MONASH University

Essays on Financial Modeling and Forecasting

A thesis submitted for the degree of

Doctor of Philosophy

by

Hong Wang

Supervisors:
Assoc. Prof. Catherine Forbes
Dr. Bonsoo Koo

Department of Econometrics and Business Statistics
Monash University

January 31, 2019

Copyright notice

Abstract

Understanding the dynamic mechanisms of some key financial and economic quantities plays a central role in an array of decision making processes such as risk management, portfolio allocation, and more generally managerial planning in response to macroeconomic forecasts. This thesis develops Bayesian inferential methodologies for dynamic hierarchically specified models relevant to certain empirical economic and financial settings. The new models are flexible and are therefore able to accommodate non-standard distributional shapes as well as nonlinear relationships between variables. Bayesian inference is obtained by sampling from the relevant posterior densities using Markov Chain Monte Carlo (MCMC) simulation techniques. In addition, the thesis also develops a novel portfolio optimization method for high-dimensional portfolio selection problems.

The thesis, then, is largely based around three distinct chapters that each contributes to different aspects of the economic and financial modeling and forecasting literatures. The first paper is concerned with modelling bank loan recovery data that has, marginally, a non-standard distributional shape, along with a collection of potential high-dimensional recovery determinants. The distributional features are accounted for using a Gaussian mixture model along with a hierarchical regression model structure coupled with a prior. In addition, a Markov-switching mechanism is incorporated as a proxy for a latent credit cycle, helping to explain differences in observed recovery rates over time. Utilizing data extracted from Moody's Ultimate Recovery Database, we are able to demonstrate how the probability of default and certain loan-specific and other variables hold different explanatory power with respect to recovery rates over ‘good’ and ‘bad’ states in the credit cycle.

The second paper proposes an extension to the Vector Auto-Regressive (VAR) model introduced into empirical economics by Sims (1980). The model is adapted to suit multivariate time series data observed with mixed frequencies, i.e., when certain variables observed only at low frequency (e.g. quarterly) while others are observed at high frequency (e.g. monthly), as well as unknown form of nonlinearity present between variables. The nonlinearities between variables are accommodated through a Gaussian process prior for the unknown function. Bayesian analysis of the model, which is specified in a state-space form, is amenable to the use of MCMC methods. Utilizing the framework of Stroud, Müller, and Polson (2003), an auxiliary mixture model is introduced to facilitate the MCMC sampling. Conditional on a vector of latent indicator variables, the auxiliary mixture model reduces to a linear Gaussian state space model, and the efficient block sampling algorithms of Carter and Kohn (1994) and Frühwirth-Schnatter (1994) is employed to jointly update all unobserved states.

The third paper develops a new framework for determining portfolios with stable out-of-sample performance in the presence of a large universe of underlying assets. The proposed approach builds upon taking the advantage of both subset resampling Shen and Wang (2017) and parameter regularization Fan, Zhang, and Yu (2012) within a unified framework. By exploiting a hierarchical clustering algorithm, subsets of assets are randomly sampled while having controlled maximum correlation. These subsets are used as regularization targets when constructing subportfolios which are then averaged to stabilize the final portfolio weights. We show that the resulting portfolio strategy compares favorably with state-of-the-art strategies across a range of different covariance structures in a simulation setup. The usefulness of the proposed method is also illustrated using the Fama-French industrial portfolios and large U.S. stock

market benchmarks.

Declaration of Authorship

I, Hong Wang, declare that this thesis titled, "Essays on Financial Modeling and Forecasting" and the work presented in it are my own. I confirm that:

This thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where reference is made in the text of the thesis.

Signature: Hong

Print Name: HONG WANG

Date: 14/09/2018

Acknowledgements

My journey traveling down the road to completion of a Ph.D. degree is a safari, and to this, I am grateful. To each and every one of my trip leaders and fellow travellers I am greatly indebted. A safari adventurer can not survive on his own, and my dissertation could not have been completed without help of many individuals. I would like to thank my supervisors Associate Professor Catherine Forbes and Dr Bonsoo Koo, for their guidance, patience and support throughout my Ph.D. studies. Your meticulous attitude has shaped my understanding of research; your generous approach builds my attitude towards learning; your philosophy and principle of how to grapple with difficulties coming from all parts of academic life have perpetually influenced my personal values and beliefs. Hereby, I would like to express my highest appreciation to their help, support and enlightenment.

Among many other people who helped me at Monash, I am especially thankful to Professor Param Silvapulle and Associate Professor Xibin Zhang, for encouraging me pursue a Ph.D degree upon my completion of the Master of Applied Econometrics. Thank you, Professor Gael Martin. Your remarkable skill with language, presentation, and collaboration have enlightened me on the role of communication. My wholehearted thanks go to my thesis panel members, Professor Farshid Vahid and Professor Param Silvapulle, for sacrificing time to painstakingly read the various drafts, and attend the milestone presentations. Besides, I appreciate the warm help from many EBSers, Professor Mervyn Silvapulle and Professor Donald Poskitt, for their intuitive teaching in the well-instructed first year Ph.D coursework units. Thank you, Professor

Colin O'Hare - for lack of a more meaningful phrase. I wish I could have talked to you more when I had the chance. You were not only a mentor guiding me through my first research project but also a liberal friend with an affable and humorous personality. I would also like to thank my dearest fellow Ph.D students and research fellows, Zhichao Liu, Rongju Zhang, Melvern Leung, Manh Cuong Pham, Kanchana Nadarajah, Yiru Wang, Alex Cooper, Lina Zhang, Yuejun Zhao, Patrick Leung, Chuhui Li, Ou Yang, Bin Jiang, for many inspiring discussions and interesting conversations we had together.

Last, but not least, I would like to thank my parents, Xiao Zhou and Lianlong Wang and my parents in law, Rong Wang and Dahe Wang for your unconditional love, patience and support. Especially to my mother: I hope I will eventually converge to your expectation although the variance is not controlled yet. I reserve my last gratitude to my wife, Jiayi Wang, for being with me over the ups and downs of my life so far. I truly look forward to the life we are going to spend together.

Contents

List of Figures

List of Tables

"In God we trust; all others must bring data."

William Edwards Deming

List of Abbreviations

AIC	**A**kaike **I**nformation **C**riterion
BIC	**B**ayesian **I**nformation **C**riterion
DGP	**D**ata **G**enerating **P**rocess
VAR	**V**ector **A**uto-**R**egression
MIDAS	**MI**xed-**DA**ta **S**ampling
GP	**G**aussian **P**rocess
MCMC	**M**arkov **C**hain **M**onte **C**arlo
MH	**M**etropolis-**H**astings
LASSO	**L**east **A**bsolute **S**hrinkage and **S**election **O**perator
GDP	**G**ross **D**omestic **P**roduct
S&P	**S**tandard & **P**oor
CV	**C**ross-**V**adlidation
OLS	**O**rdinary **L**east **S**quares
RR	**R**ecovery **R**ate
PD	**P**robability of **D**efault
GFC	**G**lobal **F**inancial **C**risis

List of Symbols

$\propto$	proportional to
$\lvert \cdot \rvert$	absolute value
$\lvert\lvert \cdot \rvert\rvert_p$	ℓ_p-norm
$\sum$	sum
$\prod$	product of numbers
$[a, b]$	closed interval bounded by a,b
(a, b)	open interval bounded by a,b
$\rightarrow$	converge
$\overset{ind}{\sim}$	independently distributed
$\overset{P}{\rightarrow}$	converge in probability
$\mathbb{E}(\cdot)$	expectation
$\mathbb{E}(\cdot \mid x)$	conditional expectation (on variable x)
$P(\cdot)$	probability
$P(\cdot \mid A)$	conditional probability (on event A)
p.d.f	probability density function
$m \vee n$	minimum of m and n
$\in$	set membership
i.i.d	independently and identically distributed
$\mathbb{N}$	set of natural number
$\mathcal{N}(\mu, \sigma^2)$	Gaussian distribution with mean μ and variance σ^2
A$\otimes$B	Kronecker product (of matrices A and B)
A$\circ$B	Hadamard product (of matrices A and B)

For Jackie

Chapter 1

Introduction

1.1 Background

Understanding the dynamic mechanisms of some key financial and economic quantities plays a central role in an array of decision making processes such as risk management, portfolio allocation, and more generally managerial planning in response to macroeconomic forecasts. Recently, a considerable amount of economic and financial literature has placed an emphasis on the validation of economic and financial theory where an empirical model's performance is assessed predominantly on its out-of-sample forecast performance. The well known adage that "simple, parsimonious models tend to be the best for out-of-sample forecasting..." (Diebold, 1998) suggests the notion that a mis-specified model can have better predictive ability than an imprecisely estimated larger encompassing model, but correctly specified, due to the bias-variance trade-off.

Regarding the development of new econometric forecasting models, one notable strand of research that has spurred a large literature and has had significant implication for forecasting price distributions involves the notion of "volatility clustering". Often described as "large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes" (Mandelbrot, 1963), the statistical modelling of this identified feature of financial price returns distributions was first explored in the seminal paper (Engle, 1982) on autoregressive conditional heteroscedastic models (ARCH).

The ARCH model, along with its close relative Generalised ARCH (GARCH), introduced by Bollerslev (1986), is a strong example of how an econometric modelling approach can provide a characterization of an important feature of observed data, leading to the delivery of improved forecasts.

The Bayesian inferential approach, often implemented via Markov chain Monte Carlo (MCMC) simulation techniques, has arguably become the primary approach for dealing with complicated model structures, particularly within a dynamic hierarchical framework. Believes about certain static parameters or the evolution of dynamic factors that impact on the observed process may be incorporated through a carefully constructed prior distribution.

Concurrent with the advancement of the use of Bayesian inference in applied settings is the growth of penalized regression methods. These methods, that promote the use of less elaborated, or 'sparse', models, trade-off bias and variance with the specific aim of improving out-of-sample forecast performance. Moreover, penalized regression (also known as 'shrinkage') methods have been shown to deliver predictive gains in high-dimensional settings, where the number of variables is of similar, or greater size than that of the sample size available.

In general, modeling and forecasting are two major tasks used by practitioners and scholars of economics and finance. The former might typically involve the use of a combination of modern economic and finance theory and expert judgment to build a mathematical representation relevant to data observed in some empirical setting.Often it is a conceptual model which is then used for multiple purposes - story telling, policy experiments and forecasting. The latter attempts to go further, exploiting statistical regularities in the available data to produce predictions of future outcomes that are then used as management tools to guide policy, manage risks and for other programmatic decision making. As a consequence of recent innovations in strict modeling assumptions previous assumed may now be re-assessed or relaxed in light of available modern statistical and econometric tools.

Amongst all, there have been enormous advances in Bayesian modeling and computational technologies useful for economic and financial data. In particular, there have been advances in MCMC techniques for hierarchical modeling strategies known as a state space model, particularly suitable for time series data. Through the use of a hierarchical structure, the various features apparent in observed data may be attributed to different model components, and can provide forecast distributions that accommodate, among other things, latent dynamics, parameter uncertainty, and model uncertainty. Another modeling advances including the development of Gaussian process where an entire functional form may be treated as known, such a model fits neatly into a hierarchical modeling framework, offering expanded opportunities to relax previous held assumptions.

More recently, research relating to high-dimensional statistics and their use in economic and financial settings has gained recognition. Amongst all, the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) that performs automated variable shrinkage and selection in a regression context where a large number of potential predictor variables is available, has featured heavily in relevant theoretical and empirical research. It is interesting to consider the success of the LASSO from a frequentist perspective, as it trades off estimation bias and variance as it forces some parameter estimates to shrunk to zero. This feature is particularly in line with the forecasting literature's emphasis on the so-called principle of parsimony. Arguably, the LASSO is amongst the most influential recent developments in statistical methodology, and has implications for both modelling and forecasting.

In keeping with the above reflections, the focus of this thesis is twofold. It first considers the development of modeling architectures that help to empirically characterize the main features of the distributional shapes, dynamic patterns and other dependence present in the observed data. Two chapters, namely Chapter 3 and 4, each details a hierarchically structured time series model incorporating a modern prior – the Bayesian LASSO prior in Chapter 3 and a Gaussian process prior in Chapter 4 – both designed to overcome distinct problems in a real-world empirical problem. These models are flexible and contain latent variables within their hierarchical structures. Inference for these models relies upon Bayesian machinery and takes full advantages of MCMC simulation techniques.

In addition to the modeling and methodological issues addressed, this thesis also investigates key empirical questions. We investigate the connection between bank loan recovery rate determinants and changing general economic environment using a dataset extracted from Moody's Ultimate Recovery Database. The recovery rate is naturally restricted to lie over the unit interval with large concentrations of observed rates at or just near each the boundary. Moreover, there is a large set of potential recovery determinants available, and the literature suggests that only a few of these seems to be relevant, depending on the state of economy, see Khieu, Mullineaux, and Yi (2012) and Hu and Perraudin (2002). We utilize an idea from Altman and Kalotay (2014) and build a hierarchical model that builds upon a finite Gaussian mixture model. This is able to capture the clustering behavior in the observed recovery rates, by attaching a latent ordered probit regression that links the recovery determinants to each of the Gaussian mixture components. In addition, and to account for systemically time-varying changes in the recovery rate distribution, the final model considered allows for time-varying coefficients that dependents upon the prevailing state of the "credit cycle". Given the state of the credit cycle, the relevance of each recovery determinant is examined through the use of the Bayesian LASSO proposed by Park and Casella (2008). The resulting model is relatively parsimonious in the sense that only a few recovery determinants are found to play a key role in explaining the observed recovery rate distribution.

Next, we develop a semi-parametric Vector Auto-Regressive model for data sampled at different frequencies. Mixed frequency data is often encountered in macroeconomic studies. For example, researchers are often concerned with so-called "Taylor rule", which is essentially a monetary policy reaction function that describes the nominal interest rate set by the central bank in reaction to the inflation rate and to the gross domestic product (GDP). For most countries, GDP measures are sampled at most quarterly, while higher frequency measures for other quantities such as the interest rate and inflation are readily available. Rather than dropping observations and modeling the joint process sampled at a common low frequency, a flexible VAR model is developed to accommodate the mixed frequency data in this framework we are also able to accommodate nonlinear co-movement between variables through a Gaussian process prior. This allows researchers to investigate the functional form of the monetary policy reaction function instead of relying on the restrictive linearity assumption.

A second contribution is made with regard to the development of modeling strategy for the challenging and high-dimensional portfolio setting (Markowitz, 1952). The nature of the problem inherent in a given dataset of large dimension has been shown to differ considerably depending on the cross-sectional correlation structure. Moreover, the intrinsic predictability of the data also plays an important role in determining a final allocation. Clearly, a sparse strategy that seeks a bias-variance trade-off is desirable in this setting. It is well-known that performance of a standard mean-variance optimal portfolio deteriorates out-of-sample, presumably due to the estimation error of the input parameters (the mean and covariances of the asset returns). On this basis, we introduce a method for controlling estimation error and thereby producing stable out-of-sample performance in a high-dimensional setting. We bridge ideas from two diversified strands of the literature, namely, the use of the random subspace method (Ho, 1998) and parameter regularization (Tibshirani, 1996). More precisely, we combine a number of subportfolios constructed using a regularization method that employs a tailored regularization target, itself is selected by exploiting the hierarchical clustering algorithms of Bühlmann et al. (2013) and in such a way that the maximum correlation between the targeted assets is controlled. Therefore, the proposed method produces what is referred to as a Targeted Regularized Portfolio (TRP). The subportfolio weights are then averaged to produce the overall portfolio weights, In a Monte Carlo simulation study, the proposed TRP achieves promising improvements over the gross exposure constrained portfolios of Fan, Zhang, and Yu (2012). Finally, we illustrated the utility of the TRP method in the context of two empirical applications covering four datasets associated with the Fama-French industrial portfolios and two large U.S. stock market benchmarks.

1.2 Outline of thesis

The thesis is comprised of six chapters. Following this introductory chapter, Chapter 2 reviews the fundamental concepts of Bayesian inference. Several commonly used MCMC methods for posterior simulation are explained, including the Gibbs sampler and Metropolis-Hasting algorithms. The general structure of a state space model and algorithms enabling efficient Bayesian posterior inference for them, are also detailed. A discussion of the LASSO penalized regression method is reviewed both from a Frequentist point of view and the corresponding Bayesian prior of Park and Casella (2008). In addition, the idea of Gaussian process priors for Bayesian nonparametric function estimation is detailed.

Next, Chapter 3 considers an empirical application that makes use of a Bayesian hierarchical model. Specifically, we develop a dynamic predictive model along with the corresponding Bayesian inferential methodology required to analyze the model for bank loan recovery rates. In particular, variable shrinkage is performed conditional on the underlying states of the credit cycle using a Bayesian LASSO. Our results illustrate the importance of using a dynamic model that can handle time-varying conditions, as there is significant variation in the explanatory power of the variables depending on such conditions, yielding new insights previously unavailable from the established literature.

Chapter 4 proposes an extension of a commonly used VAR model in the context of a particular macroeconomic study. The proposed model accommodates variables sampled at different frequencies as well as an unknown and potentially nonlinear reaction function that describes the nature of variable co-movement in the model using a Gaussian process prior. Again, the inferential methodology is developed and implemented via an innovative MCMC algorithm that is seen to be an extension of an existing method for nonlinear state space models proposed by Stroud, Müller, and Polson (2003). Unlike the existing method, the new algorithm incorporates a random nonlinear latent process.

Chapter 5 focus specifically on developing a novel approach, TRP, for high-dimensional portfolio selection. The new TRP method builds upon the foundations developed by Tibshirani (1996) who proposed the original LASSO to perform variable shrinkage and selection simultaneously, justified by prior belief in a sparse model structure. The usefulness of this approach is illustrated in a Monte Carlo study and in four empirical applications, in which found that the TRP method performs well out-of-sample.

This thesis concludes with Chapter 6, where its contributions from aforementioned chapters are reflected upon, and an outline of potential future research directions is given.

Chapter 2

Bayesian Inference in Financial Econometrics

2.1 Fundamentals of Bayesian Inference

The Bayesian inferential paradigm may be considered as a framework for updating update prior belief about the primary elements of a data generating process (DGP), i.e. the stochastic mechanism from which the data are thought to arise, and with the updating taking place after observations from the DGP are revealed. In contrast to its main alternative, the frequentist approach, Bayesian inference conditions only on the particular sample data at hand rather than with reference to many hypothetical repeated samples from the DGP. Therefore, under the Bayesian framework, probability distributions are viewed as expressions of one's degree of uncertainty (or belief) about unknown quantities, and not, as from the frequentist view, as a characterization of the relative frequency of associated with the hypothetical repeated samples. Bayes' theorem is the central mechanism for this update, with two sources of information, prior and sample information, being combined via this theorem. In its simplest form, Bayes' theorem is given by

$$p(\theta|y_{1:T}) \propto p(\theta)\mathcal{L}(\theta|y_{1:T})$$

where $p(\theta)$ denotes the prior probability density function (pdf, or density) associated with the prior distribution for the unknown parameter θ, $\mathcal{L}(\theta|y_{1:T})$ stands for the likelihood function for θ, taken as the joint pdf of the data realizations $y_{1:T} = (y_1, \cdots, y_T)'$ conditionally given θ, with $p(\theta|y_{1:T})$ then the resultant posterior pdf. This posterior pdf delivers the θ that characterizes the uncertainty about θ that remains after observing the sample evidence.

Implicitly, the update mechanism is also conditional on a specific econometric model $\mathcal{M}$ that defines the likelihood function. For simplicity of exposition, we suppress the explicit dependence on $\mathcal{M}$ in the notation when there is only one model specification involved in the discussion. Note that θ will be a finite dimensional parameter when working in a parametric model setting, while for semi- or non-parametric models, θ may be, potentially, of infinite dimension.

Hence, Bayesian inference is the formalization of a prior distribution reflecting prior belief about an unknown parameter, and then using Bayes' theorem to update that belief in light of observed data. Once the Bayesian posterior distribution has been obtained, uncertainty about a predicted future observation, y_{T+1} say, can be expended through the predictive posterior pdf, given by

$$p(y_{T+1}|y_{1:T}) = \int p(y_{T+1}, \theta|y_{1:T})d\theta. \tag{2.1}$$

The integrand in (2.1), i.e., $p(y_{T+1}, \theta|y_{1:T})$ is the joint posterior density of the future observation and the parameter, obtained via composition, with

$$p(y_{T+1}, \theta|y_{1:T}) = p(y_{T+1}|y_{1:T}, \theta) \times p(\theta|y_{1:T}),$$

where $p(y_{T+1}|y_{1:T}, \theta)$ is typically available directly from the specified model. This posterior predictive pdf can also be expressed as an expectation with respect to the posterior distribution for θ_j, i.e.,

$$p(y_{T+1}|y_{1:T}) = \mathbb{E}_{\theta|y_{1:T}}\left[p(y_{T+1}|y_{1:T}, \theta)\right]. \tag{2.2}$$

From the two expressions in (2.1) and (2.2) it can be seen that the posterior predictive density is obtained by marginalizing over the posterior uncertainty about θ, yielding a predictive distribution that will tend to have a greater degree of dispersion compared with one produced by conditioning on a single value, for example, a point estimate, of θ.

Due to the computational burden imposed by the integration involved in Bayes' theorem to produce, and ultimately use, the posterior distribution, various methods have been devised to facilitate the use of the Bayesian paradigm. Since analytic solutions are rarely available in practice, and hence simulation-based integration techniques have gained rapid recognition with the declining cost of computing power.

In what follows, we describe the basic structure of a MCMC for posterior computation, with an emphasis on the widely used Gibbs sampling and Metropolis-Hasting algorithms.

2.2 Markov chain Monte Carlo sampling

The posterior distribution, which encodes all uncertainty regarding the model unknowns, is of fundamental interest in Bayesian inference. A summary measure, κ, associated with a distribution have pdf $p(\theta)$, and expressed as

$$\kappa = \int h(\theta)p(\theta)d\theta,$$

where $h(\cdot)$ is a deterministic function. The basic idea of Monte Carlo methods is to generate a sample $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(G)}$ from the distribution p, either directly or indirectly, and then use the average of corresponding evaluations $h(\theta^{(1)}), h(\theta^{(2)}), \cdots, h(\theta^{(G)})$ to approximate κ. As $h(\cdot)$ may take different functional forms, any population characteristics or even density itself, can be obtained based on such a sample of draws from the distribution p.

Direct Monte Carlo (MC) methods, where independent draws of θ are generated from the distribution p, may be difficult to draw independent samples, particular when the dimension of θ is large. In such case, MCMC sampling methods provide an alternative approach for obtaining a sample from p. As the name suggests, MCMC produces a sequence of dependent draws from the target posterior distribution. Provided the draws are produced using an approximate Markovian dependence structure, the resulting sequence of draws will have a long-term (ergodic) distribution equal to the target posterior distribution.

Given a draw $\theta^{(i-1)}$ at iteration $i-1$, the next draw $\theta^{(i)}$ is drawn conditionally from a transition distribution (kernel), denoted by $K(\theta^{(i)}, \theta^{(i-1)})$, as

$$\theta^{(i)} \sim p(\theta^{(i)}|\theta^{(i-1)}, y_{1:T}).$$

Under mild conditions, the Markov chain will deliver (dependent) draws from the desired posterior distribution, with some arbitrary initial value $\theta^{(0)}$.

To ensure the initial value $\theta^{(0)}$ plays no crucial role in the convergence to the target posterior density, the Markov chain needs to be irreducible. Furthermore, an irreducible Markov chain is ergodic if all of its states are aperiodic and positive recurrent. See, for example, Robert and Casella (2013) for details. As a result, the empirical average based on the sample produced by the MCMC method

$$\hat{\kappa} = \frac{1}{G}\sum_{i=1}^{G} h(\theta^{(i)})$$

will be a valid approximation of the corresponding posterior expectation since

$$\hat{\kappa} \xrightarrow{P} \mathbb{E}\left[h(\theta)|\theta, y_{1:T}\right] \quad \text{as} \quad G \to \infty.$$

In practice, various alternatives exist for constructing such a Markov chain that converges to the target distribution. In this section, we review two dominating approaches used in the Bayesian computation literature, namely the

Metropolis-Hastings (MH) algorithm, whose root can be tracked back to Metropolis et al. (1953) and Hastings (1970). The MH algorithm nests the Gibbs sampler, which enjoyed an initial surge of popularity originating from the paper of Geman and Geman (1984) and Gelfand and Smith (1990).

2.2.1 Metropolis-Hastings algorithm

The idea of the Metropolis-Hastings algorithm is to accept or reject a draw from some proposal (instrumental) distribution that is easy to simulate from with an appropriate MH acceptance ratio. This accept-reject mechanism is somewhat similar to an accept-reject algorithm in an independent MC setup. However, the MH approach generates dependencies in the resulting sample. In what follows, the general MH algorithm is first introduced along with a brief discussion on the choice of the proposal density. Details for the implementation of certain variants employed in this thesis are provided subsequently.

To make the MH idea concrete, consider the case when the entirety of the posterior distribution $p(\theta|y_{1:T})$ is unknown and hence direct sampling approaches are infeasible. A MH algorithm proceeds by generating a candidate draw θ^* from a designated proposal density $q(\theta^*|\theta^{(i-1)})$ from where a sample of draws may be feasibly obtained. The candidate draw θ^* is then accepted with probability given by

$$\alpha(\theta^*, \theta^{(i-1)}) = \min\left\{\frac{p(\theta^*|y_{1:T})}{p(\theta^{(i-1)}|y_{1:T})}\frac{q(\theta^{(i-1)}|\theta^*)}{q(\theta^*|\theta^{(i-1)})}, 1\right\},$$

otherwise, the proposal θ^* is discarded with the value $\theta^{(i)}$ set equal to $\theta^{(i-1)}$. Thus, the resulting value $\theta^{(i)}$ may be a repetition of the value $\theta^{(i-1)}$. In addition, the transition kernel for the MH algorithm given by

$$K(\theta^{(i)}, \theta^{(i-1)}) = \alpha(\theta^{(i)}, \theta^{(i-1)})q(\theta^{(i)}|\theta^{(i-1)}) + (1 - \alpha(\theta^{(i-1)}))\delta_{\theta^{(i-1)}}(\theta^{(i)}),$$

where

$$\alpha(\theta^{(i-1)}) = \int \alpha(\theta^*, \theta^{(i-1)}) q(\theta^* | \theta^{(i-1)}) d\theta^*$$

and where $\delta_{\theta^{(i-1)}}(\cdot)$ denote the Dirac mass on $\theta^{(i-1)}$. Now, we formally state the MH algorithm below in Algorithm 1. Clearly, the calculation of the acceptance probability $\alpha(\theta^*, \theta^{(i-1)})$ requires the target posterior to be known up to a normalizing constant. Furthermore, the proposal draw is accepted with probability one when the ratio of $p(\theta^*|y_{1:T})/q(\theta^*|\theta^{(i-1)}, y_{1:T})$ is large relative to the value $p(\theta^{(i-1)}|y_{1:T})/q(\theta^{(i-1)}|\theta^*, y_{1:T})$.

Algorithm 1 Metropolis-Hastings algorithm

```
Inputs: $y_{1:T}$: data observations; $G$: number of iterations; $\theta^{(0)}$: initial value;
    $p(\theta^*|y_{1:T})$: target density; $q(\theta^*|\theta^{(i-1)}, y_{1:T})$: proposal density;
for $i = 1 \to G$ do
    Generate $\theta^*$ from the proposal density $q(\theta^*|\theta^{(i-1)}, y_{1:T})$;
    Calculate
        $\alpha(\theta^*, \theta^{(i-1)}) = \min\left\{ \frac{p(\theta^*|y_{1:T})}{p(\theta^{(i-1)}|y_{1:T})} \frac{q(\theta^{(i-1)}|\theta^*, y_{1:T})}{q(\theta^*|\theta^{(i-1)}, y_{1:T})}, 1 \right\};$
    Generate $u$ from a uniform distribution $U(0, 1)$;
    If $u < \alpha(\theta^*, \theta^{(i-1)})$, set $\theta^{(i)} = \theta^*$, otherwise, set $\theta^{(i)} = \theta^{(i-1)}$;
Outputs: A sample of $G$ draws from $p(\theta|y_{1:T})$.
```

In line with an accept-reject algorithm, the success of the MH algorithm is closely related to the choice of the proposal density $q(\cdot)$. A minimal necessary condition for the MH algorithm to produce draws form the target posterior is that the support of the proposal density covers the entire support of the target density. The distributional shape of an ideal proposal is that it matches closely to the target density. In the extreme case when the proposal density coincides with the target density exactly, the MH algorithm reduces to the Gibbs sampler. This case is discussed in the following section. An interesting variant of the MH algorithm, known as the *random walk Metropolis-Hastings* (RWMH) algorithm, arises when the proposal density is symmetric. Specifically, the proposals θ^* is constructed as

$$\theta^* = \theta^{(i-1)} + \varepsilon_i,$$

where ε_i are i.i.d and from a distribution that is symmetric about zero. The RWMH is outlined below in Algorithm 2. For simplicity of exposition, we consider a widely used distribution, $\varepsilon_i \sim \mathcal{N}(0, a\sigma_*^2)$, where a is a positive constant which is fixed by controlling the acceptance rate (Gelman, Roberts, and Gilks, 1996). The choice of this tuning parameter is crucial: a choice of a that is too small will result in only very small movement in the Markov chain due to proposed draws only ever deviate slightly from the previous chain value. Whereas, a choice of a that is too large will likely produce extreme proposals that are associated with extremely small acceptance probability $\alpha(\theta^*, \theta^{(i-1)})$, leading to long sequences where a single value is repeated in the Markov chain. Both of these result in slow movement of the chain towards its target. To select a a that delivers good convergence properties, Chen and So (2006) propose to select a having an acceptance probability of 25% to 50%. It has now become evident that in some cases, especially in high-dimensional problems, it can be very difficult, if not impossible, to construct an appropriate proposal distribution that approximates the joint posterior density well. Against this drawback,

Algorithm 2 Random walk Metropolis-Hastings algorithm

```
Inputs: $y_{1:T}$: data observations; $G$: number of iterations; $\theta^{(0)}$: initial value;
    $p(\theta^*|y_{1:T})$: target density; $\mathcal{N}(\theta^{(i-1)}, \sigma_*^2)$: proposal density;
for $i = 1 \to G$ do
    Generate $\theta^*$ from the proposal density $\mathcal{N}(\theta^{(i-1)}, a\sigma_*^2)$;
    Calculate
        $\alpha(\theta^*, \theta^{(i-1)}) = \min\left\{\frac{p(\theta^*|y_{1:T})}{p(\theta^{(i-1)}|y_{1:T})}, 1\right\};$
    Generate $u$ from a uniform distribution $U(0,1)$;
    If $u < \alpha(\theta^*, \theta^{(i-1)})$, set $\theta^{(i)} = \theta^*$, otherwise, set $\theta^{(i)} = \theta^{(i-1)}$;
Outputs: A sample of $G$ draws from $p(\theta|y_{1:T})$.
```

we describe a sampling scheme that does not need an accept-rejection step to be valid, namely, the Gibbs sampler.

2.2.2 Gibbs Sampling

Gibbs sampling is an efficient technique to generate from a desired posterior distribution indirectly, without having to calculate the relevant posterior density (see, e.g., Casella and George, 1992). In many data modelling settings, such as these explained in this thesis, it is impractical or impossible to simulate directly from the joint posterior distribution or to construct an appropriate proposal density that can be used in a MH algorithm. Indeed, the computational burden for evaluating multi-dimensional integrals increases exponentially as the dimension of a problem increases, which is often referred to as the "curse of dimensionality". In this instance, the Gibbs sampling algorithm, or "Gibbs sampler", turns the curse into a blessing by partitioning the unknown vector into a collection of sub-blocks. One finds, then, for each sub-block, the full conditional posterior distribution, which will ideally be of a recognizable form. As a result, a sample of draws of the entire vector of unknowns can be obtained iteratively by alternatively simulating the sub-blocks of unknowns from this corresponding full conditional posteriors. Intuitively speaking, conditional densities are used in Gibbs sampler since the joint posterior density can be derived by using only the full conditional densities, for example,

$$p(\theta_1, \theta_2|y_{1:T}) = \frac{p(\theta_1|\theta_2, y_{1:T})}{\int p(\theta_1|\theta_2, y_{1:T})/p(\theta_2|\theta_1, y_{1:T})d\theta_1},$$

while this generally does not hold for using marginal densities.

As an illustration, consider a two-stage Gibbs sampler, where the vector of unknowns may be partitioned into two sub-blocks, $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$. The two-stage

Gibbs sampler then alternatively simulates from the respective full conditionals

$$\theta_1^{(i)} \sim p(\theta_1^{(i)}|\theta_2^{(i-1)}, y_{1:T}), \text{and}$$
$$\theta_2^{(i)} \sim p(\theta_2^{(i)}|\theta_1^{(i)}, y_{1:T}).$$

It's worth noting that not only the Gibbs sequence $\left\{\theta_1^{(i)}, \theta_2^{(i)}\right\}$ is a Markov chain, whose ergodic distribution is joint pdf given by $p(\theta_1, \theta_2|y_{1:T})$. Also as a by-product, each of $\{\theta_1\}$ and $\{\theta_2\}$ is a Markov chain. As is evident now, the transition kernel of the two-stage Gibbs sampler is given by

$$K(\theta^{(i)}, \theta^{(i-1)}) = p(\theta_1^{(i)}|\theta_2^{(i-1)}, y_{1:T})p(\theta_2^{(i)}|\theta_1^{(i)}, y_{1:T}).$$

After obtaining a sufficient long Gibbs sequence, the average of the conditional densities $f(\theta_1|\theta_2^{(i)}, y_{1:T})$ and $f(\theta_2|\theta_1^{(i)}, y_{1:T})$ may be used to approximate the marginal densities $f(\theta_1|y_{1:T})$ and $f(\theta_2|y_{1:T})$, i.e.,

$$\hat{p}(\theta_1|y_{1:T}) = \frac{1}{G}\sum_{i=1}^{G} p(\theta_1|\theta_2^{(i)}, y_{1:T}) \text{ and } \hat{p}(\theta_2|y_{1:T}) = \frac{1}{G}\sum_{i=1}^{G} p(\theta_2|\theta_1^{(i)}, y_{1:T}),$$

where G is the number of draws in the generated Gibbs sequence. The rational behind such result arises from by first recognizing the transition kernel of the chain, for instance, the kernel for the sub-chain $\{\theta_1\}$ is obtained by integrating out θ_2,

$$K(\theta_1^{(i)}, \theta_1^{(i-1)}) = \int p(\theta_1^{(i)}|\theta_2^{(i-1)}, y_{1:T})p(\theta_2^{(i)}|\theta_1^{(i)}, y_{1:T})d\theta_2.$$

Then, following Robert and Casella (2013), it can be shown that $p(\theta_1|y_{1:T})$ is the invariance distribution associated with $\{\theta_1\}$, since for any $\theta_1 = \theta^*$, we have

$$\begin{aligned}
p(\theta_1^*|y_{1:T}) &= \int p(\theta_1^*|\theta_2, y_{1:T})p(\theta_2|y_{1:T})d\theta_2 \\
&= \int p(\theta_1^*|\theta_2, y_{1:T}) \int p(\theta_2|\theta_1, y_{1:T})p(\theta_1|y_{1:T})d\theta_1 d\theta_2 \\
&= \int\int p(\theta_1^*|\theta_2, y_{1:T})p(\theta_2|\theta_1, y_{1:T})d\theta_2 p(\theta_1|y_{1:T})d\theta_1 \\
&= \int K(\theta_1^*, \theta_1)p(\theta_1|y_{1:T})d\theta_1.
\end{aligned}$$

For a general multi-stage Gibbs sampler with k partition, $\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_k\}$, the algorithm is as follows. Therefore, a sample of $\{\theta_1, \cdots, \theta_k\}$ from joint posterior is produced, and any marginal sequence of $\theta^{(i)}$ converges to its marginal posterior with pdf $p(\theta_j|y_{1:T})$.

Algorithm 3 k-stage Gibbs sampling algorithm

```
Inputs: $y_{1:T}$: data observations; $G$: number of iterations; $\theta^{(0)}$: initial value; $p(\theta_1^{(i)}|\theta_2^{(i-1)}, \theta_3^{(i-1)}, \ldots, \theta_k^{(i-1)}, y_{1:T}), \cdots,$ $p(\theta_k^{(i)}|\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_{k-1}^{(i)}, y_{1:T})$: full conditional posterior distributions;
for $i = 1 \to G$ do
    Generate $\theta_1^{(i)}$ from $p(\theta_1^{(i)}|\theta_2^{(i-1)}, \theta_3^{(i-1)}, \ldots, \theta_k^{(i-1)}, y_{1:T})$;
    Generate $\theta_2^{(i)}$ from $p(\theta_2^{(i)}|\theta_1^{(i)}, \theta_3^{(i-1)}, \ldots, \theta_k^{(i-1)}, y_{1:T})$;
    $\vdots$
    Generate $\theta_k^{(i)}$ from $p(\theta_k^{(i)}|\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_{k-1}^{(i)}, y_{1:T})$;
Outputs: A sample of $G$ draws of $\{\theta_1, \cdots, \theta_k\}$ from $p(\theta_1, \cdots, \theta_k|y_{1:T})$.
```

We conclude this section by illustrating the usefulness of the Gibbs sampling along with the data-augmentation technique. As will become clear, the Gibbs sampler is closely related to the idea of data augmentation of Tanner and Wong (1987). Consider the Gaussian mixture model, a collection of Gaussian densities that are commonly used to approximate density with irregular forms. Let $y_{1:T} = (y_1, y_2, \cdots, y_T)$ be the vector of observed data. The mixture of Gaussian distribution is given by

$$p(y_t|\theta, w) = \sum_{j=1}^{k} w_j p(y_t|\mu_j, \sigma_j^2), \text{ for } t = 1, 2, \cdots, T,$$

where $p(\cdot|\mu_j, \sigma_j^2)$ is the Gaussian density having mean μ_j and variance σ_j^2 and k is the number of mixture components. The mixture weight, w_j, is associated with the j-th mixture component, for $j = 1, \ldots, k$, and with $\sum_{j=1}^{k} w_j = 1$. A Gibbs sampling strategy can be readily implemented in this setting when conditionally conjugate prior distributions are assumed. Specifically, if we assume $(\mu_j, \sigma_j^2) \sim \mathcal{NIG}(\bar{m}_j, \bar{h}_j, \bar{s}_j, \bar{\nu}_j)$ for $j = 1, \cdots, k$, corresponding to the Normal-Inverse Gamma distribution with density function proportional to

$$p(\mu_j, \sigma_j^2) \propto (\sigma_j^2)^{-\frac{\bar{v}_j}{2}-1} \exp\left(-\frac{\bar{s}_j}{2\sigma_j^2}\right) (\sigma_j^2)^{-\frac{1}{2}} \exp\left(-\frac{\bar{h}_j(\mu - \bar{m}_j)^2}{\sigma_j^2}\right).$$

In such case the data-augmentation approach will associate with every observation a latent and random indicator variable $Z_t \in (1, \ldots, k)$, whose realized value $Z_t = z_t$ indicates the mixture component to be associated with y_t. Once augmented with the indicator variable, the mixture model can be written as

$$Z_t \sim \mathcal{M}_k(1; \pi_{t,1}, \cdots, \pi_{t,k}), \quad y_t|z_t \sim p(y_t|\mu_{z_t}, \sigma_{z_t}^2),$$

where $\mathcal{M}_k(\cdot)$ denotes a k-dimensional multi-nominal distribution with probability parameter vector $(\pi_{t,1}, \cdots, \pi_{t,k})$. Then, with fixed mixture weights $w_1, \cdots, w_k$ for mixture components, the joint posterior distribution for all unknowns is given by

$$\begin{aligned}
&p\left(\left\{\mu_j, \sigma_j^2\right\}_{j=1}^{k}, \{z_t\}_{t=1}^{T} | y_{1:T}, \{w_j\}_{j=1}^{k}\right) \\
\propto &p\left(\{(\mu_j, \sigma_j)\}_{j=1}^{k}\right) p\left(\{z_t\}_{t=1}^{T} | \{w_j\}_{j=1}^{k}\right) \times p\left(y_{1:T}| \left\{(\mu_j, \sigma_j^2)\right\}_{j=1}^{k}, \{z_t\}_{t=1}^{T}\right) \\
\propto &p(\left\{(\mu_j, \sigma_j^2)\right\}_{j=1}^{k}) \prod_{t=1}^{T} p(z_t| \{w_j\}_{j=1}^{k}) p(y_t| \left\{(\mu_j, \sigma_j^2)\right\}_{j=1}^{k}).
\end{aligned}$$

In this setup, direct sampling from the joint posterior is difficult since the vector of unknowns is of large dimension. Further, while its analytical form is known, the integrating constant required to compute posterior probabilities is unknown. However, the Normal-Inverse Gamma is conditionally conjugate to the Gaussian distribution, then a set of full conditional distribution may be derived, with each available in closed-form. Specifically, we have

$$\begin{aligned}
\mu_j, \sigma_j | z_{1:T}, y_{1:T} &\sim \mathcal{NIG}(\bar{\bar{m}}_j, \bar{\bar{h}}_j, \bar{\bar{s}}_j, \bar{\bar{\nu}}_j) \text{ for each } j = 1, \cdots, k, \\
z_t | \left\{(\mu_j, \sigma_j), w_j\right\}_{j=1}^{k} &\sim \mathcal{M}_k(1, \bar{\bar{\pi}}_t) \text{ for each } t = 1, \cdots, T,
\end{aligned}$$

where

$$\begin{aligned}
\bar{\bar{m}}_j &= \bar{\bar{h}}_j^{-1}(\bar{h}_j \bar{m}_j + \sum_{t=1}^{T} \mathbb{I}_{z_t=j} y_t), \\
\bar{\bar{h}}_j &= \bar{h}_j + n_j, \\
\bar{\bar{s}}_j &= \bar{s}_j + \sum_{t=1}^{T} \mathbb{I}_{z_t=j} y_t^2 + \bar{h}_j \bar{m}_j^2 - \bar{\bar{h}}_j \bar{\bar{m}}_j^2 \\
\bar{\bar{\nu}}_j &= \bar{v}_j + n_j,
\end{aligned}$$

$n_j = \sum \mathbb{I}_{z_t=j}$ and $\bar{\bar{\pi}}_t = (\bar{\bar{\pi}}_{t,1}, \cdots, \bar{\bar{\pi}}_{t,k})$ with $\bar{\bar{\pi}}_j \propto w_j p(y_t | \mu_j, \sigma_j^2)$. Given the initial values of the static parameter vector $\boldsymbol{\theta} = \left(\{\mu_j, \sigma_j\}_{j=1}^{k}\right)$, a sample from the posterior posterior may be obtained by exploiting a two-stage Gibbs sampler, which is summarized in Algorithm 4.

If the mixture weights, $w_{1:k}$, are random and unknown a priori, a Dirichlet distribution prior

$$p_1, \cdots, p_k \sim \mathcal{D}_k(\gamma_1, \cdots, \gamma_k)$$

may be used for the mixture weights. In particular, Dirichlet distribution is conditionally conjugate to the multi-nominal distribution, and the full conditional

Algorithm 4 Gibbs sampling algorithm for Gaussian mixture models with fixed mixture weights

```
**Inputs:** $y_{1:T}$: data observations; $G$: number of iterations; $\boldsymbol{\theta}^{(0)}$: initial value; $\mathcal{NIG}(\bar{\bar{m}}_j, \bar{\bar{h}}_j, \bar{\bar{s}}_j, \bar{\bar{\nu}}_j)$, $\mathcal{M}_k(1, \bar{\bar{\pi}}_t)$: full conditional posterior distributions;
**for** $i = 1 \to G$ **do**
    Generate $Z_t$, for $t = 1, 2, \cdots, T$, from
        $Z_t | \{(\mu_j, \sigma_j), w_j\}_{j=1}^{k} \sim \mathcal{M}_k(1, \bar{\bar{\pi}}_t);$
    Generate $\{\mu_j, \sigma_j\}$, for $j = 1, 2, \cdots, k$, from
        $\mu_j, \sigma_j | z_{1:T}, y_{1:T} \sim \mathcal{NIG}(\bar{\bar{m}}_j, \bar{\bar{h}}_j, \bar{\bar{s}}_j, \bar{\bar{\nu}}_j);$
**Outputs:** A sample of $G$ draws of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|y_{1:T})$.
```

Algorithm 5 Gibbs sampling algorithm for Gaussian mixture models with random mixture weights

```
**Inputs:** $y_{1:T}$: data observations; $G$: number of iterations; $\boldsymbol{\theta}^{(0)}$: initial value; $\mathcal{NIG}(\bar{\bar{m}}_j, \bar{\bar{h}}_j, \bar{\bar{s}}_j, \bar{\bar{\nu}}_j)$, $\mathcal{M}_k(1, \bar{\bar{\pi}}_t)$, $\mathcal{D}_k(\gamma_1 + n_1, \cdots, \gamma_k + n_k)$: full conditional posterior distributions;
**for** $i = 1 \to G$ **do**
    Generate $Z_t$, for $t = 1, 2, \cdots, T$, from
        $Z_t | \{(\mu_j, \sigma_j), w_j\}_{j=1}^{k} \sim \mathcal{M}_k(1, \bar{\bar{\pi}}_t);$
    Generate $\{\mu_j, \sigma_j\}$, for $j = 1, 2, \cdots, k$, from
        $\mu_j, \sigma_j | z_{1:T}, y_{1:T} \sim \mathcal{NIG}(\bar{\bar{m}}_j, \bar{\bar{h}}_j, \bar{\bar{s}}_j, \bar{\bar{\nu}}_j);$
    Generate $w_j$, for $j = 1, 2, \cdots, k$, from
        $w_{1:k} | z_{1:T} \sim \mathcal{D}_k(\gamma_1 + n_1, \cdots, \gamma_k + n_k);$
**Outputs:** A sample of $G$ draws of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|y_{1:T})$.
```

posterior for the mixture weights are known in closed-form, which is given by

$$w_{1:k}|z_{1:T} \sim \mathcal{D}_k(\gamma_1 + n_1, \cdots, \gamma_k + n_k).$$

In such case the Gibbs-based sampler will involve an additional step to generate $w_{1:k}$. We summarized the algorithm to implement the Gibbs sampler for Gaussian mixture models with random mixture weights in Algorithm 5, where the vector of static parameter is given by $\boldsymbol{\theta} = \left(\{\mu_j, \sigma_j\}_{j=1}^k, \{\omega_j\}_{j=1}^k\right)$.

In what follows, we describe the role of prior distributions played in Bayesian inference, with an emphasis on a shrinkage prior in a regression context, and priors to be specified for unknown function.

2.3 Prior Specification

The prior distribution, $p(\theta)$, through which the investigator's initial belief regarding the unknown parameter θ is quantified, is an essential component of Bayesian inference. Traditionally, specification of the prior distribution (or simply, 'the prior') is determined (strictly) before (i.e. *prior to*) observing the data. The investigator's prior knowledge about θ, such as its plausible values and known restrictions, would be incorporated into this distributional specification. Within the context of a given parametric model, a prior distribution may be classified as belonging to one of two categories: informative and non-informative, with the classification referencing whether the distribution is 'sharp' or 'diffuse' relative to the available sample information. In relatively simple models at least, the latter will often lead to posterior parameter estimates that are numerically similar to those obtained by optimising the likelihood function alone. In contrast, when the prior information is sharp, posterior inference can be heavily influenced by the prior, and consequently the evidence from data may impact little on the resulting inferential conclusions. In reality, the aim is to produce posterior inference that is a combination of both investigator's prior belief and sample evidence. The resulting posterior distribution can be used to quantify uncertainties of interest, for instance, the marginal posterior means can be considered as the optimal Bayesian point estimator under mean-squared error loss.

For some DGPs, notably those in the exponential family class, the so-called 'natural conjugate' prior is available. Such a prior distribution provides analytic tractability of the resulting posterior due to the fact that the prior and posterior belong to same family of distributions. The hyper-parameters of the natural conjugate prior may be selected to provide relatively weak (uninformative) prior information, relative to the likelihood function, or to provide strong information. As such these priors are somewhat flexible with regard to their influence on posterior inference.

In addition, more contemporary approaches to the specification of the prior are often more pragmatic than idealogical. Rather than relying purely on subjective considerations, certain methodologies for articulating priors have been proposed in the literature that reflect an attempt to *minimize* subjectivity. Such priors, commonly referred to as 'reference priors' aim to produce inference that depends only on the data and the assumed DGP, thereby rendering posterior inference that would, it is argued, be more palatable to non-Bayesians and applied practitioners who may lack the skills to formally develop probabilistic priors consistent with subjective concerns. See Berger, Bernardo, and Sun (2009) for recent advances and references relating to earlier development.

In many situations, however, determining a suitable prior will be challenging, and particularly so when dealing with DGPs that involve many parameters. Conjugate and reference priors are often not available, leaving the analyst to grapple with trying to match true prior belief regarding potential outcomes with rich enough DGP structures to accommodate the increasing volume and complexity of data that are observed. Often these DGP structures involve nonlinearities, time-varying behavior and other complex dependencies. In such a setting, not only is the prior difficult to specify, but so also is the computational problem of obtaining, and working with, the posterior distribution.

As aforementioned, computation of the posterior distribution is a direct consequence of the choice of the prior specification. In the remainder of this section, we focus on the discussion of two particular pragmatic prior specifications for certain unknown parameters in relevant to models investigated in this thesis, and the resulting MCMC-based computation strategies needed to obtained the resulting posterior distribution.

2.3.1 The Bayesian LASSO

When a large set of predictor variables is available in a regression context, an investigator is often interested in improving the overall model predictive accuracy of the regression model (out-of-sample) as well as determining the subset of covariates most relevant to this prediction. This take is made particularly challenging when the number of covariates is large and, as is often the case, when potential covariance exists between the regressors. For this purpose, the LASSO of Tibshirani (1996) was proposed, from a frequentist perspective. The LASSO results in an ℓ_1 penalized least squares estimator of the coefficients $\boldsymbol{\beta}$ in a linear regression model, e.g.,

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.3}$$

produced through an ℓ_1 penalized least square method, where $\mathbf{y} = [y_1, \cdots, y_T]'$ is a vector of univariate responses, $X = \left[\mathbf{x}'_1, \cdots, \mathbf{x}'_p\right]$ is a $T \times p$ design matrix and $\boldsymbol{\varepsilon} = [\varepsilon_1, \cdots, \varepsilon_T]'$ is a vector of random error. Under this framework, an estimator of the parameter vector $\boldsymbol{\beta}$ is obtained by solving

$$\min_{\boldsymbol{\beta}} ||\mathbf{y} - X\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1,$$

where λ is the Lagrange multiplier that governs the size of the penalty due to the L_1 norm of $\boldsymbol{\beta}$ given by $||\boldsymbol{\beta}||_1$, which is typically called the "shrinkage" parameter. This penalty serves to off-set the usual sum of squared errors criteria, given by $||\mathbf{y} - X\boldsymbol{\beta}||_2^2$, which may be improved by the inclusion of additional regressors.

The aforementioned minimization problem can be recast as an Lagrangian dual problem which takes the following form:

$$\min_{\boldsymbol{\beta}} ||\mathbf{y} - X\boldsymbol{\beta}||_2^2, \quad \text{s.t.} \quad ||\boldsymbol{\beta}||_1 \leq c(\lambda),$$

where there exists a unique $c(\lambda)$ that corresponds to λ. In particular, the LASSO reduces to the ordinary least squares (OLS) estimator when $\lambda = 0$ and, on the other hand, leads to a constant model when $\lambda = \infty$. In practice, λ may be often selected using a cross-validation technique over a grid of plausible values of λ, such as obtained in using the Least Angle Regression (LARS) algorithm proposed by Efron et al. (2004).

Unlike the OLS estimator, which is unbiased for the parameter β when all required regressors are included in the model, the LASSO introduces a small amount of bias in order to reduce the overall variance of the estimator. Thus the LASSO retains the desirable features of both subset selection (Guyon and Elisseeff, 2003) and ridge regression (Hoerl and Kennard, 1970) while performing variable shrinkage and selection simultaneously. It is hoped that by doing so the overall predictive performance will be improved over that of OLS.

Tibshirani (1996) also gives a Bayesian interpretation for the regression analysis techniques of LASSO and its relative, ridge regression. The LASSO estimator can be interpreted as the Bayesian posterior mode under a Gaussian likelihood corresponding to (2.3) and under the assumption of independent double-exponential priors for each β_j, given by

$$p(\beta_j|\lambda) = \prod_{j=1}^{p} \frac{\lambda}{2} \exp\left(-\lambda|\beta_j|\right).$$

Park and Casella (2008), however, show that the corresponding posterior distribution is not necessarily unimodal as this result depends upon the choice of the prior for σ^2, where σ^2 denotes the noise variance. For example, the joint posterior distribution of β and σ^2 is bimodal when a scale-invariant prior $p(\sigma^2) \propto 1/\sigma^2$ will be used, regardless of the distribution assumed for $\boldsymbol{\varepsilon}$. On the other hand, ridge regression implicitly uses independent normal prior distributions for each regression coefficient having mean zero and with a variance parameter that is inversely proportional to the shrinkage parameter.

To mitigate this issue, Park and Casella (2008) propose the Bayesian LASSO based on a conditional (independent) Laplace prior of the form

$$\pi(\boldsymbol{\beta}|\sigma^2, \lambda) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\lambda|\beta_j|/\sqrt{\sigma^2}\right) \tag{2.4}$$

The resulting joint conditional prior along with the non-informative marginal prior $p(\sigma^2) \propto 1/\sigma^2$ ensures a unimodal posterior distribution for the regression parameters. This prior also has the added benefit that a sample from the posterior distribution is relatively easy to obtain due to a representation of the Laplace distribution as a scale mixture of normal distributions. That is, for scalar valued β, if $\beta \sim \mathit{Laplace}(\tau)$, then its pdf is given by

$$p(\beta \mid \tau) = \frac{\tau}{2} \exp\left(-\tau|\beta|\right),$$

for $-\infty < \beta < \infty$ and $\tau > 0$. Then, if we take β *given* variance s as having a $\mathcal{N}(0, s)$ distribution and assume that s itself is exponentially distributed, with rate parameter $\tau^2/2$, then, in fact, the *marginal* distribution of β is $\mathit{Laplace}(\tau)$ with

$$p(\beta \mid \tau) = \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\left(-\beta^2/2s\right) \frac{\tau^2}{2} \exp\left(-\tau^2 s/2\right) ds.$$

Given the produce form in (2.4), not only can a conditional Laplace prior be used for $\beta_1, \cdots, \beta_p$ but in fact independent conditional exponential priors, given the shrinkage parameter, λ, may be used.

Based on the aforementioned representation, the regression model with Gaussian errors, under the Bayesian LASSO prior, has a hierarchical structure that can be exploited when simulating from the joint posterior distribution. This hierarchical model is represented as

$$\begin{aligned} y_{1:T}|X, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2 I_T), \\ \boldsymbol{\beta}|\sigma^2, \tau_1^2, \cdots, \tau_p^2 &\sim \mathcal{N}(\mathbf{0}_p, \sigma^2 D_{\boldsymbol{\tau}}), \\ D_{\boldsymbol{\tau}} &= diag(\tau_1^2, \tau_2^2, \ldots, \tau_p^2), \\ \tau_1^2, \ldots, \tau_p^2|\sigma^2 &\sim \frac{\lambda^2}{2} \exp\left(-\lambda^2 \tau/2\right), \\ \sigma^2 &\sim \pi(\sigma^2) \end{aligned}$$

That is, the Bayesian LASSO assumes an independent (hyper) prior structure, with

$$\tau_1^2, \tau_2^2, \ldots, \tau_p^2 \mid \lambda^2 \overset{iid}{\sim} Exp(\lambda^2/2),$$

and where $Exp(s)$ denotes the exponential distribution with rate parameter s, corresponding to a mean value of $1/s$. The shrinkage parameter of the Bayesian LASSO can be chosen by the use of an appropriate hyper-prior, e.g., a Gamma distribution for λ^2, given by

$$p(\lambda^2) = \frac{\delta^r}{\Gamma(r)} (\lambda^2)^{r-1} e^{-\delta\lambda^2}.$$

If a conditionally conjugate prior is used for σ^2, the following efficient Gibbs algorithm (Algorithm 6) simulates from the joint posterior distribution of σ^2, λ and the auxiliary variables $\tau_1^2, \tau_2^2, \ldots, \tau_p^2$. The inverse-Gaussian density used in Algorithm 6 is given by

$$f(x) = \sqrt{\frac{\lambda}{2\pi}} x^{\frac{3}{2}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\},$$

Algorithm 6 Gibbs sampling algorithm for the Bayesian LASSO

```
**Inputs:** $y_{1:T}, X$: data observations; $G$: number of iterations; $\theta^{(0)} = \left(\beta^{(0)}_{1:p}, \tau^{(0)}_{1:p}, \lambda^{2(0)}, \sigma^{2(0)}\right)$: initial value; $\mathcal{IG}(\bar{\bar{a}}, \bar{\bar{b}})$, $Gamma(\bar{\bar{r}}, \bar{\bar{\delta}})$, $Inverse\text{-}Gaussian(\bar{\bar{\mu}}, \bar{\bar{\lambda}})$, $\mathcal{N}(\bar{\bar{A}}, \bar{\bar{B}})$: full conditional posterior distributions;
**for** $i = 1 \to G$ **do**
    Generate error variance parameter
        $$\sigma^2 | y_{1:T}, X \sim \mathcal{IG}(\bar{\bar{a}}, \bar{\bar{b}}),$$
    where $\bar{\bar{a}} = \frac{(n+p-1)}{2}$ and $\bar{\bar{b}} = \frac{(y_{1:T} - X\boldsymbol{\beta})'(y_{1:T} - X\boldsymbol{\beta})}{2} + \frac{\boldsymbol{\beta} D_\tau^{-1} \boldsymbol{\beta}}{2}$;
    Generate LASSO shrinkage parameter
        $$\lambda^2 | y_{1:T}, X, \sigma^2, \{\tau_j\} \sim Gamma(\bar{\bar{r}}, \bar{\bar{\delta}}),$$
    where $\bar{\bar{r}} = p + \bar{r}$ and $\bar{\bar{\delta}} = \sum_{j=1}^{p} \tau_j^2 + \bar{\delta}$;
    Generate the inverse local shrinkage parameter, for $j = 1, \cdots, p$,
        $$1/\tau_j^2 | y_{1:T}, X, \sigma^2, \lambda^2 \sim Inverse\text{-}Gaussian(\bar{\bar{\mu}}, \bar{\bar{\lambda}}),$$
    where $\bar{\bar{\mu}} = \sqrt{\frac{\lambda^2 \sigma^2}{\beta_j^2}}$ and $\bar{\bar{\lambda}} = \lambda^2$;
    Generate the regression parameters,
        $$\boldsymbol{\beta} | y_{1:T}, X, \sigma^2, \tau_{1:p} \sim \mathcal{N}(\bar{\bar{A}}, \bar{\bar{B}}),$$
    where $\bar{\bar{A}} = (X'X + D_\tau^{-1}) X' y_{1:T}$ and $\bar{\bar{B}} = \sigma^2 (X'X + D_\tau^{-1})^{-1}$;
**Outputs:** A sample of $G$ draws of $\theta$ from $p(\theta | y_{1:T}, X)$.
```

where μ is a location parameter and λ is a shape parameter.

The Bayesian LASSO has a few advantages over the ordinary frequentist LASSO. First, the turning parameter λ in ordinary LASSO is typically chosen by cross validation, which does not appear to have a strong theoretical basis (Zou, 2006). The Bayesian LASSO offers an alternative through the use of a hyperprior for λ (implied by the hyper-prior on λ^2). In contrast, for example, a single value may be selected for λ so that the resulting β coefficient estimates resemble those of the standard frequentist LASSO, for example via an empirical Bayes approach. Of course, the Bayesian can take full advantage of the existing parameter uncertainties by marginalising over the prior structures to produce, for example, Bayesian point and interval estimates for the regression coefficients.

On the other hand, however, and with probability one, the Bayesian LASSO does not permit an individual regression parameter to be set to zero in the posterior, and hence no reduction in the number of regressors is obtained. Rather, only the influence of regressors will be diminished. If reduction to a subset of covariates is desired, decision rules based on marginal Bayesian creditable intervals may be used (Park and Casella, 2008). Overall, the Bayesian LASSO is a powerful technique for regression problems when a large set of covariates is available, as is the case of the empirical applications investigated in Chapter 4 and Chapter 5 of this thesis.

In the following section, we turn to the discussion of another useful prior for modelling unknown functions.

2.3.2 The Gaussian Process Prior

We discuss now a useful prior for non-parametric function estimation. A common task in data analysis is to estimate a function $g(x)$ given some noisy observations $y_{1:T}$ at given input locations $\mathbf{x} = (x_1, \cdots, x_T)'$. A parametric approach to this problem is to select a particular class of functions, for example, polynomial class, as the model and then to estimate the parameters involved in that model class by minimizing an appropriate loss function at observed data points, e.g., mean squared error between the fitted values and the observed values associated with $\mathbf{x}$, hereby, yielding a 'best fitting' model for the given data set.

In contrast, Gaussian process is a prior may be used to produce a nonparametric Bayesian method for function estimation. Knowledge of the function values $g(x)$ is encapsulated in the prior distribution $p(g(x))$ and updated through the likelihood function $p(y_{1:T}|x, g(x))$, encompassing Bayesian nonlinear regression,

$$p(g(x)|y_{1:T}, x) = \frac{p(g(x))p(y_{1:T}|g(x), x)}{p(y_{1:T}|x)},$$

where $p(g(x)|y_{1:T}, x)$ is the posterior distribution. This approach provides a joint probability distribution over the function value $g(x_1) \cdots, g(x_k)$ given particular collection of regressors $x_1, \cdots, x_k$, conditioned on the observed data. Roughly speaking, nonparametric modeling using a Gaussian Process extends the way of expressing belief about the unknown function $g(x)$ using probability distribution over an infinite dimensional object, namely $g(x)$ itself. In particular, Gaussian process has a few advantages over parametric approaches. First, investigator can express beliefs about the functions via the prior distribution instead of having to pick a particular class of parametric functional forms. Second, not only does the Gaussian process provide a posterior distribution over possible functions but also it accounts for the uncertainty about that function.

As the name suggests, a Gaussian process is a stochastic process that is completely specified by a mean function $m(x)$ and a covariance function $k(x_i, x_j)$ for all $i, j \in (1, \cdots, n)$. It may be considered as a distribution over the space of functions where the function values (or outputs), at any finite collection of points $x_1, \cdots, x_n$ are jointly Gaussian, with

$$\begin{bmatrix} g(x_1) \\ \vdots \\ g(x_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix} \right).$$

As a result, the marginal prior distributions for each function value $g(x_i)$ is also Gaussian with a particular mean and variance given by $m(x_i)$ and $k(x_i, x_j)$. On the other hand, for given input locations x_1 and x_2, the covariance between corresponding function values is specified by the covariance function $k(x_1, x_2)$.

To demonstrate the usefulness of the Gaussian process, we consider the case where of an additive Gaussian noise,

$$y_t = g(x_t) + \varepsilon_t,$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. With the added assumption that the errors, denoted by ε, are mutually independent, the conditional likelihood function associated with observations $y_1, \cdots, y_T$ at given inputs $x_1, \cdots, x_T$ is given by

$$p(y_{1:T}|\mathbf{x}, g_{1:T}) = \prod_{t=1}^{T} \mathcal{N}(y_t; g_t, \sigma_\varepsilon^2),$$

where $g_{1:T} = g(x_1), \cdots, g(x_T)$. For additive Gaussian noise, the posterior distribution for function values is analytically tractable. Specifically, the posterior distribution can be obtained using conditional linearity property of multivariate Gaussian distribution since $y_{1:T}$ and $g_{1:T}$ are jointly Gaussian, i.e.,

$$\begin{bmatrix} y_{1:T} \\ g_{1:T} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}) \end{bmatrix}, \begin{bmatrix} K(\mathbf{x}, \mathbf{x}) + \sigma_\varepsilon^2 I_T & K(\mathbf{x}, \mathbf{x}) \\ K(\mathbf{x}, \mathbf{x}) & K(\mathbf{x}, \mathbf{x}) \end{bmatrix} \right),$$

and the conditional linearity property of multivariate Gaussian yields

$$p(g_{1:T}|y_{1:T}, \mathbf{x}) = \mathcal{N}(m(\mathbf{x}) + K\left[K + \sigma_\varepsilon^2 I_T\right]^{-1}(y_{1:T} - m(\mathbf{x})), K - K\left[K + \sigma_\varepsilon^2 I_T\right]^{-1}K),$$

where K is the $T \times T$ matrix given by $K = K(\mathbf{x}, \mathbf{x})$. Similarly, the posterior distribution of function evaluations associated with a set of test points, i.e., values comprising to $(\bar{x}_k, g(\bar{x}_k))$ not associated with any observations $y_{1:T}$ for $t = 1, \cdots, T$, denoted $\bar{\mathbf{x}}$, is also Gaussian,

$$p(\bar{g}|y_{1:T}, \mathbf{x}, \bar{\mathbf{x}}) = \mathcal{N}(\mathfrak{m}, \mathfrak{s}),$$

where

$$\begin{aligned} \mathfrak{m} &= m(\bar{\mathbf{x}}) + K(\bar{\mathbf{x}}, \mathbf{x})\left[K(\mathbf{x}, \mathbf{x}) + \sigma_\varepsilon^2 I_T\right]^{-1}(y_{1:T} - m(\mathbf{x})) \\ \mathfrak{s} &= k(\bar{\mathbf{x}}, \bar{\mathbf{x}}) - K(\bar{\mathbf{x}}, \mathbf{x})\left[K(\mathbf{x}, \mathbf{x}) + \sigma_\varepsilon^2 I_T\right]^{-1}K(\mathbf{x}, \bar{\mathbf{x}}). \end{aligned}$$

The marginal likelihood in this setting is obtained by averaging over the possible function values $g_{1:T}$, i.e.,

$$p(y_{1:T}|\mathbf{x}) = \int p(y_{1:T}|\mathbf{x}, g_{1:T})p(g_{1:T}|\mathbf{x})dg_{1:T}.$$

Hence, since $g_{1:T}$ follows a multivariate Gaussian distribution, this marginal likelihood is tractable under the assumption of Gaussian noise, in particular, the log marginal likelihood function is given by

$$\log p(y_{1:T}|\mathbf{x}) = -\frac{1}{2}(y_{1:T} - m(\mathbf{x}))'(K + \sigma_\varepsilon^2 I_T)^{-1}(y_{1:T} - m(\mathbf{x})) - \frac{1}{2}|K + \sigma_\varepsilon^2 I_T| - \frac{T}{2}\log 2\pi.$$

Notably, having a closed-form expression for the (log) likelihood is useful when there are unknown hyper-parameters in the covariance function.

The covariance function $k(\cdot, \cdot)$ is a fundamental element for a Gaussian process as it expresses belief regarding the properties of the underlying function. It generates the so called Gram matrix K whose (i, j) entry is $K_{ij} = k(x_i, x_j)$, for given input locations $\mathbf{x}$. The most commonly-used covariance function appears to be the 'squared exponential' covariance (kernel) function (Rasmussen and Williams, 2006) which takes the form

$$k_{SE}(x_i, x_j) = \sigma_f \exp\left(-\frac{(x_i - x_j)^2}{2\ell^2}\right),$$

where σ_f and ℓ are hyper-parameters. This squared exponential kernel is infinitely differentiable, so its derivatives are available at all orders. For instance, the first derivative of this kernel function respect to $x_j = \bar{\boldsymbol{\mu}}$ is given by

$$\begin{aligned}
\frac{\partial k(x_i, \bar{\boldsymbol{\mu}})}{\partial \bar{\boldsymbol{\mu}}} &= \frac{\partial}{\partial \bar{\boldsymbol{\mu}}} \left\{ \sigma_f \exp\left(-\frac{1}{2\ell^2}(x_i - \bar{\boldsymbol{\mu}})^2\right) \right\} \\
&= \frac{\partial}{\partial \bar{\boldsymbol{\mu}}} \left\{ \exp\left(-\frac{1}{2\ell^2}(x_i - \bar{\boldsymbol{\mu}})^2\right) \right\} k(x_i, \bar{\boldsymbol{\mu}}) \\
&= -\ell^{-2}(x_i - \bar{\boldsymbol{\mu}}) k(\bar{\boldsymbol{\mu}}, x_i),
\end{aligned}$$

which is a scalar.

For the purpose of illustration, we plot the posterior of Gaussian process using a squared exponential covariance function with prior mean $m(x) = 0$ for all x in Figure 2.1. Notably, the uncertainty in the function values collapses around observations and expands as we move away from data points. As is common in nonparametric function estimation, the GP posterior returns to the prior as we move away from the data, resulting in bias and large uncertainty near the boundaries of the support produced by x.

Having discussed the role of prior distributions played in Bayesian inference, we focus on the discussion of state space models in the context of Bayesian inference in the following section.

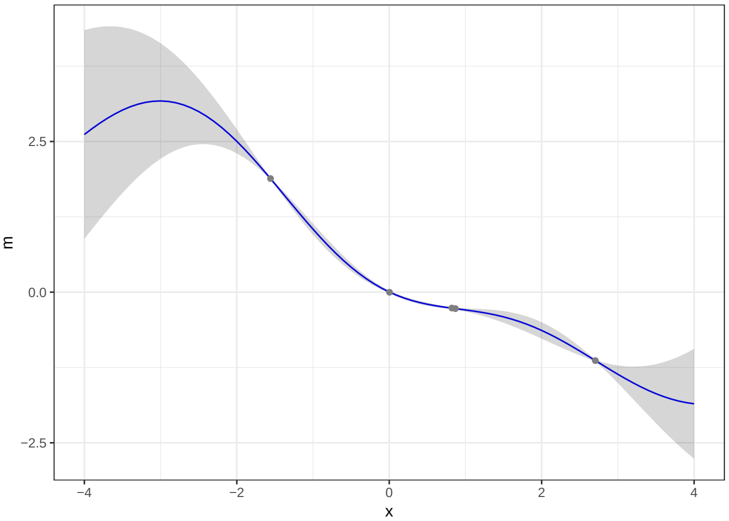

FIGURE 2.1: Gaussian process posterior using a squared exponential kernel. The blue line shows the mean of the posterior distribution, along with 95% creditable interval (gray shaded area).

2.4 Bayesian Inference for State Space Models

A state space model (SSM) provides a unified framework for time series analysis. In this approach a series of observations $y_1, \cdots, y_T$ are associated with an unobserved series of state variables $\alpha_1, \cdots, \alpha_T$ over time. The evolution of the state variable and the corresponding observations are assumed to be governed by the stochastic mechanism under study. With knowledge of the observations $y_{1:T} = y_1, \cdots, y_n$, a main purpose of a state space analysis is to infer the relevant distributions for latent variables $\alpha_1, \cdots, \alpha_T$, along with the estimation of any static parameters and the forecasting of future outcomes.

State space modeling is used in a range of research areas of broad and current interest. A well-studied example of a state space model useful to model financial time series is the stochastic volatility model (Kim, Shephard, and Chib, 1998; Jacquier, Polson, and Rossi, 2002; Eraker, Johannes, and Polson, 2003). Another example is to analyze high-dimensional time series, for example, biological data sets, using the state space representation for dynamic graphic models (Carvalho and West, 2007). Not only does the SSM framework encompass a wide range of real applications on its own, but also it is conveniently used as a workhorse facilitating many existing time series models, including moving average (MA) model, vector autoregressive (VAR) model, as well as for inference when dealing with missing data. Exploiting the state space representation, powerful MCMC simulation methods facilitate Bayesian inference in these complex settings.

The empirical applications investigated in this thesis are developed using certain state space representations arising from a Bayesian hierarchical model, and a mixed frequency VAR model, respectively. In each case, the state space representation is used to connect latent processes to the observations, with conditional inference undertaken through the development of targeted MCMC algorithms. In each cases, the challenge of the problem is essentially due to the presence if a high-dimensional latent variable. Thus, the SSM provides the foundation needed for the modeling and inferential framework used in Chapters 3 and 4.

A general discrete-time SSM may often be explicitly stated using the following form

$$y_t = f(\alpha_t, \varepsilon_t, \phi) \qquad \varepsilon_t \overset{ind}{\sim} p(\varepsilon_t|\alpha_t, \phi) \tag{2.5}$$

$$\alpha_{t+1} = g(\alpha_t, \eta_t, \phi) \qquad \eta_t \overset{ind}{\sim} p(\eta_t|\alpha_t, \phi) \tag{2.6}$$

for $t = 1, \cdots, T$, where ϕ denotes a vector of static parameters and the equations (2.5) and (2.6), associated with y_t and α_t are called the measure equation and the state transition equation respectively. Both y_t and α_t can potentially be multivariate. A feature of this model, as aforementioned is that, in the time series context, the dynamics involved are typically expended as device only through α_t, whose process usually assumed to follow a first-order Markov process starting with $\alpha_0 \sim p(\alpha_0|\phi)$. The SSM is completed with assumptions regarding the random disturbances ε_t and η_t, which will have distribution with zero mean and with possibly correlated covariance structure. For simplicity of exposition, we denote the vector of latent states $a_{0:T} = (\alpha_0, \cdots, \alpha_T)'$ for the reminder of this section.

In this generic SSM setup, the joint density function for $y_{1:T}$ and $\alpha_{0:T}$ is given by

$$p(y_{1:T}, \alpha_{1:T}|\phi) = p(\alpha_0|\phi)\prod_{t=1}^{T} p(y_t|\alpha_t, \phi)p(\alpha_t|\alpha_{t-1}, \phi),$$

and hence given a prior pdf for ϕ, $p(\phi)$, the joint posterior satisfies

$$\begin{aligned} p(\phi, \alpha_{0:T}|y_{1:T}) &\propto p(y_{1:T}|\alpha_{0:T}, \phi)p(\alpha_{0:T}|\phi)p(\phi) \\ &= \left[\prod_{t=1}^{T} p(y_t|\alpha_t, \phi)\right]\left[p(\alpha_0|\phi)\prod_{t=1}^{T} p(\alpha_t|\alpha_{t-1}, \phi)\right] p(\phi). \end{aligned}$$

In terms of actually computing this posterior distribution, a two-stage Gibbs sampling method can at least be conceptually used for posterior simulation, consisting of alternatively generating

$$\phi \sim p(\phi|y_{1:T}, \alpha_{0:T}),$$

and

$$\alpha_{0:T} \sim p(\alpha_{0:T}|\phi, y_{1:T})$$

alternatively. However, unless the SSM takes a linear Gaussian, or a discrete finite state, form, then the objective of computing the state variables rapidly becomes cumbersome as T increases. Nevertheless, the theory of state filtering and smoothing suggests how to update knowledge of a system as a new observation y_t is made available. Specifically, the conditional density, $p(y_t|y_{1:t-1}, \phi)$, as well as the filtered density, $p(\alpha_t|y_{1:t}, \phi)$, are found progressively via

$$p(y_t|y_{1:t-1}, \phi) = \int p(y_t|\alpha_t, \phi)p(\alpha_t|y_{1:t-1}, \phi)d\alpha_t$$

and

$$p(\alpha_t|y_{1:t},\phi) = \frac{p(\alpha_t|y_{1:t-1},\phi)p(y_t|\alpha_t,\phi)}{p(y_t|y_{1:t},\phi)},$$

as each observation y_t becomes available for $t = 1, 2, \cdots, T$. Moreover, the state predictive density given by

$$p(\alpha_t|y_{1:t-1},\phi) = \int p(\alpha_t|\alpha_{t-1},\phi)p(\alpha_{t-1}|y_{1:t-1},\phi)d\alpha_{t-1},$$

and the so called smoothed distribution given by

$$p(\alpha_t|y_{1:T},\phi),$$

are often of interest. The closed-form solutions for these densities that characterize uncertainties of interest, however, are only available in very limited settings.

In what follows, a class of state space models is described, namely, linear Gaussian SSM. In this special setting, several of the desired distributional quantities are readily computed.

2.4.1 Linear Gaussian state space models

The linear Gaussian SSM, also known as the dynamic linear model, consists of the joint regression equations

$$y_t = c_t + Z_t\alpha_t + \varepsilon_t \qquad \varepsilon_t \overset{ind}{\sim} \mathcal{N}(0, H_t) \tag{2.7}$$

$$\alpha_{t+1} = d_t + R_t\alpha_t + \eta_t \qquad \eta_t \overset{ind}{\sim} \mathcal{N}(0, Q_t) \tag{2.8}$$

that are each linear in α_t as well as having additive Gaussian disturbances ε_t and η_t respectively, and that are assumed to be temporally independent for all t. The vectors c_t and d_t are of dimensions k and l, being the same dimension as y_t and α_t respectively and Z_t and R_t are conformable matrices that are potentially relevant upon static parameters. For simplicity of exposition, we assume the measurement disturbance, ε_t, is independent of the state disturbance, η_t, for $t = 1, 2, \cdots, T$.

In the special case of linear Gaussian SSM and conditionally given a value of ϕ, the filtered and smoothed distributions can be derived analytically, with the imposing of the initial assumption that

$$\alpha_0 \sim \mathcal{N}(a_0, P_0).$$

Owing to the linear and additive structure of (2.7) and (2.8) along with the additive Gaussian error terms, the joint distribution of all observations and states is Gaussian. Exploiting the properties of the multivariate Gaussian conditional distribution, and marginal distributions of any state variable, α_t, conditional on any subset of observations, may be analytically obtained via a recursive update rule. In this setting, the progressive procedure to attain analytic results for the latent states is called the Kalman filter recursions (Kalman, 1960).

The idea of Kalman filtering is to revise the predicted mean and variance of α_t as y_t becomes available. Since a Gaussian distribution is fully characterized by its mean and variance, the update of the mean and variance is sufficient to fully update each distribution which occurs as follows. For each t, we denote respectively, the predictive mean and variance for α_t, conditional on $y_{1:t-1} = (y_1, \cdots, y_{t-1})$ by $a_{t|t-1}$ and $P_{t|t-1}$, i.e.,

$$a_{t|t-1} = \mathbb{E}(\alpha_t | y_{1:t-1})$$

$$P_{t|t-1} = \mathbb{V}(\alpha_t | y_{1:t-1})$$

and also denote the updated, or filtered, mean and variance for α_t, conditional on $y_{1:t}$, by $a_{t|t}$ and $P_{t|t}$, i.e.,

$$\begin{aligned} a_{t|t} &= \mathbb{E}(\alpha_t|y_{1:t}) \\ P_{t|t} &= \mathbb{V}(\alpha_t|y_{1:t}). \end{aligned}$$

Now, for given initializations, $a_0 = a_{0|0} = \mathbb{E}(\alpha_0)$ and $P_0 = P_{0|0} = \mathbb{V}(\alpha_0)$, the Kalman filter recursions, given by

$$a_{t|t-1} = d_{t-1} + R_{t-1}a_{t-1|t-1} \tag{2.9}$$

$$P_{t|t-1} = R_{t-1}P_{t-1|t-1}R'_{t-1} \tag{2.10}$$

$$v_t = y_t - c_t - Z_t a_{t|t-1} \tag{2.11}$$

$$F_t = Z_t P_t Z'_t + H_t \tag{2.12}$$

$$M_t = P_{t|t-1}Z'_t \tag{2.13}$$

$$a_{t|t} = a_{t|t-1} + M_t F^{-1} v_t \tag{2.14}$$

$$P_{t|t} = P_{t|t-1} - M_t F^{-1} M'_t, \tag{2.15}$$

for $t = 2, \cdots, T$. Having obtained via the Kalman filter each of the one-step ahead mean and variance, the likelihood function associated with the dynamic linear model may be expressed through a prediction-error decomposition, i.e.,

$$p(y_{1:T}|\phi) = p(y_1|\phi)\prod_{t=1}^{T} p(y_t|y_{1:t-1}, \phi), \tag{2.16}$$

with each component of (2.16) being a Gaussian pdf given by

$$\begin{aligned} p(y_t|y_{t-1}, \phi) &= p(v_t|\phi) \\ &= |F_t|^{\frac{1}{2}}(2\pi)^{\frac{-Tk}{2}} \exp\left\{-\frac{1}{2}v'_t F_t^{-1} v_t\right\}, \end{aligned}$$

for $t = 1, \cdots, T$, where v_t is given in (2.11) and F_t is given in (2.12). The result follows directly from the Kalman filter recursions which progressively introduce, and then marginalize over the uncertainty in each α_t analytically while conditional on each observed value of y_t. As a consequence, maximum likelihood (ML) inference for a vector of static parameters is straightforward, since the Gaussian likelihood function may be computed exactly, via (2.16), noting that both v_t and F_t are expressions of ϕ.

Also of interest are the smoothed marginal distributions given by $p(\alpha_t|y_{1:T}, \phi)$ for $t = 0, 1, 2, \cdots, T$. These marginal smoothed distribution have been fully updated and can help provide an understanding of the underlying state process in many applications. For a comprehensive discussion of the use of the marginal smoothed distributions for frequentist inference, refer to Durbin and Koopman (2012).

The focus in Bayesian inference, however, is to obtain samples from the joint smoothed distribution, namely pdf $p(\alpha_{0:T}|y_{1:T})$. In this linear Gaussian setting, it is possible to draw $\alpha_{0:T} = \alpha_0, \cdots, \alpha_T$ from its joint posterior distribution in a single block. This block sampling algorithm is referred to as *forward filtering, backward sampling* (FFBS) in the MCMC literature and is attributed to both Carter and Kohn (1994) and Frühwirth-Schnatter (1994). This result is due to the fact that the density of the joint distribution can be decomposed into the product of conditional densities

$$p(\alpha_1, \cdots, \alpha_T|y_{1:T}, \phi) = p(\alpha_T|y_{1:T}, \phi)p(\alpha_{T-1}|\alpha_T, y_{1:T}, \phi) \cdots p(\alpha_0|\alpha_{1:T}, y_{1:T}, \phi), \tag{2.17}$$

with each of the densities in the decomposition shown in (2.17) generated in reversed order, for $t = T, T-1, \cdots, 1, 0$, recursively as follows. Having completed a forward pass through the Kalman filter, as discussed above, the smoothed distribution for the terminal state, $p(\alpha_T|y_{1:T}, \phi)$, is immediately available as it is the same as the filtered distribution at final time T. Hence a draw of α_T denoted as α_T^*, may be obtained. Thus, through this initialization at time $t = T$, at each subsequent time, the previously obtained filtered distribution, proceeded with time $t-1$, i.e., $p(\alpha_{t-1}|y_{1:t-1})$, is now updated with the draw α_t^* via the transition density $p(\alpha_t^*|\alpha_{t-1})$. This yields the revised pdf $p(\alpha_{t-1}|\alpha_t^*, y_{1:T})$ from which a draw α_{t-1}^* may be obtained. Specifically, the backward sampling algorithm for the state variables $\alpha_{0:T}$ given the means and variances, $a_{t|t}$ and $P_{t|t}$, respectively, from the filtered distributions which have already been produced, via the Kalman filter and the backward sampling, proceeds as follows:

The backward sampling algorithm

1. *For* $t = T$, *generate* α_T^* *from*

$$\alpha_T|y_{1:T}, \phi \sim \mathcal{N}(a_{T|T}, P_{T|T})$$

2. *For* $t = T-1, T-2, \cdots, 1, 0$, *sequentially generate* α_t^* *from*

$$\alpha_t|\alpha_{t+1}^*, y_{1:T}, \phi \sim \mathcal{N}(a_{t|T}, P_{t|T}),$$

where

$$\begin{aligned} a_{t|T} &= a_{t|t} + P_{t|t}R_t'P_{t+1|t}^{-1}(\alpha_{t+1}^* - a_{t+1|t}) \\ P_{t|T} &= P_{t|t} + P_{t|t}R_t'P_{t+1|t}^{-1}R_tP_{t|t}. \end{aligned}$$

According to this procedure, the FFBS algorithm efficiently generates a joint draw of $\alpha_{0:T}^*$ from the joint conditional posterior distribution $p(\alpha_{0:T}|y_{1:T}, \phi)$. In conjunction with an additional step to generate a draw of ϕ conditionally given $\alpha_{0:T}^*$ and $y_{1:T}$, a two-stage Gibbs sampling approach will produce an ergodic Markov chain whose stationary distribution is the relevant joint posterior distribution of all unknowns.

In the following subsection, we introduce a class of SSMs whose state variable may take on only a finite number of distinct values.

2.4.2 Markov Switching State Space Model

Arguably, the most prominent illustration of the use of this SSM, at least in the economics and finance literature, is the Markov switching model of Hamilton (1989). It has been extensively employed to analyze business cycle and volatility, see, for example, Kim and Nelson (1999a), Eo and Kim (2016) and Chen, So, and Lin (2009) among others. This model permits multiple dependence structures that encompass distinct time series behaviors in different regimes. As the name suggests, the switching mechanism is governed by a state variable that follows a first order Markovian structure, featuring the fact that the current value of the state depends on its immediate past. To illustrate the idea, we consider a two state Markov switching model for the conditional mean in a linear model setting. Let y_t be a univariate response variable and X_t be a p-dimensional vector of covariates, both indexed by $t = 1, 2, \cdots, T$. The observation at time t is extended given the state variable S_t as

$$y_t = X_t \beta_{s_t} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

where $s_t = 0, 1$ may equal zero or one. The value of S_t is unobserved but assumed to follow a first-order Markov process with transition matrix given by

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{10} \\ p_{01} & p_{11} \end{bmatrix}, \tag{2.18}$$

where $p_{00} = \Pr(S_t = 0 | S_{t-1} = 0)$, $p_{11} = \Pr(S_t = 1 | S_{t-1} = 1)$, $p_{10} = 1 - p_{00}$ and $p_{01} = 1 - p_{11}$. This Markov process typically initializes with its long-term (ergodic) probabilities given by $\Pr(S_0 = 0) = p_0$ and $\Pr(S_0 = 1) = 1 - p_0$. Note that the normality assumption imposed on the disturbance term ε_t in (2.4.2) is not essential but is convenient as it simplifies our discussion here.

Given the first-order Markov transition matrix in (2.18) and the initial state probability p_0, the joint density of all states $S_{0:T}$ and observations $y_{1:T}$, given the static parameter $\phi = (\beta_0, \beta_1, \sigma_\varepsilon^2, p_{00}, p_{11})$ is given by

$$\Pr(S_{0:T}, y_{1:T}|\phi) = \Pr(S_0 = s_0|\phi) \prod_{t=1}^{T} \left[\Pr(S_t = s_t|S_{t-1} = s_{t-1}, \phi) p(y_t|s_t, \phi)\right].$$

As in the case detailed in Section 2.4.1, if the filtered distribution for each of the latent states, here $S_0, S_1, \cdots, S_T$, can be obtained, the calculation of the exact likelihood function is feasible. The resulting filtering algorithm often referred to as the Hamilton filter, which is used to evaluate the likelihood function for any given value of the static parameter $\phi = (\beta_0, \beta_1, \sigma_\varepsilon^2, p_{00}, p_{11})$.

The idea of the Hamilton filter is to compute the filtering probabilities, $\Pr(s_t = i|y_{1:t}, \phi)$ for each value $i = 0$ and $i = 1$. Those conditional probabilities are computable as S_t can only take on a finite number of values, here just two. With the initial values $\Pr(S_0 = i|\phi)$, for $i = 0$ and $i = 1$, the Hamilton filter progressively computes

$$\begin{aligned}
\Pr(S_t = 0|y_{1:t-1}, \phi) =& p_{00} \Pr(S_{t-1} = 0|y_{1:t-1}, \phi) + p_{10} \Pr(S_{t-1} = 1|y_{1:t-1}, \phi) \\
\Pr(S_t = 1|y_{1:t-1}, \phi) =& 1 - \Pr(S_t = 0|y_{1:t-1}, \phi) \\
p(y_t|y_{1:t-1}, \phi) =& \Pr(S_t = 0|y_{1:t-1}, \phi) p(y_t|y_{1:t-1}, S_t = 0, \phi) \\
& + \Pr(S_t = 1|y_{1:t-1}, \phi) p(y_t|y_{1:t-1}, S_t = 1, \phi) \\
\Pr(S_t = 0|y_{1:t}, \phi) =& \frac{\Pr(S_t = 0|y_{1:t-1}, \phi) p(y_t|y_{1:t-1}, s_t = 0, \phi)}{p(y_t|y_{1:t-1}, \phi)} \\
\Pr(S_t = 1|y_{1:t}, \phi) =& 1 - \Pr(S_t = 0|y_{1:t}, \phi)
\end{aligned}$$

for $t = 1, 2, \cdots, T$. Hence given data values $y_{1:T}$, the filtered probabilities $\Pr(S_t = i|y_{1:t})$, for $i = 0$ or 1, may be calculated, for any given value of ϕ.

It is also straightforward to simulate adversely the state variables $s_{0:T}$ from the joint posterior distribution using a backward sampling algorithm. Note

$$\Pr(S_t = i|y_{1:T}, s_{t+1} = j, \phi) = \frac{p_{ij}\Pr(S_t = i|y_{1:t}, \phi)}{\Pr(S_{t+1} = j|y_{1:t}, \phi)}, \tag{2.19}$$

for any $i = 0$ or 1 and any $j = 0$ or 1. Hence given the filtered probabilities $\Pr(S_t = i|y_{1:t}, \phi)$, for $i = 0, 1$, the transition probabilities p_{ij} for $i = 0, 1$ and $j = 0, 1$, as well as a draw of $S_{t+1} = s^*_{t+1}$ from $\Pr(S_{t+1} = s_{t+1}|y_{1:T}, \phi)$, a draw of S_t from (2.19) can be obtained. Now, given a draw of $S_{0:T}$, it is easy to recognize that under independent and conditionally conjugate priors, draws of the parameters β_0, β_1 and σ^2_ε may all be obtained. Finally, under independent and conditionally conjugate Beta priors, $\mathcal{B}(\bar{u}_{0,0}, \bar{u}_{0,1})$ and $\mathcal{B}(\bar{u}_{1,1}, \bar{u}_{1,0})$, for p_{00} and p_{11} respectively, given these regression parameters, the latent states $S_{0:T}$ and the data $y_{1:T}$, draws of the parameters in the Markov transition matrix may be sampled from Beta distributions, with

$$p_{00}|s_{0:T} \sim \mathcal{B}(\bar{u}_{0,0} + n_{0,0}, \bar{u}_{0,1} + n_{0,1}) \text{ and}$$
$$p_{11}|s_{0:T} \sim \mathcal{B}(\bar{u}_{1,1} + n_{1,1}, \bar{u}_{1,0} + n_{1,0}),$$

where $\mathcal{B}(a, b)$ denotes the Beta distribution on $(0, 1)$, having mean $\frac{a}{a+b}$ and variance $\frac{ab}{(a+b)^2(a+b+1)}$, here with

$$n_{1,0} = \sum_{t=1}^{T}(S_t|S_{t-1} = 0),$$
$$n_{1,1} = \sum_{t=1}^{T}(S_t|S_{t-1} = 1),$$

and $n_{0,0} = n_{S_t=0} - n_{1,0}$ and $n_{0,1} = n_{S_t=1} - n_{1,1}$. Accordingly, denote $\boldsymbol{\theta}_{/\beta}$ all the elements contained in the parameter vector $\boldsymbol{\theta}$ except β for any β, where β is a generic notation. Then, the backward sampling algorithm is summarized in Algorithm 7.

Algorithm 7 Gibbs sampling algorithm for the Markov switching model

```
Inputs: $y_{1:T}$: data observations; $G$: number of iterations; $\theta^{(0)} = \left(p_{00}^{(0)}, p_{11}^{(0)}, \beta_0^{(0)}, \beta_1^{(0)}, \sigma_\varepsilon^{2(0)}\right)$: initial value; $p(p_{00}|s_{0:T})$, $p(p_{11}|s_{0:T})$, $p(\beta_1|s_{0:T}, y_{1:T}, \phi_{/\beta_1})$, $p(\beta_0|s_{0:T}, y_{1:T}, \phi_{/\beta_0})$, $p(\sigma_\varepsilon^2|y_{1:T}\phi_{/\sigma_\varepsilon^2})$: full conditional posterior distributions;
for $i = 1 \to G$ do
    For $t = 1, 2, \cdots, T$, compute the filtered distribution $p(S_t = i|y_{1:t}, \phi)$, then generate $S_{0:T}$ from
        $$S_{1:T} \sim p(S_{0:T}|y_{1:T}, \phi)$$
    using the Hamilton filter and a backward sampler;
    Generate the vector of static parameters $\phi$ by sampling each element from
        $p_{00} \sim p(p_{00}|s_{0:T})$,   $p_{11} \sim p(p_{11}|s_{0:T})$,   $\beta_0 \sim p(\beta_0|s_{0:T}, y_{1:T}, \phi_{/\beta_0})$,
        $\beta_1 \sim p(\beta_0|s_{0:T}, y_{1:T}, \phi_{/\beta_1})$   and   $\sigma_\varepsilon^2 \sim p(\sigma_\varepsilon^2|y_{1:T}, \phi_{/\sigma_\varepsilon^2})$;
Outputs: A sample of $G$ draws of $\theta$ and $S_{0:T}$ from $p(\theta, S_{0:T}|y_{1:T})$.
```

As illustrated with the previous two SSMs, simulating the state variables requires each of the predictive, filtered and smoothed distributions to have closed-form solutions. However, these quantities are not available when an SSM involves either nonlinear functional forms or non-Gaussian disturbance terms, or both. This situation arises often, including in the modeling setting in Chapter 3 and 4. In the following section, we extend the discussion of SSMs to a nonlinear setting, and provide a description of certain MCMC algorithms for posterior simulation that will prove to be useful for the work presented in the later chapters.

2.4.3 Nonlinear state space models

In the presence of nonlinear functions or non-Gaussian error terms, the Kalman filter and corresponding backward sampling algorithm detailed in Section 2.4.1 no longer provide a means to sample from the relevant (conditional) joint state distribution. In a non-Gaussian SSM setting, closed-form solutions of the filtering and smoothing state distributions are not available, except for in the special discrete Markov switching setting as demonstrated in Section 2.4.2. A variety of methods have been proposed to try to sample the latent state vector, such as the particle filtering methods and approximation based methods, have been proposed in a range of nonlinear settings. The former, also known as sequential Monte Carlo (SMC), which employs sequential importance sampling/resampling methods first introduced by Gordon, Salmond, and Smith (1993). In particular, as the particle filter produces an unbiased estimate of the likelihood function, so called pseudo-marginal approaches have been developed in an attempt to use SMC technique within an MCMC environment (see, for example, Beaumont, 2003,Andrieu and Roberts, 2009 and Pitt et al., 2012). An alternative approach to dealing with the nonlinear SSMs is used in this thesis. In particular, an approach using an auxiliary mixture model to approximate the relevant target sampling distributions from which latent variable draws are generated using an FFBS algorithm is used. Once the candidate draw has been generated from the approximating model, it is either accepted or rejected using a MH acceptance probability.

Most relevant to this approximation strategy is to approximate the filtered distributions via a mixture of Gaussian components. Specifically, consider the case where the model under study has a state space representation, given by

$$\begin{aligned} y_t &= c + Z\alpha_t + \varepsilon_t \qquad & \varepsilon_t &\overset{ind}{\sim} \mathcal{N}(0, H) \\ \alpha_t &= d + g(\alpha_{t-1}) + \eta_t \qquad & \eta_t &\overset{ind}{\sim} \mathcal{N}(0, Q). \end{aligned}$$

Notice that here the nonlinearity occurs only in the state transition equation via the function $g(\cdot)$. Note also that even though the measurement equation is linear in α_t and the measurement and state disturbances ε_t and η_t are assumed to be Gaussian, the Kalman filter will not produce the correct filtered distribution since a nonlinear transformation (e.g., $g(\cdot)$) of a Gaussian random variable (e.g., α_{t-1}) is, in general, not Gaussian. By way of notation to simplify the discussion, we suppress any explicit dependence on the static parameter ϕ.

The approximating auxiliary mixture model approach used here was introduced by Stroud, Müller, and Polson (2003), hereafter referred as SMP. In the current setting, this approach is to replace the nonlinear model with Gaussian mixture model. The choice of the approximating Gaussian component is undertaken in an adaptive way, depends on the location of the state vector to be sampled. To facilitate the sampling, a collection of auxiliary mixture indicator variables is used. A important feature of this approach is that when conditioned on the relevant mixture indicators, the auxiliary model reduces to a linear Gaussian SSM and as such that a standard FFBS algorithm may be applied to produce a candidate draw from the joint posterior distribution of state variables. In particular, the SMP method employs a collection of predetermined grid points denoted by, $\{\bar{\mu}_1, \bar{\mu}_2, \cdots, \bar{\mu}_K\}$, each from the support of α_t, to construct a set of local linear approximations of the nonlinear function at given grid points, $\{\bar{\mu}_1, \bar{\mu}_2, \cdots, \bar{\mu}_K\}$. Each grid point, $\bar{\mu}_j$, corresponds to a state-dependent mixture weight, $p^a(u_t|\alpha_t)$, through

$$p^a(u_t = j|\alpha_t) = \psi(\alpha_t; \bar{\mu}_j, \bar{\sigma}_j^2) / \sum_{k=1}^{K} \psi(\alpha_t; \bar{\mu}_k, \bar{\sigma}_k^2),$$

where u_t denotes the mixture indicator variable attached to α_t, that can take values from $\{1, 2, \cdots, K\}$, and where $\psi(\alpha_t; \bar{\mu}_k, \bar{\sigma}_k^2)$ denotes a Gaussian density with mean $\bar{\mu}_k$ and variance $\bar{\sigma}_k^2$. The local linearization of the nonlinear function g are produced bu undertaken a first order Taylor series expansion of g around each $\bar{\mu}_k$ for $k = 1, \cdots, K$. In particular, we define

$$b_k = \frac{\partial g(\bar{\mu}_k)}{\partial \alpha_t} \qquad a_k = g(\bar{\mu}_k) - b_k\bar{\mu}_k \tag{2.20}$$

and hence define, conditional on the the mixture indicator, u_t, the mixture component by

$$p^a(\alpha_{t+1}|\alpha_t, u_t = k) \sim \mathcal{N}(a_k + b_k\alpha_t, Q),$$

resulting the approximating conditional measurement model given by

$$p^a(y_t|\alpha_t) = \sum_{k=1}^{K} p^a(y_t|\alpha_t, u_t = k)p^a(u_t = k|\alpha_t).$$

Following SMP, a hybrid Gibbs based algorithm is used to simulate from the posterior distribution of the state variables, $p(\alpha_{0:T}|y_{1:T})$. This algorithm is summarized in the following as Algorithm 8. Note that in Step 3 of of the SMP algorithm, the mixture indicators are easily sampled owing to the fact that each u_t can take on only a limited number of possible values. Once a vector of state $\alpha_{0:T}$ is generated from its fill conditional posterior distribution, an additional step can be taken within MCMC to update the vector of static parameters, ϕ given $\alpha_{0:T}$ and $y_{1:T}$.

An essential aspect required of the SMP approach is that the auxiliary mixture model approximate the underlying nonlinear SSM well. As the state vector is typically of high dimension, the approximation error could accumulate rapidly, resulting in an MCMC chain that moves very slowly due to a low MH acceptance probability. This problem is amplified when the nonlinear function is unknown, as is the case in Chapter 4. This case also requires an alternative approach to select the local linearization parameters a_k and b_k.

2.5 Conclusion

In this chapter, we have reviewed some key elements of Bayesian inference that are most relevant to the contributions made in this thesis. As is evident by now, while th posterior distribution is determined by the prior specification and the form of the likelihood function, its evaluation may rely heavily on a

Algorithm 8 The SMP algorithm

```
Inputs: $y_{1:T}$: data observations; $G$: number of iterations; $\{\bar{\mu}_1, \bar{\mu}_2, \cdots, \bar{\mu}_K\}$: a set of grid points; $\{\bar{\sigma}_1, \bar{\sigma}_2, \cdots, \bar{\sigma}_K\}$, $\{a_1, a_2, \cdots, a_K\}$, $\{b_1, b_2, \cdots, b_K\}$: tuning parameters;
for $i = 1 \to G$ do
    Generate the mixture indicator variables $u_{1:T}$ from
        $p^a(u_{1:T}|y_{1:T}, \alpha_{1:T}) \propto \prod_{t=0}^{T-1} p^a(\alpha_{t+1}|\alpha_t, u_t) p^a(u_t|\alpha_t);$
    Generate the vector of candidate state variables, $\alpha^*_{0:T}$ from the full conditional posterior distribution under the auxiliary mixture model, via FFBS,
        $p^a(\alpha_{0:T}|u_{1:T}, y_{1:T}) \propto p(\alpha_0) \prod_{t=1}^{T} p(y_t|\alpha_t) p^a(\alpha_{t+1}|\alpha_t, u_t) p^a(u_t|\alpha_t);$
    Accept the candidate draw with the Metropolis-Hastings probability
        $\alpha(\alpha^*_{0:T}, \alpha^{(i-1)}_{0:T}) = 1 \vee \prod_{t=0}^{T} \frac{p(y_t|\alpha_t^{(i-1)}) p^a(\alpha_{t+1}^{(i-1)}|\alpha_t^{(i-1)})}{p(y_t|\alpha_t^*) p^a(\alpha_{t+1}^*|\alpha_t^*)} \times \frac{p(y_t|\alpha_t^*) p^a(\alpha_{t+1}^*|\alpha_t^*)}{p(y_t|\alpha_t^{(i-1)}) p^a(\alpha_{t+1}^{(i-1)}|\alpha_t^{(i-1)})};$
Outputs: A sample of $G$ draws of $\alpha_{0:T}$ jointly from $p(\alpha_{0:T}|y_{1:T}, \phi)$.
```

computable approach, such as MCMC. The development of an effective and efficient MCMC algorithm in any given situation – particularly those where high-dimensional and time-dependent latent variables are involved – requires care. The building blocks contained in the current chapter underpin the new methodologies proposed for the more complex models presented in Chapters 4 and 5.

Chapter 3

The Determinants of Bank Loan Recovery Rates in Good times and Bad - New Evidence

In the chapter, we find that factors explaining bank loan recovery rates vary depending on the state of the economic cycle. Our modeling approach incorporates a two-state Markov switching mechanism as a proxy for the latent credit cycle, helping to explain differences in observed recovery rates over time. Using US bank default loan data from Moody's Ultimate Recovery Database and covering the pre-and post-GFC period, this paper develops a dynamic predictive model for bank loan recovery rates, accommodating the distinctive empirical features of the recovery rate data while incorporating a large number of possible determinants. We find that the probability of default and certain loan-specific and other variables hold different explanatory power with respect to recovery rates over 'good' and 'bad' times in the credit cycle, meaning that the relationship between recovery rates and certain loan characteristics, firm characteristics and the probability of default differs depending on underlying credit market conditions. Our findings demonstrate the importance of accounting for countercyclical expected recovery rates when determining capital retention levels.

3.1 Introduction

Loan defaults are inevitable events within a bank's loan book. Credit risk management processes require banks to accurately model loan default probabilities and subsequent recovery rates (RRs, hereafter). These models are a key compliance requirement for banks subscribing to the Advanced Internal Ratings Based (AIRB) models. Furthermore, the latest International Financial Reporting Standards on Financial Instruments (IFRS 9) in particular requires entities to reflect on a default probability based on best available forward looking information. An accurate understanding of RR performance over time is critical for banks, and could potentially result in more efficient use of capital. In this paper, we study the interaction of borrower characteristics, loan features and macroeconomic conditions together with other key criteria with respect to probability of default (PD, hereafter) and RRs across credit cycles.

Several previous studies have investigated the determinants of RRs. See, for example, Altman et al. (2005), Acharya, Bharath, and Srinivasan (2007), Bruche and Gonzalez-Aguado (2010), Khieu, Mullineaux, and Yi (2012) and Altman and Kalotay (2014). However, the systemically time-varying PD and RR response to different credit and economic cycles does not appear to have been captured. Most studies assume a constant association between PD and RR, potentially leading to an inaccurate assessment of RR risk (Resti, 2002; and Altman et al., 2005).

In addition, most of the contemporary literature concerning RR determinants does not focus specifically on bank loans. Altman and Kalotay (2014), using the same set of determinants to study bank loans and corporate bonds, combine bank loans with corporate bonds, whereas Mora (2015) investigates corporate bonds only. Bank loans are fundamentally different to other securities; typically, they are senior to traded corporate debt. Due to the different repayment hierarchy, this tends to make bank loan RRs higher. Furthermore, a bank generally has much greater access to customers' financial information than other types of investor, forcing covenant compliance if any financial ratios or loan covenants are breached. Therefore, given their access to non-public information, banks are more likely to enforce bankruptcy than other key stakeholders, and hold more power over borrower firms with respect to RRs. Additionally, banks may gain access to underlying assets as their fixed/floating charges allow them to be paid before other creditors.

Khieu, Mullineaux, and Yi (2012) studied bank loans, but do not examine RRs through the 2007/08 financial crisis and beyond. They also impose parametric assumptions/constraints which are embedded within the models they employ. The limitation of such an approach is that it assumes distributions for the data that may be quite different from observed RRs. Furthermore, a quasi-likelihood method is employed. The RRs are modelled using a Bernoulli likelihood that does not naturally accommodate observed RRs that fit inside the unit interval. This approach is replicated by Khieu, Mullineaux, and Yi (2012), who also employ a linear modelling approach where the errors are effectively assumed to be normal - an assumption contrary to the observed RR distributions.

In view of the above, this paper develops a dynamic predictive model for bank loan RRs, allowing for good and bad times. This enables us to ascertain whether variable relations are constant over time, while accounting for the distinctive multi-modal shape found in the empirical distribution of the data. We also account for a range of other relevant factors in a dynamic framework, as such variables have only previously been considered as RR predictors in static contexts. Here, however, the predictors are conditional upon the underlying economic and credit cycle, which in turn we characterize in line with Bruche and Gonzalez-Aguado (2010) with two distinct states - good and bad.

Utilizing data from Moody's Ultimate Recovery Database, this study focuses on defaulted bank loans between 1987 and 2015 in the United States (US).

In order to manage the latent economic states and the flexible nature of the empirical RR distribution, a Bayesian inferential methodology is developed, exploiting the hierarchical structure, along the lines of Kim and Nelson (1999b). Moreover, due to concerns regarding the large number of available predictors, and the fact that many of these regressors may be correlated, we incorporate a LASSO prior for the regression components. The inferential results from the dynamic model are subsequently compared to the available static versions, with new insights reported, contributing further to the literature.

Overall, we find more significant loan characteristics during good times. Conversely, during bad times, only certain collateral determinants are related to RRs. This finding reinforces the notion that not all of a firm's assets facilitate a full loan recovery, with inventory and accounts payable more likely to achieve such an objective. This has consequences for discounts that banks apply to assets offered as collateral. The size of the discount should, we argue, not only depend on the riskiness and liquidity of the assets being offered, but also on the credit cycle.

The type of recovery process is not directly related to RRs in either cycle, however, when the time to emerge is considered alongside prepackaged recovery processes, some interesting results are found. Banks are found to have lower RRs during a bad cycle when there is no prepackaging. Conversely, recoveries in a good cycle, result in a higher RR, suggesting banks need to be mindful of the likely resolution time involved during such processes at the origination time.

We find RRs are significantly affected by the condition of the economy's countercyclicality and the borrowers' characteristics. Khieu, Mullineaux, and Yi (2012) report no association; however, as we control for cycles, we find larger firm size is associated with reduced RRs during bad times and the opposite during good times. Our finding highlights the importance of carefully considering the definition of idiosyncratic risk, particularly the tail risk of the bank loan loss distribution. During bad times, banks need to give greater weight to their tail-risk forecasts, covering any unexpected losses from their large customers by allocating further economic capital. Once again, such findings have an impact on IFRS 9 implementation with respect to forward-looking judgments.

Finally, in line with Khieu, Mullineaux, and Yi (2012), we use the all-in-spread (AIS) measure as our proxy for PD and report a statistically negative association only during bad cycles. PD and RR have historically been found to be negatively related. Although several important advances have previously been made using credit risk modelling techniques, at the individual loan level the relation between PD and RR has remained an open question. Our study provides evidence of a negative relationship during bad times only, suggesting the presence of systemic time variation in this context.

Overall, such findings have a direct impact on a bank's loan loss distribution, which is a vital component for determining capital allocation. The default loan recovery process suggests loan, recovery, borrower, economic and PD features need to be dynamically managed for banks to optimally allocate capital across credit cycles. Clearly then, debt recovery is time-varying and the potential risk of not accurately addressing such variables in the credit risk process will lead to potentially under- or over-providing for future loan losses.

These findings support the Basel III framework's recommendations for the use of countercyclical buffers, creating an environment where the banking sector is protected from periods of excessive aggregate credit growth. So capital buffers assist against such a build-up phase and help the banks' going concern when RRs underperform. Conversely, during bad times, capital buffers are essential, as the supply of credit may be curtailed by regulatory capital requirements. Furthermore, throughout bad times, the banking system may also experience further unexpected loan losses emanating from lower RRs. This is a major issue for reporting entitles, as in line with IFRS 9, they are required to provide forward-looking judgments.

Therefore, if RRs may be accurately forecasted during different cycles, capital buffers would be an effective tool to address the countercylicality nature of the economy, managing the complex environment of absorbing any unexpected credit losses. This view ensures the banking sector applies appropriate prudential practices, including maintaining capital requirements and controls for banks to operate within a bad cycle, but flexible enough to adjust accordingly during better periods. This has important implications for the pro-cyclicality effects of credit risk models, particularly the larger banks using an AIRB approach.

The remainder of our paper proceeds as follows. In Section 3.2, we review the literature determining RRs and the proposed econometric specifications for the model calibrated with and without the latent credit cycle variable. Section 3 contains a description of the Moody's dataset used for the empirical analysis. The Bayesian inferential framework and corresponding MCMC simulation algorithms are then detailed in Section 3.4. This is followed by the results of our analysis and evaluation of our proposed models and their predictive capability in Section 3.5. Section 3.7 concludes with a discussion and suggested directions for future research.

3.2 Literature review

Financial market participants, including bank regulators, are increasingly concerned with the management of risky assets, particularly bank loans. It is crucial to consider risk factors and market conditions at the time of placing an investment, but how these factors vary over time is also becoming increasingly viewed as critically important for making lending decisions. It is of particular importance to be mindful of the PD and subsequent loan recovery prospects during different or extreme conditions, as this will be critical to achieving expected return to finance providers, such as banks. The consequences of not considering this state- dependent risk can be severe for a bank, and may reach far beyond manageable levels, as was demonstrated in the financial crisis of 2008. This profound failure in prudential regulation, and corporate governance, attributable to poor operational risk management practices (Financial Crisis Inquiry Commission, 2011), underscores the importance of understanding financial risk, and in particular, credit risk, when pricing corporate loan contracts.

When a corporate borrower defaults, the lender endeavours to recover the outstanding debt using available collateral and liquidity mechanisms. Default alone, while not ideal for the borrower in terms of their ability to establish or maintain a strong credit rating, does not imply that the outstanding balance of the loan cannot ultimately be fully recovered during the post default period. In the vast majority of cases, lenders recover their full entitlement. However, there are many occasions when none or only a part of the outstanding indebted balance is recovered. Not surprisingly, the projected RR in the event of default is one of the key inputs when determining the price of any credit-related financial contract, including the value of the fundamental investment itself. Loss Given Default (LGD) is defined as one minus the RR, where the RR represents the proportion of the borrowed funds recovered (referred to as the exposure at default, or EAD) after the borrower goes into default. Hence, it is important for lenders to understand the factors that affect the actual RR, so appropriate decisions, including loan terms, can be made. Further motivation is provided by regulators, who require financial institutions to show evidence of prudential planning and modelling, and to ensure regulated capital requirements, as specified by the Basel III or governmental financial regulators, are maintained.

There are three main variables that determine the credit risk of a credit-based financial contract: the PD, the RR in the event of default and exposure at default (EAD), which is the total value to which the lending institution is exposed. Altman, Resti, and Sironi (2004) points out that while significant attention has been paid to PD, RR and its apparent inverse relation with PD has attracted less attention. Notably, the RR is often treated as a constant variable, independent of PD. Existing studies have documented some empirical irregularities in the observed RR distribution. Schuermann (2004) finds that the concept of average recovery, a quantity often reported by rating agencies, is potentially very misleading, as the recovery distribution is restricted to exist over the unit interval. This restriction implies that the lender cannot lose or recover more than the outstanding amount at the time of default. Notably, the observed RR distribution is typically U-shaped, with the largest relative frequency occurring near or at unity, a non-negligible mass around zero, and a spread of RRs observed across the interval itself. A flexible nonlinear model is more appropriate to reflect the relationship of this RR distribution with a number of loan or firm characteristics.

Additionally, a range of econometric methods have been previously used to study RRs, including ordinary least squares (OLS), beta regression and a quasi-maximum likelihood estimation (QMLE) and a fractional regression methodology, as explored by Gupton et al. (2002), Acharya, Bharath, and Srinivasan (2007) and Khieu, Mullineaux, and Yi (2012), respectively. Such analyses provide further insight into the possible determinants of a loan's expected RR. Nevertheless, most approaches have some shortcomings. Notably, standard OLS ignores the unique distributional aspects of the observed RRs, despite the fact that the resulting RR values predicted from the model need not be bounded between zero and one. It also assumes constant marginal effects for each of the explanatory variables, a feature that is also unlikely given the constrained RR distribution. Furthermore, while the beta distributions underpinning a beta regression framework covers some variation of distributional densities over the unit interval, they cannot simultaneously accommodate a relative frequency mass in the middle of the unit interval with the relative large frequency masses observed around zero and one (De Servigny and Renault, 2004). Finally, while the model underlying the QMLE-based approach accommodates the constraints of the observed RR values, it does so at the expense of a coherent distribution model, fitting as it does a model for (binary) Bernoulli observations when in fact the values of RR may also lie inside the zero to one range, with clustering at 0 or 1.

The unsatisfactory features of the existing parametric approaches in the RR context have led to the development of more flexible models. Nonparametric methods have been shown to sometimes outperform their parametric counterparts in terms of accommodating non-linear relations between observed RRs and certain conditioning variables (Qi and Zhao, 2011). However, Bastos (2010) and Qi and Zhao (2011) find such flexible predictive models are more likely to over-fit the data and do not tend to work well in predicting future defaulted loan recoveries. Similarly, regression trees can become overly large and appear to produce results sensitive to assumed distribution and the dataset used. For example, Bastos (2010) and Qi and Zhao (2011) report very distinctive trees based on different datasets.

Modelling the determinants of RRs has shown them to be a function of individual loan characteristics, firm characteristics or fundamentals, industry variables, recovery process variables and macroeconomic factors. However, Altman et al. (2005) also demonstrates a negative association between an aggregate measure of the underlying default rate over a given period and the average RR, suggesting that changes in the underlying credit environment can also impact RRs. Hu and Perraudin (2002) observe a similar negative relationship from data covering the period 1971-2000. In response, Bruche and Gonzalez-Aguado (2010) define a two-state latent credit cycle variable and suggests that a 99% credit VaR (the Value at Risk for a portfolio consisting of bank loans and corporate bonds) is underestimated by more than 1.5% of the total outstanding amount if the credit cycle is omitted. The focus of this current research is to enhance the understanding of RR for corporate loans, addressing the limitations of the aforementioned literature. This is achieved by investigating RR in relation to appropriate firm and loan variables, in a time variant framework that allows for different economic circumstances, including major shocks such as the financial crisis of 2008. We use a less restrictive Bayesian, non-linear approach, accounting for time variation in the relationship between RR and PD.

3.3 Data from Moody's Ultimate Recovery Database (1987- 2015)

We investigate RRs from Moody's Ultimate Recovery Database over the period 1987 through to 2015. A suitable RR variable is selected, with associate loan recovery processes and borrower characteristics extracted from the same dataset. In addition, we obtain macroeconomic and industry variables, such as a credit spread variable as a proxy for the PD, to be used as the RR determinants. The definitions, data sources and features of these determinants are provided in Section 3.3.1, following information about the RR variable.

3.3.1 Data description

The recovery data and other information about the defaulting firms and instruments are extracted from Moody's Ultimate Recovery Database, resulting in a set of 1,611 defaulted bank loans of US firms originated by an array of syndicated lending institutions over the period 1987 through 2015. Consistent with other empirical studies within the recovery literature (Khieu, Mullineaux, and Yi, 2012 and Altman and Kalotay, 2014), we use Moody's Discounted Recovery Rate variable as the relevant empirical RR measure.

A complete set of loan recovery determinants associated with each of the observed RRs, as used by Khieu, Mullineaux, and Yi (2012), is also employed. Broadly speaking, the available determinants address: loan characteristics, recovery process, borrower characteristics, as well as macroeconomic, industry and PD. The observed determinants relating to borrower characteristics are obtained through manual matching of these firms with Standard & Poor's Compustat firms based on both CUSIP numbers and also firm names. Definitions for each of these determinants are provided in Table 3.6. Typically, the defaulted loans have debt values, at the time of default, of greater than \$50 million. This information, provided in Section 3.3.2, follows discussion of the key features of the RR data.

3.3.2 Recovery rates and their determinants

The RR variable we use is defined as the nominal settlement recovery amount discounted back from each settlement instrument's trading date to the last date cash was paid on the individual defaulted instrument, using the instrument's own effective interest rate. This key recovery measure takes account of the time value of money for the effective settlement period. The sample here, covering 1987 through 2015, is split between term loans (48%) and revolvers (52%), i.e., loans that can be repaid and re-drawn any number of times within a term. About 7.6% of the recoveries have some type of reorganization plan that shareholders have approved prior to, or at the time of, the bankruptcy filing.

Frequency plots (histograms) of both the raw RR and its transformed values (corresponding to (3.3.1)) for the sample period are shown in Figure 3.1. Note the data concentration on the extreme right boundary of both graphs. While extreme modes associated with RR values at zero and unity are present, the mode at zero is almost negligible compared to that associated with full recovery. This feature is in line with the tendency towards left skewness typically associated with RRs for bank loans Altman and Kalotay (2014). Given that the observed distribution of the RR is neither symmetric nor unimodal, the use of the average or median recovery as a single summary measure for the entire distribution is potentially very misleading. In particular, for this sample the average RR is 80.8%, while the median value is 100%, indicating that the distribution is indeed skewed to the left. It can be seen that both boundaries of the distribution have attracted concentrations, corresponding to the extremes of no recovery on the left and full recovery on the right, although the right boundary behavior dominates.

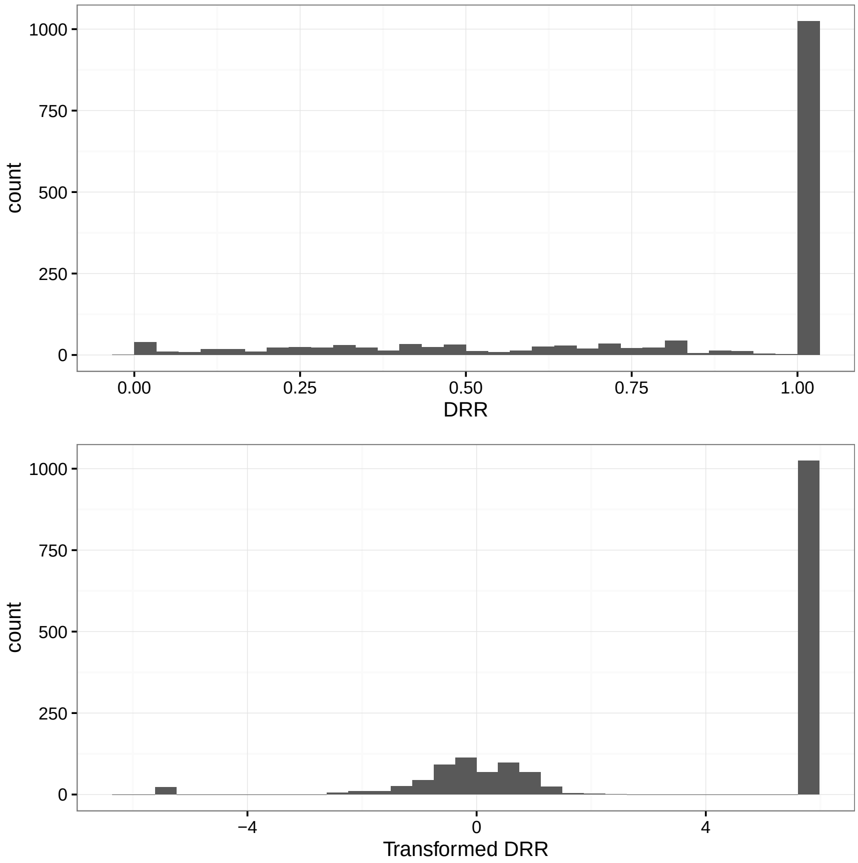

FIGURE 3.1: Histograms of the discounted recovery rates (RR) (top panel) and of the transformed discounted RRs ($\mathbf{y}$) (bottom panel), 1987-2015.

Details of each of the RR determinants considered are described in Table 3.6. Additional descriptive statistics for the determinants are provided in Table 3.1, along with those for the observed RR data. Overall, the summary statistics of the RR determinants for the sample period here are similar to those found in Emery (2007) and Khieu, Mullineaux, and Yi (2012), suggesting that in aggregate the data here are, by and large, in line with those of earlier studies. Furthermore, as per Khieu, Mullineaux, and Yi (2012), the firm

Variable	***n***	**Mean**	**Median**	**1st Qu**	**3rd Qu**	**Min**	**Max**	**Binary**
RR	1611	0.81	1.00	0.66	1.00	0.00	1.00	N
Loan characteristics								
(1) LOANSIZE($M)	1611	224.5	96.0	35.0	208.5	1.0	11150.0	N
(2) LOANTYPE	1611	0.48	0	0	1	0	1	Y
(3) LOANTYPE × FIRMSIZE	1611	813.5	0.0	0.0	692.3	0.0	60631.9	N
(4) ALLASSETCOLL	1611	0.62	1	0	1	0	1	Y
(5) INVENTRECIVECOLL	1611	0.10	0	0	0	0	1	Y
(6) OTHERCOLL	1611	0.17	0	0	0	0	1	Y
Recovery process characteristics								
(7) PREPACK	1611	0.08	0	0	0	0	1	Y
(8) RESTRUCTURE	1611	0.13	0	0	0	0	1	Y
(9) OTHERDEFAULT	1611	0.01	0	0	0	0	1	Y
(10) TIMETOEMERGE	1611	13.49	9.67	2.59	18.42	0	156.33	N
(11) TIMETOEMERGE^2	1611	427.17	93.51	6.682	339.11	0.00	24439.07	N
(12) PREPACK × TIMETOEMERGE	1611	0.21	0.00	0.00	0.00	0.00	11.87	N
Borrower characteristics								
(13) FIRMSIZE	1611	1654.8	665.5	227.4	1365.7	0	60631.92	N
(14) FIRMPPE	1611	0.54	0.44	0.13	0.82	0	9.73	N
(15) FIRMCF	1611	0.16	0.09	0.05	0.14	0	23.48	N
(16) FIRMLEV	1611	1.08	0.94	0.77	1.26	0	4.9	N
(17) EVERDEFAULTED	1611	0.15	0	0	0	0	1	Y
Macroeconomic and industry conditions								
(18) GDP	1611	2.61	2.80	0.98	4.09	0.06	4.79	N
(19) INDDISTRESS	1611	0.18	0	0	0	0	1	Y
Probability of default								
(20) AIS	1611	0.04	0.03	0.02	0.04	0	0.3	N

Table 3.1: Descriptive statistics of the discounted RR and of the determinants of bank loan recoveries, by determinant category. By column corresponding to the variable indicated in the far left-hand column, the following statistics are reported: sample size (n), sample mean (Mean), sample median (Median), 25% quantile (1st Qu), 75% quantile (3rd Qu), sample minimum (Min), sample maximum (Max) and whether variable is binary [Y] or not [N] (Binary).

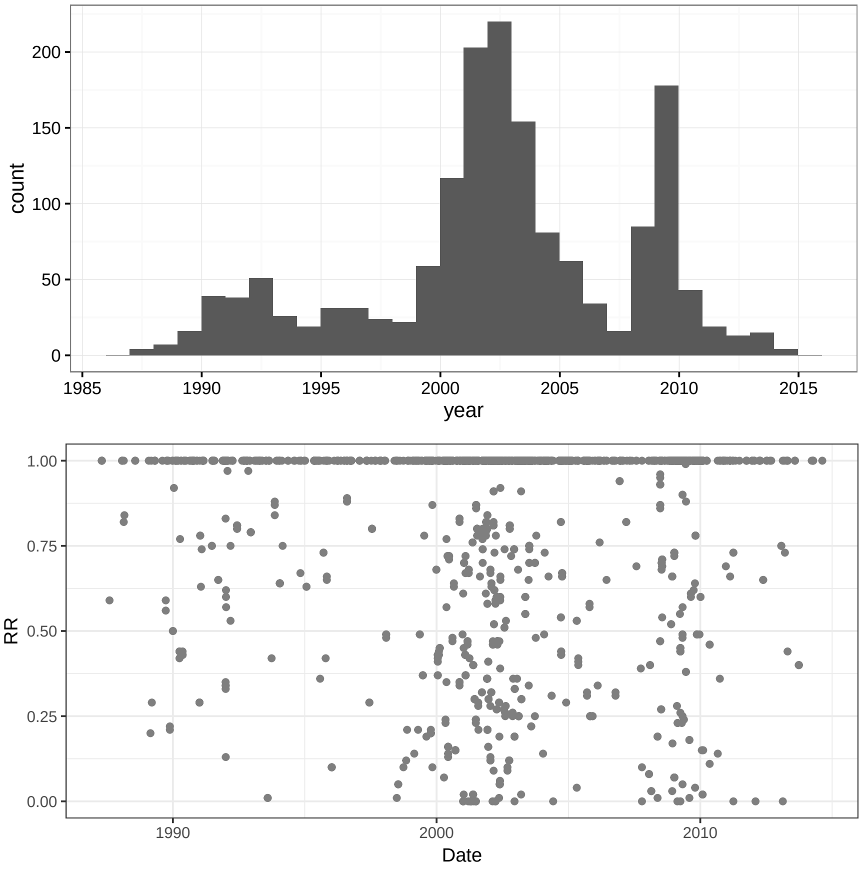

FIGURE 3.2: Number of defaults (top panel) and the RR outcomes by year (bottom panel), 1987-2015.

characteristics in our baseline analyses are measured one year before default. In terms of loan characteristics, however, our dataset reports larger average loan sizes ($224m compared to $142m) than Khieu, Mullineaux, and Yi (2012), with similar increases in the average term loan size, in the average revolver value and in the values of loans secured by all assets. With respect to recovery process characteristics, the average length of time a sample firm stays in default is 13 months with a maximum of 13 years. Most of the firms in the sample defaulted in a non-prepackaged bankruptcy, whereas 10% went through prepackaged bankruptcy and 13% had private workouts. Similar to McConnell, Lease, and Tashjian (1996) and Khieu, Mullineaux, and Yi (2012) the mean RR for loans with a reorganization plan lie between those for loans resolved in the traditional bankruptcy and those for loans going through the other forms of default resolution. With respect to borrower characteristics, the mean and median cash flows, relative to total assets for the settlement sample firm, are 16% and 9%, respectively. Figure 3.2 shows the number of defaults and recovery outcomes over time. More defaults are observed around 1993, 2003 and 2008, and more low recovery rates are also observed over these years.

3.4 A hierarchical econometric model for bank loan recovery rates

The starting point for our investigation is to determine the role of a complete set of recovery determinants, as explored by Khieu, Mullineaux, and Yi (2012). However in this research we implement a more enhanced and flexible modelling framework, and also use an extended sample that includes the GFC to ensure the capture of different economic conditions. The proposed methodology addresses two key challenges previously identified in the RR modelling literature, namely that the observed RR distribution has a distinctive (non-Gaussian) shape, and that when plotted over time the RRs appear to exhibit varying behavior - possibly owing to differences in the underlying PD. The first issue is addressed through the use of a finite Gaussian mixture model, first implemented on a combined loan and bond dataset by Altman and Kalotay (2014). This approach enables the RR determinants to be stochastically connected to the observed RRs through a latent predictive regression structure. In addition, to capture cyclical aspects such as the impact of the GFC, we augment the model structure with a Markov switching mechanism within the predictive regression model. In this framework, the regression coefficients depend on the state of the credit cycle, where the state corresponds to either a ‘credit upturn’ or a ‘credit downturn’ - i.e. a ‘good’ state or a ‘bad’ one. The coefficient estimates we obtain for each credit state provide insight into the procyclical effects of RR determinants. To combine the Gaussian mixture components, the latent predictive regression and

Markov switching mechanism, a hierarchical model is developed and estimated using a fully Bayesian inferential approach. The Bayesian approach enables a flexible hierarchical structure, which is estimated jointly and has the benefit of the consideration of each model component individually (i.e., marginally) while accounting for the uncertainty present in the remaining components.

Before detailing the form of the hierarchical model, also known as a 'state space' model, and describing the associated Bayesian inferential framework, following Altman and Kalotay (2014) we transform the observed RRs from the unit interval to the real line via the inverse of the cumulative distribution function (cdf) associated with the standard normal distribution, denoted by $\Phi^{-1}(\cdot)$. Specifically, if RR_i denotes the observed (appropriately discounted) RR value associated with defaulted loan i, we obtain the transformed RR value, denoted by y_i and given by

$$y_i = \Phi^{-1}(RR_i^*) \tag{3.1}$$

where

$$RR_i^* = \begin{cases} \epsilon & \text{if } RR_i = 0 \\ RR_i & \text{if } 0 < RR_i < 1 \\ 1-\epsilon & \text{if } RR_i = 1, \end{cases}$$

for $i = 1, 2, \ldots, n$. As is typical, before transformation is undertaken, the values of RR_i at zero are replaced with a small positive value, ϵ, and values at unity are replace with $1-\epsilon$, so that the y_i values are all finite.[1] It is the distribution of these y_i values that we model. Note that positive y_i values result whenever the original $RR_i > 0.5$. We now turn to the hierarchical model specification and the Bayesian inferential framework used to estimate it. Section 3.4.1 first details the Gaussian mixture model where membership to each mixture component is

[1] We use $\epsilon = 1 \times 10^{-8}$ since Altman and Kalotay (2014) find mixture model is not sensitive to the choice of ϵ.

predicted by a latent regression on RR determinants. The regression coefficients here are shown in their static form, without the Markov switching component, which is described later in Section 3.4.2. Section 3.4.3 then summarizes a computational strategy suitable for Bayesian inference to be conducted for the full dynamic model. Details regarding the algorithms required and implementation of the computational strategy are given in Section B of the Appendix.

3.4.1 A mixture model from recovery rate determinants

Having transformed each original RR observation, y_i is then treated as arising from one of J distinct Gaussian distributions, with the j^{th} distribution having mean and variance denoted by μ_j and σ_j^2, with $\mu_1 < \mu_2 < \cdots < \mu_J$. From the investor's perspective, recovery outcomes from a mixture component having a larger mean will be preferred, e.g. the J^{th} mixture component is preferred over the $(J-1)^{st}$, etcetera, with the first component being least desired, and therefore the ordering is imposed to retain the ability to interpret each of the categories.

Next, the connection between the mixture components and the RR determinants occurs through a latent ordered probit regression framework (Albert and Chib, 1993), which permits a range of explanatory variables, including loan, borrower, recovery process and macroeconomic or industry conditions, to characterize the probability of y_i being in component j of the Gaussian mixture. In particular, the determinants associated with loan i, denoted by $x_{1,i}, x_{2,i}, \ldots, x_{K,i}$, are related to a latent variable z_i through the regression equation

$$z_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_K x_{K,i} + \varepsilon_i, \tag{3.2}$$

with $\varepsilon_i \sim N(0,1)$. The latent (unobserved) z_i is referred to as a *predictive score* for defaulted loan i, while the vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)'$ contains the regression coefficients that describe the marginal impact of each of the determinants on this predictive score. The predictive score for loan i in (3.2) relates to each of the J Gaussian mixture components via a set of so-called 'cut-points', $\mathbf{c} = (c_0, \cdots, c_J)$, with $c_0 = -\infty, c_1 = 0, c_J = +\infty$, so that when in fact loan i belongs to group j, the j^{th} mixture probability may be calculated as $\Pr(c_{j-1} < z_i \leq c_j)$. Although the values of c_0, c_1 and c_J are fixed for identification purposes (see again Albert and Chib (1993)) the locations of the remaining cut-points (here c_2 and c_{J-1}) are treated as unknowns to be estimated. The values of μ_j and σ_j^2 are also estimated, essentially being determined by those y_i that are predicted by the regression to fall in category j.

Up to this point, the approach used here largely follows that of Altman and Kalotay (2014), apart from our use of a wider set of determinants as discussed in Appendix A. However, as described in the next section, we introduce an additional Markov switching component to the framework, so that the impact of the RR determinants is able to vary with the credit environment, whether good or bad, at the time of default. We also introduce a prior for $\boldsymbol{\beta}$, associated with the LASSO and discussed in Section 3.4.3.

3.4.2 The credit cycle

To incorporate the notion of an underlying dynamic credit cycle, a two-state Markov switching component is added to the mixture model with latent predictive regression framework outlined in Section 3.4.1. This binary credit cycle state variable for time (year) t is denoted by S_t, and takes on either the value of zero or one, depending on the underlying credit environment prevalent at the time of default.[2] The credit cycle states are normalized so that $S_t = 0$ corresponds to a low recovery period (a downturn, or 'bad' credit state), while $S_t = 1$ corresponds to a high recovery period (an upturn, or 'good' credit state). Transition to each credit cycle state at time t from a relevant state one period

[2] In this study, $t = 1$ corresponds to 1987, the first year of the available sample period.

earlier, time $t-1$, is governed by the probabilities

$$\begin{aligned} \Pr(S_t = 0|S_{t-1} = 0) &= p \\ \Pr(S_t = 1|S_{t-1} = 0) &= 1-p \\ \Pr(S_t = 1|S_{t-1} = 1) &= q \\ \Pr(S_t = 0|S_{t-1} = 1) &= 1-q. \end{aligned} \tag{3.3}$$

According to the transition probabilities in (3.3), if the credit cycle at time $t-1$ is in a low recovery state (i.e $S_{t-1} = 0$), then the chance of remaining in this 'bad' state at time t equal to p, with $0 < p < 1$, while the chance of moving to the 'good' recovery state (i.e.with $S_t = 1$) is equal to $1-p$. On the other hand, if $S_{t-1} = 1$, then the chance of remaining in the 'good' recovery state at time t is equal to q, with $0 < q < 1$, and the chance of moving to the 'bad' recovery state at time t is given by $1-q$.

Using the Markov switching device, two sets of regression coefficients are obtained for the recovery determinants: $\boldsymbol{\beta_0} = (\beta_{0,0}, \beta_{1,0}, \ldots, \beta_{K,0})'$ relating to the predictive scores in credit cycle downturns, and $\boldsymbol{\beta_1} = (\beta_{0,1}, \ldots, \beta_{K,1})'$ applicable during credit cycle upturns. To link the latent credit cycle states to the available data, let t_i denote the time associated with the default of loan i, so that S_{t_i} indicates the relevant state of the credit cycle at the time of default of loan i. The predictive regression coefficient vector $\boldsymbol{\beta_0}$ will apply for predicting z_i if $S_{t_i} = 0$, whereas the vector $\boldsymbol{\beta_1}$ will apply for predicting z_i if $S_{t_i} = 1$. Hence, by adding the Markov switching component, the regression coefficients in (3.2) become state dependent, and the predictive regression for loan i becomes

$$z_i = \beta_{0,S_{t_i}} + \beta_{1,S_{t_i}} x_{1,i} + \cdots + \beta_{K,S_{t_i}} x_{K,i} + \varepsilon_i, \tag{3.4}$$

where again $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$, for $i = 1, 2, \ldots, n$. Now that the regression coefficients in the predictive regression are state dependent, the estimated values of the vectors $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_0}$ will provide insight into the differentiated impact of RR determinants in 'good' times and 'bad'.

3.4.3 Bayesian inference

Like Altman and Kalotay (2014), we take a Bayesian approach when estimating the proposed model, an approach that offers several advantages over the perhaps more familiar frequentist strategy. The outcome of any Bayesian inferential procedure is a full joint probability distribution for all unknowns, including both parameters and latent variables. This outcome distribution, referred to the joint posterior distribution, characterizes all that is known about the parameters, and the credit states, prediction scores and Gaussian mixture allocations for each loan. From this joint posterior, the corresponding marginal distribution for any individual parameter or state variable (or indeed any subset of these) will automatically and coherently account for uncertainty in the remaining unknown variables. This is of particular importance when working with a hierarchical model, such as the one we advocate here.

An added advantage of using a hierarchical model within a Bayesian framework is that computation to produce the posterior can be undertaken efficiently using MCMC techniques. Details of this computation is provided in Appendix 3.7. As a further advantage, Bayesian inference yields a finite sample analysis, conditioning only on the available data, whereas a corresponding Frequentist inferential method would typically require assumptions about the behavior of estimators as the sample size increases without bound. This is important in empirical applications, such as the one undertaken here, where the number of RR observations are limited relative to the number of unknowns being estimated.

Prior distribution incorporating the Bayesian LASSO

We conservatively adopt a relatively non-informative prior distribution with *a priori* independence assumed between $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_J)'$, $\boldsymbol{\sigma^2} = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_J^2)'$,

$\mathbf{c}, \boldsymbol{\beta_0}, \boldsymbol{\beta_1}, p$ and q. Apart from the prior specified for the state-dependent predictive regression coefficients, $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ discussed below, the prior components are chosen from the appropriate conditionally conjugate family, thereby enabling fast computation of the posterior distribution via MCMC.

For the predictive regression vectors, $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$, we introduce the use of the Bayesian LASSO prior of Park and Casella (2008). As is now widely recognized the LASSO encourages a sparse regression model by down-weighting certain covariates when a large number of regression terms are used (Nazemi and Fabozzi (2018)), as is the case here. Effectively, the LASSO will reduce the size of the estimated regression coefficients to account for correlation

(multi-collinearity) or other dependence between the available recovery determinants, favouring putting weight on regressors whose association with the response variable (here the latent predictive scores $\mathbf{z} = (z_1, z_2, \ldots, z_n)$) can be estimated with relative certainty. In this way, predictive information shared by different determinants is not 'double counted' when fitting the model. The Bayesian LASSO achieves this reduction, or *shrinkage*, through the choice of the prior distribution for $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$. This prior distribution for each regression vector in the dynamic credit cycle context relies critically on certain additional so-called shrinkage parameters, denoted by λ_0^2 and λ_1^2, respectively, with a single shrinkage parameter, denoted by λ^2 used for the static latent regression model. These shrinkage parameters are included as unknowns, and are also estimated here within the Bayesian framework.

We note that many existing studies have considered the predictive performance for RRs. For instance, Altman and Kalotay (2014) investigate the predictive performance of different models using a set of variables for debt seniority, collateralization and industry classification. In the Bayesian paradigm, Barbieri and Berger (2004) point out that a model with highest posterior probability is not necessarily optimal for prediction, instead, optimal predictive models are 'median probability models'.

3.5 Empirical results

The results reported in this section are based on two implementations of the model described in Section 3.4, namely the *static* version, corresponding to Section 3.4.2 where the predictive regression coefficient variables are assumed to be constant over the entire sample period, and the *dynamic* version described in Section 3.4.3, where the latent time-varying credit cycle is included. The LASSO priors are used in both cases. Note that the where the term estimate is used it will generally refer to the posterior mean of the posterior distribution for the relevant quantity. Uncertainty in such an estimate will be indicated by a 95% so-called highest posterior density (HPD) interval taken as the shortest single interval associated with 95% marginal posterior probability. These Bayesian point and interval estimates are used to summarize the marginal posterior distributions, and are obtained from the MCMC output based on 100,000 MCMC draws retained following a 5,000 burn-in period.

3.5.1 Recovery mixture components

As alluded to in Section 3.3.1, given RR observations are clustered at zero (zero recovery) and one (full recovery), following Altman and Kalotay (2014), we apply a $J = 4$ Gaussian mixture model to transformed RRs. Table 3.2 provides details of the features of the estimated Gaussian mixture components that result from each of the two models fitted to the dataset considered. For each case, the estimated mean and standard deviation parameters for each Gaussian component are provided, along with its corresponding mixture weight and median RR. The estimated components labeled 1 and 4 effectively concentrate, with the same relative proportions for both the dynamic and static specifications, on point masses corresponding to RR values at zero and one, respectively. This fact confirms that the mixture specification accommodates the corresponding observed concentrations at the extremes found in the empirical RR distribution.

Component (j)	**Static model**				**Dynamic model**			
	1	*2*	*3*	*4*	*1*	*2*	*3*	*4*
a) Mean (μ_j)	-5.61 (-5.72,-5.50)	-0.85 (-1.77,-0.32)	0.21 (0.03,0.46)	5.61 (5.61,5.61)	-5.61 (-5.72,-5.50)	-1.31 (-2.00,-0.52)	0.09 (-0.02,0.24)	5.61 (5.61,5.61)
b) Std (σ_j)	0.08 (0.05,0.13)	0.54 (0.21,0.83)	0.41 (0.29,0.53)	0.00 (0.00,0.00)	0.08 (0.05,0.13)	0.44 (0.14,0.90)	0.48 (0.38,0.58)	0.00 (0.00,0.00)
c) Implied weight	0.01 (0.01,0.01)	0.10 (0.02,0.20)	0.25 (0.15,0.33)	0.64 (0.64,0.64)	0.01 (0.01,0.01)	0.04 (0.01,0.11)	0.31 (0.24,0.34)	0.64 (0.64,0.64)
d) **Mean RR**	0.00 (0.00,0.00)	0.25 (0.04,0.37)	0.63 (0.51,0.68)	1.00 (1.00,1.00)	0.00 (0.00,0.00)	0.10 (0.02,0.30)	0.54 (0.49,0.59)	1.00 (1.00,1.00)

Table 3.2: Estimated Gaussian mixture components. Posterior mean and 95% HPD interval (in parentheses) for each mixture component as indicated in the first row, and for a) the component mean parameter (Mean (μ_j)) in row two; b) the component standard deviation parameter (Std (σ_j)); c) the Implied weight, as given by the proportion of observations allocated to the mixture component; and d) Mean RR, corresponding to the inversion of (3.1); for each of the four mixture components as labeled by $j = 1, 2, 3$, and 4, for the static model (columns 2-5) and the dynamic model (columns 6-9)

However, the two interior mixture components (labeled 2 and 3) show differences across these attributes, notably in the third mixture component. As the two models correspond to different latent predictive regression structures - one static (i.e. without imposing the Markov switching credit states) and the other dynamic - two separate estimation results are shown. The mean parameter for the j^{th} component, μ_j, and the corresponding standard deviation, σ_j^2, is determined by observed RRs with predictive regressions correspond to outcomes that fall in mixture component j.

3.5.2 The latent credit cycle

The latent Markov switching states are introduced into the dynamic model to characterize the time-series variation in the observed RRs. Estimates of q, the probability of remaining in a good credit state, from one year to the next, and p, the probability of the economy remaining in a bad credit state, are given in Table 3.3, with the corresponding estimated steady state (or long-run) probabilities being 61% and 39% for the 'good' and 'bad' states, respectively, as indicated by the final row of Table 3.3. A line graph of the estimated probabilities for being in the good credit cycle state during each specific year during the given sample period is overlaid on a plot of RR outcomes in Figure 3.3, with the shaded bars shown in the figure indicating the sample proportion of fully recovered RRs reported in each calendar year. Interestingly, the troughs that appear in the estimated credit cycle reflect the well known economic downturns, namely, the early stage of the 1997 Asia financial crisis, the burst of US dot-com bubble in 2002 and global financial crisis (GFC) in 2008-2009. The indicated credit downturn corresponding to 1994-1995 may bear some connection to the Mexican peso crisis and its impact on the North American Free Trade Agreement.

To demonstrate the importance of allowing for temporal variation in economic conditions, we contrast the inferential results from the dynamic and static

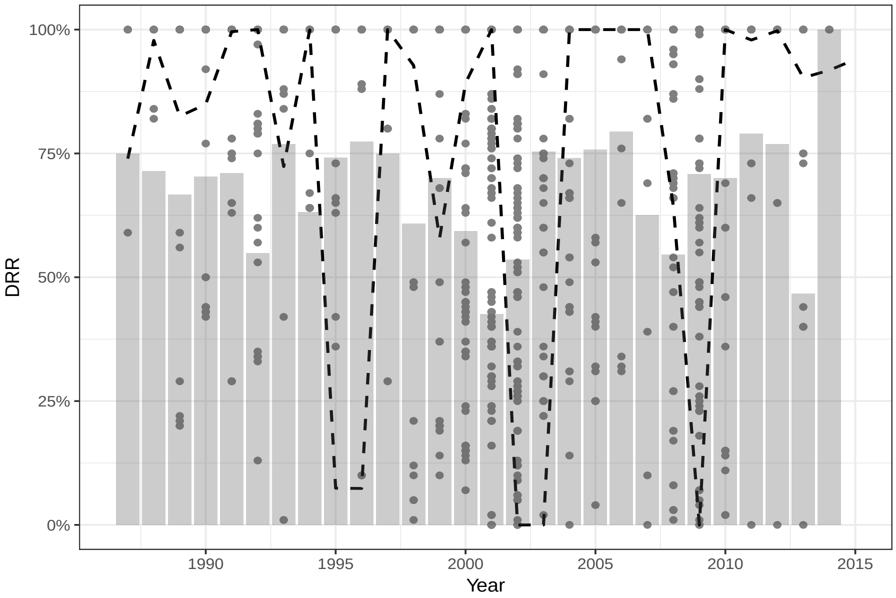

Figure 3.3: RRs (·) plotted over the 1987-2015 period and aligned by calender year, with the estimated probability of being in a 'good' credit state (dashed line), implied by the dynamic model, given by the superimposed line graph. Shaded area represents the percentage of full recovery in each corresponding year.

Bayesian models. As discussed in Section 4, the models are developed using a Bayesian LASSO to control for multi-collinearity arising from competing and highly correlated RR determinants. Figures 3.4 and 3.5 illustrate the significance of each of the variables after applying the Bayesian LASSO, with Figure 3.4 showing the significance of parameters in the static model, and Figure 3.5 showing those for the dynamic case, with the top panel of the latter figure corresponding to significance for the bad' credit state and the lower panel corresponding to the significance of determinants under the 'good' credit state. In all cases, interval estimates for variables that cross the vertical axis at zero indicate a lack of (marginal) significance for that variable in the relevant model.

	$\Pr(S_t \mid S_{t-1})$	
	$S_t = 0$	$S_t = 1$
$S_{t-1} = 0$	0.53 (0.24,0.83)	0.47 (0.17,0.76)
$S_{t-1} = 1$	0.67 (0.40,0.88)	0.33 (0.12,0.60)
Steady state: $\Pr(S_t = 1)$	0.61 (0.38,0.83)	

TABLE 3.3: Estimated posterior mean and 95% HDP (in parenthesis) for each possible transition probability associated with a one period transition from credit state S_{t-1}, shown by row in the first column, to a new credit state S_t, shown by corresponding entry in columns 2 and 3. The final row provides the estimated overall long-run probability of being in the 'good' credit state $\Pr(S_t = 1)$ resulting from the dynamic model.

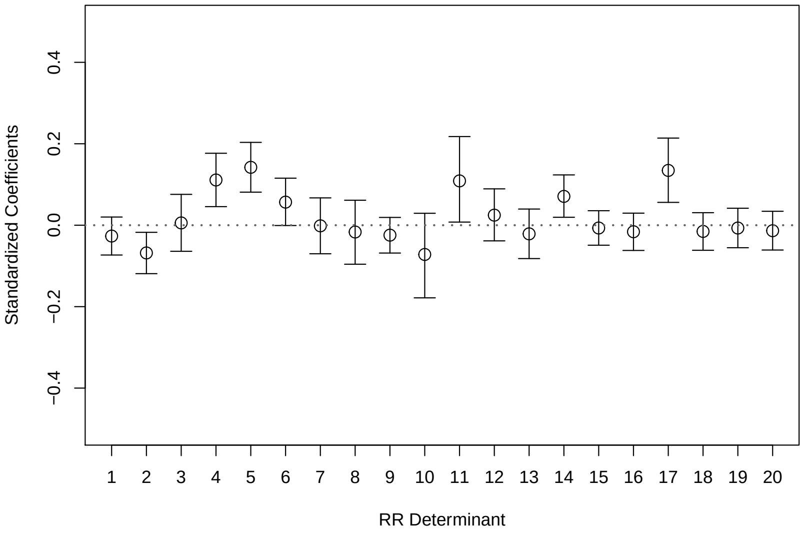

FIGURE 3.4: Static model: Posterior mean estimates (○) of individual β_k coefficients, for variable $k = 1, 2, ..., K = 20$, with corresponding 95% HPDs indicated by the vertical bars. The variables are: (1) LOANSIZE, (2) LOANTYPE, (3) LOANTYPE × FIRMSIZE, (4) ALLASSETCOLL, (5) INVENTRECIVECOLL, (6) OTHERCOLL, (7) PREPACK, (8) RESTRUCTURE, (9) OTHERDEFAULT, (10) TIMETOEMERGE, (11) TIMETOEMERGE^2, (12) PREPACK × TIMETOEMERGE, (13) FIRMSIZE, (14) FIRMPPE, (15) FIRMCF, (16) FIRMLEV, (17) EVERDEFAULTED, (18) INDUSTRESS, (19) GDP, (20) AIS.

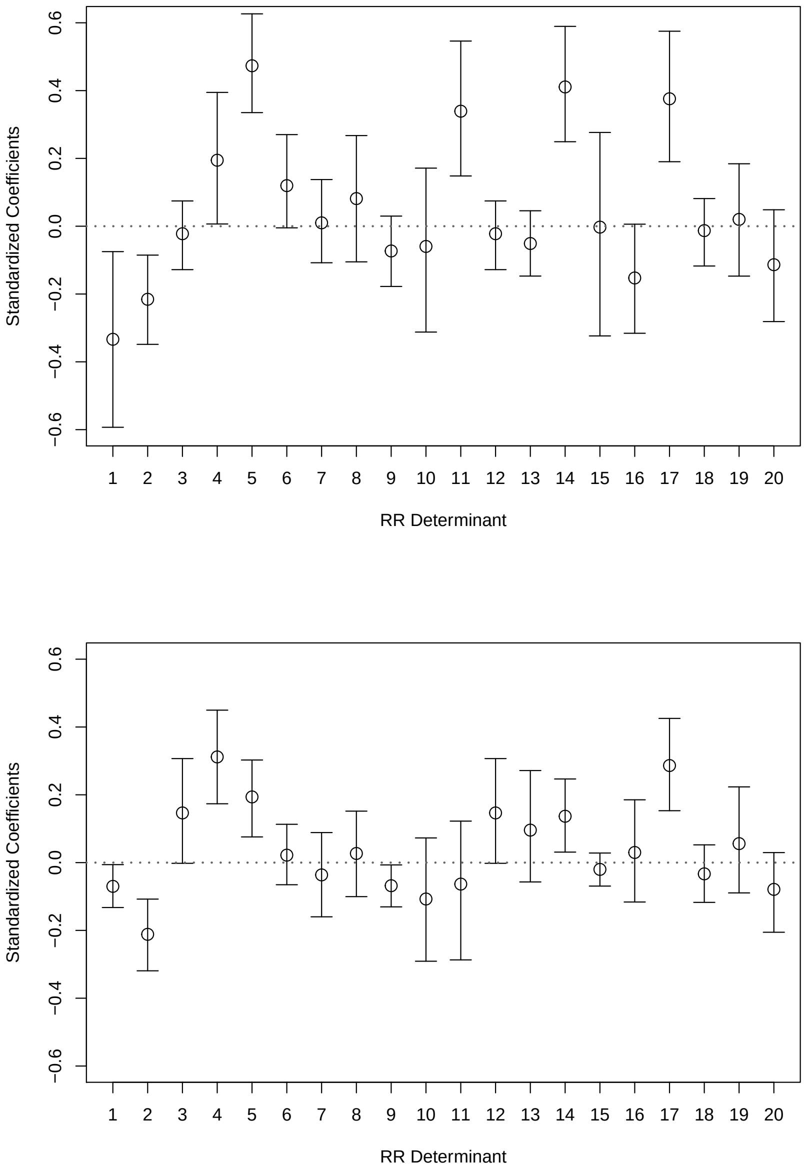

FIGURE 3.5: Dynamic model: Posterior mean estimates (○) of $\beta_{0,k}$ (top panel) and $\beta_{1,k}$ (bottom panel), for variable $k = 1, 2, ..., K = 20$, along with corresponding 95% credible intervals indicated by the vertical bars. The variables are: (1) LOANSIZE, (2) LOANTYPE, (3) LOANTYPE × FIRMSIZE, (4) ALLASSETCOLL, (5) INVENTRECIVECOLL, (6) OTHERCOLL, (7) PREPACK, (8) RESTRUCTURE, (9) OTHERDEFAULT, (10) TIMETOEMERGE, (11) TIMETOEMERGE^2, (12) PREPACK × TIMETOEMERGE, (13) FIRMSIZE, (14) FIRMPPE, (15) FIRMCF, (16) FIRMLEV, (17) EVERDEFAULTED, (18) INDUSTRESS, (19) GDP, (20) AIS.

	Static model	Dynamic model		Khieu, Mullineaux, and Yi (2012)	Khieu, Mullineaux, and Yi (2012)
	β	β_0	β_1	β	β
Recovery Determinant	**Bayes**	bad **Bayes**	good **Bayes**	**OLS**	**QMLE**
Loan characteristics					
(1) LOANSIZE		–	–		
(2) LOANTYPE	–		–	–	–
(3) LOANTYPE×FIRMSIZE			+	+	+
(4) ALLASSETCOLL	+		+	+	+
(5) INVENTRECIVECOLL	+	+	+	+	+
(6) OTHERCOLL		+	+	+	+
Recovery process characteristics					
(7) PREPACK					
(8) RESTRUCTURE					
(9) OTHERDEFAULT					
(10) TIMETOEMERGE		–	–	–	
(11) TIMETOEMERGE^2		+		+	
(12) PREPACK×TIMETOEMERGE		–	+	+	+
Borrower characteristics					
(13) FIRMSIZE		–	+		
(14) FIRMPPE	+	+			
(15) FIRMCF					
(16) FIRMLEV				–	
(17) EVERDEFAULTED	+	+	+	+	+
Macroeconomic & industry conditions					
(18) GDP				+	+
(19) INDDISTRESS		–			
Probability of default					
(20) AIS		–			

Table 3.4: The sign of significant RR determinants under each of the Bayesian models and of those from Khieu, Mullineaux, and Yi (2012).

3.5.3 The predictive regressions

In Table 3.4 we report an alternative summary of these predictive regression estimation results, again for each of the static and dynamic models, in this case showing the sign only of the significant coefficients along with those obtained previously in Khieu, Mullineaux, and Yi (2012). In column two, we report the signs of the significant RR determinants identified in column one under our static model resulting from the Bayesian approach and corresponding to data from 1987-2015. Columns three and four of Table 3.4 report the sign of significant RR determinants under the Bayesian dynamic model, with β_0 corresponding to the bad credit state (i.e. when $S_t = 0$) and with β_1 corresponding to the good credit states (i.e. when $S_t = 1$). For comparative purposes, the sign of the significant coefficients of these determinants corresponding to Frequentist inference using OLS and QMLE methodologies, and relating to data from 1997-2007 (as reported in Khieu, Mullineaux, and Yi (2012)), are provided in columns five and six. This is done to illustrate the contribution made by static vs. dynamic versions, and the need to allow for variation in the impact of RR determinants under different credit conditions. The reported models are more parsimonious relative to the existing literature due to our use of a LASSO, though both are consistent overall regarding the relevance of RR determinants. While the numerical values of the estimated Frequentist and Bayesian coefficients themselves are not directly comparable, we can compare their statistical significance and their sign.

	Static model	**Dynamic model**	
	$\boldsymbol{\beta}$	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$
Standardized Parameter	**MPM** **(95% HPD)**	**MPM** **(95% HPD)**	**MPM** **(95% HPD)**
Loan characteristics			
(1) LOANSIZE($M)	−0.026 (−0.081, 0.030)	**-0.28** (−0.45, −0.10)	**-0.093** (−0.18, −0.00054)
(2) LOANTYPE	**-0.068** (−0.13, −0.0079)	−0.27 (−0.43, 0.11)	**-0.26** (−0.41, −0.11)
(3) LOANTYPE×FIRMSIZE	0.025 (−0.050, 0.10)	0.0098 (−0.11, 0.13)	**0.35** (0.0094, 0.71)
(4) ALLASSETCOLL	**0.11** (0.032, 0.19)	0.23 (−0.016, 0.48)	**0.69** (0.49, 0.90)
(5) INVENTRECIVECOLL	**0.14** (0.069, 0.22)	**1.89** (0.57, 5.33)	**0.37** (0.24, 0.50)
(6) OTHERCOLL	0.057 (−0.011, 0.13)	**0.25** (0.095, 0.41)	**0.15** (0.028, 0.26)
Recovery process characteristics			
(7) PREPACK	−0.0024 (−0.084, 0.080)	0.14 (−0.050, 0.34)	−0.14 (−0.34, 0.057)
(8) RESTRUCTURE	−0.016 (−0.11, 0.076)	−0.19 (−0.45, 0.07)	−0.0056 (−0.21, 0.21)
(9) OTHERDEFAULT	−0.024 (−0.077, 0.028)	−0.060 (−0.18, 0.068)	0.48 (−0.18, 4.41)
(10) TIMETOEMERGE	−0.072 (−0.20, 0.048)	**-0.65** (−1.20, −0.11)	**-0.44** (−0.86, −0.0078)
(11) TIMETOEMERGE^2	0.11 (−0.0093, 0.24)	**1.14** (0.53, 1.77)	0.12 (−0.13, 0.47)
(12) PREPACK×TIMETOEMERGE	0.0064 (−0.076, 0.090)	**-0.22** (−0.39, −0.057)	**0.45** (0.16, 0.81)
Borrower characteristics			
(13) FIRMSIZE	−0.022 (−0.095, 0.050)	**-0.055** (−0.16, −0.053)	**0.23** (0.079, 0.39)
(14) FIRMPPE	**0.071** (0.011, 0.13)	**0.52** (0.29, 0.75)	−0.016 (−0.076, 0.044)
(15) FIRMCF	−0.0065 (−0.057, 0.045)	0.17 (−0.36, 0.70)	−0.17 (−0.37, 0.032)
(16) FIRMLEV	−0.016 (−0.072, 0.038)	−0.14 (−0.34, 0.065)	−0.21 (−0.42, 0.0050)
(17) EVERDEFAULTED	**0.13** (0.041, 0.22)	**0.53** (0.28, 0.80)	**0.60** (0.37, 0.87)
Macro-eco & industry conditions			
(18) GDP	−0.0072 (−0.065, 0.050)	−0.075 (−0.32, 0.18)	−0.042 (−0.25, 0.16)
(19) INDDISTRESS	−0.015 (−0.070, 0.039)	**-0.15** (−0.30, −0.0071)	0.077 (−0.027, 0.19)
Probability of default			
(20) AIS	−0.013 (−0.070, 0.043)	**-0.28** (−0.48, −0.080)	−0.018 (−0.17, 0.13)

TABLE 3.5: Bayesian estimates of the regression coefficients based on 100,000 retained MCMC draws (with 5,000 burn-in) from each marginal posterior as indicated by the column heading. Results in the dynamic case shown in column three and four correspond to estimates conditioned on the latent credit cycle state, with $\boldsymbol{\beta_0}$ corresponding to the 'bad' state, and $\boldsymbol{\beta_1}$ corresponding to the 'good' state. MPM denotes the marginal posterior mean and 95% HPD (in parentheses) denotes the 95% higher posterior density interval. For the static model, the MPM of the squared shrinkage parameter is $\lambda^2 = 3.33$ the static case, whereas for the dynamic case, the (conditional) MPMs are $\lambda_0^2 = 2.98$ and $\lambda_1^2 = 2.71$, corresponding to the bad and good states, respectively.

We note that only three loan characteristic determinants appear to be important for explaining RRs in bad times, whereas there is evidence that six are relevant during good times. We note that, like the OLS and QMLE results of Khieu, Mullineaux, and Yi (2012), the LOANSIZE determinant does not appear significant in the Bayesian static model. However once the credit cycle is incorporated this determinant does appear to be important.[3] In the case of recovery process and borrower characteristics, we find that a lesser number of variables are important in bad times and a greater number in good times. Also, and importantly, the relationships for some variables change from negative to positive. Finally, and consistent with Khieu, Mullineaux, and Yi (2012), the Bayesian posterior distribution for the static model shows no relation between RR and the PD measured by AIS. However, when we allow for different economic conditions, a negative relationship is indeed found between PD and RR, but only during bad times. This finding has clear implications for countercyclical capital allocations in operational risk modelling.

We now examine these variables more closely using our dynamic model. Table 3.5 reports fully detailed numerical summaries of the static and dynamic model Bayesian posterior distributions. The results for each type of RR determinant grouping are discussed in detail over the next several subsections.

In line with Dermine and De Carvalho (2006), we find loan size to be negatively associated with RRs. Irrespective of being in a good or bad cycle, from a bank's perspective, the larger the loan amount, the less likely the bank will be able to recover subsequent to default. Larger loans are generally organized around a syndicate banking arrangement; hence, as more providers are involved, lower RRs are realized once they enter foreclosure. This finding is contrary to those of Acharya, Bharath, and Srinivasan (2007), as banks granting

[3] Although significant in both 'good' and 'bad' states, the magnitude of the estimated marginal impact under the 'good' state is relatively small.

larger loans are meant to have less asymmetric information and more bargaining power during the bankruptcy process. However, it seems that as loan sizes increase and default occurs during a downturn, banks are less likely to recover their outstanding debts.

Loan type is not useful to explain RR levels in bad times, irrespective of whether the credit granted is a term loan or a revolver. During an upturn, however they can contribute to explaining RRs with respect to revolver loans. Khieu, Mullineaux, and Yi (2012) find a similar significant relationship and argue that since revolvers typically have a shorter duration and are therefore reviewed more often, banks are able to reassess their clients' credit profiles and seek further collateral if necessary.

The literature emphasizes the importance of collateral with respect to higher RRs emanating from secured loans where more secured loans imply higher RRs (Altman and Kishore, 1996; Araten, Jacobs, and Varshney, 2004; and Castle and Keisman, 1999). Our study contributes to the literature by showing that during good times, we report similar results to those of Khieu, Mullineaux, and Yi (2012), i.e., a significant positive relation between the RR and total assets used as collateral. While during bad times this association is not significant, our dynamic model also shows that assets such as inventory, receivables and other more liquid assets do appear to be important for recovering a higher RR across both credit cycle states, particularly during bad times.

Recovery process characteristics

The existence of pre-arranged recovery processes for bankruptcy and out-of-court restructuring in the event of default-triggered failure is examined. Pre-packaged processes do not have a significant relationship with RRs subsequent to default in either good or bad times. Although this finding is in line with those of Khieu, Mullineaux, and Yi (2012), we note that the literature finds companies that pre-package appear to be more financially sound (Ryan (2008)).

With respect to distressed exchanges, it transpires that firms undertaking pre-packaging are normally more solvent at the time of re-organization than are bankrupt firms (Franks and Torous (1994)).

While none of our static results support associations between any recovery process RR characteristics, the dynamic model results do indicate that TIMETOEMERGE, is negatively related to the ultimate RR, with banks less likely to recover when they engage further in bankruptcy proceedings. Owing to the magnitude of the estimated coefficient for this variable and that of its squared value, TIMETOEMERGE^2, it appears this relationship becomes more relevant during bad times. Furthermore, the dynamic model also finds that the constructed determinant given by PRE-PACK × TIMETOEMERGE, is also important for explaining RR outcomes irrespective of the state of credit cycle.

However, the quadratic term for time to emerge, representing the nonlinearity between TIMETOEMERGE and RR, has a significant impact on recovery outcomes only in downturns. This finding is different from that observed in bond studies such as Covitz, Han, and Wilson (2006).

Borrower characteristics

The literature is not definitive on whether firm size impacts RRs. Large firms may signal higher bankruptcy costs, resulting in lower RRs. Conversely, larger firms are expected to present less information asymmetry problems to creditors, hence facilitating any restructuring process and improving recoveries from lenders. As per Khieu, Mullineaux, and Yi (2012), we do not find a significant relation between firm size and RRs with the static model. However, our dynamic model reveals a significant negative (positive) relation with RRs and firm size during bad (good) times. During bad times, the larger the firm, the greater the negative impact on RRs. In good times, this is reversed; larger firms are associated with greater RRs. This could be a sign of loan mispricing: recoverable assets are being over-valued prior to bankruptcy in bad times. Conversely, during good times, these asset values are more likely realizable and consistent with higher RRs.

The level of a firm's tangible assets, namely property, plant and equipment (FIRMPPE), is thought to be positively related to the RR (Acharya, Bharath, and Srinivasan, 2007), that is, banks are more likely to recover outstanding loans when firms report tangible assets on their balance sheet. We find likewise, but only relating to bad times. (We note that Acharya, Bharath, and Srinivasan (2007) include bonds that are generally unsecured, in their sample.) We also find that firm cash flow and leverage are not significantly related to RRs. Unlike Khieu, Mullineaux, and Yi (2012), as our dataset excludes bonds and focuses only on loans (which are likely to be secured by tangible assets), this contrast is not surprising. Finally, consistent with the literature, we find that prior defaults as indicated by the variable EVERDEFAULTED are significant and positively related to RRs in both good and bad times.

Macroeconomic & Industry conditions

We find no significant relation between GDP and RRs, however this is due to the fact that we control for the underlying economic conditions, which coincides with the credit cycle (see Figure 3.3 and the discussion in Section 3.5.2). We do, however, obtain a significant negative relation between industry distress (measured by stock returns, via the INDDISTRESS variable) and RRs in bad times.

Probability of default

The AIS is used to measure the PD, following Khieu, Mullineaux, and Yi (2012). The literature suggests a negative relationship between the PD and the RR. Both Hu and Perraudin (2002) and Altman et al. (2005) report a negative association, although the former uses bond default data. Khieu, Mullineaux, and Yi (2012) report no relationship between the PD and the RR. We find that PD is significantly negatively related to RR, but only in bad times, arguably consistent with banks being less likely to recover under an increased PD during bad times. This finding is partially consistent with Altman et al. (2005), although the study does not distinguish between good and bad times.

3.6 Model evaluation and further investigations

We evaluate the performance of the proposed model against several benchmarks using a formal Bayesian approach in this section. Specifically, we compare the cumulative log-predictive Bayes factor (see, for example, Berger, 2013) for the proposed model and the mixture model proposed by Altman and Kalotay (2014). Let $p(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathcal{M}_A, \boldsymbol{\theta}_A)$ denote the parametric conditional density for $\mathbf{y}_t$ given the history $\mathbf{y}_{t-1}$ and prediction model $\mathcal{M}_A$ as well as the modeling parameters $\boldsymbol{\theta}_A$, where $\mathbf{y}_t$ consists of all recovery data defaulted in year t. Then, the marginal likelihood can be written as

$$p(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathcal{M}_A) = \int p(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathcal{M}_A, \boldsymbol{\theta}_A) p(\boldsymbol{\theta}_A|\mathbf{y}_{t-1}, \mathcal{M}_A) d\boldsymbol{\theta}_A$$

where $p(\boldsymbol{\theta}_A|\mathbf{y}_{t-1}, \mathcal{M}_A)$ is the posterior density. Accordingly, the Bayes factor for comparing $\mathcal{M}_A$ against the alternative, $\mathcal{M}_B$, can be computed using all the data by

$$BF = \frac{p(\mathbf{y}_{1:T}|\mathcal{M}_A)}{p(\mathbf{y}_{1:T}|\mathcal{M}_B)}.$$

However, given the hierarchical structure and the complexity of the model, the marginal likelihood is not available in closed form. Since we have used Gibbs-based MCMC method to evaluate the posterior densities for all modeling parameters, the method described in Chib (1995) is suitable for computing the marginal likelihood for the proposed model. An estimate of the marginal density can be obtained using the following decomposition, on the logarithm scale,

$$\ln \hat{p}(\mathbf{y}_t|\mathbf{y}_{t-1}) = \ln p(\mathbf{y}_t|\mathbf{y}_{t-1}, \boldsymbol{\theta}^*) - \ln p(\boldsymbol{\theta}^*) - \ln \hat{p}(\boldsymbol{\theta}^*|\mathbf{y}_{t-1}) \tag{3.5}$$

where we suppressed the explicit dependence on the model. We further decompose the last term in (3.5) as

$$\begin{aligned}\ln \hat{p}(\boldsymbol{\theta}^*|\mathbf{y}_{t-1}) = {} & \ln \hat{p}(\lambda_0^*|\mathbf{y}_{t-1}) + \ln \hat{p}(\lambda_1^*|\mathbf{y}_{t-1}) + \ln \hat{p}(\beta_0^*|\lambda_0^*, \mathbf{y}_{t-1}) + \ln \hat{p}(\beta_1^*|\lambda_1^*, \mathbf{y}_{t-1}) \\ & + \ln \hat{p}(\mathbf{c}^*|\beta_0^*, \lambda_0^*, \beta_1^*, \lambda_1^*, \mathbf{y}_{t-1}) + \ln \hat{p}(p^*, q^*|\mathbf{y}_{t-1}) + \ln \hat{p}(\sigma^{2*}|\mathbf{y}_{t-1}) \\ & + \ln \hat{p}(\mu^*|\sigma^{2*}, \mathbf{y}_{t-1}).\end{aligned}$$

To obtain each of the estimates, the Gibbs sampler need to be run for multiple times with some parameters fixed. We first estimate

$$p(\beta_0^*|\lambda_0^*, \mathbf{y}_{t-1}) \text{ and } p(\beta_1^*|\lambda_1^*, \mathbf{y}_{t-1})$$

as

$$\begin{aligned} & G^{-1}\sum(p(\beta_1^*|\lambda_1^*, \tau_1^{(j)}, \mathbf{S}^{(j)}, \mathbf{z}^{(j)}, \mathbf{z}^{*(j)}, \mu^{(j)}, \sigma^{2(j)}, \mathbf{y}_{t-1})) \\ \text{and } & G^{-1}\sum(p(\beta_0^*|\lambda_0^*, \tau_0^{(j)}, \mathbf{S}^{(j)}, \mathbf{z}^{(j)}, \mathbf{z}^{*(j)}, \mu^{(j)}, \sigma^{2(j)}, \mathbf{y}_{t-1}))\end{aligned}$$

respectively, where $\{\tau_0, \tau_1, \beta_0, \beta_1, \mathbf{z}, \mathbf{z}^*, \mu, \sigma^2, \mathbf{S}\}$ are obtained by continuing the Gibbs sampler with fixing $\lambda_0 = \lambda_0^*$ and $\lambda_1 = \lambda_1^*$. Then, additional G iterations are used in a similar strategy to obtain an estimate for each of $p(\mathbf{c}^*|\beta_0^*, \lambda_0^*, \beta_1^*, \lambda_1^*, \mathbf{y}_{t-1})$, $p(p^*, q^*|\mathbf{y}_{t-1})$, $p(\sigma^{2*}|\mathbf{y}_{t-1})$ and $p(\mu^*|\sigma^{2*}, \mathbf{y}_{t-1})$.

We plot the cumulative log-predictive Bayes factors from 1987 to 2015 for the dynamic mixture model by using the static model as the benchmark in Figure 3.6. The line indicates the difference between the dynamic model and the static model. At the end of the sample period, the calculated log marginal likelihood values are -3376.27 and -3409.86 for the dynamic model and the static model respectively. The figure shows that the dynamic model gains in predictive power over time compared to the benchmark mixture model because it incorporates time dynamics, especially during the market downturn. For instance, at the financial crisis in 2009, we observe a jump-up in the line at that period indicating the value of regime switching in density forecasts.

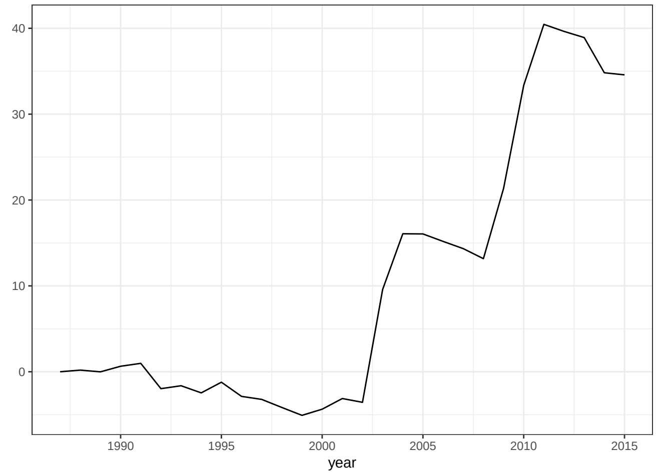

Figure 3.6: Difference in cumulative log-predictive likelihoods of the dynamic model against the static model.

3.7 Discussion and conclusion

Using US bank default loan data from Moody's Ultimate Recovery Database and covering the pre- and post-GFC period, this paper develops a dynamic predictive model for bank loan RRs, accommodating the distinctive features of the empirical RR distribution and incorporating a large number of possible determinants. Furthermore, some of the factors that are analyzed and reported in the literature have been overlooked as insignificant, due to the static model approach, which does not control for the different states of the economy. Our temporal conditioning in a hierarchical framework allows us to discriminate between good and bad states of the credit cycle. Thus, this paper contributes to the literature in different ways. The methodological approach used is Bayesian in nature and therefore is able to handle the hierarchical specification that is built to explain the complex relationship between PD, determinants and the empirical distribution of RRs. It is the first paper to incorporate time-series variation into the probabilistic modelling of bank loan RRs, proposing a Bayesian hierarchical framework that enables inference of a latent credit cycle. We also introduce the use of a LASSO prior to encourage the most relevant RR determinants to be found, despite potentially confounding evidence of correlation between observed RR determinants.

We find that some loan characteristics such as those using specific types of collateral hold different explanatory power in good times and bad. We find that certain recovery process variables, such as the length of time between default and resolution for loans with pre-packaged recoveries, differ in their importance in relation to RRs, depending on the state of the credit cycle, in this case being negatively related to RR in bad times while being positively related in good times. Only a few borrower characteristics and industry conditions appear to be relevant across the cycle. The defaulting firm's size and asset tangibility can imply different relationships with RRs depending on conditions. Finally, by allowing for variation in the level of PD, on top of the latent dynamic cycle states, we find a negative relationship with RR but one that is only significant during credit downturns.

Our results illustrate the importance of utilizing dynamic models that allow for time-varying conditions, as there is significant variation in the explanatory power of the variables analyzed depending on these conditions, yielding new insights previously unavailable from the established literature. Taking the case of PD, no relation between RR and PD is reported, yet under our dynamic model we find it is significantly and negatively related to RR during bad times. This variation in significance in variables across good and bad times occurs in several of our variables, supporting the need for a dynamic approach.

Our results also yield significant implications for the banking sector, notably providing empirical support for the latest addition to the Basel framework concerning the importance of activating countercyclical capital buffers during economic downturns. Applying such a buffer would not only enable banks to absorb increased losses but would also assist in achieving the broad macro-prudential goals of protecting the banking sector in periods of excess aggregate credit growth, and from the build-up of system-wide risk.

The notion that RR is driven by a systemic risk component that becomes more pronounced during bad times is evident from the results reported in our dynamic model. Loan size and type are also critical features, especially during bad times. Such features need to be priced within the cost of financing, as some banks are less likely to recover when the economy is entering a downturn. These important differential impacts, during bad and good times, suggests that RRs have a large element of systemic risk that needs to be factored in during the pricing of loan finance contracts. As RRs are an integral part of credit risk, this aspect should attract an additional risk premium allowing for a differentiated credit risk exposure. Under the new regime, banks are required to provide more timely and forward-looking information. It is no longer necessary for a credit event to have occurred before a credit loss is recognized. The paradigm shift is in being cognizant of the credit cycle and to update the bank's loan loss provision in line with their recovery rate expectations.

In summary, we find several variables are important for explaining RRs, depending on the state of the credit cycle. This has major implications for the countercyclicality of regulatory capital and operational risk management. The potential risk of not addressing such factors will result in either underestimating the relevant credit risk, or overestimating it. Both of these eventualities could potentially result in negative consequences, such as more expensive loans. This in turn would result in desirable customers leaving to access financing at cheaper rates from alternative institutions more effective in correctly pricing loans through the procyclical process.

Appendix A: The definitions of the recovery rate determinants

Table 3.6 details each of the twenty RR determinants considered in this paper, with the name of the determinant given in the first columns and the corresponding definition given in the second column. The determinants are clustered according broad type (loan characteristics, recovery process characteristics, borrower characteristics, macroeconomic and industry condition determinants and the probability of default) and each given a unique determinant number (in parentheses preceeding the name) that is used throughout the paper.

Name	Definition
Loan characteristics	
(1) LOANSIZE($M)	The dollar amount (in millions of dollars) of the facility at the time of issuance.
(2) LOANTYPE	A dummy variable equal to one if the loan is a term loan (fixed tenure and not recallable on demand), and equal to zero if it is a revolver (short-term revolving and recallable on demand).
(3) LOANTYPE × FIRMSIZE	The product of LOANTYPE and FIRMSIZE
(4) ALLASSETCOLL	A dummy variable equal to one if the loan is secured by all firm assets, and zero otherwise.
(5) INVENTRECIVECOLL	A dummy variable equal to one if the loan is secured by inventory, accounts receivable, or both, and zero otherwise.
(6) OTHERCOLL	A dummy variable equal to one if the loan is secured differently from the other types, and zero otherwise.
Recovery process characteristics	
(7) PREPACK	A dummy variable equal to one if the bankruptcy is through a pre-packaged bankruptcy, and zero otherwise.
(8) RESTRUCTURE	A dummy variable equal to one if default is resolved by out-of-court restructuring, including distressed exchange offers, and zero otherwise.
(9) OTHERDEFAULT	A dummy variable, equal to one if default is resolved by other methods than an out-of-court restructuring, pre-packaged formal bankruptcy, and 0 otherwise.
(10) TIMETOEMERGE	The length of time (in months) between bankruptcy or restructuring and emergence, often known as resolution time.
(11) TIMETOEMERGE^2	TIMETOEMERGE squared.
(12) PREPACK× TIMETOEMERGE	The product of PREPACK and TIMETOEMERGE.
Borrower characteristics	
(13) FIRMSIZE	The market value of firm-level assets one year before default. The market value is calculated as the book value of long-term and short-term debt plus the number of common shares outstanding.
(14) FIRMPPE	Firm asset tangibility, measured as net property, plant, and equipment over total book assets one year before default.
(15) FIRMCF	Firm cash flows, measured EBITDA (earning before interest, tax and depreciation and amortization) over total book assets one year before default.
(16) FIRMLEV	Firm leverage, measured as total long-term debt plus debt in current liabilities over total book assets one year before default.
(17) EVERDEFAULTED	A dummy variable equal to one if the firm has defaulted before, and zero otherwise.
Macro-eco & industry conditions	
(18) GDP	The annual GDP growth rate measured 1 year before default.
(19) INDDISTRESS	A dummy variable equal to one if the industry median stock returns in the year default is less than -30%, and zero otherwise. The stock returns are calculated without the defaulting firms and the industry is defined according to the three-digit SIC codes.
Probability of default	
(20) AIS	The credit spread (in percent) at the time of loan origination over LIBOR of the drawn loan that defaulted.

TABLE 3.6: Definitions of the RR determinants.

Appendix B: Implementation details for Bayesian analysis

The model detailed in Section 3.4 provides a characterization of the distribution of the observed RRs via a predictive regression of the RR mixture component on a large collection of RR determinants, with the regression coefficients in turn dependent upon the current state of an underlying credit cycle state variable. In the static regression setting, calculation of the posterior distribution via MCMC follows the approach of Altman and Kalotay (2014), who in turn rely upon the methodology details provided in Albert and Chib (1993). Our implementation here is similar, including in the dynamic case where we include the additional hierarchical layer containing the Markov switching variables, except for the use of the alternative LASSO prior specification on the latent RR regression parameter coefficient vector(s).

Appendix B1: Likelihood function

The relevant likelihood for Bayesian analysis is the joint probability density function (pdf) of the complete set of measurements, denoted by $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, together with the latent Markov switching state variables, $\mathbf{S} = (S_1, S_2, \ldots, S_T)$, all conditional upon the collection of parameters, $(\boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{\beta_0}, \boldsymbol{\beta_1})$. Unfortunately, even if the sequence of latent credit state variables, $\mathbf{S}$, were known, calculation of the likelihood function is not available in closed form, and consequently the Bayesian posterior is also not available. However, owing to the relationship between the Gaussian mixture model for each $\mathbf{y_i}$, the cut-points $\mathbf{c}$ and the latent predictive regression in (3.2), we can express the joint pdf of $\mathbf{y}$

and $\mathbf{z}$ conditional on $\mathbf{S}$, given by the product of

$$\begin{aligned} p(\mathbf{y}, \mathbf{z}, \mathbf{z}^* | \mathbf{S}, \boldsymbol{\psi}, \mathbf{x}) \propto & \prod_{i=1}^{n} \prod_{j=1}^{J} \frac{1}{\sigma_j} \phi(\frac{y_i - \mu_j}{\sigma_j}) \\ & \times \mathbb{I}(c_{j-1} < z_i \leq c_j) \times \mathbb{I}(z_i^* = j) \\ & \times \phi(z_i - \mathbf{x}_i'((1 - S_{t_i})\boldsymbol{\beta_0} + S_{t_i}\boldsymbol{\beta_1})), \end{aligned} \tag{3.6}$$

where $\phi(\cdots)$ denotes the pdf of the standard normal distribution, $\mathbb{I}(\cdot)$ is the indicator function so that $\mathbb{I}(A) = 1$ if event A is true and is equal to zero otherwise, $\boldsymbol{\psi} = (\boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \mathbf{c})$, and the joint pdf of the Markov switching states $\mathbf{S}$, given by

$$p(\mathbf{S}|\boldsymbol{\psi}) = p(S_1|p, q) \prod_{t=2}^{T} p(S_t|S_{1:t-1}, p, q), \tag{3.7}$$

where $p(S_t|S_{1:t-1}, p, q)$ is given in (3.3), and $p(S_1|p, q)$ arising from the long-run marginal probability given by

$$S_1|p, q \sim Bernoulli((1 - p)/(2 - p - q)). \tag{3.8}$$

The likelihood function is then the product of (3.6) and (3.7). Note that the factors in these equations are expressed conditionally given the parameters $(\boldsymbol{\psi}, p, q)$ and also given the regression covariates $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K]$, where $\mathbf{x}_k = (x_{k,1}, x_{k,2}, \ldots, x_{k,n})'$.

Appendix B2: Priors

To complete a fully Bayesian analysis, we must put a (joint) prior distribution over the unknown parameters. We take these priors to be relatively diffuse, so that the data will dominate the analysis. Specifically, the prior mean μ_j and variance σ_j^2 for the j^{th} mixture component of the RR distribution are taken as independent normal (N) and inverse gamma (IG) distributions, with

- $\mu_j \overset{ind}{\sim} N(\bar{\mu}_j \ , \ \bar{V}_{\mu,j})$, where $\bar{\mu}_j = 0$ and $\bar{V}_{\mu,j} = 100$ for $j = 1, \cdots, J$, and
- $\sigma_j^2 \overset{ind}{\sim} IG(\bar{a}_j \ , \ \bar{b}_j)$, where $\bar{a}_j = 3$ and $\bar{b}_j = 1$ denote the scale and shape parameters, respectively, for $j = 1, \cdots, J$.

To avoid the well known label switching problems in the finite mixture model, we impose the same identification restrictions, $\mu_1 < \cdots < \mu_J$, as in Koop, Poirier, and Tobias (2007).[4] The joint prior distribution for the cut-point vector $\mathbf{c}$ is completely diffuse, while the prior distributions for $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$, corresponding to the Bayesian LASSO for the coefficients of the RR determinants under the 'bad' and 'good' credit cycle states, respectively, are specified hierarchically using independent scale mixture of normals for each. These are given by

- $\boldsymbol{\beta_0} \mid \sigma_\varepsilon^2, \boldsymbol{\tau_0} \sim N(\mathbf{0}_K, \sigma_\varepsilon^2 \mathbf{I}_K D_{\boldsymbol{\tau_0}})$, with $D_{\boldsymbol{\tau_0}} = diag(\tau_{0,1}, \tau_{0,2}, \ldots, \tau_{0,K})$, and
- $\boldsymbol{\beta_1} \mid \sigma_\varepsilon^2, \boldsymbol{\tau_1} \sim N(\mathbf{0}_K, \sigma_\varepsilon^2 \mathbf{I}_K D_{\boldsymbol{\tau_1}})$, with $D_{\boldsymbol{\tau_1}} = diag(\tau_{1,1}, \tau_{1,2}, \ldots, \tau_{1,K})$,

with I_K denoting the $K-$ dimensional identity matrix and the mixing variables (also known as local shrinkage parameters) given by $\boldsymbol{\tau_0} = (\tau_{0,1}, \tau_{0,2}, \ldots, \ \tau_{0,K})$ and $\boldsymbol{\tau_1} = (\tau_{1,1}, \tau_{1,2}, \ldots, \tau_{1,K})$ and with the variance $\sigma_\varepsilon^2 = 1$ held fixed as used in the latent ordered probit regression. In addition, following Park and Casella (2008), we use the following independent (hyper) priors

- $\tau_{0,1}, \tau_{0,2}, \ldots, \tau_{0,K} \mid \lambda_0^2 \overset{ind}{\sim} Exp(\lambda_0^2/2)$, and
- $\tau_{1,1}, \tau_{1,2}, \ldots, \tau_{1,K} \mid \lambda_1 \overset{ind}{\sim} Exp(\lambda_1^2/2)$,

where $Exp(s)$ denotes the exponential distribution with mean value $1/s$. Then, the (hyper) prior for the two global LASSO parameters λ_0^2 and λ_1^2 is given by independent distributions

[4] For a detailed discussion of alternative solutions to the label switching problem, see Frühwirth-Schnatter (2006).

- $\lambda_0^2 \sim \text{Gamma}(\bar{r}, \bar{\delta})$, and

- $\lambda_1^2 \sim \text{Gamma}(\bar{r}, \bar{\delta})$,

where $\bar{r} = 3$ and $\bar{\delta} = 1$. Finally, we have priors for the Markov switching probabilities, corresponding to the parameters in (3.3), given by

- $p \sim \mathcal{B}(\bar{u}_{0,0}, \bar{u}_{0,1})$ and $q \sim \mathcal{B}(\bar{u}_{1,0}, \bar{u}_{1,1})$,

with $\bar{u}_{0,0}$, $\bar{u}_{0,1}$, $\bar{u}_{1,0}$ and $\bar{u}_{1,1}$ all set equal to 0.5, as per the algorithm of Kim and Nelson (2001). Collectively, the joint prior distribution is specified over the entire collection of unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\psi}, p, q, \boldsymbol{\tau_0}, \boldsymbol{\tau_1}, \lambda_0^2, \lambda_1^2)$ is given by

$$\begin{aligned} p(\boldsymbol{\theta}) &= p(\boldsymbol{\mu})\ p(\boldsymbol{\sigma^2})\ p(\mathbf{c}) \\ &\times p(\boldsymbol{\beta_0} \mid \boldsymbol{\tau_0})\ p(\boldsymbol{\beta_1} \mid \boldsymbol{\tau_1}) \\ &\times p(\boldsymbol{\tau_0} \mid \lambda_0^2)\ p(\boldsymbol{\tau_1} \mid \lambda_1^2)\ p(p)\ p(q). \end{aligned}$$

Appendix B3: MCMC Algorithm

The marginal posterior distribution is obtained using a basic Gibbs sampling approach, where the parameters and latent variables are each drawn recursively from the relevant (full) conditional posteriors. Given the prior distribution, the g^{th} iteration of the Gibbs sampler proceeds as follows:

We note that in Step 3 each shrinkage parameter, either λ_0^2 and λ_1^2 (or λ^2 in the static case), is generated by first sampling an augmentation vector, $\tau_0^{(g)}$ and $\tau_0^{(g)}$, respectively, from the corresponding distribution that conditions on the relevant previous draw of the shrinkage parameter. This approach follows as per Park and Casella (2008). The new draws of the shrinkage parameters are then sampled from the full conditional distributions that utilize the augmentation vectors, i.e. $\lambda_0^{2(g)} \sim p(\lambda_0^2 \mid \tau_0^{(g)}, \boldsymbol{\beta}_0^{(g)})$ and $\lambda_1^{2(g)} \sim p(\lambda_1^2 \mid \tau_1^{(g)}, \boldsymbol{\beta}_1^{(g)})$, respectively. The values of the $\tau_0^{(g)}$ and $\tau_1^{(g)}$ are not required for any additional part of the MCMC algorithm and may be discarded at the end of each iteration. We also

Algorithm 9 Gibbs sampling algorithm for posterior simulation

```
Inputs: $\mathbf{y}, \mathbf{x}$: data observations; $G$: number of iterations; $\theta^{(0)} = \left(\boldsymbol{\psi}^{(0)}, p^{(0)}, q^{(0)}, \boldsymbol{\tau_0}^{(0)}, \boldsymbol{\tau_1}^{(0)}, \lambda_0^{(0)}, \lambda_1^{(0)}\right)$: initial value; $\mathcal{TN}_{(c_{z_i^*-1}, c_{z_i^*})}(\mathbf{x}_i'\beta_{S_{t_i}}, 1)$, $\mathcal{N}(D_{S_{t_i}} d_{S_{t_i}}, \Gamma(\bar{\bar{r}}, \bar{\bar{\delta}})$, $\mathcal{U}(l_j, u_j)$, $\mathcal{B}(\bar{u}_{0,0} + n_{0,0}, \bar{u}_{0,1} + n_{0,1})$, $\mathcal{B}(\bar{u}_{1,1} + n_{1,1}, \bar{u}_{1,0} + n_{1,0})$, $\mathcal{TN}_{(\mu_{j-1}, \mu_{j+1})}(D_{\mu_j} d_{\mu_j}, D_{\mu_j})$ : full conditional posterior distributions;
for $i = 1 \to G$ do
    Generate the mixture indicators and the predictive scores
        $p(z_i^*|\mathbf{y}, \boldsymbol{\mu}', \boldsymbol{\sigma}^{2'}, \boldsymbol{c}, \boldsymbol{\beta_0}, \boldsymbol{\beta_1}) \propto p(z_i^*|w_i)p(w_i|\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\sigma}^{2'}, \boldsymbol{c}, \boldsymbol{\beta_0}, \boldsymbol{\beta_1})$
        $p(z_i|\mathbf{c}, \beta_0, \beta_1, z_i^*) \sim \mathcal{TN}_{(c_{z_{i-1}^*}, c_{z_i^*})}(\mathbf{x}_i'\beta_{S_{t_i}}, 1);$
    Generate the regression parameters
        $p(\beta_{S_{t_i}}|\mathbf{y}, \mathbf{z}, \mathbf{S}, \tau) \sim \mathcal{N}(D_{S_{t_i}} d_{S_{t_i}}, D_{S_{t_i}});$
    Generate the shrinkage parameters (via augmentation)
        $p(\lambda^2|\tau^2) \sim \Gamma(\bar{\bar{r}}, \bar{\bar{\delta}});$
    Generate each of the $J$ cut-points
        $p(c_j|c_{/j}, \mathbf{z}, \mathbf{z}^*) \sim \mathcal{U}(l_j, u_j);$
    Generate the latent Markov states using FFBS
        $\mathbf{S}^{(g)}|y, \mathbf{z}^{(g)}, \boldsymbol{\beta_0}^{(g)}, \boldsymbol{\beta_1}^{(g)}, p^{(g-1)}, q^{(g-1)};$
    Generate the Markov transition probabilities
        $p|\mathbf{S} \sim \mathcal{B}(\bar{u}_{0,0} + n_{0,0}, \bar{u}_{0,1} + n_{0,1})$ and $q|\mathbf{S} \sim \mathcal{B}(\bar{u}_{1,1} + n_{1,1}, \bar{u}_{1,0} + n_{1,0});$
    Generate the vector of mean parameters for the Gaussian mixture distribution, for $j = 1, \cdots, J$,
        $\mu_j|\mathbf{y}, \mathbf{z}, \sigma_j^2 \sim \mathcal{TN}_{(\mu_{j-1}, \mu_{j+1})}(D_{\mu_j} d_{\mu_j}, D_{\mu_j});$
    Generate the vector of variance parameters for the Gaussian mixture distribution, for $j = 1, \cdots, J$,
        $p(\sigma_j^2|\mathbf{y}, \mathbf{z}^*, \mu_j) \sim IG(\bar{\bar{a}}_j, \bar{\bar{b}}_j);$
Outputs: A sample of $G$ draws of static parameter $\theta$ and latent variable $S_{0:T}$ from $p(\theta, S_{0:T}|\mathbf{y}, \mathbf{x})$.
```

note for Step 4 that $\boldsymbol{c}_{/j} = \{c_0, c_1, \ldots, c_{j-1}, c_{j+1}, \ldots, c_J\}$, denoting the vector $\mathbf{c}$ but with the j^{th} element excluded. Since the priors are conditionally conjugate for all unknowns, the relevant conditional posterior distributions are all derived analytically, ensuring a fast algorithm for sampling from the full joint posterior distribution.

Mixture indicator vector $\boldsymbol{z}^*$ and the latent predictive score vector $\boldsymbol{z}$

The mixture indicator variable components $z_1^*, \cdots, z_N^*$ are conditionally independent and hence are be sampled independently from multinomial distributions with probabilities

$$p(z_i^*|\mathbf{y}, \boldsymbol{\mu}', \boldsymbol{\sigma}^{2'}, \boldsymbol{c}, \boldsymbol{\beta_0}, \boldsymbol{\beta_1}) \propto p(z_i^*|w_i)p(w_i|\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\sigma}^{2'}, \boldsymbol{c}, \boldsymbol{\beta_0}, \boldsymbol{\beta_1}),$$

with diffuse priors on $\boldsymbol{w}$, we get

$$w_{i,j} = \frac{\left[\Phi(\mathbf{x}_i'\beta_{S_{t_i}} - c_{j-1}) - \Phi(\mathbf{x}_i'\beta_{S_{t_i}} - c_j)\right] \ \phi(y_i; \mu_j, \sigma_j^2)}{\sum_{j=1}^J \left[\Phi(\mathbf{x}_i'\beta_{S_{t_i}} - c_{j-1}) - \Phi(\mathbf{x}_i'\beta_{S_{t_i}} - c_j)\right] \ \phi(y_i; \mu_j, \sigma_j^2)},$$

for $i = 1, \cdots, n$ and $j = 1, \cdots, J$, where $\Phi(\cdot)$ denotes the cdf of a standard normal random variable.

Conditional on the sampled mixture indicator variable z_i^* (as well as on the relevant latent credit state coefficient $\boldsymbol{\beta_{S_{t_i}}}$ corresponding to the latent credit state at the time of default for RR_i, the data $\mathbf{y}$, other parameters), the latent data for individual i, given by z_i, is generated from a truncated normal distribution

$$p(z_i|\mathbf{c}, \beta_0, \beta_1, z_i^*) \sim \mathcal{TN}_{(c_{z_{i-1}^*}, c_{z_i^*})}(\mathbf{x}_i'\beta_{S_{t_i}}, 1),$$

where $c_{z_{i-1}^*}$ and $c_{z_i^*}$ are the lower bound and the upper bound parameters.

State-dependent regression coefficients $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$

The latent data $\mathbf{z}$ and recovery determinants $\boldsymbol{x}$ are divided into $\mathbf{Z}_0$, $\mathbf{Z}_1$ and $\boldsymbol{X}_0 = [\boldsymbol{x}_{1,0}, \boldsymbol{x}_{2,0}, \cdots, \boldsymbol{x}_{k,0}]$, $\boldsymbol{X}_1 = [\boldsymbol{x}_{1,1}, \boldsymbol{x}_{2,1}, \cdots, \boldsymbol{x}_{k,1}]$ respectively, according the latent

Markov state S_{t_i}. Given the data, $\mathbf{y}$ and $\mathbf{z}$, Markov states, S_{t_i}, and the local shrinkage parameter, $\tau_0^2 = (\tau_{1,0}, \cdots, \tau_{k,0})$ and $\tau_1^2 = (\tau_{1,1}, \cdots, \tau_{k,1})$, the conditional posterior distribution for $\beta_{S_{t_i}} = (\beta_0, \beta_1)'$ is given by

$$p(\beta_{S_{t_i}}|\mathbf{y}, \mathbf{z}, \mathbf{S}, \tau) \sim \mathcal{N}(D_{S_{t_i}} d_{S_{t_i}}, D_{S_{t_i}}),$$

where

$$D_{S_{t_i}} = (\mathbf{x}'_{S_{t_i}} \mathbf{x}_{S_{t_i}} + \text{diag}(\tau^2)^{-1})^{-1} \text{ and } d_{S_{t_i}} = \mathbf{x}'_{S_{t_i}} y_{S_{t_i}}.$$

Shrinkage parameters τ^2 and λ^2

For each credit state, the local shrinkage parameters $\tau_1^2, \cdots, \tau_k^2$ are conditionally independent, with

$$p(1/\tau_j^2|\beta_j, \lambda^2) \sim InvGaussian(\bar{\bar{\mu}}_j, \lambda^2),$$

where *InvGaussian* denotes an Inverse Gaussian distribution and

$$\bar{\bar{\mu}}_j = \sqrt{\frac{\lambda^2}{\beta_j^2}},$$

for $j = 1, \cdots, k$. With a conjugate prior, the full conditional distribution of λ^2 is given by

$$p(\lambda^2|\tau^2) \sim \Gamma(\bar{\bar{r}}, \bar{\bar{\delta}}),$$

where Γ denotes a gamma distribution with shape parameter $K + \bar{r}$ and rate parameter $\sum_{j=1}^{K} \tau_j^2/2 + \bar{\delta}$.

Cut-points c

We follow Albert and Chib (1993) and use diffuse priors for all cut-points $c_2, \cdots, c_{J-1}$. For identification purpose, we set $c_0 = -\infty$, $c_1 = 0$ and $c_J = \infty$ as it is common in any other studies using an ordered probit model. The joint conditional posterior for the cut-points $j = 2, \cdots, J-1$ (recall that $c_0 = -\infty, c_1 = 0$, and $c_J = \infty$) is given by,

$$p(c_j | c_{/j}, \mathbf{z}, \mathbf{z}^*) \sim \mathcal{U}(l_j, u_j),$$

where $\mathcal{U}(l_j, u_j)$ denotes a uniform distribution with

$$l_j = \max\{c_{j-1}, \max\{z_i : z_i^* = j\}\},$$
$$u_j = \min\{c_{j+1}, \min\{z_i : z_i^* = j+1\}\}.$$

The latent Markov states S

We use the efficient block sampling algorithm of Carter and Kohn (1994) and Frühwirth-Schnatter (1994) to generate $\mathbf{S}$, which is known as *forward filtering, backward sampling* (FFBS). Hamilton (1989) provides the following filtering algorithm to calculate the filtered probabilities for $\mathbf{S}$. Let z_t be vector contains all z_i observed in year t, the Hamilton filter consists of, for $t = 1, \cdots, T$,

- predict

$$\begin{bmatrix} \Pr(S_t = 0 | z_t) \\ \Pr(S_t = 1 | z_t) \end{bmatrix} = \begin{bmatrix} p & 1-q \\ 1-p & q \end{bmatrix} \begin{bmatrix} \Pr(S_t = 0 | z_{t-1}) \\ \Pr(S_t = 1 | z_{t-1}) \end{bmatrix},$$

- update

$$\Pr(S_t = 0 | z_t) \propto p(z_t | S_t = 0) \Pr(S_t = 0 | z_t) \text{ and}$$
$$\Pr(S_t = 1 | z_t) \propto p(z_t | S_t = 1) \Pr(S_t = 1 | z_t),$$

where

$$p(z_t|S_t = 0) = \prod_{i=1}^{n_t} p(z_{i_t}|S_t = 0) \text{ and}$$
$$p(z_t|S_t = 1) = \prod_{i=1}^{n_t} p(z_{i_t}|S_t = 1).$$

The latent Markov states are then simulated sequentially, for $t = T, T-1, ..., 1$. Given S_{t+1}, the parameter in the Bernoulli distribution for each t is calculated by

$$\Pr(S_t = 1|z_{1:T})/(\Pr(S_t = 0|z_{1:T}) + \Pr(S_t = 1|z_{1:T})),$$

where

$$\Pr(S_t = 0|z_{1:T}) \propto \Pr(S_t = 0|z_t) \Pr(S_t = 0|S_{t+1}) \text{ and}$$
$$\Pr(S_t = 1|z_{1:T}) \propto \Pr(S_t = 1|z_t) \Pr(S_t = 1|S_{t+1}).$$

The Markov transition probabilities p and q

Conditional on $\mathbf{S}$, the transition probabilities, p and q are independent of the data. Since we have assigned beta prior distributions to the transition probabilities, the conditional posterior distributions are given by

$$p(p, q|\mathbf{S}) \propto p(p, q)p(\mathbf{S} \mid p, q),$$

where $p(\mathbf{S}|p, q)$ describes the joint probabilities associated with the latent Markov-switching states. Prior independence (of p and q) implies posterior independence here, and hence p and q may be jointly sampled according to

$$p|\mathbf{S} \sim \mathcal{B}(\bar{u}_{0,0} + n_{0,0}, \bar{u}_{0,1} + n_{0,1}) \text{ and}$$
$$q|\mathbf{S} \sim \mathcal{B}(\bar{u}_{1,1} + n_{1,1}, \bar{u}_{1,0} + n_{1,0}),$$

where $\mathcal{B}(a,b)$ denotes the Beta distribution on $(0,1)$, having mean $\frac{a}{a+b}$ and variance $\frac{ab}{(a+b)^2(a+b+1)}$, here with

$$n_{1,0} = \sum_{t=1}^{T}\sum_{i=1}^{n_t} S_{t_i}|S_{t-1} = 0,$$
$$n_{1,1} = \sum_{t=1}^{T}\sum_{i=1}^{n_t} S_{t_i}|S_{t-1} = 1,$$

and $n_{0,0} = n_{S_t=0} - n_{1,0}$ and $n_{0,1} = n_{S_t=1} - n_{1,1}$.

The Gaussian mixture means $\boldsymbol{\mu}$

Given the independent conjugate priors, the individual μ_j, for $j = 1, \cdots, J$, may be sampled independently from

$$\mu_j|\mathbf{y}, \mathbf{z}, \sigma_j^2 \sim \mathcal{TN}_{(\mu_{j-1},\mu_{j+1})}(D_{\mu_j} d_{\mu_j}, D_{\mu_j}),$$

where

$$D_{\mu_j} = \left(\sum_{i=1}^{n} \mathbb{I}(z_i^* = j)/\sigma_j^2 + \bar{V}_{\mu_j}^{-1}\right)^{-1},$$

and

$$d_{\mu_j} = \sum_{i=1}^{n} \mathbb{I}(z_i^* = j) y_i/\sigma_j^2 + \bar{V}_{\mu_j}\bar{\mu}_j.$$

The Gaussian mixture variances $\boldsymbol{\sigma}^2$

The individual variance parameters σ_j^2 for mixture components $j = 1, \cdots, J$, are sampled independently conditional on μ_j and $\mathbf{z}^*$ given by

$$p(\sigma_j^2|\mathbf{y}, \mathbf{z}^*, \mu_j) \sim IG(\bar{\bar{a}}_j, \bar{\bar{b}}_j),$$

with

$$\bar{\bar{a}}_j = \frac{\sum_{i=1}^{n} \mathbb{I}(z_i^* = j)}{2} + \bar{a}_j,$$

and

$$\bar{\bar{b}}_j = \bar{b}_j^{-1} + \frac{1}{2}\sum_{i=1}^{n} \mathbb{I}(z_i^* = j)(y_i - \mu_j)^2.$$

Chapter 4

Monitoring Macroeconomic Linkages with a semi-parametric VAR Model

This chapter proposes an extension to the Vector AutoRegressive (VAR) model for economic variables observed with mixed frequency. The modelling framework employs the notion that certain variables observed only at low frequency (e.g. quarterly) have latent (unobserved) high frequency (e.g. monthly) values that are “missing”. Such variables are nevertheless related to several high frequency observed variables through a regression framework. Bayesian analysis of the model, which is specified in hierarchical form, is amenable to the use of MCMC methods. We consider the case where the latent variable evolves nonlinearly, with the non-linearity specified non-parametrically using a Gaussian Process prior. The proposed modeling framework may be applied to progressively “nowcast” the present state of a set of macroeconomic and financial variables that are only reported at the end of each quarter. A Monte Carlo study is used to demonstrate the effectiveness of the proposed MCMC sampler.

4.1 Introduction

Investigating the relationships among economic variables is one of the most important challenges to both practitioners and scholars. Modeling such multivariate time series, however, is more challenging since relevant measures of variables are often sampled at different frequencies. In practice, high frequency measures between two consecutive low frequency measures may be temporally aggregated, averaged or even discarded. Economic analyses involve different frequencies is then conducted based on a joint process sampled at a common low frequency only. Such practice has consequences potentially mis-specifying the co-movements among variables. See, for example, Andreou, Ghysels, and Kourtellos (2010) and Ghysels (2016) for a theoretical exploration of the problem.

In tandem, the nonlinearities between macroeconomic variables are well-documented in the literature. Kim, Osborn, and Sensier (2005) investigate the apparent nonlinearities present in the monetary policy rule of the U.S. Federal Reserve and provide significant evidence of nonlinearity for the period from 1960 to 1979. Surico (2007) finds a similar nonlinearity between Fed's monetary policy rule and U.S. inflation for the period from 1960 to 2003. Concerned with real GDP and oil price, Hamilton (2003) characterizes the nonlinearity using a flexible parametric framework and concludes oil price increases are much more important than oil price decreases. However, the natural question that arises at this point is: are these relationships intrinsically nonlinear, or as suggested by Andreou, Ghysels, and Kourtellos (2010), are their conclusions consequence of an inappropriate aggregation scheme? We provide a methodological approach to answer this question.

We introduce an extension of the Vector Autoregressive (VAR) model (Sims,

1980) which presents the following novel features (1) accommodating data sampled at different frequencies and (2) enabling generic forms of nonlinearity between variables. Our approach delivers inference and 'nowcasting' as a by-product for the 'missing' observations of low-frequency variables obtained via an efficient and general MCMC algorithm developed for computing the relevant marginal posterior distributions. Earlier literature on this topic includes Harvey and Pierse (1984), Mariano and Murasawa (2003), Mariano and Murasawa (2010), Eraker et al. (2014). However, notwithstanding the considerable attentions received in the literature, none of these consider nonlinearities. The proposed mixed frequency VAR model exploits a state space representation of the VAR model, relying on nonlinear filtering and smoothing techniques to impute state variable that is used to predict future dynamics. To enable efficient inference, a simple yet intuitive filtering method for models involving nonparametric functions is introduced. Unlike the filtering method proposed by Frigola et al. (2013) in the context of Gaussian process state space models, the filtering methodology developed here does not rely on particle MCMC and instead it builds upon a modified version of the nonlinear but parametric method of Stroud, Müller, and Polson (2003).

Given the mixed frequency nature of the current setting, it is worthwhile to review the related MIxed-DAta Sampling (MIDAS) approach proposed by Ghysels, Santa-Clara, and Valkanov (2004). Unlike the state space form used here, and in its simplest form, the MIDAS approach involves a single equation in which the high frequency variable is regressed on the low frequency one through a lag polynomial function. In particular, the lag polynomial function is parsimoniously parameterized by a small set of hyperparameters and it projects the high frequency process into the low frequency one with a data consistent weighting scheme. The MIDAS regression has been used in a variety of economic studies, see, for example, Ghysels, Santa-Clara, and Valkanov (2005), Ghysels, Santa-Clara, and Valkanov (2006), Ghysels (2016), Pettenuzzo, Timmermann, and Valkanov (2016), among others. Notably, the relationship between MIDAS regression and state space models and their relative benefits and drawbacks are reviewed by Bai, Ghysels, and Wright (2013) and more recently, Ghysels (2016) also considers a mixed frequency VAR model based on a multivariate extension of the standard MIDAS regression model. Despite its recent popularity, arguably owing to its simplicity, the potential to accommodate either nonlinearities or nonparametrics, or both, remains unclear. Moreover, MIDAS regressions are estimated by nonlinear least squares and inferences are undertaken using frequentist approaches.

From a Bayesian perspective, our model contains nonparametric dynamics, avoiding the negative consequences of mis-specified nonlinearity on the estimation of co-movement. For instance, if the dynamics are mistakenly modeled linearly or via a low-order polynomial, the conditional distribution of the responses and the joint distribution of the errors may appear to be non-Gaussian even if the true error distribution is Gaussian. There is considerable work on circumventing parametric assumptions under both Bayesian and Frequentist framework, for instance, the partially linear model (Robinson, 1988), the generalized additive model (Hastie and Tibshirani, 1986; Smith and Kohn, 1996) and the semi-parametric single index model (Manski, 1988; Ichimura, 1993). In the current work, a Gaussian process, GP hereafter, prior is placed over the nonlinear functions, resulting in a flexible model able to capture complex dynamic structure. In the GP latent variable model literature, inference on the posterior distribution of the unknown functions often relies on posterior optimization approaches, e.g., variational methods (Lawrence, 2004; Titsias and Lawrence, 2010 and Ko and Fox, 2011). However, in practice it is nearly impossible to know how their approaches perform without comparing the results to a posterior simulation method. To our best knowledge, the only exact (up to simulation error) Bayesian inferential methodology available for GP involving latent variable is Frigola et al. (2013)'s method, which is based on a specially tailored particle MCMC sampling scheme. The filtering technique proposed in this paper is more general in the sense that it can be calibrated to adapt other Bayesian nonparametric methods with some simple modifications. The performance of the proposed MCMC sampler is demonstrated in a Monte Carlo simulation study, where two different forms of nonlinearities are considered in a bi-variate VAR model, with monthly-quarterly type mixed frequency data.

The reminder of the chapter is organized as follows. In Section **??**, we present the semi-parametric VAR model that accommodates mixed frequencies and nonlinearities, along with the corresponding state space representation. In Section 4.2, we discuss the Bayesian estimation method for the proposed model and the corresponding MCMC simulation scheme. A Monte Carlo simulation study is then used to evaluate the proposed MCMC scheme, along with a discussion of the simulation results, are presented in Section 4.3.

The VAR model contains equations for a set of $J = J_1 + J_2$ variables, where we assume J_1 are sampled at high frequency while J_2 other variables are sampled at a lower frequency. In particular, it is assumed that y_1 is always observed while y_2 is only observed every m-th period. The dynamic structure of y_1 and y_2 may be parametric or nonparametric, depending on whether it is believed to be of some unknown form. In the case of containing a mixture of parametric and nonparametric functions, the VAR system can be seen as a multivariate extension of the partially linear time-series model. We refer to Härdle, Liang, and Gao (2012) for a more detailed discussion of partially linear models.

In detail, suppose that a macro-economy can be represented by the following partially linear VAR(p) model:

$$y_t^* = A + \sum_{l=1}^{p} \left((\iota_J \iota_J' - C_l) \circ B_l \right) y_{t-l}^* + \sum_{l=1}^{p} C_l \circ \boldsymbol{G}_l \left(\iota_J \otimes y_{t-l}^{*'} \right) \iota_J + \varepsilon_t, \tag{4.1}$$

$$\varepsilon_t \sim \mathcal{N}(0, \Sigma)$$

where $\otimes$ denotes the Kronecker product, $\circ$ denotes the Hadamard (element-wise) product and ι_J represents a J-sized column vector of ones. The vector of disturbance, ε_t, is assumed to follow a multivariate Normal distribution with mean zero and time-constant covariance matrix Σ. In particular, y_t^* consists of a J_1-sized vector x_t and a J_2-sized vector z_t^* of which x_t is always observed and z_t is only observed every m-th period. The 'missing values' between two consecutive observation of z_t is augmented and we denote the augmented vector by $z_t^* = \{z_1, z_2^*, \cdots, z_m^*, z_{m+1}, \cdots\}$, as such $y_t^* = (x_t, z_t)' = y_t$ if z_t is observed and $y_t^* = (x_t, z_t^*)'$ otherwise. We further denote by A a vector of intercept, by B_l a $J \times J$ coefficient matrix associated with the lagged dependent variables of y_t^* that represents linear relationships between y_t^* and the lagged values of y_t^*, and by C_l a $J \times J$ selector matrix in which the lagged variables that are nonlinearly related to the dependent variable are selected. For instance, C_1 may take the form of

$$C_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

a hollow matrix in a bi-variate VAR model, if the interest is to investigate the functional forms of the evolving dynamics between x_t and z_{t-1}^*, as well as z_t^* and x_{t-1} in a bi-variate VAR(1) setting. Importantly, we denote by $\boldsymbol{G}_l$ a non-specified generic matrix function which reflects nonlinear relationships between y_t^* and the lagged y_t^*. For notional simplicity, the operator $\boldsymbol{G}_l(\cdot)$ is a matrix of functions defined in the following. For given a $J \times J$ matrix, for instance,

$$\iota_J y_{t-l}' = \begin{bmatrix} y_{t-l,1} & \cdots & y_{t-l,J} \\ \vdots & & \\ y_{t-l,1} & \cdots & y_{t-l,J} \end{bmatrix},$$

$\boldsymbol{G}_l(\iota_J y'_{t-l})$ yields

$$\boldsymbol{G}_l(\iota_J y'_{t-k}) = \begin{bmatrix} G_{l,11}(y_{t-l,1}) & \cdots & G_{l,1J}(y_{t-l,J}) \\ \vdots & \ddots & \vdots \\ G_{l,J1}(y_{t-l,1}) & \cdots & G_{l,JJ}(y_{t-l,J}) \end{bmatrix}.$$

In addition, the nonparametric functions are only identified up to a constant term in the additive nonparametric structure. For instance, $G_{ij}(\cdot) + G_{ih}(\cdot) = G^*_{ij}(\cdot) + G^*_{ih}(\cdot)$ implies likelihood is unchanged if we define $G^*_{ij}(\cdot) = G_{ij}(\cdot) + a$ and $G^*_{ih}(\cdot) = G_{ih}(\cdot) - a$ for some constant a. As a result, we restrict $G_{l,ij}(0) = 0$, as such an identification restriction is imposed in the similar spirit of Shively, Kohn, and Wood (1999), namely, A is used as the overall intercept of the regression.

It is worth noting that the normality assumption in conjunction with nonparametric functions is much more flexible than it may seem. A similar specification is used by Chib, Greenberg, and Jeliazkov (2009) in modeling data in the presence of endogeneity and sample selection. In the context of macroeconomic policy evaluation, this way of modeling provides more insight into the potential non-monotonic marginal effects, including the feedback effect of a macro-economic shock to the system. In contrast to its alternative, modeling error distribution flexibly but restrict the mean to be linear, the latter is less favorable since the marginal effects of the lagged dependent variables are always monotonic.

Finally, we illustrate the details regarding the specifications of the proposed model using a simple bi-variate VAR(1) setup in the following. Now, consider a simple partially linear VAR(1) model given by

$$y^*_t = A + ((\iota_2\iota'_2 - C) \circ B)\, y^*_{t-1} + C \circ G(\iota_2 \otimes (y^*_{t-1})')\iota_2 + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma), \tag{4.2}$$

which can be written as

$$\begin{bmatrix} x_t \\ z_t^* \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} (1-c_{11})\beta_{11} & (1-c_{12})\beta_{12} \\ (1-c_{21})\beta_{21} & (1-c_{22})\beta_{22} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ z_{t-1}^* \end{bmatrix} \\ + \begin{bmatrix} c_{11}G_{11}(x_{t-1}) + c_{12}G_{12}(z_{t-1}^*) \\ c_{21}G_{21}(x_{t-1}) + c_{22}G_{22}(z_{t-1}^*) \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}, \quad \varepsilon_t \sim \mathcal{N}(0,\Sigma), \tag{4.3}$$

with

$$y_t^* = \begin{bmatrix} x_t \\ z_t^* \end{bmatrix}, \; A = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}, \; B = \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix}, \; C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \; \varepsilon_t = \begin{bmatrix} \varepsilon_{x,t} \\ \varepsilon_{z,t} \end{bmatrix},$$

$$G = \begin{bmatrix} G_{11}(\cdot) & G_{12}(\cdot) \\ G_{21}(\cdot) & G_{22}(\cdot) \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}. \tag{4.4}$$

In the case of higher order VARs, it is straightforward to devise a similar representation by exploiting the companion form of VAR and rewrite the VAR(p) into a VAR(1).

4.1.1 The Likelihood Function

The development of the algorithm is initiated with a discussion of the likelihood function of model (4.3). For the following, we define the notations for a bivariate VAR

$$\begin{aligned} y_t^* &= (x_t, z_t^*)', \\ G_j(y_t) &= (G_{1j}(x_t), G_{2j}(z_t^*))', \\ G(y_t) &= (G_{11}(x_t), G_{12}(z_t^*), G_{21}(x_t), G_{22}(z_t^*))'. \end{aligned}$$

For simplicity of exposition, we assume $T/m \in \mathbb{N}$, that is, T is a multiple of m. Let $\mathbb{1}_t$ denote the indicator variable such that $\mathbb{1}_t = 1$ indicates that z_t is observed and $\mathbb{1}_t = 0$ indicates that z_t is non-observed. We also set $\mathcal{T}_1 = \{t : \mathbb{1}_t = 1\}$ to be the T/m observations when $y_t = (x_t, z_t)$ is fully observed and $\mathcal{T}_2 = \{t : \mathbb{1}_t = 0\}$ to be observations when z_t is non-observed. Finally, let $\boldsymbol{\theta}$ to be the set of all parameters.

The complete data likelihood function of the observed data and latent observations conditional on $\boldsymbol{\theta}$ is given by

$$\begin{aligned} f(\mathbf{y}^*|\boldsymbol{\theta}) &= \prod_{t=1}^{T} f(y_t^*|y_{t-1}^*, \boldsymbol{\theta}) \\ &= \left[\prod_{t\in\mathcal{T}_1} f(y_t|y_{t-1}^*, \boldsymbol{\theta})\right]\left[\prod_{t\in\mathcal{T}_2} f(y_t^*|y_{t-1}^*, \boldsymbol{\theta})\right], \end{aligned} \tag{4.5}$$

where the second product comes from the fact that $y_{t-1}^* = y_{t-1}$ if z_{t-1} is observed. We have, for $t \in \mathcal{T}_2$,

$$f(y_t^*|y_{t-1}^*, \boldsymbol{\theta}) \propto |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\varepsilon_t^{*\prime}\Sigma^{-1}\varepsilon_t^*\right\},$$

where $\varepsilon_t^* = \mathbf{y}_t^* - A - ((\iota_2\iota_2' - C)\circ B)\,\mathbf{y}_{t-1}^* - C \circ G\left(\iota_2' \otimes \mathbf{y}_{t-1}^*\right)\iota_2$, and for $t \in \mathcal{T}_1$,

$$f(y_t|y_{t-1}^*, \boldsymbol{\theta}) \propto |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\varepsilon_t'\Sigma^{-1}\varepsilon_t\right\},$$

where $\varepsilon_t = y_t - A - ((\iota_2\iota_2' - C)\circ B)\,y_{t-1}^* - C \circ G\left(\iota_2' \otimes (y_{t-1}^*)'\right)\iota_2$. As is standard for latent variable models, the likelihood function $f(y_t|\boldsymbol{\theta})$ is obtained by integrating $f(y_t|\mathbf{z}_t^*, \boldsymbol{\theta})$ over the latent variables $\mathbf{z}_t^*$, which is difficult since $\mathbf{z}_t^*$ is inside the nonparametric function G whose uncertainty must also be accounted for, and such integration involves nonlinear filtering and smoothing.

In order to exploit the existing literature, we introduce a state space representation of the mixed frequency VAR(1) model. It consists of the state equation given by (4.2) and the measurement equation

$$y_t = \mathbb{D}_t y_t^* + (I_2 - \mathbb{D}_t) v_t,$$

where

$$\mathbb{D}_t = \begin{bmatrix} 1 & 0 \\ 0 & \mathbb{1}_t \end{bmatrix} \text{ and } v_t = \begin{bmatrix} 0 \\ \nu_t \end{bmatrix},$$

and ν_t is a random draw from a distribution that does not depend on $\boldsymbol{\theta}$. Therefore, our discussion on the algorithm for posterior simulation is based on

$$\begin{aligned} y_t &= \mathbb{D}_t y_t^* + (I_2 - \mathbb{D}_t) v_t \\ y_t^* &= A + \left((\iota_2 \iota_2' - C) \circ B \right) y_{t-1}^* + C \circ G \left(\iota_2' \otimes (y_{t-1}^*)' \right) \iota_2 + \varepsilon_t. \end{aligned}$$

4.1.2 Prior Distributions

The prior distribution for the parameters and nonparametric functions is specified in this section. To increase the tractability of posterior distributions, conjugate priors are considered in most cases. We assume that A and B jointly follow a multivariate normal distribution with mean $\boldsymbol{b_0}$ and variance $\boldsymbol{B_0}$ and that covariance Σ has a inverse Wishard distribution with hyper-parameters $\boldsymbol{\nu}$ and $\boldsymbol{Q}$,

$$\pi(A, B, \Sigma) = \mathcal{N}(A, B | \boldsymbol{b_0}, \boldsymbol{B_0}) \mathcal{IW}(\Sigma | \boldsymbol{\nu}, \boldsymbol{Q}),$$

where $\mathcal{N}(A, B | \boldsymbol{b_0}, \boldsymbol{B_0})$ is the density for the multivariate normal distribution and $\mathcal{IW}(\Sigma | \boldsymbol{\nu}, \boldsymbol{Q})$ is the density for inverse Wishard distribution.

We model each of the unknown functions through the GP priors. Gaussian Processes (GPs) (Rasmussen and Williams, 2006; O'Hagan and Kingman, 1978) are a flexible approach for modeling functions from data. From a Bayesian perspective, the use of GPs enable a complete prior distribution to be defined over the values of a function, which is an infinite dimensional object. This prior can then be updated in light of either a noisy or perfect observation of the function over a (finite) set of locations. As its name suggests, a GP is the extension of the Gaussian distribution to functions. In a GP all the function evaluations (denoted by g) are jointly Gaussian given m, k and the data, i.e.,

$$\begin{bmatrix} g(x_1) \\ \vdots \\ g(x_{T-1}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_{T-1}) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_{T-1}) \\ \vdots & \ddots & \vdots \\ k(x_{T-1}, x_1) & \cdots & k(x_{T-1}, x_{T-1}) \end{bmatrix} \right),$$

where $x_1, \cdots, x_{T-1}$ are a set of input locations and $k(\cdot, \cdot)$ is the covariance function that takes two arguments, $k(x_1, x_2)$, and returns the covariance between their corresponding function values,

$$\mathrm{cov}(g(x_1), g(x_2)) = k(x_1, x_2).$$

The shorthand notation of GP, $g(X) \sim \mathcal{N}(m(X), K(X, X))$, is mostly used in what follows. GPs has received considerable attention and now has several variants and extensions, for instance, piecewise Gaussian processes (Kim, Mallick, and Holmes, 2005), deep Gaussian processes (Damianou and Lawrence, 2013), sparse Gaussian processes (Snelson and Ghahramani, 2006) and also a interesting extension, nonlinear principle component analysis (Lawrence, 2005). We note that other nonparametric modeling approaches could be pursued for the unknown functions without modifying the way of accommodating mixed frequencies in our setting. These include Markov process smoothness priors (Chib and Greenberg, 2007; Chib, Greenberg, and Jeliazkov, 2009), regression splines (Smith and Kohn, 1996) and b-spline priors (Silverman, 1985). See Denison (2002) for a discussion of different nonparametric modeling approaches from the Bayesian perspective.

GPs have the major advantage of producing tractable posterior distributions for function values when Gaussian likelihood is assumed. This can be seen from the fact that the posterior distribution for the function evaluations is obtained through prediction given the data

$$\begin{bmatrix} \mathbf{y} \\ \boldsymbol{g} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(X) \\ m(X) \end{bmatrix}, \begin{bmatrix} K(X,X)+\sigma_\varepsilon^2 I_{T-1} & K(X,X) \\ K(X,X) & K(X,X) \end{bmatrix} \right), \tag{4.6}$$

and the conditional linearity property of multivariate Gaussian distribution yields

$$f(\boldsymbol{g}|\mathbf{y},X) = \mathcal{N}(m(X)+K\left[K+\sigma_\varepsilon^2 I_{T-1}\right]^{-1}(\mathbf{y}-m(X)), K-K\left[K+\sigma_\varepsilon^2 I_{T-1}\right]^{-1}K), \tag{4.7}$$

where $K = K(X,X)$. Similarly, the posterior distribution of function evaluations for a set of test points, denoted $\bar{X}$, is also Gaussian,

$$f(\bar{\boldsymbol{g}}|\mathbf{y},X,\bar{X}) = \mathcal{N}(\mathfrak{m},\mathfrak{s}),$$

where

$$\begin{aligned} \mathfrak{m} &= m(\bar{X}) + \boldsymbol{k}(\bar{X},X)\left[K(X,X)+\sigma_\varepsilon^2 I_{T-1}\right]^{-1}(\mathbf{y}-m(X)) \\ \mathfrak{s} &= k(\bar{X},\bar{X}) - K(\bar{X},X)\left[K(X,X)+\sigma_\varepsilon^2 I_{T-1}\right]^{-1}K(X,\bar{X}). \end{aligned}$$

The covariance function, K, is central to the Gaussian processes, as it encodes our assumptions about the function which we wish to estimate. We consider the 'squared exponential' covariance function, a stationary covariance function that is invariant to translations in the input space. That is, the covariance between two function evaluations only depends on the distance between two points they are evaluated in the input space. This is a reasonable assumption since most of the macroeconomic variables such as, GDP growth, interest rate, involved in the VAR model are stationary. Specifically, the 'squared exponential' covariance function takes the form of

$$k_{SE}(x_i, x_j) = \sigma_f \exp\left(-\frac{(x_i - x_j)^2}{2\ell^2}\right), \tag{4.8}$$

where σ_f and ℓ are hyper-parameters. This covariance function is widely used in Gaussian processes literature. Note that, the covariance function given in (4.8) is infinitely differentiable w.r.t the inputs, which has an important implication to the filtering method proposed. Later, we exploit the differentiability and integrability of the GP posterior distribution to generate proposals of latent data in a Metropolis–Hastings scheme. Finally, the price paid for these advantages is in terms of computational time which is $\mathcal{O}(T^3)$, due to the matrix inversion, $[K(X,X) + \sigma_\varepsilon^2 I_{T-1}]^{-1}$, inside both the posterior mean and variance. Although we note that this issue can be dealt with using existing pseudo marginal MCMC (Andrieu and Roberts, 2009) in conjunction with the sparse Gaussian processes technique (Snelson and Ghahramani, 2006), it is not the main concern of this chapter and we leave it to future research.

4.2 Estimation

In Section 4.2.1, we outline a general Gibbs-based algorithm for posterior simulation. For simplicity of exposition, we present the MCMC algorithms by assuming the selector matrix, as defined in (4.2), to be $C = \iota_2 \iota_2' 1 - I_2$. The corresponding state space representation is then given by

$$\begin{aligned}
\begin{bmatrix} x_t \\ z_t \end{bmatrix} &= \begin{bmatrix} x_t \\ z_t^* \times \mathbb{1}_t \end{bmatrix} + \begin{bmatrix} 0 \\ v_t \times (1 - \mathbb{1}_t) \end{bmatrix} \\
\begin{bmatrix} x_t \\ z_t^* \end{bmatrix} &= \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} \beta_{11} & 0 \\ 0 & \beta_{22} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ z_{t-1}^* \end{bmatrix} + \begin{bmatrix} G_1(z_{t-1}^*) \\ G_2(x_{t-1}) \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix},
\end{aligned} \tag{4.9}$$

where v_t is a random draw from a distribution that does not depend on $\boldsymbol{\theta}$. Mariano and Murasawa (2003) have used this specification in a maximum likelihood setup to ensure the invertibility of the innovation covariance matrix used in the Kalman filter. This specification is also valid in Bayesian inference, that is, the likelihood function defined in (4.5) now becomes

$$f(\mathbf{y}^*|\boldsymbol{\theta}) \prod_{t \in \mathcal{T}_2} f(v_t),$$

and the posterior distributions for all unknowns are equivalent up to scale

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}^*) &= \frac{p(\boldsymbol{\theta}) f(\mathbf{y}^*|\boldsymbol{\theta}) \prod_{t \in \mathcal{T}_1} f(v_t)}{p(\mathbf{y}^*)} \\ &\propto p(\boldsymbol{\theta}) f(\mathbf{y}^*|\boldsymbol{\theta}). \end{aligned}$$

4.2.1 MCMC Algorithm for Posterior Simulation

We define $\boldsymbol{\theta} = (A, B, \Sigma)$ and let $\boldsymbol{\theta}_{/\boldsymbol{\theta}_i}$ represent the elements of $\boldsymbol{\theta}$ other than $\boldsymbol{\theta}_i$ for economy of notation. The posterior distribution for $\boldsymbol{\theta}$ and z_t^* for the semi-parametric model of 4.3 is given by

$$\begin{aligned} \pi(\boldsymbol{\theta}, &\mathbf{z}^*, \boldsymbol{g}_1, \boldsymbol{g}_2|\mathbf{y}) \\ &\propto \left[\prod_{t \in \mathcal{T}_1} f(y_t|y_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_1, \boldsymbol{g}_2)\right] \left[\prod_{t \in \mathcal{T}_2} f(y_t^*|y_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_1, \boldsymbol{g}_2)\right] \mathcal{N}(A, B|\boldsymbol{b_0}, \boldsymbol{B_0}) \\ &\quad \times \mathcal{W}(\Sigma|\boldsymbol{\nu}, \boldsymbol{Q}) \left[\prod_{i,j} \mathcal{GP}(\boldsymbol{g}_{i,j}|\sigma_f, \ell)\right], \end{aligned} \tag{4.10}$$

where , $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ are function evaluations of $G_1(\cdot)$ and $G_2(\cdot)$ for given input pints $\mathbf{z}^*$ and $\mathbf{x}^*$. The posterior distribution can be simulated by MCMC methods. We propose a hybrid Gibbs-based algorithm for posterior simulation, which is summarized in Algorithm 10. We now turn to the details of the proposed MCMC algorithm.

Algorithm 10 A Blocking Algorithm with Metropolis-Hastings update

```
Inputs: $\mathbf{y}$: data observations; $G$: number of iterations; $\boldsymbol{\theta}^{(0)} = \big(A^{(0)}, B^{(0)}, \Sigma^{(0)}\big)$: initial value; $\mathcal{N}(\mathbf{b}_1, \mathbf{B}_1)$, $\mathcal{IW}(\nu + T + 1, \boldsymbol{R})$, $\mathcal{N}(\mathfrak{m}_1, \mathfrak{s}_1)$: full conditional posterior distributions;
for $i = 1 \to G$ do
    Generate regression parameters
        $$A, B|\mathbf{y}, \mathbf{z}^*_{t\in\mathcal{T}_2}, \boldsymbol{\theta}_{/A,B}, \boldsymbol{g}_1, \boldsymbol{g}_2 \sim \mathcal{N}(\mathbf{b}_1, \mathbf{B}_1);$$
    Generate error covariance matrix
        $$\Sigma|\mathbf{y}, \mathbf{z}^*_{t\in\mathcal{T}_2}, \boldsymbol{\theta}_{/\Sigma}, \boldsymbol{g}_1, \boldsymbol{g}_2 \sim \mathcal{IW}(\nu + T + 1, \boldsymbol{R});$$
    Generate nonparametric function evaluations
        $$\boldsymbol{g}_1|\mathbf{y}, \mathbf{z}^*_{t\in\mathcal{T}_2}, \boldsymbol{\theta}, \boldsymbol{g}_2;$$
    Generate $\mathbf{z}^*_{t\in\mathcal{T}_2}$ and $\boldsymbol{g}_1$ in a block from $\mathbf{z}^*_{t\in\mathcal{T}_2}, \boldsymbol{g}_1|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{g}_2$ by first sample $\mathbf{z}^*_{t\in\mathcal{T}_2}$ from $\mathbf{z}^*_{t\in\mathcal{T}_1}|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{g}_2$ with a Metropolis-Hastings update based on an auxiliary mixture model, then sample $\boldsymbol{g}_1$ from $\boldsymbol{g}_1|\mathbf{y}, \mathbf{z}^*_{t\in\mathcal{T}_2}, \boldsymbol{\theta}, \boldsymbol{g}_2$;
Outputs: A sample of $G$ draws of $\boldsymbol{\theta}$ and $\boldsymbol{g_1}$ and $\boldsymbol{g_2}$ from $p(\boldsymbol{\theta}, \boldsymbol{g_1}, \boldsymbol{g_2}|\mathbf{y})$.
```

4.2.2 Sampling A, B

The posterior distribution for $A, B|\mathbf{y}, \mathbf{z}^*_{t\in\mathcal{T}_1}, \boldsymbol{\theta}_{/A,B}, \boldsymbol{g}_1, \boldsymbol{g}_2 \sim \mathcal{N}(\mathbf{b}_1, \mathbf{B}_1)$, where

$$\boldsymbol{b}_1 = \boldsymbol{B}\left(\boldsymbol{B}_{\mathbf{0}}^{-1}\boldsymbol{b}_0 + \sum_{t=1}^{T}\text{diag}(y^*_{t-1})'\Sigma^{-1}\left(y^*_t - C \circ G\left(\iota_2 \otimes y^{*\prime}_{t-1}\right)\iota_2\right)\right)$$

$$\boldsymbol{B}_1 = \left(\boldsymbol{B}_0^{-1} + \sum_{t=1}^{T}\text{diag}(y^*_{t-1})'\Sigma^{-1}\text{diag}(y^*_{t-1})\right)^{-1},$$

and

$$\text{diag}(y^*_{t-1}) = \begin{bmatrix} x_t & 0 \\ 0 & z^*_t \end{bmatrix},$$

where C and G are defined in (4.4). This step proceeds with the imputed 'missing' data of low frequency variable and computes the conditional mean and covariance of A and B at a joint high frequency.

4.2.3 Sampling Σ

The posterior distribution for error covariance matrix, $\Sigma|\mathbf{y}, \mathbf{z}^*_{t\in\mathcal{T}_0}, \boldsymbol{\theta}_{/\Sigma}, \boldsymbol{g}_1, \boldsymbol{g}_2 \sim \mathcal{IW}$, is given by

$$\pi\left(\Sigma|\mathbf{y}, \mathbf{z}^*_{t\in\mathcal{T}_0}, \boldsymbol{\theta}_{/\Sigma}, \boldsymbol{g}_1, \boldsymbol{g}_2\right) \propto |\Sigma|^{-(\boldsymbol{\nu}+T+1)/2}\exp\left[\frac{1}{2}\text{tr}(\Sigma^{-1}\boldsymbol{R})\right],$$

where $\boldsymbol{R} = \boldsymbol{Q} + \sum_{t=2}^{T}\varepsilon'_t\varepsilon_t$, and ε_t is defined in (4.2). The sampling can proceed by exploiting the conjugate nature of prior distribution to derive a closed-form solution for the full conditional posterior distribution of Σ, conditional on the latent data.

4.2.4 Sampling the Nonparametric Functions

The nonparametric functions are simulated one at a time, conditional on all remaining functions, parameters and latent data. Based on the discussion of the Gaussian process in the previous section, the posterior distribution for function evaluations is obtained through a Gaussian prediction. In general, for the i-th equation in the VAR, we first isolate the j-th function by defining

$$\xi_{i,j,t} \equiv y^*_{i,t} - \alpha_i - \sum_{l=1}^{J}(1-c_{i,l})\beta_{i,l}y^*_{i,t-1} - \sum_{l=1,l\neq j}^{J} c_{i,l}G_{i,l}(y^*_{l,t-1}) - \mathbb{E}(\varepsilon_{i,t}|\varepsilon_{\setminus i}) \quad (4.11)$$

and the posterior distribution of function evaluations is $\boldsymbol{g}_{i,j}|\mathbf{y}, \mathbf{z}^*_{t\in\mathcal{T}_1}, \boldsymbol{\theta}_{/\boldsymbol{g}_{i,j}} \sim \mathcal{N}(\mathfrak{m}_1, \mathfrak{s}_1)$, where

$$\begin{aligned}\mathfrak{m}_1 &= m(\mathbf{y}^*_i) + K(\mathbf{y}^*_i, \mathbf{y}^*_i)\left[K(\mathbf{y}^*_i, \mathbf{y}^*_i) + \mathbb{E}(\sigma^2_i|\varepsilon_{\setminus i})I_{T-1}\right]^{-1}(\xi_{i,j} - m(\mathbf{y}^*_i))\\ \mathfrak{s}_1 &= K(\mathbf{y}^*_i, \mathbf{y}^*_i) - K(\mathbf{y}^*_i, \mathbf{y}^*_i)\left[K(\mathbf{y}^*_i, \mathbf{y}^*_i) + \mathbb{E}(\sigma^2_i|\varepsilon_{\setminus i})I_{T-1}\right]^{-1}K(\mathbf{y}^*_i, \mathbf{y}^*_i).\end{aligned}$$

The idea of data-argumentation has led to tractable posterior distribution for each of A, B, Σ and $\{\boldsymbol{g}_i\}$ and now we turn to the sampling scheme for the latent data $\mathbf{z}_{t\in\mathcal{T}_1}$.

4.2.5 Sampling the Latent Data

Conditional on $\boldsymbol{\theta}$, in particular $\{\boldsymbol{g_i}\}$, the efficient block sampling algorithm known as *forward filtering backward sampling* (FFBS) can be used to impute $z^*|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{g}_1, \boldsymbol{g}_2$. Specifically, the FFBS algorithm generates a vector of $\mathbf{z}^*$ from the smoothed distribution

$$p(\mathbf{z}^*|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{g}_1, \boldsymbol{g}_2) \propto p(y_1)\prod_{t=2}^{T} p(y_t|y^*_t, \mathbb{1}_t)p\left(y^*_t|y^*_{t-1}, \boldsymbol{\theta}, \boldsymbol{g}_1, \boldsymbol{g}_2\right). \quad (4.12)$$

A drawback of directly sampling $\mathbf{z}^*|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{g}_1, \boldsymbol{g}_2$ from its full conditional posterior is that the produced Markov chain could be persistent and subject to slow convergence to the invariant distribution. We introduce a efficient blocking scheme which simulates the latent data $\mathbf{z}^*$ and $\boldsymbol{g}_1$ in a block in the following section.

4.2.6 An Efficient Blocking Algorithm for Sampling the Latent Data

The main idea of the proposed approach is to approximate the equation containing $G_1(z_t^*)$ by an auxiliary mixture model. The mixture model components, denoted by $p^a(y_t^*|y_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2, u_{t-1})$, are each assumed to be linear and Gaussian given the state-dependent mixture indicator, u_t. Since only $G_1(z_{t-1}^*)$ involves latent inputs, the construction of auxiliary mixture model is focused on approximating $G_1(\cdot)$ conditional on data, other static parameters and, in particular, $\boldsymbol{g}_2$ throughout the remainder.

An Auxiliary Mixture Model for GP

Following Stroud, Müller, and Polson (2003), hereafter denoted as SMP, let $\mathbf{u} = (u_1, \cdots, u_{T-1})$ be a vector of mixture indicators that takes integer values, i.e., $u_t \in (1, \cdots, K)$. The auxiliary model is defined as

$$p^a(y_t^*|y_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2) = \sum_{k=1}^{K} p^a(y_t^*|y_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2, u_{t-1} = k)\pi_k(z_{t-1}^*), \tag{4.13}$$

where $\pi_k(z_{t-1}^*) = p^a(u_{t-1} = k|z_{t-1}^*)$ are state-dependent weights. Conditional the mixture indicator, the SMP method defines an auxiliary mixture model that is linear and has state-dependent variance. In the VAR context, it can be expressed as

$$p^a(y_t^*|y_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2, u_{t-1} = k) = \mathcal{N}(A_k + B_k y_{t-1}^*, \Sigma), \tag{4.14}$$

where

$$A_k = A + \begin{bmatrix} a_k \\ g_{2,t-1} \end{bmatrix} \text{ and } B_k = B + \begin{bmatrix} 0 & b_k \\ 0 & 0 \end{bmatrix}.$$

The mixture weights are chosen to be standardized Gaussian kernels,

$$p^a(u_{t-1}=k|z^*_{t-1}) = \frac{\phi(z^*_{t-1};\bar{\mu}_k,\bar{s}_k)}{\sum_{l=1}^{K}\phi(z^*_{t-1};\bar{\mu}_l,\bar{s}_l)}, \tag{4.15}$$

where $\phi(x;\bar{\mu},\bar{s})$ denotes a normal density with mean $\bar{\mu}$ and variance $\bar{s}^2$. The regression parameters a_k and b_k defined in (4.14) are obtained from the first order Taylor series approximation of the nonlinear function at $\bar{\mu}_k$. That is, conditional on $\boldsymbol{\theta}$, a_k and b_k are given by

$$b_k(\boldsymbol{\theta}) = \frac{\partial G_1(\bar{\mu}_k)}{\partial z^*_{t-1}} \text{ and } a_k(\boldsymbol{\theta}) = G_1(\bar{\mu}_k) - b_k(\boldsymbol{\theta})\bar{\mu}_k. \tag{4.16}$$

We suppress the explicit dependence on $\boldsymbol{\theta}$ hereafter for simplicity of notation. The above first order Taylor series expansion permits a linear approximation of the nonlinear function G_1 locally at $\bar{\mu}_k$. In conjunction with the choice of Gaussian kernels in (4.15), efficient sampling schemes such as FFBS, are readily at disposal to generate a proposal of the latent states in a block which is then used in a Metropolis-Hastingsss update.

Such an approach obviously requires calculation of a_k, the derivative b_k in (4.16) and the state-dependent variance τ_k^2. In the context of Gaussian process models, however, the random function are not directly observable, they are only seen through the function evaluations at a given set of grid points. Although biased estimators of the gradient b_k, such as estimators based on finite-differencing techniques, are possible, the approximation errors due to the bias may accumulate rapidly in a block sampling scheme, resulting in an MCMC chain that moves slowly with highly correlated draws which gives imprecise estimators of integrals over the posterior distribution.

We propose an unbiased gradient estimator of a_k and b_k that marginalizes over the random function evaluation at $\bar{\mu}_k$,

$$\mathbb{E}_{g_{\bar{\mu}_k}|\boldsymbol{\theta},\bar{\mu}_k}\left(a_k + b_k(\bar{\mu}_k - z^*_{t-1})\right) = \mathbb{E}_{g_{\bar{\mu}_k}|\boldsymbol{\theta},\bar{\mu}_k}\left(g_{\bar{\mu}_k} + \frac{\partial G_1(\bar{\mu}_k)}{\partial z^*_{t-1}} z^*_{t-1}\right).$$

Then, write the expectation as an integral yields

$$\begin{aligned}&\mathbb{E}_{g_{\bar{\mu}_k}|\boldsymbol{\theta},\bar{\mu}_k}\left(g_{\bar{\mu}_k} + \frac{\partial G_1(\bar{\mu}_k)}{\partial z^*_{t-1}} z^*_{t-1}\right)\\ &= \int g_{\bar{\mu}_k} p(g_{\bar{\mu}_k}|\boldsymbol{\theta},\bar{\mu}_k) dg_{\bar{\mu}_k} + \int \frac{\partial G_1(\bar{\mu}_k)}{\partial z^*_{t-1}} z^*_{t-1} p(g_{\bar{\mu}_k}|\boldsymbol{\theta},\bar{\mu}_k) dg_{\bar{\mu}_k}\end{aligned}$$

where the first integral $\int g_{\bar{\mu}_k} p(g_{\bar{\mu}_k}|\boldsymbol{\theta},\bar{\mu}_k) dg_{\bar{\mu}_k}$ can be recognized as the posterior mean for $G_1(\bar{\mu}_k)$. Since the posterior distribution of $p(g_{\bar{\mu}_k}|\boldsymbol{\theta},\bar{\mu}_k)$ is known in closed form, we can write the posterior mean which is given by

$$\bar{g}_{\bar{\mu}_k} = m(\bar{\mu}_k) + K(\bar{\mu}_k,\mathbf{z}^*)\left[K(\mathbf{z}^*,\mathbf{z}^*) + \mathbb{E}(\sigma_1^2|\varepsilon_{\setminus 1}) I_{T-1}\right]^{-1}(\xi_{1,2} - m(\mathbf{z}^*)).$$

Furthermore, if $G_1(\cdot)$ is smooth enough to allow the exchange of differentiation and integration operators, the second integral reduces to

$$\begin{aligned}\int \frac{\partial G_1(\bar{\mu}_k)}{\partial z^*_{t-1}} z^*_{t-1} p(g_{\bar{\mu}_k}|\boldsymbol{\theta},\bar{\mu}_k) dg_{\bar{\mu}_k} &= z^*_{t-1}\frac{\partial}{\partial z^*_{t-1}} \int g_{\bar{\mu}_k} p(g_{\bar{\mu}_k}|\boldsymbol{\theta},\bar{\mu}_k) dg_{\bar{\mu}_k}\\ &= z^*_{t-1}\frac{\partial}{\partial z^*_{t-1}}\bar{g}_{\bar{\mu}_k},\end{aligned}$$

that is, z^*_{t-1} multiplied by the first derivative of the posterior mean for $g_{\bar{\mu}_k}$. hence in our auxiliary mixture model, we used regression parameters $\tilde{a}_k$ and $\tilde{b}_k$ which are given by

$$\tilde{b}_k(\boldsymbol{\theta}) = \frac{\partial}{\partial z^*_{t-1}}\bar{g}_{\bar{\mu}_k} \text{ and } \tilde{a}_k(\boldsymbol{\theta}) = \bar{g}_{\bar{\mu}_k} - \tilde{b}_k(\boldsymbol{\theta})\bar{\mu}_k.$$

For GP models with squared exponential kernel, the first derivative of the kernel between $\bar{\boldsymbol{\mu}} = (\bar{\mu}_1, \cdots, \bar{\mu}_K)'$ and $z_t^* \in \mathbf{z}^* = (z_1^*, \cdots, z_T^*)'$ is given by

$$\begin{aligned}
\frac{\partial k(\bar{\boldsymbol{\mu}}, z_t^*)}{\partial \bar{\boldsymbol{\mu}}} &= \frac{\partial}{\partial \bar{\boldsymbol{\mu}}} \left\{ \sigma_f \exp\left(-\frac{1}{2\ell^2} (\bar{\boldsymbol{\mu}} - z_t^*)'(\bar{\boldsymbol{\mu}} - z_t^*) \right) \right\} \\
&= \frac{\partial}{\partial \bar{\boldsymbol{\mu}}} \left\{ \exp\left(-\frac{1}{2\ell^2} (\bar{\boldsymbol{\mu}} - z_t^*)'(\bar{\boldsymbol{\mu}} - z_t^*) \right) \right\} k(\bar{\boldsymbol{\mu}}, z_t^*) \\
&= -\ell^{-2} (\bar{\boldsymbol{\mu}} - z_t^*) k(\bar{\boldsymbol{\mu}}, z_t^*)
\end{aligned}$$

which is a $K \times 1$ vector. The derivative for each of z_t^* need to be concatenated to compute the derivative of the posterior mean. Let $\tilde{\bar{\boldsymbol{\mu}}} = (\bar{\boldsymbol{\mu}} - z_1^*, \cdots, \bar{\boldsymbol{\mu}} - z_T^*)$, we have that

$$\begin{aligned}
\frac{\partial \bar{\boldsymbol{g}}_{\bar{\mu}}}{\partial \bar{\boldsymbol{\mu}}} &= \frac{\partial \boldsymbol{k}(\bar{\boldsymbol{\mu}}, \mathbf{z}^*)}{\partial \bar{\boldsymbol{\mu}}} (K(\mathbf{z}^*, \mathbf{z}^*) + \sigma^2 I_{T-1})^{-1} \xi_{1,2} \\
&= -\ell^{-2} \tilde{\bar{\boldsymbol{\mu}}}' \left(\boldsymbol{k}(\bar{\boldsymbol{\mu}}, \mathbf{z}^*)' \circ (K(\mathbf{z}^*, \mathbf{z}^*) + \sigma^2 I_{T-1})^{-1} \xi_{1,2} \right),
\end{aligned}$$

a $K \times 1$ vector, where $\xi_{1,2}$ is defined in (4.11).

Simulating From the Smoothing Distribution

Having constructed the auxiliary mixture model, the latent data $\mathbf{z}^*$ is then simulated using a three steps procedure, with each iteration comprises steps summarized as followings: (1) generate the vector of latent mixture indicators $\mathbf{u}$ based on the currently imputed latent variable $\mathbf{z}^*$, (2) propose new values for the latent data $\mathbf{z}^*$ from the constructed conditional linear Gaussian model by exploiting a tailored version of FFBS and (3) accept the candidate draw $\tilde{\mathbf{z}}^*$ with an appropriate Metropolis-Hastingsss acceptance probability. In what follows, we outline each of the three steps in detail.

1. Generating mixture indicators $\mathbf{u} = (u_1, \cdots, u_{T-1})'$. Given $\mathbf{z}^*$, $\mathbf{y}$, $\boldsymbol{g}_2$ and $\boldsymbol{\theta}$, the full conditional posterior distribution of the mixture indicators is given by

$$p^a(\mathbf{u}|\mathbf{z}^*, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{g}_2) \propto \prod_{t=2}^{T} p^a\left(y_t^*|y_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2, u_{t-1}\right) p^a(u_{t-1}|z_{t-1}^*).$$

The indicator variables $u_1, \cdots, u_{T-1}$ are conditionally independent and hence can be sampled independently from multinomial distributions with probabilities of the form $p^a(u_{t-1}|x_{t-1}, z_{t-1}^*, y_t^*, \boldsymbol{\theta}, \boldsymbol{g}_2) \propto p^a\left(y_t^*|y_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2, u_{t-1}\right) p^a(u_{t-1}|z_{t-1}^*)$.

2. Generating a proposal draw of $\tilde{\mathbf{z}}^*$ from the auxiliary mixture model, where the full conditional posterior distribution $p^a(\mathbf{z}^*|\mathbf{y}, \mathbf{u}, \boldsymbol{\theta}, \boldsymbol{g}_2)$ can be written as

$$p^a(\mathbf{z}^*|\mathbf{y}, \mathbf{u}, \boldsymbol{\theta}, \boldsymbol{g}_2) \propto p(y_1) \prod_{t=2}^{T} p(y_t|y_t^*, \mathbb{1}_t) p^a\left(y_t^*|y_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2, u_{t-1}\right) p^a(u_{t-1}|z_{t-1}^*).$$

Following SMP, this conditional posterior distribution is decomposed into two components. The first term includes all necessary components that are linear in the latent states with Gaussian disturbances, which allows us to devise an efficient proposal distribution. The second term is used in the acceptance probability in a Metropolis-Hastings step. The conditional distribution $p^a(\mathbf{z}^*|\mathbf{y}, \mathbf{u}, \boldsymbol{\theta}, \boldsymbol{g}_2)$ can be written as

$$\begin{aligned} &p^a(\mathbf{z}^*|\mathbf{y}, \mathbf{u}, \boldsymbol{\theta}, \boldsymbol{g}_2) \\ &\quad \propto \underbrace{p(y_1) \prod_{t=2}^{T} p(y_t|y_t^*, \mathbb{1}_t) p^a\left(y_t^*|y_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2, u_{t-1}\right) \phi(z_{t-1}^*; \bar{\mu}_{u_{t-1}}, \bar{s}_{u_{t-1}})}_{q(\mathbf{z}^*|\mathbf{y}, \mathbf{u}, \boldsymbol{\theta}, \boldsymbol{g}_2)} \times \frac{1}{c(z_{t-1}^*)} \end{aligned}$$

where $c(z_{t-1}^*) = \sum_{l=1}^{K} \phi(z_{t-1}^*; \bar{\mu}_l, \bar{s}_l)$ is the denominator in (4.15). The first term $q(\mathbf{z}^*|\mathbf{y}, \mathbf{u}, \boldsymbol{\theta}, \boldsymbol{g}_2)$ is proportional to the smoothing joint distribution of a linear, Gaussian state space model.

By combining the linearized state equation (4.14) with the Gaussian weighting kernel (4.15), the following state space model has a smoothed joint distribution given by $q(\mathbf{z}^*|\mathbf{y}, \mathbf{u}, \boldsymbol{\theta}, \boldsymbol{g}_2)$:

$$\tilde{y}_t = D_t y_t^* + \tilde{D}_t v_t \quad v_t \sim \mathcal{N}(\mathbf{0}, H)$$

$$y_t^* = A_{u_{t-1}} + B_{u_{t-1}} y_{t-1}^* + \varepsilon_t \quad \varepsilon_t \sim \mathcal{N}(\mathbf{0}, \Sigma),$$

where

$$\tilde{y}_t^* = \begin{bmatrix} x_t \\ z_t \\ \bar{\mu}_{u_t} \end{bmatrix}, \; y_t^* = \begin{bmatrix} x_t \\ z_t^* \end{bmatrix}, \; D_t = \begin{bmatrix} 1 & 0 \\ 0 & \mathbb{1}_t \\ 0 & 1 \end{bmatrix}, \; \tilde{D}_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 - \mathbb{1}_t \\ 0 & 1 \end{bmatrix}, \; H = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \sigma_\nu^2 & 0 \\ 0 & 0 & \bar{s}_{u_t}^2 \end{bmatrix}$$

$$A_{u_{t-1}} = \begin{bmatrix} \alpha_1 + a_{u_{t-1}} \\ \alpha_2 + G_2(x_{t-1}) \end{bmatrix}, \; B_{u_{t-1}} = \begin{bmatrix} \beta_{11} & b_{u_{t-1}} \\ 0 & \beta_{22} \end{bmatrix}, \text{ and } \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}.$$

Algorithm 11 Forward Filtering

```
Inputs: u: mixture indicator; $D_t$: missing observation indicator; $\boldsymbol{\theta}$: static parameters; $p^a(y_t^*|y_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2, u_{t-1})$: auxiliary model; $T-1$: number of iterations; $\tilde{y}_1^*$: initial state;
for $t = 1 \to T-1$ do
    Predict:
    State mean: $\mathbb{E}_{t|t-1}(y_t^*|\tilde{y}_{t-1}) = A_{u_{t-1}} + B_{u_{t-1}}\mathbb{E}_{t-1|t-1}(y_{t-1}^*|\tilde{y}_{t-1})$
    State variance: $\mathbb{V}_{t|t-1}(y_t^*|\tilde{y}_{t-1}) = B_{u_{t-1}}\mathbb{V}_{t-1|t-1}(y_{t-1}^*|\tilde{y}_{t-1})B'_{u_{t-1}} + \Sigma$;
    Update:
    $v_t = \tilde{y}_t - D_t\mathbb{E}_{t|t-1}(y_t^*|\tilde{y}_{t-1})$
    $F_t = D_t\mathbb{V}_{t|t-1}(y_t^*|\tilde{y}_{t-1})D_t' + D_t H$
    $M_t = \mathbb{V}_{t|t-1}(y_t^*|\tilde{y}_{t-1})D_t'$;
    Revise:
    Filtered mean: $\mathbb{E}_{t|t}(y_t^*|\tilde{y}_t) = \mathbb{E}_{t|t-1}(y_t^*|\tilde{y}_{t-1}) + M_t F_t^{-1} v_t$
    Filtered variance: $\mathbb{V}_{t|t}(y_t^*|\tilde{y}_t) = \mathbb{V}_{t|t-1}(y_t^*|\tilde{y}_{t-1}) - M_t F^{-1} M_t'$;
Outputs: Filtering distribution $\mathcal{N}\left(\mathbb{E}_{t|t}(y_t^*|\tilde{y}_t), \mathbb{V}_{t|t}(y_t^*|\tilde{y}_t)\right)$ for $t = 1, 2, \cdots, T$.
```

Finally, we can provide a brief exposition of the tailored FFBS algorithm, which efficiently generates candidate values of $\tilde{\mathbf{z}}^*$ from the joint posterior distribution of of the latent state. The Kalman filtering algorithm is summarized in Algorithm 11. Note that it yields the filtered distribution, at each time t, for each of components of the state vectors $\mathbf{y}^* = \left(\mathbf{x}', \mathbf{z}^{*'}\right)$, however, the distribution is degenerate at observed values whenever the observation is available. Furthermore, the importance of augmenting our model with v_t now becomes clear. It ensures that the innovation covariance matrix F_t used in the Kalman gain is always invertible. Algorithm 12 then simulates $\mathbf{y}^*$ from the smoothing distribution in a block, bringing efficiency to the sampling scheme. The above algorithm generates the observed values as its realizations for y_t when it is observed and imputes the unobserved ones from the smoothing distribution under the auxiliary linear, Gaussian model. The imputed values for $\mathbf{z}^*$, called $\tilde{\mathbf{z}}^*$, is then used as a proposal in the following step.

Algorithm 12 Backward Sampling

```
1: Inputs: $\mathbf{u}$: mixture indicator; $D_t$: missing observation indicator; $\boldsymbol{\theta}$: static parameters; $p^a(y_t^*|y_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2, u_{t-1})$: auxiliary model; $T-1$: number of iterations; $\mathcal{N}\left(\mathbb{E}_{t|t}(y_t^*|y_t), \mathbb{V}_{t|t}(\tilde{y}_t^*|\tilde{y}_t)\right)$: filtering distribution;
2: for $t = 1 \to T-1$ do
3:     Calculate:
       $v_t^* = y_{t+1}^* - A_{u_t} - B_{u_t}\mathbb{E}_{t|t}(y_t^*|\tilde{y}_t)$
       $F_t^* = B_{u_t}\mathbb{V}_{t|t}(y_t^*|\tilde{y}_t)B_{u_t}' + \Sigma$
       $M_t^* = \mathbb{V}_{t|t}(y_t^*|\tilde{y}_t)B_{u_t}'$;
4:     Update:
       $\mathbb{E}_{t|T}(y_t^*|\tilde{y}_T) = \mathbb{E}_{t|t}(y_t^*|\tilde{y}_t) + M_t^* F_t^{*-1} v_t^*$
       $\mathbb{V}_{t|T}(y_t^*|\tilde{y}_T) = \mathbb{V}_{t|t}(y_t^*|\tilde{y}_t) + M_t^* F_t^{*-1} M_t^{*'}$;
5:     Generate:
       $y_t^* \sim \mathcal{N}\left(\mathbb{E}_{t|T}(y_t^*|\tilde{y}_T), \mathbb{V}_{t|T}(y_t^*|\tilde{y}_T)\right)$;
6: Outputs: A draw from the joint posterior distribution of $\mathbf{y}^* \sim q(\mathbf{y}^*|\mathbf{y}, \mathbf{u}, \boldsymbol{\theta}, \boldsymbol{g}_2)$.
```

Metropolis-Hastings Rejection Step

The currently imputed proposal $\tilde{\mathbf{z}}^*$ is retained and replaces $\mathbf{z}^*$ with MH acceptance probability

$$a(\mathbf{z}^*, \tilde{\mathbf{z}}^*) = 1 \vee \left(\prod_{t=2}^{T} \frac{p(\tilde{y}_t^*|x_{t-1}, \tilde{z}_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2)}{c(\tilde{z}_{t-1}^*)p^a(\tilde{y}_t^*|x_{t-1}, \tilde{z}_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2)} \frac{c(z_{t-1}^*)p^a(y_t^*|x_{t-1}, z_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2)}{p(y_t^*|x_{t-1}, z_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2)} \right),$$

otherwise, the proposal $\tilde{\mathbf{z}}^*$ is discarded and $\mathbf{z}^*$ is left unchanged. In particular, note that $p^a(y_t^*|x_{t-1}, z_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2)$ is an approximation to the state transition density $p(y_t^*|x_{t-1}, z_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2)$ implied by the auxiliary mixture model, that is, $c(z_{t-1}^*)p^a(y_t^*|x_{t-1}, z_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2) = \sum_{k=1}^K \phi(z_{t-1}^*; \bar{\mu}_k, \bar{s}_k)p^a(y_t^*|y_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2, u_{t-1} = k)$. The marginal likelihood $p(y_t^*|x_{t-1}, z_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2)$ can be computed by integrating out the function evaluations, $\boldsymbol{g}_1$,

$$p(y_t^*|x_{t-1}, z_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2) = \int p(y_t^*|x_{t-1}, z_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_1, \boldsymbol{g}_2)p(\boldsymbol{g}_1|x_{t-1}, z_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2)d\boldsymbol{g}_1,$$

which is Gaussian by recognizing the joint distribution of the data and function evaluations from the marginal likelihood given by (4.6).

As such, the above Metropolis-Hastings acceptance probability ensures an ergodic Markov chain converge to the invariant distribution $p(\mathbf{z}^*, \mathbf{u}|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{g}_2)$. Note that it can be seen that

$$\begin{aligned}
p(\mathbf{z}^*|\mathbf{y}, \mathbf{u}, \boldsymbol{\theta}, \boldsymbol{g}_2) \propto & p(\mathbf{z}^*|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{g}_2)p^a(\mathbf{u}|\mathbf{z}^*, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{g}_2) \\
& \propto p(y_1)\prod_{t=2}^{T} p(y_t|y_t^*, \mathbb{1}_t)p(y_t^*|y_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2) \\
& \quad \times \frac{p^a(y_t^*|y_{t-1}^*, u_{t-1}, \boldsymbol{\theta}, \boldsymbol{g}_2)\phi(z_{t-1}^*; \bar{\mu}_{u_{t-1}}, \bar{s}_{u_{t-1}})}{\sum_{l=1}^K p^a(y_t^*|y_{t-1}^*, u_{t-1} = l, \boldsymbol{\theta}, \boldsymbol{g}_2)\phi(z_{t-1}^*; \bar{\mu}_{u_l}, \bar{s}_{u_l})},
\end{aligned}$$

and in tandem,

$$q(\mathbf{z}^*|\mathbf{y}, \mathbf{u}, \boldsymbol{\theta}, \boldsymbol{g}_2) = p(y_1)\prod_{t=2}^{T} p(y_t|y_t^*, \mathbb{1}_t)p^a\left(y_t^*|y_{t-1}^*, \boldsymbol{\theta}, \boldsymbol{g}_2, u_{t-1}\right)\phi(z_{t-1}^*; \bar{\mu}_{u_{t-1}}, \bar{s}_{u_{t-1}}),$$

implying the previously stated form of the MH acceptance probability.

4.3 A Monte Carlo Simulation Study

We conduct a simulation study to evaluate the effectiveness of the proposed sampler in estimating static parameters and nonparametric functions. The model considered is the bi-variate VAR(1) in (4.9): each equation contains an intercept, a slope coefficient and a non-parametric function. We graph the non-parametric functions in Figure 4.1. The respective functions are sigmoidal, $G_2(z_{t-1}) = 2\Phi(\mu) - 1$ where $\Phi(\cdot)$ is the standard normal cdf; and quadratic, $G_1(x_{t-1}) = 0.5\mu^2$. The static parameters are set to be $\alpha_1 = -0.2$, $\alpha_2 = -0.2$, $\beta_1 = 0.2$, $\beta_2 = 0.2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$ and $\sigma_{12} = 0.2$, which implies stationarity of both $\mathbf{x}$ and $\mathbf{z}$ as well as non-zero means and median correlation of 0.2 between the errors in the individual equations.

We assume x_t is observed for all t, while z_t is only observed every third period with sample size $T = 200, 400, 800$ respectively. Each function is then evaluated at nine equally spaced points about zero. The performance of sampler is expected to improve, in terms of decreased estimation bias and variances, as the sample size is increased. The prior distributions for the equations are: $(A', B')' \sim \mathcal{N}(\mathbf{0}, 100 \times I)$, $\Sigma \sim \mathcal{IW}(J+2, 2 \times I)$ and the length scale parameter in GP prior is set to be $\ell = 2$. Due to the randomness in obtaining simulated dataset, we analyses the posterior distributions based on a typical realization. Here, the typical sample refers to the one that has the median value of the sample mean of $\mathbf{z}$ among 100 simulations.

For an illustrative purpose, the true and estimated functions for sample size $T = 200, 400, 800$ are plotted in Figure 4.2, 4.6 and 4.10 respectively. As is well known in the literature, non-parametric estimations are subject to worse biased near the boundaries then interiors and bias correction method may be used to improve estimation at boundaries. In the simulation study, the estimated functions are biased towards zero, the prior mean, at all sample sizes due to lack of observations at boundaries. The plots for function estimates show that the method recovers the true functions well, especially when sample size is large.

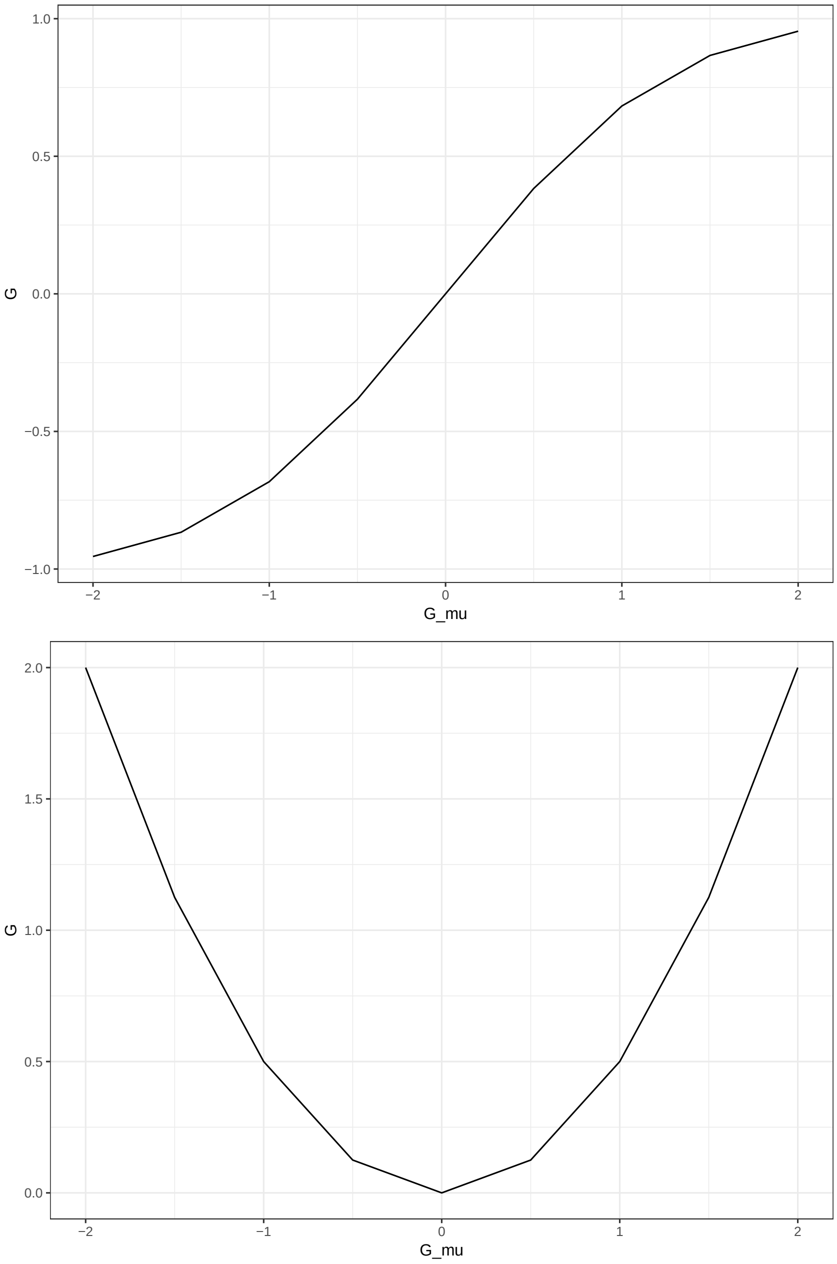

Figure 4.1: Functions used in the simulation study.

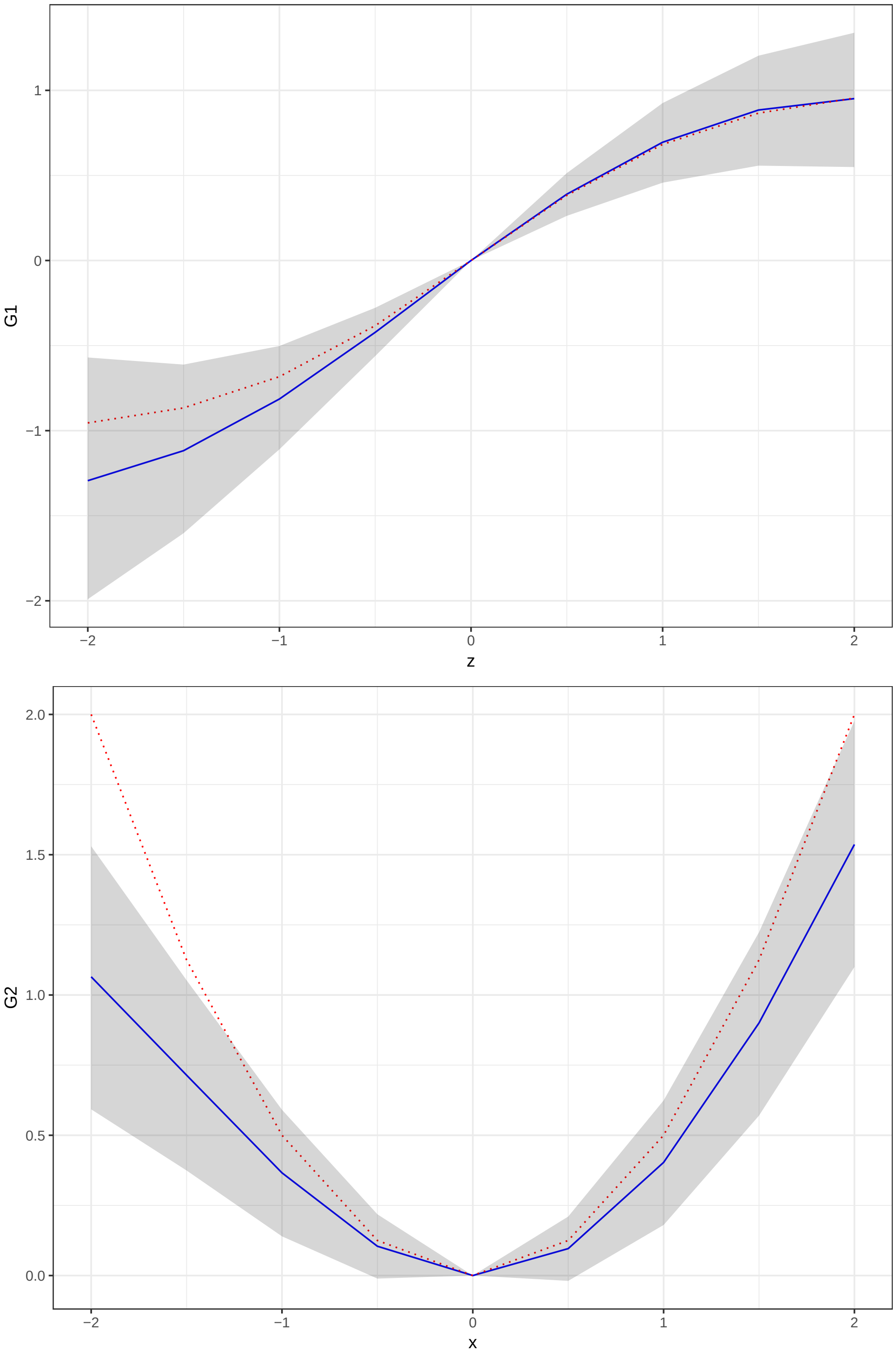

FIGURE 4.2: Posterior mean estimates of G_1 and G_2 ($T = 200$), along with 90% credible intervals. The red dot line represents the function used to generate the data, blue solid line represents the posterior mean and the black shaded area represents the corresponding 90% creditable interval.

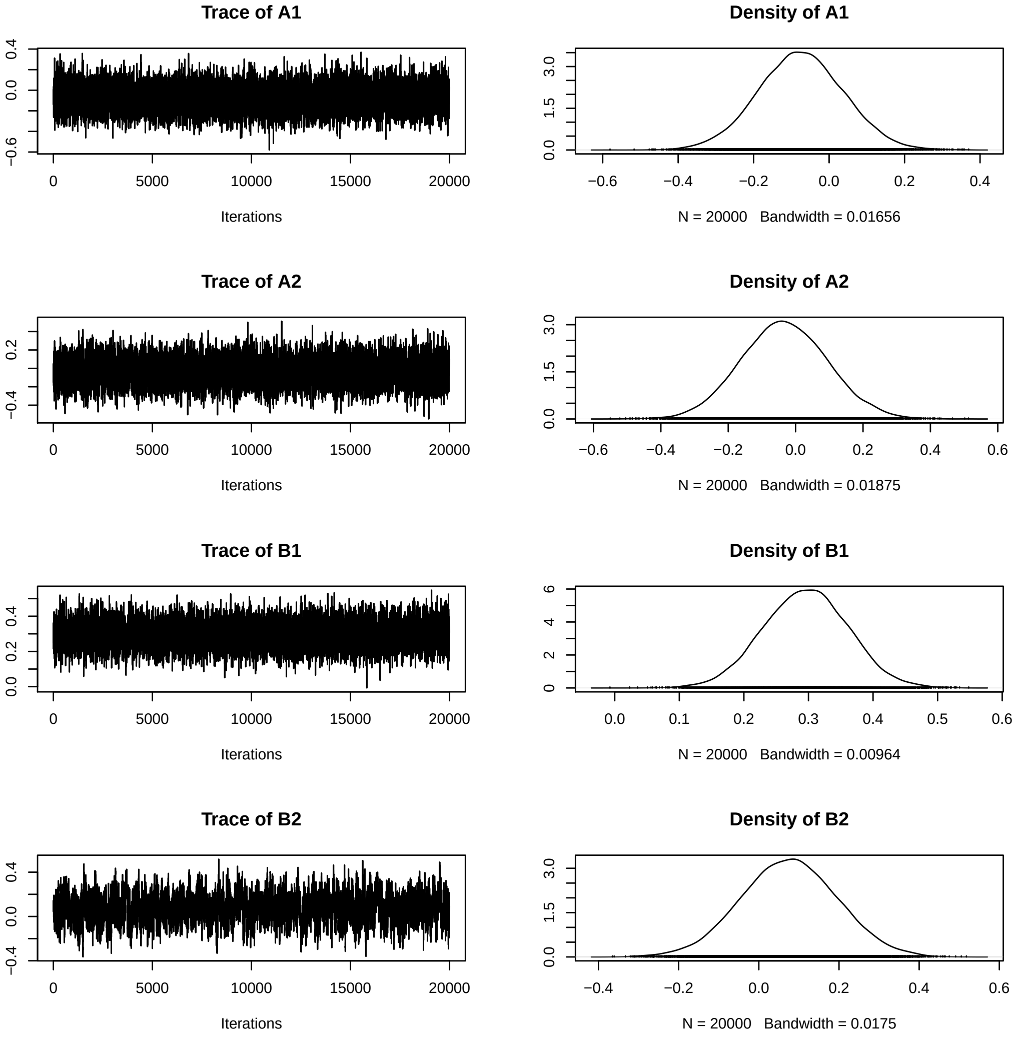

FIGURE 4.3: Trace plot (left penal) and kernel density estimation (right penal) of Bayesian posterior distributions for $\alpha_1, \alpha_2, \beta_1$ and β_2 ($T = 200$) based on 20,000 retained MCMC draws (with 5,000 burn-in).

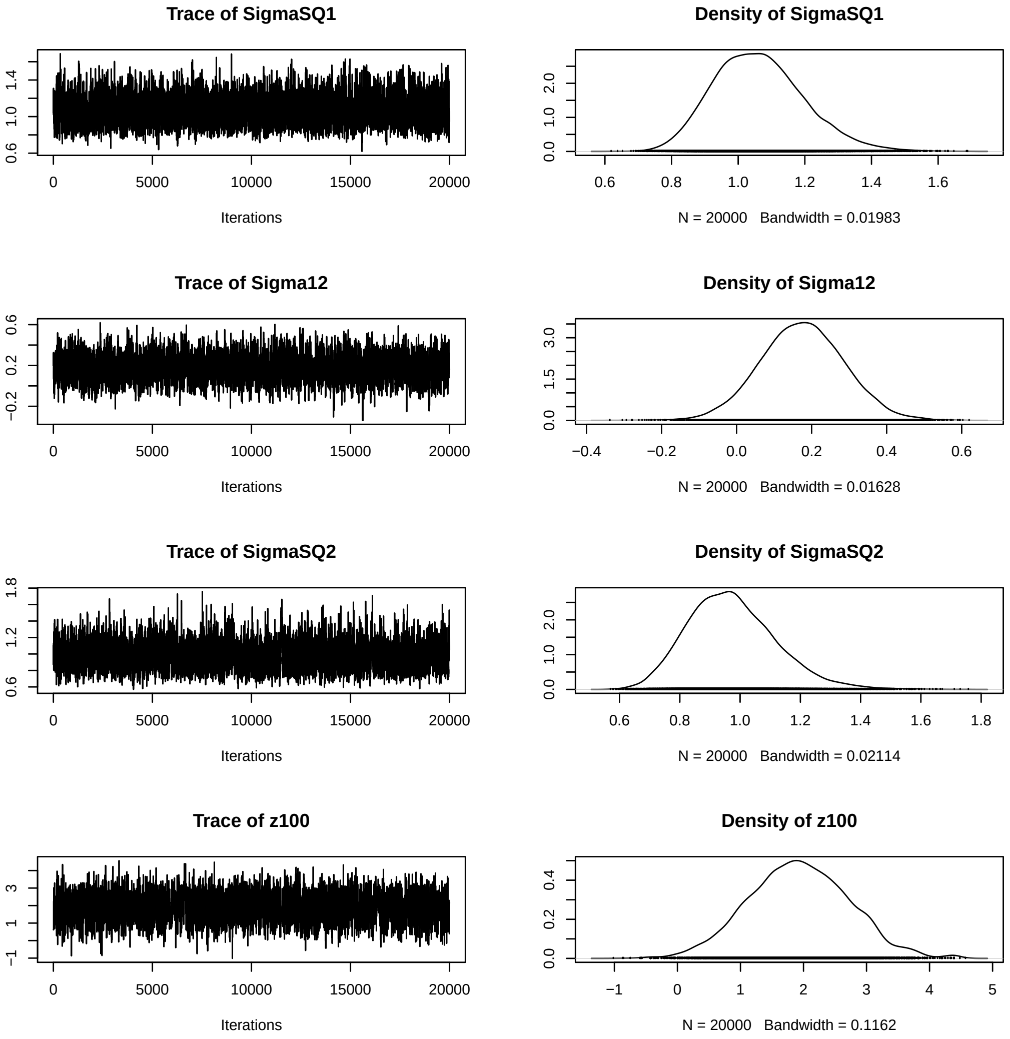

Figure 4.4: Trace plot (left penal) and kernel density estimation (right penal) of Bayesian posterior distributions for $\sigma_1^2, \sigma_{12}, \sigma_2^2$ and z_{100} ($T = 200$) based on 20,000 retained MCMC draws (with 5,000 burn-in).

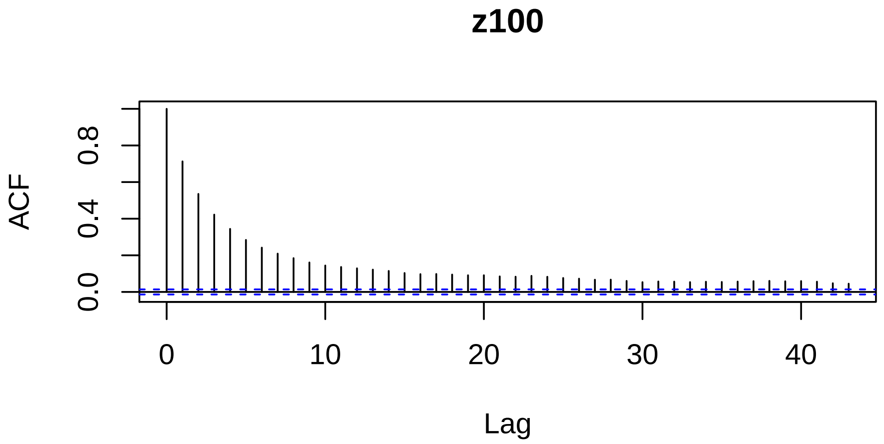

Figure 4.5: Autocorrelation plot for z_{100} ($T = 200$) based on 20,000 retained MCMC draws (with 5,000 burn-in).

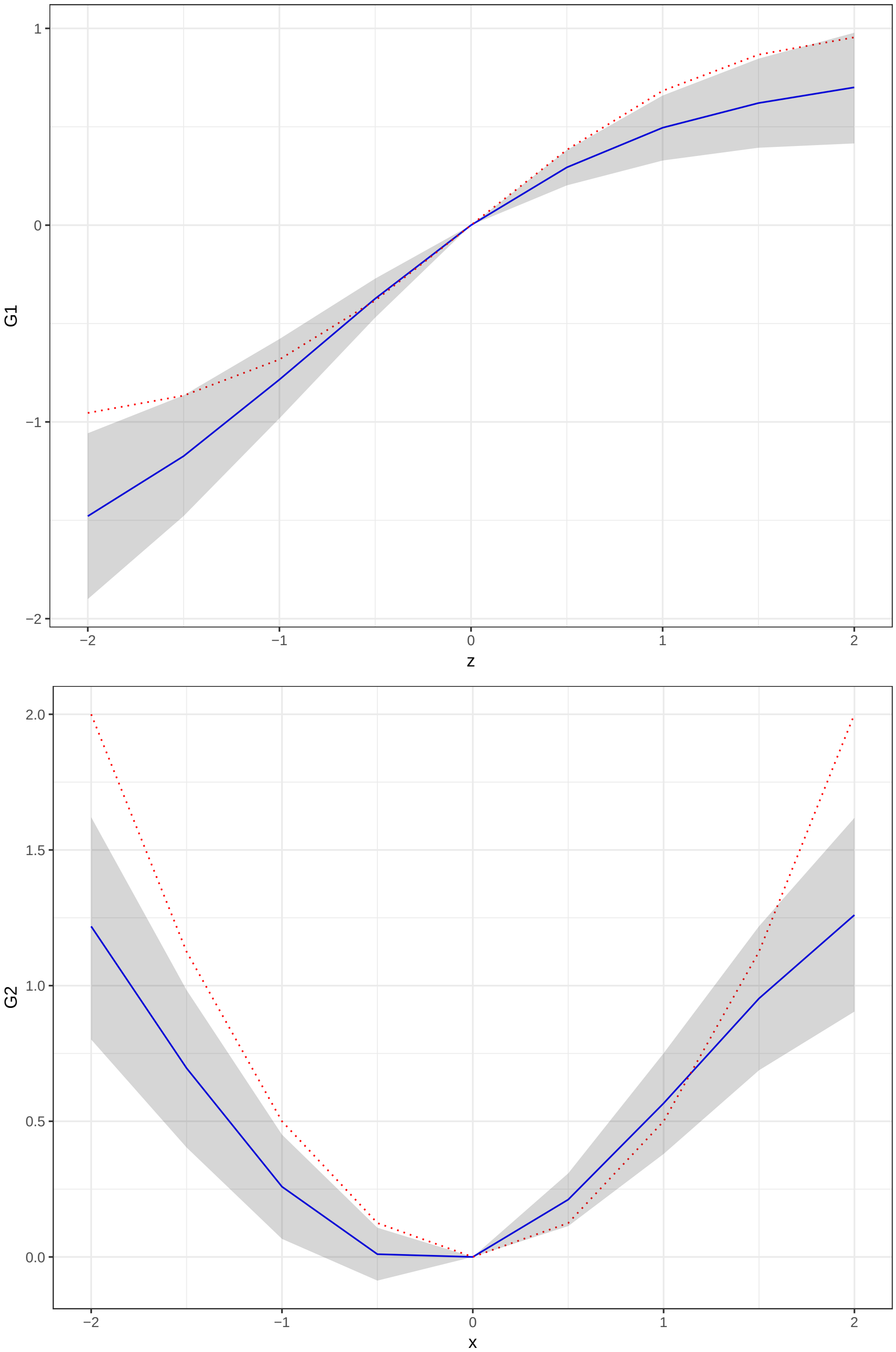

Figure 4.6: Posterior mean estimates of G_1 and G_2 ($T = 400$), along with 90% credible intervals. The red dot line represents the function used to generate the data, blue solid line represents the posterior mean and the black shaded area represents the corresponding 90% creditable interval.

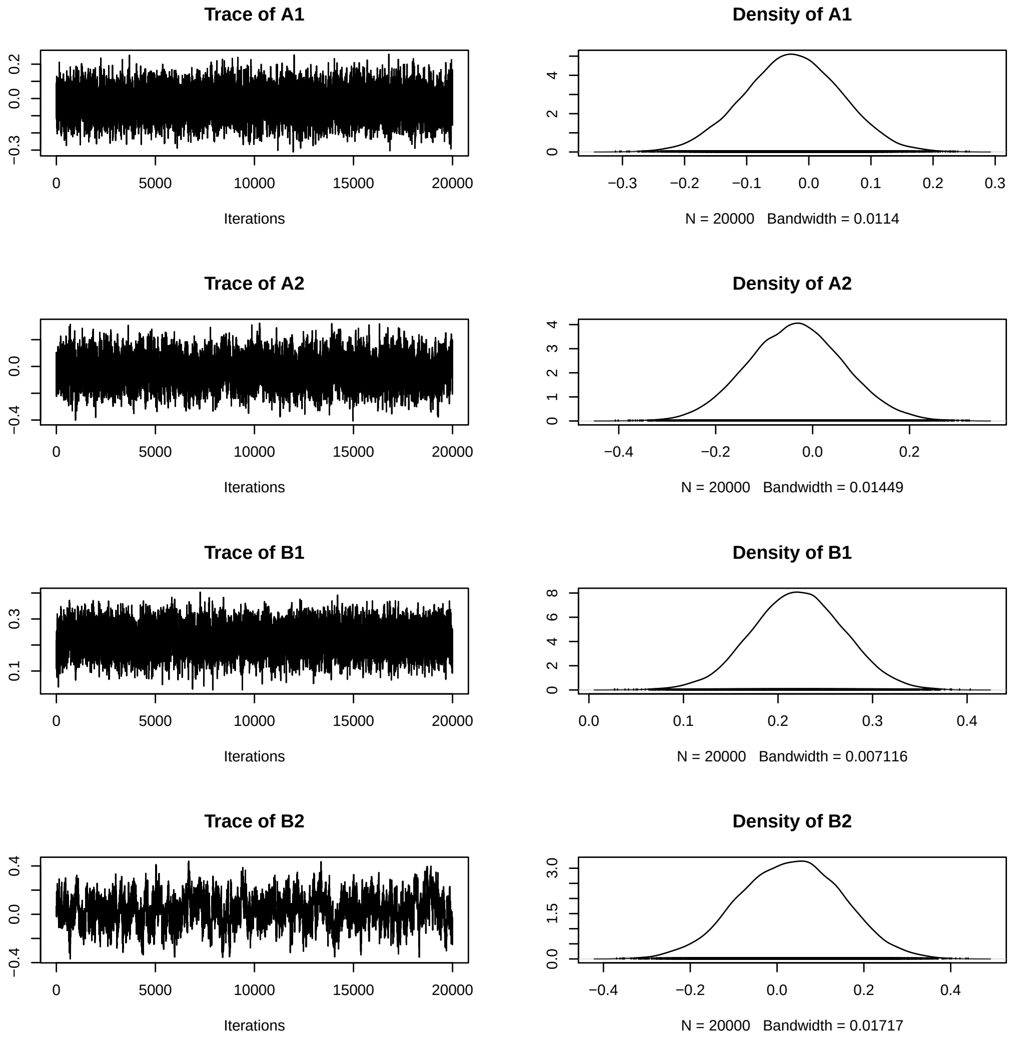

Figure 4.7: Trace plot (left penal) and kernel density estimation (right penal) of Bayesian posterior distributions for $\alpha_1, \alpha_2, \beta_1$ and β_2 ($T = 400$) based on 20,000 retained MCMC draws (with 5,000 burn-in).

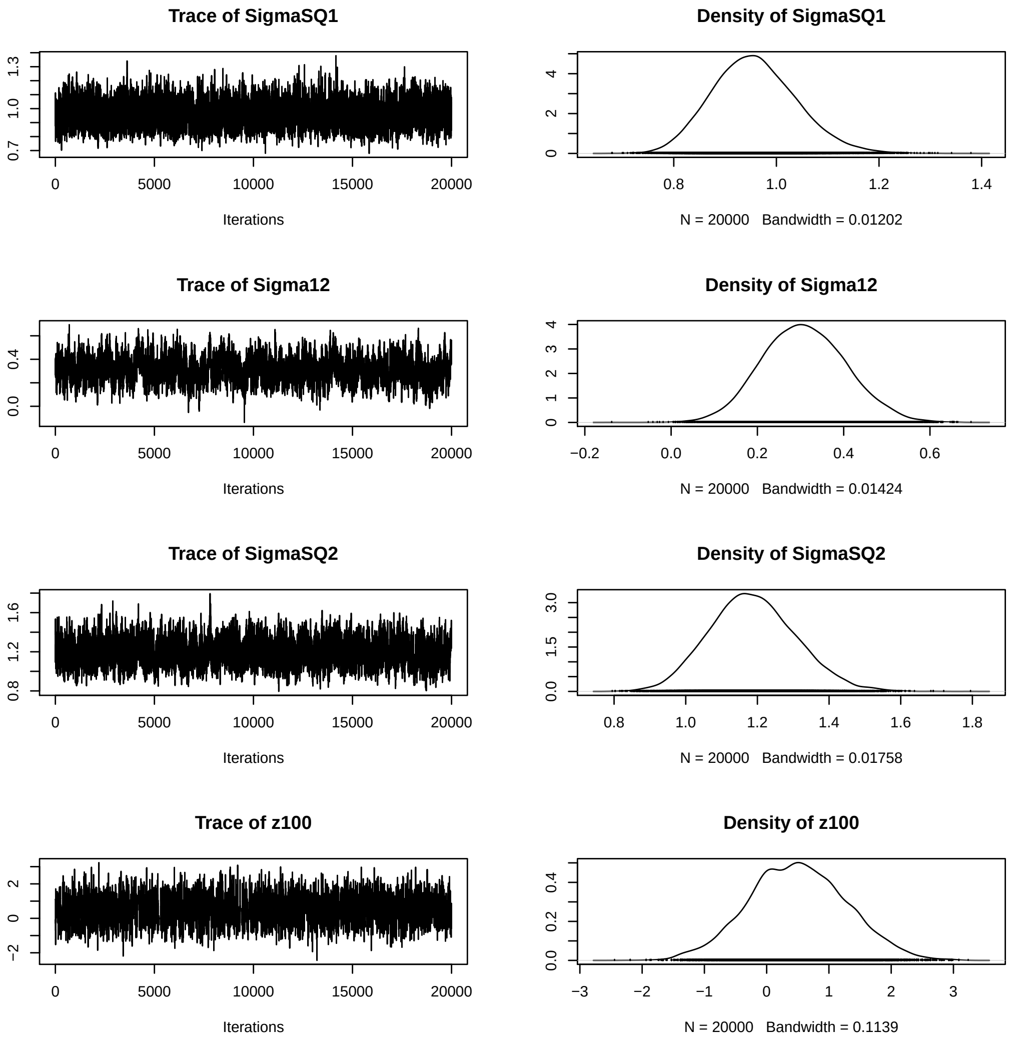

FIGURE 4.8: Trace plot (left penal) and kernel density estimation (right penal) of Bayesian posterior distributions for $\sigma_1^2, \sigma_{12}, \sigma_2^2$ and z_{100} ($T = 400$) based on 20,000 retained MCMC draws (with 5,000 burn-in).

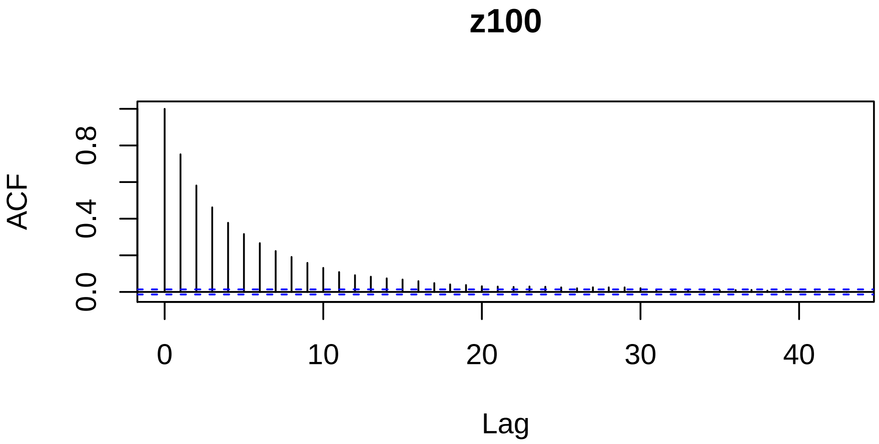

FIGURE 4.9: Autocorrelation plot for z_{100} ($T = 400$) based on 5,000 retained MCMC draws (with 2,000 burn-in).

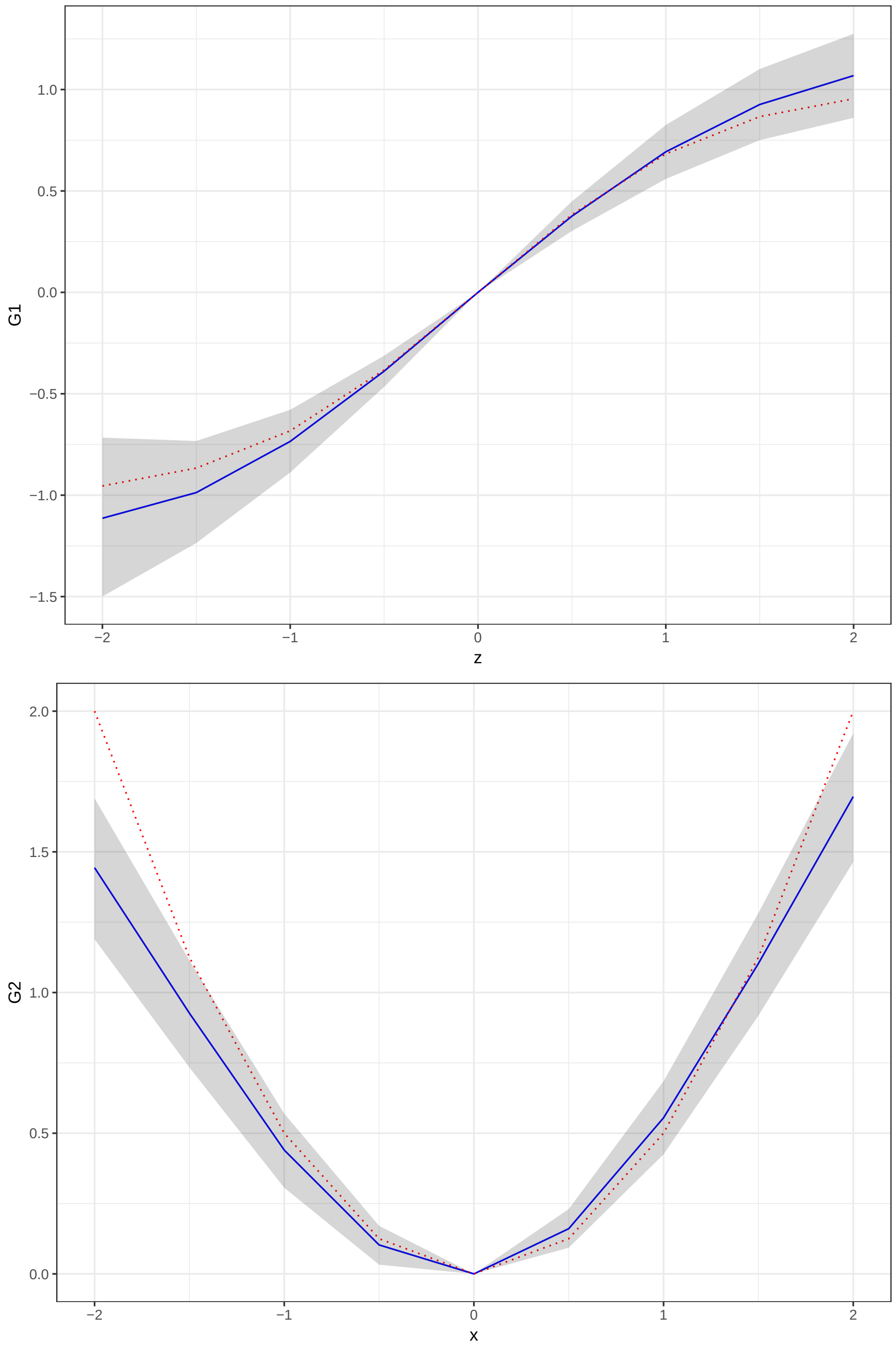

Figure 4.10: Posterior mean estimates of G_1 and G_2 (T = 800), along with 90% credible intervals. The red dot line represents the function used to generate the data, blue solid line represents the posterior mean and the black shaded area represents the corresponding 90% creditable interval.

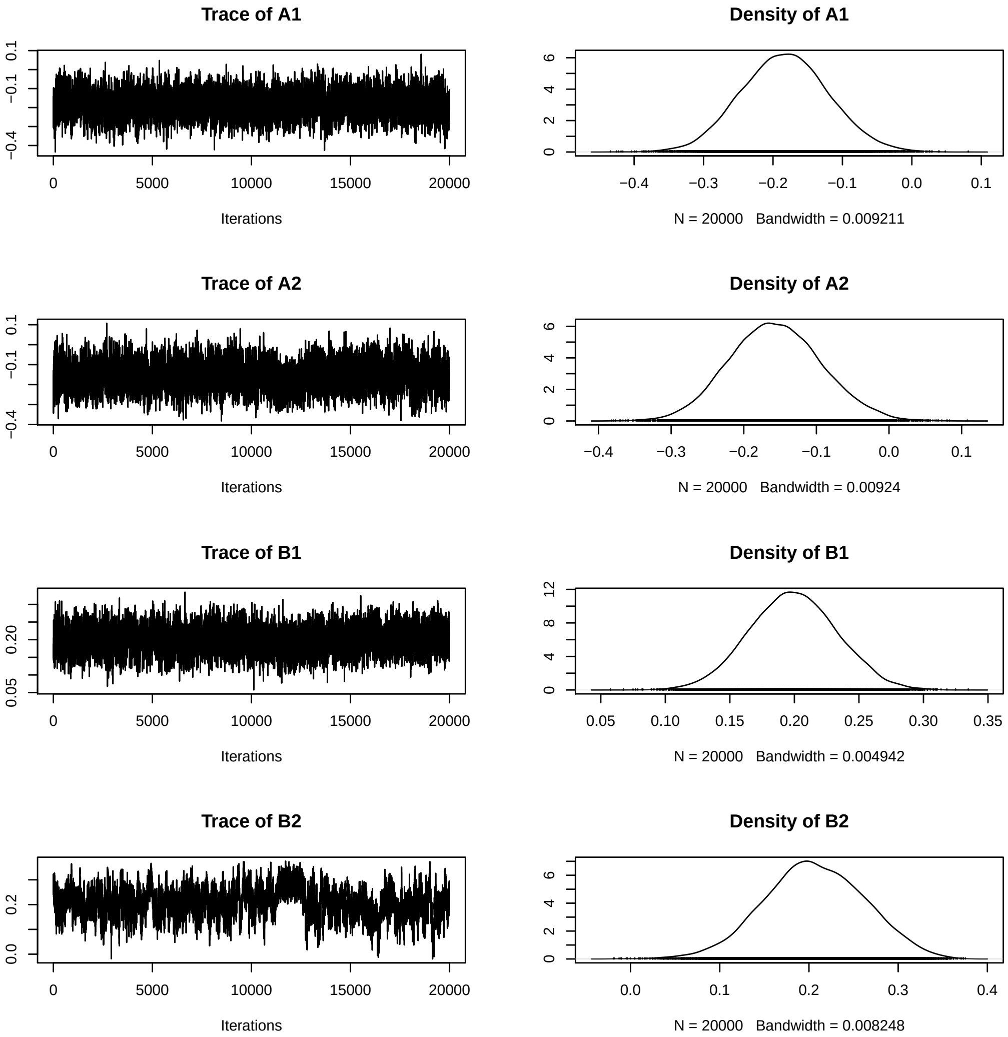

FIGURE 4.11: Trace plot (left penal) and kernel density estimation (right penal) of Bayesian posterior distributions for $\alpha_1, \alpha_2, \beta_1$ and β_2 ($T = 800$) based on 20,000 retained MCMC draws (with 5,000 burn-in).

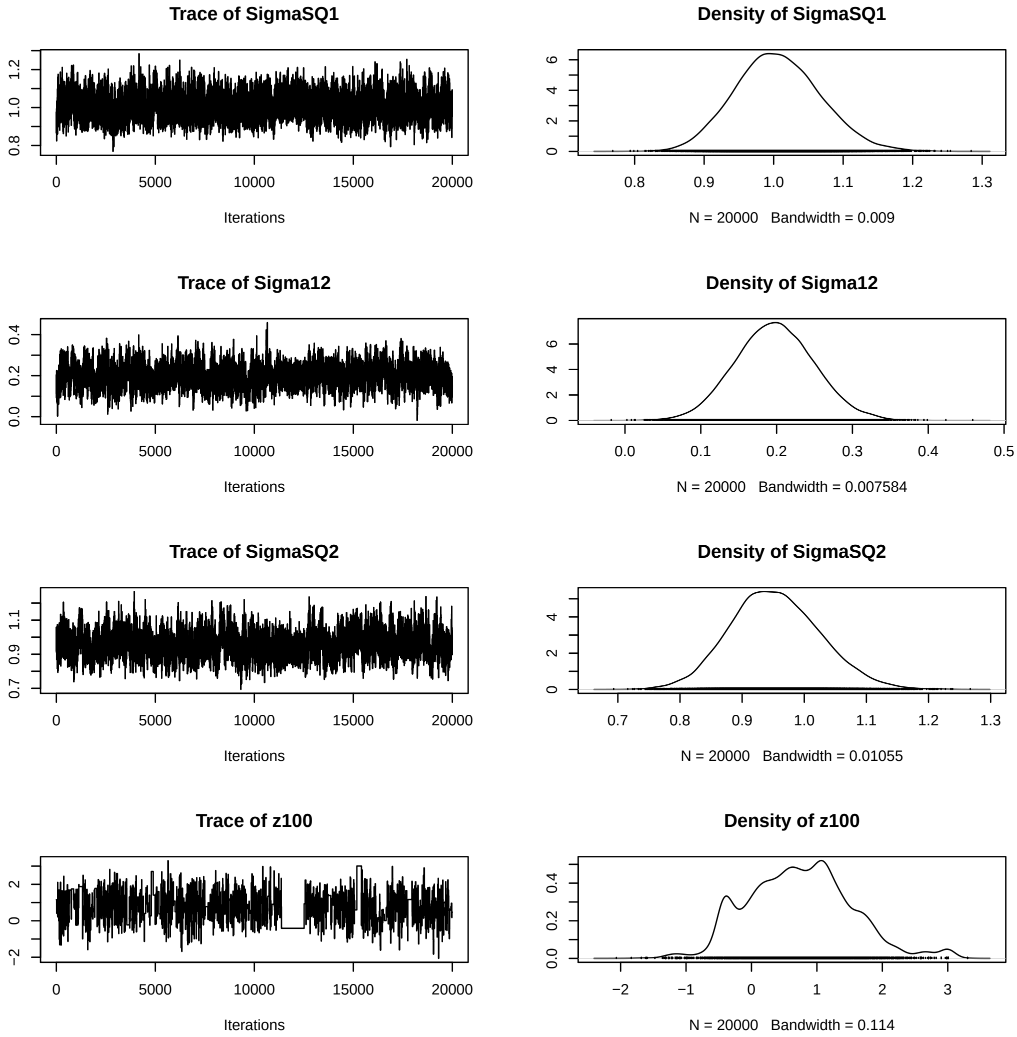

FIGURE 4.12: Trace plot (left penal) and kernel density estimation (right penal) of Bayesian posterior distributions for $\sigma_1^2, \sigma_{12}, \sigma_2^2$ and z_{100} ($T = 800$) based on 20,000 retained MCMC draws (with 5,000 burn-in).

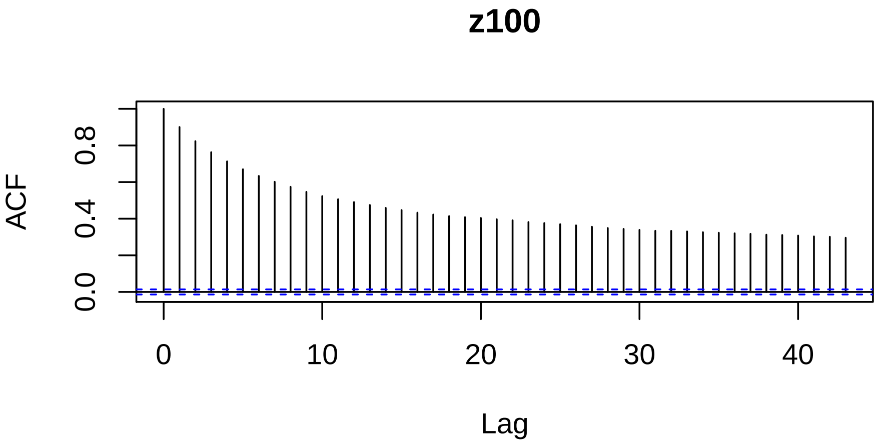

FIGURE 4.13: Autocorrelation plot for z_{100} ($T = 800$) based on 20,000 retained MCMC draws (with 5,000 burn-in).

We graph the trace plot and the kernel density estimation for static parameters based on retained MCMC draws. The trace plot exhibits increasing persistence in Markov chains as sample size increases, most notably, for β_2, σ_{12}, σ_2^2 and z^*_{100}. The parameters related to the second equation is estimated less efficient than the other parameters due to the MH acceptance ratio of $\mathbf{z}^*$ decreases as the dimension of the state vector increase, resulting in slow convergence of the chain of parameters involved in the equation of $\mathbf{z}^*$. As one important feature of the model, larger sample size implies larger amounts of latent variables z^* to impute, consequently slowing down the convergence of the Markov chain. The kernel density estimation of the posterior distributions of the static parameters exhibits decreasing variances around true parameter values for larger samples. Although the estimated posterior mean is inaccurate for smaller samples, the true parameter value still well lies within the 90% creditable interval.

Parameters	$T = 200$	$T = 400$	$T = 800$
α_1	4.84	8.99	10.80
α_2	4.54	7.01	9.47
β_1	5.05	7.92	8.07
β_2	17.26	26.92	27.34
σ_1^2	7.14	6.51	10.70
σ_{12}	10.14	20.18	21.63
σ_2^2	9.07	12.32	21.07
z^*_{100}	8.74	9.88	26.04
$\bar{\alpha}$	41.49%	33.31%	14.59%

Table 4.1: Inefficiency factors for all static parameter and z_{100}. Last row reports the percentage of accepted MH draws ($\bar{\alpha}$) from 20,000 MCMC iterations.

The performance of an MCMC algorithm is often gauged by the inefficiency factors (IFs) associated with the unknown quantities. An IF for a given parameter α, say, is calculated by $IF(\alpha) = 1 + 2\sum_{l=1}^{L} \rho(l)$, where $\rho(l)$ is the l-th order sample autocorrelation associated with the MCMC draws of α. To accommodate a high degree of autocorrelation in consecutive parameter draws, we set $L = 50$. The resulting IFs for a collection of unknows for the semi-parametric VAR example are summarized in Table 4.1, suggesting several conclusions. First, although the inefficient factor for z^*_{100} increases fast as the sample size increase, it still lies in the range of $5\% - 95\%$ suggested by the rule of thumb. Second, the inefficient factors for β_2 are the largest and hence may require longer Markov chain to obtain more accurate estimation of β_2. Finally, Table 4.1 reveals that the inefficiency factors rises with larger T, partially related to the decreased HM acceptance ratio. To achieve reasonable MH acceptance ratio, we recommend to partition the state vector into sub-blocks to control accumulated approximation errors within the MH scheme.

4.4 Conclusion

This chapter introduces an efficient approach to analyzing data sampled at different frequencies with a semi-parametric VAR model. The model includes both linear and nonparametric components, where low frequency variables may enter the equation of high-frequency response variables non-parametrically. As such, the semi-parametric VAR model may be used to investigate nonlinear relationship between variables and perform multivariate forecasting simultaneously.

For the proposed semi-parametric VAR model, a MCMC algorithm is developed by exploiting the idea of SMP method. An important feature of this algorithms is that the filtering distribution is approximated by a mixture of Gaussian components. In particular, conditional on the mixture indicators, the nonlinear model reduces to a linear Gaussian state space model. The "missing" observations are then imputed using the efficient FFBS algorithm and accepted with an appropriate MH acceptance probability. The proposed MCMC sampler can be modified to accommodate standard nonlinear state space models with latent variable enters a non-parametric function. The proposed semi-parametric MFVAR model is useful in macroeconomic studies such as investigating central bank response function to GDP and inflation. Finally, a Monte Carlo simulation study shows that the proposed MCMC sampler performs well in terms of estimating both static parameters and nonparametric functions.

Chapter 5

Portfolio Selection via Targeted Regularization

Recent developments in high-dimensional statistics and machine learning have led to significant improvements in classical mean-variance portfolio selection techniques. In this chapter, we adopt methods popular in machine learning to propose a novel approach to portfolio selection. Our method takes advantage of both subset resampling (Shen and Wang, 2017) and parameter regularization (Fan, Zhang, and Yu, 2012) within a unified framework. By exploiting a hierarchical clustering algorithm, we randomly sample subsets of assets with controlled maximum correlation. These subsets are used as regularization targets in constructing subportfolios which are then averaged to stabilize the final portfolio weight estimates. We show that the resulting portfolio strategy compares favorably with state-of-the-art strategies across a range of portfolio performance evaluation criteria.

5.1 Introduction

Portfolio selection lies at the heart of optimal investment decisions made by individual investors and financial institutions. Recent developments in machine learning and high-dimensional statistics have led to major considerable advances in empirical portfolio selection methods. Investors typically evaluate portfolios according to a range of criteria according to their own investment goals, including the Sharpe ratio, volatility, turnover rate and so on. Of the recently introduced methods for optimal portfolio selection, none has emerged a clear winner across all performance evaluation criteria.

Markowitz's mean-variance framework laid the foundations of modern portfolio theory (Markowitz, 1952). For a given desired return, the framework prescribes that investors minimize portfolio risk, as measured by return variance. This framework is among the most comprehensively studied. It has spurred an expansive theoretical literature, including the celebrated capital asset pricing model (CAPM). Yet its empirical results have been disappointing, and a large literature documents its shortcomings. See for instance Brandt (2009).

In empirical finance, the mean-variance framework requires estimation of input parameters (i.e. mean and variance of asset return) to determine optimal portfolio weights. As portfolio size (and hence the data dimension) increases, estimates of input parameters becomes less and less reliable. This renders estimated optimal weights sub-optimal, and results in poor out-of-sample performance. This problem is exacerbated by the unstable nature of the underlying model and parameters. In practice, training sample period must remain relatively short.

Many authors have attempted to address the problems associated with poor finite-sample performance of mean-variance based portfolio weights. Two popular 'schools of thought' include subset resampling and regularization. Shen and Wang (2017) provide one example of a resampling approach, in turn inspired by

Michaud (1989). The authors devise a subset resampling method for portfolio choice based on ensemble learning methods popular in machining leaning. They estimate the optimal mean-variance portfolio weights for many small subsets of risky assets drawn at random from the portfolio, to control for the error caused by high-dimensionality. The subportfolio weights are then averaged to estimate overall portfolio weights. This method is in line with ensemble methods like random subspace methods and random forests, in that the resulting portfolio is constructed by averaging over many random subspaces of the feature space. The size of the subsets is a hyperparameter that plays a crucial role in this approach.

The second school of thought imposes regularization constraints on portfolio weights. This approach reflects the trade-off between bias and variance. In portfolio choice with large-scale portfolios, the shift in bias induced by imposing (possibly incorrect) constraints on portfolio weights is rather small, whereas the associated reduction in estimation error is substantial. Jagannathan and Ma (2003) show that the no-short-sale constraint in the Markowitz mean-variance portfolio can be considered as a form of regularization. Fan, Zhang, and Yu (2012) generalize the no-short-sale constraint to gross exposure by relaxing the no-short-sale constraint to no-extreme-short-sale-or-long-position, since the no-short-sale constraint leads to portfolios that are not diversified enough in practice. More recently, Shen, Wang, and Ma (2014) expand Fan, Zhang, and Yu (2012) by employing double regularization for portfolio choice, obtaining a portfolio that consists of a limited set of assets and controls changes in asset positions. However, these approaches become less reliable with strongly correlated assets.

Standing on the success of these two separate schools of thought, we propose a novel 'target-based' regularization method for portfolio choice. We proceed in two steps. First, we divide the assets into correlated clusters and construct subsets by randomly drawing one element from each cluster. This procedure follows the approach of Bühlmann et al. (2013) and is determined by the min-max of the canonical correlation between clusters. Optimal portfolio weights are computed for each subset (called 'targets'). Next, we impose regularization constraints determined by the target portfolio weights. Results for each of the subsets are averaged to achieve our final estimated optimal portfolio weights. The resulting portfolio strategy combines the advantages of both approaches, within a unified framework. In fact, this strategy nests the subset resampling of Shen and Wang (2017) as well as the gross exposure constraints in Fan, Zhang, and Yu (2012).

Our approach also addresses the drawbacks of both methods, while simultaneously offering additional benefits. On regularization, because the targets are given by subset elements drawn from different clusters and are therefore less correlated, we achieve much more robust estimates of optimal portfolio weights. On subset resampling, we offer a data-driven means of selecting the size and composition of each subset by making use of the correlation structure between assets rather than selecting arbitrarily-sized subsets at random. Empirically, this is a useful technique because in many large-scale portfolios the risky assets can be divided into just two or three highly correlated clusters.

In addition, we promote better diversification and cushion the effects of induced biases in two ways. First, because subsets are composed of elements drawn from clusters of highly correlated assets, each element will be somewhat representative of its respective cluster. The bias induced by focusing on a small subset is reduced. Second, diversification is further improved as the regularization will not entirely reduce the portfolio to any particular subset. Rather, some of the weights will survive in each case. As such, there is a distinct improvement in diversification compared to simply taking the optimal weights for the subset itself.

We conduct simulation studies to show how the correlation structure could affect the finite sample performance of various portfolio strategies. Our analysis compares our strategy with a range of alternatives using Fama and French, Russel 200, and the S&P 500 dataset.

5.2 Relevant Literature

The need to mitigate estimation risk arising from parameter uncertainty has long been a focus of the machine learning community. Ensemble methods (Zhou, 2012) have gained recognition in a variety of disciplines, mainly due to their ability to improve the prediction performance of weak learners. As in portfolio selection, the accuracy of an ensemble of models is characterized by the diversity of its generation mechanism (Rokach, 2010). Diverse ensembles of learners often exhibit less correlated predictions, which in turn improves prediction accuracy (Hu, 2001). One class of ensemble methods, random subspace methods (Ho, 1998), is particularly appealing for improving diversity in ensembles by training weak learners with data comprising by various feature subsets (Polikar, 2006). Consequently, methods based on random subspaces have recently drawn much attention. For example, Breiman (1996) proposes a innovative variant of random space method based on decision trees called random forest. Another recent example is the Bag of Little Bootstraps (BLB) of Kleiner et al. (2014), which assess the quality of estimators when dealing with very large datasets. More recently, Shen and Wang (2017) consider subset resampling for portfolio selection. The subset resampling method is concurrently and independently developed by Elliott, Gargano, and Timmermann (2013) in econometrics for economic forecasting.

Meanwhile, regularization methods have been particularly effective for controlling the accumulation of estimation error in the optimization procedures DeMiguel et al. (2009) and Fan, Zhang, and Yu (2012). Chief among them is ℓ_1 norm regularization, which achieves structured sparsity (Huang, Zhang, and

Metaxas, 2011). For example, Shen, Wang, and Ma (2014) apply ℓ_1 norm regularization to allocate invested wealth sparsely across assets. Building on this, Fan, Zhang, and Yu (2012) propose portfolio selection by imposing the gross exposure constraint and Shen, Wang, and Ma (2014) take into account both sparse selection and portfolio turnover by forming a doubly regularized formulation. In addition, Meinshausen and Bühlmann (2010) demonstrate the use of randomized regularization methods for stable feature selection in the presence of high correlation. Statistically speaking, the idea behind norm regularization is to shrink an unbiased estimator towards a lower variance target, which is in line with the notion that bias toward simplicity can improve learning generalizations in machine learning (Mitchell, 1980). However, it is also well-known that norm regularization is sensitive to the choice of the regularization parameter, which in turn needs to account for the correlation structure of features (Dalalyan, Hebiri, and Lederer, 2017). Therefore, to construct a portfolio that performs well out-of-sample with a small training sample size, we resort to exploiting the aforementioned advantages of both ensemble and regularization methods.

5.3 Methodology

This section introduces the Markowitz portfolio selection model and discusses the properties and characteristics of our proposed portfolio strategy.

Suppose we have n risky assets with their respective returns $\{R_1, R_2, ..., R_n\}$ and for each asset we have τ observations, where τ is the time block considered for the optimal portfolio construction. Due to the probable presence of time varying behaviors of the underlying models and parameters, τ is set to be slightly larger than the number of risky assets. The return vector of the n risky assets at time t is denoted as $\boldsymbol{R}_t$. Its mean ($\mathbb{E}(\boldsymbol{R}_t)$) and variance-covariance matrix ($\mathbb{E}(\boldsymbol{R}_t\boldsymbol{R}_t')$), within the time block τ, are denoted as $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively.

The quantity of primary interest is $\boldsymbol{\omega}$, the portfolio allocation vector satisfying $\boldsymbol{\omega}'\iota = 1$ where ι is an n-sized vector of ones.

5.3.1 Mean-variance optimization and its variants

Under the Markowitz mean-variance theory, taking returns on assets as random variables, the portfolio choice problem boils down to selecting an optimal set of weights for the available assets. The Markowitz mean-variance criterion can be formulated as

$$\min_{\boldsymbol{\omega}} \boldsymbol{\omega}'\hat{\boldsymbol{\Sigma}}\boldsymbol{\omega} \text{ s.t } \boldsymbol{\omega}'\iota = 1 \text{ and } \boldsymbol{\omega}'\boldsymbol{\mu} \geq \bar{R} \tag{5.1}$$

where $\hat{\boldsymbol{\Sigma}}$ is the estimated variance-covariance matrix of risky assets and $\bar{R}$ is a target return.

In what follows, the focus is on risk minimization as in (5.2) since it is well-documented that expected return $\boldsymbol{\mu}$ is hard to estimate accurately and therefore, in empirical finance, the primary focus is shifted to variance only, treating the target return condition as given (Merton, 1980; Rapach and Zhou, 2013).

$$\min_{\boldsymbol{\omega}} \boldsymbol{\omega}'\hat{\boldsymbol{\Sigma}}\boldsymbol{\omega} \text{ s.t. } \boldsymbol{\omega}'\boldsymbol{\iota} = 1 \tag{5.2}$$

The optimization in (5.2) can be easily computed using a relatively simple quadratic programming method.

Nevertheless, the solution to (5.2) is disappointing at best, the primary culprit for which being the accumulation of estimation errors amplified by the presence of a large number of assets. Consequently, variance estimation errors could be unbounded unless their accumulation is controlled in some way. One way to do this is by imposing constraints, i.e.

$$\min_{\boldsymbol{\omega}}[\boldsymbol{\omega}'\hat{\boldsymbol{\Sigma}}\boldsymbol{\omega} + P_{\lambda}(\boldsymbol{\omega})] \text{ s.t. } \boldsymbol{\omega}'\boldsymbol{\iota} = 1, \tag{5.3}$$

where $P_\lambda(\boldsymbol{\omega})$ is a penalty function.

Most regularization methods can be characterized as variants of (5.3). For instance, for gross exposure constraints in Fan, Zhang, and Yu, 2012, $P_\lambda(\boldsymbol{\omega}) = \lambda\|\boldsymbol{\omega}\|_1$, where $\|\cdot\|_p$ is the ℓ_p norm corresponding to $\|\boldsymbol{\omega}\|_1 \leq c$, where c is the gross exposure parameter. For double regularization in Shen, Wang, and Ma (2014), $P_\lambda(\boldsymbol{\omega}) = \lambda_1\|\boldsymbol{\omega}\|_1 + \lambda_2\|\boldsymbol{\omega} - \boldsymbol{\omega}_-\|_2^2$ where $\boldsymbol{\omega}_- = (\boldsymbol{\omega}_{-1} \circ \boldsymbol{R})/\boldsymbol{R}'\boldsymbol{\omega}_{-1}$ is the re-normalized portfolio weight vector before rebalancing and $\circ$ denotes the Hadamard product. The no-short-sale constraint in Jagannathan and Ma (2003) is considered as a form of a regularization such that $\|\boldsymbol{\omega}\|_1 = 1$ as explained in Brandt (2009).

Another approach to controlling the accumulation of estimation errors, along the lines of ensemble methods, could be formulated as

$$\min_{\boldsymbol{\omega}} \boldsymbol{\omega}'_{\mathbb{S}_j}\hat{\boldsymbol{\Sigma}}_{\mathbb{S}_j}\boldsymbol{\omega}_{\mathbb{S}_j} \text{ s.t. } \boldsymbol{\omega}'_{\mathbb{S}_j}\boldsymbol{\iota} = 1, \tag{5.4}$$

where $\mathbb{S}_j, j = 1, ..., J$ is a subset of the assets chosen by random draws from the uniform distribution once the size of the subset is determined. From each subset $j = 1, ..., J$, $\hat{\boldsymbol{\omega}}_j$ is obtained. Then, the subset resampling optimal portfolio weights becomes $\hat{\boldsymbol{\omega}} = 1/J\sum_{j=1}^{J}\hat{\boldsymbol{\omega}}_j$.

5.3.2 Target-based Regularized Portfolio

The Target-based Regularized Portfolio (TRP) is the average of J sets of estimated weights $\hat{\boldsymbol{\omega}}_j$ for $j = 1, ..., J$. Each set of weights is obtained by solving, with respect to $\boldsymbol{\omega}$,

$$\hat{\boldsymbol{\omega}}_j = \arg\min \boldsymbol{\omega}'\hat{\boldsymbol{\Sigma}}\boldsymbol{\omega} + \lambda_j||\boldsymbol{\omega}'S_j||_1 \quad \text{s.t.} \quad \boldsymbol{\omega}'\iota = 1, \tag{5.5}$$

where S_j is a selection matrix and λ_j denotes the corresponding regularization coefficient. S_j is constructed using subsets $\mathbb{S}_j$ as below. We refer to the assets chosen in each subset $\mathbb{S}_j$ as the targeted assets.

An important feature of the TRP comes down to the formulation of the S_j matrix, which is an $n \times n$ diagonal matrix with ones in cells corresponding to targeted assets and $1/\alpha$ for the rest where $\alpha \in (0, 1]$ imposes a perturbation on the regularization weights, expediting shrinkage towards the targeted assets as λ increases. Note that this formulation is connected to the existing literature for particular choices of λ and α. For example, when $\alpha = 1$, the TRP reduces to the gross exposure constrained portfolio proposed by Fan, Zhang, and Yu (2012), while small λ and infinitesimal α leads to a formulation of Shen and Wang (2017).

Before performing the optimization, the targeted assets must first be chosen. We employ the data-driven hierarchical clustering algorithm proposed by Bühlmann et al. (2013), which effectively determines the target set using canonical correlation (Anderson, 1958). The algorithm allocates assets into clusters such that the maximum canonical correlations between clusters are minimized. It proceeds by searching for such a partition $\mathcal{G}$ of q disjoint clusters $\{G_1, ..., G_q\}$ that satisfies this property. More precisely, we construct $\hat{\mathcal{G}}$ based on the following criterion: for $q = 1, ..., n$,

$$\begin{aligned}\hat{\mathcal{G}}(\hat{q}) = \text{Partition } \hat{\mathcal{G}} \text{ consisting of } (\hat{G}_1, \cdots, \hat{G}_{\hat{q}}),\ \text{with}\\ \hat{q} = \arg\min_{q} \rho_{\max}(\hat{\mathcal{G}}(q)) \text{ such that}\\ \rho_{\max}(\hat{\mathcal{G}}(q)) = \max\{\hat{\rho}_{\text{can}}(G_l, G_k); l, k \in \{1, \cdots, q\}, l \neq k\}\end{aligned}$$

where $\hat{\rho}_{\text{can}}(G_l, G_k)$ represents the empirical canonical correlation between assets from the l-th and k-th clusters respectively. The procedure is summarized in Algorithm 13.

After the partition $\hat{\mathcal{G}}$ is obtained via Algorithm 13, to ensure the maximum correlation between assets in a target set is controlled, a target set can be formed by choosing a subset indexed by $\mathbb{S}_j = \{s_1, \cdots, s_{\hat{q}}\}$ such that $R_{s_i} \in G_i$

Algorithm 13 Hierarchical Clustering

```
**Inputs:** $\{\boldsymbol{R}_t\}_{t=1}^{\tau}$: historical returns data;
$\hat{\mathcal{G}}^n$ = single asset as $q = n$ clusters
**for** $q = (n-1) \to 1$ **do**
    $\hat{\mathcal{G}}^q$ = a partition with $q$ clusters by merging $G_l^{q+1}$ and $G_k^{q+1}$ for $\hat{\rho}_{\text{can}}(G_l^{q+1}, G_k^{q+1}) = \hat{\rho}_{\max}(\hat{\mathcal{G}}^{q+1})$;
$\hat{\mathcal{G}} = \hat{\mathcal{G}}^{\hat{q}}$, where $\hat{q} = \arg\min \hat{\rho}_{\max}(\hat{\mathcal{G}}(q))$
**Outputs:** A partition $\hat{\mathcal{G}}$ of assets.
```

for $i = 1, \cdots, \hat{q}$, where $\hat{q} = \|\mathbb{S}_j\|_0$, taking one asset from each cluster. Therefore, q is the number of clusters and at the same time, the size of each subset. Once a target portfolio is determined, a more diversified portfolio can be constructed based on the portfolio weights obtained by solving (5.5).

Algorithm 14 Target-based Regularized Portfolio

```
**Inputs:** $\tau$: number of period for estimation; $\{\boldsymbol{R}_t\}_{t=1}^{\tau}$: historical returns; $\boldsymbol{R}_{\tau+1}$: out-of-sample returns; $n$: number of assets; $\hat{\mathcal{G}} = \{G_1, \cdots, G_{\hat{q}}\}$: a partition of assets; $J$: number of sampled target portfolios;
**for** $j = 1 \to J$ **do**
    Sample an index set $\mathbb{S}_j = \{s_1^j, \cdots, s_{\hat{q}}^j\}$ uniformly at random such that $R_{\mathbb{S}_{ji}} \in G_i$ and construct $S_j$;
    Compute the optimal portfolio weight vector, $\boldsymbol{\omega}_\tau^j$, by solving (5.5);
Aggregate the preliminary portfolio weights based on $J$ samples $\hat{\boldsymbol{\omega}}_\tau = J^{-1}\sum_{j=1}^{J}\hat{\boldsymbol{\omega}}_\tau^j$;
Compute the out-of-sample portfolio net return $\hat{r}_{\tau+1} = \boldsymbol{R}'_{\tau+1}\hat{\boldsymbol{\omega}}_\tau$;
**Outputs:** A vector of portfolio weights $\hat{\boldsymbol{\omega}}_\tau$ and the corresponding portfolio net return $\hat{r}_{\tau+1}$.
```

If applied with only one target set, the target-based regularization is not particularly useful even if the target portfolio is well-constructed. However, by applying the target-based regularization many times with randomly sampled target sets and taking the average, we obtain a very stable final portfolio estimate. The procedure is summarized in Algorithm 14.

5.3.3 Discussion

TRP differs from existing strategies in two main aspects: (1) the subset selection technique and (2) regularization towards targeted assets. There are some chapters on portfolio selection based on pre-determined targets, such as Ledoit and Wolf (2003) and Ledoit and Wolf (2004), however the choice of targets has a significant impact on performance and in practice it is difficult to select targets a priori. Instead of relying on targets based on some prior assumptions, we use hierarchical clustering for the TRP to control correlation between elements in each subset and that the size of subsets is determined in a data-driven way as the size of subsets is always equal to the number of clusters.

The benefits of the TRP in this respect are twofold. The proposed strategy is robust due to the low level of correlation between targeted assets when constructing an optimal portfolio via regularization. In addition, the TRP promotes diversification in line with portfolio selection theory in two distinct ways. First, the number of nonzero elements of $\hat{\boldsymbol{\omega}}_j$ from the TRP is greater than or equal to the number of nonzero elements of $\hat{\boldsymbol{\omega}}_{\mathbb{S}_j}$, where $\hat{\boldsymbol{\omega}}_{\mathbb{S}_j}$ is obtained using the subset of assets $\mathbb{S}_j$ without regularization. Second, the TRP precludes the possibility that assets are chosen from the same cluster, and therefore the coverage of the TRP target will be at least equal to any randomly chosen subset, if not larger.

The TRP depends on the hyperparameter α, which is loosely related to the randomized lasso Meinshausen and Bühlmann (2010) and weak greedy algorithms Temlyakov (2000), in which it is referred to as 'weakness'. In our simulations and empirical studies, the choice of α does not particularly affect the performance of the TRP. Figure 1 illustrates the performance of the TRP for different α values.

In a simulation study, the log ℓ_1 norm of the differences in weight vectors (top row) and the log differences in actual risk (bottom row) between the TRP estimate and the true weights, along with the 10^{th} and 90^{th} percentiles, based on different perturbation parameter values $\alpha = (0.01, 0.2, 0.4, 0.6, 0.8, 1)$ across different correlation structures are plotted in Figure 5.1 to Figure 5.4. Asset returns are assumed to follow a multivariate normal distribution with mean

$\boldsymbol{\mu}_{\text{sim}} = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{\text{sim}}$ with $\tau = 120$ and $n = 100$, where we assume $\boldsymbol{\Sigma}_{\text{sim}}$ has a block-diagonal structure (10 blocks) with equi-correlation within blocks (10 assets within each block) Bühlmann et al. (2013) and the correlations are 0.9 within each block and the between-block correlation parameter ρ takes value from each of the vector $(0, 0.5, 0.75, 0.8)$.

Two theoretical metrics are used to evaluate the performance of the TRP as compared to the true optimal weights: (i) the ℓ_1 distance, i.e., $||\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}||_1$ and (ii) the difference in actual risk, i.e. $\hat{\boldsymbol{\omega}}'\boldsymbol{\Sigma}_{\text{sim}}\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}'\boldsymbol{\Sigma}_{\text{sim}}\boldsymbol{\omega}$. These two metrics are calculated based on the theoretical quantities $\boldsymbol{\omega}$ and $\boldsymbol{\Sigma}_{\text{sim}}$. These parameters are unknown in practice, but in simulations can effectively characterize the estimation error of the weight vector and the associated portfolio risk. Notably, with an appropriately chosen α, there are promising improvements over the gross exposure constrained portfolios, namely, the case when $\alpha = 1$. Furthermore, our simulation study indicates that a smaller α is more favorable when ρ is small, while the performances become indifferent for $\alpha \in (0, 1)$ when ρ is large. In particular, Meinshausen and Bühlmann (2010) illuminates that there is an intrinsic trade-off between a well-posted design matrix (large α) and better chance for sparse recovery (small α).

As pointed out by Shen, Wang, and Ma (2014), the optimization problem in (5.5) is not straightforward to solve due to the sum-to-one constraint. For given λ, the problem can be solved using an efficient first-order algorithm such as the Alternating Direction Method of Multipliers (ADMM) (Boyd et al., 2011). However in general, to the best of our knowledge, a solution path algorithm for this type of constrained formulation has only been recently studied by Gaines, Kim, and Zhou (2018) in the context of regression. The optimization problem in (5.5) can be recast as an ℓ_1 norm regularized least-square problem:

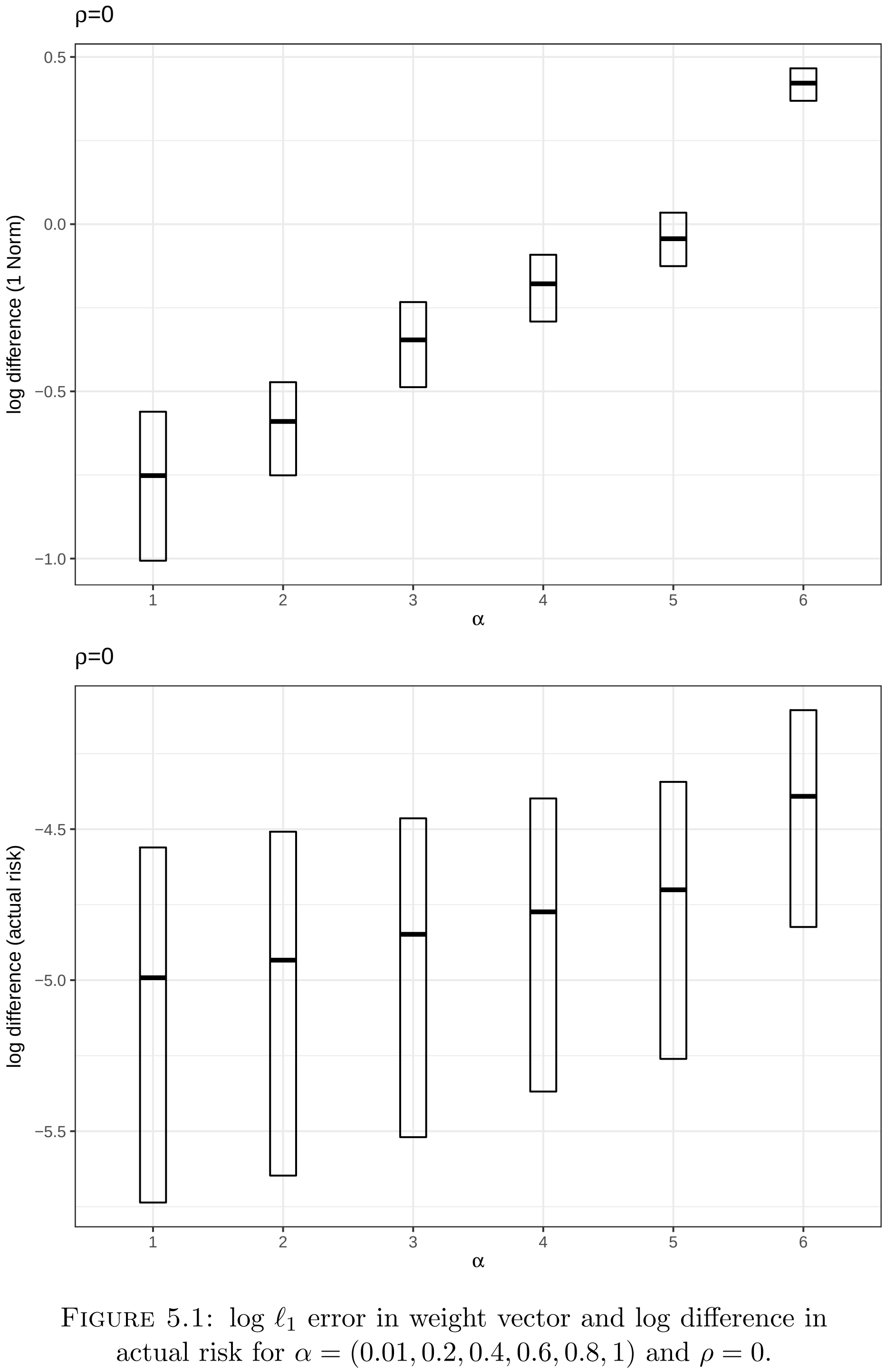

FIGURE 5.1: log ℓ_1 error in weight vector and log difference in actual risk for $\alpha = (0.01, 0.2, 0.4, 0.6, 0.8, 1)$ and $\rho = 0$.

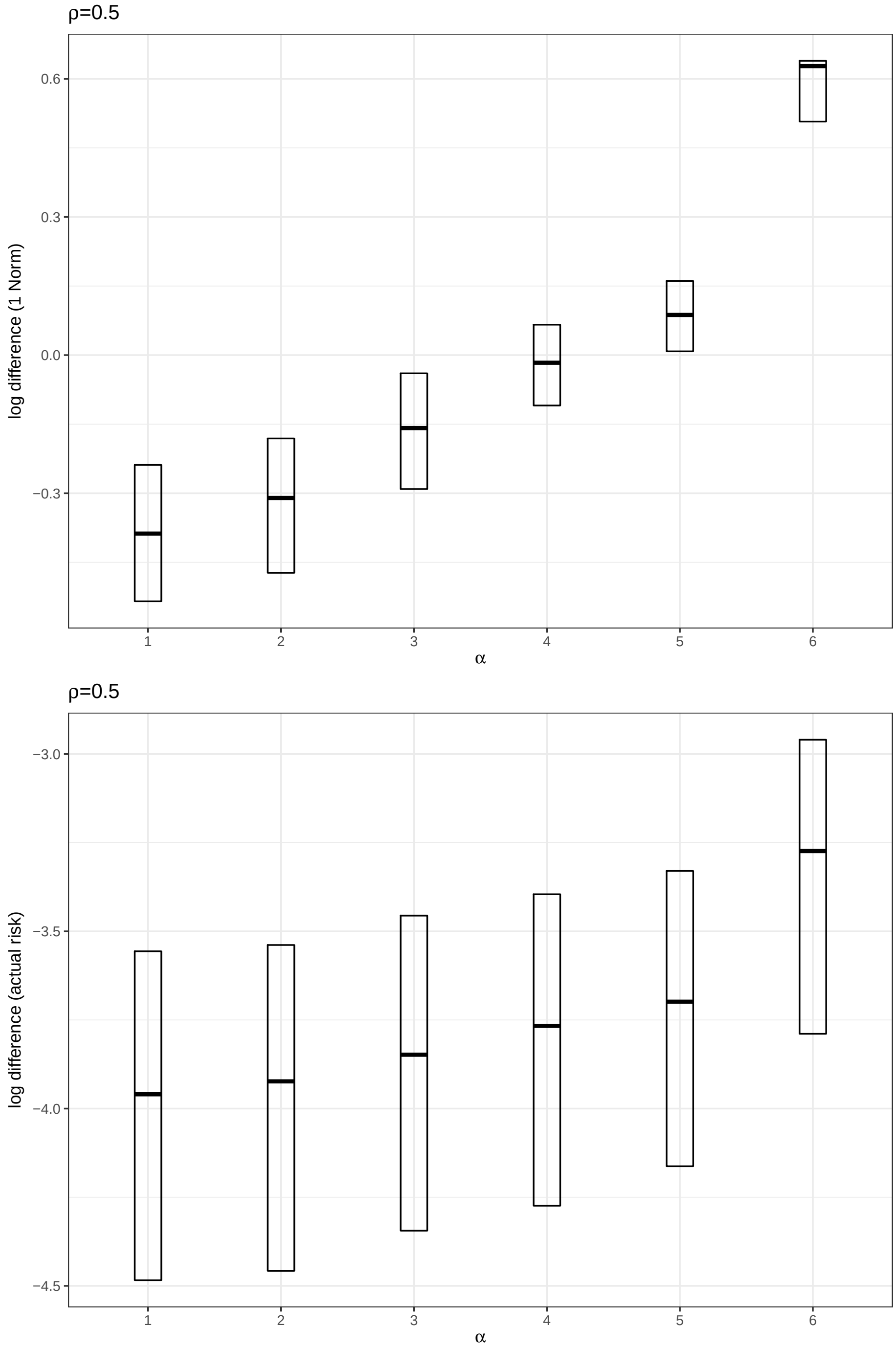

FIGURE 5.2: log ℓ_1 error in weight vector and log difference in actual risk for $\alpha = (0.01, 0.2, 0.4, 0.6, 0.8, 1)$ and $\rho = 0.5$.

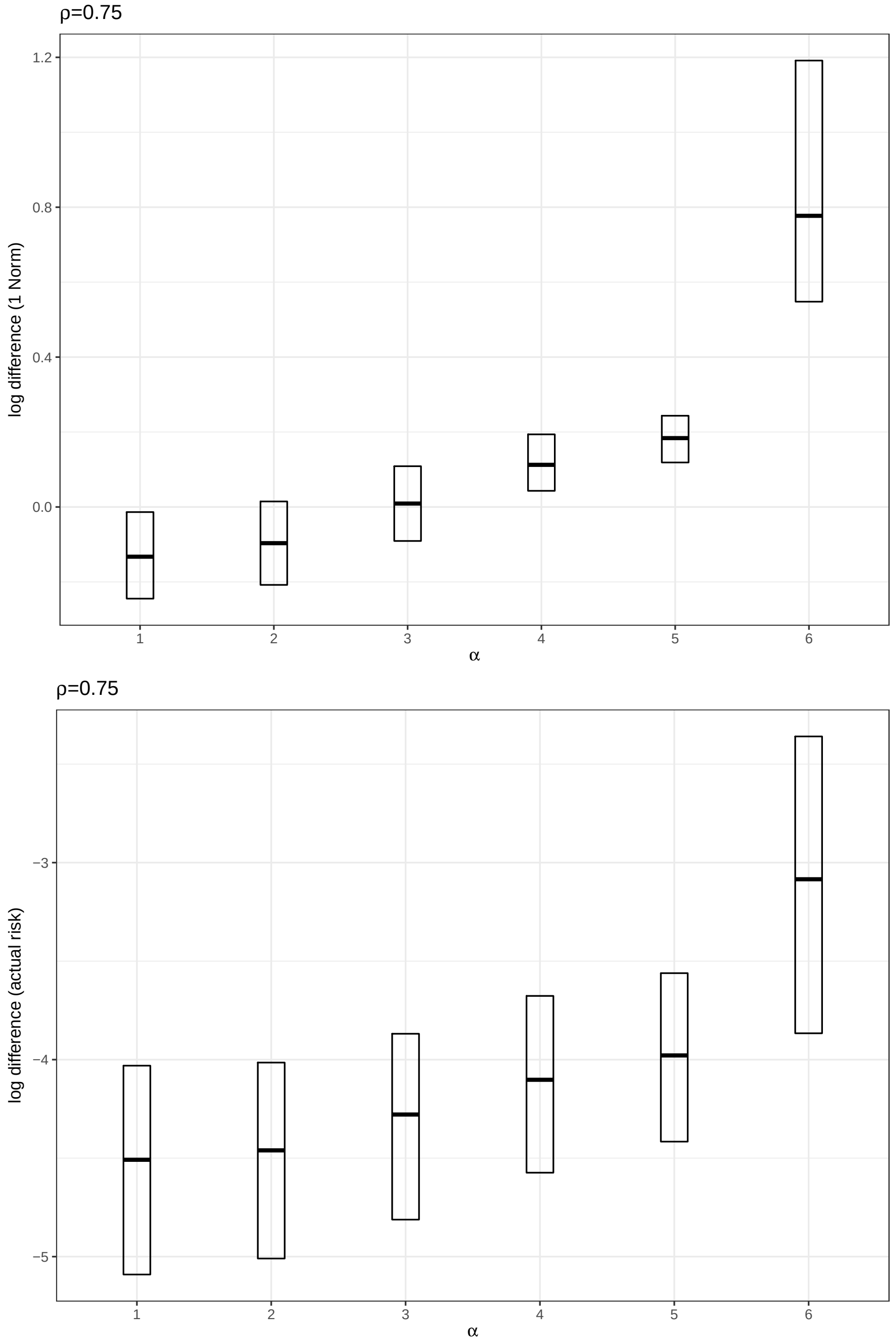

FIGURE 5.3: log ℓ_1 error in weight vector and log difference in actual risk for $\alpha = (0.01, 0.2, 0.4, 0.6, 0.8, 1)$ and $\rho = 0.75$.

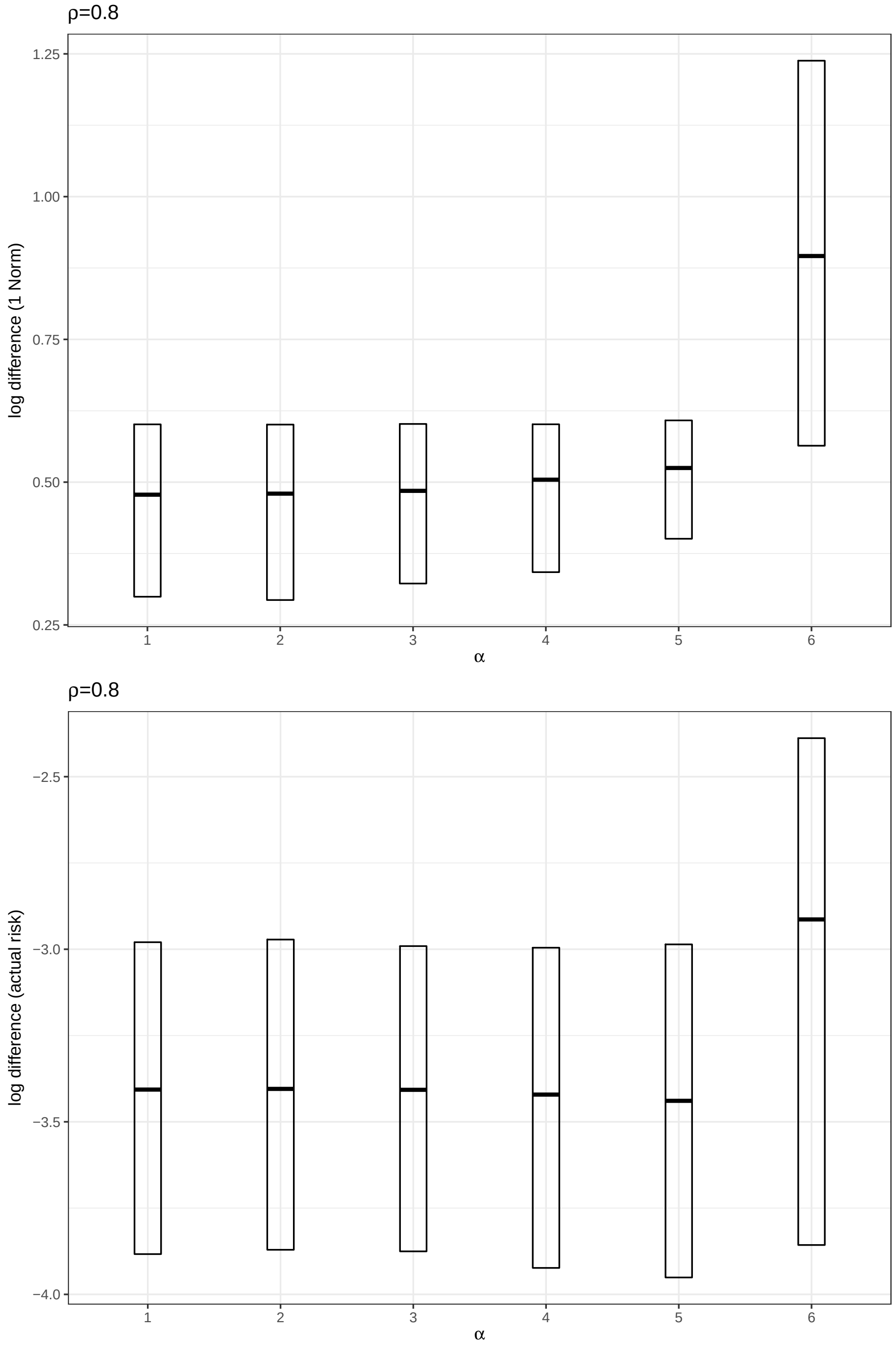

FIGURE 5.4: log ℓ_1 error in weight vector and log difference in actual risk for $\alpha = (0.01, 0.2, 0.4, 0.6, 0.8, 1)$ and $\rho = 0.8$.

$$\begin{aligned} &\min \mathbb{E}(\boldsymbol{\omega}'\boldsymbol{\Sigma}\boldsymbol{\omega}) + \lambda||\boldsymbol{\omega}'S||_1 \text{ s.t. } \boldsymbol{\omega}'\iota = 1 \\ =&\min \mathbb{E}(\boldsymbol{X}'\boldsymbol{\beta})^2 + \lambda||\boldsymbol{\beta}||_1 \text{ s.t. } \boldsymbol{\beta}S^{-1}\iota = 1, \end{aligned} \tag{5.6}$$

where $\boldsymbol{X} = S^{-1}\boldsymbol{R}$ and $\boldsymbol{\beta} = \boldsymbol{\omega}'S$. A solution path algorithm can efficiently solve for $\boldsymbol{\omega}(\lambda)$ along the path of $\lambda \in [0, \infty)$. Another approximation method has been proposed by Fan, Zhang, and Yu (2012), but its performance and theoretical soundness have not yet been fully vindicated. Similar problems, without the sum-to-one constraint, have been studied by Efron et al. (2004) resulting in what they call the LARS algorithm. In our empirical experiments, we use the ADMM method with a given λ for simplicity.

Moreover, the constrained optimization problem (5.6) for each subportfolio can be written into a Lagrangian, whose solution is obtained by minimizing

$$\boldsymbol{\beta}S^{-1}\hat{\boldsymbol{\Sigma}}S^{-1}\boldsymbol{\beta}'/2 + \lambda_1(||\boldsymbol{\beta}||_1 - c) + \lambda_2(1 - \boldsymbol{\beta}S^{-1}\iota), \tag{5.7}$$

where c is the constraint parameter that corresponds to λ in the Lagrangian dual problem of (5.6). Let $\boldsymbol{g}$ be the subgradient vector of the function $||\boldsymbol{\beta}||_1$, whose elements are in $[-1, 1]$. The Karush-Kuhn-Tucker conditions for the constrained optimization are

$$S^{-1}\hat{\boldsymbol{\Sigma}}S^{-1}\boldsymbol{\beta}' + \lambda_1\boldsymbol{g} - \lambda_2 S^{-1}\iota = 0, \tag{5.8}$$

$$\lambda_1(c - ||\boldsymbol{\beta}||_1) = 0, \ \lambda_1 \geq 0, \tag{5.9}$$

in addition to the constraints $\boldsymbol{\beta}S^{-1}\iota = 1$ and $||\boldsymbol{\beta}||_1 \leq c$, whose corresponding solution is denoted by $\tilde{\boldsymbol{\beta}}$.

Theorem 1 *The constrained optimization problem* (5.5) *is equivalent to the mean variance optimization*

$$\min \boldsymbol{\omega}'\tilde{\boldsymbol{\Sigma}}_S\boldsymbol{\omega} \ \textit{s.t.}\ \boldsymbol{\omega}'\iota = 1 \tag{5.10}$$

with the regularized covariance matrix

$$\tilde{\boldsymbol{\Sigma}}_S = \hat{\boldsymbol{\Sigma}} + \lambda_1(S_j\tilde{\boldsymbol{g}}\iota' + \iota\tilde{\boldsymbol{g}}'S_j),$$

where $\tilde{\boldsymbol{g}}$ *is the subgradient evaluated at* $\tilde{\boldsymbol{\omega}}$ *and* λ_1 *is the Lagrange multiplier defined in* (5.8) *and* (5.9).

Proof. The solution to the problem (5.10) is given by

$$\boldsymbol{\omega}_{\text{opt}} = \tilde{\boldsymbol{\Sigma}}_S^{-1}\iota/\iota'\tilde{\boldsymbol{\Sigma}}_S^{-1}\iota.$$

By using $\tilde{\boldsymbol{\beta}}S^{-1}\iota = 1$ and $\tilde{g}'\tilde{\boldsymbol{\beta}} = ||\tilde{\boldsymbol{\beta}}||_1$, we have

$$\begin{aligned} S^{-1}\tilde{\boldsymbol{\Sigma}}_S S^{-1}\tilde{\boldsymbol{\beta}}' &= S^{-1}\hat{\boldsymbol{\Sigma}}S^{-1}\tilde{\boldsymbol{\beta}}' + \lambda_1\tilde{\boldsymbol{g}} + \lambda_1||\boldsymbol{\beta}||_1 S^{-1}\iota \\ &= \lambda_2 S^{-1}\iota + \lambda_1 c S^{-1}\iota. \end{aligned}$$

Thus, $\tilde{\boldsymbol{\beta}}' = (\lambda_2 + \lambda_1 c)S\tilde{\boldsymbol{\Sigma}}_S^{-1}\iota$ and $S^{-1}\tilde{\boldsymbol{\beta}}' = \tilde{\boldsymbol{\omega}} = (\lambda_2 + \lambda_1 c)\tilde{\boldsymbol{\Sigma}}_S^{-1}\iota$. Note that $\boldsymbol{\omega}_{\text{opt}}$ and $\tilde{\boldsymbol{\omega}}$ are equivalent up to a constant, i.e., $\boldsymbol{\omega}_{\text{opt}} = \kappa\tilde{\boldsymbol{\omega}}$, where κ is some constant. Then, since $\kappa = 1$, they must be equal. This completes the proof. ∎

Therefore, each subportfolio is also equivalent to the optimal portfolio constructed using a regularized covariance matrix. This result is of the same spirit of Jagannathan and Ma (2003), DeMiguel et al. (2009) and Fan, Zhang, and Yu (2012).

Note that the perturbation parameter α in the formulation of (5.5) would shrink all the assets except for the targeted ones more by factor $1/\alpha$ as λ increase. While in gross exposure constrained portfolio of Fan, Zhang, and Yu, 2012, the portfolio weight vector is shrunk towards to that of the no-short-sell portfolio. To see this, without loss of generality, let the first element ω_1 be non-negative. The other cases where $\omega_j; j = 2, ..., n$ are analogous. Then,

$$\begin{aligned} ||\boldsymbol{\omega}||_1 &= \omega_1 + \sum_{j=2}^{n} |\omega_j| \qquad (5.11) \\ &= 1 + \sum_{j=2}^{n} (|\omega_j| - \omega_j) \\ &= 1 + \sum_{j=2}^{n} 2\,|\omega_j|\,1\,\{\omega_j < 0\}\,. \end{aligned}$$

Given the penalty function, the preceding penalty shrinks the negative weights toward zero even more by factor 2 than in the regular lasso estimation with the same λ, resulting in the no-short-sell portfolio for sufficiently large λ.

5.4 Experiment

In this section we will demonstrate the usefulness of the TRP by an extensive empirical experiment on several real-world benchmark datasets.

5.4.1 Data and Settings

Data: In our experiments, we intentionally choose high-dimensional datasets with a large n and a relatively small τ to fairly validate the proposed approach. For the evaluation of performance of various strategies, we consider returns over the past 20 years for out-of-sample evaluation.

TABLE 5.1: Summary of testing datasets

#	Dataset	Frequency	τ	n	Sample Period
1	FF48	Monthly	60	48	04/01/1993 - 29/06/2018
2	FF100	Daily	252	100	02/01/1998 - 29/03/2018
3	EQ141	Daily	504	141	02/01/1998 - 03/08/2018
4	SP352	Daily	504	352	02/01/1998 - 03/08/2018

Two types of datasets are considered in our experiments: (1) Fama-French Portfolios and (2) large U.S. stock market benchmarks.

Fama-French Portfolios: The Fama and French datasets have enjoyed a surge of popularity since their initial use by Fama and French (1992). Briefly speaking, the benchmarks consist of portfolios formed for representing different financial exposures. For example, FF100 consists of daily returns of 100 portfolios formed by the two-way sort of stocks according to market equity and the ratio of book equity to market equity with 10 categories in each factor. Meanwhile, FF48 contains monthly returns from 48 different industrial sectors.

U.S. stock market benchmarks: The two large U.S. stock market benchmark datasets considered in our experiments are the Russell Top 200 and the S&P 500. The Russell Top 200 index contains the 200 largest stocks based on the market capitalizations of the Russell 3000. In contrast, the S&P 500 consists of a more comprehensive range of 500 large common stocks listed on the NYSE or NASDAQ. After eliminating stocks with missing historical data, we obtain a total of 141 and 352 assets for the Russell Top 200 (EQ141) and S&P 500 (SP352) respectively.

Table 5.1 outlines the aforementioned benchmark datasets. Note that in order to gauge the performance of each strategy in the long-run and to understand the performance of each strategy under different market conditions, the out-of-sample evaluation periods are spanned over 20 years to cover the early 2000s recession and the more recent 2007 Global Financial Crisis (GFC). Also, to avoid having an estimation window that spans over a long time period, which will adversely affect the parameter estimation due to the time-varying behavior of stock returns, we employ daily data for datasets involving large-scale portfolios (FF100, EQ141 and SP352).

Competing strategies: Table 5.2 summarizes five types of portfolio strategies from the extant literature compared against in our experiments. Essentially, all the portfolio strategies considered can be viewed as shrinkage portfolios with

Table 5.2: Summary of portfolio strategies evaluated

#	Model	Abbreviation
Simple benchmarks		
1	Equally-weighted $(1/n)$ portfolio	EW
2	Value-weighted (market) portfolio	VW
Minimum-variance portfolios		
3	Minimum-variance portfolio with shortsales unconstrained	MV
4	Minimum-variance portfolio with shortsales constrained Jagannathan and Ma (2003)	JM
5	Minimum-variance portfolio with gross exposure constraint Fan, Zhang, and Yu (2012)	FZY
Portfolios based on blending strategy or weighted averages		
6	Weighted average of sample covariance and identity matrix Ledoit and Wolf (2004)	LW
7	Blending of mean-variance optimal and equally-weighted portfolio Tu and Zhou (2011)	TZ
Minimum-variance portfolios based on ensemble methods		
8	Averaging resampling portfolios Shen and Wang (2017)	SSR
9	Averaging target-based regularized portfolios	TRP
On-Line machine learning algorithms		
10	Passive aggressive mean reversion strategy Li et al. (2012)	PAMR
11	Correlation-driven nonparametric learning Li, Hoi, and Gopalkrishnan (2011)	CORN

different shrinkage targets and amounts except for MV, TZ, PAMR and CORN. In particular, JM is the well-known optimal no-short-sale portfolio, where the ℓ_1 norm of the weight vector is constrained to be one. EW and VW can be viewed as two specific types of no-short-sale portfolio, which are also the standard benchmarks widely used in the market. FZY aims to improve no-short-sale portfolios by considering a wider range of exposure coefficients. Somewhat similarly, LW is the result of shrinking the covariance matrix towards the identity matrix instead of shrinking the portfolio-weight vector as in JM. On the other hand, SW employs subsets of data, resulting in shrinkages depending on the full covariance matrix with the degree of shrinkage determined by the pair of (n, q). In addition, two innovative on-line machine learning algorithms are also compared: the on-line passive aggressive mean reversion portfolio, which uses the mean reversion behavior of asset returns, and the correlation-driven non-parametric portfolio, which exploits the correlation structure of assets.

5.4.2 Performance Metrics

To mitigate the non-stationarity of modeling parameters, we perform estimation based on rolling windows resulting in a high-dimensional setup where the number of assets is close to the sample size of the data. For $s \in \{\tau, \cdots, T-1\}$, we first determine portfolio weights $\hat{\boldsymbol{\omega}}_t$ using return data $\{\boldsymbol{R}_t\}_{t=s-\tau+1}^{s}$. The out-of-sample net return $\hat{r}_{s+1}$ is computed based on the realized gross returns, i.e., $\hat{r}_{s+1} = \boldsymbol{\omega}'_{s+1}\boldsymbol{R}_{s+1} - 1$. Accordingly, the resulting allocation is held until the next rebalancing. For each subsequent s between two consecutive rebalances, we simply set $\hat{\boldsymbol{\omega}}_s = \hat{\boldsymbol{\omega}}_{s^-}$. In our experiments, the rebalancing frequency is assumed to be monthly, that is, the portfolios are held for one month and rebalanced at the beginning of the next month. We compare out-of-sample empirical performance across the 10 portfolios using five performance metrics as follows.

(i) **Sharpe Ratio** (SR): SR has gained considerable popularity since the seminal chapter of Sharpe (1964). It summarizes the mean and variance with a simple measure of risk-adjusted return, also known as "reward-to-variability ratio". For out-of-sample evaluation, SR is calculated by $SR = \bar{r}/\bar{\sigma}$, where $\bar{r}$ denotes the mean of realized net returns and $\bar{\sigma}$ denotes the mean of realized standard deviations:

$$\bar{r} = \frac{1}{T-\tau}\sum_{t=\tau+1}^{T}\hat{r}_t, \;\; \bar{\sigma} = \sqrt{\frac{1}{T-\tau}\sum_{t=\tau+1}^{T}(\bar{r}-\hat{r}_t)^2}. \tag{5.12}$$

The annualized Sharpe ratio, $\sqrt{H}SR$, is reported, wherein H is 252 since the datasets used have daily frequency.

(ii) **Volatility** (VO): VO measures the risk associated with mean-variance portfolios for given expected return. For strategies implementing the minimum-variance portfolio, VO is the only measure that underlines the effectiveness of risk minimization. In line with SR, we report the annualized volatility, $\sqrt{H}\bar{\sigma}$.

(iii) **Turnover Rate** (TO): TO indicates the volume of portfolio rebalancing. It is an important performance metric as a high TO inevitably leads to high transition costs. Given the renormalized weight vector before rebalancing, TO is calculated by

$$\text{TO} = \frac{1}{T-\tau-1}\sum_{t=\tau+1}^{T-1}||\hat{\boldsymbol{\omega}}_t - \hat{\boldsymbol{\omega}}_{t^-}||_1, \tag{5.13}$$

that is, the average ℓ_1 norm of the difference in the weight vector across trading periods.

(iv) **Maximum Drawdown** (MDD): MDD is the maximum cumulative loss from a peak to a following trough Magdon-Ismail and Atiya (2004), representing the persistence of investment loss. During market downturn, large drawdown often leads to panic selling and subsequent fund redemption. Denoting the cumulative wealth at time j as W_j, MDD is computed by

$$\text{MDD} = \max_{t\in[\tau,T]}(M_t - W_t),$$

with

$$M_t = \max_{j\in[\tau,t]} W_j \text{ and } W_j = \prod_{l=\tau+1}^{j}(1+\hat{r}_l).$$

(v) **Gross Exposure** (GE): GE measures the total investment amount at risk in both the long position and the short position. Specifically, GE is computed as the average ℓ_1 norm of the portfolio weight vector across a tested time period:

$$\text{GE} = \frac{1}{T-\tau-1}\sum_{t=\tau+1}^{T-1}||\hat{\boldsymbol{\omega}}_t||_1, \tag{5.14}$$

In practice, although the preference of GE depends on investor's risk profile, allowing short positions is a powerful tool to mitigate risk and produce returns when the allocation of positions is appropriately chosen.

5.4.3 Results

Table 5.3 summarizes the performance of the compared portfolio strategies across the tested benchmark datasets. We denote the best performing strategy under each metric, except for GE, in bold. The proposed TRP strategy has produced the highest SR in every dataset amongst all 11 impugned portfolio selection strategies. In particular, the TRP outperforms the two passively managed benchmarks, EW and VW, with higher SRs, lower VO and lower MDD. The TRP achieves very stable performance compared to other actively managed strategies across the four datasets, illuminating its robustness when constructing portfolios using assets with different empirical characteristics. The TRP also has moderate GE and low MDD, which are indicative of its efficient risk minimization in market downturn.

Table 5.3: Summary of portfolio performance

Dataset	Metrics	TRP	EW	VW	MV	JM	FZY	LW	TZ	SSR	PAMR	CORN
FF48 (Industry)	SR	**1.09**	0.77	0.22	0.03	1.04	0.95	0.97	0.76	0.96	0.03	0.77
	p-value	0.16	1.00	0.18	0.13	0.08	0.35	0.59	0.75	0.05	0.05	1.00
	VO (%)	10.23	15.66	36.46	18.44	10.61	10.72	10.32	14.10	**9.80**	33.96	15.66
	TO (%)	110.97	24.90	**0.00**	280.51	56.21	60.14	153.12	41.77	97.58	142.44	249.01
	MDD (%)	**16.03**	26.64	44.09	33.38	20.27	18.75	19.30	25.44	20.54	51.11	26.64
	GE	1.45	1.00	1.00	9.90	1.00	1.04	2.92	1.45	2.15	1.00	1.00
FF100 (ME&BM)	SR	**2.58**	0.65	0.87	2.46	1.74	2.57	2.32	2.57	2.32	0.54	0.65
	p-value	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.99
	VO (%)	**8.37**	21.13	18.47	9.24	11.81	8.51	8.58	9.18	8.86	27.19	24.14
	TO (%)	130.34	2.55	**0.00**	319.64	25.78	184.57	211.10	287.86	67.13	158.76	17.78
	MDD (%)	8.15	19.56	17.02	**7.30**	12.56	8.22	8.58	7.88	8.15	24.06	25.74
	GE	3.12	1.00	1.00	7.14	1.00	3.49	5.68	6.42	3.12	1.00	1.00
EQ141	SR	**0.59**	0.40	0.40	0.47	0.44	0.58	0.50	0.45	**0.59**	0.38	0.40
	p-value	0.29	1.00	0.98	0.78	0.67	0.36	0.65	0.36	0.27	0.83	0.86
	VO (%)	11.79	19.00	19.00	12.44	16.34	12.16	11.96	15.39	**11.77**	25.05	19.03
	TO (%)	92.82	0.05	**0.00**	118.93	7.34	167.57	87.89	37.35	24.14	43.65	1.71
	MDD (%)	11.56	19.04	19.04	10.89	18.36	13.37	**10.50**	15.51	12.45	20.71	19.04
	GE	1.77	1.00	1.00	2.23	1.00	1.51	1.95	1.08	1.25	1.00	1.00
SP 352	SR	**0.90**	0.38	0.38	0.66	0.36	0.87	0.81	0.46	0.83	0.43	0.38
	p-value	0.02	1.00	0.45	0.29	0.50	0.03	0.08	0.02	0.10	0.03	0.89
	VO (%)	11.38	19.95	19.93	14.71	19.25	11.39	11.27	17.35	**10.26**	30.33	20.03
	TO (%)	162.29	0.06	**0.00**	478.40	0.92	168.37	193.80	85.91	37.27	67.35	1.84
	MDD (%)	11.97	19.41	19.41	16.66	19.41	11.73	**11.27**	17.83	11.35	24.68	19.41
	GE	2.30	1.00	1.00	10.15	1.00	2.25	6.00	1.99	2.18	1.00	1.00

Note: The best performing strategy for all criterion is in bold. TRP, MV, JM, FZY, LW and SSR are implemented minimum variance portfolio without constraint on the expected return. For strategies requires tuning parameters, we have used for: TRP: $\alpha = 0.8$, $J = 100$, TZ: $\gamma = 3$, SW: $q = n^{0.7}$ and $J = 2{,}000$, PAMR: $\epsilon = 0.5$, CORN: $w = 5$ and $\rho = 0.1$. The *p-values* under SR quantify the SR difference between the corresponding strategy and EW using the approach by Ledoit and Wolf (2008) based on the studentized circular block bootstrapping methodology with 5000 bootstrap resamples and a block with the size of 5.

Compared other norm regularized methods, FZY and LW, and their corresponding regularization targets, EW and JM, we see that the TRP typically has higher SR and lower VO, TO and MDD. The improvement over other norm regularized methods is attributed to the idea of averaging over a number of regularization targets, as it is unlikely to be the case that the no-short-sale portfolio or the equally weighted portfolio is the theoretical optimal portfolio over a long investment horizon. Our simulation study shows that averaging over a number of plausible targets often leads to more stabilized portfolio weights and improved risks.

Notably, the other ensemble based method, SSR, often achieves similar VO to the TRP but has substantially lower TO. In practice, low TO often leads to low transaction cost and hence is more indicative of before-cost returns. However, as illuminated by Grinblatt, Titman, and Wermers (1995), “momentum investors” who actively buys assets that were past winners have realized significantly better performance than others, mainly due to the time-varying behavior of asset returns and the decreased trading costs. [1] In particular, the TRP has higher SR which in turn implies higher cumulative wealth as the corresponding volatilities are similar. Accordingly, this observation highlights the difference between the formulation of subportfolios between the TRP and SSR: individual assets with strong historical performances can be selected repeatedly in TRP while only being selected at most a fixed proportion for a given size of the subsets in SSR.

5.5 Conclusion

In this chapter we propose a target-based regularization method for portfolio selection by combining ensemble methods with regularization. The proposed strategy addresses the stability concerns of regularization methods by targeting

[1] Since the termination of fixed commissions in May 1975.

subsets of less correlated assets without the sacrifice of diversification benefits common to ensemble methods while retaining the advantages each approach offers. We conduct extensive experiments on various widely used datasets and provide comparison studies to demonstrate the robust and strong performance of the proposed strategy compared to related extant strategies. Our future research involves exploring different types of clustering such as classification methods using factor-based clustering as well as finding a data-driven way to choose the perturbation parameter α.

Chapter 6

Conclusions

6.1 Conclusions of the thesis

The rapid growth of data accessibility in economics and finance urges the implementation of state-of-the-art analytics tools. With this in mind, the thesis aims to investigate three important topics, namely, the determinants of loan recovery rates, uncertainties about the functional forms of evolving dynamics in macroeconomic forecasting and portfolio selection with risk minimization. While exploring these critical empirically motivated problems, the thesis also contributes to methodologies for modeling and inference when dealing with mixed frequency data or data which consists of a large set of covariates.

Following an introductory chapter, Chapter 2 reviews the fundamentals of Bayesian econometrics, including the role of the prior distribution to produce posterior inference. Technical aspects such as simulation-based MCMC techniques required to undertake in non-standard setting are covered, with a particular focus given to filtering-based methods that are suitable for state space models. In particular, a detailed discussion regarding MCMC sampling methods for Gaussian mixture and Markov switching models are provided, along with a discussion of modern priors such as the Bayesian LASSO and Gaussian process priors. These techniques are all employed in the thesis as various empirical problems are explored.

The first main contribution of the thesis is made in Chapter 3. A Gaussian mixture model is proposed to accommodate the observed clustering behavior found in bank loan recovery rates. A latent ordered probit structure is used to connect the mixture component to a large set of potential recovery determinants discussed in the reference, e.g., Khieu, Mullineaux, and Yi (2012). Owing to the apparent time-varying behavior of recovery rates thereof to vary according different states of the economy, a Markov switching mechanism is introduced to index the latent ordered probit regression coefficients to differentiate between "good" time and "bad". In addition, the so-called Bayesian LASSO prior is used to help selecting an appropriate subset of determinants from the large universe of available, but often correlated predictors. We use the proposed method and the developed Bayesian methodology to investigate recovery determinants for defaulted large U.S. bank loans using a dataset extracted from Moody's Ultimate Recovery Database. It is found that the behavior of some recovery rate determinants show very different relevance to recovery outcomes depending on whether the market is in a downward or in a more expansionary state. The result of the empirical study highlights the importance of accounting for countercyclical expected recovery rates when determining required capital buffers, especially when the risk of a market downturn is non-negligible.

A second main contribution of the thesis is contained in Chapter 4, where a new Bayesian method for estimating a VAR model in the presence mixed-frequency data is proposed. In this setting, certain variables are observed only at low frequency (e.g. quarterly) and are modeled as having corresponding have higher frequency (e.g. monthly) values that are "missing" from the observation set. The model is cast into a state-space representation and is then augmented with the "missing" observations. Importantly, a non-linear dependence structure for the latent observation is proposed, through the use of a Gaussian process prior. Exploiting the existing literature on nonlinear filtering techniques for state-space models, a filtering scheme for the proposed VAR model is developed.

Simulation study illustrates the usefulness of the proposed MCMC scheme in terms of characterizing both the uncertainties associated with the nonlinear function and the "missing" variables associated with the low-frequency data.

The third main contribution of the thesis is presented in Chapter 5. Here a novel ensemble-based approach for portfolio selection under the mean-variance framework is proposed, aiming to provide stable out-of-sample performance in the presence of estimation error. More precisely, we first construct regularization targets using subsets of assets with controlled maximum correlation by exploiting a hierarchical clustering algorithm. These subsets are used as regularization targets in the construction of subportfolios which are themselves averaged to stabilize the final portfolio weight. In a Monte Carlo simulation setup, it is shown that the proposed approach delivers promising improvements over the gross exposure constrained portfolios proposed by Fan, Zhang, and Yu (2012). Using four benchmark datasets, we show that the resulting portfolio selection strategy compares favorably against competing state-of-the-art strategies when evaluated using a range of portfolio performance evaluation criteria.

6.2 Future work

A number of future research directions are apparent following the research detailed in the thesis.

First, it is of interest to extend the finite mixture model employed in Chapter 3 to an infinite mixture model so that the assumption on the number of mixture components can be relaxed. Due to the clustering behavior of recovery rates and the presence of a large set of recovery determinants, a clustering-based nonparametric model along with a single index structure would be desirable to retain flexibility in capturing the complex distributional shape while at the same time mitigating the curse of dimensionality.

Another direction of research is to incorporate the sparse Gaussian process of Snelson and Ghahramani (2006) to improve on the computationally efficiency of the MCMC algorithm proposed in Chapter 4. The current Gaussian process involves $\mathcal{O}(n^3)$ operations, where n is the number of observations, with the complexity driven mainly due to the inversion of the covariance matrix required to calculate the Gaussian process posterior distribution inside each MCMC iteration. Since the posterior distribution is evaluated repeatedly within the MCMC scheme, the approach will become infeasible when the number of time periods becomes large. A plausible solution would be to use the sparse Gaussian process as a proposal within a pseudo-marginal MCMC framework, ultimately corrected via an additional MH step.

Finally, the theoretical ground of the proposed method in Chapter 5 could be further investigated. For instance, the associated predictive squared error loss of the proposed method could be analytically studied in relation to the correlations of the covariates. In light of the theory developed in Meinshausen and Bühlmann (2010) for variable selection with strongly correlated covariates, it would be of interest to investigate the prediction performance of the proposed approach, as it can be viewed as a type of randomized LASSO. Finally, the current methodology is computationally expensive since the tuning parameter λ has to be determined repeatedly for each and every subportfolio. It would be desirable to have a theoretically justified rule for choosing an universal λ to be used for all subportfolios, along with a data-driven method for tuning α.

Bibliography

Acharya, Viral V, Sreedhar T Bharath, and Anand Srinivasan (2007). “Does industry-wide distress affect defaulted firms? Evidence from creditor recoveries”. In: *Journal of Financial Economics* 85.3, pp. 787–821.

Albert, James H and Siddhartha Chib (1993). “Bayesian analysis of binary and polychotomous response data”. In: *Journal of the American Statistical Association* 88.422, pp. 669–679.

Altman, Edward, Andrea Resti, and Andrea Sironi (2004). “Default recovery rates in credit risk modelling: a review of the literature and empirical evidence”. In: *Economic Notes* 33.2, pp. 183–208.

Altman, Edward I and Egon A Kalotay (2014). “Ultimate recovery mixtures”. In: *Journal of Banking & Finance* 40, pp. 116–129.

Altman, Edward I and Vellore M Kishore (1996). “Almost everything you wanted to know about recoveries on defaulted bonds”. In: *Financial Analysts Journal* 52.6, pp. 57–64.

Altman, Edward I et al. (2005). “The link between default and recovery rates: theory, empirical evidence, and implications”. In: *Journal of Business* 78.6, p. 2203.

Anderson, Theodore Wilbur (1958). *An introduction to multivariate statistical analysis.* Vol. 2. Wiley New York.

Andreou, Elena, Eric Ghysels, and Andros Kourtellos (2010). “Regression models with mixed sampling frequencies”. In: *Journal of Econometrics* 158.2, pp. 246–261.

Andrieu, Christophe and Gareth Roberts (2009). "The pseudo-marginal approach for efficient Monte Carlo computations". In: *The Annals of Statistics* 37.2, pp. 697–725.

Araten, Michel, Michael Jacobs, and Peeyush Varshney (2004). "Measuring LGD on commercial loans: an 18-year internal study". In: *RMA JOURNAL* 86.8, pp. 96–103.

Bai, Jennie, Eric Ghysels, and Jonathan H Wright (2013). "State space models and MIDAS regressions". In: *Econometric Reviews* 32.7, pp. 779–813.

Barbieri, Maria Maddalena and James O Berger (2004). "Optimal predictive model selection". In: *Annals of Statistics*, pp. 870–897.

Bastos, João A (2010). "Forecasting bank loans loss-given-default". In: *Journal of Banking & Finance* 34.10, pp. 2510–2517.

Beaumont, Mark A (2003). "Estimation of population growth or decline in genetically monitored populations". In: *Genetics* 164.3, pp. 1139–1160.

Berger, James O (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Berger, James O, José M Bernardo, and Dongchu Sun (2009). "The formal definition of reference priors". In: *The Annals of Statistics* 37.2, pp. 905–938.

Bollerslev, Tim (1986). "Generalized autoregressive conditional heteroskedasticity". In: *Journal of econometrics* 31.3, pp. 307–327.

Boyd, Stephen et al. (2011). "Distributed optimization and statistical learning via the alternating direction method of multipliers". In: *Foundations and Trends® in Machine learning* 3.1, pp. 1–122.

Brandt, Michael W (2009). "Portfolio choice problems". In: *Handbook of financial econometrics* 1, pp. 269–336.

Breiman, Leo (1996). "Bagging predictors". In: *Machine learning* 24.2, pp. 123–140.

Bruche, Max and Carlos Gonzalez-Aguado (2010). “Recovery rates, default probabilities, and the credit cycle”. In: *Journal of Banking & Finance* 34.4, pp. 754–764.

Bühlmann, Peter et al. (2013). “Correlated variables in regression: clustering and sparse estimation”. In: *Journal of Statistical Planning and Inference* 143.11, pp. 1835–1858.

Carter, Chris K and Robert Kohn (1994). “On Gibbs sampling for state space models”. In: *Biometrika* 81.3, pp. 541–553.

Carvalho, Carlos M and Mike West (2007). “Dynamic matrix-variate graphical models”. In: *Bayesian analysis* 2.1, pp. 69–97.

Casella, George and Edward I George (1992). “Explaining the Gibbs sampler”. In: *The American Statistician* 46.3, pp. 167–174.

Castle, Karen Van de and David Keisman (1999). “Recovering your money: Insights into losses from defaults”. In: *Standard & Poor’s Credit Week* 16, p. 1999.

Chen, Cathy WS and Mike KP So (2006). “On a threshold heteroscedastic model”. In: *International Journal of Forecasting* 22.1, pp. 73–89.

Chen, Cathy WS, Mike KP So, and Edward MH Lin (2009). “Volatility forecasting with double Markov switching GARCH models”. In: *Journal of Forecasting* 28.8, pp. 681–697.

Chib, Siddhartha (1995). “Marginal likelihood from the Gibbs output”. In: *Journal of the american statistical association* 90.432, pp. 1313–1321.

Chib, Siddhartha and Edward Greenberg (2007). “Semiparametric modeling and estimation of instrumental variable models”. In: *Journal of Computational and Graphical Statistics* 16.1, pp. 86–114.

Chib, Siddhartha, Edward Greenberg, and Ivan Jeliazkov (2009). “Estimation of semiparametric models in the presence of endogeneity and sample selection”. In: *Journal of Computational and Graphical Statistics* 18.2, pp. 321–348.

Covitz, Daniel M, Song Han, and Beth Anne Wilson (2006). "Are longer bankruptcies really more costly?" In:

Dalalyan, Arnak S, Mohamed Hebiri, and Johannes Lederer (2017). "On the prediction performance of the lasso". In: *Bernoulli* 23.1, pp. 552–581.

Damianou, Andreas and Neil Lawrence (2013). "Deep gaussian processes". In: *Artificial Intelligence and Statistics*, pp. 207–215.

De Servigny, Arnaud and Olivier Renault (2004). *Measuring and managing credit risk*. McGraw Hill Professional.

DeMiguel, Victor et al. (2009). "A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms". In: *Management Science* 55.5, pp. 798–812.

Denison, David GT et al. (2002). *Bayesian methods for nonlinear classification and regression*. Vol. 386. John Wiley & Sons.

Dermine, Jean and C Neto De Carvalho (2006). "Bank loan losses-given-default: A case study". In: *Journal of Banking & Finance* 30.4, pp. 1219–1243.

Diebold, Francis X (1998). *Elements of forecasting*. Citeseer.

Durbin, James and Siem Jan Koopman (2012). *Time series analysis by state space methods*. Vol. 38. Oxford University Press.

Efron, Bradley et al. (2004). "Least angle regression". In: *The Annals of statistics* 32.2, pp. 407–499.

Elliott, Graham, Antonio Gargano, and Allan Timmermann (2013). "Complete subset regressions". In: *Journal of Econometrics* 177.2, pp. 357–373.

Emery, Kenneth (2007). *Moody's Ultimate Recovery Database, Moody's*.

Engle, Robert F (1982). "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". In: *Econometrica: Journal of the Econometric Society*, pp. 987–1007.

Eo, Yunjong and Chang-Jin Kim (2016). "Markov-switching models with evolving regime-specific parameters: Are postwar booms or recessions all alike?" In: *Review of Economics and Statistics* 98.5, pp. 940–949.

Eraker, Bjørn, Michael Johannes, and Nicholas Polson (2003). “The impact of jumps in volatility and returns”. In: *The Journal of Finance* 58.3, pp. 1269–1300.

Eraker, Bjørn et al. (2014). “Bayesian mixed frequency VARs”. In: *Journal of Financial Econometrics* 13.3, pp. 698–721.

Fama, Eugene F and Kenneth R French (1992). “The cross-section of expected stock returns”. In: *The Journal of Finance* 47.2, pp. 427–465.

Fan, Jianqing, Jingjin Zhang, and Ke Yu (2012). “Vast portfolio selection with gross-exposure constraints”. In: *Journal of the American Statistical Association* 107.498, pp. 592–606.

Financial Crisis Inquiry Commission (2011). *The financial crisis inquiry report: Final report of the national commission on the causes of the financial and economic crisis in the United States.* Washington DC: US Government Printing Office.

Franks, Julian R and Walter N Torous (1994). “A comparison of financial recontracting in distressed exchanges and Chapter 11 reorganizations”. In: *Journal of financial economics* 35.3, pp. 349–370.

Frigola, Roger et al. (2013). “Bayesian inference and learning in Gaussian process state-space models with particle MCMC”. In: *Advances in Neural Information Processing Systems*, pp. 3156–3164.

Frühwirth-Schnatter, Sylvia (1994). “Data augmentation and dynamic linear models”. In: *Journal of time series analysis* 15.2, pp. 183–202.

— (2006). *Finite mixture and Markov switching models.* Springer Science & Business Media.

Gaines, Brian R, Juhyun Kim, and Hua Zhou (2018). “Algorithms for fitting the constrained lasso”. In: *Journal of Computational and Graphical Statistics* just-accepted.

Gelfand, Alan E and Adrian FM Smith (1990). "Sampling-based approaches to calculating marginal densities". In: *Journal of the American statistical association* 85.410, pp. 398–409.

Gelman, Andrew, Gareth O Roberts, Walter R Gilks, et al. (1996). "Efficient Metropolis jumping rules". In: *Bayesian statistics* 5.599-608, p. 42.

Geman, Stuart and Donald Geman (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". In: *IEEE Transactions on pattern analysis and machine intelligence* 6, pp. 721–741.

Ghysels, Eric (2016). "Macroeconomics and the reality of mixed frequency data". In: *Journal of Econometrics* 193.2, pp. 294–314.

Ghysels, Eric, Pedro Santa-Clara, and Rossen Valkanov (2004). "The MIDAS touch: Mixed data sampling regression models". In:

— (2005). "There is a risk-return trade-off after all". In: *Journal of Financial Economics* 76.3, pp. 509–548.

— (2006). "Predicting volatility: getting the most out of return data sampled at different frequencies". In: *Journal of Econometrics* 131.1, pp. 59–95.

Gordon, Neil J, David J Salmond, and Adrian FM Smith (1993). "Novel approach to nonlinear/non-Gaussian Bayesian state estimation". In: *IEE Proceedings F (Radar and Signal Processing)*. Vol. 140. 2. IET, pp. 107–113.

Grinblatt, Mark, Sheridan Titman, and Russ Wermers (1995). "Momentum investment strategies, portfolio performance, and herding: A study of mutual fund behavior". In: *The American economic review*, pp. 1088–1105.

Gupton, Greg M et al. (2002). "LossCalcTM: Model for predicting loss given default (LGD)". In: *Moody's KMV, New York*.

Guyon, Isabelle and André Elisseeff (2003). "An introduction to variable and feature selection". In: *Journal of machine learning research* 3.Mar, pp. 1157–1182.

Hamilton, James D (1989). “A new approach to the economic analysis of nonstationary time series and the business cycle”. In: *Econometrica: Journal of the Econometric Society* 57.2, pp. 357–384.

— (2003). “What is an oil shock?” In: *Journal of econometrics* 113.2, pp. 363–398.

Härdle, Wolfgang, Hua Liang, and Jiti Gao (2012). *Partially linear models.* Springer Science & Business Media.

Harvey, Andrew C and Richard G Pierse (1984). “Estimating missing observations in economic time series”. In: *Journal of the American Statistical Association* 79.385, pp. 125–131.

Hastie, Trevor and Robert Tibshirani (1986). “Generalized Additive Models”. In: *Statistical Science* 1.3, pp. 297–310.

Hastings, WK (1970). “Monte Carlo sampling methods using Markov chains and their applications”. In: *Biometrika* 57, p. 97.

Ho, Tin K (1998). “The random subspace method for constructing decision forests”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.8, pp. 832–844.

Hoerl, Arthur E and Robert W Kennard (1970). “Ridge regression: Biased estimation for nonorthogonal problems”. In: *Technometrics* 12.1, pp. 55–67.

Hu, Xiaohua (2001). “Using rough sets theory and database operations to construct a good ensemble of classifiers for data mining applications”. In: *icdm.* IEEE, p. 233.

Hu, Yen-Ting and William Perraudin (2002). “The dependence of recovery rates and defaults”. In: *Birbeck College and Bank of England, Working Paper.*

Huang, Junzhou, Tong Zhang, and Dimitris Metaxas (2011). “Learning with structured sparsity”. In: *Journal of Machine Learning Research* 12.Nov, pp. 3371–3412.

Ichimura, Hidehiko (1993). “Semiparametric least squares (SLS) and weighted SLS estimation of single-index models”. In: *Journal of Econometrics* 58.1-2, pp. 71–120.

Jacquier, Eric, Nicholas G Polson, and Peter E Rossi (2002). “Bayesian analysis of stochastic volatility models”. In: *Journal of Business & Economic Statistics* 20.1, pp. 69–87.

Jagannathan, Ravi and Tongshu Ma (2003). “Risk reduction in large portfolios: Why imposing the wrong constraints helps”. In: *The Journal of Finance* 58.4, pp. 1651–1683.

Kalman, Rudolph Emil (1960). “A new approach to linear filtering and prediction problems”. In: *Journal of Fluids Engineering* 82.1, pp. 35–45.

Khieu, Hinh D, Donald J Mullineaux, and Ha-Chin Yi (2012). “The determinants of bank loan recovery rates”. In: *Journal of Banking & Finance* 36.4, pp. 923–933.

Kim, Chang-Jin and Charles R Nelson (1999a). “Has the US economy become more stable? A Bayesian approach based on a Markov-switching model of the business cycle”. In: *Review of Economics and Statistics* 81.4, pp. 608–616.

— (1999b). *State-space models with regime switching: classical and Gibbs-sampling approaches with applications.* Vol. 2. MIT press Cambridge, MA.

— (2001). “A Bayesian approach to testing for Markov-switching in univariate and dynamic factor models”. In: *International Economic Review* 42.4, pp. 989–1013.

Kim, Dong Heon, Denise R Osborn, and Marianne Sensier (2005). “Nonlinearity in the Fed’s monetary policy rule”. In: *Journal of applied Econometrics* 20.5, pp. 621–639.

Kim, Hyoung-Moon, Bani K Mallick, and CC Holmes (2005). “Analyzing nonstationary spatial data using piecewise Gaussian processes”. In: *Journal of the American Statistical Association* 100.470, pp. 653–668.

Kim, Sangjoon, Neil Shephard, and Siddhartha Chib (1998). “Stochastic volatility: likelihood inference and comparison with ARCH models”. In: *The review of economic studies* 65.3, pp. 361–393.

Kleiner, Ariel et al. (2014). “A scalable bootstrap for massive data”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.4, pp. 795–816.

Ko, Jonathan and Dieter Fox (2011). “Learning GP-BayesFilters via Gaussian process latent variable models”. In: *Autonomous Robots* 30.1, pp. 3–23.

Koop, Gary, Dale J Poirier, and Justin L Tobias (2007). *Bayesian econometric methods.* Cambridge University Press.

Lawrence, Neil (2005). “Probabilistic non-linear principal component analysis with Gaussian process latent variable models”. In: *Journal of machine learning research* 6.Nov, pp. 1783–1816.

Lawrence, Neil D (2004). “Gaussian process latent variable models for visualisation of high dimensional data”. In: *Advances in neural information processing systems*, pp. 329–336.

Ledoit, Oliver and Michael Wolf (2008). “Robust performance hypothesis testing with the Sharpe ratio”. In: *Journal of Empirical Finance* 15.5, pp. 850–859.

Ledoit, Olivier and Michael Wolf (2003). “Improved estimation of the covariance matrix of stock returns with an application to portfolio selection”. In: *Journal of empirical finance* 10.5, pp. 603–621.

— (2004). “A well-conditioned estimator for large-dimensional covariance matrices”. In: *Journal of Multivariate Analysis* 88.2, pp. 365–411.

Li, Bin, Steven CH Hoi, and Vivekanand Gopalkrishnan (2011). “Corn: Correlation-driven nonparametric learning approach for portfolio selection”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.3, p. 21.

Li, Bin et al. (2012). “PAMR: Passive aggressive mean reversion strategy for portfolio selection”. In: *Machine learning* 87.2, pp. 221–258.

Magdon-Ismail, M and A Atiya (2004). *Maximum drawdown. Risk Magazine.*

Mandelbrot, B. B. (1963). "The variation of certain speculative prices". In: *The Journal of Business* 36.4, pp. 394–419.

Manski, Charles F (1988). "Identification of binary response models". In: *Journal of the American statistical Association* 83.403, pp. 729–738.

Mariano, Roberto S and Yasutomo Murasawa (2003). "A new coincident index of business cycles based on monthly and quarterly series". In: *Journal of applied Econometrics* 18.4, pp. 427–443.

— (2010). "A coincident index, common factors, and monthly real GDP". In: *Oxford Bulletin of Economics and Statistics* 72.1, pp. 27–46.

Markowitz, Harry (1952). "Portfolio selection". In: *The journal of finance* 7.1, pp. 77–91.

McConnell, John J, Ronald C Lease, and Elizabeth Tashjian (1996). "Prepacks as a mechanism for resolving financial distress: The evidence". In: *Journal of Applied Corporate Finance* 8.4, pp. 99–106.

Meinshausen, Nicolai and Peter Bühlmann (2010). "Stability selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4, pp. 417–473.

Merton, Robert C (1980). "On estimating the expected return on the market: An exploratory investigation". In: *Journal of Financial Economics* 8.4, pp. 323–361.

Metropolis, Nicholas et al. (1953). "Equation of state calculations by fast computing machines". In: *The journal of chemical physics* 21.6, pp. 1087–1092.

Michaud, Richard O (1989). "The Markowitz optimization enigma: Is â˜optimizedâ™ optimal?" In: *Financial Analysts Journal* 45.1, pp. 31–42.

Mitchell, Tom M (1980). *The need for biases in learning generalizations.* Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ. New Jersey.

Mora, Nada (2015). "Creditor recovery: The macroeconomic dependence of industry equilibrium". In: *Journal of Financial Stability* 18, pp. 172–186.

Nazemi, Abdolreza and Frank J Fabozzi (2018). "Macroeconomic Variable Selection for Creditor Recovery Rates". In: *Journal of Banking & Finance* 89, pp. 14–25.

O'Hagan, Anthony and JFC Kingman (1978). "Curve fitting and optimal design for prediction". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–42.

Park, Trevor and George Casella (2008). "The Bayesian LASSO". In: *Journal of the American Statistical Association* 103.482, pp. 681–686.

Pettenuzzo, Davide, Allan Timmermann, and Rossen Valkanov (2016). "A MIDAS approach to modeling first and second moment dynamics". In: *Journal of Econometrics* 193.2, pp. 315–334.

Pitt, Michael K et al. (2012). "On some properties of Markov chain Monte Carlo simulation methods based on the particle filter". In: *Journal of Econometrics* 171.2, pp. 134–151.

Polikar, Robi (2006). "Ensemble based systems in decision making". In: *IEEE Circuits and systems magazine* 6.3, pp. 21–45.

Qi, Min and Xinlei Zhao (2011). "Comparison of modeling methods for loss given default". In: *Journal of Banking & Finance* 35.11, pp. 2842–2855.

Rapach, David and Guofu Zhou (2013). "Forecasting stock returns". In: *Handbook of economic forecasting.* Vol. 2. Elsevier, pp. 328–383.

Rasmussen, Carl Edward and Christopher KI Williams (2006). *Gaussian process for machine learning.* MIT press.

Resti, Andrea (2002). *The New Basel Capital Accord: Structure Possible Changes and Micro-and Macroeconomic Effects.* 30. Ceps.

Robert, Christian and George Casella (2013). *Monte Carlo statistical methods.* Springer Science & Business Media.

Robinson, Peter M (1988). "Root-N-consistent semiparametric regression". In: *Econometrica: Journal of the Econometric Society*, pp. 931–954.

Rokach, Lior (2010). “Ensemble-based classifiers”. In: *Artificial Intelligence Review* 33.1-2, pp. 1–39.

Ryan, Stephen G (2008). “Fair value accounting: Understanding the issues raised by the credit crunch”. In: *Council of Institutional Investors* July, 2008, pp. 1–24.

Schuermann, Til (2004). “What do we know about Loss Given Default?” In: *In: Shimko, D. (Ed.), Credit Risk Models and Management. London, UK.*

Shen, Weiwei and Jun Wang (2017). “Portfolio Selection via Subset Resampling.” In: *AAAI*, pp. 1517–1523.

Shen, Weiwei, Jun Wang, and Shiqian Ma (2014). “Doubly Regularized Portfolio with Risk Minimization.” In: *AAAI*, pp. 1286–1292.

Shively, Thomas S, Robert Kohn, and Sally Wood (1999). “Variable selection and function estimation in additive nonparametric regression using a data-based prior”. In: *Journal of the American Statistical Association* 94.447, pp. 777–794.

Silverman, Bernhard W (1985). “Some aspects of the spline smoothing approach to non-parametric regression curve fitting”. In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–52.

Sims, Christopher A (1980). “Macroeconomics and reality”. In: *Econometrica: Journal of the Econometric Society*, pp. 1–48.

Smith, Michael and Robert Kohn (1996). “Nonparametric regression using Bayesian variable selection”. In: *Journal of Econometrics* 75.2, pp. 317–343.

Snelson, Edward and Zoubin Ghahramani (2006). “Sparse Gaussian processes using pseudo-inputs”. In: *Advances in neural information processing systems*, pp. 1257–1264.

Stroud, Jonathan R, Peter Müller, and Nicholas G Polson (2003). “Nonlinear state-space models with state-dependent variances”. In: *Journal of the American Statistical Association* 98.462, pp. 377–386.

Surico, Paolo (2007). "The Fed's monetary policy rule and US inflation: The case of asymmetric preferences". In: *Journal of Economic Dynamics and Control* 31.1, pp. 305–324.

Tanner, Martin A and Wing Hung Wong (1987). "The calculation of posterior distributions by data augmentation". In: *Journal of the American statistical Association* 82.398, pp. 528–540.

Temlyakov, Vladimir N (2000). "Weak greedy algorithms". In: *Advances in Computational Mathematics* 12.2-3, pp. 213–227.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Titsias, Michalis and Neil D Lawrence (2010). "Bayesian Gaussian process latent variable model". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 844–851.

Tu, Jun and Guofu Zhou (2011). "Markowitz meets Talmud: A combination of sophisticated and naive diversification strategies". In: *Journal of Financial Economics* 99.1, pp. 204–215.

Zhou, Zhi-Hua (2012). *Ensemble Methods: Foundations and Algorithms.* CRC Press.

Zou, Hui (2006). "The adaptive lasso and its oracle properties". In: *Journal of the American statistical association* 101.476, pp. 1418–1429.