SOME ADVANCES IN BAYESIAN REGRESSION PROBLEMS

A THESIS SUBMITTED TO

THE UNIVERSITY OF KENT AT CANTERBURY

IN THE SUBJECT OF STATISTICS

FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY BY RESEARCH

By

Sotirios Prevenas

October 2023

Word count: 26022

Some Advances in Bayesian Regression Problems

Sotirios Prevenas

Abstract

In the first part of the thesis, we propose a novel objective Bayesian method with an application in vector autoregressive models when the normal-Wishart prior, with ν degrees of freedom, is considered. In particular, we depart from the current approach of setting $\nu = m + 1$ in the Wishart prior of an $m-$dimensional covariance matrix by setting a loss-based prior on ν. By doing so, we have been able to exploit any information about ν in the data and achieve better predictive performance than the method currently used in the literature. We show how this works well on both simulated and real data sets, where, in the latter case, we used data of macroeconometric nature as well as viral data. In addition, we explain why we believe we achieve better performance by showing that the data appear to suggest a value of ν far from the canonical $m + 1$ value.

In the second part of the thesis, we introduce a variational Bayesian inference algorithm that performs variable selection in linear regression models. The algorithm is based on a combination of a spike-and-slab prior with a normal-gamma prior on the coefficients. The spike-and-slab prior is widely considered the "gold standard" for sparse Bayesian problems, and the normal-gamma prior is a hierarchical shrinkage prior that generalises the Bayesian LASSO. The algorithm also combines two types of updates, one for the initial iterations and one for the later ones. This hybrid updating scheme results in both high accuracy and speed of convergence. Simulations and real data examples demonstrate the competitiveness of the algorithm against recent variable selection methods. We also modify our method to perform variable selection on two types of generalised linear models.

Covid-19 Impact Statement

Dear Examiners,

The events of March 2020, i.e. the implementation of lockdown and work-from-home orders issued by the UK Government in response to the COVID-19 pandemic outbreak, occurred when I was in the second year of my PhD studies. During the same period, another mishap occurred: my main supervisor, whom I was greatly influenced by and considered a mentor, stopped working for the University of Kent. Although this would have happened regardless of the pandemic, the two events had a combined effect on my progress. I found myself in a situation where I lacked a proper workplace (access to a desktop computer and printer is vital for me) while finishing up a research project and searching for a new one. Eventually, I started working with a new supervisor in September 2020, but it was not until spring 2021, a year into the pandemic, that I started doing actual research on a new topic. This delay was due to the need to adjust my interests to new areas of Statistics, very different from those I had been dealing with during the first two years of my studies. This adjustment had to be made during a period of strict measures, which included working from home—something I found very difficult to cope with. Inevitably, the slow progress had a detrimental psychological impact on me and also created friction between me and my supervisors. At the end of my third year, I relocated to my private residence outside the UK so I could at least have my own proper workplace without the fear of not being allowed to use it.

After some time, I took a few months of intermission to counter the accumulated stress that had been inflicted on me.

All things considered, I believe that a combination of unfortunate events prevented me from meeting my initial expectations in terms of quantity. However, I did everything in my power to ensure that the quality of the thesis meets your high standards.

Declaration

The work in this thesis is based on research completed at the School of Mathematics, Statistics and Actuarial Science at the University of Kent. This thesis, nor any part of it, has been submitted elsewhere for any other degree or qualification.

Acknowledgements

I would like to thank Prof. Rachel McCrea and Dr. Cristiano Villa for being patient, supportive, and motivating, and for giving me the opportunity to study at the University of Kent. I also thank Prof. Jian Zhang for his guidance and support in the latter part of my studies and in the finalisation of my thesis. Special thanks go to Claire Carter for her assistance as a student officer and for her kindness towards me.

I consider myself lucky to have made a few friends in Canterbury, to whom I am very grateful. Dimitris, Nikitas, Guillerme, and Oz, thank you for withstanding my whining during the pandemic period. My studies also gave me the opportunity to come closer to my relatives Tanya, Manolis, Sofia, and Elli, who are permanent UK residents.

Finally, I would like to thank EPSRC for funding my PhD studies.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Motivation

Hierarchical models are an essential element of Bayesian statistics. Model parameters are assigned prior distributions parametrised by hyperparameters. The hierarchy can be extended by assigning priors to the hyperparameters and so forth. The hyperparameters at the end of the hierarchy can either be specified using subjective information or estimated from the data to allow for a more objective approach to our analysis. In Bayesian regression, hierarchical models are very flexible tools, as different choices of priors and their hyperparameters lead to different marginal priors imposed on model parameters. This can be very useful, depending on our goal; for instance, we might want to induce shrinkage on weak coefficients, and a prior with density concentrated at the origin and heavy tails is desired.

The motivation behind this thesis stems from contemporary statistical challenges, such as estimating and forecasting multivariate time series models in macroeconomic studies or finding the smallest set of features that can accurately predict an outcome in a regression problem, when both the number of true predictors and the sample size are far less than the total number of available predictors.

Since we approach statistics from a Bayesian point of view, we are particularly interested in examining how the structure and parametrisation of hierarchical models influence performance, with an emphasis on improving prediction accuracy, inference speed, and objectivity.

Therefore, with our thesis, we attempt to complement statistical literature in two ways. Firstly, we present a new prior for a hyperparameter of a very popular hierarchical prior that has so far been given a fixed value. Secondly, we demonstrate how a combination of hierarchical priors can help us construct an efficient algorithm for performing Bayesian variable selection in regression.

1.2 General overview

This thesis makes contributions to two quite different areas of Bayesian statistics. We begin with the construction of a novel prior for the degrees of freedom of the Wishart distribution. Until this thesis, there was no other treatment of this parameter in the related literature other than setting it equal to the dimension of the matrix being modelled by the Wishart distribution plus one. Since it is uncommon for a practitioner to have any prior beliefs regarding the degrees of freedom, we follow existing methodology for constructing objective prior distributions, and the resultant prior has no tunable parameters. Our work draws motivation from Bayesian vector autoregressive (BVAR)[1] models, where the Wishart distribution is used as a prior for the precision matrix of the vector of errors, which is assumed to follow a zero-mean multivariate normal distribution. The degrees of freedom control the mean and variance of the Wishart distribution, and thus control the uncertainty around the precision matrix. We have found the standard practice of fixing the degrees of freedom to a specific value to be rather restricting when

[1] See (Koop et al., 2010) for an introduction on BVAR models.

analysing real datasets. For example, lengthy macroeconomic time series spanning several decades might contain periods with varying uncertainty around the errors, and the degrees of freedom should be allowed to vary. We support this position by using our prior to estimate the degrees of freedom of a BVAR model fit on a US macroeconomy dataset and monitor how our estimates change over the years using a rolling window analysis.

We then turn our focus to the widely addressed problem of variable selection in regression. In particular, we study problems in which there are many more predictors than observations ($D \gg N$), samples are correlated, and each predictor has a weak effect. Typical examples are gene expression and genetic association studies, where we might take measurements of tens of thousands of predictors while observations might be in the hundreds. These types of problems are becoming increasingly challenging in our era as more high-throughput measurement tools become available, allowing us to collect massive amounts of data. To this end, we propose a Bayesian model that is inferred by approximating the posterior distribution of its parameters using variational inference and test our method on highly demanding scenarios.

1.3 Organisation of the thesis

- **Chapter 2** contains theoretical background on topics related to this thesis. It provides a basic review of objective Bayesian methods for constructing prior distributions for both continuous and discrete parameters, an introduction to Bayesian variable selection through shrinkage and spike-and-slab priors, and variational inference. It also contains a literature review of methods that use spike-and-slab priors and are relevant and competitive with our work in Chapter 4.

- **Chapter 3** introduces a loss-based prior distribution for the degrees of freedom of the Wishart distribution. Necessary theoretical results about the construction of the prior are contained in the chapter. BVAR models are used as a testing ground, where our novel prior is used as a hyperprior in the independent normal-Wishart prior assigned to the coefficients. By assigning a prior to the degrees of freedom, we are able to learn this parameter from the data and make comparisons against a fixed value, which has been the popular treatment of this parameter so far.

- **Chapter 4** presents a variable selection method that is a direct antagonist to the methods reviewed in the last section of Chapter 2. It begins with the specification of a hierarchical prior for a Bayesian linear regression model, combining the spike-and-slab and the normal-gamma shrinkage prior (contrary to other methods that employ a normal or a double exponential slab distribution). The resulting model has three global parameters that govern all predictors and regulate overall sparsity and coefficient shrinkage. The contribution of Chapter 4 is a variational expectation-maximisation algorithm that approximates the posteriors of the local parameters using known distributions and estimates the three global parameters from the data. It does so by optimising a multi-parameter objective function. Our focus is on sparse $D \gg N$ scenarios. In such scenarios, as we explain in Section 4.3, the variational algorithm faces an increased risk of getting trapped in poor local optima[2], especially during its initial iterations. To reduce that risk, we devise a simple strategy. We propose an updating scheme that begins by optimising the objective function with respect to variational parameters using costly updates that contain matrix operations until a criterion is met, and then switches to faster updates until convergence. We demonstrate the

[2] For a discussion on the problem of variational methods being subject to poor local optima we refer the reader to Altosaar et al. (2017). Examples of this problem can also be found in Beal (2003) and in MacKay (2003).

efficiency of our algorithm in challenging simulation studies, where it outperforms three modern Bayesian variable selection methods. The algorithm is then modified to perform variable selection in two generalised linear models via gradient ascent optimisation of the variational objective function, and additional simulated and real data analysis is performed.

- **Chapter 5** concludes the thesis and presents some ideas on how the prior distribution presented in Chapter 3 can be useful to other models, how to extend the work in Chapter 4, and directions for further research on variable selection.

- **Appendix** contains additional simulation studies for Chapter 3 and derivation of expressions for both Chapter 3 and Chapter 4.

Chapter 2

Literature review

In this chapter, we review some very relevant methodologies to our work in the remainder of the thesis. We begin by reviewing literature on constructing objective Bayesian priors, which is the core of Chapter 3. Sufficient theoretical background on Bayesian vector autoregressive models is contained within Chapter 3. We then review basic theory on variable selection and variational inference, which are fundamental to our work in Chapter 4, as well as some very relevant and competitive methods to the one we propose.

2.1 Objective Bayesian methods

Traditionally, objective Bayesian analysis is performed with the use of objective prior distributions. Such distributions are designed to be minimally informative in some sense. Over the last few decades, several methodologies have been proposed on how to construct noninformative priors. We discuss some of them briefly.

The Jeffreys prior (Jeffreys, 1939) is defined in terms of the Fisher information matrix:

$$\pi(\theta) \propto |\mathcal{I}(\theta)|^{\frac{1}{2}},$$

where $\mathcal{I}(\theta)$ is the Fisher information matrix with entries:

$$\big[\mathcal{I}(\theta)\big]_{ij} = -E_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(X;\theta)\right].$$

Jeffreys priors work well for single-parameter models, but they can exhibit problematic behaviour in multi-parameter models.

Reference priors (Bernardo, 1979; Berger and Bernardo, 1992) are equivalent to Jeffreys priors for one-dimensional parameters, and they have been more successful for higher-dimensional problems. As mentioned in Berger et al. (2009), reference priors have been rigorously defined in specific contexts and heuristically defined in general, but a rigorous general definition has been missing from the statistical literature. To understand the idea of reference priors, one can think the construction of a reference prior as an optimisation problem. The prior is chosen in a way that maximises the expected Kullback-Leibler divergence[1] (Kullback and Leibler, 1951) between the posterior and the prior. We can define the optimisation problem as choosing the prior $\pi^*(\theta)$ that maximises the expected information:

$$\pi^*(\theta) = \arg\max_{\pi(\theta)} I(\theta, \mathbf{x}),$$

where $I(\theta, \mathbf{x})$ is the information to be expected from the observation vector $\mathbf{x}$:

$$\begin{aligned} I(\theta, \mathbf{x}) &= \int\int p(\theta|\mathbf{x})p(\mathbf{x})\log\frac{p(\theta|\mathbf{x})}{\pi(\theta)}d\mathbf{x}d\theta \\ &= \int p(\mathbf{x})\mathrm{KL}(p(\theta|\mathbf{x})||\pi(\theta))d\mathbf{x}. \end{aligned}$$

By maximising the expected KL divergence, the prior has a minimum impact while at the same time the data have a maximum effect on the posterior. For an explicit form of the reference prior, see Bernardo and Smith (1994) or Berger

[1] Other suitable divergences or distances can be chosen as well. We denote the Kullback-Leibler divergence by KL for the remaining of the thesis.

et al. (2009).

Another concept is the maximum entropy prior; information on this methodology can be found in Jaynes (2003). The idea behind the maximum entropy prior is that the amount of information contained in a probability distribution can be measured by the Shannon entropy (Shannon, 1948). The larger the entropy of a probability distribution is, the less information it contains and thus the less informative it is considered. Maximum entropy prescribes choosing the prior with the maximal entropy, among a set of possible priors. As an example, for a discrete parameter, we would choose the prior $\pi(\theta)$ that maximises $H(\theta) = -\sum_{\theta\in\Theta} \pi(\theta) \log \pi(\theta)$ subject to $\sum_{\theta\in\Theta} \pi(\theta) = 1$.

The interested reader can resort to Kass and Wasserman (1996) for a comprehensive review of methods for constructing noninformative priors for continuous parameter spaces.

Objective priors can also be constructed for discrete parameter spaces. Direct application of standard reference prior theory to a discrete parameter yields a constant (uniform) prior. To this end, Berger et al. (2012) proposed a methodology that embeds the discrete structure of the parameter space into a continuous one, and then applies the reference prior methodology to the continuous structure. They proposed four approaches for the embedding. In the first approach, the discrete parameter is simply treated as continuous and a reference prior is found. In the second approach, a hyperprior is introduced in the modelling hierarchy and a reference prior is found for its continuous hyperparameter. In the third approach, the asymptotic distribution of a consistent estimator of θ is considered, and then a reference prior is found for the parameter in the asymptotic distribution, treating it as continuous. In the fourth approach, a limiting operation is used in order to transform the problem into a continuous one, and then the ordinary reference prior methodology is applied. Not all approaches are suitable for all problems. The resulting continuous prior might be appropriately discretised if necessary.

Another approach, which is also the one that is adopted in this thesis, is presented in Villa and Walker (2015). A prior constructed via this method, assigns to each parameter $\theta \in \Theta$ a mass given by the formula:

$$\pi(\theta) \propto \exp\left\{\min_{\theta' \neq \theta \in \Theta} \mathrm{KL}\left(f(x|\theta) \| f(x|\theta')\right)\right\} - 1. \tag{2.1.1}$$

The term inside the exponential in (2.1.1) is the KL divergence measured from the model parametrised by θ, to its nearest model, parametrised by θ'. This divergence represents the utility or worth that each parameter θ has. If θ is the true parameter and is removed, the posterior distribution accumulates asymptotically at the nearest $\theta' \in \Theta$ in terms of KL divergence (Berk, 1966). Thus, we can quantify the loss in information we would incur if we remove θ while it is the true parameter. This method is suitable for the needs of our thesis, since the KL divergence for Wishart distributions is available in closed form. It has also been applied successfully on the degrees of freedom of a Student t distribution (Villa and Walker, 2014) as well as the multivariate t distribution and the t-copula (Villa and Rubio, 2018).

2.2 Penalised regression and its Bayesian version

In regression analysis, penalised regression is one of the standard tools in the practitioner's apparatus that is used to identify and estimate the effect of relevant variables for predicting an outcome in cases where there are many predictors and (possibly) few data. From a frequentist point of view, penalised regression usually amounts to the following optimisation problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta_0, \boldsymbol{\beta}} \left\{ \frac{1}{N} \|\boldsymbol{y} - \beta_0 \mathbf{1} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_q \right\}, \tag{2.2.1}$$

where $\|\boldsymbol{\beta}\|_q = \left(\sum_{j=1}^{D} |\beta_j|^q\right)^{\frac{1}{q}}$ is the penalisation term, $\boldsymbol{y} = (y_1, \ldots, y_N)^\top$ is an N-dimensional response vector, $\boldsymbol{X}$ is the design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_D)^\top$ is the vector of regression coefficients, and β_0 is the intercept. The parameter λ regulates the shrinkage level, with larger values causing more shrinkage towards zero, while $\lambda = 0$ results in ordinary least squares (OLS) estimation. The penalty term, governed by parameter q, decides what type of penalisation is imposed on $\boldsymbol{\beta}$. For example, $q = 1$ leads to the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) penalty, and $q = 2$ leads to the ridge regression penalty (Hoerl and Kennard, 1970). Other penalty terms include the elastic net (Zou and Hastie, 2005) penalty:

$$\lambda_2\|\boldsymbol{\beta}\|_2 + \lambda_1\|\boldsymbol{\beta}\|_1,$$

which is a combination of LASSO and ridge penalties; the adaptive LASSO (Zou, 2006):

$$\lambda \sum_{j=1}^{D} |\hat{\beta}_j^0| \cdot |\beta_j|,$$

where $\hat{\beta}_j^0$ is an initial estimate (e.g. OLS) and λ is a positive parameter; the group LASSO penalty (Yuan and Lin, 2006):

$$\sum_{g=1}^{G} \lambda_g \|\boldsymbol{\beta}_g\|_2,$$

which is useful when predictors are grouped in some way, for example in linear regression with categorical explanatory variables, where each one may have several levels and can be expressed through a group of dummy variables; or the fused LASSO penalty (Tibshirani et al., 2005):

$$\lambda_1 \sum_{j=1}^{D} |\beta_j| + \lambda_2 \sum_{j=2}^{D} |\beta_j - \beta_{j-1}|,$$

which is useful in problems with variables that can be ordered in some meaningful way, and we want neighbouring coefficients to be similar to each other, besides being sparse. The first penalty term promotes sparsity in the coefficients, and the second one promotes sparsity in their differences.

Most estimations obtained using (2.2.1) also have a Bayesian equivalent as the posterior mode under an appropriate prior on $\boldsymbol{\beta}$. For example, the $\hat{\boldsymbol{\beta}}$ obtained via LASSO penalisation is equivalent to the posterior mode under independent Laplace priors on each β_j, with a common hyperparameter. If we also condition on the error variance σ^2, the Laplace becomes the Bayesian LASSO prior (Park and Casella, 2008):

$$p(\boldsymbol{\beta}|\lambda, \sigma^2) = \prod_{j=1}^{D} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}}. \tag{2.2.2}$$

Conditioning on σ^2 guarantees a unimodal full posterior. Similarly, Kyung et al. (2010) proposed Bayesian equivalents to grouped and fused LASSO penalties, and Li and Lin (2010) proposed the Bayesian version of elastic net. These are typical examples of Bayesian penalisation techniques, and the priors they use are referred to as *shrinkage priors*, as they aim to shrink small effects towards zero. Many of these shrinkage priors are presented as scale mixtures of normal distributions (see, for example, Andrews and Mallows, 1974; West, 1987). In a normal scale mixture, a normal prior is assigned to β_j, i.e. $\beta_j \sim \mathcal{N}(0, \tau_j)$, and then a prior (or a hierarchy of priors) is assigned to the variance τ_j, which serves as the mixing density. The type of prior and its parameters govern the level of shrinkage and decide the shape of the marginal prior that is imposed on β_j.

There is an extensive literature on shrinkage priors, and we briefly present one example here. The hyperlasso (Griffin and Brown, 2007, 2011) is a Bayesian analogue to the adaptive LASSO, obtained through the following normal mixture:

$$\begin{aligned} \beta_j | \tau_j &\sim \mathcal{N}(0, \tau_j), \\ \tau_j | \gamma_j &\sim \text{Expon}(\gamma_j), \\ \gamma_j | \nu, \lambda &\sim \text{Ga}(\nu, \lambda^{-2}). \end{aligned} \tag{2.2.3}$$

This normal-exponential-gamma prior belongs to the class of global-local shrinkage priors (Polson and Scott, 2011), since it has a local variance component τ_j, which is specific to each predictor, and two global parameters, ν and λ, that are common for all predictors and control the heaviness of the tails and the scale of the prior. The density of the marginal distribution of β_j, the normal-exponential-gamma probability density function, can be expressed as

$$\pi(\beta_j) = \frac{\nu}{\sqrt{\pi}} 2^{\nu} \lambda \Gamma(\nu + 1/2) \exp\left(\frac{1}{4}\lambda^2 \beta_j^2\right) D_{-2(\nu+1/2)}(\lambda|\beta_j|),$$

where $D_{\nu}(x)$ is the parabolic cylinder function (Olver et al., 2010, Ch. 12). Figure 2.2.1 presents a graphical model representation of hyperlasso, along with contour plots of the penalty (negative log density of the joint prior for two variables) induced by hyperlasso. This penalty is an example of a non-convex penalty. This can be seen by connecting with a line two points in the contour — the line will lie outside of the contour.

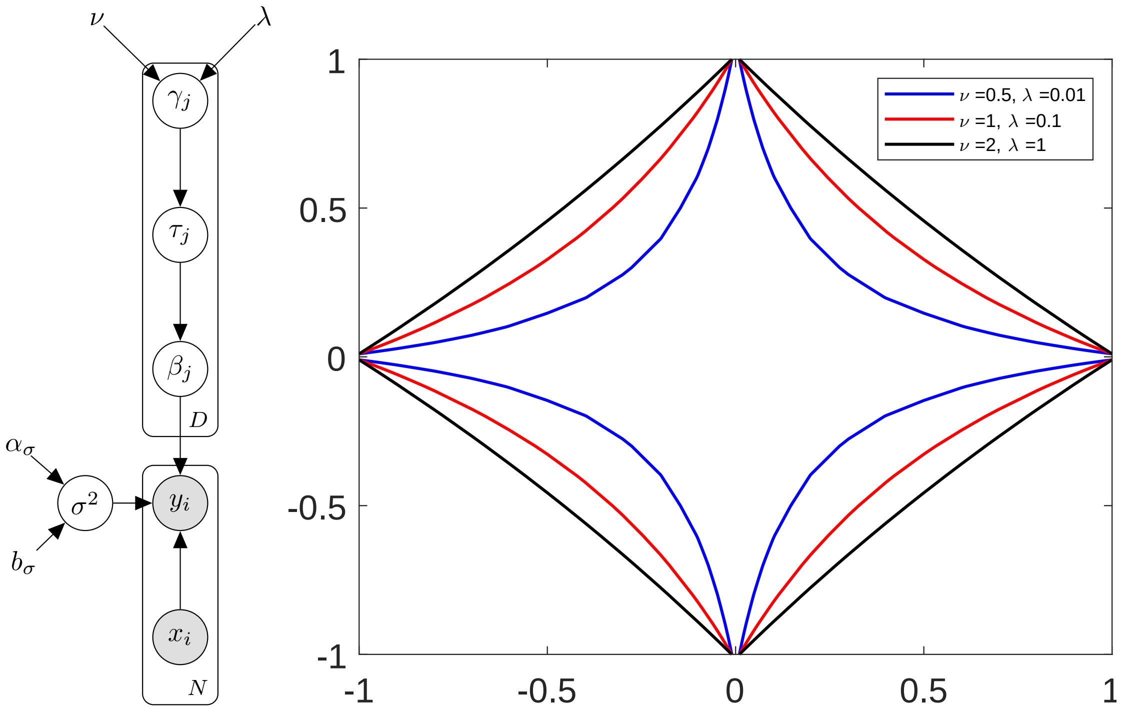
Another popular prior for $\boldsymbol{\beta}$ is the spike-and-slab prior (Mitchell and Beauchamp, 1988; George and McCulloch, 1993). We consider two types of spike-and-slab priors:

$$\beta_j | z_j, \theta_1 \sim (1 - z_j)\delta_0(\beta_j) + z_j p(\beta_j | \theta_1), \tag{2.2.4}$$

and

$$\beta_j | z_j, \theta_0, \theta_1 \sim (1 - z_j) p(\beta_j | \theta_0) + z_j p(\beta_j | \theta_1). \tag{2.2.5}$$

We refer to priors of the forms in (2.2.4) and (2.2.5) as discrete spike-and-slab



(a) (b)

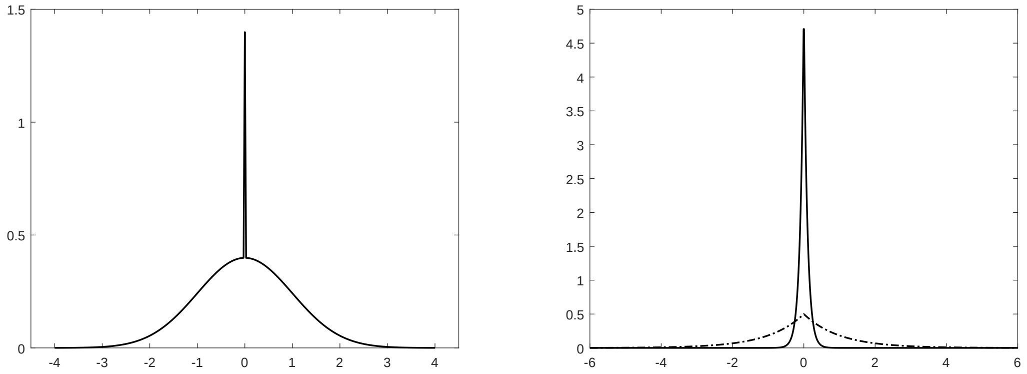
Figure 2.2.1: (a) Graphical model representation of hyperlasso. Notice that, contrary to the Bayesian LASSO prior in (2.2.2), the prior on β_j is not conditioned on the error variance σ^2. (b) Bivariate contour plots of the hyperlasso penalty (negative log-prior) for three combinations of ν and λ.

and continuous spike-and-slab priors, respectively. $\delta_0(\beta_j)$ is a point mass at zero, $p(\beta_j|\theta_1)$ is a continuous density serving as the slab distribution, and in (2.2.5), $p(\beta_j|\theta_0)$ is a continuous density highly concentrated about zero. Both formulations include the latent variable z_j, which is a binary indicator that takes the values

$$z_j = \begin{cases} 1 & \text{if predictor j is included in the model,} \\ 0 & \text{otherwise.} \end{cases}$$

The discrete spike-and-slab prior has the ability to shrink a coefficient exactly to zero and thus to effectively exclude a covariate from the model. A typical choice of a prior distribution on z_j is a Bernoulli distribution

$$p(z_j|\rho) = \rho^{z_j}(1-\rho)^{1-z_j}. \tag{2.2.6}$$



(a)

(b)

Figure 2.2.2: (a) A discrete spike-and-slab prior, with a normal slab and point mass at zero. (b) Two Laplace densities $p(\beta_j|\lambda) = \frac{\lambda}{2}e^{-\lambda|\beta_j|}$. Solid line ($\lambda = 10$) represents a spike distribution, while dashed line ($\lambda = 1$) represents a slab distribution in a continuous spike-and-slab prior.

Parameter ρ can be either fixed, given a hyperpior (e.g. a beta distribution), estimated empirically, or integrated out to simplify a Monte Carlo simulation.

Spike-and-slab priors offer great flexibility in the choices of the spike and the slab densities; Figure 2.2.2 shows combinations of normal and Laplace densities. This flexibility is not restricted to continuous symmetric densities. In the context of Bayesian nonparametrics (BNP), a Dirichlet process (Ferguson, 1973) or even a hierarchical Dirichlet process (Teh et al., 2006) can be used as a slab, along with the point mass at the spike. Also, again in the context of BNP, a spike-and-slab prior can be used to create a base measure for a Dirichlet process prior (Barcella et al., 2016, 2017).

Bayesian computation in the context of Bayesian regression with spike-and-slab priors can be conducted via Gibbs sampling algorithms (see, for example, George and McCulloch, 1997; Narisetty et al., 2019). However, such approaches do not scale well to large datasets with many variables, and we discuss some alternative methods in Section 2.4.

2.3 Overview of variational inference

The following review of basic variational inference theory is based on introductory readings such as Blei et al. (2017),Zhang et al. (2019),Ormerod and Wand (2010), as well as Chapter 10 of Bishop (2006).

2.3.1 A peek at variational methods

In Bayesian statistics, we are interested in computing the conditional (or posterior) density of latent variables, given observations. Let $\mathbf{x}$ be a set of observed variables, and $\mathbf{z}$ be a set of latent variables, which may include hidden variables and model parameters, with joint density $p(\mathbf{z}, \mathbf{x})$. Bayesian inference is then based on the posterior density:

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}.$$

The denominator $p(\mathbf{x}) = \int p(\mathbf{z}, \mathbf{x}) d\mathbf{z}$ is the marginal likelihood of the data and is also known as the evidence. This quantity is available in closed form only for simple models. When not analytically available, practitioners resort to approximate inference methods, such as Markov chain Monte Carlo (MCMC, see, for example, Robert et al. (2004)), Laplace approximation (Tierney et al., 1989; MacKay, 1992), or expectation-maximisation (EM, Dempster et al. (1977)). An alternative method is variational inference (VI), which transforms statistical inference into an optimisation problem. The idea is to approximate the intractable posterior with a simpler distribution of known form.

In VI, we select a variational distribution $q(\mathbf{z})$ parameterised by a set of variational parameters. The goal is to find the best set of variational parameters that make the Kullback-Leibler divergence between $q(\mathbf{z})$ and $p(\mathbf{z}|\mathbf{x})$ as small as possible. If we denote the variational parameters by $\boldsymbol{\lambda}$, then variational inference can be described as solving the optimisation problem:

$$q^*(\mathbf{z}) = \arg\min_{\boldsymbol{\lambda}} \mathrm{KL}(q(\mathbf{z}|\boldsymbol{\lambda})||p(\mathbf{z}|\mathbf{x})),$$

where $q^*(\mathbf{z})$ is the best approximation of $p(\mathbf{z}|\mathbf{x})$. We have that:

$$\begin{aligned}
\mathrm{KL}(q(\mathbf{z}|\boldsymbol{\lambda})||p(\mathbf{z}|\mathbf{x})) &= \int q(\mathbf{z}|\boldsymbol{\lambda}) \log \frac{q(\mathbf{z}|\boldsymbol{\lambda})}{p(\mathbf{z}|\mathbf{x})} d\mathbf{z} \\
&= \int q(\mathbf{z}|\boldsymbol{\lambda}) \log q(\mathbf{z}|\boldsymbol{\lambda}) d\mathbf{z} - \int q(\mathbf{z}|\boldsymbol{\lambda}) \log p(\mathbf{z}|\mathbf{x}) d\mathbf{z} \\
&= \int q(\mathbf{z}|\boldsymbol{\lambda}) \log q(\mathbf{z}|\boldsymbol{\lambda}) d\mathbf{z} - \int q(\mathbf{z}|\boldsymbol{\lambda}) \log p(\mathbf{z}, \mathbf{x}) d\mathbf{z} + \log p(\mathbf{x}).
\end{aligned} \tag{2.3.1}$$

The intractable log-evidence term $\log p(\mathbf{x})$ in (2.3.1) does not depend on the variational parameters, and the negative of the other two terms can be used as an alternative objective function:

$$\mathcal{L}(q) = \int q(\mathbf{z}) \log \frac{p(\mathbf{z}, \mathbf{x})}{q(\mathbf{z})} d\mathbf{z}, \tag{2.3.2}$$

where $\boldsymbol{\lambda}$ is omitted to ease notation. The three quantities, $\mathcal{L}(q)$, $\mathrm{KL}(q||p)$ (as defined in (2.3.1) without the arguments) and $\log p(\mathbf{x})$ are connected via the following equation, which also gives a decomposition of the log marginal likelihood for an arbitrary density q:

$$\log p(\mathbf{x}) = \mathcal{L}(q) + \mathrm{KL}(q||p).$$

Clearly, maximising $\mathcal{L}(q)$ is equivalent to minimising $\mathrm{KL}(q||p)$. The objective function $\mathcal{L}(q)$ is also referred to as the evidence lower bound (ELBO) as it lower bounds the log-evidence: $\log p(\mathbf{x}) \geq \mathcal{L}(q)$ for any q, since $\mathrm{KL}(q||p)$ is always a non-negative quantity.

Another way to derive the ELBO is by using Jensen's inequality (Jensen, 1906) and the concavity of the logarithm:

$$\begin{aligned}\log p(\mathbf{x}) &= \log \int p(\mathbf{z}, \mathbf{x}) d\mathbf{z} \\ &= \log \int \frac{p(\mathbf{z}, \mathbf{x}) q(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\ &= \log \mathbb{E}_q \left[\frac{p(\mathbf{z}, \mathbf{x})}{q(\mathbf{z})} \right] \\ &\geq \mathbb{E}_q \left[\log \frac{p(\mathbf{z}, \mathbf{x})}{q(\mathbf{z})} \right] \equiv ELBO(q(\mathbf{z})) = \mathcal{L}(q).\end{aligned}$$

VI provides the means to approximate the posterior distribution as well as the marginal likelihood by optimising the ELBO. The lower bound is attained when KL($q||p$) is zero, which only happens when $q(\mathbf{z})$ equals the posterior $p(\mathbf{z}|\mathbf{x})$. This never occurs in non-trivial problems, as the intractability of the posterior is the reason why one resorts to approximate inference in the first place.

2.3.2 Mean-field approximations

Variational distribution $q(\mathbf{z})$ should be chosen in a way that it is complex enough to approximate the posterior as closely as possible, while at the same time being simple enough to provide a tractable approximation. A very common and very simple choice is a fully factorised distribution, referred to as a mean-field distribution. An approximation of this kind assumes that all latent variables are independent, and each is governed by its own set of variational parameters:

$$q(\mathbf{z}|\boldsymbol{\lambda}) = \prod_{i=1}^{M} q_i(z_i|\lambda_i). \tag{2.3.3}$$

This factorised form is the only assumption made about the variational distribution, and there are no further restrictions imposed on the individual factors $q_i(z_i|\lambda_i)$. Mean-field variational inference (MFVI) has its origins in an approximation framework developed in the field of statistical mechanics called *mean-field* theory (Parisi, 1988).

2.3.3 Coordinate ascent equations

From all possible distributions of the form of (2.3.3), we need to find the one for which the lower bound $\mathcal{L}(q)$ is maximised. We therefore need to perform a free-form optimisation of $\mathcal{L}(q)$ with respect to all the individual factors $q_i(z_i)$ (where again dependence on λ_i's is omitted), which is achieved by optimising each factor in turn, in an iterative scheme. This is accomplished by inserting (2.3.3) into (2.3.2), which gives the following expression of the ELBO:

$$\begin{aligned}\mathcal{L}(q) &= \int \prod_{i=1} q_i(z_i)\{\log p(\mathbf{z}, \mathbf{x}) - \sum_i \log q_i(z_i)\}dz_i \\ &= \int q_j(z_j)\Big\{\int \log p(\mathbf{z}, \mathbf{x}) \prod_{i \neq j} q_i(z_i) dz_i\Big\}dz_j - \int q_j(z_j) \log q_j(z_j) dz_j + \text{const} \\ &= \int q_j(z_j) \mathbb{E}_{i \neq j}\big[\log p(\mathbf{z}, \mathbf{x})\big] dz_j - \int q_j(z_j) \log q_j(z_j) dz_j + \text{const}.\end{aligned} \tag{2.3.4}$$

The notation $\mathbb{E}_{i \neq j}$ denotes an expectation with respect to the variational distribution over all latent variables in $\mathbf{z}$ except for z_j. If we define a new distribution $\tilde{p}(\mathbf{x}, z_j)$ by the relation $\log \tilde{p}(\mathbf{x}, z_j) = \mathbb{E}_{i \neq j}\big[\log p(\mathbf{z}, \mathbf{x})\big] + \text{const}$, then equation (2.3.4) assumes the form of a negative KL divergence between $q_j(z_j)$ and $\tilde{p}(\mathbf{x}, z_j)$. This can be optimised with respect to $q_j(z_j)$ by keeping all distributions $\{q_{i \neq j}\}$ fixed, which yields the optimal solution $q_j^*(z_j)$ as:

$$\log q_j^*(z_j) = \mathbb{E}_{i \neq j}\big[\log p(\mathbf{z}, \mathbf{x})\big] + \text{const}. \tag{2.3.5}$$

Equation (2.3.5) can be exponentiated to give:

$$q_j^*(z_j) \propto \exp\big\{\mathbb{E}_{i \neq j}\big[\log p(\mathbf{z}, \mathbf{x})\big]\big\}, \tag{2.3.6}$$

which can then be normalised by dividing with $\int \exp\big\{\mathbb{E}_{i \neq j}\big[\log p(\mathbf{z}, \mathbf{x})\big]\big\} dz_j$. This turns out to be a very useful result in practice, as it says that for each factor in

(2.3.3) we simply take the log of the joint distribution over all variables, observed and latent, and then average with respect to those variables not in that factor.

Cycling through equations (2.3.5) (or 2.3.6) and updating factors $q_j(z_j)$ in turn underlies the coordinate ascent variational inference (CAVI), presented in Algorithm 1.

Algorithm 1 CAVI for obtaining the optimal densities under mean-field approximation.

```
Initialise: $q_1(z_1), \dots q_M(z_M)$
while ELBO has not converged do
    for $j \in \{1, ..., M\}$ do
        Set $q_j^*(z_j) \propto \exp\big\{\mathbb{E}_{i \neq j}\big[\log p(\mathbf{z}, \mathbf{x})\big]\big\}$
    end for
    Compute ELBO
end while
```

Practical implementation of Algorithm 1 requires that factors $q_j(z_j)$ are available in closed form. This is the case when we are working with conditionally conjugate models within the exponential family. Working with exponential family models greatly simplifies variational inference, and this simplicity has motivated extensive research on variational methods. Two examples are the study of theoretical properties of variational approximations (Wang and Titterington, 2012) and stochastic variational inference (SVI, Hoffman et al., 2013), a method that scales VI up to massive data. In the case of non-conjugate models, a very typical example being Bayesian logistic regression, CAVI cannot be directly applied, and other solutions have been studied (e.g. Wang and Blei, 2013).

All in all, literature on VI is quite extensive, and the reader is referred to (Blei et al., 2017) for a comprehensive review.

2.4 Computational methods for spike-and-slab models

In this section, we review the most relevant methods that employ spike-and-slab priors on the coefficients of Bayesian linear regression models in high-dimensional settings. These methods, conduct inference using either VI or EM algorithms and have appeared in the literature in recent years.

Logsdon et al. (2010) introduced variational inference in genome-wide association (GWA) studies. They partitioned the covariates into having either positive, negative, or zero effects and proposed the following prior with one spike and two slabs:

$$\begin{gathered}
\beta_j \sim (1-\rho_{\beta_+}-\rho_{\beta_-})\delta_0(\beta_j) + \rho_{\beta_+}\mathcal{N}_+(0,\sigma^2_{\beta_+}) + \rho_{\beta_-}\mathcal{N}_-(0,\sigma^2_{\beta_-}),\\
(\rho_{\beta_+},\rho_{\beta_-},1-\rho_{\beta_+}-\rho_{\beta_-}) \sim \mathrm{Dir}(1,1,1),\\
\sigma^2_{\beta_+},\sigma^2_{\beta_-} \sim \chi_1^{-2},
\end{gathered}$$

where $\mathcal{N}_+,\mathcal{N}_-$ are positive and negative truncated normal distributions respectively, Dir denotes the Dirichlet distribution, and χ_1^{-2} denotes the inverse chi-squared distribution with one degree of freedom.

Carbonetto and Stephens (2012) applied VI to a Bayesian linear model with a discrete spike-and-slab prior with a normal slab and assessed the performance of their algorithm in GWA studies. They computed the low-dimensional posterior distribution of the hyperparameters using importance sampling, where they replaced the intractable marginal likelihood in the importance weights with its corresponding evidence lower bound obtained from the variational approximation. They compared the results of their method with posterior inference via MCMC and also inferred a Bayesian logistic regression model using a modified version of their algorithm.

You et al. (2014) studied the variational posterior for a Bayesian linear regression model with a normal prior on the coefficients and an inverse gamma prior on the error variance. They provided theoretical results on the consistency of their estimators and introduced two variational information criteria: the variational Akaike and the variational Bayesian information criteria. Ormerod et al. (2017) extended their work and applied variational inference to the Bernoulli-Gaussian model and provided similar consistency results. The Bernoulli-Gaussian model was introduced by Kuo and Mallick (1998) and is also known as the binary mask model. It is similar to a linear model with a discrete spike-and-slab prior, in the sense that it also introduces binary indices. It has the following hierarchical representation:

$$
\begin{aligned}
y_i|\mathbf{x_i}, \boldsymbol{\beta}, \mathbf{z}, \sigma^2 &\sim \mathcal{N}(\sum_j z_j \beta_j x_{ij}, \sigma^2), \\
z_j &\sim \text{Bern}(\rho_0), \\
\beta_j &\sim \mathcal{N}(0, \sigma_\beta^2).
\end{aligned}
$$

Ray and Szabó (2022) studied a variational approximation to a discrete spike-and-slab model with Laplace slabs and provided theoretical results. They used the following hierarchical prior on β_j:

$$
\begin{aligned}
\beta_j|z_j &\sim z_j \text{Lap}(\lambda) + (1 - z_j)\delta_0(\beta_j), \\
z_j|\rho_j &\sim \text{Bern}(\rho_j), \\
\rho_j &\sim \text{Beta}(a_0, b_0).
\end{aligned}
$$

This setting is very close to the one proposed in this thesis. We briefly present here five key points of their work which we treat differently in Chapter 4. Firstly, probabilities ρ_j vary for each j and the prior inclusion probability equals $p(z_j = 1) = a_0/(a_0 + b_0)$, which is the mean of the beta prior on ρ_j. Secondly, regarding the error variance σ^2, instead of assigning a prior to it, they calculate an empirical estimation $\hat{\sigma}^2$ as in Reid et al. (2016):

$$\hat{\sigma}^2 = \frac{1}{N - M} \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\hat{\lambda}}\|_2^2,$$

where $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ is the vector of coefficients estimated by LASSO, $\hat{\lambda}$ is the regularisation parameter selected via cross-validation, and M is the number of non-zero elements in $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$. Thirdly, due to the Laplace slab, the ELBO contains the expectation of $|\beta_j|$ with respect to β_j's variational distribution ($\mathcal{N}(\mu_j, s_j^2)$), which is the mean of a folded normal distribution: $s_j\sqrt{2/\pi}e^{-\mu_j^2/(2s_j^2)} + \mu_j(1 - 2\Phi(-\mu_j/s_j))$, where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution. Since there are variational parameters inside $\Phi(\cdot)$, direct optimisation of the ELBO is not possible and they resort to numerical methods which inevitably make the process slower. A fourth point is that they use the maximal change in binary entropy of the posterior inclusion probabilities to monitor convergence. The problem with this quantity is that it does not change monotonically per iteration. Finally, a fifth feature of their work is their proposed initialisation scheme. They proposed taking a preliminary ridge regression estimation of the coefficients, upon which the ordering of the updating process is determined. They do so to deal with situations where all non-zero coefficients are gathered in one part of the signal. We have found this scheme unsuitable for non-trivial examples, and this has motivated us to develop an alternative initialisation process. Finally, the authors follow the same procedure for a Bayesian logistic regression model (Ray et al., 2020).

Ročková and George (2014) developed an expectation-maximisation algorithm that fits linear models with a continuous spike-and-slab prior, which is a mixture of normal distributions, on the coefficients. Ročková and George (2018) proposed the spike-and-slab LASSO (SSLASSO), which employs a spike-and-slab mixture of double-exponential distributions and provided an implementation via EM and coordinate-wise optimisation.

Tang et al. (2017a) extended the SSLASSO framework to generalised linear models in order to analyse large-scale molecular data. Tang et al. (2018) further extended the framework to include group-specific priors and Tang et al. (2017b) to a Cox proportional hazard model. These methods are implemented in the BhGLM (Yi et al., 2019) package in R (R Core Team, 2021).

Chapter 3

Loss-based prior for the degrees of freedom of the Wishart distribution

3.1 Introduction

Vector autoregressive (VAR) models are widely used by macroeconomists and central bankers for analysing and forecasting time series related to finance or economic problems. In particular, VAR models are flexible multivariate time series models that capture interdependencies among multiple variables of interest. They were introduced by Sims (1980) and have more recently been used to analyse the dynamic properties of economic variables.

In order to avoid overparametrisation and overfitting issues, Bayesian inference for VAR models has been successfully introduced (e.g. Doan et al., 1984; Litterman, 1986; Sims and Zha, 1998). In the recent literature, different prior specifications for the matrix of coefficients of the VAR models have been developed. In particular, Doan et al. (1984) and Litterman (1986) introduced a set of prior distributions that centre the whole system on a multivariate random walk. To deal with high-dimensional models and the forecasting of macroeconomic scenarios (e.g. Bańbura et al., 2010; Carriero et al., 2015, 2019; Huber and Feldkircher, 2019), hierarchical models have been introduced to address dimensionality issues. For instance, George et al. (2008) introduced stochastic search variable selection (SSVS) priors, while recent papers show improvements in inference and forecasting when dealing with parametric (e.g. normal-gamma or Dirichlet-Laplace, see Huber and Feldkircher (2019) and Cross et al. (2020)) and nonparametric (e.g. Bayesian additive regression trees (BART), see Huber and Rossini (2020), and Dirichlet process, see Billio et al. (2019)) shrinkage priors.

In the case of constant volatility models, all the approaches described previously assign a Wishart prior distribution to the precision matrix of the error terms. Thus, a very common choice for the prior of the parameters of the VAR model is the normal-Wishart distribution, where the vector of coefficients follows a priori a normal distribution and the error precision matrix follows a Wishart distribution. The advantage of this representation is the possibility of having closed-form conditional posterior distributions, which have the same form as the prior distributions.

In the usual representation of the Wishart distribution, we infer the degrees of freedom and the scale matrix. In particular, in the VAR representation, the degrees of freedom are assumed a priori to be equal to the size of the variable of interest plus one. As far as we are aware, no investigation has been done regarding this assumption, and in this chapter, we address this issue. In particular, we assume a hyperprior for the degrees of freedom by following the literature on loss-based priors (see Villa and Walker, 2015).

As stated in the simulation and empirical studies, the use of a hyperprior on the degrees of freedom leads to better results compared to fixing the value of the degrees of freedom a priori. This result provides the opportunity to investigate different values for the degrees of freedom. The aim of this chapter is to contribute to the literature on multivariate time series by introducing a novel hyperprior on the degrees of freedom.

We illustrate the improved performance of our hyperprior compared to fixing the degrees of freedom by applying our approach to two different datasets. In the first application, to assess how our model performs in a typical real-data application, we forecast different macroeconomic variables in the US using the FRED dataset (McCracken and Ng, 2020). We study the merits of our approach by considering three datasets of differing sizes, namely, a small (with $m = 3$), a medium (with $m = 7$), and a large (with $m = 15$) dataset. The second dataset analysed is the Google Dengue Trends (GDT) for ten different countries (Carneiro and Mylonakis, 2009; Strauss et al., 2017). In both applications, we perform out-of-sample predictions to measure the forecasting accuracy of our approach compared to the normal-Wishart prior with fixed degrees of freedom. Comparing the two priors reveals that our approach improves forecasting accuracy in terms of point and density forecasting measures.

The chapter is organised as follows. Section 3.2 describes the VAR model and the normal-Wishart prior setup. Section 3.3 focuses on the derivation of the loss-based prior for the degrees of freedom of the Wishart distribution. In Section 3.4, we compare the proposed hyperprior with the benchmark assumption using data simulated from a multivariate time series. Section 3.5 deals with real data; in particular we forecast macroeconomic and Google Trends data. Final discussion points and conclusions are presented in Section 3.6.

3.2 Preliminaries

Let $\boldsymbol{y}_t$ for $t = 1, \ldots, T$ be the $m \times 1$ vector of observations, then we can define a VAR model with p lags as

$$\boldsymbol{y}_t = \sum_{j=1}^{p} A_j \boldsymbol{y}_{t-j} + \boldsymbol{\varepsilon}_t, \tag{3.2.1}$$

where A_j is an $m \times m$ matrix of coefficients and $\boldsymbol{\varepsilon}_t$ an $m \times 1$ vector of errors. We assume that $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{mt})^\top$ are i.i.d. for $t = 1, \ldots, T$ with distribution $\mathcal{N}(0, \Sigma)$. Notice that we have not included an intercept term in (3.2.1); in all of our applications, the variables are transformed so that $\mathbb{E}(\boldsymbol{y}_t) = \mathbf{0}$.

Following Kadiyala and Karlsson (1997), we write the VAR model as a system of multivariate regressions

$$Y = XA + E, \tag{3.2.2}$$

where Y is a $T \times m$ matrix constructed as $Y = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T)^\top$, X is a $T \times k$ matrix constructed as

$$X = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_T \end{bmatrix},$$

where $\boldsymbol{x}_t = (\boldsymbol{y}_{t-1}^\top, \boldsymbol{y}_{t-2}^\top, \ldots, \boldsymbol{y}_{t-p}^\top)$ is a $1 \times k$ vector that contains the lagged response variables, and $k = mp$. Some authors (e.g. Carriero et al. (2015); Cross et al. (2020)) suggest transforming the variables (columns of Y and X) to make them stationary. Notice that both equations (3.2.1) and (3.2.2) imply that each of the m variables depends upon p lags of itself and p lags of all the other $m-1$ variables. Moreover, A is a $k \times m$ matrix of coefficients constructed as $A = (A_1, A_2, \ldots, A_p)^\top$ and E is a $T \times m$ matrix of errors constructed as $E = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_T)^\top$. As a further step, we can apply the vec$(\cdot)$ operator on both sides of (3.2.2) and get the vectorised form

$$\boldsymbol{y} = (I_m \otimes X)\,\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{3.2.3}$$

where $\boldsymbol{y} = \text{vec}(Y)$, $\boldsymbol{\alpha} = \text{vec}(A)$, $\boldsymbol{\varepsilon} = \text{vec}(E)$ with distribution $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma \otimes I_T)$ and $\otimes$ is the Kronecker product. Vectors $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ are $mT \times 1$, $\boldsymbol{\alpha}$ is $mk \times 1$ and matrix $I_m \otimes X$ is $mT \times mk$.

The Kronecker product $I_m \otimes X$ in (3.2.3) implies that all $\boldsymbol{y_t}$s are modelled using the same set of explanatory variables. This feature does not allow for the imposition of restrictions on specific coefficients; for example, a practitioner might want to restrict specific coefficients on the lagged dependent variables to zero. We do not impose restrictions in any of the applications in this chapter, and we refer the reader to Koop et al. (2010) for an alternative formulation of the VAR model that allows for restrictions.

The likelihood, as a function of $\boldsymbol{\alpha}$ and Σ, can be shown (Kadiyala and Karlsson, 1997) to be proportional to the product of a normal distribution for $\boldsymbol{\alpha}$ that depends on Σ, and a Wishart distribution for Σ^{-1}. In particular, the likelihood function is

$$L(\boldsymbol{\alpha}, \Sigma^{-1}) \propto \mathcal{N}(\boldsymbol{\alpha} | \hat{\boldsymbol{\alpha}}, \Sigma \otimes (X^\top X)^{-1}) \times \mathcal{W}\big(\Sigma^{-1} | S^{-1}, T - k - m - 1\big),$$

where $\hat{A} = (X^\top X)^{-1} X^\top Y$ is the OLS estimate of A[1], $\hat{\boldsymbol{\alpha}} = \text{vec}(\hat{A})$ and $S = (Y - X\hat{A})^\top (Y - X\hat{A})$. This form suggests that the natural conjugate prior for the parameters of the VAR model is the normal-Wishart prior:

$$\begin{aligned} \boldsymbol{\alpha} | \Sigma &\sim \mathcal{N}(\underline{\boldsymbol{\alpha}}, \Sigma \otimes \underline{V}), \\ \Sigma^{-1} &\sim \mathcal{W}(\underline{\nu}, \underline{S}^{-1}). \end{aligned}$$

In order to explain a drawback of the natural conjugate prior, we introduce the

[1] The OLS estimate requires the assumption that there is no perfect multicollinearity in X, so that $X^\top X$ is invertible.

following equation for the ith variable:

$$\boldsymbol{y}_i = X\boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{y}_i$, $\boldsymbol{\alpha}_i$ and $\boldsymbol{\epsilon}_i$ are the ith columns of Y,A and E respectively, and $i = 1, \ldots, m$. Thus, $\boldsymbol{\alpha}_i$ is a $k \times 1$ column vector that contains the coefficients of the response variables in the ith equation, and by stacking $\boldsymbol{\alpha}_i$s we recover $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_k \end{bmatrix}.$$

The Kronecker product $\Sigma \otimes \underline{V}$ in the covariance of the conditional prior of $\boldsymbol{\alpha}$ implies that the prior covariance matrices of the coefficients in any two equations, only differ by a scale factor since they are both proportional to $\underline{V}$. To avoid this restrictive feature, we adopt a normal-Wishart prior in which $\boldsymbol{\alpha}$ is independent of Σ:

$$\boldsymbol{\alpha} \sim \mathcal{N}(\underline{\boldsymbol{\alpha}}, \underline{V}),$$
$$\Sigma^{-1} \sim \mathcal{W}(\underline{\nu}, \underline{S}^{-1}).$$

This prior is referred to as the conditionally conjugate (Murphy, 2012) or independent (Koop et al., 2010) normal-Wishart prior, and leads to conditional posteriors of the same forms, namely a normal for $\boldsymbol{\alpha}$ and a Wishart for Σ^{-1}. These conditional posteriors can be used to conduct Bayesian inference via Gibbs sampling

(Geman and Geman, 1984; Gelfand and Smith, 1990). In particular, the conditional posterior distribution of the vectorised matrix of coefficients is

$$\boldsymbol{\alpha}|\boldsymbol{y}, \Sigma^{-1} \sim \mathcal{N}(\overline{\boldsymbol{\alpha}}, \overline{V}),$$

where $\overline{V} = (\underline{V}^{-1} + Z^\top \Lambda^{-1} Z)^{-1}$ and $\overline{\boldsymbol{\alpha}} = \overline{V}(\underline{V}^{-1}\underline{\boldsymbol{\alpha}} + Z^\top \Lambda^{-1}\boldsymbol{y})$, where $Z = I_m \otimes X$ and $\Lambda = \Sigma \otimes I_T$. On the other hand, the conditional posterior distribution for the precision matrix is

$$\Sigma^{-1}|\boldsymbol{\alpha}, \boldsymbol{y} \sim \mathcal{W}\left(\overline{\nu}, \overline{S}^{-1}\right),$$

where $\overline{\nu} = \underline{\nu} + T$ and $\overline{S} = \underline{S} + (Y - XA)^\top (Y - XA)$. Derivations of the conditional posterior distributions can be found in Appendix A.1, where we have used both the matrix representation on the VAR model in (3.2.2) and its vectorised version in (3.2.3) as well as the assumption of normality of the errors.

3.3 Loss-based hyperprior

In this section, we derive the loss-based prior distribution for the number of degrees of freedom of the Wishart distribution. We consider the parameter ν as discrete, and to construct the prior, we employ the objective method introduced in Villa and Walker (2015).

To illustrate the method, let us consider a Bayesian model with sampling distribution $f(x|\theta)$, characterised by the discrete parameter $\theta \in \Theta$, and prior $\pi(\theta)$. The idea is to assign a mass to each value of the parameter that is proportional to the Kullback–Leibler divergence between the model defined by θ and the nearest one. In other words, if $f(x|\theta)$ is the true model, and it is not chosen, then the loss in information that one would incur is represented by the Kullback–Leibler divergence from $f(x|\theta)$ and $f(x|\theta')$, where the latter is the nearest model to the true one. Thus, the prior on θ will be given by Eq. (3.3.1).

$$\pi(\theta) \propto \exp\left\{\min_{\theta' \neq \theta \in \Theta} \mathrm{KL}\left(f(x|\theta) \| f(x|\theta')\right)\right\} - 1, \tag{3.3.1}$$

where $\mathrm{KL}\left(f(x|\theta) \| f(x|\theta')\right)$ represents the Kullback–Leibler divergence between the two models.

The probability density function of a Wishart distribution with parameters the positive semi-definite scale matrix V and degrees of freedom ν, is given by:

$$\begin{aligned} \mathcal{W}(X|V,\nu) &\triangleq \frac{1}{Z}|X|^{(\nu-m-1)/2}\exp(-\operatorname{Tr}(V^{-1}X)/2), \\ Z &\triangleq |V|^{-\nu/2}2^{\frac{\nu m}{2}}\Gamma_m(\nu/2), \end{aligned}$$

where $\Gamma_m(x) = \pi^{m(m-1)/4}\prod_{j=1}^{m}\Gamma\big(x + \frac{1-j}{2}\big)$ is the multivariate gamma function and $\operatorname{Tr}(\cdot)$ is the trace function. The density function is well-defined only if $\nu > m-1$, since $\Gamma_m(\nu/2)$ in the normalisation constant Z has either simple poles or is negative if $\nu \leq m-1$. If $m = 1$ and V is a scalar, the Wishart reduces to the gamma distribution with shape parameter $\frac{\nu}{2}$ and rate parameter $\frac{1}{2V}$. In this chapter, we use the Wishart distribution to model (positive definite) matrices, and thus we only consider the case $\nu \geq m \geq 2$.

The Kullback–Leibler divergence between two Wishart distributions (Beal, 2003) that share the same $m \times m$ scale matrix and differ only in the number of degrees of freedom, say $\mathcal{W}_\nu$ and $\mathcal{W}_{\nu+c}$, is given by:

$$\mathrm{KL}(\mathcal{W}_\nu \| \mathcal{W}_{\nu+c}) = \log\left\{\frac{\Gamma_m\left(\frac{\nu+c}{2}\right)}{\Gamma_m\left(\frac{\nu}{2}\right)}\right\} - \frac{c}{2}\psi_m\left(\frac{\nu}{2}\right), \tag{3.3.2}$$

where ψ_m is the multivariate digamma function defined as $\psi_m(x) = \sum_{i=1}^{m}\psi(x + (1-i)/2)$, $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function. If $\mathcal{W}_\nu$ is well-defined, $\mathcal{W}_{\nu+c}$ is also well-defined if $c > m - \nu - 1$.

Application of (3.3.1) to the Wishart distribution requires finding the value of c that minimises (3.3.2). In doing so, it is useful to consider a real-valued c, i.e. $c \in \mathbb{R} : c > m - \nu - 1$, and examine the convexity of (3.3.2) as a function of c. The second derivative of (3.3.2) with respect to c is $\sum_{j=1}^{m} \frac{\partial^2}{\partial c^2} \log \Gamma\left(\frac{\nu+c+1-j}{2}\right)$. The second derivative of the gamma function is known as the trigamma function ($\psi^{(1)}(x)$), and is a special case of the polygamma function defined as: $\psi^{(n)}(x) = (-1)^{n+1} \int_0^\infty \frac{t^n e^{-xt}}{1-e^{-t}} dt$ (Abramowitz and Stegun, 1972, Eq. 6.4.1). For odd n, as in the case of trigamma, the polygamma function is strictly positive. Thus, the second derivative of (3.3.2) is positive and hence, (3.3.2) is a strictly convex function of c. Equation (3.3.2) has a global minimum at $c = 0$ since $\mathrm{KL}(\mathcal{W}_\nu \| \mathcal{W}_\nu) = 0$ and the Kullback–Leibler divergence is always non-negative. The two properties, strict convexity and global minimum at $c = 0$, imply that if we restrict c to integer values and exclude 0, i.e. $c \in \mathbb{Z}^{\neq} : c > m - \nu - 1$, $\mathrm{KL}(\mathcal{W}_\nu \| \mathcal{W}_{\nu+c})$ will be minimised for either $c = -1$ or $c = 1$. Theorem 1 shows that the Kullback–Leibler divergence between $\mathcal{W}_\nu$ and $\mathcal{W}_{\nu+c}$ is minimised for $c = 1$.

Theorem 1. *Let $\mathcal{W}_\nu$ and $\mathcal{W}_{\nu+c}$ denote two Wishart distributions with the same $m \times m$ scale matrix, where ν and $\nu + c$ represent their degrees of freedom, and $c \in \{-1, 1\}$. The Kullback–Leibler divergence between $\mathcal{W}_\nu$ and $\mathcal{W}_{\nu+c}$ is minimised when $c = 1$, for any ν and m such that $\nu \geq m \geq 2$.*

Proof. For $c = 1$, the Kullback–Leibler divergence is

$$\mathrm{KL}(\mathcal{W}_\nu \| \mathcal{W}_{\nu+1}) = \log \Gamma_m\left(\frac{\nu+1}{2}\right) - \log \Gamma_m\left(\frac{\nu}{2}\right) - \frac{1}{2}\psi_m\left(\frac{\nu}{2}\right),$$

and for $c = -1$, it is

$$\mathrm{KL}(\mathcal{W}_\nu \| \mathcal{W}_{\nu-1}) = \log \Gamma_m\left(\frac{\nu-1}{2}\right) - \log \Gamma_m\left(\frac{\nu}{2}\right) + \frac{1}{2}\psi_m\left(\frac{\nu}{2}\right).$$

By taking the difference of the two divergences, we have

$$
\begin{aligned}
\mathrm{KL}(\mathcal{W}_\nu \| \mathcal{W}_{\nu+1}) - \mathrm{KL}(\mathcal{W}_\nu \| \mathcal{W}_{\nu-1}) &= \log\Gamma_m\left(\frac{\nu+1}{2}\right) - \log\Gamma_m\left(\frac{\nu-1}{2}\right) - \psi_m\left(\frac{\nu}{2}\right) \\
&= \log\frac{\Gamma(\nu)}{2^m\Gamma(\nu-m)} - \psi_m\left(\frac{\nu}{2}\right).
\end{aligned}
$$

We will prove that this difference is always negative for any ν, m such that $\nu \geq m \geq 2$.

If $\nu = m$, c can only be equal to 1; if $c = -1$, $\mathcal{W}_{m-1}$ is not a well-defined density and the difference equals $-\infty$ because $\log\frac{\Gamma(m)}{2^m\Gamma(0)} = -\infty$ since $\frac{1}{\Gamma(0)} = 0$ (Abramowitz and Stegun, 1972, Eq. 6.1.3).

If $\nu > m$, then we can write ν as $\nu = m + k$, with $k = 1, 2, \ldots$ and $m \geq 2$. By proving the following inequality, we prove that the minimum Kullback–Leibler divergence is achieved at $c = 1$:

$$
\log\left\{\frac{\Gamma(m+k)}{2^m\Gamma(k)}\right\} < \psi_m\left(\frac{m+k}{2}\right). \tag{3.3.3}
$$

Two results that enable us to prove the inequality in (3.3.3), are the following:

$$
\psi_{m+1}\left(\frac{\nu+1}{2}\right) = \psi_m\left(\frac{\nu}{2}\right) + \psi\left(\frac{\nu+1}{2}\right), \tag{3.3.4}
$$

and

$$
\log\left(x - \tfrac{1}{2}\right) < \psi(x). \tag{3.3.5}
$$

Equation (3.3.4) comes from the definition of the multivariate digamma function. Inequality (3.3.5) can be deduced from the following inequality which is found in Elezovic et al. (2000):

$$
\log\left(x + \tfrac{1}{2}\right) < \psi(x+1),
$$

for $x > 0$, which by using the recurrence formula $\psi(x+1) = \psi(x) + \frac{1}{x}$ (Abramowitz and Stegun, 1972, Eq. 6.3.5), becomes

$$\log\left(x+\tfrac{1}{2}\right)-\frac{1}{x}<\psi(x). \tag{3.3.6}$$

From (Mitrinovic et al., 1964, Eq. 2.79) we have that:

$$\frac{1}{a}<\log(a+1)-\log a,$$

for $a>0$, and obviously $\frac{1}{a+1/2}<\frac{1}{a}$, thus we have:

$$\frac{1}{a+\frac{1}{2}}<\log(a+1)-\log a.$$

By setting $a=x-1/2$, we get:

$$\log\left(x-\tfrac{1}{2}\right)<\log\left(x+\tfrac{1}{2}\right)-\tfrac{1}{x}. \tag{3.3.7}$$

From (3.3.6) and (3.3.7) we have that $\log\left(x-\frac{1}{2}\right)<\psi(x)$. Notice that $\log\left(x-\frac{1}{2}\right)$ exists only for $x>1/2$, a condition that is not violated when we use the inequality later on.

At first, we make the assumption that (3.3.3) holds for a particular m and prove that it holds for $m+1$.

$$\begin{aligned}\log\left\{\frac{\Gamma(m+1+k)}{2^{m+1}\Gamma(k)}\right\}&<\psi_{m+1}\left(\frac{m+1+k}{2}\right)\\ \log\left\{\frac{\Gamma(m+k)(m+k)}{2^{m}\Gamma(k)2}\right\}&<\psi_{m}\left(\frac{m+k}{2}\right)+\psi\left(\frac{m+1+k}{2}\right)\\ \log\left\{\frac{\Gamma(m+k)}{2^{m}\Gamma(k)}\right\}+\log\left\{\frac{m+k}{2}\right\}&<\psi_{m}\left(\frac{m+k}{2}\right)+\psi\left(\frac{m+1+k}{2}\right),\end{aligned}$$

where in the last step we have that $\log\frac{m+k}{2}<\psi\left(\frac{m+1+k}{2}\right)$ as a consequence of the result in (3.3.5). Thus, if the inequality (3.3.3) holds for m, then it holds for $m+1$ and, more importantly, it holds for any k. The smallest value m can have is $m = 2$, for which we have

$$\log\left\{\frac{\Gamma(2+k)}{2^2\Gamma(k)}\right\} < \psi_2\left(\frac{2+k}{2}\right)$$
$$\log\left\{\frac{(k+1)k}{2^2}\right\} < \psi\left(\frac{2+k}{2}\right) + \psi\left(\frac{1+k}{2}\right).$$

Again, we have that $\log\left(\frac{k}{2}\right) < \psi\left(\frac{k+1}{2}\right)$ and $\log\left(\frac{k+1}{2}\right) < \psi\left(\frac{k+2}{2}\right)$ due to (3.3.5). Thus, inequality (3.3.3) holds for $m = 2$ and, subsequently, it holds for any m. □

Now we can define the objective prior distribution for ν, following Eq. (3.3.1), as

$$\pi(\nu) \propto \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu+1-m}{2}\right)} e^{-\frac{1}{2}\sum_{i=1}^{m}\psi\left(\frac{\nu+1-i}{2}\right)} - 1. \tag{3.3.8}$$

3.3.1 Properness of the posterior for ν

Let us assume that we observe one random matrix Σ^{-1} from the Wishart distribution $\mathcal{W}(S_0^{-1}, \nu)$. By using the objective prior for ν in Eq. (3.3.8), we obtain the following posterior distribution for the number of degrees of freedom:

$$\begin{aligned} p(\nu|\Sigma^{-1}) &\propto \pi(\nu)\pi\left(\Sigma^{-1}|\nu, S_0^{-1}\right) \\ &\propto \left\{\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu+1-m}{2}\right)} e^{-\frac{1}{2}\sum_{i=1}^{m}\psi\left(\frac{\nu+1-i}{2}\right)} - 1\right\} \times \\ &\quad \left\{\frac{|\Sigma^{-1}|^{(\nu-m-1)/2} e^{\left(-\operatorname{Tr}(S_0\Sigma^{-1})/2\right)}}{2^{\frac{\nu m}{2}}|S_0^{-1}|^{\nu/2}\Gamma_m\left(\frac{\nu}{2}\right)}\right\}. \end{aligned} \tag{3.3.9}$$

An important aspect when using an objective prior is to ensure that the yielded posterior is proper. The following Theorem 2 shows that the marginal posterior for ν is proper.

Theorem 2. *The posterior distribution for the number of degrees of freedom ν in Eq. (3.3.9) is proper.*

Proof. We prove that $\sum_{\nu=m}^{\infty} p(\nu|\Sigma^{-1}) < \infty$ using Abel's test of convergence. The sequence $\{\pi(\nu)\}$ is bounded: $\pi(m) > \pi(\nu) > 0$ and is also monotone (decreasing). We show that $\sum_{\nu=m}^{\infty} \pi(\Sigma^{-1}|\nu, S_0^{-1}) < \infty$ using the ratio test:

$$\begin{aligned} R_\nu &= \frac{|\Sigma^{-1}|^{(\nu-m)/2}}{2^{\frac{(\nu+1)m}{2}}|S_0^{-1}|^{(\nu+1)/2}\Gamma_m\left(\frac{\nu+1}{2}\right)} \cdot \frac{2^{\frac{\nu m}{2}}|S_0^{-1}|^{\nu/2}\Gamma_m\left(\frac{\nu}{2}\right)}{|\Sigma^{-1}|^{(\nu-m-1)/2}} \\ &= \frac{|\Sigma^{-1}|^{1/2}\Gamma_m(\frac{\nu}{2})}{2^{m/2}|S_0^{-1}|^{1/2}\Gamma_m\left(\frac{\nu+1}{2}\right)} \\ &= \frac{|\Sigma^{-1}|^{1/2}\Gamma\left(\frac{\nu+1-m}{2}\right)}{2^{m/2}|S_0^{-1}|^{1/2}\Gamma\left(\frac{\nu+1}{2}\right)}. \end{aligned}$$

It follows that

$$\lim_{\nu\to\infty}\left\{\frac{\Gamma(\frac{\nu+1-m}{2})}{\Gamma(\frac{\nu+1}{2})}\right\} = 0,$$

so $\lim_{\nu\to\infty} R_\nu = 0$ (Abramowitz and Stegun, 1972) and, therefore, the posterior for ν is a proper distribution. □

We can draw a sample from $p(\nu|\Sigma^{-1})$ using a Metropolis step (Metropolis et al., 1953) with the following simple proposal. If $\nu = m$, propose $\nu^* = m + 1$. Otherwise, propose either $\nu^* = \nu + 1$ or $\nu^* = \nu - 1$ with equal probability. Accept ν^* with probability $\min\{1, \frac{p(\nu^*|\Sigma^{-1})}{p(\nu|\Sigma^{-1})}\}$.

3.4 Simulation studies

In this section, we compare the performance of our loss-based hyperprior in different simulation studies. In particular, we use a Bayesian VAR model of order 1 with different matrix dimensions and time lengths. Since our interest in this simulation study is to evaluate our prior when estimating the error covariance matrix, we assign an inverse-Wishart prior to Σ, which is equivalent to assigning a Wishart prior to Σ^{-1}, and consider two scenarios of how ν is treated. More specifically, the prior for the coefficients is a zero-mean normal distribution with a diagonal covariance matrix (with diagonal entries equal to 10 to induce a weak prior), while the error covariance matrix follows an inverse-Wishart with a scale matrix equal to I_m and degrees of freedom equal to $\nu = m+1$ (in the first scenario) and $\nu \sim \pi(\nu)$ (in the second scenario).

As stated above, we have created different synthetic datasets for various dimension of the matrix in the VAR. In particular, we have studied small, medium and large size VARs, where m is equal to 5, 10 and 20, respectively. For each dataset, we have considered a small time dimension ($T = 30$) and a medium time dimension ($T = 100$).

Moreover, we have considered different combinations for the choice of the degrees of freedom when generating the data. In particular, for each dimension m, we have chosen ν equal to $\{5, 10, 15\}$ for $m = 5$, $\{10, 15, 20\}$ for $m = 10$, and $\{20, 24, 26\}$ for $m = 20$, respectively.

For each dataset, we run the Gibbs sampler that sequentially draws from the conditional posterior distributions $p(\alpha|y, \Sigma^{-1})$, $p(\Sigma^{-1}|y, \alpha)$ for the first scenario, and a second Gibbs sampler that draws from $p(\alpha|y, \Sigma^{-1})$, $p(\Sigma^{-1}|y, \alpha)$ and $p(\nu|\Sigma^{-1})$ for the second scenario. Then, for each simulated Markov chain, we estimate the posterior means of the elements of Σ and calculate the mean absolute deviation (MAD) between these estimates and the true values that generated each respective dataset as:

$$\text{MAD} = \frac{1}{N}\|\theta - \hat{\theta}\|_1,$$

where $N = m(m+1)/2$ is the number of unique elements of the covariance matrix, θ is the vector containing the unique elements of the true covariance matrix, and $\hat{\theta}$ the respective posterior means. However, all MADs calculated in this section are close to zero and there are several extreme outliers, so we report results in terms of the square root of MAD

$$\text{RMAD} = \sqrt{\text{MAD}},$$

where RMAD is the abbreviation for root mean absolute deviation, in order to make boxplots more legible.

This process is repeated 250 times, and the Gibbs sampler is run for 6000 iterations, with a burn-in of 1000 iterations. For each batch of RMADs, we create different boxplots for the two different scenarios: the fixed ν and the estimated degrees of freedom. In particular, since we are interested in the covariance matrix, we have reported the RMADs for the covariance matrices for each dataset. Figure 3.4.1 shows the results for the RMAD for the case with $m = 5$ and with $T = 30$[2]. The left panel explains the results when the data are generated with $\nu = 5$, the central panel when the data are generated with $\nu = 10$ and the right panel with $\nu = 15$. From Figure 3.4.1, once we move from a dataset generated with a ν equal to m to a dataset with higher degrees of freedom, the differences between the two priors increase. In fact, the left panel shows no difference between the two priors, except for the outliers, which are smaller for our proposed hyperprior. On the other hand, increasing ν to 10 and 15 leads to improvements in the evaluation of our hyperprior.

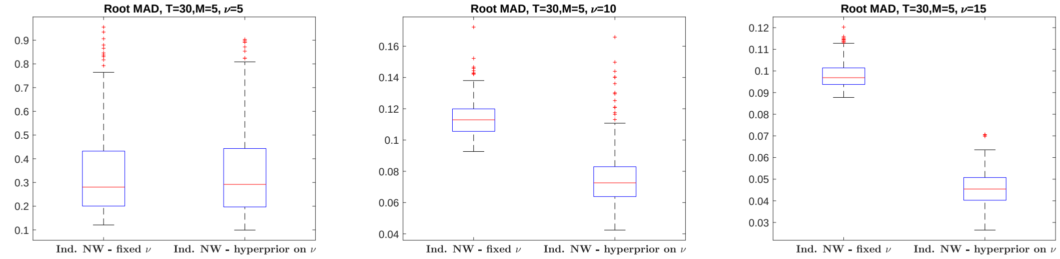


Figure 3.4.1: Monte Carlo simulation — root mean absolute deviations of the covariance matrices of dimension $m = 5$. These empirical distributions are obtained by simulating 250 VAR(1) of sample size $T = 30$. Results are reported separately for data generated from a Wishart with $\nu = 5$ (left panel), $\nu = 10$ (central panel), and $\nu = 15$ (right panel).

These results are also confirmed in high-dimensional cases as shown in Figures 3.4.2 and 3.4.3 for the ten-dimensional and twenty-dimensional cases, respectively. In Figure 3.4.2, we compare our loss-based hyperprior with the fixed ν for the data generated from a Wishart with degrees of freedom equal to 10 (left

[2] The results for $T = 100$ are available in Appendix A.2.

panel), 15 (centre), and 20 (right). As stated above, in this scenario the results indicate that our loss-based prior performs better than the fixed ν for the cases of 15 and 20 degrees of freedom. Regarding the smallest degrees of freedom, we observe small differences between the two prior scenarios, but again the loss-based prior shows better results concerning the outlier values.

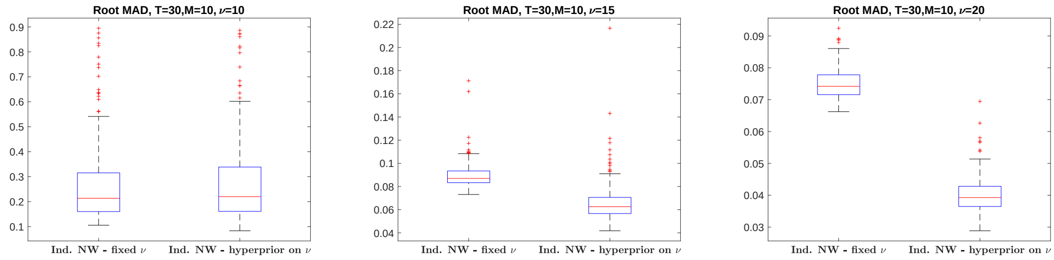


Figure 3.4.2: Monte Carlo simulation — root mean absolute deviations of the covariance matrices of dimension $m = 10$. These empirical distributions are obtained by simulating 250 VAR(1) of sample size $T = 30$. Results are reported separately for data generated from a Wishart with $\nu = 10$ (left panel), $\nu = 15$ (central panel), and $\nu = 20$ (right panel).

In conclusion, Figure 3.4.3 shows the results for the twenty-dimensional case. Also in this case, we have reported three different figures for data generated from a Wishart with 20 degrees of freedom (left panel), 24 (centre), and 26 (right) with T equal to 30. In this case, the improvements are less evident than in the previous case for the left and centre panels. However, in the case of 24 degrees of freedom, our loss-based hyperprior outperforms the fixed ν prior. This result is strong when we use data generated from a Wishart with 26 degrees of freedom, as shown in the right panel of Figure 3.4.3.

These results are confirmed for a higher number of observations, namely T equal to 100, as shown in Appendix A.2. We use the same combinations of ν and m with the only difference being in T. As expected, in this simulation study with a larger sample size, the impact of the objective hyperprior is less extensive and overall RMAD is lower for both priors.

Similarly to the sample size of $T = 30$, in the cases when $\nu = m$ (left panels

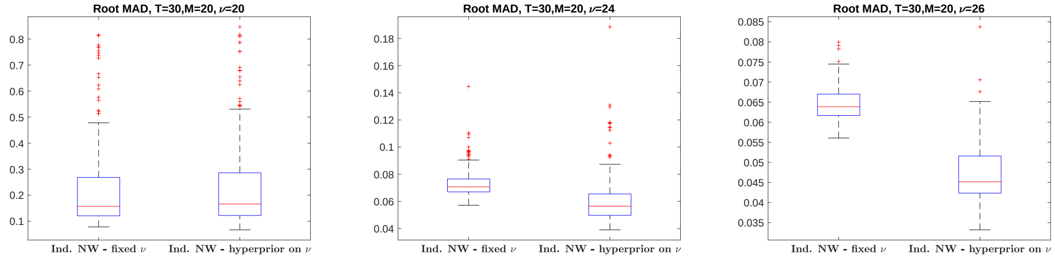


Figure 3.4.3: Monte Carlo simulation — root mean absolute deviations of the covariance matrices of dimension $m = 20$. These empirical distributions are obtained by simulating 250 VAR(1) of sample size $T = 30$. Results are reported separately for data generated from a Wishart with $\nu = 20$ (left panel), $\nu = 24$ (central panel), and $\nu = 26$ (right panel).

of Figures A.2.1,A.2.2,A.2.3) there are subtle differences between the two priors. When we move to data generated with higher degrees of freedom (central panels), the improvement in performance is less evident than in the $T = 30$ case. Median RMAD might be lower with the objective prior, but boxplots exhibit a larger interquartile range and more distant outliers. Finally, in the higher ν datasets that we examine (right panels), we get a clearer picture. There are less outliers, and the interquartile ranges for the two priors overlap only in the $m = 20$ dimension.

3.5 Forecasting of real data

In this section, we compare the results of our loss-based hyperprior with the fixed ν prior in two different real data applications. The first real data application deals with macroeconomic variables from the FRED dataset, and we consider three different sizes: small (3 variables), medium (7 variables), and large (15 variables) datasets. On the other hand, in the second application, we use Dengue fever data from 10 different countries across the world.

In both applications, we evaluate the performance of our loss-based hyperprior by forecasting one-step-ahead values. We compare the predictive ability of the two priors using both a point and a density forecasting measure. Regarding the point forecasting measure, we use the root mean square error (denoted by RMSE), which is

$$\mathrm{RMSE}_i = \left[\frac{1}{T-R}\sum_{t=R}^{T-1}(\hat{y}_{i,t+1} - y_{i,t+1})^2\right]^{\frac{1}{2}}, \tag{3.5.1}$$

where R is the length of the rolling window, $y_{i,t+1}$ is the observation for the ith variable and $\hat{y}_{i,t+1}$ is the one-step-ahead prediction for the ith variable.

In addition, we evaluate the density forecasting accuracy using the average continuous ranked probability score (CRPS) introduced by Matheson and Winkler (1976). The use of the CRPS has some advantages over other scoring functions, as it does a better job of rewarding values from the predictive density that are close to the outcome and is less sensitive to outliers (see Gneiting and Raftery (2007) for a comparison with other score functions). The CRPS is defined such that a lower number is a better score and is given by:

$$\begin{aligned}\mathrm{CRPS}_t(y_{t+1}) &= \int_{-\infty}^{+\infty}(F(z) - \mathbb{1}(y_{t+1} \leq z))^2 dz \\ &= \mathbb{E}_f|Y_{t+1} - y_{t+1}| - 0.5\mathbb{E}_f|Y_{t+1} - Y'_{t+1}|,\end{aligned} \tag{3.5.2}$$

where F is the cumulative distribution function associated with the posterior predictive density, f, $\mathbb{1}(y_{t+1} \leq z)$ is an indicator function taking the value 1 if $y_{t+1} \leq z$ and 0 otherwise, and Y_{t+1}, Y'_{t+1} are independent random draws from the posterior predictive density.

3.5.1 Macroeconomic data

In this experiment, we use a subset of the FRED-QD dataset from McCracken and Ng (2020). The data are at a quarterly frequency and the time period spans from 1959Q2 to 2019Q3. All variables are transformed to be stationary by following McCracken and Ng (2020). In terms of forecasting, this choice leads to more stable predictive densities. Following Huber and Feldkircher (2019), we use three sets of variables to estimate a small-scale, medium-scale, and a large-scale VAR.

The small-scale VAR only includes the real GDP growth, the GDP deflator, and the Federal Funds Rate (FFR). The medium-scale VAR additionally uses data on investment, consumption growth, and hours worked. The large-scale VAR includes 15 different variables. Given the quarterly frequency of our data, we include $p = 5$ lags for all the models considered in this section.

We estimate the models by running the Gibbs sampler for 6000 iterations and discarding the first 1000 iterations as burn-in. Regarding the forecasting exercise, Huber and Feldkircher (2019) use an expanding window with an initial length of 144 quarters that expands until the end of this dataset. We, instead, follow Clark and McCracken (2010), and use a rolling window of 60 quarters (fixed length) in order to compare the two priors across a wider horizon, and we make one-step ahead forecasts. The lags, the rolling window length, as well as the data transformations, all impact when we begin making forecasts. In the first seven rows of the data, there are five rows with missing values caused by lags ($p = 5$) and two more rows with missing values caused by the data transformations. Thus, all three quarters of 1959 and all four quarters of 1960 are lost. Since the rolling window is 60 quarters, or equivalently 15 years, the first forecast is made for 1976Q1.

Before examining the forecasting exercise, we briefly present the results of the estimated degrees of freedom using a rolling window estimation. Both the forecasts and the degrees of freedom estimates are made for the same period. Figure 3.5.1 shows the results of the estimated degrees of freedom together with the 95% highest posterior density (HPD) and the degrees of freedom used in the normal-Wishart scenario with fixed ν (in red). The left panel shows the results for the small-scale VAR, the central panel for the medium-scale VAR, and the right panel for the large-scale VAR.

In Figure 3.5.1, we find strong evidence of changes in the estimated degrees of freedom over time. The left panel shows a "jump" in the degrees of freedom after 2000 and a fall around 2009, strongly related to the Lehman Brothers failure. This "jump" also occurs in the medium-scale and large-scale VARs, though in the latter case it is smaller.

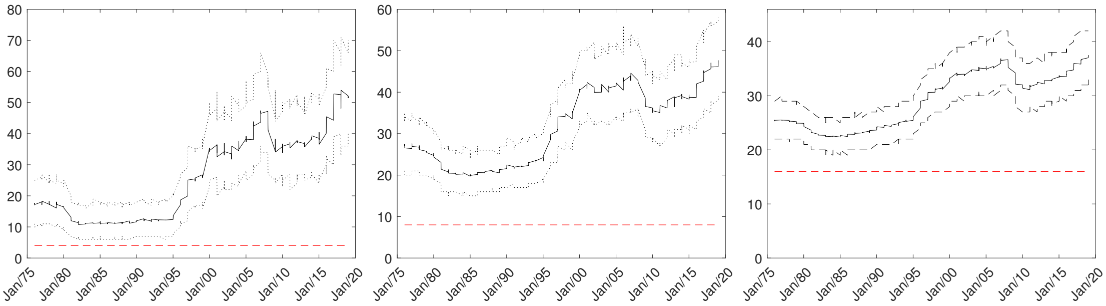


Figure 3.5.1: Estimated degrees of freedom (solid line) for the loss based hyperprior by using a rolling window of 60 quarters with the 95% HPD (dotted lines) and the fixed ν (red dashed line) for the macroeconomic data. Left panel for the small-case; centre for the medium-scale and right for the large-scale.

Moving to the forecasting analysis, we report the point and density forecasting measures in the same table for the three different cases. Table 1 shows the RMSE (left panel) and the CRPS (right panel) for the small-case VAR. In this table, we report the ratio between the two models. When the ratio is less than 1, it means that the model with the loss-based hyperprior is outperforming the benchmark model with fixed ν. In the point forecasting measure, the benchmark model is outperforming our hyperprior for the GDP and the Federal Reserve Fund by a very small margin. On the other hand, for the GDP deflator, our hyperprior leads to an improvement of around 20% compared to the benchmark. This result is confirmed for the average CRPS, where our loss-based hyperprior outperforms the benchmark model by 15% for real GDP growth and the GDP deflator, and by 2% for the Federal Reserve Funds Rate.

Moving to the medium-scale model, Table 2 presents the point and density measures for the 7 different variables analysed. In this scenario, the situation is similar to the small-case VAR regarding both point and density forecasting measures. The main change is related to the GDP, as the model with the hyperprior is outperforming the benchmark one by around 4% in point forecasting,

	RMSE			CRPS		
	fixed ν	$\nu \sim \pi(\nu)$	ratio	fixed ν	$\nu \sim \pi(\nu)$	ratio
GDPC1	0.010	0.010	1.006	0.045	0.039	0.866
GDPCTPI	0.004	0.003	0.811	0.045	0.039	0.868
FEDFUNDS	0.946	0.953	1.006	0.388	0.381	0.981

Table 1: RMSE and average CRPS for the small-case VAR for each prior. The first column refers to the variable, the second to the Gibbs sampler with fixed ν and the third with our loss-based hyperprior.

while for the Federal Funds Rate, the situation does not change in terms of point forecasting. As in the small-scale VAR, the model with the hyperprior shows better results compared to the benchmark model across the 7 variables in terms of CRPS. In particular, for the three variables of interest (the GDP, GDP deflator, and the Federal Funds Rate), the hyperprior model outperforms the benchmark by margins between 3% and 15%. This is also confirmed for the other variables analysed.

	RMSE			CRPS		
	fixed ν	$\nu \sim \pi(\nu)$	ratio	fixed ν	$\nu \sim \pi(\nu)$	ratio
GDPC1	0.008	0.007	0.959	0.058	0.049	0.857
GDPCTPI	0.004	0.004	0.971	0.058	0.050	0.865
FEDFUNDS	0.941	0.947	1.006	0.391	0.381	0.974
PCECC96	0.006	0.006	1.018	0.058	0.050	0.860
GPDIC1	0.032	0.032	1.003	0.061	0.053	0.866
AWHMAN	0.264	0.260	0.984	0.150	0.144	0.957
CES2000000008x	0.007	0.007	0.909	0.059	0.051	0.859

Table 2: RMSE and average CRPS for the medium-case VAR for each prior. The first column refers to the variable, the second to the Gibbs sampler with fixed ν and the third with our loss-based hyperprior.

In conclusion, we report the results for the large-scale VAR with 15 variables in Table 3. In this scenario, the three main variables of interest follow the improvements shown in the medium-scale VAR. Regarding the point forecasting measure, the loss-based hyperprior improves performance compared to the benchmark by around 10% for GDP and its deflator, while for the FFR, the two priors show very similar results. The average CRPS demonstrates that the improvement is stronger for every variable, particularly for the FFR. Looking at all 15 variables analysed, the loss-based hyperprior model always outperforms the benchmark in a density forecasting scenario.

	RMSE			CRPS		
	fixed ν	$\nu \sim \pi(\nu)$	ratio	fixed ν	$\nu \sim \pi(\nu)$	ratio
GDPC1	0.009	0.008	0.910	0.077	0.070	0.913
GDPCTPI	0.005	0.004	0.909	0.078	0.070	0.908
FEDFUNDS	0.949	0.952	1.002	0.397	0.391	0.983
PCECC96	0.007	0.007	0.973	0.078	0.071	0.911
GPDIC1	0.032	0.031	0.984	0.080	0.073	0.914
AWHMAN	0.247	0.246	0.994	0.150	0.144	0.964
CES2000000008x	0.008	0.007	0.880	0.079	0.072	0.906
PRFIx	0.040	0.040	0.999	0.082	0.075	0.916
INDPRO	0.014	0.013	0.908	0.081	0.073	0.905
CUMFNS	1.028	1.018	0.991	0.544	0.533	0.980
SRVPRD	0.007	0.006	0.944	0.081	0.074	0.913
PCECTPI	0.008	0.007	0.904	0.082	0.075	0.912
GPDICTPI	0.007	0.008	1.117	0.082	0.075	0.907
CPIAUCSL	0.009	0.009	1.003	0.083	0.076	0.912
SP500	0.067	0.066	0.993	0.090	0.083	0.919

Table 3: RMSE and average CRPS for the large-case VAR for each prior. The first column refers to the variable, the second to the Gibbs sampler with fixed ν and the third with our loss-based hyperprior.

3.5.2 Dengue data

In the second application, we show the performance of our loss-based hyperprior using the GDT data. Dengue fever is a viral infection transmitted by mosquitoes and is particularly present in South American and Asian countries. Recently, Google has developed a query-based reporting system for infectious diseases (Carneiro and Mylonakis, 2009; Strauss et al., 2017), while previously its accuracy has been assessed for flu (Polgreen et al., 2008; Davis et al., 2016).

Google Trends tracks Dengue fever incidence based on internet search patterns and clusters weekly queries for key terms related to the disease. We use GDT data from January 2011 to December 2014 for Argentina, Bolivia, Brazil, India, Indonesia, Mexico, the Philippines, Singapore, Thailand, and Venezuela, thus having 10 response variables. Following Davis et al. (2016), we examine a vector autoregressive model with two lags and run a forecasting exercise. The data exhibit annual seasonality, and internet searches about Dengue fever peak at different periods throughout the year across the 10 countries. Thus, it is reasonable to use a rolling window of at least one year to estimate the VAR parameters, taking into account all the different patterns in the data. We run the forecasting analysis using three different rolling window lengths: $R = \{52, 104, 156\}$. The two priors compared very similarly in these three settings, and we only report the results for $R = 104$ here. We take first-order differences of the variables to achieve stationarity. This, together with the lags, causes the loss of the first three rows of the data. Thus, in the rolling window analysis, the first forecast is made for the fourth week of 2013.

Prior to running the forecasting exercise, we examine the in-sample analysis for the ten countries and, in particular, observe the movement of the posterior mean of the degrees of freedom using our loss-based hyperprior. Again, the estimation of the degrees of freedom is considered for the same period as the forecasts. As stated above, we have run a rolling window estimation and, in Figure 3.5.2, we report the posterior means of the degrees of freedom with our hyperprior (solid black line), the 95% HPD in dotted lines, and the fixed value of ν equal to 11 (red dashed line), which is used in the benchmark model. The estimated number of degrees of freedom has some fluctuations at the beginning of the forecasting period, then decreases and remains stationary.

Moving to the forecasting exercise, we evaluate the forecasting accuracy using a point and a density forecasting measure. Table 4 shows the results for each country, and except for Argentina, the proposed loss-based hyperprior leads to better results in terms of RMSE compared to the benchmark prior in all the countries. The improvements are greater when we look at the average CRPS for

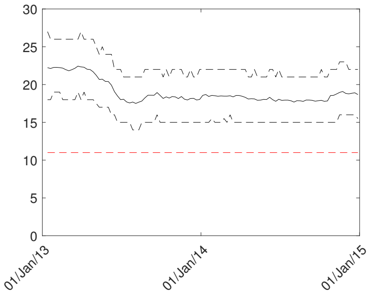


Figure 3.5.2: Estimated degrees of freedom (solid line) for the loss based hyperprior by using a rolling window of 104 weekly data with the 95% HPD (dotted lines) and the fixed ν (red dashed line) for the Dengue data.

each country, ranging from 1% for Argentina to 3% for India and Brazil.

	RMSE			CRPS		
	fixed ν	$\nu \sim \pi(\nu)$	ratio	fixed ν	$\nu \sim \pi(\nu)$	ratio
Argentina	0.328	0.329	1.002	0.174	0.172	0.992
Bolivia	0.228	0.227	0.993	0.145	0.144	0.989
Brazil	0.279	0.277	0.991	0.152	0.149	0.979
India	0.226	0.224	0.992	0.147	0.144	0.977
Indonesia	0.235	0.233	0.994	0.140	0.138	0.986
Mexico	0.273	0.271	0.994	0.150	0.148	0.989
Philippines	0.231	0.230	0.996	0.134	0.131	0.979
Singapore	0.687	0.685	0.997	0.326	0.323	0.991
Thailand	0.372	0.372	0.997	0.189	0.187	0.987
Venezuela	0.291	0.289	0.992	0.161	0.160	0.991

Table 4: RMSE and average CRPS for the GDT data for each prior. The first column refers to the variable, the second to the Gibbs sampler with fixed ν and the third with our loss-based hyperprior.

3.6 Discussions

We have presented a novel method to perform forecasting in Bayesian VAR models, where a hyperprior is assigned to the number of degrees of freedom of the covariance matrix for the normal-Wishart prior. In particular, our hyperprior can be considered objective, in the sense that it takes into consideration only the intrinsic properties of the model, i.e. the sampling distribution plus the prior. The method has been compared with what is currently used in the literature when no information about the parameter values of the normal-Wishart prior is available, which is by setting $\nu = m + 1$.

The analysis of simulated data has shown that when the true value of ν is close to $m+1$, both approaches perform in a similar way. However, as one would expect, the farther the true ν is from $m + 1$, the better the performance of the proposed method is. To illustrate in practice the advantage of having a loss-based prior on ν, we have analysed its performance in terms of prediction on two datasets. One concerns macroeconomic variables from the FRED dataset, and the other the analysis of Dengue fever data. For both datasets, it appears that the proposed method outperforms the one currently widely used, particularly when the CRPS index is considered.

In support of our results, we have estimated the number of degrees of freedom on rolling windows. This analysis has shown that the data contain information for a value of ν always above the value of $m + 1$, which justifies the use of the proposed method. In other words, as the data appear to have been "generated" by a Bayesian model with a relatively large number of degrees of freedom, by setting $\nu = m + 1$, one impacts the predictive performance of the model. On the other hand, by assuming uncertainty in the value of ν, i.e. assigning an objective prior to it, the model is free to "choose" the most appropriate value of the parameter and, even considering the extra uncertainty that this implies, the results are better than the previous method.

Chapter 4

A generalised Laplace spike-and-slab model for Bayesian variable selection

4.1 Introduction

In this chapter, we focus on the problem of estimating a sparse regression coefficient vector in a high-dimensional setting. Dealing with such a problem is of central interest in statistics, machine learning, and related fields.

Consider the classical multiple linear regression model where we assume that a vector of responses $\mathbf{y} = (y_1, \dots, y_N)^\top$ is modelled as a linear function of D predictor variables that form the columns of a design matrix $\mathbf{X}$ of dimensions $N \times D$, as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.1.1}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_D)^\top$ is a D-dimensional vector of coefficients, and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_N)^\top$ is an N-dimensional vector of errors, assumed to be generated from a normal distribution $\mathcal{N}(0, \sigma^2 \mathbf{I}_N)$. We assume that the response and the predictors are zero-centred, so that we do not need to include an intercept term in (4.1.1), and that the predictors are also standardised to have unit variance.

The task is to estimate the vector of coefficients $\boldsymbol{\beta}$. The problem becomes more challenging when D is large and at the same time only a small proportion of β_js are expected to be non-zero. Our goal is to find a sparse solution in which the majority of coefficients are exactly zero or very close to it, and to successfully identify those features that can truly explain the response variable $\mathbf{y}$ and correspond to non-zero coefficients.

There are two main approaches to the variable selection problem. One approach is through likelihood penalisation, where we minimise a function that comprises the model's log-likelihood plus a penalty (or regularisation) term on $\boldsymbol{\beta}$, with respect to the regression parameters (see, for example, Tibshirani (1996, 2011); Zou and Hastie (2005)). The other approach to this problem is through Bayesian methods, which is the scope of this chapter.

Bayesian variable selection is performed using sparsity-inducing priors. In Section 2.2, we discussed shrinkage priors, as well as discrete and continuous spike-and-slab priors. Discrete spike-and-slab priors are methodologically more ideal for sparse problems, since they put a probability mass on $\beta_j = 0$, but they pose significant computational challenges for standard simulation-based inference methods when D is too large. Shrinkage priors are more convenient computationally — most of them can be expressed as scale mixtures of normals (see Van Erp et al. (2019) for a review of shrinkage priors), and efficient Gibbs samplers can be constructed. However, the probability $p(\beta_j = 0)$ is never positive and a truly sparse solution can be achieved by ad hoc treatment such as the credible interval criterion or the scaled neighbourhood criterion (Li and Lin, 2010).

In this chapter, we examine the use of a hybrid prior specification, that uses a hierarchical shrinkage prior as a slab in a discrete spike-and-slab prior for the model coefficients of the linear model in (4.1.1). The hierarchical prior, in particular, is a normal-Gamma prior, and the overall prior can be expressed as:

$$\beta_j, \tau_j | z_j \sim \Big\{ z_j \mathcal{N}(0, \tau_j) + (1 - z_j)\delta_0(\beta_j) \Big\} \mathrm{Ga}(\tau_j | \lambda, 1/2\gamma^2), \tag{4.1.2}$$

where z_j is a binary variable that is 1 if the jth feature is included in the model and 0 otherwise, and $\delta_0(\beta_j)$ is a point mass at zero. Brown and Griffin (2010) proposed the normal-gamma as a shrinkage prior in linear regression and briefly discussed its potential use as a slab distribution in a discrete spike-and-slab prior. However, they did not pursue this line of research further. The marginal prior that is induced on the coefficients that are in a slab has the following density function:

$$\begin{aligned} \mathrm{NG}(\beta_j | \lambda, \gamma) &= \int \mathcal{N}(0, \tau_j) \mathrm{Ga}(\tau_j | \lambda, 1/2\gamma^2) d\tau_j \\ &= \frac{\gamma^{-\lambda - 1/2}}{\sqrt{\pi} 2^{\lambda - 1/2} \Gamma(\lambda)} |\beta_j|^{\lambda - 1/2} \mathcal{K}_{\lambda - 1/2}(|\beta_j| / \gamma). \end{aligned} \tag{4.1.3}$$

This distribution was introduced as the variance-gamma distribution in the financial literature (Madan and Seneta, 1990) and it is also known as the generalised Laplace or Bessel function distribution. Its negative log-density resembles that of hyperlasso (see Figure 2.2.1b). If the mixing distribution, in this Gaussian scale mixture, is a $\mathrm{Ga}(\tau_j | 1, \gamma^2/2)$ distribution, then the marginal prior is equivalent to the Laplace or double exponential distribution. $\mathcal{K}$ is the modified Bessel function of the second kind, and by using equations 10.27.3 and 10.30.2 of Olver et al. (2010), we can see that the density at 0 is $\frac{\Gamma(|\lambda - 0.5|)}{2\sqrt{\pi}\Gamma(\lambda)}$. Brown and Griffin (2010) specify priors for λ and γ, in which case, the marginal prior on β_j is no longer the one in (4.1.3). In our method, we compute empirical Bayes estimates of λ and γ instead of assigning priors to them.

MCMC methods for computing the posterior distribution corresponding to a spike-and-slab prior can be highly inefficient when D is large enough, due to the high number of possible submodels they have to search through (Castillo et al., 2015).

In our approach, we conduct posterior inference using variational methods (see, for example, Jordan et al., 1999; Beal, 2003; Blei et al., 2017). Carbonetto and Stephens (2012) were among the first to apply variational inference to a model with a spike-and-slab prior; they assumed a common prior variance τ for all β_js and sampled from its posterior using importance sampling. We instead assume an individual τ_j for each β_j and approximate each posterior distribution via a variational approximation. The resulting algorithm, based on the aforementioned prior setup, has all variational updates in closed form which makes it very competitive in terms of speed. It is also very accurate compared to other similar algorithms that have recently appeared in the literature. Empirical Bayes estimates for the three global parameters (λ, γ and the prior inclusion probability ρ: $z_j \sim \text{Bern}(\rho)$) can also be easily obtained via a variational-EM scheme (see, for example, Blei et al. (2003) for a variational-EM scheme applied to a topic model, or Braun and McAuliffe (2010) for an example of variational EM in a multi-level discrete choice modelling problem).

The chapter is structured as follows. Section 4.2 describes the optimal variational posterior densities that approximate the posterior when a normal-gamma prior is used, the modifications required when the prior in (4.1.2) is used, as well as the variational objective function and the resulting coordinate ascent algorithm. Section 4.3 describes an initialisation scheme. Section 4.4 presents some simulation studies and comparisons with competitive methods. Section 4.5 presents an application on a high-dimensional gene expression dataset. Section 4.6 provides extensions of our method to two generalised linear models, along with additional simulated and real data applications. Section 4.7 concludes the chapter.

4.1.1 Notation

We symbolise vectors with boldface lower-case letters (e.g. $\mathbf{y}$), matrices with boldface upper-case letters (e.g. $\mathbf{X}$), and scalars by either lower-case or upper-case letters (e.g. d, D). Angled brackets denote the expectation of the random variable in their argument with respect to the variational posterior density, e.g. $\langle x \rangle = \mu$ if $q(x)$ is a normal distribution $\mathcal{N}(\mu, \sigma^2)$. $\|\cdot\|_2$ denotes the L_2 norm, e.g. $\|\boldsymbol{\beta}\|_2 = \left(\sum_{j=1}^{D} |\beta_j|^2\right)^{\frac{1}{2}}$. The operator $\mathrm{diag}(\cdot)$ denotes both a vector-to-matrix and a matrix-to-vector operator depending on the argument; if the argument is a vector, then it denotes a diagonal matrix with diagonal entries the elements of the argument, and if the argument is a matrix, then it denotes a column vector containing the diagonal entries of the argument. Finally, the symbols $\odot$ and $\oslash$ denote the Hadamard (also known as element-wise) product and division, respectively, of two matrices with the same dimensions.

4.2 Variational posterior approximations

4.2.1 Model with normal-gamma prior

We first examine the optimal variational densities for the case where a normal-gamma prior is employed, before we move into the spike-and-slab model. In this case, all β_js are modelled jointly with a multivariate normal prior with diagonal covariance:

$$\begin{aligned}
&\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N)\\
&\boldsymbol{\beta} \sim \mathcal{N}(0, \boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_D)^\top))\\
&\tau_j \sim \mathrm{Ga}(\lambda, 1/(2\gamma^2))\\
&\sigma^2 \sim \mathrm{IG}(c_0, d_0)
\end{aligned} \tag{4.2.1}$$

This hierarchical model is depicted in a directed acyclic graph, in Figure 4.2.1.

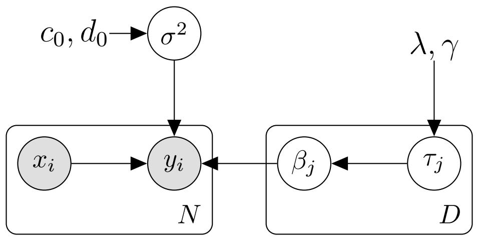


Figure 4.2.1: Graphical representation of the hierarchical model in (4.2.1).

The joint distribution of all variables is given by

$$p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma^2|\mathbf{X}) = p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)p(\sigma^2)p(\boldsymbol{\beta}|\boldsymbol{\Lambda})\prod_{j=1}^{D} p(\tau_j) \tag{4.2.2}$$

and we approximate the posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma^2|\mathbf{y}, \mathbf{X})$ with the variational posterior distribution given in the factorised form:

$$q(\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma^2) = q(\sigma^2)q(\boldsymbol{\beta})\prod_{j=1}^{D} q(\tau_j).$$

For each factor, we apply Eq. (2.3.5), i.e. we take the logarithm of the joint distribution over all variables and then take the expectation of it with respect to the variables not in that factor in order to obtain the optimal variational densities. Starting by τ_j:

$$\begin{aligned}\log q(\tau_j) &= \log p(\tau_j) + \langle \log p(\beta_j|\tau_j) \rangle + \text{const} \\ &= (\lambda - 1)\log(\tau_j) - \frac{1}{2\gamma^2}\tau_j - \frac{1}{2}\log \tau_j - \frac{1}{2}\frac{1}{\tau_j}\langle \beta_j^2 \rangle + \text{const},\end{aligned}$$

where $p(\beta_j|\tau_j)$ is a marginal of $p(\boldsymbol{\beta}|\boldsymbol{\Lambda})$ i.e. is a zero-mean normal with variance τ_j, or equivalently:

$$q(\tau_j) \propto \tau_j^{\lambda-1} \exp^{-\frac{1}{2\gamma^2}\tau_j} \frac{1}{\sqrt{\tau_j}} \exp^{-\frac{1}{2}\frac{1}{\tau_j}\langle \beta_j^2 \rangle},$$

which is recognised as a generalised inverse-Gaussian distribution parametrised: $\mathrm{GIG}(\nu = \lambda - \frac{1}{2}, g = \frac{1}{\gamma^2}, h_j = \langle \beta_j^2 \rangle)$. In the same way, for $\boldsymbol{\beta}$:

$$\begin{aligned}\log q(\boldsymbol{\beta}) &= \langle \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) \rangle + \langle \log p(\boldsymbol{\beta}|\boldsymbol{\Lambda}) \rangle + \text{const} \\ &= \left\{ -\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\langle \sigma^{-2} \rangle \mathbf{I})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} + \left\{ -\frac{1}{2}\boldsymbol{\beta}^\top \langle \boldsymbol{\Lambda}^{-1} \rangle \boldsymbol{\beta} \right\} + \text{const},\end{aligned}$$

which results in a normal distribution after completing the square:

$$q(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

with

$$\boldsymbol{\mu} = \langle \sigma^{-2} \rangle \big(\langle \boldsymbol{\Lambda}^{-1} \rangle + \langle \sigma^{-2} \rangle \mathbf{X}^\top \mathbf{X} \big)^{-1} \mathbf{X}^\top \mathbf{y},$$

$$\boldsymbol{\Sigma} = \big(\langle \boldsymbol{\Lambda}^{-1} \rangle + \langle \sigma^{-2} \rangle \mathbf{X}^\top \mathbf{X} \big)^{-1},$$

and finally, it is $q(\sigma^2) = \mathrm{IG}(c, d)$ with

$$c = c_0 + \frac{N}{2},$$

$$d = d_0 + \frac{1}{2}\big\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\mu})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\mu}) + \mathrm{Tr}(\mathbf{X}^\top \mathbf{X} \boldsymbol{\Sigma}) \big\}.$$

The required moments are: $\langle \beta_j^2 \rangle = \mu_j^2 + \boldsymbol{\Sigma}_{jj}$, $\langle \sigma^{-2} \rangle = \frac{c}{d}$, and $\langle \boldsymbol{\Lambda}^{-1} \rangle = \mathrm{diag}(\langle \boldsymbol{\tau}^{-1} \rangle)$, where $\langle \boldsymbol{\tau}^{-1} \rangle = (\langle \tau_1^{-1} \rangle, \langle \tau_2^{-1} \rangle, \ldots, \langle \tau_D^{-1} \rangle)^\top$; all moments related to the GIG distribution can be found in Appendix B.1.

In summary, the optimal variational posteriors for $\boldsymbol{\beta}$, τ_j, and σ^2 are a multivariate normal, a generalised inverse-Gaussian, and an inverse-gamma respectively. These distributions, though with different parameters, appear as the conditional posteriors of $\boldsymbol{\beta}$, τ_j, and σ^2 in the Gibbs sampler used in Brown and Griffin (2010). The relationship between mean-field variational inference and Gibbs sampling is discussed in Blei and Jordan (2006) and also in Blei et al. (2017).

4.2.2 Model with hybrid prior

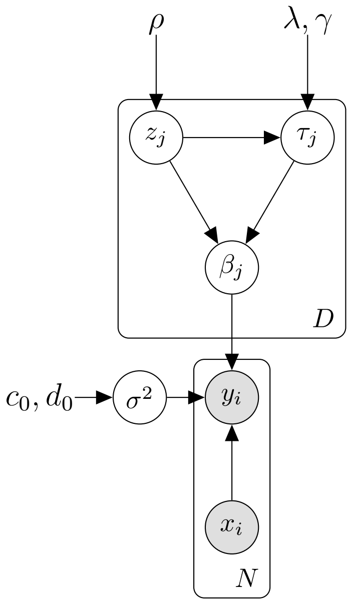


Figure 4.2.2: Graphical representation of the hierarchical model in (4.2.3) .

In order to fit the model with the spike-and-slab prior in (4.1.2), the following hierarchical model is used, in which each β_j is modelled independently, and a Bernoulli variable is introduced:

$$\begin{aligned}
&\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N) \\
&\beta_j|\tau_j, z_j \sim z_j\mathcal{N}(0, \tau_j) + (1 - z_j)\delta_0(\beta_j) \\
&\tau_j \sim \text{Ga}(\lambda, 1/(2\gamma^2)) \\
&z_j \sim \text{Bern}(\rho) \\
&\sigma^2 \sim \text{IG}(c_0, d_0)
\end{aligned} \tag{4.2.3}$$

with the joint distribution being:

$$p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma^2, \mathbf{z}|\mathbf{X}) = p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)p(\sigma^2)\prod_{j=1}^{D} p(\beta_j, \tau_j|z_j, \lambda, \gamma)\prod_{j=1}^{D} p(z_j|\rho). \tag{4.2.4}$$

In the spike-and-slab model, equations (2.3.5) and (2.3.6) for finding the optimal variational densities are not easily applicable due to the prior not being absolutely continuous, and we also need to avoid a multivariate normal variational posterior with the costly matrix operations associated with the covariance matrix. To this end, we select variational factors that resemble the ones in Section 4.2.1, and optimise the ELBO with respect to their individual variational parameters using standard calculus. Thus, the new factorised distribution has one additional set of factors for the binary variables:

$$q(\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{z}, \sigma^2) = q(\sigma^2) \prod_{j=1}^{D} q(\beta_j, \tau_j, z_j). \tag{4.2.5}$$

Similarly to Carbonetto and Stephens (2012), we approximate the joint posterior of β_j and z_j with a distribution that resembles the discrete spike-and-slab prior:

$$q(\beta_j, z_j) = \big(\alpha_j \mathcal{N}(\mu_j, s_j^2)\big)^{z_j} \big((1-\alpha_j)\delta_0(\beta_j)\big)^{1-z_j},$$

which implies that $q(z_j)$ is a Bernoulli:

$$q(z_j) = \begin{cases} \alpha_j & \text{if } z_j = 1, \\ 1-\alpha_j & \text{if } z_j = 0. \end{cases}$$

While in the model (4.2.3) all z_js share a global parameter ρ, which is the prior inclusion probability, in the variational posterior each z_j has its own individual parameter α_j that serves as the posterior inclusion probability. We use this posterior inclusion probability to conduct variable selection: if $\alpha_j > 0.5$ then predictor j is included in the model, otherwise it is not.

The variational density for the τ_j's retains the same form of a generalised inverse Gaussian distribution with parameters $\nu = \lambda - \frac{1}{2}, g = \frac{1}{\gamma^2}, h_j = \langle \beta_j^2 \rangle_{z_j=1}$, but now the last parameter is $h_j = \langle \beta_j^2 | z_j = 1 \rangle = \mu_j^2 + s_j^2$.

The density $q(\sigma^2)$ remains an inverse-gamma with parameters:

$$c = c_0 + \frac{N}{2},$$

and

$$d = d_0 + 0.5\big((\mathbf{y} - \mathbf{X}\boldsymbol{w})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{w}) + \mathrm{Tr}(\mathbf{X}^\top\mathbf{X}\boldsymbol{W})\big),$$

where $\boldsymbol{w}$ is a vector with entries $w_j = \langle\beta_j\rangle = \alpha_j\mu_j$ and $\boldsymbol{W} = \mathrm{cov}(\boldsymbol{\beta})$ is a diagonal matrix with diagonal entries $\mathrm{var}(\beta_j) = \alpha_j(\mu_j^2 + s_j^2) - (\alpha_j\mu_j)^2$.

For the calculation of the variational objective function in Section 4.2.3, it is convenient to have KL divergences between variational posteriors and priors analytically. Since $q(\beta_j, z_j) = q(\beta_j|z_j)q(z_j)$, we have the following:

$$\mathrm{KL}(q(\beta_j|z_j)||p(\beta_j|\tau_j, z_j)) = \frac{z_j}{2}\big(\log\tau_j - \log s_j^2 + \tau_j^{-1}(\mu_j^2 + s_j^2) - 1\big), \qquad (4.2.6)$$

which is a KL divergence between two normal distributions if $z_j = 1$, and 0 if $z_j = 0$ since in that case both distributions are identical (point masses at 0). Also useful is the KL divergence between two inverse-gamma distributions:

$$\mathrm{KL}(q(\sigma^2)||p(\sigma^2)) = (c - c_0)\psi(c) - \log\frac{\Gamma(c)}{\Gamma(c_0)} + c_0\log\frac{d}{d_0} + c\frac{d_0 - d}{d}. \qquad (4.2.7)$$

4.2.3 Variational lower bound

Following (2.3.2), we construct the ELBO, denoted by $\mathcal{L}$, which is the expectation of the logarithm of the joint distribution in (4.2.4) with respect to the variational distribution in (4.2.5). The ELBO breaks down to the expected log-likelihood plus negative KL divergences between the variational densities and the priors:

$$\begin{aligned}
\mathcal{L} &= \sum_{\mathbf{z}} \int\int\int q(\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{z}, \sigma^2) \log \frac{p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{z}, \sigma^2 | \mathbf{X})}{q(\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{z}, \sigma^2)} d\boldsymbol{\beta} d\boldsymbol{\tau} d\sigma^2 \\
&= \langle \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) \rangle - \sum_{j=1}^{D} \left[\langle \log q(\beta_j | z_j) \rangle - \langle \log p(\beta_j | \tau_j, z_j) \rangle \right] \\
&- \sum_{j=1}^{D} \left[\alpha_j \big(\langle \log q(\tau_j) \rangle - \langle \log p(\tau_j) \rangle \big) \right] - \langle \log q(\sigma^2) \rangle + \langle \log p(\sigma^2) \rangle \\
&- \sum_{j=1}^{D} \left[\langle \log q(z_j) \rangle + \langle \log p(z_j) \rangle \right],
\end{aligned}$$

and by inserting only the expectations that contain the variational parameters whose updates we do not have, namely μ_j, s_j^2, and α_j (terms $\boldsymbol{w}$ and $\boldsymbol{W}$ also contain μ_j, s_j^2, and α_j), we get:

$$\begin{aligned}
\mathcal{L} = &-\frac{1}{2} N \big(\log 2\pi + \langle \log \sigma^2 \rangle \big) - \frac{\langle \sigma^{-2} \rangle}{2} \big((\mathbf{y} - \mathbf{X}\boldsymbol{w})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{w}) + \mathrm{Tr}(\mathbf{X}^\top \mathbf{X} \boldsymbol{W}) \big) \\
&- \sum_{j=1}^{D} \frac{\alpha_j}{2} \big(\langle \log \tau_j \rangle - \log s_j^2 + \langle \tau_j^{-1} \rangle (\mu_j^2 + s_j^2) - 1 \big) \\
&- \sum_{j=1}^{D} \alpha_j \big(\langle \log q(\tau_j) \rangle - \langle \log p(\tau_j) \rangle \big) \\
&- \langle \log q(\sigma^2) \rangle + \langle \log p(\sigma^2) \rangle \\
&- \sum_{j=1}^{D} \alpha_j \log \frac{\alpha_j}{\rho} - \sum_{j=1}^{D} (1 - \alpha_j) \log \frac{1 - \alpha_j}{1 - \rho}.
\end{aligned} \tag{4.2.8}$$

The second line in (4.2.8) follows from (4.2.6), the third line contains the entropy of a GIG distribution and the negative cross-entropy of a gamma distribution relative to a GIG distribution, which can be found in Appendix B.1, the fourth line follows from (4.2.7), and the fifth line is a negative KL divergence between $q(z_j)$ and $p(z_j)$. The explicit form of the ELBO, with all expectations substituted by their analytic expressions, can be found in Appendix B.2.

4.2.4 Coordinate ascent variational inference

Optimising (4.2.8) with respect to μ_j, s_j^2, and α_j gives the respective updates for the three variational parameters. Derivations can be found in Appendix B.3. Algorithm 2 provides the detailed steps of the CAVI for obtaining the optimal approximations. The variational parameter c, which remains constant per iteration of CAVI, is computed once at the beginning of the initialisation process (see Section 4.3). In the E-step of this variational-EM scheme, we update the variational parameters μ_j ,s_j^2, α_j, h_j, and d. In the M-step, we compute empirical Bayes estimates of the global parameters λ, γ, and ρ (see Appendix B.4). After the M-step, the variational parameters ν and g must be updated since they depend on λ and γ, respectively. The algorithm requires input of λ, γ, ρ, c_0, and d_0. In all our examples we have used: $\lambda = 1$, $\gamma = 1/\sqrt{2}$ and $\rho = 0.05$, which serve as starting points rather than expressing any prior beliefs.

We can also optimise the ELBO with respect to c_0 and d_0, which are global parameters but do not govern the coefficients (see Figure 4.2.2), by adding the updates

$$c_0 \leftarrow \arg\max_{c_0} \big\{ -c_0\psi(c) + \log\Gamma(c_0) + c_0(\log d - \log d_0) \big\},$$

$$d_0 \leftarrow \frac{d}{c} c_0,$$

$$c \leftarrow c_0 + \frac{N}{2},$$

to the M-step. We conducted part of the experiments of this chapter using the aforementioned modification in the M-step, and noticed no substantial difference in the outcome while the algorithm ran for considerably more iterations until stopping. In particular, in the real data analysis of Section 4.5, the algorithm selected the same variables as it did without the updates for c_0 and d_0, while it ran for roughly four times more iterations. We also tested this modification on some scenarios of the simulations studies (see Section 4.4) and again we had longer computation times with no improvement in accuracy. Therefore, we chose not to include the two updates in the final version of our algorithm. Instead, we set $c_0 = d_0 = 0.01$ to induce a weak prior on σ^2.

Algorithm 2 CAVI for linear model with spike-and-slab - normal - gamma prior

```
    Input: Design matrix $\mathbf{X}$, response vector $\mathbf{y}$, parameters $\lambda$,$\gamma$,$\rho$,$c_0$,$d_0$.
    Output: Variational parameters: $g, c, d, \nu$, and $\mu_j, s_j^2, \alpha_j, h_j$, for $j = 1, \ldots, D$.
 1: Run Algorithm 3 for initialisation.
 2: $n \leftarrow 1$
 3: while $R_n >$ tol1 do
    Perform the E-step
 4:     for $j \in \{1, \ldots D\}$ do
 5:         $s_j^2 \leftarrow \big(\frac{c}{d}(\mathbf{X}^\top\mathbf{X})_{jj} + \langle\tau_j^{-1}\rangle\big)^{-1}$
 6:         $\mu_j \leftarrow \frac{c}{d}\big((\mathbf{X}^\top\mathbf{y})_j - \sum_{k\neq j}(\mathbf{X}^\top\mathbf{X})_{kj}\alpha_k\mu_k\big)s_j^2$
 7:         $h_j \leftarrow \mu_j^2 + s_j^2$
 8:         Update $\log\frac{\alpha_j}{1-\alpha_j}$ according to Eq. (4.2.9)
 9:     end for
10:     $d \leftarrow d_0 + 0.5\big((\mathbf{y} - \mathbf{X}\boldsymbol{w})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{w}) + \mathrm{Tr}(\mathbf{X}^\top\mathbf{X}\boldsymbol{W})\big)$
    Perform the M-step
11:     $\gamma \leftarrow \left(\frac{1}{2}\frac{\sum_{j=1}^D \alpha_j\langle\tau_j\rangle}{\lambda\sum_{j=1}^D \alpha_j}\right)^{\frac{1}{2}}$
12:     $\lambda \leftarrow \arg\max_\lambda\big\{-\sum_{j=1}^D \alpha_j\big(\log\Gamma(\lambda) + \lambda\log(2\gamma^2) - \lambda\log\sqrt{h_j/g}\big)\big\}$
13:     $\rho \leftarrow \frac{\sum_{j=1}^D \alpha_j}{D}$
    Update $\nu$ and $g$ and check for convergence
14:     $\nu \leftarrow \lambda - \frac{1}{2}$
15:     $g \leftarrow \frac{1}{\gamma^2}$
16:     Compute $R_n$ according to Eq. (4.3.2)
17:     $n \leftarrow n + 1$
18: end while
```

In line 8 of Algorithm 2, α_j is updated in terms of its log-odds:

$$\begin{aligned}\log\frac{\alpha_j}{1-\alpha_j} &= \log\frac{\rho}{1-\rho} + \frac{c}{d}\left((\mathbf{X}^\top\mathbf{y})_j\mu_j - (\mathbf{X}^\top\mathbf{X})_{jj}\frac{\mu_j^2 + s_j^2}{2} - \mu_j\sum_{k\neq j}(\mathbf{X}^\top\mathbf{X})_{kj}\alpha_k\mu_k\right)\\ &\quad - \frac{1}{2}\big(\langle\log\tau_j\rangle - \log s_j^2 + \langle\tau^{-1}\rangle(\mu_j^2 + s_j^2) - 1\big)\\ &\quad - \big(\langle\log q(\tau_j)\rangle - \langle\log p(\tau_j)\rangle\big).\end{aligned} \tag{4.2.9}$$

4.3 Initialisation and stopping

The main interest rests in the parameters $(\mu_j)_{j=1}^D$ since they determine the estimates for the regression coefficients. In Algorithm 2, these parameters are updated sequentially, and the update of each μ_j takes into account all other parameter values $(\mu_k)_{k \neq j}$. A potential issue can arise when D is very large and the non-zero coefficients are located towards the end of the coefficient vector. Consider an example in which $D = 10000$ and the first relevant predictor is at $j = 9000$, thus $\beta_{9000} \neq 0$. Algorithm 2 will update each μ_j starting from $j = 1$. Until it reaches the first relevant feature, it will have updated all prior μ_js having used an inaccurate value of μ_{9000}. Inevitably it will update μ_{9000} using poorly updated parameters $(\mu_j)_{j=1}^{8999}$. Eventually, it will update the remaining parameters based on a poorly updated μ_{9000}. In a situation like the one we just described, CAVI can get trapped in poor local optima and provide an inaccurate solution. We can alleviate this problem with the following initialisation procedure.

The objective function, ELBO, can be optimised with respect to the vector $\boldsymbol{\mu}$ instead of each μ_j. By expressing $\mathcal{L}$ in terms of $\boldsymbol{\mu}$ and keeping only the terms that contain it, we get:

$$\mathcal{L}_{[\boldsymbol{\mu}]} = -\langle \sigma^{-2} \rangle \frac{1}{2} \big((\mathbf{y} - \mathbf{X}\boldsymbol{w})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{w}) + \operatorname{Tr}(\mathbf{X}^\top \mathbf{X} \boldsymbol{W}) \big) - \frac{1}{2} \boldsymbol{\mu}^\top \mathbf{B} \boldsymbol{\mu},$$

where $\boldsymbol{w} = \mathbf{A}\boldsymbol{\mu}$, $\mathbf{B} = \mathbf{A}\langle \boldsymbol{\Lambda}^{-1} \rangle$ and $\mathbf{A}$ is a diagonal matrix with diagonal elements the variational parameters α_j. Then, the derivative with respect to $\boldsymbol{\mu}$ is:

$$\frac{\partial}{\partial \boldsymbol{\mu}} \mathcal{L}_{[\boldsymbol{\mu}]} = \langle \sigma^{-2} \rangle \big(\mathbf{A}\mathbf{X}^\top \mathbf{y} - \mathbf{A}\mathbf{X}^\top \mathbf{X} \mathbf{A} \boldsymbol{\mu} - \mathbf{X}^\top \mathbf{X} \odot \mathbf{A}(\mathbf{I} - \mathbf{A}) \boldsymbol{\mu} \big) - \mathbf{B}\boldsymbol{\mu},$$

where we have used that: $\operatorname{Tr}(\mathbf{X}^\top \mathbf{X} \mathbf{A} \odot \boldsymbol{\mu}\boldsymbol{\mu}^\top \odot (\mathbf{I} - \mathbf{A})) = \boldsymbol{\mu}^\top \mathbf{X}^\top \mathbf{X} \odot \mathbf{A}(\mathbf{I} - \mathbf{A})\boldsymbol{\mu}$, which is the part of the term $\operatorname{Tr}(\mathbf{X}^\top \mathbf{X} \boldsymbol{W})$ that depends on $\boldsymbol{\mu}$. Setting this derivative equal to 0 and solving for $\boldsymbol{\mu}$ gives the update:

$$\boldsymbol{\mu} = \langle\sigma^{-2}\rangle\big(\langle\sigma^{-2}\rangle\mathbf{A}\mathbf{X}^\top\mathbf{X}\mathbf{A} + \langle\sigma^{-2}\rangle\mathbf{X}^\top\mathbf{X}\odot\mathbf{A}(\mathbf{I}-\mathbf{A}) + \mathbf{B}\big)^{-1}\mathbf{A}\mathbf{X}^\top\mathbf{y}. \qquad (4.3.1)$$

The matrix inversion in (4.3.1) can be done via the *Woodbury identity* (Woodbury, 1950).

This update can be used in a preliminary CAVI (Algorithm 3) that provides an initialisation of the variational parameters, which are then passed into Algorithm 2. In Algorithm 3, we also introduce the vectors $\mathbf{h} = (h_1, \dots, h_d)^\top$ and $\mathbf{s} = (s_1^2, \dots, s_d^2)^\top$ to facilitate vectorised updates for the parameters included. The transition from Algorithm 3 to 2 is determined by the relative change in $\mathcal{L}$ at nth iteration:

$$R_n = \frac{|\mathcal{L}_n - \mathcal{L}_{n-1}|}{|\mathcal{L}_n|}, \qquad (4.3.2)$$

which is also the criterion to monitor convergence in Algorithm 2. When R_n falls below a tolerance level tol2, Algorithm 3 stops and we transition to Algorithm 2, and when R_n falls below a tolerance level tol1, we declare convergence and Algorithm 2 stops. Obviously, tol2 $\gg$ tol1. It is important to choose tol2 so that Algorithm 3 climbs the ELBO high enough but at the same time does not spend too much time doing the costly updates of (4.3.1). Empirically, through the simulation studies we conducted, we have found that setting tol2 $= 0.1$ serves this purpose well. We also set tol1 $= 10^{-5}$; with this choice, the process does not stop too early and does not waste iterations either.

Figure 4.3.1 demonstrates the initialisation scheme on the high-dimensional Bardet–Biedl Syndrome dataset, which we analyse in Section 4.5, by showing how the ELBO and R_n evolve until convergence is reached. In this example, initialisation takes 5 iterations, and then the updates in Algorithm 2 are used until convergence.

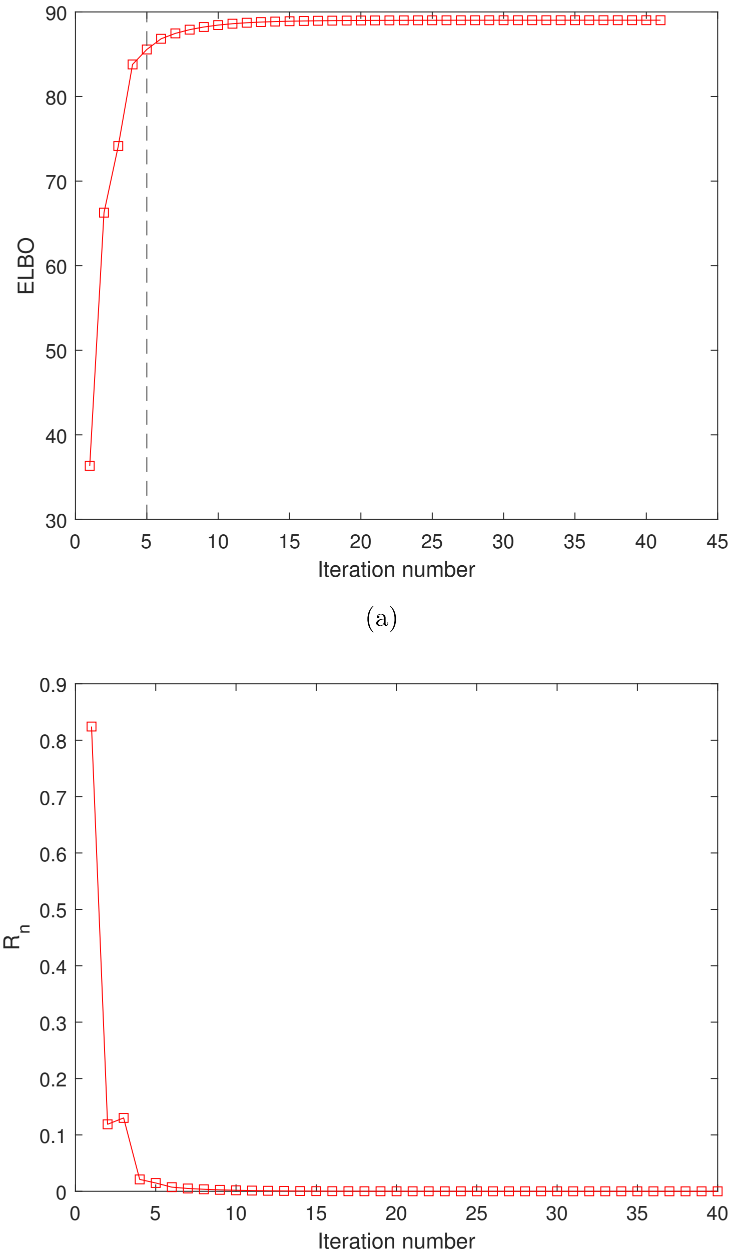
It can be reasonable to consider using Algorithm 3 solely. However, the transitioning scheme is more preferable for two reasons. Firstly, it has shown better performance in both simulated and real data applications. Secondly, in a $D \gg N$ scenario, when both D and N are large (or even when just D is large), the matrix inversion in (4.3.1) can be time-demanding even via the Woodbury identity, which only needs inversion of an $N \times N$ matrix, as it requires many matrix multiplications.

Regarding the initial values of the variational parameters, we set all α_js equal to 1, i.e. $\mathbf{A} = \mathbf{I}$; other values we tried caused the matrix that is inverted in (4.3.1) to be singular. We also set all μ_js equal to 0 as we expect most model coefficients to be 0. The two choices are contradictory, as $\alpha_j = 1$ means that the jth covariate should be included in the model and $\mu_j = 0$ means that its coefficient is 0. However, convergence is not affected by these choices. Lastly, we use $(s_j^2)_{j=1}^D = 0.1$; practically, any small number is suitable since these variational parameters will almost certainly become even smaller than this initial choice.

For the remainder of the thesis, we refer to Algorithm 2 as SSNG, after the initials of the two priors: spike-and-slab - normal-gamma. As mentioned in Section 4.1, the input data $(\mathbf{y}, \mathbf{X})$ are transformed as follows: $\mathbf{y}$ is zero-centred, and each column in $\mathbf{X}$ is zero-centred and also standardised to have unit variance. After SSNG has converged, we calculate the estimate for the coefficient vector as: $\hat{\boldsymbol{\beta}} = (\alpha_j \mu_j / \sigma_{\mathbf{x}_j})_{j=1}^D$, where $\sigma_{\mathbf{x}_j}$ is the standard deviation of the jth column of $\mathbf{X}$ prior to standardisation. If we want to calculate the fitted values of the model or to make out-of-sample predictions, we need to insert an intercept term in $\boldsymbol{\beta}$ and add a column of ones to the matrix of the predictors. The intercept term is created as: $\beta_0 = \bar{y} - \bar{\mathbf{x}}^\top \boldsymbol{\beta}$, where $\bar{y}$ is the mean of $\mathbf{y}$ and $\bar{\mathbf{x}}$ is a $D \times 1$ vector of the means of the columns of $\mathbf{X}$ prior to them being zero-centred.

Algorithm 3 CAVI for initial iterations

```
**Input:** Design matrix $\mathbf{X}$, response vector $\mathbf{y}$.
**Output:** Variational parameters: $g, c, d, \nu$, and $\mu_j, s_j^2, \alpha_j, h_j$, for $j = 1, \ldots, D$.
$c \leftarrow c_0 + \frac{N}{2}$
$\nu \leftarrow \lambda - \frac{1}{2}$
$g \leftarrow \frac{1}{\gamma^2}$
$n \leftarrow 1$
Perform the E-step
while $R_n >$ tol2 do
    $\boldsymbol{\mu} \leftarrow \langle \sigma^{-2} \rangle \big( \langle \sigma^{-2} \rangle \mathbf{A}\mathbf{X}^\top \mathbf{X}\mathbf{A} + \langle \sigma^{-2} \rangle \mathbf{X}^\top \mathbf{X} \odot \mathbf{A}(\mathbf{I} - \mathbf{A}) + \mathbf{B} \big)^{-1} \mathbf{A}\mathbf{X}^\top \mathbf{y}$
    $\mathbf{s} \leftarrow \mathbf{1} \oslash \big( \langle \sigma^{-2} \rangle \mathrm{diag}(\mathbf{X}^\top \mathbf{X}) + \langle \boldsymbol{\tau}^{-1} \rangle \big)$
    $\mathbf{h} \leftarrow \boldsymbol{\mu} \odot \boldsymbol{\mu} + \mathbf{s}$
    for $j \in \{1, ...D\}$ do
        Update $\log \frac{\alpha_j}{1-\alpha_j}$ according to Eq. 4.2.9
    end for
    $d \leftarrow d_0 + 0.5\big( (\mathbf{y} - \mathbf{X}\boldsymbol{w})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{w}) + \mathrm{Tr}(\mathbf{X}^\top \mathbf{X}\boldsymbol{W}) \big)$
Perform the M-step
    $\gamma \leftarrow \left( \frac{1}{2} \frac{\sum_{j=1}^{D} \alpha_j \langle \tau_j \rangle}{\lambda \sum_{j=1}^{D} \alpha_j} \right)^{\frac{1}{2}}$
    $\lambda \leftarrow \arg\max_\lambda \big\{ -\sum_{j=1}^{D} \alpha_j \big( \log \Gamma(\lambda) + \lambda \log(2\gamma^2) - \lambda \log \sqrt{h_j/g} \big) \big\}$
    $\rho \leftarrow \frac{\sum_{j=1}^{D} \alpha_j}{D}$
Update $\nu$ and $g$ and check for convergence
    $\nu \leftarrow \lambda - \frac{1}{2}$
    $g \leftarrow \frac{1}{\gamma^2}$
    Compute $R_n$ according to Eq. (4.3.2)
    $n \leftarrow n + 1$
end while
```



(a)

(b)

Figure 4.3.1: (a) ELBO vs iteration for the Bardet–Biedl Syndrome data analysis. The vertical dashed line indicates the fifth iteration, on which R_n falls below tol2. (b) Relative change vs iteration for the same analysis.

4.4 Simulation studies

In this section, we evaluate the proposed algorithm in three scenarios of simulated data and compare it against three competitive methods. The three scenarios, all with $D \gg N$, differ in how the non-zero coefficients are located in the signal. In the first case, the non-zero coefficients are spread randomly across the signal; in the second case, they form mini-batches that are spread across the signal; and in the third case, they form a single batch randomly located in the signal. They also differ in other characteristics such as correlation among covariates and error variance.

The competitors fit linear regression models with spike-and-slab priors. However, they conduct inference using different algorithms and employ different slab distributions.

- **varbvs:** An R package (Carbonetto et al., 2017) that implements the methods in Carbonetto and Stephens (2012). It employs a discrete spike-and-slab prior with a normal slab and combines variational inference with importance sampling.

- **sparsevb:** An R package that performs variational inference on a discrete spike-and-slab model with Laplace slabs (Ray and Szabó, 2022). It addresses the ordering issue explained in Section 4.3 by obtaining an initial estimation of the coefficients using the glmnet package (Friedman et al., 2010), based on which it determines the updating order.

- **SSLASSO:** An R package that implements the methods in Ročková and George (2018). It utilises expectation-maximisation and coordinate-wise optimisation. Sparsity is induced by a continuous spike-and-slab prior with both the spike and the slab being Laplace distributions. We set the variance argument as “unknown”, so the error variance is considered unknown and estimated from the data. This version of SSLASSO is described in Moran et al. (2019).

For each of the scenarios, 100 datasets are simulated from the model in (4.1.1). Design matrices are drawn from a multivariate normal distribution with a zero mean and a covariance matrix constructed to match the correlation requirement of each scenario. Each competitive method is fitted to each dataset and a number of statistics are calculated. We calculate the false discovery rate:

$$\text{FDR} = \frac{\text{FP}}{\text{FP} + \text{TP}},$$

and the true positive rate:

$$\text{TPR} = \frac{\text{TP}}{\text{P}},$$

where FP is the number of false positives, TP is the number of true positives, and P is the total number of real positives. The Euclidean distance (or L^2 norm) between real and estimated model coefficients is calculated as:

$$L_2 = \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2 = \left(\sum_{j=1}^{D} (\beta_j - \hat{\beta}_j)^2\right)^{\frac{1}{2}},$$

and the root mean square prediction error between fitted values and the response is:

$$\text{RMSPE} = \left(\frac{1}{N}\sum_{i=1}^{N} (\hat{\boldsymbol{\beta}}^\top \boldsymbol{x}_i - y_i)^2\right)^{\frac{1}{2}},$$

where $\mathbf{x}_i$ is a column vector with the entries of the ith row of $\mathbf{X}$. We also count the time in seconds that each method takes per dataset. Comparisons in time can be considered to be somewhat biased since our algorithm is implemented in Matlab, while the competitors are implemented in R using the powerful Rcpp interface (Eddelbuettel and François, 2011).

4.4.1 Scenario 1

In the first scenario, there are $N = 200$ samples and $D = 800$ covariates. There are 20 non-zero coefficients, which are randomly located in the coefficient vector and are all equal to 10. The error variance is $\sigma^2 = 25$, and the covariates are weakly correlated with a Pearson correlation coefficient of $r = 0.3$. Table 5 reports the average statistics over 100 simulated datasets, with parentheses showing the respective standard deviations.

Method	FDR	TPR	L_2	RMSPE	Time
SSNG	0.00(0.01)	1	2.11(0.39)	4.62(0.23)	0.48(0.09)
varbvs	0.00(0.01)	1	2.11(0.38)	4.65(0.24)	5.16(0.89)
sparsevb	0.04(0.04)	1	2.69(0.55)	4.70(0.22)	2.25(0.68)
SSLASSO	0.00(0.01)	0.56(0.05)	36.80(3.40)	29.35(2.73)	0.03(0.01)

Table 5: Comparison statistics for the first scenario.

SSNG and varbvs perform very similarly, both in terms of variable selection and parameter estimation, while sparsevb's performance is slightly inferior. The three methods achieve a perfect TPR score in all 100 simulated datasets. SSNG and varbvs have a near-zero FDR, while sparsevb has a slightly higher one at 0.04. Of the three methods, SSNG is by far the fastest, taking an average of 0.48 seconds per dataset.

Despite this being a trivial example, with the non-zero coefficients being very distant from zero, SSLASSO performs considerably worse than the other three methods. It is the fastest among the competitors; however, its L_2 and RMSPE statistics are too high. It also exhibits a near-zero FDR, like SSNG and varbvs, but this comes together with a TPR of 0.56, which means that SSLASSO is very conservative in including variables in the model. Possible reasons behind SSLASSO's much inferior performance to the other methods are the high error variance ($\sigma^2 = 25$) and SSLASSO's sensitivity to initialisation (i.e. its default settings are not suitable for this scenario).

4.4.2 Scenario 2

This example is a modified version of the one used in Ročková and George (2014). In the initial example ($N = 100, D = 1000$), the data are generated with a vector of coefficients specified as $\boldsymbol{\beta} = (3, 2, 1, 0, 0, \dots, 0)^\top$. To make the example more challenging, the batch of non-zero coefficients $(3, 2, 1)$ appears 10 times at random locations, resulting in a total of 30 non-zero coefficients.. Similarly to the initial example, predictor values for each observation are simulated from $\mathcal{N}(0, \Sigma)$, where $\Sigma = (r_{ij})_{i,j=1}^{D}$ with $r_{ij} = 0.6^{|i-j|}$. The error variance is $\sigma^2 = 3$. Table 6 reports the averaged statistics over the 100 datasets generated for this scenario.

Method	FDR	TPR	L_2	RMSPE	Time
SSNG	0.02(0.03)	0.69(0.11)	4.76(1.59)	2.22(0.69)	0.47(0.05)
varbvs	0.07(0.08)	0.32(0.10)	10.61(1.30)	5.28(1.65)	1.17(0.27)
sparsevb	0.33(0.18)	0.51(0.13)	9.42(5.11)	2.16(0.85)	1.75(3.39)
SSLASSO	0.02(0.10)	0.32(0.13)	9.90(1.20)	5.62(1.53)	0.16(0.03)

Table 6: Comparison statistics for the second scenario.

This is a far more challenging example than the previous one. The main difficulty lies in identifying the coefficients equal to 1, since they are adjacent to higher values and neighbouring variables are correlated. Again, SSNG performs best overall; it has the lowest L_2 and the highest TPR, and it ties with SSLASSO for the lowest FDR, though with lower standard deviation. Of the four competitors, sparsevb has the lowest RMSPE, which is attributed to the many false discoveries it makes (FDR=0.33). Lastly, varbvs and SSLASSO perform very similarly, with SSLASSO being the fastest overall.

4.4.3 Scenario 3

In the third scenario, the non-zero elements of the coefficient vector form a single batch: $\{3, \dots, 3, 2.5, \dots, 2.5, 2, \dots, 2, 1.5, \dots, 1.5, 1, \dots, 1\}$. Each value appears 4 times, resulting in a total of 20 non-zero coefficients. The design matrix is generated from a standard normal distribution with dimensions $N = 100$, $D = 600$. The error variance is $\sigma^2 = 0.25$. We conduct the experiment three times, using different correlation coefficients among the covariates: 0, 0.4 and 0.8.

This simulation study is specifically designed to demonstrate the value of the initialisation process of SSNG. Of the four competitors, only SSNG and sparsevb have some sort of mechanism to deal with the issue discussed in Section 4.3, when all non-zero coefficients are gathered in one location of the coefficient vector. When there is no correlation (Table 7), SSNG and varbvs perform comparably, and sparsevb follows closely. In the case of mild correlation ($r = 0.4$, Table 8), SSNG retains the highly polarised FDR and TPR scores and remains superior overall in parameter estimation and prediction, while sparsevb outperforms varbvs in all aspects. In the extreme case where all covariates are highly correlated ($r = 0.8$, Table 9), we see that the two methods that don't take into account the ordering of the updates, i.e. varbvs and SSLASSO, show very poor performance. The two methods are heavily outperformed by sparsevb and SSNG, with the latter, once again, providing the best results overall.

Method	FDR	TPR	L_2	RMSPE	Time
SSNG	0.00(0.01)	0.99(0.08)	0.33(0.78)	0.46(0.49)	0.17(0.02)
varbvs	0.00(0.03)	0.96(0.17)	0.59(1.65)	0.64(1.12)	0.40(0.11)
sparsevb	0.18(0.29)	0.91(0.13)	1.97(2.46)	0.57(0.39)	0.47(0.15)
SSLASSO	0	0.71(0.07)	3.08(0.69)	2.40(0.46)	0.35(0.08)

Table 7: Comparison statistics for the third scenario($r = 0$).

Method	FDR	TPR	L_2	RMSPE	Time
SSNG	0.00(0.08)	0.99(0.01)	0.33(0.28)	0.45(0.07)	0.18(0.02)
varbvs	0.31(0.30)	0.63(0.35)	5.47(4.96)	1.78(1.32)	4.28(1.14)
sparsevb	0.28(0.31)	0.80(0.20)	3.85(3.72)	0.79(0.34)	0.92(0.20)
SSLASSO	0.28(0.21)	0.21(0.06)	12.16(1.51)	7.38(0.82)	0.07(0.01)

Table 8: Comparison statistics for the third scenario($r = 0.4$).

Method	FDR	TPR	L_2	RMSPE	Time
SSNG	0.22(0.12)	0.87(0.07)	2.45(1.11)	0.68(0.18)	0.25(0.06)
varbvs	0.74(0.15)	0.08(0.06)	16.58(2.32)	4.74(0.90)	10.66(3.25)
sparsevb	0.43(0.18)	0.43(0.11)	8.22(1.91)	1.85(0.29)	1.51(0.30)
SSLASSO	0.52(0.28)	0.11(0.06)	17.75(1.96)	5.65(0.76)	0.03(0.00)

Table 9: Comparison statistics for the third scenario($r = 0.8$).

4.5 A real data application

We further assess the SSNG algorithm by analysing a microarray dataset of gene expression measurements from the eye tissue of 120 laboratory rats (*Rattus norvegicus*). The aim of this analysis is to identify genes that are associated with the gene TRIM32 (Tripartite motif-containing protein 32). TRIM32 has been identified as a gene that causes the Bardet–Biedl syndrome (BBS), a pleiotropic, autosomal recessive disorder characterised by obesity, pigmentary retinopathy, polydactyly, renal abnormalities, learning disabilities, and hypogenitalism (Chiang et al., 2006).

The complete dataset, which consists of 31,099 probe sets, has also been used in Bai et al. (2020) in order to compare the SSLASSO methodology with the LASSO penalty, the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), and the minimax concave penalty (MCP) (Zhang, 2010). We perform the analysis as in Bai et al. (2020) for a fair comparison. We include the 10,000 probe sets whose logarithms exhibit the highest variances, and thus, $N = 120$ and $D = 10,000$. On this reduced dataset, we run SSNG and also assess predictive accuracy via cross-validation. We randomly split the data into a training dataset $(\mathbf{y}_{train}, \mathbf{X}_{train})$ consisting of 90 observations and a testing dataset $(\mathbf{y}_{test}, \mathbf{X}_{test})$ consisting of the remaining 30 observations. We run SSNG on the training data, estimate $\hat{\beta}_{train}$, and compute the mean squared prediction error (MSPE) on the testing set: MSPE $= \frac{1}{30}\sum_{i=1}^{30}(\hat{\beta}_{train}^{\top}\mathbf{x}_{i,test} - y_{i,test})^2$, where $y_{i,test}$ is the ith entry of $\mathbf{y}_{test}$ and $\mathbf{x}_{i,test}$ is a column vector with the entries of the ith row of $\mathbf{X}_{test}$. We conduct 100 repetitions of this process and take the average of the MSPEs. Table 10 shows the resulting number of selected probe sets and average MSPE for this analysis, along with the respective results for SSLASSO, LASSO, SCAD, and MCP taken from Bai et al. (2020).

	MSPE	**Number of selected probe sets**
SSNG	0.013	2
SSLASSO	0.011	28
LASSO	0.012	32
SCAD	0.015	44
MCP	3.699	9

Table 10: Average mean square prediction errors and the numbers of selected probe sets for the BBS data analysis. The results for SSLASSO, LASSO, SCAD and MCP are taken from Bai et al. (2020); the same analysis also appears in Tadesse and Vannucci (2021).

According to the GeneCards (`https://www.genecards.org`) database of human genes, there are 25 proteins that are known to interact with TRIM32. With the aforementioned dimensionality reduction imposed on the original dataset of 31,099 genes, out of these 25 proteins, only the gene SCGB1A1 is included in the 10,000 that take part in the analysis. SSNG is able to identify SCGB1A1 plus one more gene with probe ID "1390539_at", about which no information is available. Bai et al. (2020) mention that SSLASSO selects SCGB1A1 plus 27 additional genes, and that LASSO selects 32 genes in total, without commenting on which these genes are. By selecting only 2 genes, SSNG manages to achieve an MSPE of 0.013, which is slightly higher than the MSPE of SSLASSO (0.011) and LASSO (0.012). Thus, by selecting the most parsimonious model, SSNG is able to predict the response roughly as well as models with many more predictors.

4.6 Extensions to generalised linear models

Generalised linear models (GLMs), first introduced by Nelder and Wedderburn (1972), build on linear regression to provide a flexible class of predictive models. GLMs consist of three components: the conditional probability model of response $\mathbf{y}$ given the covariates, the linear predictor, and the link function. The response density is assumed to be in the exponential family; examples include the binomial, the Poisson, the multinomial distribution, and others (see McCullagh and Nelder (1989)).

Here, we extend SSNG to perform variable selection when the response is assumed to follow either a gamma or a Poisson distribution. Sparse gamma regression can be useful when the response variable is right-skewed and non-negative, e.g. when modelling insurance claims (De Jong and Heller, 2008). Sparse Poisson regression is used in image reconstruction and deblurring applications (see, for example, Jia (2019) and Guastavino and Benvenuto (2019)).

This is easily accomplished by simply changing the expected log-likelihood part of the ELBO in (4.2.8) with the expected log-likelihood of the response under a gamma or Poisson distribution. We begin with the case of gamma.

We use the shape-scale parametrisation of the gamma density function:

$$f(y_i; \kappa, \theta_i) = \frac{y_i^{\kappa-1} e^{-y_i/\theta_i}}{\theta_i^{\kappa} \Gamma(\kappa)} \qquad \text{for } y_i > 0 \text{ and } \kappa, \theta_i > 0,$$

where κ is the shape and θ_i is the scale parameter. The expectation of y_i is $E(y_i) = \kappa\theta_i$. In a GLM, the mean of the response is related to the linear predictor via the link function. In the case of gamma-GLM, we use the logarithmic link function:

$$\log(E(y_i)) = \log(\kappa\theta_i) = \beta^{\top}\mathbf{x}_i, \tag{4.6.1}$$

where $\boldsymbol{\beta}$ is the vector of coefficients. Thus, the observation-specific scale parameter can be written as:

$$\theta_i = e^{\boldsymbol{\beta}^\top \mathbf{x}_i}/\kappa.$$

Then we can write the log-likelihood function as:

$$LL = \sum_{i=1}^{N} - \log \Gamma(\kappa) - \kappa \boldsymbol{\beta}^\top \mathbf{x}_i + \kappa \log \kappa + (\kappa - 1) \log y_i - \kappa y_i e^{-\boldsymbol{\beta}^\top \mathbf{x}_i}.$$

In order to obtain the expectation of the above log-likelihood, we need the expectations of $\boldsymbol{\beta}^\top \mathbf{x}_i$ and $e^{-\boldsymbol{\beta}^\top \mathbf{x}_i}$ under the variational distribution. The first expectation is straightforward: $\langle \boldsymbol{\beta}^\top \mathbf{x}_i \rangle = \sum_{j=1}^{D} \alpha_j \mu_j x_{ij}$, where x_{ij} is the jth element of $\mathbf{x}_i$. The second term can be written as: $e^{-\boldsymbol{\beta}^\top \mathbf{x}_i} = e^{-\sum_{j=1}^{D} \beta_j x_{ij}} = \prod_{j=1}^{D} e^{-\beta_j x_{ij}}$. Since the β_js are i.i.d. under the variational distribution, we have: $\langle e^{-\boldsymbol{\beta}^\top \mathbf{x}_i} \rangle = \langle \prod_{j=1}^{D} e^{-\beta_j x_{ij}} \rangle = \prod_{j=1}^{D} \langle e^{-\beta_j x_{ij}} \rangle$. For the jth term it is:

$$\begin{aligned}
\langle e^{-\beta_j x_{ij}} \rangle &= \alpha_j \langle e^{-\beta_j x_{ij}} \rangle_{\text{ slab }} + (1 - \alpha_j) \langle e^{-\beta_j x_{ij}} \rangle_{\text{ spike}} \\
&= \alpha_j e^{-x_{ij}\mu_j + x_{ij}^2 s_j^2/2} + (1 - \alpha_j) \\
&= \alpha_j \big(e^{-x_{ij}\mu_j + x_{ij}^2 s_j^2/2} - 1 \big) + 1.
\end{aligned}$$

This is simply a weighted mean of the expectation of $e^{-\beta_j x_{ij}}$ at the slab (where $\beta_j \sim N(\mu_j, s_j^2)$; thus, $e^{-\beta_j x_{ij}}$ follows a log-normal distribution with mean $e^{-x_{ij}\mu_j + x_{ij}^2 s_j^2/2}$) and at the spike (where $\beta_j = 0$; thus $e^0 = 1$). We can now write the ELBO with the expected log-likelihood for the gamma-GLM:

$$\begin{aligned}\mathcal{L} = &-N \log \Gamma(\kappa) - \kappa \sum_{i=1}^{N} \sum_{j=1}^{D} \alpha_j \mu_j x_{ij} + \kappa \log \kappa + (\kappa - 1) \log y_i \\ &- \kappa \sum_{i=1}^{N} y_i \prod_{j=1}^{D} \Big[\alpha_j \big(e^{-x_{ij}\mu_j + x_{ij}^2 s_j^2/2} - 1 \big) + 1 \Big] \\ &- \sum_{j=1}^{D} \frac{\alpha_j}{2} \big(\langle \log \tau_j \rangle - \log s_j^2 + \langle \tau^{-1} \rangle (\mu_j^2 + s_j^2) - 1 \big) \\ &- \sum_{j=1}^{D} \alpha_j \big(\langle \log q(\tau_j) \rangle - \langle \log p(\tau_j) \rangle \big) \\ &- \sum_{j=1}^{D} \alpha_j \log \frac{\alpha_j}{\rho} - \sum_{j=1}^{D} (1 - \alpha_j) \log \frac{1 - \alpha_j}{1 - \rho}. \end{aligned} \tag{4.6.2}$$

We optimise (4.6.2) with respect to μ_j and s_j^2 with numerical methods. We update the vector $\boldsymbol{\mu}$ with a gradient ascent algorithm. Using gradient methods to optimise an objective function in mean-field variational inference is very common; for example, Blei and Lafferty (2007) use conjugate gradient and Newton-Raphson algorithms for optimising an ELBO with exponential terms similar to (4.6.2). At each iteration of CAVI, we compute the gradient $\boldsymbol{g}_n = \nabla_{\boldsymbol{\mu}} \mathcal{L}$ with elements:

$$\frac{\partial \mathcal{L}}{\partial \mu_j} = -\kappa \sum_{i=1}^{N} x_{ij} + \kappa \sum_{i=1}^{N} y_i x_{ij} e^{-x_{ij}\mu_j + x_{ij}^2 s_j^2/2} \prod_{n \neq j} \big[\alpha_n \big(e^{-x_{in}\mu_n + x_{in}^2 s_n^2/2} - 1 \big) + 1 \big] - \mu_j \langle \tau_j^{-1} \rangle,$$

and stepsize:

$$\zeta_n = \frac{|(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n-1})^\top (\boldsymbol{g}_n - \boldsymbol{g}_{n-1})|}{\|\boldsymbol{g}_n - \boldsymbol{g}_{n-1}\|_2^2},$$

which is known as the Barzilai–Borwein stepsize (Barzilai and Borwein, 1988), and update $\boldsymbol{\mu}$ as:

$$\boldsymbol{\mu}_n \leftarrow \boldsymbol{\mu}_{n-1} + \zeta_n \boldsymbol{g}_n,$$

starting from $\boldsymbol{\mu}_1 = \boldsymbol{\mu}$, until the change in (4.6.2) is too small, for a maximum of 100 iterations ($n = 1, \ldots, 100$).

To update s_j^2, we first isolate the terms of (4.6.2) that contain it:

$$\mathcal{L}_{[s_j^2]} = -\kappa \sum_{i=1}^{N} y_i e^{-x_{ij}\mu_j + x_{ij}^2 s_j^2/2} \prod_{n \neq j} \big[\alpha_n \big(e^{-x_{in}\mu_n + x_{in}^2 s_n^2/2} - 1\big) + 1\big] - \frac{1}{2}\big(-\log s_j^2 + s_j^2 \langle \tau_j^{-1} \rangle\big),$$

and then apply the variable transformation $\eta_j = \log s_j^2 \Rightarrow s_j^2 = e^{\eta_j}$:

$$\mathcal{L}_{[\eta_j]} = -\kappa \sum_{i=1}^{N} y_i e^{-x_{ij}\mu_j + x_{ij}^2 s_j^2/2} B_j - \frac{1}{2}\big(- \eta_j + e^{\eta_j} \langle \tau_j^{-1} \rangle\big), \tag{4.6.3}$$

where $B_j = \prod_{n \neq j} \big[\alpha_n \big(e^{-x_{in}\mu_n + x_{in}^2 s_n^2/2} - 1\big) + 1\big]$. We then optimise (4.6.3) with the Nelder–Mead method (Nelder and Mead, 1965). The parameter α_j is updated by updating its log-odds:

$$\begin{aligned} \log \frac{\alpha_j}{1-\alpha_j} = \log \frac{\rho}{1-\rho} &- \kappa \mu_j \sum_{i=1}^{N} x_{ij} - \kappa \sum_{i=1}^{N} y_i \big(e^{-x_{ij}\mu_j + x_{ij}^2 s_j^2/2} - 1\big) B_j \\ &- \frac{1}{2}\big(\langle \log \tau_j \rangle - \log s_j^2 + \langle \tau^{-1} \rangle (\mu_j^2 + s_j^2) - 1\big) \\ &- \big(\langle \log q(\tau_j) \rangle - \langle \log p(\tau_j) \rangle\big). \end{aligned}$$

We can add one more update in the M-step and optimise (4.6.2) with respect to the shape parameter κ. We can write the ELBO keeping only terms with κ as:

$$\mathcal{L}_{[\kappa]} = -N \log \Gamma(\kappa) + N\kappa \log \kappa - \kappa C,$$

where $C = \sum_{i=1}^{N} \sum_{j=1}^{D} \alpha_j \mu_j x_{ij} - \sum_{i=1}^{N} \log y_i + \sum_{i=1}^{N} y_i \prod_{j=1}^{D} \big[\alpha_j \big(e^{-x_{ij}\mu_j + x_{ij}^2 s_j^2/2} - 1\big) + 1\big]$. The first derivative is:

$$\frac{\partial \mathcal{L}}{\partial \kappa} = -N\psi(\kappa) + N \log \kappa + N - C.$$

We can find the root of this derivative using the algorithm of Brent (1971), which searches for the root in a given interval. We first bound the derivative using the following inequality from Alzer (1997):

$$\log \kappa - \frac{1}{\kappa} < \psi(\kappa) < \log \kappa - \frac{1}{2\kappa},$$

and then calculate the roots of the lower and upper bound. The two roots are available in closed form and thus, we obtain the interval: $\left(\frac{N}{2(C-N)}, \frac{N}{C-N}\right)$.

The procedure with the Poisson-GLM is very similar. First, we substitute the expected log-likelihood part of (4.6.2) with:

$$\sum_{i=1}^{n} \left[y_i \sum_{j=1}^{D} \alpha_j \mu_j x_{ij} - \prod_{n \neq j} \left[\alpha_n \left(e^{x_{in}\mu_n + x_{in}^2 s_n^2/2} - 1 \right) + 1 \right] - \log(y_i!) \right].$$

Similarly, we update $\boldsymbol{\mu}$ with a gradient ascent algorithm, based on derivatives:

$$\frac{\partial \mathcal{L}}{\partial \mu_j} = \sum_{i=1}^{N} y_i x_{ij} - x_{ij} e^{x_{ij}\mu_j + x_{ij}^2 s_j^2/2} \prod_{n \neq j} \left[\alpha_n \left(e^{x_{in}\mu_n + x_{in}^2 s_n^2/2} - 1 \right) + 1 \right] - \mu_j \langle \tau_j^{-1} \rangle.$$

We update s_j^2 by optimising the objective function under a variable transformation:

$$\mathcal{L}_{[\eta_j]} = -\sum_{i=1}^{N} e^{x_{ij}\mu_j + x_{ij}^2 s_j^2/2} B_j - \frac{1}{2}\left(-\eta_j + e^{\eta_j} \langle \tau_j^{-1} \rangle \right), \tag{4.6.4}$$

where $\eta_j = \log s_j^2$ and $B_j = \prod_{n \neq j} \left[\alpha_n \left(e^{x_{in}\mu_n + x_{in}^2 s_n^2/2} - 1 \right) + 1 \right]$. We then optimise (4.6.4) with the Nelder–Mead method (Nelder and Mead, 1965). Finally, we update α_j by updating its log-odds:

$$\begin{aligned} \log \frac{\alpha_j}{1-\alpha_j} = \log \frac{\rho}{1-\rho} - \mu_j \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} \left(e^{x_{ij}\mu_j + x_{ij}^2 s_j^2/2} - 1 \right) B_j \\ - \frac{1}{2}\left(\langle \log \tau_j \rangle - \log s_j^2 + \langle \tau^{-1} \rangle (\mu_j^2 + s_j^2) - 1 \right) \\ - \left(\langle \log q(\tau_j) \rangle - \langle \log p(\tau_j) \rangle \right). \end{aligned}$$

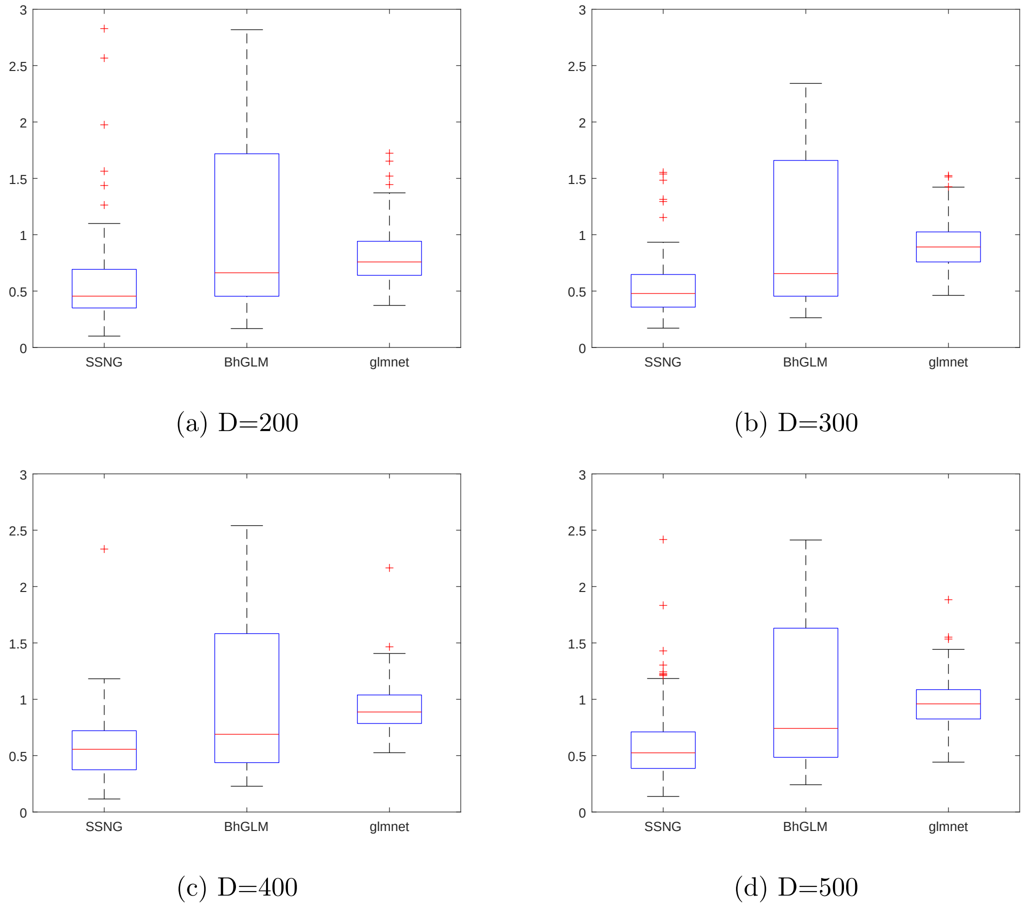
4.6.1 A low dimensional simulation study

We compare the SSNG version for Poisson regression with the BhGLM (Yi et al., 2019) and glmnet (Friedman et al., 2010) packages in R. We generate 100 datasets of size $N = 100$ and $D = 200$, and all design matrices are drawn from a standard normal distribution. There are 10 non-zero elements in $\boldsymbol{\beta}$ which are drawn from a $\mathcal{N}(0, 0.25)$ distribution and the response vector $\mathbf{y}$ is drawn from a $\text{Poi}(\exp(\mathbf{X}\boldsymbol{\beta}))$ distribution, where Poi denotes the Poisson distribution. For each dataset, we fit the three competitive methods and calculate the coefficient estimation error $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2$. We repeat the simulations for $D = 300$, $D = 400$ and $D = 500$.

Figure 4.6.1 shows the boxplots of the estimation errors for each setup. SSNG has the lowest median estimation error overall. BhGLM has a lower median estimation error than glmnet, but with much higher variance. SSNG does produce some outliers, but most of them lie below the upper whiskers of the other two methods. In terms of convergence time, SSNG and glmnet require about 1 second on average per dataset, for all dimensions, while BhGLM averages 13,16,22 and 29 seconds for $D = 200, 300, 400$ and 500 respectively. For higher dimensions, e.g. $D = 1000$, BhGLM fails to converge in a reasonable time.

4.6.2 Motif data analysis

The data consist of expression ratios of $N = 4443$ *Saccharomyces cerevisiae* genes and the corresponding $D = 2155$ motif-matching scores. The dataset was used in Conlon et al. (2003) in a motif-regression approach and also in Bühlmann and Van De Geer (2011) to demonstrate penalised regression techniques. Khalili et al. (2011) and Zhang (2017) applied mixtures of sparse regression models in order to deal with heterogeneity that appears in the data. Here, we use the motif data to compare the SSNG and its gamma-GLM version, which we refer to as SSNG-Gamma.



(a) D=200

(b) D=300

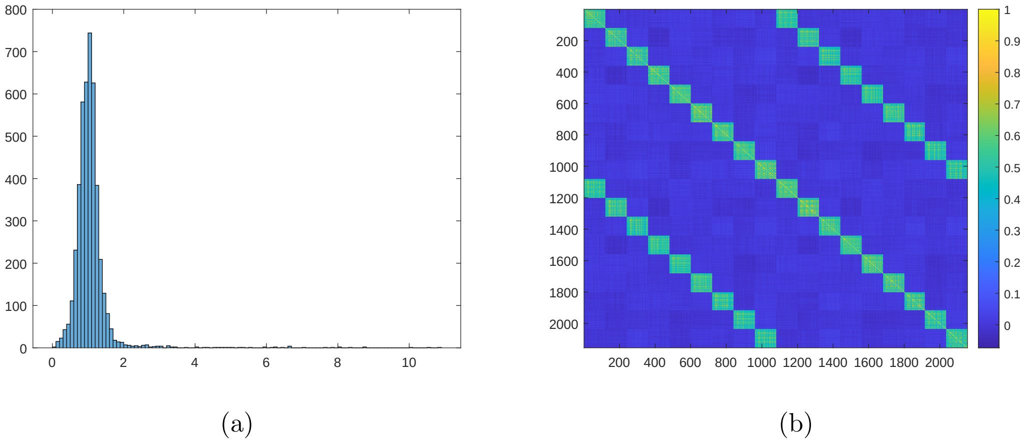
(c) D=400

(d) D=500

Figure 4.6.1: Estimation errors of SSNG, BhGLM and glmnet for four different sizes of datasets.

We use the expression ratios as the response variable for SSNG-Gamma and their logarithm as the response for SSNG. The expression ratios are non-negative and positive skewed, as shown in Figure 4.6.2a, which makes the use of gamma regression reasonable. Although this is not a $D \gg N$ scenario, the motif data are high-dimensional and highly correlated, which makes variable selection a challenging task for this dataset. The covariates exhibit a block-diagonal sample correlation matrix, and there is also correlation among blocks, as shown in Figure 4.6.2b.

SSNG converges after 280 iterations and selects 142 motifs, while SSNG-Gamma converges after 37 iterations and selects 115 motifs. Figure 4.6.3a shows



(a)

(b)

Figure 4.6.2: Distribution of expression ratios and correlation matrix of the motif data.

the evolution of ELBO per iteration for the two methods until convergence is reached. To assess and compare predictive accuracy, a leave-1000-out cross-validation is performed. The data are randomly split into 3443 training observations and 1000 test observations. For each training set, we fit the two versions of SSNG and compute the RMSPE on the left-out set. This process is repeated 100 times. Since in SSNG-Gamma we use the logarithmic link (4.6.1) and in SSNG we use the logarithm of the expression ratios as the response variable, we calculate the RMSPE in both cases as:

$$\text{RMSPE} = \left(\frac{1}{1000} \sum_{i=1}^{1000} (\hat{\boldsymbol{\beta}}^\top \mathbf{x}_i - \log y_i)^2 \right)^{\frac{1}{2}}.$$

We also make comparisons using the negative log-likelihood (NLL), calculated on the left-out data. For the normal-linear model, it is:

$$\text{NLL} = \frac{1}{2} N \big(\log 2\pi + \log \hat{\sigma}^2 \big) + \frac{1}{2\hat{\sigma}^2} \big((\mathbf{y} - \mathbf{X}\hat{\boldsymbol{w}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{w}}) + \text{Tr}(\mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{W}}) \big),$$

where the estimates $\hat{\sigma}^2$, $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{W}}$ are variational posterior means computed using the variational parameters obtained from running SSNG on the training data. For the gamma-GLM, it is:

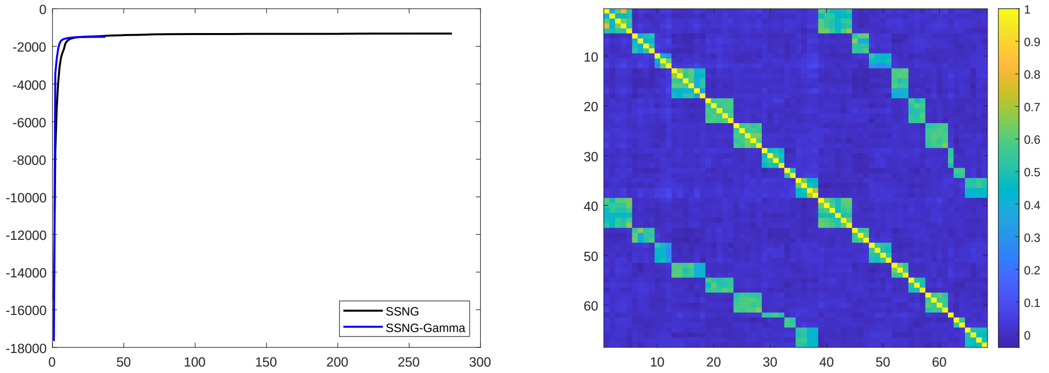
$$\begin{aligned} \text{NLL} = N \log \Gamma(\hat{\kappa}) + \hat{\kappa} \sum_{i=1}^{N} \sum_{j=1}^{D} \alpha_j \mu_j x_{ij} + \hat{\kappa} \log \hat{\kappa} + (\hat{\kappa} - 1) \log y_i \\ + \hat{\kappa} \sum_{i=1}^{N} y_i \prod_{j=1}^{D} \Big[\alpha_j \big(e^{-x_{ij}\mu_j + x_{ij}^2 s_j^2/2} - 1 \big) + 1 \Big], \end{aligned}$$

where, likewise, we use variational parameters obtained from running SSNG-Gamma on the training data, and $\hat{\kappa}$ is the empirical Bayes estimate of the shape parameter, obtained from the M-step of SSNG-Gamma. Figure 4.6.4 shows the RMSPE and NLL results for the cross-validation comparison. SSNG-Gamma's performance is slightly inferior to that of SSNG, considering that it included 27 fewer variables (115 vs 142).

Finally, we perform a brief analysis on the 68 motifs that are selected by both SSNG and SSNG-Gamma. These motifs exhibit a block-diagonal sample correlation matrix (see Figure 4.6.3b), which this time has blocks of uneven sizes and weaker correlations than the complete data. We compare the coefficient estimates of the 68 motifs from SSNG, which is run on the complete dataset, with the ordinary least squares estimates using only these 68 predictors in Figure 4.6.5a. Five of these variables have p-values above the 5% significance level, as shown in Figure 4.6.5b.

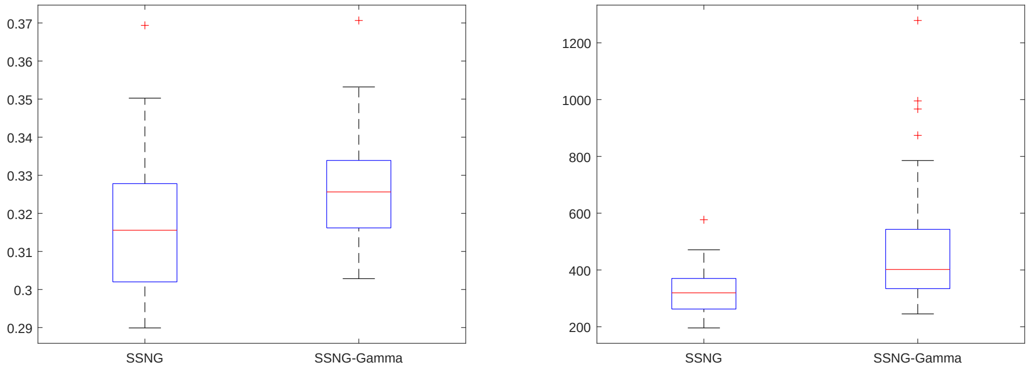
4.7 Discussions

We have proposed SSNG, a scalable algorithm for estimating sparse signals in high-dimensional problems. SSNG is based on the combination of a discrete spike-and-slab prior with a hierarchical shrinkage prior and uses mean-field variational inference to approximate the intractable posterior. We compared our work to similar methods that use a normal or a Laplace distribution as a slab. Our approach, instead, uses a normal-gamma prior which is also known as the generalised



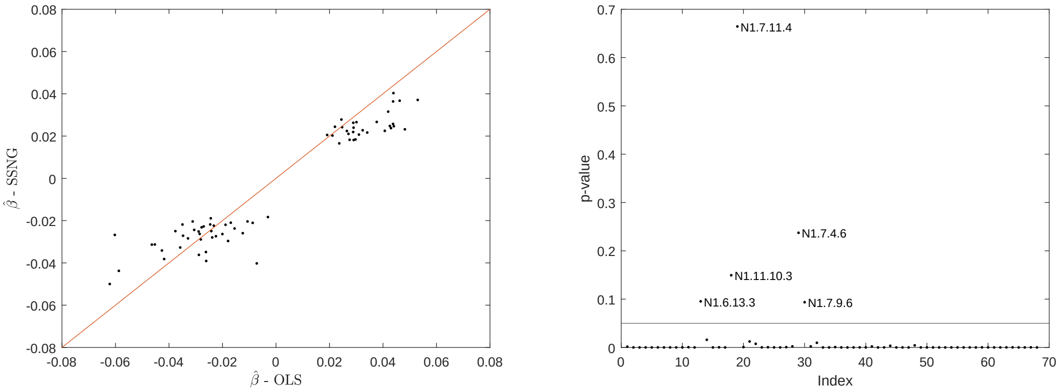
(a) (b)

Figure 4.6.3: (a) ELBO versus iteration for SSNG and SSNG-Gamma. (b) Sample correlation matrix of the 68 motifs that are commonly selected by both SSNG and SSNG-Gamma (see Table 11 for the 68 motifs).



(a) (b)

Figure 4.6.4: (a) Root mean squared prediction errors for SSNG and SSNG-Gamma respectively. (b) Negative log-likelihoods for SSNG and SSNG-Gamma respectively. The two statistics are calculated in each of the 100 repetitions in the leave-1000-out cross validation conducted on the motif dataset.



(a)

(b)

Figure 4.6.5: (a) Scatter plot of estimated model coefficients by OLS estimation and SSNG. (b) p-values for the t-tests that each variable is statistically significant in the model. Solid black horizontal line indicates the 5% significance level.

Laplace distribution and has the Laplace as a special case. The benefits of using a Gaussian scale mixture as a slab are that the evidence lower bound and the updates for the variational parameters are available in closed form, and that we were able to devise a very simple, yet very efficient, initialisation scheme without compromising speed.

Not integrating out the variance parameter (τ_j) in the normal-gamma prior of each model coefficient results in additional variational distributions for these variables ($q(\tau_j)$). Conveniently enough, the optimal $q(\tau_j)$ turns out to be a generalised inverse-Gaussian, with two of its three parameters being fixed and the third one being a function of two other variational parameters. Thus, the addition of D variational densities to our approximation only adds minimal computational cost to the process. Another convenience the GIG offers is that the expectations of $\tau_j, \tau_j^{-1}, \log \tau_j$ as well as the entropy and cross-entropy, are available in closed form.

There are a few possible ways in which our work can be improved. Mean-field variational methods are known to underestimate posterior variance. Indeed, in all the examples presented in this chapter, s_j^2, which is the variance of $q(\beta_j)$, is always

N1.9.3	N1.6.13.3	N1.8.8.5	N1.12.3.8	P1.11.3.2	P1.12.14.5
N1.9.8	N1.7.6.3	N1.9.12.5	N1.12.4.8	P1.11.14.2	P1.8.12.6
N1.9.9	N1.7.15.3	N1.10.10.5	P1.6.4	P1.12.15.2	P1.8.11.7
N1.11.2	N1.10.7.3	N1.12.12.5	P1.11.7	P1.9.11.3	P1.12.11.6
N1.11.15	N1.11.1.3	N1.7.4.6	P1.11.13	P1.11.14.3	P1.7.8.8
N1.8.6.1	N1.11.10.3	N1.7.9.6	P1.12.1	P1.11.15.3	P1.11.9.8
N1.9.1.1	N1.7.11.4	N1.8.15.6	P1.12.11	P1.12.9.4	P1.12.4.8
N1.12.2.1	N1.7.14.4	N1.11.15.6	P1.12.15	P1.12.11.4	P1.12.6.8
N1.12.6.1	N1.8.4.4	N1.9.13.7	P1.9.14.1	P1.12.13.4	
N1.7.5.2	N1.8.13.4	N1.12.14.7	P1.11.11.1	P1.11.7.5	
N1.11.11.2	N1.12.11.4	N1.6.7.8	P1.11.14.1	P1.12.7.5	
N1.12.1.2	N1.7.11.5	N1.8.14.8	P1.6.14.2	P1.12.11.5	

Table 11: Names of 68 motifs that are commonly selected by both SSNG and SSNG-Gamma.

a very small number after the convergence of our variational algorithm. One way to counter this phenomenon is to incorporate the linear response variational Bayes method of Giordano et al. (2015).

Lastly, it should be noted that it can be challenging to create a binary response version of SSNG, that is efficient in $D \gg N$ scenarios. In a logistic regression model, it is assumed that the responses $y_i \in \{0, 1\}$, for $i = 1, \ldots, N$, are conditionally independent draws from a $\text{Bern}(\text{sigm}(x_i^\top \beta))$ distribution, where $\text{sigm}(x) = (1 + e^{-x})^{-1}$ is the sigmoid function. The log-likelihood function of the logistic model thus contains the term $\log \text{sigm}(x_i^\top \beta)$. The expectation of this term with respect to β is intractable, and thus, having the ELBO in closed form is not possible.

Jaakkola and Jordan (2000) made variational inference for a Bayesian logistic regression model possible by introducing a tangent quadratic lower bound of the logistic log-likelihood, which is conjugate to the Gaussian priors they assigned to the regression coefficients. In the process of making this thesis, we were able to create variational logistic regression versions of Algorithms 2 and 3 by using this quadratic lower bound. However, the performance in simulation studies as demanding as those we conducted in Section 4.4 was not satisfactory, and we did not pursue this matter any further.

It can be interesting to try to infer sparse binary regression models using alternative techniques that have appeared in the statistical or machine learning literature. Marlin et al. (2011) and Knowles and Minka (2011) proposed tighter bounds than the quadratic bound of Jaakkola and Jordan (2000), which can be used in logistic regression. Paisley et al. (2012) proposed an alternative algorithm for directly optimising the ELBO via stochastic optimisation, with Bayesian logistic regression used as an example of their method. More recently, Durante et al. (2019) proposed another variational approach that combines the quadratic bound of Jaakkola and Jordan (2000) with the highly popular data augmentation technique of Polson et al. (2013). Finally, it is worth examining the possibility of conducting variational inference for a probit model (Albert and Chib, 1993) with a spike-and-slab prior on its coefficients.

Chapter 5

Discussion and future work

In Chapter 3, we created an objective prior for the degrees of freedom of a Wishart distribution. The prior is a function of the Kullback–Leibler divergence between Wishart distributions that share the the same scale matrix and whose degrees of freedom differ by one. The Wishart distribution is related to the inverse-Wishart via the following relation: if $X \sim \mathcal{W}(V, \nu)$ then $X^{-1} \sim \mathcal{W}^{-1}(V^{-1}, \nu)$, where $\mathcal{W}^{-1}$ denotes the inverse-Wishart distribution. Since the Kullback–Leibler divergence is invariant under variable transformations, the same prior can be used as an objective prior for the degrees of freedom of an inverse-Wishart distribution.

This enables us to use the prior in any hierarchical prior specification in which a Wishart or an inverse-Wishart distribution is used. One such example is the matrix-F distribution (Mulder and Pericchi, 2018), which can be used as a prior for covariance matrices. The matrix-F distribution is available in closed form, but it can also be conveniently presented as a Wishart mixture of inverse-Wishart distributions or a Wishart mixture of Wishart distributions.

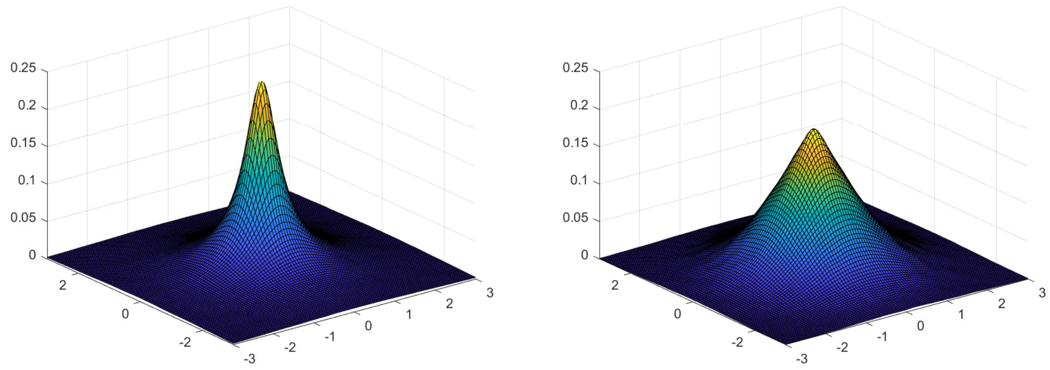
In a sparse regression problem, we would assign a multivariate normal prior to the model parameters (θ) and the matrix-F to its covariance matrix. As per

Mulder and Pericchi (2018), the marginal prior that is induced on θ, which is

$$\pi(\theta) = \int\int \mathcal{N}(\theta; \mu, \Sigma) \times \mathcal{W}^{-1}(\Sigma; \delta + k - 1, \Psi) \times \mathcal{W}(\Psi; \nu, B) d\Psi d\Sigma,$$

for $\nu = k$ (here k is the dimension of θ) and $\delta = 1$ has a pole at the origin: $\pi(\theta = \mu) = \infty$ and tails heavier than a multivariate Cauchy distribution. This is an example of a horseshoe-type prior (Carvalho et al., 2009, 2010); i.e. a prior that has a pole at the origin and heavy tails. Such priors are suitable for modelling sparse signals, as the pole causes small signals to approach 0 in the posterior while at the same time, the heavy tails allow large signals to remain large in the posterior.

Figure 5.0.1 shows how the marginal prior on θ changes when we assign our objective prior to ν. The tails, which are heavier than those of a Cauchy distribution, become even heavier. An interesting topic for future research would be to examine whether the pole at the origin is preserved in the new marginal prior and how such a prior can be useful in VAR models or other regression problems.



(a)

(b)

Figure 5.0.1: (a) A bivariate marginal prior with $\delta = 1$, $\nu = k$. (b) A bivariate marginal prior with $\delta = 1$, $\nu \sim \pi(\nu)$. Plots are kernel density estimates based on 10^6 simulated samples for each scenario.

In Chapter 4, we presented an algorithm that performs variable selection in linear and generalised linear models, based on variational approximations of posterior distributions. Although we thoroughly examined its performance in simulated and real data applications, we did not establish any theoretical results. The topic of theoretical studies of variational approximations is relatively new and active. For example, Ormerod et al. (2017) and Yang et al. (2020) studied the consistency properties of linear models with discrete spike-and-slab priors, while Wang and Blei (2019) explored these properties for other models, including Bayesian generalised linear models. Such studies rely on treating the mean-field variational posterior mean (or mode) as a point estimate and examining its consistency in the frequentist sense. Zhou and Pati (2021) reviewed some theoretical developments regarding discrete spike-and-slab priors, including results on variational inference. These theoretical guarantees do not hold in general, and each existing result pertains to a particular model with a specific variational approximation. Therefore, our work can be expanded by establishing similar theoretical studies.

Another direction regarding variable selection and variational inference is the construction of algorithms that perform variable selection while taking into account a possible clustering of observations, thereby allowing for different regression models across clusters of data. This approach to variable selection has been investigated in recent years in the context of Bayesian nonparametrics. Most of the methods that have been proposed follow Müller et al. (1996) and treat the covariates as random variables, specifying an appropriate probability model on $\mathbf{Z} = (\mathbf{y}, \mathbf{X})$. Barcella et al. (2016) exploited this idea and proposed a covariate-dependent Dirichlet process mixture model with a base measure defined by a discrete spike-and-slab prior for the regression coefficients. They performed posterior inference using an MCMC algorithm (Neal, 2000). More recently, Ding and Karabatsos (2021) used slice sampling (Walker, 2007; Kalli et al., 2011) to infer a similar model with a base measure defined by absolutely continuous shrinkage priors (horseshoe and normal-gamma). A review of earlier methods on covariate-dependent Dirichlet process mixture models can be found in Barcella et al. (2017). An issue with such methods is that they rely on MCMC techniques for posterior inference, which limits their applications to datasets of low to moderate dimensionality. Therefore, examining how variational methods for Dirichlet process mixtures (Blei and Jordan, 2006) can help scale up these methods to higher dimensions is of interest, both from practical and theoretical perspectives.

Appendix A

Appendix to Chapter 3

A.1 Derivation of conditional posteriors

Derivation of the two conditional posteriors for $\boldsymbol{\alpha}$ and Σ^{-1}.

$$\boldsymbol{\alpha}|\boldsymbol{y}, \Sigma^{-1} \sim \mathcal{N}(\overline{\boldsymbol{\alpha}}, \overline{V}),$$

$$\Sigma^{-1}|\boldsymbol{\alpha}, \boldsymbol{y} \sim \mathcal{W}\left(\overline{\nu}, \overline{S}^{-1}\right),$$

$$\begin{aligned}
p(\boldsymbol{\alpha}) &\propto \exp\bigl\{-\frac{1}{2}(\boldsymbol{\alpha} - \underline{\boldsymbol{\alpha}})^\top \underline{V}^{-1}(\boldsymbol{\alpha} - \underline{\boldsymbol{\alpha}})\bigr\} \\
&\propto \exp\bigl\{-\frac{1}{2}\bigl(\boldsymbol{\alpha}^\top \underline{V}^{-1}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^\top \underline{V}^{-1}\underline{\boldsymbol{\alpha}} + \underline{\boldsymbol{\alpha}}^\top \underline{V}^{-1}\underline{\boldsymbol{\alpha}}\bigr)\bigr\} \\
&\propto \exp\bigl\{-\frac{1}{2}\bigl(\boldsymbol{\alpha}^\top \underline{V}^{-1}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^\top \underline{V}^{-1}\underline{\boldsymbol{\alpha}}\bigr)\bigr\} \\
&\propto \exp\bigl\{-\frac{1}{2}\bigl(\boldsymbol{\alpha}^\top \overline{V_0}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^\top \overline{\boldsymbol{\alpha}}_0\bigr)\bigr\}
\end{aligned}$$

where $\overline{V_0} = \underline{V}^{-1}$ and $\overline{\boldsymbol{\alpha}}_0 = \underline{V}^{-1}\underline{\boldsymbol{\alpha}}$. Let $Z = I_m \otimes X$ and $\Lambda = \Sigma \otimes I_T$.

$$\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{\alpha},\Sigma) &\propto \exp\big\{-\frac{1}{2}(\boldsymbol{y} - Z\boldsymbol{\alpha})^\top \Lambda^{-1}(\boldsymbol{y} - Z\boldsymbol{\alpha})\big\} \\
&\propto \exp\big\{-\frac{1}{2}\big(\boldsymbol{y}^\top\Lambda^{-1}\boldsymbol{y} - \boldsymbol{y}^\top\Lambda^{-1}Z\boldsymbol{\alpha} - (Z\boldsymbol{\alpha})^\top\Lambda^{-1}\boldsymbol{y} + (Z\boldsymbol{\alpha})^\top\Lambda^{-1}Z\boldsymbol{\alpha}\big)\big\} \\
&\propto \exp\big\{-\frac{1}{2}\big(\boldsymbol{y}^\top\Lambda^{-1}\boldsymbol{y} - \boldsymbol{y}^\top\Lambda^{-1}Z\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top Z^\top\Lambda^{-1}\boldsymbol{y} + \boldsymbol{\alpha}^\top Z^\top\Lambda^{-1}Z\boldsymbol{\alpha}\big)\big\} \\
&\propto \exp\big\{-\frac{1}{2}\big(-2\boldsymbol{\alpha}^\top Z^\top\Lambda^{-1}\boldsymbol{y} + \boldsymbol{\alpha}^\top Z^\top\Lambda^{-1}Z\boldsymbol{\alpha}\big)\big\} \\
&\propto \exp\big\{-\frac{1}{2}\big(\boldsymbol{\alpha}^\top\overline{V_1}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^\top\overline{\boldsymbol{\alpha}}_1\big)\big\}
\end{aligned}$$

where $\overline{V_1} = Z^\top\Lambda^{-1}Z$ and $\overline{\boldsymbol{\alpha}}_1 = Z^\top\Lambda^{-1}\boldsymbol{y}$. Then, the conditional posterior of $\boldsymbol{\alpha}$ will be:

$$\begin{aligned}
p(\boldsymbol{\alpha}|\boldsymbol{y},\Sigma) &\propto p(\boldsymbol{y}|\boldsymbol{\alpha},\Sigma)p(\boldsymbol{\alpha}) \\
&\propto \exp\big\{-\frac{1}{2}\big(\boldsymbol{\alpha}^\top\overline{V_1}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^\top\overline{\boldsymbol{\alpha}}_1\big)\big\}\exp\big\{-\frac{1}{2}\big(\boldsymbol{\alpha}^\top\overline{V_0}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^\top\overline{\boldsymbol{\alpha}}_0\big)\big\} \\
&\propto \exp\big\{\boldsymbol{\alpha}^\top\overline{\boldsymbol{\alpha}}_1 - \frac{1}{2}\boldsymbol{\alpha}^\top\overline{V_1}\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^\top\overline{V_0}\boldsymbol{\alpha} + \boldsymbol{\alpha}^\top\overline{\boldsymbol{\alpha}}_0\big\} \\
&\propto \exp\big\{-\frac{1}{2}\big(\boldsymbol{\alpha}^\top(\overline{V_0} + \overline{V_1})\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^\top(\overline{\boldsymbol{\alpha}}_0 + \overline{\boldsymbol{\alpha}}_1)\big)\big\}
\end{aligned}$$

which is recognised as a normal distribution with variance $\overline{V} = (\overline{V_0} + \overline{V_1})^{-1} = (\underline{V}^{-1} + Z^\top\Lambda^{-1}Z)^{-1}$ and mean $\overline{\boldsymbol{\alpha}} = \overline{V}(\underline{V}^{-1}\underline{\boldsymbol{\alpha}} + Z^\top\Lambda^{-1}\boldsymbol{y})$.

For the derivation of the second conditional posterior, the following form of the likelihood is more convenient:

$$\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{\alpha},\Sigma) &= (2\pi)^{-\frac{mT}{2}}|\Sigma|^{-\frac{T}{2}}\exp\big\{-\frac{1}{2}(\boldsymbol{y} - (I_m\otimes X)\boldsymbol{\alpha})^\top(\Sigma\otimes I_T)^{-1}(\boldsymbol{y} - (I_m\otimes X)\boldsymbol{\alpha})\big\} \\
&\propto |\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\operatorname{Tr}\left[(Y - XA)^\top(Y - XA)\Sigma^{-1}\right]\right\}
\end{aligned}$$

Obtaining the second form of the likelihood, with the trace inside the exponential term, requires a few steps of linear algebra. Firstly, we notice that $\operatorname{vec}(Y - XA) =$

$\boldsymbol{y} - (I_m \otimes X)\boldsymbol{\alpha}$. Set $\boldsymbol{u} = \boldsymbol{y} - (I_m \otimes X)\boldsymbol{\alpha})$ and $U = Y - XA$, so that:

$$\boldsymbol{u} = \text{vec}(U) = \text{vec}(UI_m) = (I_m \otimes U)\text{vec}(I_m).$$

Thus, we have:

$$\begin{aligned}
\boldsymbol{u}'(\Sigma \otimes I_T)^{-1}\boldsymbol{u} &= \text{vec}(I_m)^\top (I_m \otimes U)^\top (\Sigma \otimes I_T)^{-1}(I_m \otimes U)\text{vec}(I_m) \\
&= \text{vec}(I_m)^\top (I_m \otimes U^\top)(\Sigma^{-1} \otimes I_T)(I_m \otimes U)\text{vec}(I_m) \\
&= \text{vec}(I_m)^\top \left[(I_m\Sigma^{-1}) \otimes (U^\top I_T)\right](I_m \otimes U)\text{vec}(I_m) \\
&= \text{vec}(I_m)^\top (\Sigma^{-1} \otimes U^\top)(I_m \otimes U)\text{vec}(I_m) \\
&= \text{vec}(I_m)^\top \big((\Sigma^{-1}I_m) \otimes (U^\top U)\big)\text{vec}(I_m) \\
&= \text{vec}(I_m)^\top (\Sigma^{-1} \otimes U^\top U)\text{vec}(I_m) \\
&= \text{Tr}\left[\text{vec}(I_m)^\top (\Sigma^{-1} \otimes U^\top U)\text{vec}(I_m)\right] \\
&= \text{Tr}\left[\text{vec}(I_m)\text{vec}(I_m)^\top (\Sigma^{-1} \otimes U^\top U)\right] \\
&= \text{Tr}\left[\Sigma^{-1}U^\top U\right] \\
&= \text{Tr}\left[U^\top U\Sigma^{-1}\right]
\end{aligned}$$

Then, the conditional posterior of Σ^{-1} will be:

$$\begin{aligned}
p(\Sigma^{-1}|\boldsymbol{y}, \boldsymbol{\alpha}, \underline{\nu}, \underline{S}^{-1}) &\propto p(\boldsymbol{y}|\boldsymbol{\alpha}, \Sigma)p(\Sigma^{-1}|\underline{\nu}, \underline{S}^{-1}) \\
&\propto |\Sigma|^{-\frac{T}{2}}|\Sigma|^{-\frac{\underline{\nu}-m-1}{2}} \exp\left\{-\frac{1}{2}\,\text{Tr}\left[(Y - XA)^\top (Y - XA)\Sigma^{-1}\right]\right\} \times \\
&\exp\left\{-\frac{1}{2}\,\text{Tr}(\underline{S}\Sigma^{-1})\right\} \\
&\propto |\Sigma^{-1}|^{-\frac{\underline{\nu}+T-m-1}{2}} \times \\
&\exp\left\{-\frac{1}{2}\,\text{Tr}\left[(Y - XA)^\top (Y - XA)\Sigma^{-1}\right] - \frac{1}{2}\,\text{Tr}(\underline{S}\Sigma^{-1})\right\} \\
&\propto |\Sigma^{-1}|^{-\frac{\underline{\nu}+T-m-1}{2}} \exp\left\{-\frac{1}{2}\,\text{Tr}\left[(Y - XA)^\top (Y - XA)\Sigma^{-1} + \underline{S}\Sigma^{-1}\right]\right\} \\
&\propto |\Sigma^{-1}|^{-\frac{\underline{\nu}+T-m-1}{2}} \exp\left\{-\frac{1}{2}\,\text{Tr}\left[\left((Y - XA)^\top (Y - XA)\Sigma^{-1} + \underline{S}\right)\Sigma^{-1}\right]\right\}
\end{aligned}$$

which is recognised as a $\mathcal{W}\left(\overline{\nu}, \overline{S}^{-1}\right)$ with $\overline{S} = \underline{S} + (Y - XA)^{\top}(Y - XA)$ and $\overline{\nu} = \underline{\nu} + T$.

A.2 Further simulation results

In this section we report simulation results for the case with $T = 100$. In particular, Figure A.2.1 shows the RMADs for the covariance matrix for the five-dimensional case, where the data are generated from a Wishart distribution with degrees of freedom equal to 5 (left), 10 (centre) and 15 (right). The same arises in Figure A.2.2, where we have a ten-dimensional case for the covariance matrix, and the data are generated with 10 (left), 15 (centre) and 20 (right) degrees of freedom, respectively. In conclusion, Figure A.2.3 shows the results for the twenty-dimensional case, where the data are generated from a Wishart distribution with 20 (left), 24 (centre) and 26 (right) degrees of freedom, respectively.

As stated in Chapter 3, the results show improvements in the use of our loss-based prior compared to a fixed ν prior when the data are generated with degrees of freedom higher than the dimension.

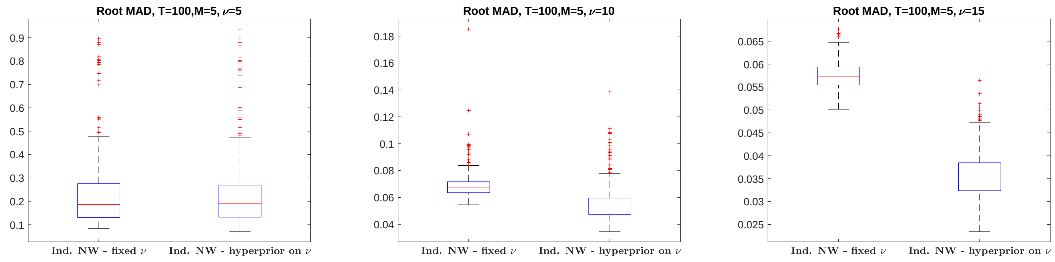


Figure A.2.1: Monte Carlo simulation — root mean absolute deviations of the covariance matrices of dimension $m = 5$. These empirical distributions are obtained by simulating 250 VAR(1) of sample size $T = 100$. Results are reported separately for data generated from a Wishart with $\nu = 5$ (left panel), $\nu = 10$ (central panel), and $\nu = 15$ (right panel).

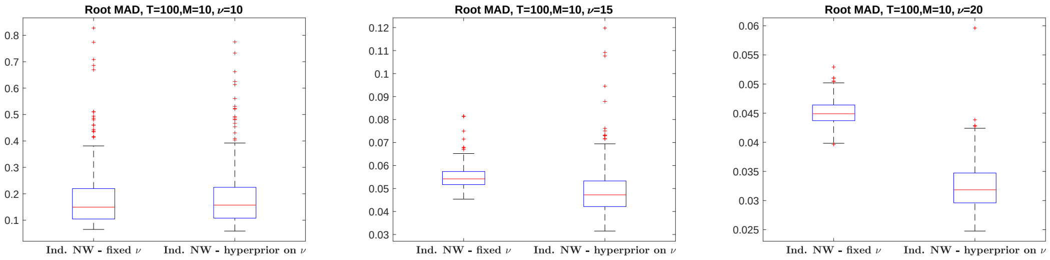


Figure A.2.2: Monte Carlo simulation — root mean absolute deviations of the covariance matrices of dimension $m = 10$. These empirical distributions are obtained by simulating 250 VAR(1) of sample size $T = 100$. Results are reported separately for data generated from a Wishart with $\nu = 10$ (left panel), $\nu = 15$ (central panel), and $\nu = 20$ (right panel).

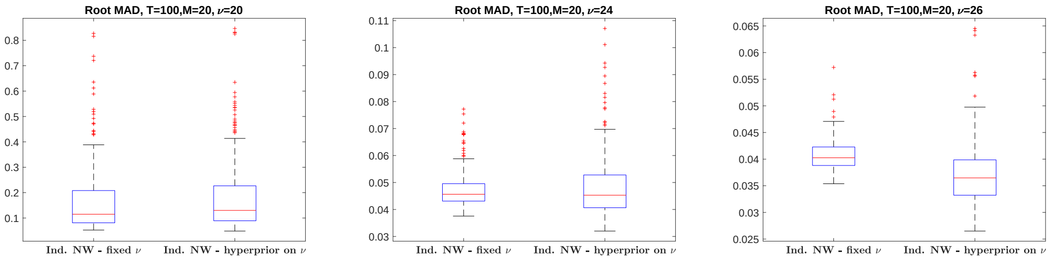


Figure A.2.3: Monte Carlo simulation — root mean absolute deviations of the covariance matrices of dimension $m = 20$. These empirical distributions are obtained by simulating 250 VAR(1) of sample size $T = 100$. Results are reported separately for data generated from a Wishart with $\nu = 20$ (left panel), $\nu = 24$ (central panel), and $\nu = 26$ (right panel).

Appendix B

Appendix to Chapter 4

B.1 The Generalised inverse Gaussian distribution and its moments

Variational distribution $q(\tau)$ is a $GIG(\nu, g, h)$ with density function:

$$\frac{(g/h)^{\nu/2}}{2\mathcal{K}_\nu(\sqrt{gh})}\tau^{\nu-1}e^{-(g\tau+h\tau^{-1})/2}.$$

The moments $\langle\tau\rangle$, $\langle\tau^{-1}\rangle$ and $\langle\log\tau\rangle$ can be found in Jørgensen (1982):

$$\langle\tau\rangle = \frac{\sqrt{h}\,\mathcal{K}_{\nu+1}(\sqrt{gh})}{\sqrt{g}\,\mathcal{K}_\nu(\sqrt{gh})},$$

$$\langle\tau^{-1}\rangle = \frac{\sqrt{g}\,\mathcal{K}_{\nu-1}(\sqrt{gh})}{\sqrt{h}\,\mathcal{K}_\nu(\sqrt{gh})},$$

$$\langle\log\tau\rangle = \log\frac{\sqrt{h}}{\sqrt{g}} + \frac{\partial}{\partial m}\log\mathcal{K}_m(\sqrt{gh})\Big|_{m=\nu},$$

Then, the calculation of the entropy is straightforward:

$$\begin{aligned}
-\langle \log q(\tau)\rangle &= -\left\langle \frac{\nu}{2}\log\left(\frac{g}{h}\right) - \log\left(2\mathcal{K}_\nu\left(\sqrt{gh}\right)\right) + (\nu-1)\log\tau - \frac{1}{2}(g\tau + h\tau^{-1})\right\rangle \\
&= \frac{1}{2}\log\left(\frac{h}{g}\right) + \log\left(2\mathcal{K}_\nu\left(\sqrt{gh}\right)\right) - (\nu-1)\frac{\partial}{\partial m}\log\mathcal{K}_m(\sqrt{gh})\Big|_{m=\nu} \\
&\quad + \frac{\sqrt{gh}}{2\mathcal{K}_\nu\left(\sqrt{gh}\right)}\left(\mathcal{K}_{\nu+1}\left(\sqrt{gh}\right) + \mathcal{K}_{\nu-1}\left(\sqrt{gh}\right)\right),
\end{aligned}$$

$$\langle \log p(\tau)\rangle = \left\langle \log\frac{\lambda_2^{\lambda_1}\tau^{\lambda_1-1}e^{-\lambda_2\tau}}{\Gamma(\lambda_1)}\right\rangle = \lambda_1\log\lambda_2 + (\lambda_1-1)\langle\log\tau\rangle - \lambda_2\langle\tau\rangle - \log\Gamma(\lambda_1).$$

The derivative of the Bessel function with respect to the order, that appears in the logarithmic derivative $\frac{\partial}{\partial m}\log\mathcal{K}_m(z)\Big|_{m=\nu} = \frac{\frac{\partial}{\partial m}\mathcal{K}_m(z)}{\mathcal{K}_m(z)}\Big|_{m=\nu}$, can be calculated with the following simple algorithm which is based on results from Olver et al. (2010) and Brychkov and Geddes (2005).

- if $\nu \in \mathbb{Z}$
  - $\frac{\partial}{\partial m}\mathcal{K}_m(z)\Big|_{m=\pm\nu} = \pm\frac{\nu!}{2(\frac{1}{2}z)^\nu}\sum_{k=0}^{\nu-1}\frac{(\frac{1}{2}z)^k\mathcal{K}_k(z)}{k!(\nu-k)}$, for $\nu = 1, 2, 3\ldots$
  - $\frac{\partial}{\partial m}\mathcal{K}_m(z)\Big|_{m=0} = 0$
- if $\nu \notin \mathbb{Z}$
  - $\frac{\partial}{\partial\nu}\mathcal{K}_\nu(z) = \frac{1}{2}\pi\csc(\nu\pi)\left[\frac{\partial}{\partial\nu}\mathcal{I}_{-\nu}(z) - \frac{\partial}{\partial\nu}\mathcal{I}_\nu(z)\right] - \pi\cot(\nu\pi)\mathcal{K}_\nu(z)$

where $\mathcal{I}_\nu(z)$ is the modified Bessel function of the first kind, of order ν, and its formulation is:

$$\mathcal{I}_\nu(z) = \left(\frac{1}{2}z\right)^\nu\sum_{k=0}^{\infty}\frac{(\frac{1}{4}z^2)^k}{k!\Gamma(\nu+k+1)}$$

and the two derivatives with respect to order are:

$$\frac{\partial}{\partial\nu}\mathcal{I}_\nu(z) = \mathcal{I}_\nu(z)\log\left(\frac{1}{2}z\right) - \left(\frac{1}{2}z\right)^\nu\sum_{k=0}^{\infty}\frac{(\frac{1}{4}z^2)^k\psi(\nu+k+1)}{k!\Gamma(\nu+k+1)}$$

$$\frac{\partial}{\partial \nu}\mathcal{I}_{-\nu}(z) = -\mathcal{I}_{-\nu}(z)\log\left(\frac{1}{2}z\right) + \left(\frac{1}{2}z\right)^{-\nu}\sum_{k=0}^{\infty}\frac{(\frac{1}{4}z^2)^k\psi(-\nu+k+1)}{k!\Gamma(-\nu+k+1)}$$

B.2 Evidence lower bound in explicit form

By substituting all expectations in (4.2.8) and gathering the terms that contain the logarithmic derivative of the Bessel function, we obtain the following explicit form of the ELBO:

$$\begin{aligned}
\mathcal{L} = &-\frac{1}{2}N\big(\log 2\pi + \log d - \psi(c)\big) - \frac{c}{2d}\big((\mathbf{y}-\mathbf{X}\boldsymbol{w})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{w}) + \mathrm{Tr}(\mathbf{X}^\top\mathbf{X}\boldsymbol{W})\big)\\
&-\sum_{j=1}^{D}\frac{\alpha_j}{2}\left(-\log s_j^2 + \left(\frac{\sqrt{g}\,\mathcal{K}_{\nu+1}(\sqrt{gh_j})}{\sqrt{h_j}\,\mathcal{K}_\nu(\sqrt{gh_j})} - \frac{2\nu}{h_j}\right)(\mu_j^2+s_j^2) - 1\right)\\
&-\sum_{j=1}^{D}\alpha_j\left(-\frac{1}{2}\log\frac{h_j}{g} - \log(2\mathcal{K}_\nu(\sqrt{gh_j})) - \frac{\sqrt{gh_j}}{2\mathcal{K}_\nu\left(\sqrt{gh_j}\right)}\left(\mathcal{K}_{\nu+1}\left(\sqrt{gh_j}\right) + \mathcal{K}_{\nu-1}\left(\sqrt{gh_j}\right)\right)\right.\\
&-\lambda_1\log\lambda_2 + \lambda_2\frac{\sqrt{h_j}\,\mathcal{K}_{\nu+1}(\sqrt{gh_j})}{\sqrt{g}\,\mathcal{K}_\nu(\sqrt{gh_j})} + \log\Gamma(\lambda_1) + (1.5-\lambda_1)\log\sqrt{\frac{h_j}{g}}\\
&\left.+\big(\frac{1}{2}-\lambda_1+\nu\big)\frac{\partial}{\partial m}\log\mathcal{K}_m(\sqrt{gh_j})\Big|_{m=\nu}\right)\\
&-(c-c_0)\psi(c) + \log\frac{\Gamma(c)}{\Gamma(c_0)} - c_0\log\frac{d}{d_0} - c\frac{d_0-d}{d}\\
&-\sum_{j=1}^{D}\alpha_j\log\frac{\alpha_j}{\rho} - \sum_{j=1}^{D}(1-\alpha_j)\log\frac{1-\alpha_j}{1-\rho}.
\end{aligned}$$

Notice the term $\big(\frac{1}{2}-\lambda_1+\nu\big)\frac{\partial}{\partial m}\log\mathcal{K}_m(\sqrt{gh_j})\Big|_{m=\nu}$. This term always equals 0 since $\nu = \lambda_1 - \frac{1}{2}$, and thus, we never need to calculate the logarithmic derivative of the Bessel function. This fact enables the use of the ELBO to monitor convergence since, despite its lengthy analytical expression, it can be calculated after every iteration of the CAVI at a very low computational cost. In particular, according to the Matlab Profiler, calculations of the ELBO account for less than 1% of total computation time until convergence is reached in all examples examined in Chapter 4. The same is also true for the Poisson and gamma GLMs.

B.3 Derivation of Algorithm 2

In order to find the update for μ_j, we first isolate the terms of ELBO (4.2.8) that contain the elements of the vector $\boldsymbol{\mu}$:

$$\mathcal{L}_{[\boldsymbol{\mu}]} = -\frac{c}{2d}\big((\mathbf{y}-\mathbf{X}\boldsymbol{w})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{w}) + \mathrm{Tr}(\mathbf{X}^\top\mathbf{X}\boldsymbol{W})\big) - \frac{1}{2}\sum_{j=1}^{D}\alpha_j\langle\tau_j^{-1}\rangle\mu_j^2.$$

Variational parameters $(\mu_j)_{j=1}^D$ appear in the last summation term, inside vector $\boldsymbol{w}$, and inside matrix $\boldsymbol{W}$. Thus, it is easier to break down the differentiation in parts. Each μ_j is the expectation of β_j, under the variational distribution, conditioned on being on a slab: $\mu_j = \mathbb{E}_q[\beta_j|z_{j=1}]$. If $z_j = 1$, then $\alpha_j = 1$. The jth term of the trace $\mathrm{Tr}(\mathbf{X}^\top\mathbf{X}\boldsymbol{W})$ becomes $(\mathbf{X}^\top\mathbf{X})_{jj}(\alpha_j(\mu_j^2+s_j^2) - (\alpha_j\mu_j)^2) = (\mathbf{X}^\top\mathbf{X})_{jj}s_j^2$, which does not contain μ_j. We also need the derivative of the quadratic term $(\mathbf{y}-\mathbf{X}\boldsymbol{w})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{w})$ with respect to μ_j. For convenience with this differentiation, we rewrite the vector $\boldsymbol{w}$ as $\boldsymbol{w} = \mathcal{A}\boldsymbol{\mu}$, where $\mathcal{A}$ is a diagonal matrix with diagonal elements $\alpha_j = 1$ and the rest being α_k for $k \neq j$. The derivative of the quadratic term is:

$$\frac{\partial}{\partial\boldsymbol{\mu}}(\mathbf{y}-\mathbf{X}\mathcal{A}\boldsymbol{\mu})^\top(\mathbf{y}-\mathbf{X}\mathcal{A}\boldsymbol{\mu}) = -2(\mathcal{A}^\top\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\mathcal{A}\boldsymbol{\mu})) = -2(\mathcal{A}^\top\mathbf{X}^\top\mathbf{y}-\mathcal{A}^\top\mathbf{X}^\top\mathbf{X}\mathcal{A}\boldsymbol{\mu}).$$

The jth element of that vector is $-2(\mathcal{A}^\top\mathbf{X}^\top\mathbf{y} - \mathcal{A}^\top\mathbf{X}^\top\mathbf{X}\mathcal{A}\boldsymbol{\mu})_j = -2((\mathbf{X}^\top\mathbf{y})_j - \sum_{k\neq j}(\mathbf{X}^\top\mathbf{X})_{kj}\alpha_k\mu_k - \mathbf{X}^\top\mathbf{X}\mu_j)$. We can then calculate the derivative of the ELBO with respect to μ_j as:

$$\frac{\partial\mathcal{L}}{\partial\mu_j} = \frac{c}{d}\big((\mathbf{X}^\top\mathbf{y})_j - \sum_{k\neq j}(\mathbf{X}^\top\mathbf{X})_{kj}\alpha_k\mu_k - \mathbf{X}^\top\mathbf{X}\mu_j\big) - \mu_j\langle\tau_j^{-1}\rangle.$$

Setting this derivative equal to 0 and solving for μ_j yields:

$$\mu_j = \frac{c}{d}\Bigg((\mathbf{X}^\top\mathbf{y})_j - \sum_{k\neq j}(\mathbf{X}^\top\mathbf{X})_{kj}\alpha_k\mu_k\Bigg)\Bigg(\frac{c}{d}(\mathbf{X}^\top\mathbf{X})_{jj} + \langle\tau_j^{-1}\rangle\Bigg)^{-1},$$

which, after finding the update for s_j^2, becomes:

$$\mu_j = \frac{c}{d}\Bigg((\mathbf{X}^\top\mathbf{y})_j - \sum_{k\neq j}(\mathbf{X}^\top\mathbf{X})_{kj}\alpha_k\mu_k\Bigg)s_j^2,$$

which is the required update.

Similarly, $s_j^2 = \mathrm{Var}[\beta_j|z_j = 1]$, and by isolating terms with s_j^2 we get:

$$\mathcal{L}_{[s_j^2]} = -\frac{c}{2d}(\mathbf{X}^\top\mathbf{X})_{jj}s_j^2 + \frac{1}{2}\log s_j^2 - \frac{1}{2}s_j^2\langle\tau_j^{-1}\rangle,$$

and the derivative with respect to s_j^2:

$$\frac{\partial\mathcal{L}}{\partial s_j^2} = -\frac{c}{2d}(\mathbf{X}^\top\mathbf{X})_{jj} + \frac{1}{2}\frac{1}{s_j^2} - \frac{1}{2}\langle\tau_j^{-1}\rangle.$$

By setting this derivative equal to 0 and solving for s_j^2, we get the update:

$$s_j^2 = \Bigg(\frac{c}{d}(\mathbf{X}^\top\mathbf{X})_{jj} + \langle\tau_j^{-1}\rangle\Bigg)^{-1}.$$

To update α_j, we write the vector $\boldsymbol{w}$ as $\boldsymbol{w} = \boldsymbol{M}\boldsymbol{\alpha}$, where $\boldsymbol{M} = \mathrm{diag}(\boldsymbol{\mu})$. The derivative of the quadratic term, with respect to $\boldsymbol{\alpha}$ is:

$$\frac{\partial}{\partial\boldsymbol{\alpha}}(\mathbf{y} - \mathbf{X}\boldsymbol{M}\boldsymbol{\alpha})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{M}\boldsymbol{\alpha}) = -2(\boldsymbol{M}^\top\mathbf{X}^\top\mathbf{y} - \boldsymbol{M}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{M}\boldsymbol{\alpha}).$$

The jth element of that vector is $-2(\boldsymbol{M}^\top\mathbf{X}^\top\mathbf{y} - \boldsymbol{M}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{M}\boldsymbol{\alpha})_j = -2((\mathbf{X}^\top\mathbf{y})_j\mu_j -$

$\mu_j \sum_{k \neq j} (\mathbf{X}^\top \mathbf{X})_{kj} \alpha_k \mu_k - \alpha_j \mu_j^2 (\mathbf{X}^\top \mathbf{X})_{jj})$. The derivative of the trace term with respect to α_j is:

$$\frac{\partial}{\partial \alpha_j} (\mathbf{X}^\top \mathbf{X})_{jj} (\alpha_j (\mu_j^2 + s_j^2) - (\alpha_j \mu_j)^2) = (\mathbf{X}^\top \mathbf{X})_{jj} (\mu_j^2 + s_j^2) - 2\alpha_j (\mathbf{X}^\top \mathbf{X})_{jj} \mu_j^2.$$

Finally, we get the complete derivative of the ELBO with respect to α_j:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \alpha_j} = -\frac{c}{2d} \Big(& -2\big((\mathbf{X}^\top \mathbf{y})_j \mu_j - \mu_j \sum_{k \neq j} (\mathbf{X}^\top \mathbf{X})_{kj} \alpha_k \mu_k - \alpha_j \mu_j^2 (\mathbf{X}^\top \mathbf{X})_{jj}\big) + (\mathbf{X}^\top \mathbf{X})_{jj} (\mu_j^2 + s_j^2) \\
& - 2\alpha_j (\mathbf{X}^\top \mathbf{X})_{jj} \mu_j^2 \Big) \\
& - \frac{1}{2} \big(\langle \log \tau_j \rangle - \log s_j^2 + \langle \tau^{-1} \rangle (\mu_j^2 + s_j^2) - 1 \big) \\
& - \big(\langle \log q(\tau_j) \rangle - \langle \log p(\tau_j) \rangle \big) \\
& + \log \frac{1 - \alpha_j}{\alpha_j} + \log \frac{\rho}{1 - \rho}.
\end{aligned}$$

By setting $\frac{\partial \mathcal{L}}{\partial \alpha_j} \equiv 0$ and solving for $\log \frac{1-\alpha_j}{\alpha_j}$, we obtain equation 4.2.9.

B.4 Parameter estimation

A run of the CAVI algorithm moves the variational lower bound as close as possible to the intractable marginal log-likelihood. We can then use this tractable lower bound as a substitute for the marginal log-likelihood and, for fixed values of the variational parameters, optimise this substitute with respect to the model parameters $\Theta = (\lambda_1, \lambda_2, \rho)$, where we have re-parametrised: $\lambda_1 = \lambda$ and $\lambda_2 = 1/(2\gamma^2)$. This way, we obtain approximate empirical Bayes estimates for the model parameters. This optimisation is straightforward:

$$\frac{\partial \mathcal{L}}{\partial \lambda_1} = -\sum_{j=1}^{D} \alpha_j \big(\psi(\lambda_1) - \log \lambda_2 - \log \sqrt{h_j / g} \big).$$

Setting this to 0, we get:

$$\psi(\lambda_1) = \log \lambda_2 + \frac{\sum_{j=1}^{D} \alpha_j \log \sqrt{h_j/g}}{\sum_{j=1}^{D} \alpha_j}.$$

The inverse of the digamma function is not available in closed form, so we instead approximate numerically the root of the equation $\psi(\lambda_1) - \log \lambda_2 - \frac{\sum_{j=1}^{D} \alpha_j \log \sqrt{h_j/g}}{\sum_{j=1}^{D} \alpha_j} = 0$ using a simple Newton-Raphson algorithm based on the derivative of the digamma function, $\psi^{(1)}(x) = \frac{\partial}{\partial x}\psi(x)$, which is also known as the trigamma function. For the general case of finding the root of $\psi(x) - y = 0$, the following Newton update is used:

$$x_{n+1} = x_n - \frac{\psi(x) - y}{\psi^{(1)}(x)},$$

using the following initialisation as found in Minka (2000):

$$x_0 = \begin{cases} \exp(y) + 0.5 & \text{if } y \geq -2.2, \\ -\frac{1}{y - \psi(1)} & \text{if } y \leq -2.2. \end{cases}$$

For the estimates of λ_2 and ρ, we get exact solutions:

$$\frac{\partial \mathcal{L}}{\partial \lambda_2} = -\sum_{j=1}^{D} \alpha_j \Big(-\lambda_1 \frac{1}{\lambda_2} + \langle \tau_j \rangle \Big),$$

which gives the solution:

$$\lambda_2 = \frac{\lambda_1 \sum_{j=1}^{D} \alpha_j}{\sum_{j=1}^{D} \alpha_j \langle \tau_j \rangle}.$$

And finally,

$$\frac{\partial \mathcal{L}}{\partial \rho} = \frac{1}{\rho} \sum_{j=1}^{D} \alpha_j - \frac{1}{1-\rho} \sum_{j=1}^{D} (1 - \alpha_j),$$

giving:

$$\rho = \frac{\sum_{j=1}^{D} \alpha_j}{\sum_{j=1}^{D}(1 - \alpha_j) + \sum_{j=1}^{D} \alpha_j} = \frac{\sum_{j=1}^{D} \alpha_j}{D}.$$

Bibliography

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Mineola, NY: Dover Publications.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

Altosaar, J., Ranganath, R., and Blei, D. M. (2017). Proximity variational inference. *arXiv preprint arXiv:1705.08931.*

Alzer, H. (1997). On some inequalities for the gamma and psi functions. *Mathematics of Computation*, 66(217):373–389.

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1):99–102.

Bai, R., Rockova, V., and George, E. I. (2020). Spike-and-slab meets lasso: A review of the spike-and-slab lasso. *arXiv preprint arXiv:2010.06451.*

Bańbura, M., Giannone, D., and Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92.

Barcella, W., De Iorio, M., and Baio, G. (2017). A comparative review of variable selection techniques for covariate dependent Dirichlet process mixture models. *Canadian Journal of Statistics*, 45(3):254–273.

Barcella, W., Iorio, M. D., Baio, G., and Malone-Lee, J. (2016). Variable selection in covariate dependent random partition models: An application to urinary tract infection. *Statistics in Medicine*, 35(8):1373–1389.

Barzilai, J. and Borwein, J. M. (1988). Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148.

Beal, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference.* PhD thesis, Gatsby Computational Neuroscience Unit, University College London.

Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors. In *Bayesian Statistics 4*, pages 35–60. Oxford, England: Oxford University Press.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, 37(2):905–938.

Berger, J. O., Bernardo, J. M., and Sun, D. (2012). Objective priors for discrete parameter spaces. *Journal of the American Statistical Association*, 107(498):636–648.

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):113–128.

Bernardo, J. M. and Smith, A. F. (1994). *Bayesian theory.* New York, NY: Wiley.

Billio, M., Casarin, R., and Rossini, L. (2019). Bayesian nonparametric sparse VAR models. *Journal of Econometrics*, 212(1):97–115.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(1):993–1022.

Braun, M. and McAuliffe, J. (2010). Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335.

Brent, R. P. (1971). An algorithm with guaranteed convergence for finding a zero of a function. *The Computer Journal*, 14(4):422–425.

Brown, P. J. and Griffin, J. E. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188.

Brychkov, Y. A. and Geddes, K. (2005). On the derivatives of the Bessel and Struve functions with respect to the order. *Integral Transforms and Special Functions*, 16(3):187–198.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Berlin, Germany: Springer.

Carbonetto, P. and Stephens, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–108.

Carbonetto, P., Zhou, X., and Stephens, M. (2017). varbvs: Fast variable selection for large-scale regression. *arXiv preprint arXiv:1709.06597*.

Carneiro, H. A. and Mylonakis, E. (2009). Google trends: A web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases*, 49(10):1557–1564.

Carriero, A., Clark, T. E., , and Marcellino, M. (2015). Bayesian VARs: Specification choices and forecast accuracy. *Journal of Applied Econometrics*, 30(1):46–73.

Carriero, A., Clark, T. E., and Marcellino, M. (2019). Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors. *Journal of Econometrics*, 212(1):137–154.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pages 73–80.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018.

Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E., Nishimura, D. Y., Braun, T. A., Kim, K.-Y. A., Huang, J., et al. (2006). Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet–Biedl syndrome gene (BBS11). *Proceedings of the National Academy of Sciences*, 103(16):6287–6292.

Clark, T. E. and McCracken, M. W. (2010). Averaging forecasts from VARs with uncertain instabilities. *Journal of Applied Econometrics*, 25(1):5–29.

Conlon, E. M., Liu, X. S., Lieb, J. D., and Liu, J. S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Sciences*, 100(6):3339–3344.

Cross, J. L., Hou, C., and Poon, A. (2020). Macroeconomic forecasting with large Bayesian VARs: Global-local priors and the illusion of sparsity. *International Journal of Forecasting*, 36(3):899–915.

Davis, R. A., Zang, P., and Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25(4):1077–1096.

De Jong, P. and Heller, G. Z. (2008). *Generalized linear models for insurance data*. Cambridge, England: Cambridge University Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Ding, D. and Karabatsos, G. (2021). Dirichlet process mixture models with shrinkage prior. *Stat*, 10(1):e371.

Doan, T., Litterman, R., and Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, 3(1):1–100.

Durante, D., Rigon, T., et al. (2019). Conditionally conjugate mean-field variational Bayes for logistic models. *Statistical Science*, 34(3):472–485.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Elezovic, N., Giordano, C., and Pecaric, J. (2000). The best bounds in Gautschi's inequality. *Mathematical Inequalities and Applications*, 3(2):239–252.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7(2):339–373.

George, E. I., Sun, D., and Ni, S. (2008). Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics*, 142(1):553–580.

Giordano, R. J., Broderick, T., and Jordan, M. I. (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 1441–1449.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Griffin, J. E. and Brown, P. J. (2007). Bayesian adaptive lassos with non-convex penalization. Technical report, Centre for Research in Statistical Methodology, University of Warwick.

Griffin, J. E. and Brown, P. J. (2011). Bayesian hyper-lassos with non-convex penalization. *Australian and New Zealand Journal of Statistics*, 53(4):423–442.

Guastavino, S. and Benvenuto, F. (2019). A consistent and numerically efficient variable selection method for sparse Poisson regression with applications to learning and signal recovery. *Statistics and Computing*, 29(3):501–516.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14(5):1303–1347.

Huber, F. and Feldkircher, M. (2019). Adaptive shrinkage in Bayesian vector autoregressive models. *Journal of Business and Economic Statistics*, 37(1):27–39.

Huber, F. and Rossini, L. (2020). Inference in Bayesian additive vector autoregressive tree models. *arXiv preprint arXiv:2006.16333.*

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37.

Jaynes, E. T. (2003). *Probability theory: The logic of science.* Cambridge, England: Cambridge University Press.

Jeffreys, H. (1939). *Theory of probability.* Oxford, England: Clarendon Press.

Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1):175–193.

Jia, J. (2019). Sparse Poisson regression with penalized weighted score function. *Electronic Journal of Statistics*, 13(2):2898.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Jørgensen, B. (1982). *Statistical properties of the generalized inverse Gaussian distribution.* New York, NY: Springer.

Kadiyala, K. R. and Karlsson, S. (1997). Numerical methods for estimation and inference in Bayesian VAR-models. *Journal of Applied Econometrics*, 12(2):99–132.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370.

Khalili, A., Chen, J., and Lin, S. (2011). Feature selection in finite mixture of sparse normal linear models in high-dimensional feature space. *Biostatistics*, 12(1):156–172.

Knowles, D. A. and Minka, T. P. (2011). Non-conjugate variational message passing for multinomial and binary regression. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 1701–1709.

Koop, G., Korobilis, D., et al. (2010). Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends in Econometrics*, 3(4):267–358.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, 60(1):65–81.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–412.

Li, Q. and Lin, N. (2010). The Bayesian elastic net. *Bayesian Analysis*, 5(1):151–170.

Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions – five years of experience. *Journal of Business and Economic Statistics*, 4(1):25–38.

Logsdon, B. A., Hoffman, G. E., and Mezey, J. G. (2010). A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics*, 11(58):1–13.

MacKay, D. J. (1992). Bayesian interpolation. *Neural Computation*, 4(3):415–447.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms.* Cambridge, England: Cambridge University Press.

Madan, D. B. and Seneta, E. (1990). The variance gamma (V.G.) model for share market returns. *Journal of Business*, 63(4):511–524.

Marlin, B. M., Khan, M. E., and Murphy, K. P. (2011). Piecewise bounds for estimating Bernoulli-logistic latent Gaussian models. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 633–640.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096.

McCracken, M. W. and Ng, S. (2020). FRED-QD: A Quarterly database for macroeconomic research. *Federal Reserve Bank of St. Louis Review.*

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London, England: Chapman and Hall.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Minka, T. (2000). Estimating a Dirichlet distribution. Technical report, Massachusetts Institute of Technology.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Mitrinovic, D. S., Barnes, E., Marsh, D., and Radok, J. (1964). *Elementary inequalities.* Groningen, Netherlands: Noordhoff.

Moran, G. E., Ročková, V., and George, E. I. (2019). Variance prior forms for high-dimensional Bayesian variable selection. *Bayesian Analysis*, 14(4):1091–1119.

Mulder, J. and Pericchi, L. R. (2018). The matrix-F prior for estimating and testing covariance matrices. *Bayesian Analysis*, 13(4):1193–1214.

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83(1):67–79.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective.* Cambridge, MA: MIT Press.

Narisetty, N. N., Shen, J., and He, X. (2019). Skinny Gibbs: A consistent and scalable Gibbs sampler for model selection. *Journal of the American Statistical Association*, 114(527):1205–1217.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Olver, F. W., Lozier, D. W., Boisvert, R. F., and Clark, C. W. (2010). *The NIST handbook of mathematical functions.* New York, NY: Cambridge University Press.

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, 64(2):140–153.

Ormerod, J. T., You, C., and Müller, S. (2017). A variational Bayes approach to variable selection. *Electronic Journal of Statistics*, 11(2):3549–3594.

Paisley, J., Blei, D. M., and Jordan, M. I. (2012). Variational Bayesian Inference with Stochastic Search. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1363–1370.

Parisi, G. (1988). *Statistical field theory.* Redwood City, CA: Addison-Wesley.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., and Weinstein, R. A. (2008). Using internet searches for influenza surveillance. *Clinical Infectious Diseases*, 47(11):1443–1448.

Polson, N. G. and Scott, J. G. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics 9*, pages 501–538. Oxford, England: Oxford University Press.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ray, K. and Szabó, B. (2022). Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 117(539):1270–1281.

Ray, K., Szabo, B., and Clara, G. (2020). Spike and slab variational Bayes for high dimensional logistic regression. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 14423–14434.

Reid, S., Tibshirani, R., and Friedman, J. (2016). A study of error variance estimation in lasso regression. *Statistica Sinica*, 26(1):35–67.

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods* (2nd ed.). New York, NY: Springer.

Ročková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Sims, C. A. and Zha, T. (1998). Bayesian methods for dynamic multivariate models. *International Economic Review*, 39(4):949–968.

Strauss, R. A., Castro, J. S., Reintjes, R., and Torres, J. R. (2017). Google dengue trends: An indicator of epidemic behavior. The Venezuelan case. *International Journal of Medical Informatics*, 104:26–30.

Tadesse, M. G. and Vannucci, M. (2021). *Handbook of Bayesian Variable Selection*. Boca Raton, FL: Chapman and Hall/CRC.

Tang, Z., Shen, Y., Li, Y., Zhang, X., Wen, J., Qian, C., Zhuang, W., Shi, X., and Yi, N. (2018). Group spike-and-slab lasso generalized linear models for disease prediction and associated genes detection by incorporating pathway information. *Bioinformatics*, 34(6):901–910.

Tang, Z., Shen, Y., Zhang, X., and Yi, N. (2017a). The spike-and-slab lasso generalized linear models for prediction and associated genes detection. *Genetics*, 205(1):77–88.

Tang, Z., Shen, Y., Zhang, X., and Yi, N. (2017b). The spike-and-slab lasso Cox model for survival prediction and associated genes detection. *Bioinformatics*, 33(18):2799–2807.

Teh, Y., Jordan, M. I., Beal, M., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, 84(407):710–716.

Van Erp, S., Oberski, D. L., and Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, 89:31–50.

Villa, C. and Rubio, F. J. (2018). Objective priors for the number of degrees of freedom of a multivariate t distribution and the t-copula. *Computational Statistics and Data Analysis*, 124:197–219.

Villa, C. and Walker, S. G. (2014). Objective prior for the number of degrees of freedom of a t distribution. *Bayesian Analysis*, 9(1):197–220.

Villa, C. and Walker, S. G. (2015). An objective approach to prior mass functions for discrete parameter spaces. *Journal of the American Statistical Association*, 110(511):1072–1082.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation*, 36(1):45–54.

Wang, B. and Titterington, D. (2012). Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. *arXiv preprint arXiv:1207.4159*.

Wang, C. and Blei, D. M. (2013). Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(4):1005–1031.

Wang, Y. and Blei, D. M. (2019). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648.

Woodbury, M. A. (1950). *Inverting modified matrices*. Statistical Research Group.

Yang, Y., Pati, D., and Bhattacharya, A. (2020). α-variational inference with statistical guarantees. *Annals of Statistics*, 48(2):886–905.

Yi, N., Tang, Z., Zhang, X., and Guo, B. (2019). BhGLM: Bayesian hierarchical GLMs and survival models, with applications to genomics and epidemiology. *Bioinformatics*, 35(8):1419–1421.

You, C., Ormerod, J. T., and Mueller, S. (2014). On variational Bayes estimation and variational information criteria for linear regression models. *Australian and New Zealand Journal of Statistics*, 56(1):73–87.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. (2019). Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(08):2008–2026.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, J. (2017). Screening and clustering of sparse regressions with finite non-Gaussian mixtures. *Biometrics*, 73(2):540–550.

Zhou, S. and Pati, D. (2021). Recent theoretical advances with the discrete spike-and-slab priors. In *Handbook of Bayesian Variable Selection*, pages 25–56. Boca Raton, FL: Chapman and Hall/CRC.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.