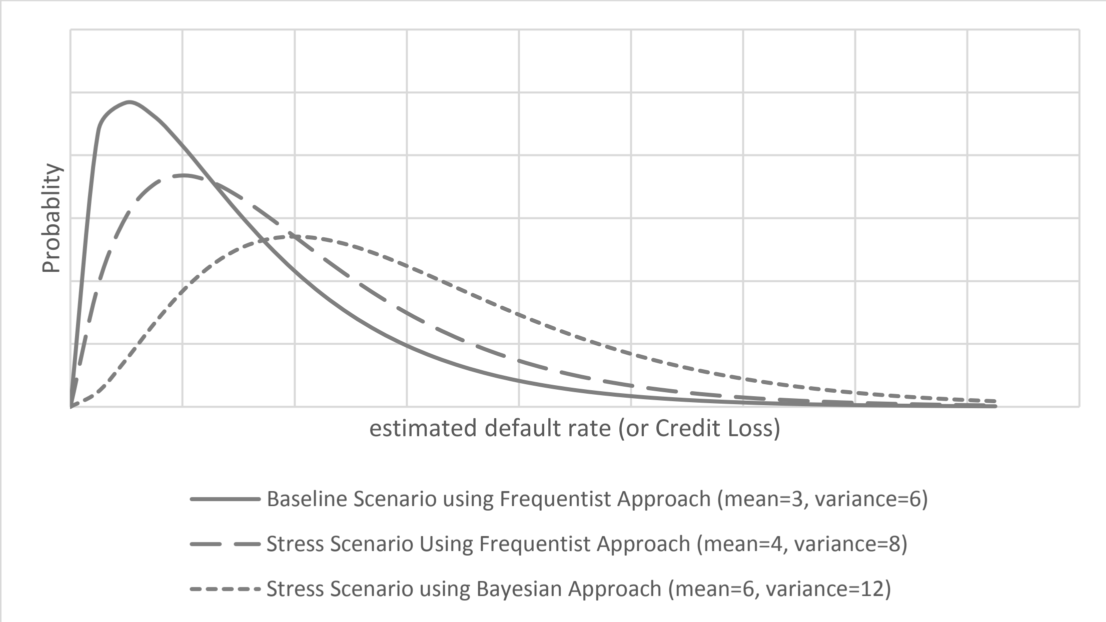

Fig. 5.2 Potential relative positions of hypothetical simulated default rate distributions using different stress testing methods

5.3 Data, variables

We use data from the same Freddie Mac single-family loan level dataset[14] that was used in the earlier chapters to illustrate our methods. We use the mortgage accounts that originated during the 12 months in 2007 as a training sample for a model developed during the crisis period. For the training sample for the tranquil period we use accounts that originated during the 12 months in 2010. We use accounts that originated during the 12 months in 2014 as the test sample. We take December 2009 and December 2012 as observation dates for the crisis and tranquil training samples respectively. For the test sample, we take December 2016 as the observation date. For each year, we use a random sample of 50000 loans.

We consider that an account is in default if it has in its payment history record an episode of at least 60 days delinquency. Again, we arrange the training and test

[14] Dataset url: http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.page

samples such that the observations after the first default of any accounts are set to missing values. This ensures that the econometric model is parameterised using data up until the first default, which is what is required for the single event hazard distribution. Table 5.1 illustrates the number of accounts and defaults in each sample.

Table 5.1 Training and test samples

	Train crisis (2007)	Train tranquil (2010)	Test (2014)
Observation date	December 2009	December 2012	December 2016
Number of accounts	50000	50000	50000
Defaults of accounts	4,201	307	415

Fig. 5.3 shows the monthly observed default rate between Jan 2002 and Dec 2016 using a pooled sample of accounts originated between 1999 and 2016, with 50000 accounts in each year following the way Freddie Mac database calculates its monthly default rates[15].

The choice of the crisis training sample is based on the time period of the 2007/2008 financial crisis. It is also decided according to the rapid rise in the default rate between the start of 2007 and the end of 2009 shown in Fig. 5.3, representing a fast deteriorating economic environment. Therefore we use accounts originated in 2007 as the crisis period training sample and use the end of 2009 as the observation date. Since the start of 2010, the observed default rate gradually decreased, which is a sign that economic conditions improved. Therefore we use accounts originated in 2010 as the post-crisis training sample. Its observation date is chosen so that the accounts in the two training samples have durations of similar lengths.

[15] That is, all the account at risk are pooled together regardless of the year of loan origination. The default rate at each month is calculated as the total number of defaults at that particular month over the total number of accounts at risk.

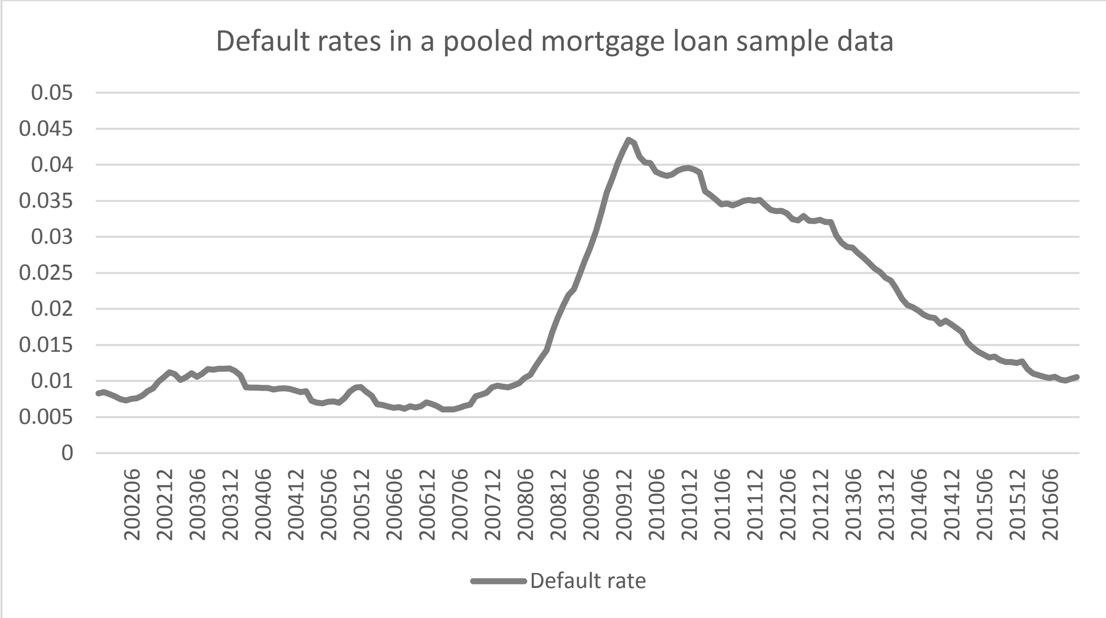

Fig. 5.3 Observed default rates in each month between Jan 2002 – Dec 2016 using 50000 accounts originated in each year between 1999 – 2016

5.4 Results and discussion

5.4.1 Frequentist and Bayesian estimation results

We estimate the discrete time hazard model on the training samples using the frequentist approach and the Bayesian approach with non-informative priors. Table 5.2 and Table 5.3 illustrate the estimation results based on training samples of tranquil and crisis time periods, respectively.

Table 5.2 Estimation Results using frequentist approach and Bayesian approach with non-informative priors based on the training sample during a tranquil period

	Training sample originated during 2010 with December 2012 as the observation date	
Variables	Frequentist Estimate (std.error)	Bayesian Posterior mean (std.dev)
Intercept	-17.8828393 *** (1.5044700)	-17.9747859 (1.4517682)
d_l_tbill_3m	3.4491426 * (1.9523948)	3.4715360 (1.9250224)
d_l_unemployment_rate	-0.4784589 (0.4866302)	-0.4780704 (0.4859698)
d_l_CPI	0.0304362 (0.1751539)	0.0326520 (0.1775234)
d_l_consumer_confidence	0.0059500 (0.0109468)	0.0059030 (0.0109317)
d_l_retail_sales	9.7222647 (14.2301412)	9.3395679 (14.0812482)
d_l_personal_earnings	-24.8523099 (17.5742103)	-24.9321081 (17.5094413)
d_l_IPI	0.2213348 (0.1597368)	0.2131014 (0.1612776)
d_l_dowjones_index	1.9529397 (1.6452997)	2.0458691 (1.6602687)
d_l_CS_houseprice_index	0.0981218 (0.1292146)	0.0928184 (0.1295023)
original_debt_to_income_ratio	0.0543894 *** (0.0062024)	0.0547142 (0.0062451)
original_loan_to_value	0.0109754 *** (0.0042073)	0.0110507 (0.0041958)
l_current_actual_upb	-0.0767203 (0.1017471)	-0.0769602 (0.0972451)
l_current_interest_rate	1.4724575 *** (0.1598832)	1.4754349 (0.1571467)
l_remaining_months	-0.0008054 (0.0010939)	-0.0007627 (0.0011046)
loan_age	0.0980407 *** (0.0340109)	0.1005526 (0.0338325)
loan_age_sq	-0.0020541 ** (0.0010050)	-0.0021122 (0.0009969)
	Log likelihood = -2644.404 Prob > chi2 = 0.0000	Number of draws in MCMC = 100000 Burn-in = 200000

Table 5.3 Estimation Results using frequentist approach and Bayesian approach with non-informative priors based on the training sample during a crisis period

	Training sample originated during 2007 with December 2009 as the observation date	
Variables	Frequentist Estimate (std.error)	Bayesian Posterior mean (std.dev)
Intercept	-23.6584688 *** (0.5081902)	-23.661662 (0.4916477)
d_l_tbill_3m	0.0317284 (0.0666831)	0.031562 (0.0655450)
d_l_unemployment_rate	0.1126435 (0.1369818)	0.116241 (0.1366128)
d_l_CPI	-0.0900573 *** (0.0213571)	-0.090223 (0.0213631)
d_l_consumer_confidence	0.0015748 (0.0029222)	0.001591 (0.0028842)
d_l_retail_sales	-2.3613185 (1.6943453)	-2.375036 (1.6982706)
d_l_personal_earnings	-6.4253837 *** (1.7700427)	-6.514378 (1.7815710)
d_l_IPI	-0.0610231 *** (0.0183981)	-0.061175 (0.0184346)
d_l_dowjones_index	-1.1240714 ** (0.4077305)	-1.106491 (0.3969147)
d_l_CS_houseprice_index	0.3142990 *** (0.0425264)	0.314196 (0.0417874)
original_debt_to_income_ratio	0.0232151 *** (0.0013464)	0.023267 (0.0013443)
original_loan_to_value	0.0161019 *** (0.0011818)	0.016113 (0.0012001)
l_current_actual_upb	0.5368529 *** (0.0313212)	0.536867 (0.0305348)
l_current_interest_rate	1.1721138 *** (0.0372275)	1.171451 (0.0368473)
l_remaining_months	0.0024178 *** (0.0004931)	0.002425 (0.0004999)
loan_age	0.1163065 *** (0.0120738)	0.116171 (0.0120432)
loan_age_sq	-0.0015617 *** (0.0003259)	-0.001559 (0.0003234)
	Log likelihood = -24854.8 Prob > chi2 = 0.0000	Number of draws in MCMC = 100000 Burn-in = 200000

In the frequentist estimation results based on the crisis time periods, more covariates are significant than in its tranquil counterparts. The consumer price, industrial production index, personal earnings, and stock price index are all significantly negative which is consistent with expectations since the better the economy is, the wealthier each borrower is on average. Therefore they are less likely to default. The house price index has a significant positive sign which may be explained by arguing that the more expensive properties are, the more difficult it is for homeowners to repay mortgage loans. The original debt to income ratio and loan to value ratio are both significantly positive since the higher the amount of debt compared to income and relative to the value of the property the less able and possibly less keen the borrowers are to repay. The current interest rate has a significantly positive impact on default since the higher the interest rates the greater the repayments are. The loan age variable and its squared term have significant positive and negative signs, respectively, representing a nonlinear relationship between default probability and time. In the frequentist estimation results based on the tranquil time periods, the covariates that are significant have the same signs as in the estimation based on the crisis time period. As for the Bayesian estimation results, since we use non-informative priors, the estimates when using the frequentist approach are close in both samples to those from the Bayesian approach.

Table 5.4 presents the Geweke diagnostic results for the Bayesian estimations based on the training samples of the tranquil and crisis time periods respectively. Initial draws from the first half of the MCMC are compared with draws from the second half. The z-scores for the coefficients are within the [-2, 2] range. As an example, Fig. 5.4 gives the trace and density plots of the marginal posterior distribution of one coefficient which demonstrate good mixing. The marginal posterior distributions of the other coefficients have similar characteristics. Therefore based on the Bayesian coefficients convergence diagnostics, the Markov chain converges well and the estimation results are reliable.

Table 5.4 Geweke diagnostic of convergence for Bayesian estimation based on training samples during tranquil and crisis time periods

Geweke diagnostic of convergence	z-scores	
Intercept	0.711684	0.2186
d_l_tbill_3m	-1.242724	-0.8681
d_l_unemployment_rate	-1.019658	1.3373
d_l_CPI	-0.510085	1.5070
d_l_consumer_confidence	0.918201	1.4735
d_l_retail_sales	0.107416	-1.0123
d_l_personal_earnings	-0.005511	-0.6370
d_l_IPI	-1.546842	0.1960
d_l_dowjones_index	-1.940418	-0.8504
d_l_CS_houseprice_index	-0.802466	-0.2536
original_debt_to_income_ratio	0.155242	1.8230
original_loan_to_value	0.667617	0.5878
l_current_actual_upb	-0.472571	-0.1459
l_current_interest_rate	-0.823602	-0.9056
l_remaining_months	0.774328	0.4999
loan_age	-0.451672	-1.7938
loan_age_sq	0.390750	1.7045

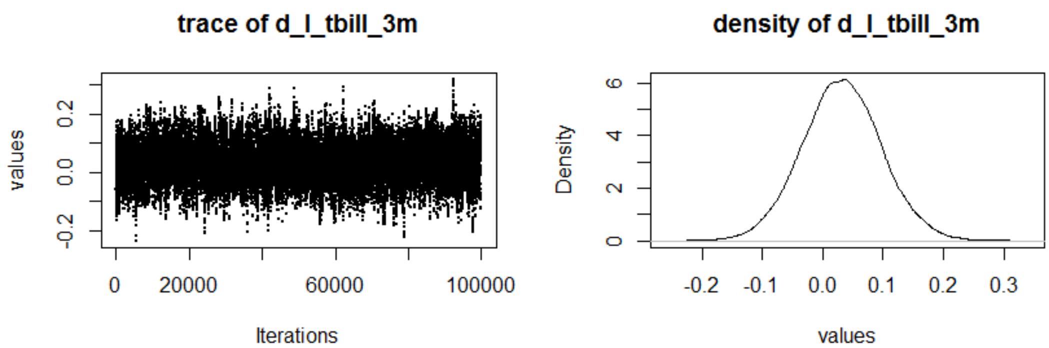

Fig. 5.4 Trace and density plots for the marginal posterior distribution of a coefficient based on the crisis sample

5.4.2 Frequentist and Bayesian model performances

Table 5.5 demonstrates the model performance results based on the training samples and test sample for the duration of the first 12 months for each account using the frequentist and Bayesian approaches with models built on the tranquil and crisis datasets. That is, we predict the probability an account will default in any of the first 12 months and use this and the observed instances of default over the same period to compute the predictive accuracies. The results from the test sample suggest the model trained on the tranquil time period is more accurate than that trained on the crisis time period. We consider this is firstly because the test sample is closer in default rate value to the tranquil training sample than to the crisis training sample, and secondly because the test sample is closer in time to the tranquil training sample than to the crisis training sample.

Table 5.5 Performance results using frequentist and Bayesian methods for the duration of the first 12 months based on the training samples during tranquil and crisis periods

Method	Sample	H-measure	GINI	AUC	AUCH	K-S
Frequentist tranquil	Train	0.2111560	0.5342676	0.7671338	0.7755307	0.4048441
	Test	0.1015387	0.3752977	0.6876489	0.6958642	0.2889237
Frequentist crisis	Train	0.1320903	0.4066508	0.7033254	0.7052240	0.2973461
	Test	0.0689591	0.2719505	0.6359752	0.6474961	0.2068731
Bayesian tranquil	Train	0.2118400	0.5343156	0.7671578	0.7759671	0.4071112
	Test	0.1014931	0.3749754	0.6874877	0.6956585	0.2876491
Bayesian crisis	Train	0.1320524	0.4066781	0.7033391	0.7052371	0.2974764
	Test	0.0691076	0.2719540	0.6359770	0.6476068	0.2072070

5.4.3 Frequentist and Bayesian stress testing

We apply our stress testing method to the test sample. We use June 2016 as t_s, the time that stress testing is applied to. For computational efficiency, we take a random sample of 50% of the 50000 accounts in the test data. We perform stress testing on all the accounts in this sample that live to June 2016.

To address the impact macroeconomic stress has on model parameters, models built on datasets of the crisis and tranquil periods are used for the stressed and normal scenarios respectively. We consider the normal scenario based on the tranquil period dataset and using the frequentist coefficient mean estimates without considering estimation risk as the baseline scenario.

To address estimation risk, we use the Bayesian stress testing approach and take 100 random draws from the Bayesian coefficient posterior distributions obtained from Bayesian credit risk models built on datasets of the crisis and tranquil periods. For the coefficients of the macroeconomic variables and the constant, we use draws from the posterior distribution. For the account specific variables, we use the Bayesian coefficient mean estimates. In the frequentist stress testing approach without addressing estimation risk, we use the frequentist coefficient mean estimates for all the variables.

To consider macroeconomic risk, we take random draws of the past values of the macroeconomic variables, before the stress testing time period, between Jan 1999 to May 2016. The values of all of the macroeconomic variables are drawn jointly and simultaneously from each time period. The simultaneity maintains the observed covariance structure between the variables. For computational efficiency, each month is taken once in the Bayesian stress testing. That is, each of the 209 vectors of macroeconomic values is combined with 100 draws from the posterior distribution. In the frequentist stress testing methods, to ensure the same number of scenarios among using all stress testing approaches for comparison purposes, 20900 random draws are taken, with replacements, of the vectors of macroeconomic values.

To avoid sampling bias, we apply bootstrapping for stress testing computations. For each of the four methods (frequentist and Bayesian for each of the crisis and tranquil periods), the stress testing procedure is repeated 100 times, each with random simulation for the samplings of macroeconomic scenarios, the Bayesian coefficient posterior distribution, and the error terms. We then collect all the estimated default rates obtained in the 100 computations to build the empirical simulated default rate distribution for each of the four stress testing methods.

Fig. 5.5 and 5.6 illustrate the histogram plots of the simulated default rate distributions using the frequentist estimation method based on the training samples during tranquil and crisis time periods.

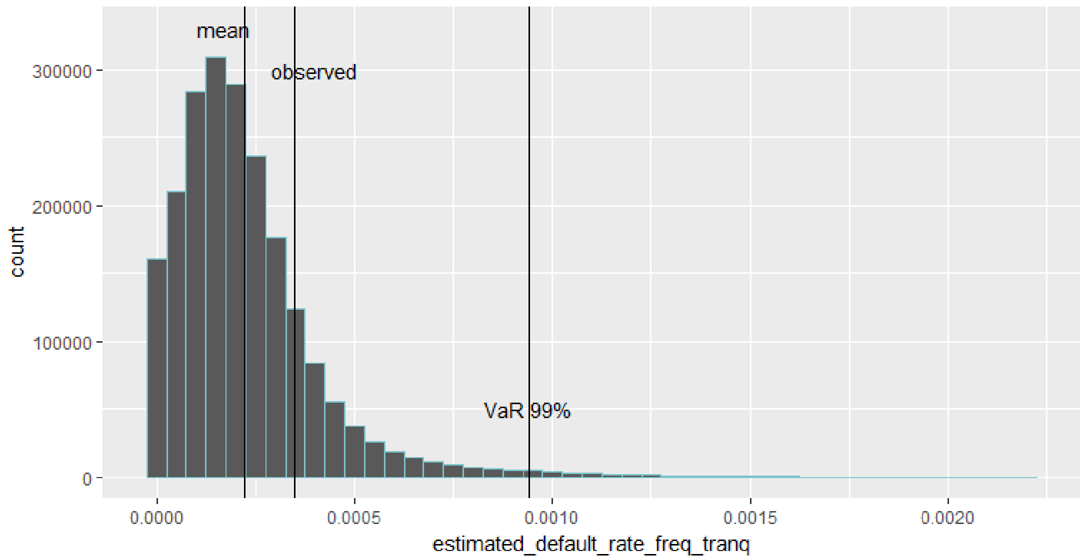

Fig. 5.5 Histogram plot of the simulated DR distribution obtained using the frequentist approach based on the training sample during a tranquil time period

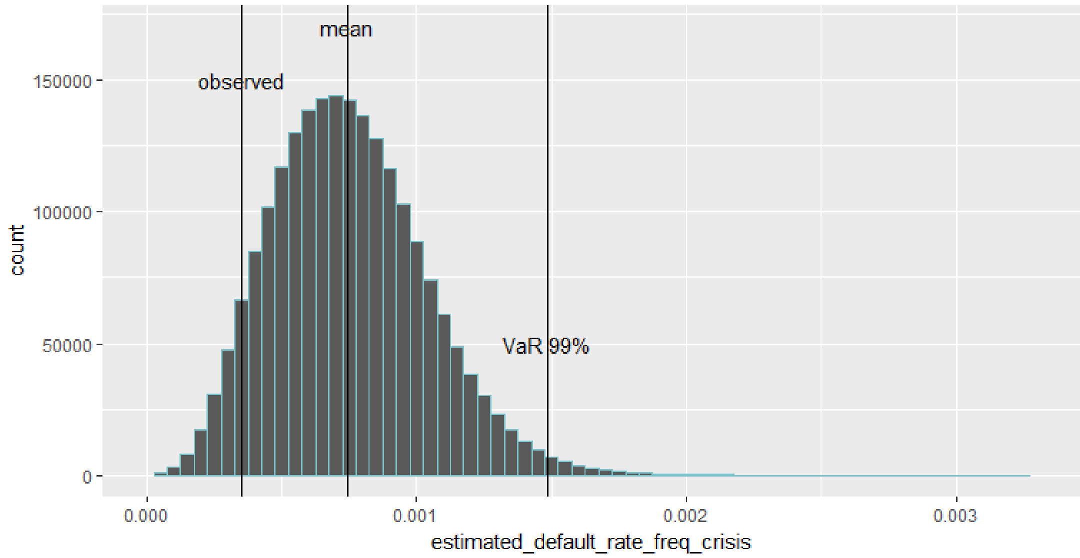

Fig. 5.6 Histogram plot of the simulated DR distribution obtained using the frequentist approach based on the training sample during a crisis time period

Fig. 5.7 and 5.8 show the histogram plots of the simulated default rate distributions using the Bayesian approach with random draws from the Bayesian posterior distributions based on the training samples during tranquil and crisis periods.

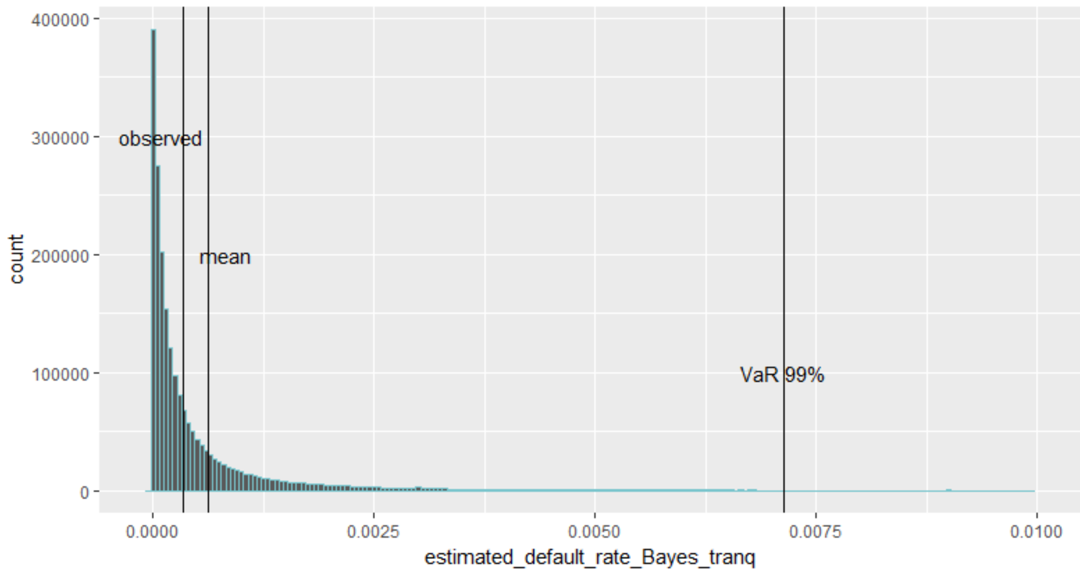

Fig. 5.7 Histogram plot of the simulated DR distribution obtained using the Bayesian posterior distribution approach based on the training sample during a tranquil time period

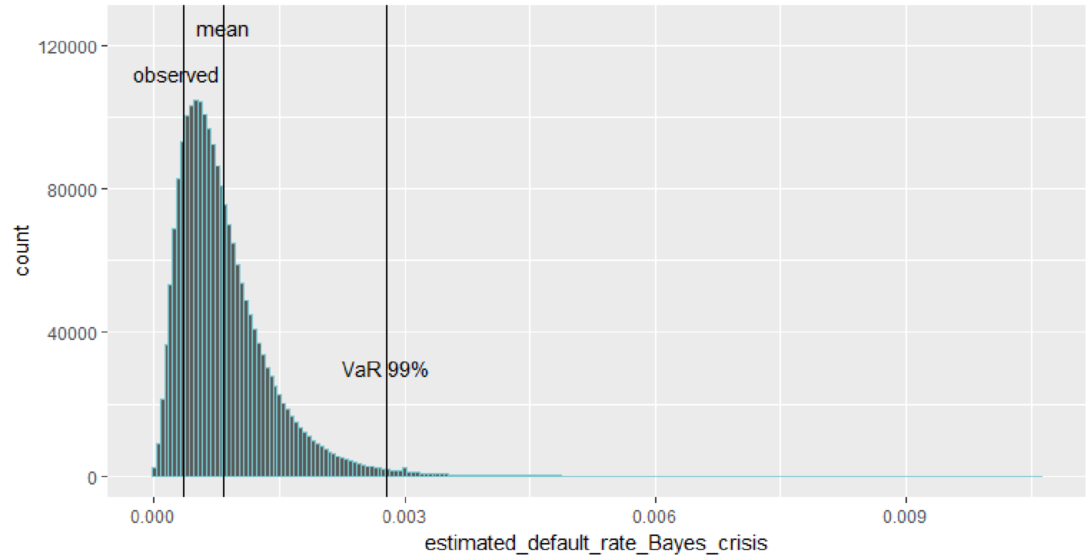

Fig. 5.8 Histogram plot of the simulated DR distribution obtained using the Bayesian posterior distribution approach based on the training sample during a crisis time period

Table 5.6 shows different VaRs, means, and standard deviations of the simulated DR distributions using the frequentist and Bayesian approaches based on models built on tranquil and crisis time period datasets as well as the observed default rate in the stress testing period June 2016 which we compare our stress testing results against.

Table 5.6 Statistics of the simulated default rate distributions compared with the observed default rate

statistics	Frequentist tranquil	Frequentist crisis	Bayesian tranquil	Bayesian crisis
95% VaR	0.00055	0.00124	0.00248	0.00188
99% VaR	0.00094	0.00149	0.00714	0.00278
Mean	0.00022	0.00074	0.00063	0.00083
St.d	0.00019	0.00028	0.00190	0.00056
Observed DR in June 2016	0.00035			

From Fig. 5.5 and Table 5.6, the observed default rate in the stress testing time period June 2016 (0.00035) is between the distribution mean (0.00022) and the 99% VaR (0.00094) of the simulated DR distribution using the frequentist approach based on tranquil time period (2010) dataset. From Fig. 5.6 and Table 5.6, the mean (0.00074) of the frequentist simulated DR distribution based on the training sample of the crisis period (2007) is more than twice the observed default rate in June 2016 (0.00035). The 99% VaR (0.00149) of the distribution is higher than 4 times the observed default rate in June 2016. From Fig. 5.7 and Table 5.6, the observed default rate in June 2016 (0.00035) is close to the distribution mean (0.00063) of the simulated DR distribution using the Bayesian approach with random draws from the posterior distribution based on the training sample during a tranquil period (2010). The 99% VaR of the distribution is more than 20 times the observed default rate in June 2016. From Fig. 5.8 and Table 5.6, the distribution mean (0.00083) of the simulated default rate distribution using the Bayesian approach with random draws from the parameter posterior distribution based on the training sample during a crisis time period (2007) is around 2.4 times the observed default rate in June 2016 (0.00035). The 99% VaR (0.00278) of the distribution is about 8 times the observed default rate. In summary, the 95% and 99% VaRs of the four stress testing methods all successfully include the observed default rate. Except for the baseline DR distribution, all DR distribution means are larger than the observed default rate.

Fig. 5.9 illustrates a comparison of the simulated DR distributions based on models built on crisis and tranquil time period datasets using the frequentist and Bayesian approaches. Fig. 5.10 presents the tails of the simulated DR distributions using the four approaches.

As expected, both parameter instability between stressed and normal scenarios and estimation risk cause the simulated default rate distributions to have larger VaRs, distribution means and standard deviations compared to the baseline scenario. Based on Table 5.6 and Fig. 5.9, out of all four simulated distributions, the baseline scenario has the smallest distribution statistics such as the VaRs, distribution mean and variance as it takes into account neither the macroeconomic shocks on model parameters nor estimation risk. The difference between the parameters of the models built on the crisis period and tranquil period cause the simulated DR distribution based on the crisis period to shift to the right with higher VaRs and distribution variance than the simulated DR distribution based on the tranquil period, both distributions obtained using the frequentist approach without estimation risk.

As including coefficient uncertainty is adding an additional source of variation, for stress testing within a given scenario (crisis or tranquil), the DR distributions with estimation risk included using the Bayesian stress testing method have larger distribution variances and VaRs than when estimation risk is neglected using the frequentist parameter mean estimates, as is shown in Table 5.6. This result is consistent with Omlin and Reichert (1999) who found that including estimation uncertainty leads to a wider prediction confidence interval than neglecting it. It can be seen from Fig. 5.9, at midrange default rates, the simulated DR distribution has higher frequencies using the mean estimate approach than when using the Bayesian distribution approach. At other default rates, the reverse is true. From Fig. 5.10, given the same scenario, the simulated DR distributions have fatter and longer tails when using the Bayesian posterior distribution approach compared to when using the frequentist approach.

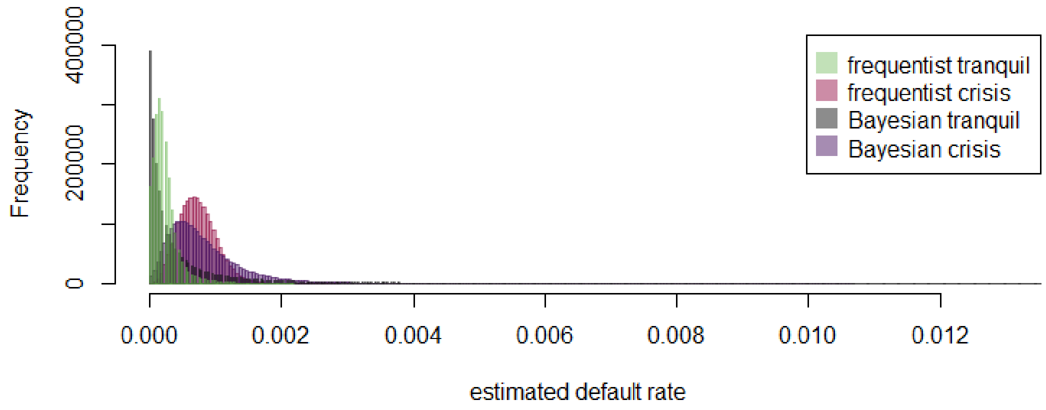

Fig. 5.9 Simulated default rate distributions using the frequentist and Bayesian approaches trained on crisis and tranquil datasets

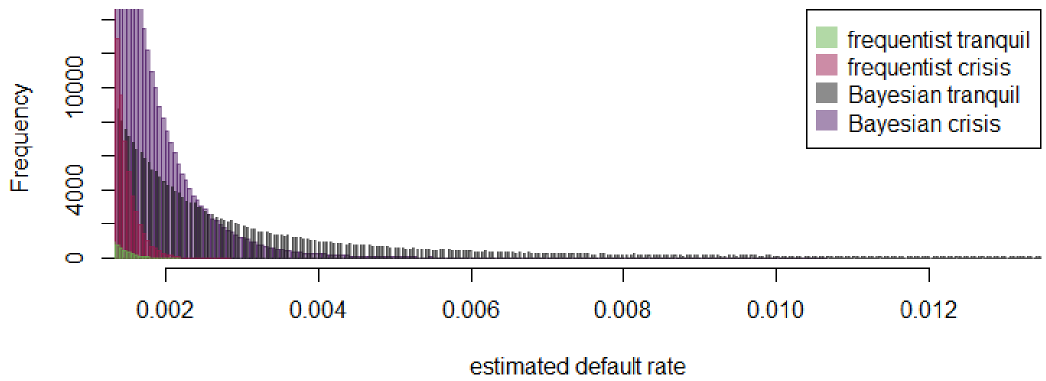

Fig. 5.10 Tails of the simulated default rate distributions using the frequentist and Bayesian approaches trained on crisis and tranquil datasets

The difference between the DR distributions simulated using the frequentist method based on the tranquil time period (VaR: 0.00094; St.d: 0.00019) and the DR distribution simulated using the Bayesian method based on the tranquil time period (VaR: 0.00714; St.d: 0.00190) is larger than the difference between the former (VaR: 0.00094; St.d: 0.00019) and the DR distribution simulated using the frequentist method based on the crisis time period (VaR: 0.00149; St.d: 0.00028). Similarly, the difference between the DR distribution simulated using the frequentist method based on the crisis time period (VaR: 0.00149; St.d: 0.00028) and the DR distribution simulated using the Bayesian method based on the crisis time period (VaR: 0.00278; St.d: 0.00056) is larger than the difference between the former (VaR: 0.00149; St.d: 0.00028) and the DR distribution simulated using the frequentist method based on the tranquil time period (VaR: 0.00094; St.d: 0.00019). These results suggest that estimation risks, for models built on either the crisis or tranquil time period dataset, have a larger influence than the parameter instability has on the simulated DR distribution.

The inclusion of estimation risk has a larger influence on DR distribution changes in stress testing based on a tranquil training sample (VaR difference: 0.00620; St.d difference: 0.00171) than in stress testing based on a crisis training sample (VaR difference: 0.00129; St.d difference: 0.00028). We consider this is because the posterior distribution obtained from the Bayesian estimation based on the crisis period has a narrow dispersion, with small standard deviations compared to mean estimates, whereas the posterior distribution obtained from the training sample of the tranquil period has large standard deviations compared to mean estimates. Since we use random draws from the posterior distributions in the Bayesian stress testing exercises, a small standard deviation compared to the posterior mean estimate implies that a great proportion of the sample draws from the posterior are not far from the mean estimates. In other words, the estimation risk is smaller in this case. Therefore the difference of simulated DR distributions between using the frequentist and Bayesian approaches is small for stress testing models built on the crisis period. This result is consistent with Omlin and Reichert (1999) who found that a wider posterior distribution leads to a wider prediction confidence interval.

Table 5.7 demonstrates a comparison of the 99% VaRs and distribution means based on models built on crisis and tranquil time period datasets using the frequentist and the Bayesian approaches.

Table 5.7 the 99% VaRs and distribution means of the simulated default rate distributions based on the training samples of crisis and tranquil periods using the frequentist and Bayesian approaches

	Mean	99% VaR
frequentist tranquil	0.00022	0.00094
frequentist crisis	0.00074	0.00149
Bayesian tranquil	0.00063	0.00714
Bayesian crisis	0.00083	0.00278

From Table 5.7, the simulated DR distribution using parameters of a model built on the crisis period and using the Bayesian approach has the largest distribution mean, taking into consideration both estimation risk and the impact macroeconomic stress has on model parameters. The simulated DR distribution obtained using the Bayesian approach and estimated on the tranquil time period has the largest VaR.

Table 5.8 demonstrates the monetary values of the 99% VaRs assuming (a) the exposure at default being the total current unpaid balance ($ 152.4 billion) of accounts originated in 2014 and alive in June 2016 in the Freddie Mac database population, (b) probability of default being the 99% VaRs obtained based on models built on crisis and tranquil period sample datasets using the frequentist and the Bayesian approaches, and (c) loss given default being 100%. It also demonstrates the expected losses assuming the probability of default being the distribution means based on models built on crisis and tranquil period samples using the frequentist and the Bayesian approaches, and loss given default being 100%. The capital required is the unexpected loss which is the difference between the monetary value of 99% VaR and the expected loss.

Table 5.8 The monetary values of the 99% VaRs, expected loss and required capital based on training samples of crisis and tranquil periods using the frequentist and Bayesian approaches

	Expected loss	Monetary value of 99% VaR	Required capital
frequentist tranquil	$ 33.5 million	$ 143.3 million	$ 109.8 million
frequentist crisis	$ 112.8 million	$ 227.1 million	$ 114.3 million
Bayesian tranquil	$ 90.0 million	$ 1088.1 million	$ 998.1 million
Bayesian crisis	$ 126.5 million	$ 423.7 million	$ 297.2 million

To compensate expected losses, financial institutions charge interest on loans. The higher the expected loss, the higher the interest rates. The amount of required capital maintained is determined by the unexpected losses. The higher the unexpected loss, the more capital banks need to keep.

Based on Table 5.8, the expected loss in the baseline scenario which only considers changes in macroeconomic values while ignoring both estimation risk and the effect the crisis has on model parameters is about $ 30 million. The monetary value of 99% VaR in the baseline scenario is about $ 140 million. The required capital in the baseline scenario that ignores both estimation risk and parameter instability is about $ 110 million. The expected loss is around $ 130 million when macroeconomic stress, the effect the crisis had on model parameters, and estimation risk are all addressed. The loss based on 99% VaR is around $ 420 million and the required capital is around $ 300 million. That is, the required capital is underestimated by nearly $ 200 million and the expected loss is underestimated by over $ 90 million. These results show that neglect of model parameter changes between scenarios and coefficient estimation uncertainty may considerably underestimate credit loss and capital required. Financial institutions would undercharge interest and keep insufficient capital in this case as the stress scenario is not extreme enough with only one source of risk being addressed.

Out of the four approaches, when 3 sources of risks (changes in the macroeconomic variables, impact of the crisis on model parameters, and estimation uncertainty when modelling on a crisis period) are taken into account, the DR distribution has the highest expected loss of nearly $ 130 million and the second highest required capital of around $ 300 million. The expected loss obtained using a Bayesian method and estimated on the tranquil period dataset is $ 90 million, and the required capital is close to $ 1 billion, which is the highest among the four approaches. These results are all much larger than those obtained from the baseline distribution. Our stress testing results show that including either or both parameter instability and estimation risk increase expected loss and required capital above those values implied by the baseline scenario.

5.5 Conclusion

The problem of insufficient bank capital manifested itself during the financial crisis. Consequently, banks had to enhance their capital base, and governments of most of the major economic entities had to bail out banks to improve their solvency and liquidity and stabilise the economy, such as the America's Capital Purchase Program and the UK's bank rescue package. As the financial crisis has induced huge amounts of economic and social costs, the important role of more prudential and comprehensive stress tests to provide adequate required capital estimates has been highlighted. Stress tests need to be severe and plausible. The fact that many banks seemed well capitalised and passed their respective capital requirements but still had insufficient capital to counter unexpected stress shows that the stress tests that banks and regulators implemented may not be adequately severe to give sufficient required capital estimates. Therefore, to better cope with potential losses caused by extreme events, the required capital needs to be reasonably and systematically increased.

Many reasons were proposed to explain the phenomenon of inadequate capital during the crisis. For instance, the Basel III package suggested that insufficient minimum capital ratio requirements, the definition of capital being too wide, high leverage, procyclical amplification of financial shocks, inadequate liquidity requirements were some of the main reasons. This research proposed another possibility which is that the types of risks being considered in the stress tests were too narrow. Conventionally most banks and regulators consider risk types such as credit risk, operational risk, and market risk in deciding bank capital. In recent years, especially after the 2008 financial crisis, the importance of different kinds of model risks has been increasingly discussed. However, the uncertainty of model parameters such as parameter estimation risk and parameter structural breaks between scenarios has not been well studied in the stress testing literature. Therefore in this research we incorporate parameter uncertainty in stress tests to give more prudential and conservative estimates of predicted losses and capital required.

In the conventional stress tests, the model parameters were treated as fixed values. Modellers vary the values of the covariates using different scenarios to provide predicts of losses and subsequently required capital in different situations. For stress tests constructed after the financial crisis, this is still the case. In contrast, in our stress testing procedure, we not only account for the uncertainty of the covariates, but also incorporate the uncertainty of the model parameters, including estimation uncertainty and instability between scenarios. As more sources of risks are included in the stress tests, higher estimates of losses are provided, hence systematically giving more conservative estimates of required capital. In other words, the stress tests using our method with more types of risks included can better predict and cope with extreme losses, such as those caused by the financial crisis, than the conventional method which addresses only the uncertainty of the variables. It shows that banks would be safer against losses with higher estimates of required capital obtained if more sources of uncertainty are considered.

In this chapter, we examined the influence that parameter instability between models built in the financial crisis and post-crisis tranquil time periods has on stress testing results. Using a discrete time hazard model, with both the frequentist and Bayesian approaches, we estimate and stress test account default rates based on U.S. mortgage loan data. The contribution of this research is two-fold. Firstly we study the impact that differences in parameters between stress and normal scenarios have on default forecasts in a Bayesian framework. Secondly, we incorporate estimation risk, in addition to parameter instability, into stress testing by using the Bayesian parameter posterior distribution, instead of point estimates, in the Bayesian stress testing framework.

Our main finding is that differences in parameter estimates between credit risk models built on the crisis time period data and those built on the non-crisis time period data cause the simulated DR distribution based on the former to have larger Values at Risk, distribution mean and distribution variances than those built on the latter. The inclusion of estimation risk also has the same effect. When both shocks to model parameters and coefficient uncertainty are included in stress testing, the 99% VaR obtained is around 3 times the 99% VaR of the baseline simulated DR distribution which takes neither model parameter changes when the economy moves to a crisis scenario nor parameter estimation risk into account.

We also calculate the expected losses and required capital based on the total current unpaid balance at the time period stress testing is applied to, the distribution means and VaRs of the simulated DR distributions using the frequentist and Bayesian methods with models built on crisis and tranquil datasets, and assuming loss given default is 100%. We find that the required capital computed when using our method that considers macroeconomic risk, estimation risk, and parameter instability is about 170% more than using the conventional stress testing method that only considers macroeconomic risk.

The results in this paper suggest that if a stress testing exercise only considers changes in macroeconomic variables but ignores estimation uncertainty as well as parameter changes between scenarios, it may considerably underestimate credit losses, interest rates that banks should charge; and the amount of capital banks would keep based on the stress testing result will be insufficient. Therefore financial institutions may consider these risks in their stress testing models to prevent credit loss underestimation.

Chapter 6
Conclusion

6.1 Summary

Stress testing is an area of considerable interest to academics, industry practitioners, and regulators, especially after the 2008 financial crisis. Since the Basel II Accord, banks are allowed to estimate credit risk parameters such as the probability of default using their own internal models. Banks that use an internal ratings based approach are required to stress test the risk parameters. The Value at Risk, unexpected loss, and required capital in a stressed scenario can be calculated through stress testing.

One gap in the stress testing literature is that papers only employ model coefficient point estimates thus neglecting the estimation error surrounding the point estimates. Therefore, one of the objectives of this thesis has been to include coefficient estimation risk in stress testing modelling. Another gap in the stress testing literature is that most papers only consider changes in macroeconomic variables while neglecting changes in model parameters between scenarios. The few papers that do address parameter instability between models built on different scenarios are mainly in the frequentist framework without considering coefficient estimation uncertainty. Therefore the second objective of this thesis has been to study model parameter changes between stressed and normal scenarios while also incorporating model coefficient estimation uncertainty into the stress testing model.

Researchers frequently try to find ways to enhance the classification accuracy of credit scoring models and predict loan defaults better. In the literature, improving model predictive accuracy is often achieved by adding more useful information into the modelling process, such as additional covariates and better data samples. However, the possibility of adding available information through using Bayesian informative priors to improve the performance of PD models has not been fully explored. The use of Bayesian estimation in PD modelling is also relatively underdeveloped compared to the frequentist estimation. Therefore the third objective of this thesis has been to improve model predictive accuracy by including available useful information through the use of Bayesian informative priors.

In Chapter 1 we provide the context for this thesis, such as the necessity yet lack of bank capital during the financial crisis and thus the importance of stress tests for banks to preserve enough capital during stressed periods, the guidelines of the Basel Accords regarding stress testing and capital requirements, the macro stress tests conducted in practice, and the lack of addressing coefficient estimation risk in the stress tests in use. It then gives an overview of this thesis outlining its research motivations, aims, contributions, methods, and findings, etc.

In Chapter 2 we present a literature review on PD models and stress testing. We review the state of the art classification algorithms for the probability of default modelling and various model performance measures in the literature of credit scoring. On comparing model performances of the most popular individual and ensemble classifiers, we find that support vector machines, neural networks, and logistic regression are the best performing models. Ensemble classifiers generally have good model performances. In reviewing stress testing literature, we summarise the stress testing methods into a 3-step procedure starting with a scenarios building stage, then a stress test modelling stage and finally the outcome stage. We also categorise existing papers into macro and micro levels, using either distribution simulation or point forecast approaches and with either balance sheet elements or probability of default as credit risk indicators.

This thesis has three main chapters following the introduction and literature review chapters. The three main chapters make several contributions. In Chapter 3 we construct an innovative Bayesian informative prior selection method and improve the predictive performance of the credit scoring model through adding available useful information by using the Bayesian informative priors obtained from our Bayesian informative prior selection method.

In this method, we treat a model parameter in the frequentist credit scoring models built on consecutive time periods as a time series variable and forecast its value in a future time period using an ARIMA model. We then use these ARIMA forecasts of all the model parameters as informative priors in Bayesian credit scoring models. We use various frequentist and Bayesian models and train them on sample data of pre-, during, and post-crisis time periods to compare their model predictive accuracies with the models that use the prior selection method we propose.

Our main conclusion of chapter 3 is that the Bayesian models with our method of prior selection using ARIMA forecasts of coefficients as informative priors outperform all other frequentist and Bayesian models used in this chapter based on all training samples in terms of model performance, regardless of the time periods or the economic environments on which the models are built. This result shows that by using a new informative prior selection method we included additional useful available information. Consequently, by reducing the risk of neglecting useful information we improved model performance of the PD models.

In Chapter 4, to reduce coefficient estimation risk in stress testing exercises, we contribute a new stress testing method that employs the Bayesian coefficient posterior distribution instead of point estimate values as the source of coefficients. Since only the mean estimates are used in the conventional stress testing methods, the estimation errors of the coefficient estimates are not addressed. In contrast, in our method we include full ranges of possible values of the coefficients through the use of the coefficient posterior distribution, hence incorporating estimation errors. As an additional source of risk, i.e. estimation risk, in addition to macroeconomic stress, is incorporated into stress testing, we can obtain more conservative estimates of the predicted loss compared to when only shocks to macroeconomic covariates are addressed.

We use a discrete time hazard model based on a panel dataset to obtain the coefficient estimates. We use both the frequentist method and the Bayesian estimation methods to obtain the frequentist coefficient point estimates and the Bayesian posterior distribution. We use a latent variable interpretation of the logistic regression as the stress testing model to predict the default rate at the stress testing time period. The coefficient posterior distribution obtained in the Bayesian approach and the point estimates obtained in the frequentist approach are applied to the stress testing models to simulate the Bayesian and frequentist estimated DR distributions. In the default rate distribution simulation in the Bayesian approach, we not only simulate the macroeconomic scenarios but also simulate from the coefficient posterior distribution to include other possible coefficient estimates so we do not ignore the estimation errors surrounding the coefficient mean estimates. We then compare the Bayesian simulated DR distribution with the frequentist simulated DR distribution.

Our main conclusion of chapter 4 is that compared to using conventional stress testing methods, if we use the stress testing method that we propose, more extreme predictions of losses are simulated since an extra source of risk, estimation risk, is included. The monetary value of loss calculated based on the 99% VaR when estimation risk is included is about 6.5 times as much as that when estimation risk is neglected. Therefore using our stress testing method, financial institutions need to maintain higher levels of capital and so would be safer against stressed conditions.

Our main contribution in Chapter 5 is that we incorporate parameter instability in models built on stressed and tranquil scenarios into a Bayesian stress testing methodology that considers coefficient estimation uncertainty. The use of Bayesian estimation and a Bayesian coefficient posterior distribution as the source of coefficient estimates in our method allows for the introduction of estimation risk in a stress testing approach that addresses parameter structural breaks between scenarios.

We model the probability of default and simulate the estimated default rate distributions individually for the financial crisis and post-crisis tranquil time periods to study the influence of coefficient changes on stress testing results. With the stress testing method proposed in Chapter 4, we use both frequentist coefficient point estimates and Bayesian coefficient posterior distribution in default rate simulation to study the impact estimation risk has on stress testing results given the same scenario.

Our main conclusion of Chapter 5 is that the influence of macroeconomic stress on model parameters causes the simulated default rate distributions to have higher VaRs and variances. The required capital obtained from the simulated DR distribution that considers a combined influence of changes in covariate values, changes in model parameters between scenarios, and estimation risk is approximately 170% higher than the required capital when only changes in covariates values are considered. Therefore, if a financial institution's stress testing model neglects parameter instability and coefficient uncertainty, the required capital could be severely underestimated and would be insufficient to absorb losses in a stressed condition. Therefore it is essential to address these sources of variations, including parameter instability and estimation risk, in addition to macroeconomic shocks, in stress testing methods.

6.2 Limitations

In this section, we address the limitations of this thesis. Firstly, our research mainly concentrates on the generic methodological contributions that we consider would apply to various data, variables, and models. Therefore fewer efforts were made on looking for the most suitable data, variables, etc. to apply our methods to. For instance, the set of explanatory variables is important in modelling the probability of default and the accuracy of default forecasting. The covariates chosen in this study are the conventionally commonly used variables based on the literature and data availability. We use macroeconomic variables, microeconomic variables including application variables, behavioural variables and duration variables. However, no further attempt was made to collect other potentially important covariates such as the Meso-level variables, which are rarely used in the stress testing literature and could potentially be a contribution.

Another limitation is that we concentrate on modelling the probability of default while assuming LGD and EAD, which are beyond the scope of this research, as constants. Since we focus on PD in modelling loss distributions, we also assume independence between the three risk parameters without considering the impact of LGD and EAD on PD.

Another limitation is that in the probability of default modelling of mortgage loans the data is unbalanced as the number of defaulted accounts are normally much less than the non-default accounts, which is a common problem in modelling credit default. When classes are imbalanced, there may not be sufficient patterns belonging to the minority class to adequately represent its distribution. On the other hand, our samples are large with ample numbers of cases belonging to both the majority and minority classes.

6.3 Future study

In this section, we discuss several potential topics for future study.

1. Apply our methods to other models. For instance, apply the Bayesian informative prior selection method that we proposed to other models, apart

from the logistic regression employed in Chapter 3, to improve their model performances, such as the extreme value models.

Since in our thesis, our main goal is to explore the use of Bayesian parameters in improving model performance and reducing estimation uncertainty, we apply our prior selection and stress testing methods to models that are most commonly used in the literature and the industry. Due to its good model interpretability and high predictive accuracy, logistic regression is one of the most commonly used credit scoring models in practice and in the literature. Therefore, we apply our prior selection method to a logistic regression model to improve model performance. In chapters 4 and 5, we apply our stress testing method to discrete time hazard models to address parameter uncertainty risk.

As a potential topic for future research, our Bayesian informative prior selection method using ARIMA forecasts of coefficient variables as priors and our stress testing method that addresses coefficient estimation risk can be extended to models other than logistic regression, and to topics and areas other than PD modelling as long as there are model coefficients involved. To give more details, we show that with our prior selection method, the predictive accuracy of our PD model is improved. This method should also be of use for other statistical PD classifiers. For instance, as described in the limitation section, loan defaults are rare events compared to non-defaults. If the link function is symmetric, the predicted probability of default approaches zero and one at the same rate, which implies that characteristics of events and non-events are assigned with the same importance. Consequently, binary response models with symmetric link functions, such as the logistic regression and the probit model, may underestimate the probability of rare events (Calabrese & Giudici, 2015). Extreme value models are a type of models that addresses the rarity of bank failures. As the link function for extreme value models is asymmetric which lets the predicted PD approach zero faster than it approaches one, they assign more importance to the characteristics of rare events and concentrate estimation efforts on the tail of the default distribution. Therefore, one possible route of future research is to use our prior selection and stress testing methodology on models that can better cope with rare events, such as the extreme value models.

Secondly, the U.S is a geographically large country with numerous states and cities. The degree of wealth and development of the numerous states and cities at which the accounts are situated is also diverse. Therefore, the accounts in different areas in the country may be considerably different. To improve the predictive accuracy of credit scoring models, we may include additional information such as using additional explanatory variables or try alternative models. Random effects models are a type of model where some of the parameters are random variables. A random effects model can help control for unobserved heterogeneity uncorrelated with the explanatory variables. Therefore, apart from collecting Meso-level variables, another possibility is to try our stress testing and prior selection methods on other models such as a random effects model with a city specific or individual specific random effect to better address the variability among accounts in different areas.

Furthermore, it would be an interesting extension to model the dependence structure between the credit failure probabilities over time and between units of different characteristics, such as accounts in different areas. Therefore, another potential route of future research is to account for the dependence between probabilities of default in a stress testing methodology that addresses estimation uncertainty. The copula approach is especially appropriate in modelling various types of dependence structure (Smith, 2015; Calabrese et al. 2017). For instance, the D-vine copula has been employed to model the dependence structure between statistical units based on longitudinal data over time in several geographic regions (Kim, Kim, Liao, & Jung, 2013; Smith, 2015). Therefore, we may incorporate a copula approach in a Bayesian stress testing framework to reduce estimation risk and at the same time better address the cross-sectional and serial dependence between default probabilities.

2. Use informative priors in stress testing to include additional useful information

Secondly, in our research we use non-informative priors in stress testing to ensure the differences between the loss distributions using the Bayesian posterior distribution approach and the frequentist point estimate approach are due to the inclusion of estimation risk. That is we single out the influence of estimation risk while controlling for other influences on the simulated distributions. As the Bayesian coefficient posterior distribution tends to move towards the informative prior distribution from the likelihood distribution if informative priors are used, the loss distribution obtained using informative priors will undoubtedly shift from the loss distribution obtained using non-informative priors. We have also shown in Chapter 3 that good informative priors can improve model performance. Therefore one potential development is to use informative priors on top of including estimation risk in stress testing to study their combined influence on stress testing results. Another potential study is to single out the influence of informative priors while using Bayesian point estimates to exclude estimation risk. Then we can compare the loss distributions obtained using these methods with the one obtained using the frequentist point estimates to see the impact of just additional information or a combination of additional information and estimation risk on predicted loan loss. The informative priors can be elicited using the prior selection method proposed in Chapter 3 using ARIMA forecasts as priors.

3. Use a unified model that incorporates parameter instability between stressed and tranquil scenarios in a Bayesian framework. Use a Bayesian time series model to search for informative priors.

In Chapter 5 we model the probability of default individually for models built on stressed and tranquil time periods to obtain two separate sets of model parameters.

We then substitute the two sets of parameters in the stress testing models individually to simulate two loss distributions and to study the influence of parameter changes on predicted loss. One possible future study is to use a unified model that includes both sets of model parameters for the stressed and tranquil scenarios. Consequently, we can form a unified simulated loss distribution that on the one hand addresses parameter instability while on the other hand produces a single distribution instead of two.

In chapter 3, we use frequentist ARIMA models to identify the patterns of model coefficients over time and use the forecasts of their future values as informative priors for the coefficients in the Bayesian PD models. As a potentially more unified and coherent way of prior selection, an alternative approach we consider worth exploring is to use Bayesian inference for time series models to fit the data of PD model coefficients, and use Bayesian forecasts of these coefficients as the informative priors for the Bayesian PD models. Furthermore, the use of Bayesian analysis of time series models as opposed to using the classical frequentist analysis also enables the use of Bayesian computations, Bayesian model averaging, and the inclusion of prior information, etc. (Smith, 2015; Steel, 2010).

4. Include more types of uncertainty, other than estimation risk and model parameter instability, into the stress testing method

In this research, we mainly focus on the effect of coefficient estimation uncertainty and coefficient instability between scenarios on stress testing. That is, we concentrate on the relationship between the model coefficients and the loss distribution. A potential topic of future research could be the effect of model variable selection uncertainty on stress testing. We use logistic regression as our stress testing model. The possibility of using our stress testing methodology, which includes coefficient uncertainty risk, but applied to other models, such as machine learning classifiers, could also be explored.

6.4 Policy Implications

This research gives policy implications for practitioners and regulators. Firstly, this work provides an innovative perspective to the banks as to improving model performance. In the past, a lot of efforts were made in developing machine learning techniques and in combining models, which could enhance model performance but may themselves have disadvantages such as the black box characteristics and would be difficult to explain to customers. This research suggests that banks can make better predictions for future defaults through recognising and utilising hidden patterns in the available information using statistical methods. Banks should make better use of any available information to find hidden patterns in the information. The hidden pattern in the information can then be used to make better predictions for future defaults. This research suggests that one way of recognising hidden patterns is through the application of time series models to the coefficient variables, and one way of utilising the hidden patterns is through the use of Bayesian informative priors. Therefore, this research also sheds new light on the application of Bayesian methods to credit scoring.

Secondly, this work gives one additional and possible reason that banks did not have sufficient capital during the financial crisis: the stress testing models in use did not address estimation uncertainty. Moreover, this work not only points out the problem but also provides a stress testing model that includes estimation risk and provides more conservative estimates of credit loss and required capital. It strongly suggests that it is essential to address estimation risk in stress testing since neglecting it could considerably underestimate credit loss. Our work also provides new insight into the application of the Bayesian approach for stress testing. Since in the Bayesian approach model coefficients are treated as random variables instead of fixed values, with the use of a Bayesian approach, we accommodate uncertainty in coefficient estimates in the stress testing model apart from uncertainty in covariate values. This work shows that it is important to take more types of uncertainty into account in stress tests as higher losses are predicted and more capital is required using this method. Since more capital can absorb more loss, banks and depositors can, therefore, be safer against stress.

Thirdly, we suggest that for different scenarios, the stress testing models, specifically model coefficients explored in this research, should also be different. Usually in the literature, the same sets of model parameters are used for models built on the stressed and non-stressed scenarios. We show with empirical evidence in Chapter 5 that parameter instability between scenarios makes a considerable impact on loss distributions and should be accounted for in the stress testing models that banks use in practice. Our research strongly supports the inclusion of model differences between different scenarios in stress testing.

References

Abdou, H., Pointon, J., & El-Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert systems with applications, 35*(3), 1275-1292.

Abdou, H. A. (2009). Genetic programming for credit scoring: The case of Egyptian public sector banks. *Expert systems with applications, 36*(9), 11402-11417.

Abedini, M., Ahmadzadeh, F., & Noorossana, R. (2016). Customer credit scoring using a hybrid data mining approach. *Kybernetes, 45*(10), 1576-1588.

Abellán, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert systems with applications, 73*, 1-10.

Abellán, J., & Mantas, C. J. (2014). Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. *Expert systems with applications, 41*(8), 3825-3830.

Akkoç, S. (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research, 222*(1), 168-178.

Ala'raj, M., & Abbod, M. F. (2016a). Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems, 104*, 89-105.

Ala'raj, M., & Abbod, M. F. (2016b). A new hybrid ensemble credit scoring model based on classifiers consensus system approach. *Expert systems with applications, 64*, 36-55.

Altman, E. I., Sabato, G., & Wilson, N. (2010). The value of non-financial information in small and medium-sized enterprise risk management. *The Journal of Credit Risk, 6*(2), 95-127.

Bae, J. K., & Kim, J. (2015). A personal credit rating prediction model using data mining in smart ubiquitous environments. *International Journal of Distributed Sensor Networks, 11*(9), 1-6.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), 627-635.

Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. *Expert systems with applications, 42*(19), 6609-6619.

Bangia, A., Diebold, F. X., Kronimus, A., Schagen, C., & Schuermann, T. (2002). Ratings migration and the business cycle, with application to credit portfolio stress testing. *Journal of banking & finance, 26*(2-3), 445-474.

Bank of England (2017). *Stress testing the UK banking system: 2017 results*. London: Bank of England.

Bank of England (2018). *Financial Stability Report*. London: Bank of England.

Basel Committee on Banking Supervision (1988). *International convergence of capital measurement and capital standards*. Basel: The Bank for International Settlements.

Basel Committee on Banking Supervision (1996). *Amendment to the capital accord to incorporate market risks*. Basel: The Bank for International Settlements.

Basel Committee on Banking Supervision (2001). *Overview of The New Basel Capital Accord*. Basel: The Bank for International Settlements.

Basel Committee on Banking Supervision (2004). *Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework*. Basel: The Bank for International Settlements.

Basel Committee on Banking Supervision (2011). *Basel III: A global regulatory framework for more resilient banks and banking systems - revised version June 2011*. Basel: The Bank for International Settlements.

Basel Committee on Banking Supervision (2017). *Basel III: Finalising post-crisis reforms*. Basel: The Bank for International Settlements.

Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert systems with applications, 36*(2), 3302-3308.

Bellotti, T., & Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting, 29*(4), 563-574.

Bellotti, T., & Crook, J. (2014). Retail credit stress testing using a discrete hazard model with macroeconomic factors. *Journal of the Operational Research Society, 65*(3), 340-350.

Berkowitz, J. (1999). *A coherent framework for stress-testing* (FEDS Working Paper No. 99-29). Retrieved from http://dx.doi.org/10.2139/ssrn.181931

Bijak, K., & Matuszyk, A. (2017). *Bayesian models of car lease frauds*. Paper presented at the Credit Scoring and Credit Control XV, Edinburgh, United Kingdom.

Bijak, K., & Thomas, L. C. (2015). Modelling LGD for unsecured retail loans using Bayesian methods. *Journal of the Operational Research Society, 66*(2), 342-352.

Bikker, J. A., & Hu, H. (2002). Cyclical patterns in profits, provisioning and lending of banks and procyclicality of the new Basel capital requirements. *PSL Quarterly Review, 55*(221), 143-175.

Board of Governors of the Federal Reserve System (2017). *Comprehensive Capital Analysis and Review 2017: Assessment Framework and Results*. Washington DC: Federal Reserve.

Board of Governors of the Federal Reserve System (2018). *Comprehensive Capital Analysis and Review 2018: Assessment Framework and Results*. Washington DC: Federal Reserve.

Borio, C., Drehmann, M., & Tsatsaronis, K. (2014). Stress-testing macro stress testing: does it live up to expectations? *Journal of Financial Stability, 12*, 3-15.

Bravo, C., Thomas, L. C., & Weber, R. (2015). Improving credit scoring by differentiating defaulter behaviour. *Journal of the Operational Research Society, 66*(5), 771-781.

Breeden, J. L. (2016). Incorporating lifecycle and environment in loan-level forecasts and stress tests. *European Journal of Operational Research, 255*(2), 649-658.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert systems with applications,* *39*(3), 3446-3453.

Busch, R., Koziol, P., & Mitrovic, M. (2018). Many a little makes a mickle: Stress testing small and medium-sized German banks. *The Quarterly Review of Economics and Finance, 68*, 237-253.

Calabrese, R., & Giudici, P. (2015). Estimating bank default with generalised extreme value regression models. *Journal of the Operational Research Society, 66*(11), 1783-1792.

Chen, W., Ma, C., & Ma, L. (2009). Mining the customer credit using hybrid support vector machine technique. *Expert systems with applications, 36*(4), 7611-7616.

Crone, S. F., & Finlay, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting, 28*(1), 224-238.

Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research, 173*(3), 781-800.

Crook, J., & Banasik, J. (2012). Forecasting and explaining aggregate consumer credit delinquency behaviour. *International Journal of Forecasting, 28*(1), 145-160.

Crook, J., Leow, M., & Bellotti, T. (2015). *A Framework of Credit Risk Stress Testing Methodologies*. Paper presented at the Stress Testing Conference, Edinburgh

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, 183*(3), 1447-1465.

de Melo Junior, L. S., Nardini, F. M., Renso, C., & de Macêdo, J. A. F. (2019). *An Empirical Comparison of Classification Algorithms for Imbalanced Credit Scoring Datasets.* Paper presented at the 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, Florida, USA.

Delgado, J., & Saurina, J. (2004). *Credit risk and loan loss provisions. An analysis with macroeconomic variables*. (Banco de Espana Working paper, No. 12). Madrid, Spain: Bank of Spain.

Dillon, W., & Goldstein, M. (1984). *Multivariate Analysis*. New York, NY: John Wiley & Sons.

Drehmann, M., Patton, A., & Sorensen, S. (2005). *Corporate defaults and large macroeconomic shocks*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.494.4570&rep=rep1&type=pdf

Dua, P., & Kapur, H. (2018). Macro stress testing and resilience assessment of Indian banking. *Journal of Policy Modeling, 40*(2), 452-475.

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

Escanciano, J. C., & Olmo, J. (2010). Backtesting parametric value-at-risk with estimation risk. *Journal of Business & Economic Statistics, 28*(1), 36-51.

European Banking Authority (2018). *2018 EU-wide Stress Test Results*. Paris, France: European Banking Authority.

Fallahpour, S., Lakvan, E. N., & Zadeh, M. H. (2017). Using an ensemble classifier based on sequential floating forward selection for financial distress prediction problem. *Journal of Retailing and Consumer Services, 34*, 159-167.

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research, 210*(2), 368-378.

Fitzpatrick, T., & Mues, C. (2016). An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market. *European Journal of Operational Research, 249*(2), 427-439.

Florez-Lopez, R., & Ramon-Jeronimo, J. M. (2015). Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal. *Expert systems with applications, 42*(13), 5737-5753.

Forrest, A. (2011). *Managing Model Risk in Practice*. Paper presented at the Credit Scoring and Credit Control XII, Edinburgh.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics, 19*(1), 1-67.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis, 38*(4), 367-378.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning, 29*(2-3), 131-163.

Giudici, P. (2001). Bayesian data mining, with application to benchmarking and credit scoring. *Applied Stochastic Models in Business and Industry, 17*(1), 69-81.

Gourieroux, C., & Zakoïan, J.-M. (2013). Estimation-adjusted var. *Econometric Theory, 29*(4), 735-770.

Greene, W. H. (2011). *Econometric Analysis*. Essex, England: Pearson Education.

Hahn, E. D., & Soyer, R. (2005). Probit and logit models: Differences in the multivariate realm. *The Journal of the Royal Statistical Society, Series B*, 1-12.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning, 77*(1), 103-123.

Harrell, F. E., & Lee, K. L. (1985). A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality. *Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences', North-Holland, New York, United States*, 333-343.

Henry, J., & Kok, C. (2013). *A macro stress testing framework for assessing systemic risks in the banking sector*. (ECB Occasional Paper No. 152). Frankfurt am Main, Germany: European Central Bank.

Hens, A. B., & Tiwari, M. K. (2012). Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method. *Expert systems with applications, 39*(8), 6774-6781.

Ho, T. K. (1995). *Random decision forests.* Paper presented at the 3rd international conference on document analysis and recognition, Montreal, Quebec, Canada.

Hoggarth, G., Sorensen, S., & Zicchino, L. (2005). *Stress tests of UK banks using a VAR approach* (Working Paper Series No. 282). London: Bank of England.

Hu, Y.-C., & Ansell, J. (2007). Measuring retail company performance using credit scoring techniques. *European Journal of Operational Research, 183*(3), 1595-1606.

Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications, 33*(4), 847-856.

Huang, G.-B., Chen, L., & Siew, C. K. (2006). Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. Neural Networks, 17*(4), 879-892.

Jacobs Jr, M. (2018). The validation of machine-learning models for the stress testing of credit risk. *Journal of Risk Management in Financial Institutions, 11*(3), 218-243.

Jacobs Jr, M., & Sensenbrenner, F. J. (2018). A comparison of methodologies in the stress testing of credit risk–alternative scenario and dependency constructs. *Quantitative Finance and Economics, 2 (2)*, 294-324.

Jacobs, M., Karagozoglu, A. K., & Sensenbrenner, F. (2015). Stress testing and model validation: application of the Bayesian approach to a credit risk portfolio. *Journal of Risk Model Validation, 9*(3), 41-70.

Jacobs, M., & Kiefer, N. M. (2010). *The Bayesian approach to default risk: a guide* (CAE Working Paper, No. 10-01). Retrieved from https://www.econstor.eu/bitstream/10419/70446/1/634836307.pdf

Janes, H., & Pepe, M. S. (2008). Matching in studies of classification accuracy: implications for analysis, efficiency, and assessment of incremental value. *Biometrics, 64*(1), 1-9.

Jokivuolle, E., & Viren, M. (2013). Cyclical default and recovery in stress testing loan losses. *Journal of Financial Stability, 9*(1), 139-149.

Ju, Y., Jeon, S. Y., & Sohn, S. Y. (2015). Behavioral technology credit scoring model with time-dependent covariates for stress test. *European Journal of Operational Research, 242*(3), 910-919.

Kanas, A., & Molyneux, P. (2018). Macro stress testing the US banking system. *Journal of International Financial Markets, Institutions and Money, 54*, 204-227.

Kanno, M. (2015). Macro stress test for credit risk. *The Journal of Risk Finance, 16*(5), 554-574.

Kapinos, P., & Mitnik, O. A. (2016). A top-down approach to stress-testing banks. *Journal of Financial Services Research, 49*(2-3), 229-264.

Kapinos, P. S., Martin, C., & Mitnik, O. A. (2018). Stress testing banks: Whence and whither? *Journal of Financial Perspectives, 5*(1), 1-20.

Kiefer, N. M. (2009). Default estimation for low-default portfolios. *Journal of Empirical Finance, 16*(1), 164-173.

Kim, D., Kim, J.-M., Liao, S.-M., & Jung, Y.-S. (2013). Mixture of D-vine copulas for modeling dependence. *Computational statistics & data analysis, 64*, 1-19.

Kim, H. S., & Sohn, S. Y. (2010). Support vector machines for default prediction of SMEs based on technology credit. *European Journal of Operational Research, 201*(3), 838-846.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis, 9*(2), 137-163.

Ko, A. H. R., Sabourin, R., & Britto Jr, A. S. (2008). From dynamic classifier selection to dynamic ensemble selection. *Pattern recognition, 41*(5), 1718-1731.

Koop, G., Poirier, D. J., & Tobias, J. L. (2007). *Bayesian econometric methods*: Cambridge University Press.

Koutanaei, F. N., Sajedi, H., & Khanbabaei, M. (2015). A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services, 27*, 11-23.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert systems with applications, 40*(13), 5125-5131.

Kruppa, J., Ziegler, A., & König, I. R. (2012). Risk estimation and risk prediction using machine-learning methods. *Human genetics, 131*(10), 1639-1654.

Laeven, L., & Majnoni, G. (2003). Loan loss provisioning and economic slowdowns: too much, too late? *Journal of financial intermediation, 12*(2), 178-197.

Lee, T.-S., & Chen, I. F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert systems with applications, 28*(4), 743-752.

Lee, T.-S., Chiu, C.-C., Chou, Y.-C., & Lu, C.-J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational statistics & data analysis, 50*(4), 1113-1130.

Leong, C. K. (2016). Credit risk scoring with bayesian network models. *Computational Economics, 47*(3), 423-446.

Leow, M., & Crook, J. (2016). The stability of survival model parameter estimates for predicting the probability of default: Empirical evidence over the credit crisis. *European Journal of Operational Research, 249*(2), 457-464.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124-136.

Li, J., Wei, L., Li, G., & Xu, W. (2011). An evolution strategy-based multiple kernels multi-criteria programming approach: The case of credit decision making. *Decision Support Systems, 51*(2), 292-298.

Li, K., Niskanen, J., Kolehmainen, M., & Niskanen, M. (2016). Financial innovation: Credit default hybrid model for SME lending. *Expert systems with applications, 61*, 343-355.

Li, S.-T., Shiue, W., & Huang, M.-H. (2006). The evaluation of consumer loans using support vector machines. *Expert systems with applications, 30*(4), 772-782.

Li, S., Tsang, I. W., & Chaudhari, N. S. (2012). Relevance vector machine based infinite decision agent ensemble learning for credit risk analysis. *Expert systems with applications, 39*(5), 4947-4953.

Liang, D., Tsai, C.-F., & Wu, H.-T. (2015). The effect of feature selection on financial distress prediction. *Knowledge-Based Systems, 73*, 289-297.

Louzis, D. P. (2017). Macroeconomic and credit forecasts during the Greek crisis using Bayesian VARs. *Empirical Economics, 53*(2), 569-598.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York, NY, USA: Springer Science & Business Media.

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert systems with applications, 42*(10), 4621-4631.

Malhotra, R., & Malhotra, D. K. (2003). Evaluating consumer loans using neural networks. *Omega, 31*(2), 83-96.

Maltritz, D., & Molchanov, A. (2008). *Economic determinants of country credit risk: a Bayesian approach.* Paper presented at the 12th New Zealand Finance Colloquium, Palmerston North, New Zealand.

Marqués, A. I., García, V., & Sánchez, J. S. (2012a). Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert systems with applications, 39*(11), 10244-10250.

Marqués, A. I., García, V., & Sánchez, J. S. (2012b). Two-level classifier ensembles for credit risk assessment. *Expert systems with applications, 39*(12), 10916-10922.

Miguéis, V. L., Benoit, D. F., & Van den Poel, D. (2013). Enhanced decision support in credit scoring using Bayesian binary quantile regression. *Journal of the Operational Research Society, 64*(9), 1374-1383.

Mira, A., & Tenconi, P. (2004). Bayesian estimate of credit risk via MCMC with delayed rejection. *Stochastic Analysis, Random Fields and Applications IV*, 277-291.

Misina, M., Tessier, D., & Dey, S. (2006). *Stress testing the corporate loans portfolio of the Canadian banking sector* (Working paper No. 47). Ottawa: Bank of Canada.

Mujalli, R. O., López, G., & Garach, L. (2016). Bayes classifiers for imbalanced traffic accidents datasets. *Accident Analysis & Prevention, 88*, 37-51.

Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert systems with applications, 36*(2), 3028-3033.

Omlin, M., & Reichert, P. (1999). A comparison of techniques for the estimation of model prediction uncertainty. *Ecological modelling, 115*(1), 45-59.

Paleologo, G., Elisseeff, A., & Antonini, G. (2010). Subagging for credit scoring models. *European Journal of Operational Research, 201*(2), 490-499.

Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert systems with applications, 36*(1), 2-17.

Papadopoulos, G., Papadopoulos, S., & Sager, T. (2016). *Credit risk stress testing for EU15 banks: a model combination approach* (Working paper No. 203). Athens:

Bank of Greece.

Park, I., Amarchinta, H. K., & Grandhi, R. V. (2010). A Bayesian approach for quantification of model uncertainty. *Reliability Engineering & System Safety, 95*(7), 777-785.

Partalas, I., Tsoumakas, G., & Vlahavas, I. (2009). Pruning an ensemble of classifiers via reinforcement learning. *Neurocomputing, 72*(7-9), 1900-1909.

Partalas, I., Tsoumakas, G., & Vlahavas, I. (2010). An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. *Machine learning, 81*(3), 257-282.

Pesaran, M. H., Schuermann, T., Treutler, B.-J., & Weiner, S. M. (2006). Macroeconomic dynamics and credit risk: a global perspective. *Journal of Money, Credit and Banking, 38*(5), 1211-1261.

Petropoulos, A., Siakoulis, V., Mylonas, D., Klamargias, A., Alexakis, P. D., Samantas, I. G., . . . Thomaidou, E. (2018). *A combined statistical framework for forecasting default rates of Greek Financial Institutions' credit portfolios* (Working paper No. 243). Athens: Bank of Greece.

Philippon, T., Pessarossi, P., & Camara, B. (2017). *Backtesting european stress tests* (NBER Working Paper No. 23083). Cambridge, MA: National Bureau of Economic Research.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine, 6*(3), 21-45.

Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence, 28*(10), 1619-1630.

Rösch, D., & Scheule, H. (2004). Forecasting retail portfolio credit risk. *The Journal of Risk Finance, 5*(2), 16-32.

Salas, V., & Saurina, J. (2002). Credit risk in two institutional regimes: Spanish commercial and savings banks. *Journal of Financial Services Research, 22*(3), 203-224.

Saritas, M. M., & Yasar, A. (2019). Performance Analysis of ANN and Naive Bayes

Classification Algorithm for Data Classification. *International Journal of Intelligent Systems and Applications in Engineering, 7*(2), 88-91.

Sawilowsky, S. S. (2003). You think you've got trivials? *Journal of Modern Applied Statistical Methods, 2*(1), 218-225.

Schechtman, R., & Gaglianone, W. P. (2012). Macro stress testing of credit risk focused on the tails. *Journal of Financial Stability, 8*(3), 174-192.

Schuermann, T. (2014). Stress testing banks. *International Journal of Forecasting, 30*(3), 717-728.

Siddiqi, N. (2012). *Credit risk scorecards: developing and implementing intelligent credit scoring* (Vol. 3). New York: John Wiley & Sons.

Siemsen, T., & Vilsmeier, J. (2018). *On a quest for robustness: About model risk, randomness and discretion in credit risk stress tests* (Discussion Paper No. 31/2018). Frankfurt, Germany: Deutsche Bundesbank.

Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of educational statistics, 18*(2), 155-195.

Sinha, A. P., & May, J. H. (2004). Evaluating and tuning predictive data mining models using receiver operating characteristic curves. *Journal of Management Information Systems, 21*(3), 249-280.

Sinha, A. P., & Zhao, H. (2008). Incorporating domain knowledge into data mining classifiers: An application in indirect lending. *Decision Support Systems, 46*(1), 287-299.

Smith, M. S. (2015). Copula modelling of dependence in multivariate time series. *International Journal of Forecasting, 31*(3), 815-833.

Sohn, S. Y., Kim, D. H., & Yoon, J. H. (2016). Technology credit scoring model with fuzzy logistic regression. *Applied Soft Computing, 43*, 150-158.

Sorge, M., & Virolainen, K. (2006). A comparative analysis of macro stress-testing methodologies with application to Finland. *Journal of financial stability, 2*(2), 113-151.

Starnes, D. S., Moore, D. S., Yates, D., & Tabor, J. (2014). *The practice of statistics*. New

York: Macmillan Higher Education.

Steel, M. F. J. (2010). Bayesian time series analysis. In *Macroeconometrics and time series analysis* (pp. 35-45). Berlin, Germany: Springer.

Šušteršič, M., Mramor, D., & Zupan, J. (2009). Consumer credit scoring models with limited data. *Expert systems with applications, 36*(3), 4736-4744.

Thomas, L., Crook, J., & Edelman, D. (2017). *Credit scoring and its applications* (2 ed.). Philadelphia: SIAM.

Tinoco, M. H., & Wilson, N. (2013). Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis, 30*, 394-419.

Tomczak, J. M., & Zięba, M. (2015). Classification restricted Boltzmann machine for comprehensible credit scoring model. *Expert systems with applications, 42*(4), 1789-1796.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge, England: Cambridge university press.

Tsai, C.-F. (2014). Combining cluster analysis with classifier ensembles to predict financial distress. *Information Fusion, 16*, 46-58.

Tsai, C.-F., & Hung, C. (2014). Modeling credit scoring using neural network ensembles. *Kybernetes, 43*(7), 1114-1123.

Tsai, C.-F., & Wu, J.-W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert systems with applications, 34*(4), 2639-2649.

Tsai, M.-C., Lin, S.-P., Cheng, C.-C., & Lin, Y.-P. (2009). The consumer loan default predicting model–An application of DEA–DA and neural network. *Expert systems with applications, 36*(9), 11682-11690.

Tsukahara, F. Y., Kimura, H., Sobreiro, V. A., & Zambrano, J. C. A. (2016). Validation of default probability models: A stress testing approach. *International Review of Financial Analysis, 47*, 70-85.

Twala, B. (2010). Multiple classifier application to credit risk assessment. *Expert systems with applications, 37*(4), 3326-3336.

van den End, J. W. (2006). *Indicator and boundaries of financial stability* (No. 097). Amsterdam, Netherlands: Netherlands Central Bank.

Van Sang, H., Nam, N. H., & Nhan, N. D. (2016). A novel credit scoring prediction model based on Feature Selection approach and parallel random forest. *Indian Journal of Science and Technology, 9*(20), 1-6.

Vazquez, F., Tabak, B. M., & Souto, M. (2012). A macro stress test model of credit risk for the Brazilian banking sector. *Journal of Financial Stability, 8*(2), 69-83.

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications, 38*(1), 223-230.

Wang, G., & Ma, J. (2012). A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine. *Expert systems with applications, 39*(5), 5325-5331.

West, D., Dellana, S., & Qian, J. (2005). Neural network ensemble strategies for financial decision applications. *Computers & operations research, 32*(10), 2543-2559.

Wilson, N., & Altanlar, A. (2014). Company failure prediction with limited information: newly incorporated companies. *Journal of the Operational Research Society, 65*(2), 252-264.

Wong, J., Choi, K.-f., & Fong, T. P.-W. (2008). A framework for stress testing banks' credit risk. In *The Banking Sector in Hong Kong* (pp. 240-260). London: Palgrave Macmillan.

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert systems with applications, 78*, 225-241.

Xiao, W., Zhao, Q., & Fei, Q. (2006). A comparative study of data mining methods in consumer loans credit scoring management. *Journal of Systems Science and Systems Engineering, 15*(4), 419-435.

Yao, P., & Lu, Y. (2011). Neighborhood rough set and SVM based hybrid credit scoring classifier. *Expert systems with applications, 38*(9), 11300-11304.

Yap, B. W., Ong, S. H., & Husain, N. H. M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert systems with applications, 38*(10), 13274-13283.

Yu, L., Wang, S., & Lai, K. K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. *Expert systems with applications, 34*(2), 1434-1444.

Yu, L., Wang, S., & Lai, K. K. (2009). An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: The case of credit scoring. *European Journal of Operational Research, 195*(3), 942-959.

Yu, L., Yao, X., Wang, S., & Lai, K. K. (2011). Credit risk evaluation using a weighted least squares SVM classifier with design of experiment for parameter selection. *Expert systems with applications, 38*(12), 15392-15399.

Yu, L., Yue, W., Wang, S., & Lai, K. K. (2010). Support vector machine based multiagent ensemble learning for credit risk evaluation. *Expert systems with applications, 37*(2), 1351-1360.

Zhang, C., Liu, C., Zhang, X., & Almpanidis, G. (2017). An up-to-date comparison of state-of-the-art classification algorithms. *Expert systems with applications, 82*, 128-150.

Zhang, D., Zhou, X., Leung, S. C. H., & Zheng, J. (2010). Vertical bagging decision trees model for credit scoring. *Expert systems with applications, 37*(12), 7838-7843.

Zhang, Y., Jia, H., Diao, Y., Hai, M., & Li, H. (2016). Research on credit scoring by fusing social media information in online peer-to-peer lending. *Procedia Computer Science, 91*, 168-174.

Zhou, L., Lai, K. K., & Yu, L. (2010). Least squares support vector machines ensemble models for credit scoring. *Expert systems with applications, 37*(1), 127-133.

Zhu, Y., Xie, C., Wang, G.-J., & Yan, X.-G. (2017). Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. *Neural Computing and Applications, 28*(1), 41-50.