UNIVERSITÀ DEGLI STUDI DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXV

Variational inference for high-dimensional dynamic models

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Mauro Bernardi

Co-supervisore: Prof. Daniele Bianchi

Dottorando: Nicolas Bianco

Padova, 11 January 2023

Abstract

One of the most frequently discussed aspects of research in Bayesian statistics is the estimation of posterior distributions in the context of *big data analysis*. The main problems practitioners face are the computational limitations of classical approaches and the difficulty of selecting subsets of relevant variables from all possible ones. Markov chain Monte Carlo algorithms (MCMC) have been for a long time the most adopted tool to carry out Bayesian inference. However, when introducing ad hoc priors to regularize estimates and perform authomatic variable selection, the models became more complex and MCMC methods turn out to be computationally inefficient. Therefore, the interest moved towards *variational approximations* increased in the last decades. The latter are a family of deterministic approximations that allows for Bayesian inference in complex models in a reasonable amount of time, while preserving good accuracy in the posterior estimates.

This PhD thesis focuses on combining regularization and variable selection procedures with variational approximation techniques leading to novel algorithms to efficiently estimate high-dimensional models in the context of time series data. The first Chapter provides an overview of the existing literature on variational approximations. The second Chapter undertake the problem of estimation and prediction in multivariate high-dimensional dynamic regressions in low signal to noise ratio scenarios. The third Chapter shows an interesting use of parametric variational Bayes to introduce regularization and smoothness in the posterior estimates of a univariate stochastic volatility model. We emphasize the possible importance of this approach within an empirical finance application to portfolio management. The fourth, and last, Chapter proposes a novel semi-parametric variational Bayes algorithm to perform accurate dynamic variable selection in time-varying parameter regressions. The combination between the proposed variational approximation and the model specification yield remarkable theoretical properties, not achievable through classical MCMC methods.

Sommario

Uno degli aspetti più discussi negli ultimi anni dalla ricerca in statistica bayesiana è la stima delle distribuzioni a posteriori nel contesto dell'analisi dei *big data*. I problemi principali che i ricercatori devono affrontare sono i limiti computazionali degli approcci classici e la difficoltà di selezionare dei sottoinsiemi di variabili importanti tra tutte quelle disponibili. Gli algoritmi Markov chain Monte Carlo (MCMC) sono stati per lungo tempo lo strumento più adottato per l'inferenza bayesiana. Tuttavia, con l'introduzione di distribuzioni a priori ad hoc per regolarizzare le stime ed eseguire la selezione automatica delle variabili, i modelli sono diventati sempre più più complessi e i metodi MCMC si sono rivelati computazionalmente inefficienti. Per questo motivo, negli ultimi decenni è aumentato l'interesse verso le *approssimazioni variazionali*. Quest' ultime rappresentano una famiglia di approssimazioni deterministiche che consentono l'inferenza bayesiana anche per modelli complessi in un tempo ragionevole, ma, allo stesso tempo, preservando una buona accuratezza nelle stime a posteriori.

Questa tesi di dottorato si concentra sulla combinazione di procedure di regolarizzazione e selezione delle variabili con tecniche di approssimazione variazionale. L'accostamento di queste procedure dà vita nuovi algoritmi per stimare in modo efficiente modelli ad alta dimensionalità nel contesto delle serie temporali. Il primo Capitolo fornisce una panoramica della letteratura esistente sulle approssimazioni variazionali. Il secondo Capitolo affronta il problema della stima e della previsione nelle regressioni dinamiche multivariate ad alta dimensionalità in scenari con basso rapporto segnale/rumore. Il terzo Capitolo si focalizza su un interessante utilizzo delle tecniche variazionali parametriche per introdurre la regolarizzazione nelle stime a posteriori, ottenendo traiettorie più lisce, in un modello univariata con volatilità stocastica. L'importanza di questo approccio viene sottolineata nell'ambito di un'applicazione di finanza empirica alla gestione del portafoglio. Il quarto, e ultimo, Capitolo propone un nuovo algoritmo variazionale semi-parametrico per eseguire un'accurata selezione dinamica delle variabili nelle regressioni con parametri variabili nel tempo. La combinazione tra l'approssimazione variazionale proposta e la specificazione del modello produce notevoli proprietà teoriche, non ottenibili con l'utilizzo dei classici metodi MCMC.

To Stefano, Sabrina,
Chiara and Martina.

Contents

List of Figures

List of Tables

List of Algorithms

Introduction

Overview

Since the early 2000s, there has been an increasing interest in *big data* and *complex data analysis*. The motivation can be found in the exponential increase in the amount of data (Elezaj and Tole, 2018), both structured and unstructured, the advancement of technological infrastructure to store the databases, and the increasing computing power to process them (see the recent AlphaTensor software of Fawzi *et al.*, 2022). The complexity of the analysis lies in the different nature and source of the information. For example, data can be collected over time, space, or even more complex domains, and it can come as a continuous stream. Each of the above scenarios brings with it complications. Therefore, approaching the problem with appropriate statistical methods is essential in order to obtain reliable and clear results that are useful in practice. In this thesis we will focus on time series data, i.e., sequences of observations that are collected over a period of interest and whose dynamic characterizes a temporal dependence structure. This type of data is common to many areas of scientific interest since most phenomena develop over time. Understanding the factors that drive their temporal evolution is crucial to draw qualitative conclusions as well as provide predictions for future dynamics. Some topical examples include the forecasting of macroeconomical indicators such as inflation or unemployement (Fulton and Hubrich, 2021), climate change and environmental data (Mudelsee, 2019), genomics and neuroscience (Aguayo-Orozco *et al.*, 2018; Pourahmadi and Noorbaloochi, 2016), and financial investments analysis (Sezer *et al.*, 2020).

Although the most widely used methods for approaching the analysis, especially predictive analysis, of these phenomena of interest lie in the areas of machine learning, deep learning, and artificial intelligence, simpler statistical models can be just as effective. In particular, if the researcher's interest goes beyond prediction, the latter are more easily interpreted, while machine learning tools are often identified as *black boxes*. In fact, one of the biggest challenges in recent years is to make machine learning and artificial intelligence tools more interpretable. In this thesis, we focus on Bayesian estimation of univariate and multivariate regression models in the case of high-dimensional time series data. Dimensionality grows in two directions: we can deal with many variables (*fat data*) or we can collect a huge number of observations, i.e., long time series (*tall data*). In both cases, two major issues arise. Firstly, the computational time to perform an accurate inference increases a lot and could be prohibitive. Secondly, the over-parametrization of the model leads to lack of interpretability and poor prediction performances, therefore suitable model selection procedures are necessary to identify only the relevant variables to include in the analysis.

Bayesian statistics also suffers from these problems. To tackle the over-parametrization issue, Bayesian literature proposes to make use of ad hoc prior distributions. A first class consists of the so called continuous shrinkage priors. The latter are continuous distribution functions that concentrate a large part of the probability mass around zero, while preserving large tails. The effect of such priors is to shrink towards zero the posterior distribution of unimportant parameters, while providing accurate estimates for the relevant ones. Well known and widely used continuous shrinkage priors are the Bayesian lasso of Park and Casella (2008), the Normal-Gamma of Griffin and Brown (2010), and the Horseshoe of Carvalho *et al.* (2010). Beside the latter category, discrete priors can be defined to separate the true signal from the noise. The famous Spike-and-Slab prior (George and McCulloch, 1993; Ročková and George, 2018) and its variants fall into this family. Those approaches aim to specify a mixture of two components, namely *spike* and *slab* distributions. The first can be either a Dirac at zero or a continuous distribution centered at zero with small variance, while the second one is usually a vague continuous distribution. More recent developments in Bayesian model selection are the L_1–ball prior of Xu and Duan (2020), and methods for dynamic shrinkage (see e.g., Kalli and Griffin, 2014; Kowal *et al.*, 2019) and dynamic variable selection (Ročková and McAlinn, 2021; Koop and Korobilis, 2020). The first contribution considers a widely used regularization in high-dimensional statistics, namely the L_1, and it changes the problem of choosing the subset of non-zero parameters, which is a large combinatorial problem, into a simple continuous optimization. The second stream of research is proper of time series data. In these works, the authors assume that the importance of a variable can vary across time. This means that the set of relevant variables is not the same in all the time series.

While an effective solution has been proposed for the over-parametrization problem, traditional Bayesian estimation algorithms still suffer from reduced computational efficiency. The latter usually rely on Markov chain Monte Carlo (MCMC) methods, which provides a *stochastic approximation* of the posterior distribution (see Hastings, 1970;

Gelfand and Smith, 1990; Robert and Casella, 2004, for an extensive discussion). This means that its approximation can be made arbitrarily accurate by increasing the number of draws from the target distribution. Although this is an appealing property, MCMC methods are recognized to be computationally intensive and do not scale efficiently for both fat and tall data. Moreover, its convergence should be carefully checked. These main issues gave rise to the diffusion of *deterministic approximations*, that represent a valid alternative to perform Bayesian inference. Among them, *variational approximation* have attracted the attention of the statistical community so far.

Variational methods, owe the name to variational calculus, a mathematical discipline which treats the optimization of functionals, and they consist of a set of algorithms which aim to approximate probability distributions through deterministic optimization rather than stochastic simulation. Two remarkable paradigms are variational Bayes (Ormerod and Wand, 2010; Blei *et al.*, 2017) and expectation propagation (Minka, 2001). On the one hand, variational methods usually provide faster solutions than simulation based methods, but, on the other hand, they have a bounded accuracy compared to MCMC, which can be made more and more accurate by simply increasing the Monte Carlo sample size. Variational approximations require to define a family of approximating densities and to find the element that minimizes a divergence with the target distribution. The most famous variational inference technique is probably variational Bayes (VB), which consists of minimizing the Kullback–Leibler ($\mathcal{KL}$) divergence (Kullback and Leibler, 1951). Alternative divergence measures have been proposed by (Dieng *et al.*, 2017; Minka *et al.*, 2005) within the context of expectation propagation (EP).

Variational inference has been used in a wide range of applications, ranging from statistics (Rustagi, 1976) to quantum mechanics (Sakurai, 1994), statistical mechanics (Parisi, 1988), machine learning (Hinton and Van Camp, 1993) and then generalized to many probabilistic models, taking advantage of the graphical models' representation (Jordan *et al.*, 1999). An up to date comprehensive introduction to the topic of variational methods from a statistical perspective is provided by Ormerod and Wand (2010), Blei *et al.* (2017), and more recently Zhang *et al.* (2019), which refers to all the aforementioned works. Another interesting research field consists of investigating the statistical properties of variational methods in order to make them theoretically grounded. In this direction, recent papers by Wang and Blei (2018) and Zhang and Gao (2020) provide a complete and interesting discussion on this topic.

In this thesis we focus on variational Bayes technique. Beside the specification of a divergence, which in this case is the Kullback–Leibler, this paradigm require to define a suitable family of approximating densities, namely $\mathcal{Q}$. According to the specification of $\mathcal{Q}$, we fall into different approaches. If the element $q \in \mathcal{Q}$ is assumed to be a prespecified parametric distribution, then parametric variational Bayes emerges (Ormerod and Wand, 2012). A widely used parametric approximation requires that q belongs to the family of multivariate Gaussian distributions. This gives rise to Gaussian variational approximations (GVA) method (Wand, 2014). Conversely, suppose no parametric assumption on the elements in $\mathcal{Q}$ is defined, but let $\mathcal{Q}$ be the set of probability density functions which factorize according to a given partition of the parameter set; in that case, a non-parametric approximation is assumed. The latter paradigm is called mean–field variational Bayes (MFVB). Its name is inspired by the mean–field approximations developed in statistical physics (Parisi, 1988). Note that one can also mix the two different approaches presented above and derive a semi–parametric variational Bayes algorithm (Rohde and Wand, 2016).

Although the literature on Bayesian variable selection and variational approximations is already vast, the high level of research interest in this direction enlarges it day-by-day.

Main contributions of the thesis

This thesis focuses on two main stream of literature in Bayesian analysis: the problem of regularization and variable selection, and the recent advances in variational approximations, in particular within the variational Bayes paradigm. These arguments are explored together in the context of high-dimensional time series data. The thesis is organized into four, self-contained, main Chapters, together with an extensive appendix. Chapter 1 serves as an introduction to variational approximations within the context of Bayesian inference, focusing in particular on mean–field approach and parametric approximations.

In Chapter 2 we propose a new variational algorithm to estimate large-scale multivariate linear regressions with continuous shrinkage priors. We develop both the theory and the algorithms to estimate large multivariate regression models under the assumption of several, and widely used in literature, continuous shrinkage priors, leveraging a variational Bayes approach to the inference. The motivation of this project is related to the computational issues that state-of-the-art approaches face in high-dimensional settings. In particular, when the dimension of the response vector (d) and the number of covariates (p) increases, commonly used MCMC algorithms are computationally slow and the inference may be prohibitive in a reasonable amount of time. To overcome the latter issue, state-of-the-art approaches relies on simplifications of the model through suitable re-parametrization of the matrix of regression coefficients (see Gefang *et al.* (2019) and

Chan and Yu (2022) within the context of vector autoregressions). These assumptions have the merit of simplifying the estimation, making feasible the implementation of standard MCMC algorithms. However, this procedure reduces the flexibility of the model, leading to unreliable estimates and poor prediction performances. Unlike existing methods, our approach does not rely on a transformation of the original parameters space to reduce the computational burden, but it exploits variational Bayes approximations. This allows us to elicit continuous shrinkage priors directly on the parameters of interest and perform more accurate inference. Our proposed algorithm is available for several commonly used priors such as the Bayesian lasso of Park and Casella (2008) and its adaptive extension (Leng *et al.*, 2014), the normal-gamma of Griffin and Brown (2010) and its recent improvements (Griffin and Brown, 2017; Bitto and Frühwirth-Schnatter, 2019), as well as the horseshoe prior of Carvalho *et al.* (2010). An extensive simulation study provides evidence that our approach produces more accurate estimates of the regression coefficients under different sparsity assumptions. This leads to higher interpretability of the results and precise estimate of the conditional mean, which is of key importance when making forecasts. To validate the proposed methodology we consider an application in finance, where large datasets are available. Specifically, we investigate both the statistical and economic significance of our estimation approach within the context of a representative investor who faces the choice of investing in a large set of different industry portfolios. Both the simulation and empirical results hold across different prior specification and model dimensions.

Chapter 3 focuses on a flexible estimation of the latent process that governs the heteroschedasticity in univariate regressions. The latter is known as stochastic volatility model. The common assumption in literature for the latent process is a random walk dynamic or an autoregressive process of order one. This choice is justified by empirical evidence. The novelty of this Chapter consists in the implementation of a parametric variational Bayes algorithm to provide accurate *global* approximation of the latent volatility process. The latter has two main advantages. Firstly, it provides complete posterior Bayesian inference in a reduced amount of time with respect to classical MCMC approaches. Second, the proposed method generalizes the recent approximation scheme of Chan and Yu (2022), thus providing higher accuracy in approximating the posterior distribution. In addition, we propose a general formulation for the mean vector of the variational distribution, so that it can be customised according to the research interest. In particular, we are interested in obtaining arbitrary smooth estimates for the volatility process. The motivation behind the need of smooth estimates comes from a practical problem in empirical finance. Volatility managed portfolios (Moreira and Muir, 2017;

Cederburg *et al.*, 2020) is a recent studied topic in empirical finance and it aims to reduce the losses an investor faces by re-scaling the expected return of a portfolio by the volatily. This permits to avoid dangerous dropdowns. However, the estimate of the volatility impacts the performance of the strategy. Although commonly used realized volatility (RV) provides promising results in simple scenarios, in practice it represent a non-viable solution. This is due to the fact that this estimate is highly sensible to peaks. Therefore the scaling factor, which can be seen as the weight of the portfolio, is too variable, leading to high turnover which implies unsustainable transaction costs. Our contribution of providing an unified framework to estimate volatility measure with arbitrary smoothness helps the investor to reduce transaction costs, while maintaining benefits from a volatility managed portfolio investment strategy. One point we want to stress is the following. Within our variational Bayes approach the smoothness is introduced in the posterior estimates during the inference through a parametric approximation. The latter is not the case in standard Bayesian inference: in this case one should change the model, perhaps introducing a misspecification, or assume informative prior distribution on the parameters of the latent process.

The last contribution is covered in Chapter 4. We propose a semi–parametric variational Bayes algorithm to deal with dynamic variable selection in time varying parameters regressions with many covariates. The Bernoulli-Gaussain model of Ormerod *et al.* (2017) is a valid alternative to discrete Spike-and-Slab prior in Bayesian variable selection. It has been proven to well separate the true signal from the noise compared to both Bayesian and frequentist approaches, and, surprisingly, it does not require to cross-validate fixed hyper-parameters. However, the framework considered in Ormerod *et al.* (2017) is a static linear regression. Turning into a more complex dynamic regressions with time varying parameters, the works of Kalli and Griffin (2014) and, more recently, Kowal *et al.* (2019) consider shrinkage rather than sparsity, while en effective variable selection has been proposed in Koop and Korobilis (2020) and Ročková and McAlinn (2021). The latter rely on Spike-and-Slab type priors, whose main drawback is the tuning of hyper-parameters controlling the variance of the spike and slab component. This procedure is difficult in high-dimensional setting and it may strongly affect the results. In this Chapter we extend the Bernoulli-Gaussian model to deal with time varying coefficients in dynamic linear regressions. The name of the model arises from the fact that the data generating process consists of two equations that control the evolution of the regression parameters and the inclusion variable indicator, respectively. The first one follows a Gaussian random walk, while the second is build on a sequence of correlated Bernoulli distributions. The product of the two processes tells us the value of the coefficient at time t and whether it is included or not in the current model. We study the theoretical properties of the algorithm and we show that this framework is particularly suitable for regressions with many predictors. In fact, we recover two main results: one concerns the achievement of sparse estimates, and the other one ensures dimensionality reduction of the problem. A comparison with MCMC methods revealed that the latter properties only hold under a variational Bayes paradigm. Therefore, not only VB represents a computationally faster alternative to MCMC, it even provides more accurate and efficient estimates in this scenario. We investigate the performance of the model to separate the true signal from the noise, i.e., dynamically select the correct subset of active variables, compared with established methods (Ročková and McAlinn, 2021; Koop and Korobilis, 2020), for different dimensions and signal-to-noise ratio. We also show the empirical importance of our method through real-data applications.

Three appendices complete this PhD thesis and each one refers to a corresponding Chapter. Appendices contains derivations of the optimal variational densities and evidence lower bound (ELBO), and the proofs of propositions. Together with analytical results, appendices show additional simulations and empirical analysis insights.

Chapter 1

An introduction to variational Bayes

The aim of this Chapter is to present the general concept of the variational approximation techniques that will be used in the continuation of the thesis. Starting from the basics, we first spend few lines to recall the Bayesian paradigm. Let $\mathbf{y}$ be an observed data vector of realizations from a random variable Y with probability density function $p(\mathbf{y}|\boldsymbol{\vartheta})$, where $\boldsymbol{\vartheta} \in \Theta$ is a p-dimensional parameter vector with values in $\Theta \subseteq \mathbb{R}^p$. Moreover, let $p(\boldsymbol{\vartheta})$ be the *prior* distribution summarizing all the a priori knowledge about model parameters. The Bayes theorem allows to update the prior beliefs to account for the evidence coming from the observed data into the *posterior* density function:

$$p(\boldsymbol{\vartheta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta})}{p(\mathbf{y})} = \frac{p(\mathbf{y},\boldsymbol{\vartheta})}{p(\mathbf{y})}, \tag{1.1}$$

where $p(\mathbf{y},\boldsymbol{\vartheta})$ is a joint density function and $p(\mathbf{y})$ is called *marginal likelihood*, or *model evidence*, and it is defined as:

$$p(\mathbf{y}) = \int_{\Theta} p(\mathbf{y},\boldsymbol{\vartheta})\, d\boldsymbol{\vartheta}. \tag{1.2}$$

The computation of $p(\mathbf{y})$ requires the marginalization over the parameter space, whose dimension makes the above integral intractable in many situations. As a consequence, an analytical expression for the posterior $p(\boldsymbol{\vartheta}|\mathbf{y})$ cannot be derived for most statistical models. The rise of MCMC methods, and its variants, (Hastings, 1970; Duane *et al.*, 1987; Gelfand and Smith, 1990; Robert and Casella, 2004) provided a way to sample from the posterior distributions regardless of the model complexity and parameter dimensions. Although these methods are appealing since they are able to approximate the true posterior distribution, they face some relevant issues in practical implementations. In fact, they are usually computationally intensive even for moderately complex models, with slow and not guaranteed convergence, therefore compromising the validity of inference. The latter issues are even more emphasized with large sample size datasets. Gibbs sampling (Casella and George, 1992) represents a viable solution to make inference faster, but its effective implementation is conditional on convenient prior specifications, and therefore an efficient estimation is not always possible. Moreover, even in compatible cases, it may require computationally expensive procedures such as sampling from complicated distributions.

Variational approximations represent a family of deterministic methods for approximating the posterior distribution. The latter rely on optimization rather than sampling and are usually faster than MCMC algorithms, but on the other hand suffer from bounded accuracy in approximating the true posterior distribution. Variational approximations do not aim to estimate the true posterior density, but a manageable distribution which is close to the target one. The closeness is measured according to a divergence measure $\mathcal{D}(p||q) : \mathcal{P} \times \mathcal{Q} \to \mathbb{R}^+$, where p and q are generic density functions belonging to the sets $\mathcal{P}$ and $\mathcal{Q}$, respectively. Note that $\mathcal{D}(\cdot||\cdot)$ is not necessarily symmetric, i.e., $\mathcal{D}(p||q) \neq \mathcal{D}(q||p)$. The variational approximations considered in the next Chapters of this thesis have the common goal to find the distribution $q(\boldsymbol{\vartheta}) \in \mathcal{Q}$ such that it is close to the posterior $p(\boldsymbol{\vartheta}|\mathbf{y})$ according to a divergence measure:

$$q^*(\boldsymbol{\vartheta}) = \arg \min_{q(\boldsymbol{\vartheta}) \in \mathcal{Q}} \mathcal{D}(p(\boldsymbol{\vartheta}|\mathbf{y})||q(\boldsymbol{\vartheta})), \tag{1.3}$$

where $q^*(\boldsymbol{\vartheta})$ is called *optimal variational density* and it represents the best approximation to the true posterior distribution among the possible, given the set of candidates $\mathcal{Q}$. The accuracy of the approximation can be quantified through the measure proposed in Wand *et al.* (2011):

$$\mathcal{ACC}(\boldsymbol{\vartheta}) = 100 \left\{ 1 - 0.5 \int |q(\boldsymbol{\vartheta}) - p(\boldsymbol{\vartheta}|\mathbf{y})| \, d\boldsymbol{\vartheta} \right\} \%, \tag{1.4}$$

where $q(\boldsymbol{\vartheta})$ is the proposed approximation and $p(\boldsymbol{\vartheta}|\mathbf{y})$ is the true posterior distribution. Notice that in practice $p(\boldsymbol{\vartheta}|\mathbf{y})$ is determined as the sampled distribution via MCMC with a large number of draws.

It is immediate to understand that changes in $\mathcal{D}$ and $\mathcal{Q}$ lead to different paradigms. All the methodologies used in this thesis assume $\mathcal{D}$ to be the well known Kullback-Leibler divergence (Kullback and Leibler, 1951):

$$\mathcal{KL}(p||q) = \int_\Theta p(\boldsymbol{\vartheta}|\mathbf{y}) \log \left\{ \frac{p(\boldsymbol{\vartheta}|\mathbf{y})}{q(\boldsymbol{\vartheta})} \right\} d\boldsymbol{\vartheta}. \tag{1.5}$$

In what follows we present different approaches, namely *mean–field* and *parametric* variational Bayes. As regards notation, we denote with $\boldsymbol{\mu}_{q(\vartheta)} = \mathbb{E}_q(\boldsymbol{\vartheta}) = \int_\Theta \boldsymbol{\vartheta} q(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}$ and $\boldsymbol{\Sigma}_{q(\vartheta)} = \mathsf{Var}_q(\boldsymbol{\vartheta})$ the expectation and variance of the optimal variational density, while, more in general, $\boldsymbol{\mu}_{q(f(\vartheta))} = \mathbb{E}_q(f(\boldsymbol{\vartheta})) = \int_\Theta f(\boldsymbol{\vartheta}) q(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}$ denotes the expected value of the random variable $f(\boldsymbol{\vartheta})$ with respect to the variational density q.

Variational Bayes (VB) approach to inference requires to minimize the Kullback-Leibler ($\mathcal{KL}$) divergence between an approximating density $q(\boldsymbol{\vartheta})$ and the true posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$, (see, e.g., Ormerod and Wand, 2010; Blei *et al.*, 2017). Note that the $\mathcal{KL}$ divergence in (1.5) cannot be directly minimized with respect to $\boldsymbol{\vartheta}$ because it involves the expectation with respect to the unknown true posterior distribution. Ormerod and Wand (2010) show that

$$\mathcal{KL}(p||q) = \log p(\mathbf{y}) - \log \underline{p}\left(\mathbf{y}; q(\boldsymbol{\vartheta})\right), \quad \underline{p}\left(\mathbf{y}; q(\boldsymbol{\vartheta})\right) = \mathbb{E}_q\left[\log p(\mathbf{y}, \boldsymbol{\vartheta})\right] - \mathbb{E}_q\left[\log q(\boldsymbol{\vartheta})\right],$$

where $\underline{p}\left(\mathbf{y}; q(\boldsymbol{\vartheta})\right)$ is called variational (or evidence) lower bound (ELBO). Therefore, the problem of minimizing $\mathcal{KL}$ can be equivalently stated as the maximization of the ELBO.

Now, depending on the assumption on the space $\mathcal{Q}$, we fall into different variational paradigms.

Mean-field variational Bayes. The MFVB approach only assumes a non-parametric restriction for the variational density

$$q(\boldsymbol{\vartheta}) = \prod_{i=1}^{m} q_i(\boldsymbol{\vartheta}_i), \tag{1.6}$$

for a partition $\{\boldsymbol{\vartheta}_1, \dots, \boldsymbol{\vartheta}_m\}$ of the parameter vector $\boldsymbol{\vartheta}$. Under the MFVB restriction, a closed form expression for the optimal variational density of each component $q(\boldsymbol{\vartheta}_j)$ is defined as:

$$q^*(\boldsymbol{\vartheta}_j) \propto \exp\left\{\mathbb{E}_{q^\star(\boldsymbol{\vartheta}\setminus\boldsymbol{\vartheta}_j)}\Big[\log p(\mathbf{y}, \boldsymbol{\vartheta})\Big]\right\}, \qquad q^\star(\boldsymbol{\vartheta} \setminus \boldsymbol{\vartheta}_j) = \prod_{\substack{i=1 \\ i\neq j}}^{p} q_i(\boldsymbol{\vartheta}_i), \tag{1.7}$$

where the expectation is taken with respect to the joint approximating density with the j-th element of the partition removed $q^\star(\boldsymbol{\vartheta}\setminus\boldsymbol{\vartheta}_j)$. This allows to implement an coordinate ascent variational inference (CAVI) algorithm, an iterative procedure to estimate the optimal density $q^*(\boldsymbol{\vartheta})$. A valid alternative to (1.7) is given by:

$$q(\boldsymbol{\vartheta}_j) \propto \exp\left\{\mathbb{E}_{-\boldsymbol{\vartheta}_j}\left[\log p(\boldsymbol{\vartheta}_j|\text{rest})\right]\right\}, \tag{1.8}$$

where $p(\boldsymbol{\vartheta}_j|\text{rest})$ denotes the full conditional distribution of $\boldsymbol{\vartheta}_j$, i.e., the distribution of $\boldsymbol{\vartheta}_j$ given all the other parameters.

Parametric variational Bayes. An alternative to mean-field approach consists in imposing a parametric family of distributions for the set $\mathcal{Q}$. In this case, $q(\boldsymbol{\vartheta}) = q(\boldsymbol{\vartheta}|\boldsymbol{\theta}_{q(\vartheta)})$, where $\boldsymbol{\theta}_{q(\vartheta)}$ is called variational parameter. Under this parametric approach, the aim is to find the best $\boldsymbol{\theta}_{q(\vartheta)}$, namely $\hat{\boldsymbol{\theta}}_{q(\vartheta)}$ such that $q(\boldsymbol{\vartheta})$ is close to the posterior and belongs to the pre-specified parametric family. The complexity of the optimization depends on the size of the vector $\boldsymbol{\theta}_{q(\vartheta)}$.

Among the all possible choices of the parametric family, the Gaussian distribution is the mostly used, so that, under the variational approximation, $\boldsymbol{\vartheta} \sim \mathsf{N}_p(\boldsymbol{\mu}_{q(\vartheta)}, \boldsymbol{\Sigma}_{q(\vartheta)})$. This setting has been extensively studied in the literature. Wand (2014) derived a fixed-point iteration update scheme, while more recently Rohde and Wand (2016) provide an alternative iterative updating algorithm for the variational parameters when the approximating density is a multivariate gaussian $\mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$\boldsymbol{\Sigma}^{new} = \left[\nabla^2_{\boldsymbol{\mu},\boldsymbol{\mu}} S(\boldsymbol{\mu}^{old}, \boldsymbol{\Sigma}^{old})\right]^{-1} \tag{1.9}$$

$$\boldsymbol{\mu}^{new} = \boldsymbol{\mu}^{old} + \boldsymbol{\Sigma}^{new} \nabla_{\boldsymbol{\mu}} S(\boldsymbol{\mu}^{old}, \boldsymbol{\Sigma}^{old}), \tag{1.10}$$

where $\nabla_{\boldsymbol{\mu}} S(\boldsymbol{\mu}^{old}, \boldsymbol{\Sigma}^{old})$ and $\nabla^2_{\boldsymbol{\mu},\boldsymbol{\mu}} S(\boldsymbol{\mu}^{old}, \boldsymbol{\Sigma}^{old})$ denote the first and second derivative of $S(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\mu}$ and evaluated at $(\boldsymbol{\mu}^{old}, \boldsymbol{\Sigma}^{old})$. The function S is called *non-entropy function* which is computed as the expected value of the logarithm of the joint distribution of the data and parameters with respect to the variational density q, i.e., $\mathbb{E}_q(\log p(\mathbf{y}, \boldsymbol{\vartheta}))$.

The two variational Bayes paradigms illustrated above are non strictly alternatives. In fact, one can combine mean-field and parametric approximations to end up with a semi-parametric approximation approach (see, e.g., Wand, 2014; Menictas and Wand, 2015; Luts and Wand, 2015).

Chapter 2

Large-scale multivariate regressions with shrinkage priors

2.1 Introduction

Within a Bayesian linear regression context, parameters regularization is often based on continuous shrinkage priors (see, for instance Park and Casella, 2008; Griffin and Brown, 2010; Carvalho *et al.*, 2010; Korobilis, 2013; Bhattacharya *et al.*, 2015; Hahn and Carvalho, 2015; Griffin and Brown, 2017, among others). In multivariate settings, the use of these priors often relies on a Cholesky decomposition of the residuals covariance matrix. This allows to break down a potentially large system of equations into a sequence of linear univariate regressions. Linearity is preserved assuming a tight parametrization of the regression coefficients based on the Cholesky factor. While this greatly simplifies posterior inference, it potentially prevents to directly recover the structure of the original regression parameters.

In this Chapter, we take a different approach towards the identification of the regression coefficients in multivariate regressions. More specifically, we propose a novel variational Bayes inference procedure which allows for fast and accurate posterior estimates of the regression parameters under hierarchical shrinkage priors. Our approach still leverages on the computational convenience of the Cholesky factorisation, but does not build upon a linearized system of equations. This allows to elicit standard hierarchical shrinkage priors directly on the matrix of regression coefficients, unlike in existing Bayesian inference schemes.

We first investigate the performance of our estimation procedure based on an extensive simulation study. We compare our variational Bayes approach (`VB` henceforth), against two alternative methods which are representative of state-of-the-art Bayesian estimation in the context of multivariate time series models. The first approach is a

Markov Chain Monte Carlo (MCMC) algorithm as in Cross *et al.*, 2020. The second competing approach is based on a *linearised* variational Bayes (LVB henceforth) method as originally proposed by Gefang *et al.* (2019); Chan and Yu (2022). Both approaches rely on a Cholesky-based transformation of the original regression parameters.

In addition to a standard normal prior, we consider several hierarchical shrinkage priors, such as the Bayesian adaptive lasso proposed by Leng *et al.* (2014), an adaptive version of the normal-gamma prior of Griffin and Brown (2010), and the horseshoe prior as originally proposed by Carvalho *et al.* (2009, 2010).

The simulation results show that our variational Bayes estimation procedure outperforms competing approaches, both in a mean squared sense and when it comes to identify the "true" signals: select those covariates which carry some significant predictive power. Perhaps more interestingly, our approach provides posterior estimates for the regression coefficients which are invariant to permutation of variables. On the other hand, the simulation results show that the performance of both MCMC and LVB is not permutation-invariant. The latter is a consequence of the fact that hierarchical shrinkage priors are elicited on a non-linear transformation of the regression parameters rather than on the original parameters.

Intuitively, a more accurate estimate of the regression coefficients should be of first-order importance for forecasting. We investigate both the statistical and economic value of the forecasts from our variational Bayes approach within the context of a representative investor who faces the choice of investing in a large set of different industry portfolios. Although the model is general and can be applied to any type of asset returns, as far as data are stationary, our focus on industry portfolios is motivated by keen interests from researchers (see, e.g., Fama and French, 1997) and practitioners alike.

Perhaps surprisingly, while there is a vast literature examining the out-of-sample performance of the aggregate or individual stock excess returns, the question of whether industry portfolios can be predicted has received relatively little attention so far. However, the implications of industry returns predictability are far from trivial. If all industries are unpredictable, then the market return, which is a weighted average of the industry portfolios, should also be unpredictable. As a result, the abundant evidence of aggregate market return predictability, implies that at least some industry portfolio returns is predictable. This could have important implications for asset pricing models and the efficient allocation of capital across sectors.

The empirical results show that more accurate estimates of the regression coefficients translate into better out-of-sample forecasts. This is reflected not only in higher out-of-sample R^2_{oos}s, calculated comparing against a naive rolling mean forecast as proposed Campbell and Thompson (2007), but also in higher economic performances. The latter is shown by larger certainty equivalent returns for the vast majority of industry portfolios. This result supports our view that by a more accurate identification of weak correlations in asset returns, one can significantly improve, both statistically and economically, the out-of-sample performance of investment decisions based on large-scale regression models.

This Chapter connects to two main streams of literature. The first relates to the use of Bayesian methods to estimate large-scale linear regression models. A non-exhaustive list of works on the topic contains Zou (2006); Park and Casella (2008); Carvalho, Polson and Scott (2009, 2010); Griffin and Brown (2010); Polson and Scott (2011); Leng, Tran and Nott (2014); Bhattacharya, Chakraborty and Mallick (2016); Tang, Ghosh, Xu and Ghosh (2018); Bitto and Frühwirth-Schnatter (2019), among others. We extend this literature and provide an accurate variational Bayes estimation method which allows to elicit existing hierarchical shrinkage priors without relying on a Cholesky-based transformation of the regression parameters.

A second strand of literature we contribute to is related to the predictability of stock returns (see, e.g., Goyal and Welch, 2008; Rapach *et al.*, 2010; Dangl and Halling, 2012; Johannes *et al.*, 2014; Pettenuzzo *et al.*, 2014; Smith and Timmermann, 2021, among others). More specifically, we contribute to the ongoing struggle to understand the dynamics of risk premiums by looking at industry-based portfolios. As highlighted by Lewellen *et al.* (2010), the time series variation of industry portfolios is particularly problematic to measure, since conventional risk factors do not seem to capture significant comovements and cross-signals which might improve out-of-sample predictability (see, e.g., Bianchi and McAlinn, 2020). Early exceptions are Ferson and Harvey (1991), Ferson and Korajczyk (1995) and Ferson and Harvey (1999), which use a set of industry portfolio as test assets to look at the in-sample explanatory power of macroeconomic risk factors. Using a standard Bayesian approach, Avramov (2004) explores the predictive content of standard Fama-French risk factors for a handful of industry portfolios and investigate the implications for asset allocation decisions. We extend this literature by investigating the out-of-sample predictability of industry portfolios through the lens of a novel estimation method for large-scale multivariate predictive regressions.

2.2 Choosing the model parametrization

Let $\mathbf{y}_t = (y_{1,t}, \dots, y_{d,t})^\intercal \in \mathbb{R}^d$ be a multivariate Gaussian random variable and let $\mathbf{x}_t = (1, x_{1,t}, \dots, x_{r,t})^\intercal \in \mathbb{R}^{(r+1)}$ be a vector of exogenous covariates observed at time t. A multivariate predictive regression model with constant volatility is defined as follows:

$$\mathbf{y}_t = \boldsymbol{\Theta}\mathbf{z}_{t-1} + \mathbf{u}_t, \qquad \mathbf{u}_t \sim \mathsf{N}_d\left(\mathbf{0}_d, \boldsymbol{\Omega}^{-1}\right), \qquad t = 2, 3, \dots, \tag{2.1}$$

with $\boldsymbol{\Theta} = (\boldsymbol{\Gamma}, \boldsymbol{\Phi})$ being a $d \times p$ matrix of regression coefficients where $p = d + r + 1$. In particular, $\boldsymbol{\Gamma} \in \mathbb{R}^{d\times(r+1)}$ is the matrix of regression parameters for the exogenous predictors and $\boldsymbol{\Phi} \in \mathbb{R}^{d\times d}$ is the transition matrix containing the autoregressive parameters, so that $\mathbf{z}_{t-1} = (\mathbf{x}_{t-1}^\intercal, \mathbf{y}_{t-1}^\intercal)^\intercal$. Here, $\mathbf{u}_t \in \mathbb{R}^d$ is a sequence of uncorrelated stochastic innovation terms such that $\mathbf{u}_{t-k} \perp \mathbf{u}_{t-j}$ for $k \neq j$ and $k, j = \pm1, \pm2, \dots$ and covariance matrix equal to $\boldsymbol{\Omega}^{-1}$, with $\boldsymbol{\Omega} \in \mathbb{S}^d_{++}$ being a symmetric and positive definite precision matrix.

The modified Cholesky factorization of the precision matrix $\boldsymbol{\Omega}$ can be conveniently exploited to re-write the model in (2.1) with orthogonal innovations, (see, e.g., Rothman *et al.*, 2010). Let $\boldsymbol{\Omega} = \mathbf{L}^\intercal\mathbf{V}\mathbf{L}$, where $\mathbf{L} \in \mathbb{R}^{d\times d}$ is uni-lower-triangual and $\mathbf{V} \in \mathbb{S}^d_{++}$ is diagonal. Multiply both sides of (2.1) by $\mathbf{L} = \mathbf{I}_d - \mathbf{B}$. After some simple algebra one can obtain two alternative parametrizations of the same model:

$$\mathbf{y}_t = \mathbf{B}(\mathbf{y}_t - \boldsymbol{\Theta}\mathbf{z}_{t-1}) + \boldsymbol{\Theta}\mathbf{z}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathsf{N}_d(\mathbf{0}_d, \mathbf{V}^{-1}), \tag{2.2a}$$

$$\mathbf{y}_t = \mathbf{B}\mathbf{y}_t + \mathbf{A}\mathbf{z}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathsf{N}_d(\mathbf{0}_d, \mathbf{V}^{-1}), \tag{2.2b}$$

where $\mathbf{A} = \mathbf{L}\boldsymbol{\Theta}$ and $\mathbf{B}$ has a strict-lower-triangular structure with elements $\beta_{j,k} = -l_{j,k}$ for $j = 2, \dots, d$ and $k = 1, \dots, j-1$. The key difference is that (2.2a) shows non-linearity in the parameters, while (2.2b) is linear. More importantly, (2.2b) is the parametrization that is often used in state-of-the-art MCMC and variational Bayes estimations methods (see, e.g., Gefang *et al.*, 2019; Chan and Yu, 2022), whereas (2.2a) is the parametrization at the core of our variational Bayes approach. From (2.2) one can obtain an equation-by-equation representation in which the j-th component of $\mathbf{y}_t$ becomes:

$$y_{j,t} = \boldsymbol{\beta}_j\mathbf{r}_{j,t} + \boldsymbol{\vartheta}_j\mathbf{z}_{t-1} + \varepsilon_{j,t}, \qquad \varepsilon_{j,t} \sim \mathsf{N}(0, 1/\nu_j), \tag{2.3a}$$

$$y_{j,t} = \boldsymbol{\beta}_j\mathbf{y}_t^j + \mathbf{a}_j\mathbf{z}_{t-1} + \varepsilon_{j,t}, \qquad \varepsilon_{j,t} \sim \mathsf{N}(0, 1/\nu_j), \tag{2.3b}$$

for all $j = 1, \dots, d$ and $t = 2, 3, \dots,$ where $\boldsymbol{\beta}_j \in \mathbb{R}^{j-1}$ is a row vector containing the non-null elements in the j-th row of $\mathbf{B}$, $\boldsymbol{\vartheta}_j$ and $\mathbf{a}_j$ denote the j-th row of $\boldsymbol{\Theta}$ and $\mathbf{A}$

respectively. For any $j = 1, \ldots, d$, let $\mathbf{r}_{j,t} = \mathbf{y}_t^j - \mathbf{\Theta}^j \mathbf{z}_{t-1}$ denotes the the vector of residuals up to the $(j-1)$-th regression, with $\mathbf{y}_t^j = (y_{1,t}, \ldots, y_{j-1,t})^\intercal \in \mathbb{R}^{j-1}$ being the sub-vector of $\mathbf{y}_t$ collecting the variables up to the $(j-1)$-th and $\mathbf{\Theta}^j \in \mathbb{R}^{(j-1)\times d}$ is the sub-matrix containing the first $j-1$ rows of $\mathbf{\Theta}$.

Notice that although it is not strictly needed for the development of the variational approximation, we assume the data generating process to be weakly stationary and ergodic. In addition, since we are primarily interested in the identification of the regression matrix $\mathbf{\Theta}$, we consider for simplicity that each of the elements in $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)$ are time invariant (see, e.g., Smith and Timmermann, 2021). This assumption can be relaxed by assuming each $\nu_j^{-1}, j = 1, \ldots, d$ as a latent process and leverage standard stochastic volatility modeling (see, e.g., Clark, 2011; Chan and Eisenstat, 2018; Carriero, Clark and Marcellino, 2019). We leave this development for future research.

Existing Bayesian inference approaches for high-dimensional models usually rely on the linear parametrization in (2.2b), and therefore consider the elements in $\mathbf{A}$ as the parameters of interest. This has the merit of simplifying the estimation procedure making feasible the efficient implementation of standard MCMC (see, e.g., Chan and Eisenstat, 2018) and linearized variational Bayes (`LVB`) algorithms (see, e.g., Chan and Yu, 2022). Under the parametrization $\mathbf{A} = \mathbf{L\Theta}$, each element $\vartheta_{i,j}$, which denotes the (i,j)-entry of $\mathbf{\Theta}$, is computed as a linear combination $\vartheta_{i,j} = a_{i,j} + \sum_{k=1}^{i-1} c_{i,k} a_{k,j}$, where $a_{i,j}$ and $c_{i,j}$ denote the (i,j)-entry of $\mathbf{A}$ and $\mathbf{L}^{-1}$, respectively.

However, this raises two main issues: *i)* $a_{i,j} = 0$ does not imply $\vartheta_{i,j} = 0$, i.e., a shrinkage prior on $\mathbf{A}$ does not preserve the true structure of $\mathbf{\Theta}$; and *ii)* the estimate $\mathbf{\Theta} = \mathbf{L}^{-1}\mathbf{A}$ is not permutation invariant, which is a direct consequence of the Cholesky factorization. Figure 2.1 provides a numerical representation of this argument. We compare the posterior estimates obtained from a `LVB` method based on (2.2a) versus our `VB` approach based on (2.2b), for two different permutations of $\mathbf{y}_t$.

The evidence confirms that the estimates based on the transformation $\mathbf{\Theta} = \mathbf{L}^{-1}\mathbf{A}$ do not match with the true $\mathbf{\Theta}$. In addition, the estimates are influenced by the variables permutation. Instead, our `VB` approach provides a more accurate, permutation invariant, identification of $\mathbf{\Theta}$. Before taking this intuition to task both in simulation and on real stock returns, in the next Section we provide details of our estimation approach with different hierarchical shrinkage priors.

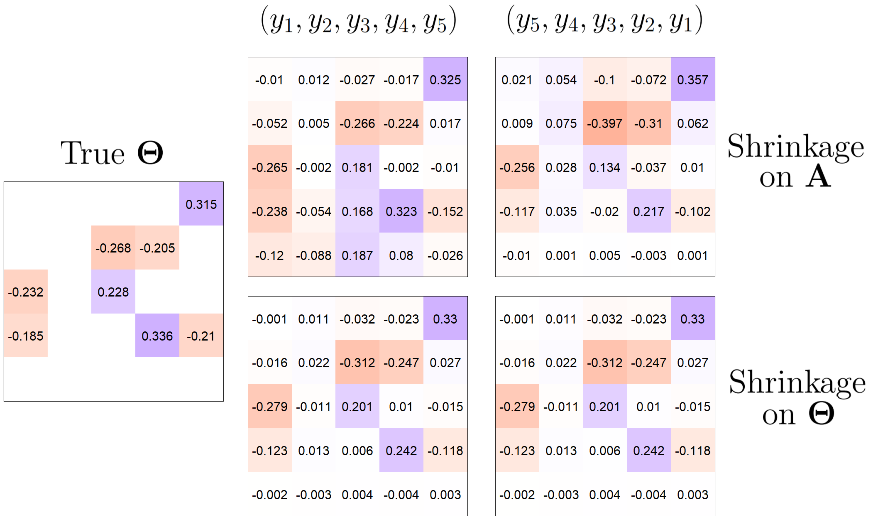


FIGURE 2.1: Comparison between the posterior inference for the linear representation $\mathbf{A} = \mathbf{L}\mathbf{\Theta}$ (first row) and the original parametrization $\mathbf{\Theta}$ (second row), for two different permutations of $\mathbf{y}_t$.

2.3 Variational Bayes inference

In this Chapter we adopt a mean–field variational Bayes approach to carry out approximate inference on the posterior distribution of the parameters. As described in Chapter 1, the factorization of the joint variational density q should be specified and it plays a central role in developing a MFVB algorithm. In the following, we present a factorization of the variational density for a non-informative prior, as well as the three alternative hierarchical shrinkage priors for $\mathbf{\Theta}$. In the main text we will summarize the optimal approximating density $q^\star$, show how to perform approximate inference on $\mathbf{\Omega}$, and illustrate how to make predictions within this framework. For the interested reader, in Appendix A.2 we provide the full set of derivations of the optimal variational densities together with the analytical form of the lower bound.

2.3.1 Shrinkage priors and optimal variational densities

To begin, we consider a non-informative normal prior for the regression coefficients. In particular, for each entry of $\mathbf{\Theta}$, let $\vartheta_{j,k} \sim \mathsf{N}(0, \upsilon)$, for $j = 1, \ldots, d$ and $k = 1, \ldots, p$. In addition, let $\nu_j \sim \mathsf{Ga}(a_\nu, b_\nu)$ for $j = 1, \ldots, d$, and $\beta_{j,k} \sim \mathsf{N}(0, \tau)$, for $j = 2, \ldots, d$ and $k = 1, \ldots, j-1$. Here, $\mathsf{Ga}(\cdot, \cdot)$ denotes the gamma distribution, and $a_\nu > 0$, $b_\nu > 0$, $\tau \gg 0$ and $\upsilon \gg 0$ are the related hyper-parameters. Let $\boldsymbol{\xi} = (\boldsymbol{\beta}^\intercal, \boldsymbol{\nu}^\intercal, \boldsymbol{\vartheta}^\intercal)^\intercal$ be the collection of the involved parameters, the variational density $q(\boldsymbol{\xi})$ can be factorised as follows:

$$q(\boldsymbol{\xi}) = q(\boldsymbol{\nu})q(\boldsymbol{\beta})q(\boldsymbol{\vartheta}), \quad q(\boldsymbol{\nu}) = \prod_{j=1}^{d} q(\nu_j), \quad q(\boldsymbol{\beta}) = \prod_{j=2}^{d} q(\boldsymbol{\beta}_j), \quad q(\boldsymbol{\vartheta}) = \prod_{j=1}^{d} q(\boldsymbol{\vartheta}_j). \tag{2.4}$$

Propositions 2.1 and 2.2 provide the optimal variational density and variational lower bound (ELBO) for this default normal prior specification, respectively.

Proposition 2.1. *The optimal variational densities under this specification for ν_j and $\boldsymbol{\beta}_j$ are $q^*(\nu_j) \equiv \mathsf{Ga}(a_{q(\nu_j)}, b_{q(\nu_j)})$, where $a_{q(\nu_j)} = a_\nu + T/2$ and $b_{q(\nu_j)} = b_\nu + \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}_q\left[\varepsilon^2_{j,t}\right]$ such that:*

$$\begin{aligned}\mathbb{E}_q\left[\varepsilon^2_{j,t}\right] = &\left(y_{j,t} - \boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})} - \boldsymbol{\mu}_{q(\vartheta_j)}\mathbf{z}_{t-1}\right)^2 + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\vartheta_j)}\mathbf{z}_{t-1}\mathbf{z}^\intercal_{t-1}\right\} \\ &+ \mathsf{tr}\left\{\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)} + \boldsymbol{\mu}^\intercal_{q(\boldsymbol{\beta}_j)}\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}\right)\mathbf{K}_{\vartheta,t}\right\} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}^\intercal_{q(\mathbf{r}_{j,t})}\right\},\end{aligned}$$

where $\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})} = \mathbf{y}^j_t - \boldsymbol{\mu}_{q(\boldsymbol{\Theta}^j)}\mathbf{z}_{t-1}$, and, for $i = 1, \ldots, j-1$ and $k = 1, \ldots, j-1$, the elements in the matrix $\mathbf{K}_{\boldsymbol{\vartheta},t}$ are $\left[\mathbf{K}_{\vartheta,t}\right]_{i,k} = \mathsf{tr}\left\{\mathsf{Cov}(\boldsymbol{\vartheta}_i, \boldsymbol{\vartheta}_k)\mathbf{z}_{t-1}\mathbf{z}^\intercal_{t-1}\right\}$. Moreover, for the rows of $\mathbf{B}$, $q^*(\boldsymbol{\beta}_j) \equiv \mathsf{N}_{j-1}(\boldsymbol{\mu}_{q(\beta_j)}, \boldsymbol{\Sigma}_{q(\beta_j)})$ where:

$$\begin{aligned}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)} &= \left(\mu_{q(\nu_j)}\sum_{t=1}^{T}\left(\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}^\intercal_{q(\mathbf{r}_{j,t})} + \mathbf{K}_{\vartheta,t}\right) + 1/\tau\mathbf{I}_{j-1}\right)^{-1}, \\ \boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)} &= \boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}\mu_{q(\nu_j)}\sum_{t=1}^{T}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}(y_{j,t} - \boldsymbol{\mu}_{q(\vartheta_j)}\mathbf{z}_{t-1})^\intercal.\end{aligned} \tag{2.5}$$

The optimal variational density for the j-th row of the parameter matrix $\boldsymbol{\Theta}$ is a multivariate Gaussian $q^*(\boldsymbol{\vartheta}_j) \equiv \mathsf{N}_p(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$ with optimal parameters:

$$\begin{aligned}\boldsymbol{\Sigma}_{q(\vartheta_j)} &= \left(\boldsymbol{\mu}_{q(\omega_{j,j})}\sum_{t=1}^{T}\mathbf{z}_{t-1}\mathbf{z}^\intercal_{t-1} + 1/\upsilon\mathbf{I}_p\right)^{-1}, \\ \boldsymbol{\mu}_{q(\vartheta_j)} &= \boldsymbol{\Sigma}_{q(\vartheta_j)}\left(\sum_{t=1}^{T}\left(\boldsymbol{\mu}_{q(\omega_j)} \otimes \mathbf{z}_{t-1}\right)\mathbf{y}_t - \left(\boldsymbol{\mu}_{q(\omega_{j,-j})} \otimes \sum_{t=1}^{T}\mathbf{z}_{t-1}\mathbf{z}^\intercal_{t-1}\right)\boldsymbol{\mu}_{q(\vartheta_{-j})}\right),\end{aligned} \tag{2.6}$$

where we denote with $\boldsymbol{\omega}_j$ the j-th row of $\boldsymbol{\Omega}$ and

$$\boldsymbol{\vartheta} = \begin{pmatrix} \boldsymbol{\vartheta}_j \\ \boldsymbol{\vartheta}_{-j} \end{pmatrix}, \qquad \boldsymbol{\Omega} = \begin{pmatrix} \omega_{j,j} & \boldsymbol{\omega}_{j,-j} \\ \boldsymbol{\omega}_{-j,j} & \boldsymbol{\Omega}_{-j,-j} \end{pmatrix}.$$

Proof. See Appendix A.2.1. □

Note from Proposition 2.1 that despite the multivariate model is reduced to a sequence of univariate regressions, the analytical form of the optimal mean $\boldsymbol{\mu}_{q(\vartheta_j)}$ depends on all the other rows through $\boldsymbol{\mu}_{q(\vartheta_{-j})}$. As a result, the posterior estimates of $\boldsymbol{\vartheta}_j$ are explicitly conditional on $\boldsymbol{\vartheta}_{-j}$. This addresses the error in the MCMC algorithm of Carriero *et al.* (2019), which has been discussed by Bognanni (2022) and revised by Carriero *et al.* (2022).

Proposition 2.2. *The variational lower bound for the non-sparse multivariate regression model can be derived analytically and it is equal to:*

$$\begin{aligned}\log \underline{p}(\mathbf{y}; q) = \text{const} &- \sum_{j=1}^{d} \left(a_{q(\nu_j)} \log b_{q(\nu_j)} - \log \Gamma(a_{q(\nu_j)})\right) - \frac{1}{2} \sum_{j=2}^{d} \sum_{k=1}^{j-1} \mu_{q(\beta^2_{j,k})}/\tau \\ &+ \frac{1}{2} \sum_{j=2}^{d} \log |\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}| - \frac{1}{2} \sum_{j=1}^{d} \sum_{k=1}^{p} \upsilon \mu_{q(\vartheta^2_{j,k})} + \frac{1}{2} \sum_{j=1}^{d} \log |\boldsymbol{\Sigma}_{q(\boldsymbol{\vartheta}_j)}|,\end{aligned} \tag{2.7}$$

where const *is a constant term depending on prior parameters.*

Proof. See the proof A.5 in Appendix A.2.1. □

Starting from the quantities derived in Propositions 2.1 and 2.2, we can implement the following iterative Algorithm 2.1 to estimate the optimal variational densities.

Algorithm 2.1: Mean-field variational Bayes for multivariate predictive regressions with non-informative prior.

```
Initialize: $q^*(\boldsymbol{\xi})$, $\Delta_\xi$, $\Delta_{\text{ELBO}}$
while $\left(\widehat{\Delta}_{ELBO} > \Delta_{ELBO}\right) \vee \left(\widehat{\Delta}_\xi > \Delta_\xi\right)$ do
    Update $(a_{q(\nu_1)}, b_{q(\nu_1)})$ and get $q^*(\nu_1)$;
    Update $(\boldsymbol{\mu}_{q(\vartheta_1)}, \boldsymbol{\Sigma}_{q(\vartheta_1)})$ and get $q^*(\boldsymbol{\vartheta}_1)$;
    for $j = 2, \dots, d$ do
        Update $(a_{q(\nu_j)}, b_{q(\nu_j)})$ and get $q^*(\nu_j)$
        Update $(\boldsymbol{\mu}_{q(\beta_j)}, \boldsymbol{\Sigma}_{q(\beta_j)})$ and get $q^*(\boldsymbol{\beta}_j)$;
        Update $(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$ and get $q^*(\boldsymbol{\vartheta}_j)$;
    end
    Compute $\log \underline{p}\,(\mathbf{y}; q)$ as in (2.7);
    Compute $\widehat{\Delta}_{\text{ELBO}} = \log \underline{p}\,(\mathbf{y}; q)^{(\text{iter})} - \log \underline{p}\,(\mathbf{y}; q)^{(\text{iter}-1)}$;
    Compute $\widehat{\Delta}_\xi = q^*(\boldsymbol{\xi})^{(\text{iter})} - q^*(\boldsymbol{\xi})^{(\text{iter}-1)}$ ;
end
```

In Appendix A.2 we present a more general case in which $\boldsymbol{\vartheta}$ is approximated jointly by $q(\boldsymbol{\vartheta})$. The latter is expected to provide better approximation, while increasing the computational cost of the update from $O(dp^2)$ to $O(p^3)$.

Bayesian adaptive lasso

The Bayesian adaptive lasso of Leng *et al.* (2014) extends the original Bayesian lasso of Park and Casella (2008) by imposing a different shrinkage for each parameter. This prior assumes a Laplace distribution with a different scaling parameter $\vartheta_{j,k}|\lambda_{j,k} \sim \mathsf{Lap}(\lambda_{j,k})$. The latter can be represented as a scale mixture of Gaussians with an exponential mixing density, i.e., $\vartheta_{j,k}|\upsilon_{j,k} \sim \mathsf{N}(0,\upsilon_{j,k})$, and $\upsilon_{j,k}|\lambda^2_{j,k} \sim \mathsf{Exp}(\lambda^2_{j,k}/2)$. The choice of the scaling parameters $\lambda^2_{j,k}$ is crucial to recover the underlying signal and it is certainly non trivial in a high-dimensional setting. A common strategy is to infer their values from the data by assuming a common prior distribution $\lambda^2_{j,k} \sim \mathsf{Ga}(h_1,h_2)$, where $h_1, h_2 > 0$ are fixed hyper-parameters. Let $\boldsymbol{\xi}_{\mathrm{L}} = (\boldsymbol{\xi}^\intercal, \boldsymbol{\upsilon}^\intercal, (\boldsymbol{\lambda}^2)^\intercal))^\intercal$ be the vector of parameters $\boldsymbol{\xi}$ augmented with the adaptive lasso prior, then:

$$q(\boldsymbol{\xi}_{\mathrm{L}}) = q(\boldsymbol{\xi})q(\boldsymbol{\upsilon},\boldsymbol{\lambda}^2), \qquad q(\boldsymbol{\upsilon},\boldsymbol{\lambda}^2) = \prod_{j=1}^{d}\prod_{k=1}^{p} q(\upsilon_{j,k})q(\lambda^2_{j,k}), \tag{2.8}$$

Propositions 2.3 and 2.4 provide the optimal variational densities and variational lower bound for the Bayesian adaptive lasso prior.

Proposition 2.3. *The optimal variational densities for ν_j and $\boldsymbol{\beta}_j$ are the same as in Proposition 2.1. The distribution $q^*(\boldsymbol{\vartheta}_j)$ is a multivariate Gaussian as in Proposition 2.1, but with covariance matrix $\boldsymbol{\Sigma}_{q(\vartheta_j)} = (\boldsymbol{\mu}_{q(\omega_{j,j})}\sum_{t=1}^{T}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal + \mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon)}))^{-1}$. For the scaling parameters we have that $q^*(\lambda^2_{j,k}) \equiv \mathsf{Ga}(a_{q(\lambda^2_{j,k})}, b_{q(\lambda^2_{j,k})})$ with $a_{q(\lambda^2_{j,k})} = h_1 + 1$ and $b_{q(\lambda^2_{j,k})} = \mu_{q(\upsilon_{j,k})}/2 + h_2$, while $q^*(1/\upsilon_{j,k}) \equiv \mathsf{IG}(a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})})$ is an inverse-Gaussian distribution where $a_{q(1/\upsilon_{j,k})} = \mu_{q(\vartheta^2_{j,k})}$ and $b_{q(1/\upsilon_{j,k})} = \mu_{q(\lambda^2_{j,k})}$.*

Proof. See Appendix A.2.2. □

Proposition 2.4. *The variational lower bound for the multivariate regression model with adaptive Bayesian lasso prior can be derived analytically and it is equal to:*

$$\begin{aligned}\log \underline{p}(\mathbf{y};q) = \mathrm{const} &- \sum_{j=1}^{d}\left(a_{q(\nu_j)}\log b_{q(\nu_j)} - \log\Gamma(a_{q(\nu_j)})\right) - \frac{1}{2}\sum_{j=2}^{d}\sum_{k=1}^{j-1}\mu_{q(\beta^2_{j,k})}/\tau \\ &+ \frac{1}{2}\sum_{j=2}^{d}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}| + \frac{1}{2}\sum_{j=1}^{d}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\vartheta}_j)}| + \sum_{j=1}^{d}\sum_{k=1}^{p}\frac{1}{2}\mu_{q(\lambda^2_{j,k})}\mu_{q(\upsilon_{j,k})} \\ &- \sum_{j=1}^{d}\sum_{k=1}^{p}(1/4\log(b_{q(1/\upsilon_{j,k})}/a_{q(1/\upsilon_{j,k})}) - \log K_{1/2}(\sqrt{b_{q(1/\upsilon_{j,k})}a_{q(1/\upsilon_{j,k})}})) \\ &- \sum_{j=1}^{d}\sum_{k=1}^{p}\left(a_{q(\lambda^2_{j,k})}\log b_{q(\lambda^2_{j,k})} - \log\Gamma(a_{q(\lambda^2_{j,k})})\right),\end{aligned} \tag{2.9}$$

where const *is a constant term depending on prior parameters.*

Proof. See the proof A.9 in Appendix A.2.2. □

An iterative Algorithm for posterior approximate inference can be set-up following the pseudo-code in Algorithm 2.2.

Algorithm 2.2: Mean-field variational Bayes for multivariate predictive regressions with Bayesian adaptive lasso prior.

```
Initialize: $q^*(\boldsymbol{\xi}_L)$, $\Delta_{\xi_L}$, $\Delta_{\mathrm{ELBO}}$
while $\left(\widehat{\Delta}_{ELBO} > \Delta_{ELBO}\right) \vee \left(\widehat{\Delta}_{\xi_L} > \Delta_{\xi_L}\right)$ do
    Update $(a_{q(\nu_1)}, b_{q(\nu_1)})$ and get $q^*(\nu_1)$;
    Update $(\boldsymbol{\mu}_{q(\vartheta_1)}, \boldsymbol{\Sigma}_{q(\vartheta_1)})$ and get $q^*(\boldsymbol{\vartheta}_1)$;
    for $j = 2, \dots, d$ do
        Update $(a_{q(\nu_j)}, b_{q(\nu_j)})$ and get $q^*(\nu_j)$
        Update $(\boldsymbol{\mu}_{q(\beta_j)}, \boldsymbol{\Sigma}_{q(\beta_j)})$ and get $q^*(\boldsymbol{\beta}_j)$;
        Update $(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$ and get $q^*(\boldsymbol{\vartheta}_j)$;
    end
    for $j = 1, \dots, d$ do
        for $k = 1, \dots, p$ do
            Update $(a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})})$ and get $q^*(\upsilon_{j,k})$;
            Update $(a_{q(\lambda^2_{j,k})}, b_{q(\lambda^2_{j,k})})$ and get $q^*(\lambda^2_{j,k})$;
        end
    end
    Compute $\log \underline{p}\,(\mathbf{y}; q)$ as in (2.9);
    Compute $\widehat{\Delta}_{\mathrm{ELBO}} = \log \underline{p}\,(\mathbf{y}; q)^{(\mathsf{iter})} - \log \underline{p}\,(\mathbf{y}; q)^{(\mathsf{iter}-1)}$;
    Compute $\widehat{\Delta}_{\xi_L} = q^*(\boldsymbol{\xi}_L)^{(\mathsf{iter})} - q^*(\boldsymbol{\xi}_L)^{(\mathsf{iter}-1)}$ ;
end
```

Adaptive normal-gamma

An extension of the Bayesian lasso prior is the normal-gamma prior proposed by Griffin and Brown (2010). Similar to the adaptive lasso we assume different shrinkage parameters in order to make the normal-gamma prior adaptive as well. The hierarchical specification for the elements of $\boldsymbol{\Theta}$ requires that $\vartheta_{j,k}|\upsilon_{j,k} \sim \mathsf{N}(0, \upsilon_{j,k})$, and $\upsilon_{j,k}|\eta_j, \lambda_{j,k} \sim \mathsf{Ga}\left(\eta_j, \eta_j\lambda_{j,k}/2\right)$ for $j = 1, \dots, d$ and $k = 1, \dots, p$. Note that by restricting $\eta_j = 1$ one could obtain the adaptive lasso prior. Marginalization over the variance $\upsilon_{j,k}$ leads to $p(\vartheta_{j,k}|\eta_j, \lambda_{j,k})$ which corresponds to the density of a variance-gamma distribution.

The hyper-parameters η_j and $\lambda_{j,k}$ determine the amount of shrinkage and should be carefully calibrated. Unfortunately, a careful calibration of these parameters is non trivial in high-dimensional parameters spaces. In order to avoid this calibration we impose a further level of hierarchy in the prior structure by assuming $\lambda_{j,k} \sim \mathsf{Ga}(h_1, h_2)$ and $\eta_j \sim \mathsf{Exp}(h_3)$, where (h_1, h_2, h_3) is a fixed vector of prior hyper-parameters with $h_l > 0$ for $l = 1, 2, 3$.

Let $\boldsymbol{\xi}_{\text{NG}} = (\boldsymbol{\xi}^\intercal, \boldsymbol{\upsilon}^\intercal, \boldsymbol{\lambda}^\intercal, \boldsymbol{\eta}^\intercal)^\intercal$ be the vector of parameters $\boldsymbol{\xi}$ augmented with the adaptive normal-gamma prior. The joint distribution $q(\boldsymbol{\xi}_{\text{NG}})$ can be factorised as:

$$q(\boldsymbol{\xi}_{\text{NG}}) = q(\boldsymbol{\xi})q(\boldsymbol{\upsilon}, \boldsymbol{\lambda}, \boldsymbol{\eta}), \qquad q(\boldsymbol{\upsilon}, \boldsymbol{\lambda}, \boldsymbol{\eta}) = \prod_{j=1}^{d} q(\eta_j) \prod_{k=1}^{p} q(\upsilon_{j,k})q(\lambda_{j,k}). \tag{2.10}$$

Proposition 2.5 provides the optimal variational densities for the adaptive normal-gamma prior, and Proposition 2.6 shows the analytical form of the variational lower bound in this case.

Proposition 2.5. *The optimal variational densities for ν_j and $\boldsymbol{\beta}_j$ are the same as in Proposition 2.1. The distribution of $q^*(\boldsymbol{\vartheta}_j)$ is a multivariate Gaussian as in Proposition 2.1 with covariance matrix $\boldsymbol{\Sigma}_{q(\vartheta_j)} = (\boldsymbol{\mu}_{q(\omega_{j,j})} \sum_{t=1}^{T} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal + \mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon)}))^{-1}$. For the scaling parameters we have that $q^*(\lambda_{j,k}) \equiv \mathsf{Ga}(a_{q(\lambda_{j,k})}, b_{q(\lambda_{j,k})})$ with $a_{q(\lambda_{j,k})} = \mu_{q(\eta_j)} + h_1$ and $b_{q(\lambda_{j,k})} = \mu_{q(\eta_j)}\mu_{q(\upsilon_{j,k})}/2 + h_2$, and $q^*(\upsilon_{j,k}) \equiv \mathsf{GIG}(\zeta_{q(\upsilon_{j,k})}, a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})})$ is a generalized inverse-Gaussian distribution with $\zeta_{q(\upsilon_{j,k})} = \mu_{q(\eta_j)} - 1/2$, $a_{q(\upsilon_{j,k})} = \mu_{q(\eta_j)}\mu_{q(\lambda_{j,k})}$ and $b_{q(\upsilon_{j,k})} = \mu_{q(\vartheta^2_{j,k})}$. The optimal density for the parameter η_j is not a known distribution function. An analytical approximation of its moments is calculated via numerical integration as in (A.15) in Appendix A.2.3.*

Proof. See Appendix A.2.3. □

Proposition 2.6. *The variational lower bound for the multivariate regression model with adaptive normal-gamma prior can be derived analytically and it is equal to:*

$$\begin{aligned}
\log \underline{p}(\mathbf{y}; q) = \text{const} &- \sum_{j=1}^{d} \left(a_{q(\nu_j)} \log b_{q(\nu_j)} - \log \Gamma(a_{q(\nu_j)})\right) - \frac{1}{2}\sum_{j=2}^{d}\sum_{k=1}^{j-1} \mu_{q(\beta^2_{j,k})}/\tau \\
&+ \frac{1}{2}\sum_{j=2}^{d} \log |\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}| + \frac{1}{2}\sum_{j=1}^{d} \log |\boldsymbol{\Sigma}_{q(\boldsymbol{\vartheta}_j)}| - \sum_{j=1}^{d}\sum_{k=1}^{p} h(\zeta_{q(\upsilon_{j,k})}, a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})}) \\
&- \sum_{j=1}^{d}\sum_{k=1}^{p} \left(a_{q(\lambda_{j,k})} \log b_{q(\lambda_{j,k})} - \log \Gamma(a_{q(\lambda_{j,k})})\right) + \sum_{j=1}^{d} \log c_{\eta_j} \\
&+ \sum_{j=1}^{d} \mu_{q(\eta_j)} \sum_{k=1}^{p} \left(\mu_{q(\lambda_{j,k})}\mu_{q(\upsilon_{j,k})} - \mu_{q(\log \lambda_{j,k})} - \mu_{q(\log \upsilon_{j,k})}\right),
\end{aligned} \tag{2.11}$$

where c_{η_j} is the normalizing constant of $q^*(\eta_j)$ and const *is a constant term depending on prior parameters.*

Proof. See the proof A.13 in Appendix A.2.3. □

Once the variational updates are available, the implementation of Algorithm 2.3 is straightforward.

Algorithm 2.3: Mean-field variational Bayes for multivariate predictive regressions with adaptive normal-gamma prior.

```
Initialize: $q^*(\boldsymbol{\xi}_{NG})$, $\Delta_{\xi_{NG}}$, $\Delta_{\mathrm{ELBO}}$
while $\left(\widehat{\Delta}_{ELBO} > \Delta_{ELBO}\right) \vee \left(\widehat{\Delta}_{\xi_{NG}} > \Delta_{\xi_{NG}}\right)$ do
    Update $(a_{q(\nu_1)}, b_{q(\nu_1)})$ and get $q^*(\nu_1)$;
    Update $(\boldsymbol{\mu}_{q(\vartheta_1)}, \boldsymbol{\Sigma}_{q(\vartheta_1)})$ and get $q^*(\boldsymbol{\vartheta}_1)$;
    for $j = 2, \ldots, d$ do
        Update $(a_{q(\nu_j)}, b_{q(\nu_j)})$ and get $q^*(\nu_j)$
        Update $(\boldsymbol{\mu}_{q(\beta_j)}, \boldsymbol{\Sigma}_{q(\beta_j)})$ and get $q^*(\boldsymbol{\beta}_j)$;
        Update $(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$ and get $q^*(\boldsymbol{\vartheta}_j)$;
    end
    for $j = 1, \ldots, d$ do
        for $k = 1, \ldots, p$ do
            Update $(a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})})$ and get $q^*(\upsilon_{j,k})$;
            Update $(a_{q(\lambda^2_{j,k})}, b_{q(\lambda^2_{j,k})})$ and get $q^*(\lambda^2_{j,k})$;
        end
        Update $q^*(\eta_j)$ via numerical integration methods;
    end
    Compute $\log \underline{p}\,(\mathbf{y}; q)$ as in (2.11);
    Compute $\widehat{\Delta}_{\mathrm{ELBO}} = \log \underline{p}\,(\mathbf{y}; q)^{(\mathsf{iter})} - \log \underline{p}\,(\mathbf{y}; q)^{(\mathsf{iter}-1)}$;
    Compute $\widehat{\Delta}_{\xi_{NG}} = q^*(\boldsymbol{\xi}_{NG})^{(\mathsf{iter})} - q^*(\boldsymbol{\xi}_{NG})^{(\mathsf{iter}-1)}$ ;
end
```

Horseshoe prior

Finally we consider an horseshoe prior originally as proposed by Carvalho *et al.* (2009, 2010). This is based on the hierarchical specification $\vartheta_{j,k}|\upsilon^2_{j,k}, \gamma^2 \sim \mathsf{N}(0, \gamma^2\upsilon^2_{j,k})$, and $\gamma \sim \mathsf{C}^+(0,1)$ and $\upsilon_{j,k} \sim \mathsf{C}^+(0,1)$, where $\mathsf{C}^+(0,1)$ denotes the standard half-Cauchy distribution with probability density function equal to $f(x) = 2/\{\pi(1+x^2)\}\mathbb{I}_{(0,\infty)}(x)$. The horseshoe is a global-local shrinkage prior (Polson and Scott, 2011; Bhattacharya *et al.*, 2016; Tang *et al.*, 2018) that retrieves aggressive shrinkage of unimportant coefficients without affecting the largest ones. We follow Wand *et al.* (2011) and utilise a scale mixture representation of the half-Cauchy distribution as follows:

$$\begin{aligned} \vartheta_{j,k}|\upsilon^2_{j,k}, \gamma^2 \sim \mathsf{N}(0, \gamma^2\upsilon^2_{j,k}), \quad \gamma^2|\eta &\sim \mathsf{IGa}(1/2, 1/\eta), \quad \upsilon^2_{j,k}|\lambda_{j,k} \sim \mathsf{IGa}(1/2, 1/\lambda_{j,k}), \\ \eta &\sim \mathsf{IGa}(1/2, 1), \qquad \lambda_{j,k} \sim \mathsf{IGa}(1/2, 1), \end{aligned} \tag{2.12}$$

where $\mathsf{IGa}(\cdot,\cdot)$ denotes the inverse-gamma distribution, and the local and global shrinkage are determined by $\upsilon_{j,k}^2$ and γ^2 respectively. Let $\boldsymbol{\xi}_{\text{HS}} = (\boldsymbol{\xi}^\intercal, (\boldsymbol{\upsilon}^2)^\intercal, \gamma^2, \boldsymbol{\lambda}^\intercal, \eta)^\intercal$ be the vector of parameters $\boldsymbol{\xi}$ augmented with the horseshoe prior. The joint distribution $\boldsymbol{\xi}_{\text{HS}}$ can be factorized as:

$$q(\boldsymbol{\xi}_{\text{HS}}) = q(\boldsymbol{\xi})q(\boldsymbol{\upsilon}^2, \gamma^2, \boldsymbol{\lambda}, \eta), \qquad q(\boldsymbol{\upsilon}^2, \gamma^2, \boldsymbol{\lambda}, \eta) = q(\gamma^2)q(\eta)\prod_{j=1}^{d}\prod_{k=1}^{p} q(\upsilon_{j,k}^2)q(\lambda_{j,k}). \tag{2.13}$$

Proposition 2.7 provides the optimal variational densities for the horseshoe prior.

Proposition 2.7. *The optimal variational densities for ν_j and $\boldsymbol{\beta}_j$ are the same as in Proposition 2.1. The distribution family of $q^*(\boldsymbol{\vartheta}_j)$ is a multivariate Gaussian as in Proposition 2.1. Within this setting the optimal variance-covariance matrix is equal to $\boldsymbol{\Sigma}_{q(\vartheta_j)} = (\boldsymbol{\mu}_{q(\omega_{j,j})}\sum_{t=1}^{T}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal + \mu_{q(1/\gamma^2)}\mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon^2)}))^{-1}$. For the global shrinkage parameters we have that $q^*(\gamma^2) \equiv \mathsf{IGa}(a_{q(\gamma^2)}, b_{q(\gamma^2)})$ with $a_{q(\gamma^2)} = (dp+1)/2$ and $b_{q(\gamma^2)} = \mu_{q(1/\eta)} + \frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{p}\mu_{q(1/\upsilon_{j,k}^2)}\mu_{q(\vartheta_{j,k}^2)}$, and $q^*(\eta) \equiv \mathsf{IGa}(1, b_{q(\eta)})$ where the optimal parameter is $b_{q(\eta)} = 1 + \mu_{q(1/\gamma^2)}$. For the local shrinkage parameters we have that $q^*(\upsilon_{j,k}^2) \equiv \mathsf{IGa}(1, b_{q(\upsilon_{j,k}^2)})$ where $b_{q(\upsilon_{j,k}^2)} = \mu_{q(1/\lambda_{j,k})} + \mu_{q(\vartheta_{j,k}^2)}\mu_{q(1/\gamma^2)}/2$, while, for $\lambda_{j,k}$, we have $q^*(\lambda_{j,k}) \equiv \mathsf{IGa}(1, b_{q(\lambda_{j,k})})$ with $b_{q(\lambda_{j,k})} = 1 + \mu_{q(1/\upsilon_{j,k}^2)}$.*

Proof. See Appendix A.2.4. □

In the following Proposition we show the variational lower bound computed under the horseshoe prior.

Proposition 2.8. *The variational lower bound for the multivariate regression model with horseshoe prior can be derived analytically and it is equal to:*

$$\begin{aligned}\log \underline{p}(\mathbf{y}; q) = \text{const} &- \sum_{j=1}^{d}\left(a_{q(\nu_j)}\log b_{q(\nu_j)} - \log\Gamma(a_{q(\nu_j)})\right) - \frac{1}{2}\sum_{j=2}^{d}\sum_{k=1}^{j-1}\mu_{q(\beta_{j,k}^2)}/\tau \\ &+ \frac{1}{2}\sum_{j=2}^{d}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}| + \frac{1}{2}\sum_{j=1}^{d}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\vartheta}_j)}| - a_{q(\gamma^2)}\log b_{q(\gamma^2)} - \log b_{q(\eta)} \\ &+ \mu_{q(1/\gamma^2)}\left(\mu_{q(1/\eta)} + \sum_{j=1}^{d}\sum_{k=1}^{p}\mu_{q(\vartheta_{j,k}^2)}\mu_{q(1/\upsilon_{j,k}^2)}\right) \\ &+ \sum_{j=1}^{d}\sum_{k=1}^{p}\left(\mu_{q(1/\upsilon_{j,k}^2)}\mu_{q(1/\lambda_{j,k})} - \log b_{q(\upsilon_{j,k}^2)} - \log b_{q(\lambda_{j,k})}\right),\end{aligned} \tag{2.14}$$

where const *is a constant term depending on prior parameters.*

Proof. See the proof A.19 in Appendix A.2.4. □

Algorithm 2.4 shows how to perform in practice approximate posterior inference with this prior specification.

Algorithm 2.4: Mean-field variational Bayes for multivariate predictive regressions with horseshoe prior.

```
Initialize: $q^*(\boldsymbol{\xi}_{HS})$, $\Delta_{\xi_{HS}}$, $\Delta_{\text{ELBO}}$
while $\left(\widehat{\Delta}_{ELBO} > \Delta_{ELBO}\right) \vee \left(\widehat{\Delta}_{\xi_{HS}} > \Delta_{\xi_{HS}}\right)$ do
    Update $(a_{q(\nu_1)}, b_{q(\nu_1)})$ and get $q^*(\nu_1)$;
    Update $(\boldsymbol{\mu}_{q(\vartheta_1)}, \boldsymbol{\Sigma}_{q(\vartheta_1)})$ and get $q^*(\boldsymbol{\vartheta}_1)$;
    for $j = 2, \dots, d$ do
        Update $(a_{q(\nu_j)}, b_{q(\nu_j)})$ and get $q^*(\nu_j)$
        Update $(\boldsymbol{\mu}_{q(\beta_j)}, \boldsymbol{\Sigma}_{q(\beta_j)})$ and get $q^*(\boldsymbol{\beta}_j)$;
        Update $(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$ and get $q^*(\boldsymbol{\vartheta}_j)$;
    end
    for $j = 1, \dots, d$ do
        for $k = 1, \dots, p$ do
            Update $(a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})})$ and get $q^*(\upsilon_{j,k})$;
            Update $(a_{q(\lambda^2_{j,k})}, b_{q(\lambda^2_{j,k})})$ and get $q^*(\lambda^2_{j,k})$;
        end
    end
    Update $(a_{q(\gamma^2)}, b_{q(\gamma^2)})$ and get $q^*(\gamma^2)$;
    Update $(a_{q(\eta)}, b_{q(\eta)})$ and get $q^*(\eta)$;
    Compute $\log \underline{p}\,(\mathbf{y}; q)$ as in (2.14);
    Compute $\widehat{\Delta}_{\text{ELBO}} = \log \underline{p}\,(\mathbf{y}; q)^{(\text{iter})} - \log \underline{p}\,(\mathbf{y}; q)^{(\text{iter}-1)}$;
    Compute $\widehat{\Delta}_{\xi_{HS}} = q^*(\boldsymbol{\xi}_{HS})^{(\text{iter})} - q^*(\boldsymbol{\xi}_{HS})^{(\text{iter}-1)}$ ;
end
```

2.3.2 Inference on the precision matrix

Variational inference allows to obtain an approximating density for the regression parameters in $\boldsymbol{\Theta}$, for the Cholesky factor $\mathbf{B}$, and therefore for $\mathbf{L}$, and for the elements on the diagonal of $\mathbf{V}$. However, to obtain a complete inference on the parameters of the original model in (2.1) we still need an approximating density for the precision matrix $\boldsymbol{\Omega} = \mathbf{L}^{\intercal}\mathbf{V}\mathbf{L}$.

Proposition 2.9 shows that, conditional on $\mathbf{L}$ and $\mathbf{V}$, the distribution of $\boldsymbol{\Omega}$ can be approximated by a d-dimensional whishart distribution $\mathsf{Wishart}_d(\delta, \mathbf{H})$, where δ and $\mathbf{H}$ are the degrees of freedom and the scaling matrix, respectively. The proof is based on the Expectation Propagation (EP) variational procedure of Minka (2001), which has the goal of minimizing the $\mathcal{KL}$ divergence between the true and unknown distribution $p(\boldsymbol{\Omega})$ and the approximating density $q(\boldsymbol{\Omega})$. Notice that the order of the density p and the approximating q is reversed in the divergence with respect to variational Bayes approach presented in Chapter 1. This leads to a different paradigm since the $\mathcal{KL}$ divergence is not symmetric, i.e., $\mathcal{KL}(q||p) \neq \mathcal{KL}(p||q)$.

Proposition 2.9. *The optimal approximate distribution q^* of $\mathbf{\Omega}$ is $\mathsf{Wishart}_d(\widehat{\delta}, \widehat{\mathbf{H}})$, where the scaling matrix is given by $\widehat{\mathbf{H}} = \widehat{\delta}^{-1}\mathbb{E}_p\left[\mathbf{\Omega}\right]$ and $\widehat{\delta}$ can be obtained numerically as the solution of a convex optimization problem.*

Proof. The Kullback-Leibler divergence between $p(\mathbf{\Omega})$ and the approximating distribution $q(\mathbf{\Omega})$ is $\mathcal{KL}(p(\mathbf{\Omega})\|q(\mathbf{\Omega})) \propto -\mathbb{E}_p(\log q(\mathbf{\Omega}))$, where the expectation is taken with respect to the distribution $p(\mathbf{\Omega})$. The optimal parameters are $(\widehat{\delta}, \widehat{\mathbf{H}}) = \arg\min_{\delta,\mathbf{H}} \psi(\delta, \mathbf{H})$, where $\psi(\delta, \mathbf{H}) = -\mathbb{E}_p(\log q(\mathbf{\Omega}))$:

$$\psi(\delta, \mathbf{H}) \propto \frac{d\delta}{2}\log 2 + \frac{\delta}{2}\log|\mathbf{H}| + \log\Gamma_d(\delta/2) - \frac{\delta}{2}\mathbb{E}_p\left[\log|\mathbf{\Omega}|\right] + \frac{1}{2}\mathsf{tr}\left\{\mathbf{H}^{-1}\mathbb{E}_p\left[\mathbf{\Omega}\right]\right\}. \quad (2.15)$$

Notice that $\mathbb{E}_p\left[\log|\mathbf{\Omega}|\right] = \mathbb{E}_{q(V)}\left[\log|\mathbf{V}|\right] = \sum_{j=1}^d \mu_{q(\log \nu_j)}$ and $\mathbb{E}_p\left[\mathbf{\Omega}\right] = \mathbb{E}_{q(L),q(V)}\left[\mathbf{L}^\intercal\mathbf{V}\mathbf{L}\right]$ are available as byproduct of the mean-field Variational Bayes algorithm. Differentiating (2.15) with respect to the scaling matrix $\mathbf{H}$, solving $\partial\psi(\delta, \mathbf{H})/\partial\mathbf{H} = 0$ provides $\widehat{\mathbf{H}}_\delta = \delta^{-1}\mathbb{E}_p\left[\mathbf{\Omega}\right]$ that depends on the degrees of freedom δ. Plugging-in the latter in the objective function $\psi(\delta, \widehat{\mathbf{H}}_\delta)$ and proceeding with the minimization of the resulting functional with respect to δ provides $\widehat{\delta}$, which completes the proof. □

In order to assess the goodness of the proposed approximation, we sample from $q(\mathbf{L})$ and $q(\mathbf{V})$ and then obtain values from $q(\mathbf{\Omega})$ exploiting the modified Cholesky decomposition. Table 2.1 compares the sampled distributions with the marginals of the Wishart with parameters $(\widehat{\delta}, \widehat{\mathbf{H}})$ in terms of the approximation accuracy $\mathcal{ACC}(\omega)$ (1.4) presented in Section 2.1, where ω is a generic element of $\mathbf{\Omega}$. The assessment is made for different cross-sectional dimensions $d = 15, 30, 50, 100$ and separately for the diagonal (ω_{jj}) and off-diagonal (ω_{jk}) elements of $\mathbf{\Omega}$.

	$d = 15$		$d = 30$		$d = 50$		$d = 100$	
	ω_{jj}	ω_{jk}	ω_{jj}	ω_{jk}	ω_{jj}	ω_{jk}	ω_{jj}	ω_{jk}
Median	98.41	98.46	98.56	98.35	98.43	98.28	97.42	98.14
Min	97.66	97.13	97.60	96.69	96.76	94.80	94.47	90.66
Max	99.02	99.03	99.34	99.18	99.21	99.24	99.35	99.24

Table 2.1: Accuracy of the Wishart approximation.

The simulation results confirm that our variational Bayes method provides an accurate approximation of the posterior distribution of $\mathbf{\Omega}$, even in a large-dimensional regression setting.

2.3.3 From shrinkage to sparsity

Shrinking rather than selecting is a defining feature of the hierarchical priors outlined in the previous section, in addition to their computational tractability. However, the posterior estimates are non-sparse, meaning one still needs to provide a clear-cut identification of significant predictors. This is key in our simulation and empirical analysis since we ultimately want to assess the accuracy of our variational Bayes approach, versus existing MCMC and `LVB` algorithms.

In this Chapter, we build upon Ray and Bhattacharya (2018) and implement a Signal Adaptive Variable Selector (SAVS) algorithm to induce sparsity in the posterior estimates of the regression matrix $\mathbf{\Theta}$, based on different shrinkage priors. The SAVS is a post-processing algorithm which divides signals and nulls on the basis of the magnitude of the regression coefficients estimates (see, e.g., Hauzenberger, Huber and Onorante, 2021). In particular, consider a regression parameter ϑ_j and the associated vector of covariates $\mathbf{z}_j$, then if $|\widehat{\vartheta}_j|\, ||\mathbf{z}_j||^2 \leq |\widehat{\vartheta}_j|^{-2}$ we set $\widehat{\vartheta}_j = 0$, where $||\cdot||$ denotes the euclidean norm.

The reason why the SAVS may be attractive for large-scale regression models is threefold. First, it is an automatic procedure in which the amount of sparsity imposed uniquely depends on the accuracy of the posterior estimates. Second, the SAVS can be implemented regardless the type of shrinkage prior. Third, it is decision theoretically motivated as it grounds on the idea of minimizing the posterior expected loss (see, e.g., Huber, Koop and Onorante, 2021). Thus, the SAVS represents a convenient alternative compared to post-estimation heuristics based on posterior confidence intervals.

2.3.4 Prediction

Consider the posterior distribution of $p(\boldsymbol{\xi}|\mathbf{z}_{1:t})$ given the information set up to time t, $\mathbf{z}_{1:t} = \{\mathbf{y}_{1:t}, \mathbf{x}_{1:t}\}$, and $p(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\xi})$ the likelihood for the new observation $\mathbf{y}_{t+1}$. The predictive density then takes the familiar form,

$$p(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = \int p(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\xi}) p(\boldsymbol{\xi}|\mathbf{z}_{1:t}) d\boldsymbol{\xi}. \tag{2.16}$$

Given an optimal variational density $q^*(\boldsymbol{\xi})$ that approximates $p(\boldsymbol{\xi}|\mathbf{z}_{1:t})$, we follow Gunawan *et al.* (2021) and obtain the variational predictive distribution

$$q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = \int p(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\xi}) q^*(\boldsymbol{\xi}) d\boldsymbol{\xi} = \int\int p(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\vartheta}, \mathbf{\Omega}) q(\boldsymbol{\vartheta}) q(\mathbf{\Omega}) d\boldsymbol{\vartheta}\, d\mathbf{\Omega}. \tag{2.17}$$

An analytical expression for the above integral is not available. A simulation-based estimator for the variational predictive distribution $q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t})$ can be obtained through Monte Carlo integration by averaging the likelihood of the new observations $p(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\xi}^{(i)})$ over the draws $\boldsymbol{\xi}^{(i)} \sim q^*(\boldsymbol{\xi})$, such that $\widehat{q}(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = N^{-1}\sum_{i=1}^{N} p(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\xi}^{(i)})$. Note that we can further simplify (2.17) by integrating $\boldsymbol{\Omega}$ such that:

$$q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = \int h(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\vartheta})q(\boldsymbol{\vartheta})d\boldsymbol{\vartheta}, \tag{2.18}$$

where $h(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\vartheta})$ denotes the density function of a multivariate Student-t distribution $\mathrm{t}_v(\mathbf{m}, \mathbf{S})$ with mean $\mathbf{m} = \boldsymbol{\Theta}\mathbf{z}_t$, scaling matrix $\mathbf{S} = (v\widehat{\mathbf{H}})^{-1}$, and $v = \widehat{\delta} - d + 1$ degrees of freedom.

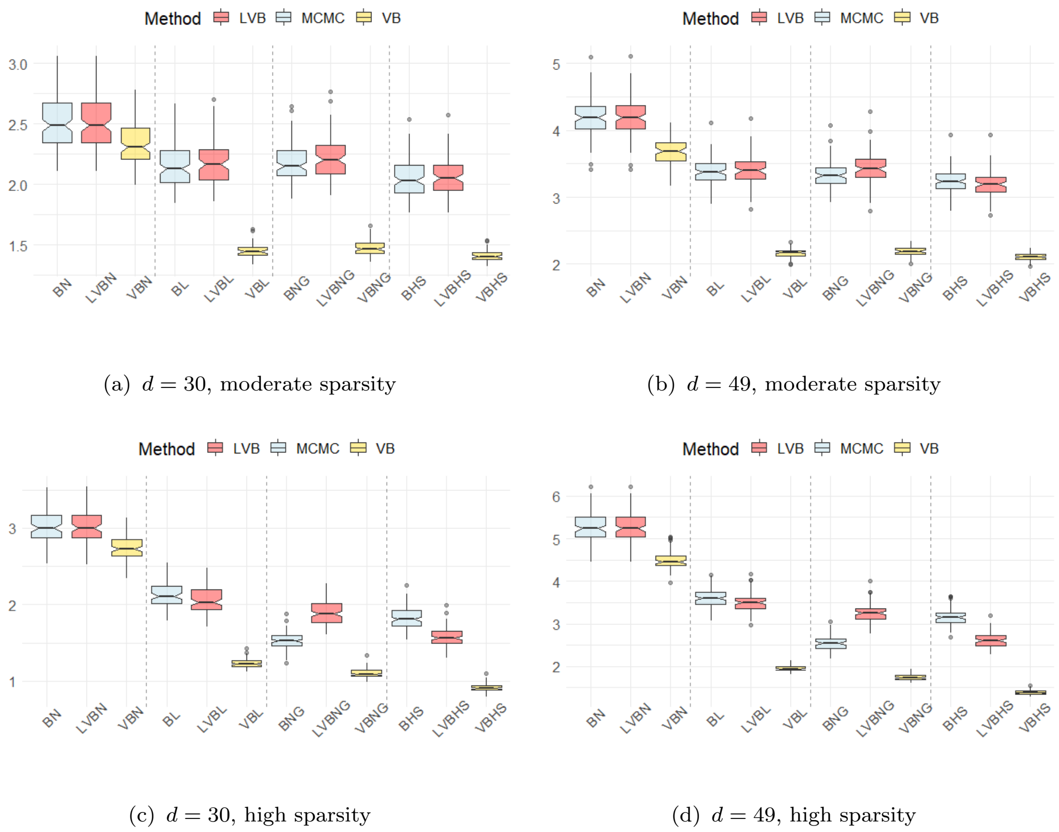
As a result, the predictive distribution can be approximated by averaging the density of the multivariate Student-t $h(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\vartheta}^{(i)})$ over the draws $\boldsymbol{\vartheta}^{(i)} \sim q^*(\boldsymbol{\vartheta})$, for $i = 1, \ldots, N$, such that $\widehat{q}(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = N^{-1}\sum_{i=1}^{N} h(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\vartheta}^{(i)})$. This allows for a more efficient sampling from the predictive density since we only need to sample values of $\boldsymbol{\vartheta}$ from a Gaussian distribution. A further simplification is available up to a second-level variational approximation which minimizes the Kullback-Leibler divergence between the multivariate Student-t and a multivariate Gaussian distribution. The latter is discussed in detail in Appendix A.3.

2.4 Simulation study

We now perform an extensive simulation study to compare the properties of our `VB` approach against standard MCMC and `LVB` approaches. Consistent with the empirical application, we set the length of the time series $T = 360$ and the cross-sectional dimension of the data generating process $d = 30, 49$. Additional results with $d = 15$ are reported in Appendix A.4. We further assume either moderate (50% of true zeros) and high level of sparsity (90% of true zeros).

Without loss of generality, the true matrix $\boldsymbol{\Theta}$ is generated in the following way: we fix to zero sd^2 entries at random, where $s = 0.5, 0.9$, while the remaining non zero coefficients are sampled from a mixture of two Gaussian distributions with means equal to ± 0.08, and standard deviation 0.1. Appendix A.4 provides additional details on the simulation setting.

We compare each estimation method across $N = 100$ replications and for all different prior specifications outlined in Section 2.3.1. Recall that while our approach is based on the parametrization in (2.2a), both the competing MCMC and linearized variational



(a) $d = 30$, moderate sparsity

(b) $d = 49$, moderate sparsity

(c) $d = 30$, high sparsity

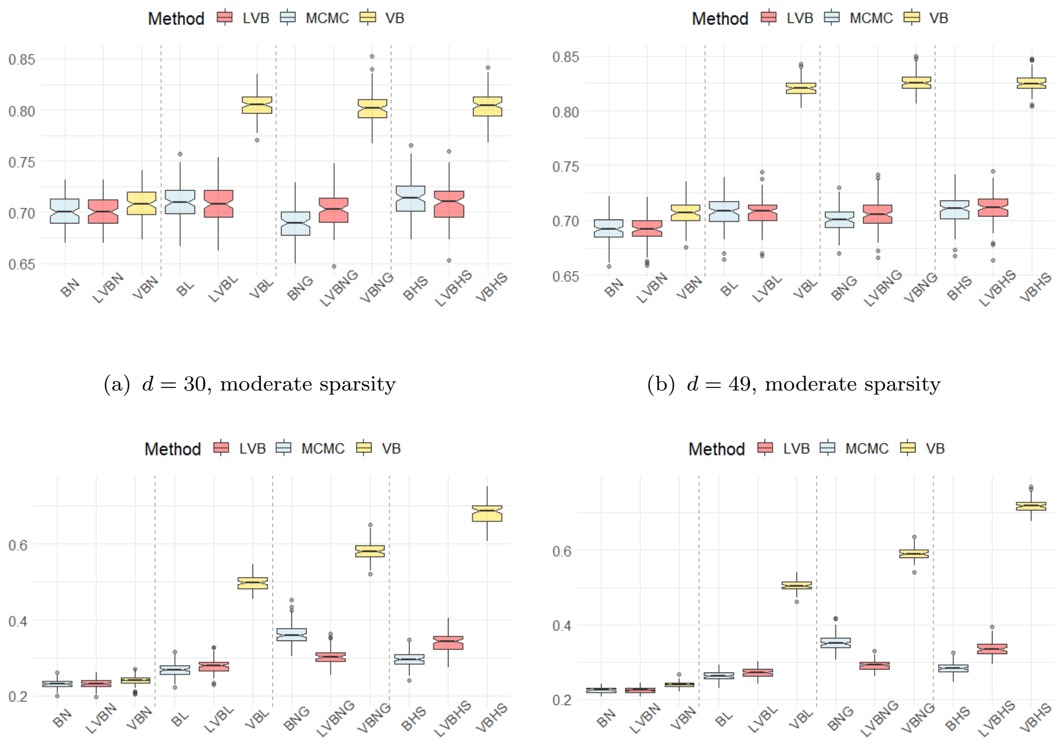
(d) $d = 49$, high sparsity

FIGURE 2.2: Frobenius norm of $\mathbf{\Theta} - \widehat{\mathbf{\Theta}}$ for different hierarchical shrinkage priors and different inference approaches.

Bayes approach are built upon the linearized system of equations implied by (2.2b). As a result, the hierarchical shrinkage priors in our setting can be directly elicited on the matrix $\mathbf{\Theta}$, whereas is imposed on the elements of $\mathbf{A}$ for both the MCMC and `LVB` approach (see, e.g., Gefang *et al.*, 2019; Cross *et al.*, 2020; Chan and Yu, 2022).

As a measure of point estimation accuracy of the regression matrix $\mathbf{\Theta}$, we first look at the Frobenius norm, denoted by $\| \cdot \|_F$. This measure the difference between the true matrix $\mathbf{\Theta}$, which is observed at each simulation, and the corresponding estimate $\widehat{\mathbf{\Theta}}$. Figure 2.2 shows the box charts for the $N = 100$ replications. Depending on the prior specification, we add to the labeling of each estimation method the extension `N` for the normal prior, `L` for the Bayesian adaptive lasso, `NG` for the normal-gamma, and `HS` for the horseshoe. For instance, with `BL`, `LVBL` and `VBL` we indicate the MCMC, the linearized variational Bayes, and our `VB` approach, respectively, under the adaptive lasso prior.

Beginning with the moderate sparsity case (i.e., 50% of zeros in $\mathbf{\Theta}$), the simulated results show that the MCMC and `LVB` approaches tend to perform equally, conditionally on the hierarchical shrinkage prior. This holds for both $d = 30$ and $d = 49$,



(a) $d = 30$, moderate sparsity

(b) $d = 49$, moderate sparsity

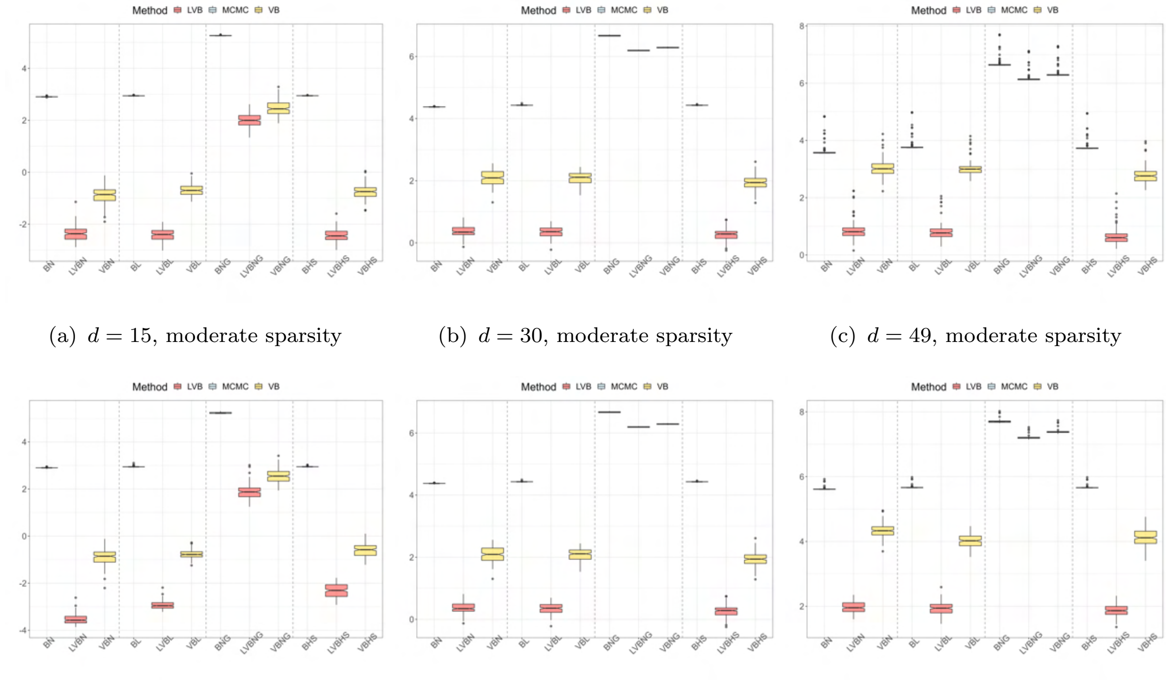
(c) $d = 30$, high sparsity

(d) $d = 49$, high sparsity

FIGURE 2.3: F1 score computed looking at the true non-null parameters in $\boldsymbol{\Theta}$ and the non-null parameters in the estimated matrix $\widehat{\boldsymbol{\Theta}}$, for different hierarchical shrinkage priors and different inference approaches.

and is reassuring, considering both approaches are built upon the same Cholesky-based parametrization. When sparsity is more pervasive (90% of zeros in $\boldsymbol{\Theta}$), there is some discrepancy between the competing approaches; the LVB approach that tends to perform on par with MCMC only under the Bayesian adaptive lasso prior. Perhaps more importantly, the simulation results show that, by eliciting shrinkage priors directly on $\boldsymbol{\Theta}$ rather than on $\mathbf{A}$, the accuracy of the estimates significantly improves. As a matter of fact, the frobenius norm obtained from our VB approach is lower compared to both MCMC and LVB irrespective of the shrinkage prior specification and the level of sparsity in the true regression matrix.

Figure 2.3 reports the F1-score across different estimation methods and for different simulation scenarios. The F1-score quantifies the type I and type II errors in the identification of the significant predictors. For each different prior specification and estimation strategy, the sparsification of the posterior estimates $\widehat{\boldsymbol{\Theta}}$ is implemented by using the SAVS algorithm proposed by Ray and Bhattacharya (2018) (see Section 2.3 for more details).



(a) $d = 15$, moderate sparsity (b) $d = 30$, moderate sparsity (c) $d = 49$, moderate sparsity

(d) $d = 15$, high sparsity (e) $d = 30$, high sparsity (f) $d = 49$, high sparsity

FIGURE 2.4: Boxplots of the computational time required by the algorithms MCMC against the variational methods (VBL and VB) for different hierarchical shrinkage priors. The time is expressed on the logarithmic scale.

Interestingly, under moderate levels of sparsity in the regression coefficients, shrinkage priors provide estimates similar to the non-sparse priors for both the MCMC and the LVB approach. This is likely due to the fact that by recovering $\mathbf{\Theta} = \mathbf{L}^{-1}\mathbf{A}$ from the estimated $\widehat{\mathbf{A}}$ leads to a dense $\widehat{\mathbf{\Theta}}$, and therefore a lower identification accuracy. This is not the case for our variational Bayes approach which directly shrinks the regression matrix $\mathbf{\Theta}$, and therefore result in a much more accurate identification of the significant predictors. This result holds across different hierarchical shrinkage priors and for different dimensions, i.e., for both $d = 30$ and $d = 49$. A set of additional results in Appendix A.4 show that the higher accuracy of our framework is preserved in smaller-dimensional settings (i.e., $d = 15$).

Another advantage of the variational methods is the computational time (see Figure 2.4). For example, in dimension $d = 49$, the VB algorithms with non-sparse, the adaptive lasso and horseshoe priors are 3.31 times faster than the MCMC counterpart. The algorithms with the normal-gamma prior are quite slower than the others and this is due to the fact that we have to sample from a complex generalized inverse Gaussian distribution and moreover we also have a metropolis-Hastings step in the MCMC approach which translates in a numerical integration in the VB algorithm.

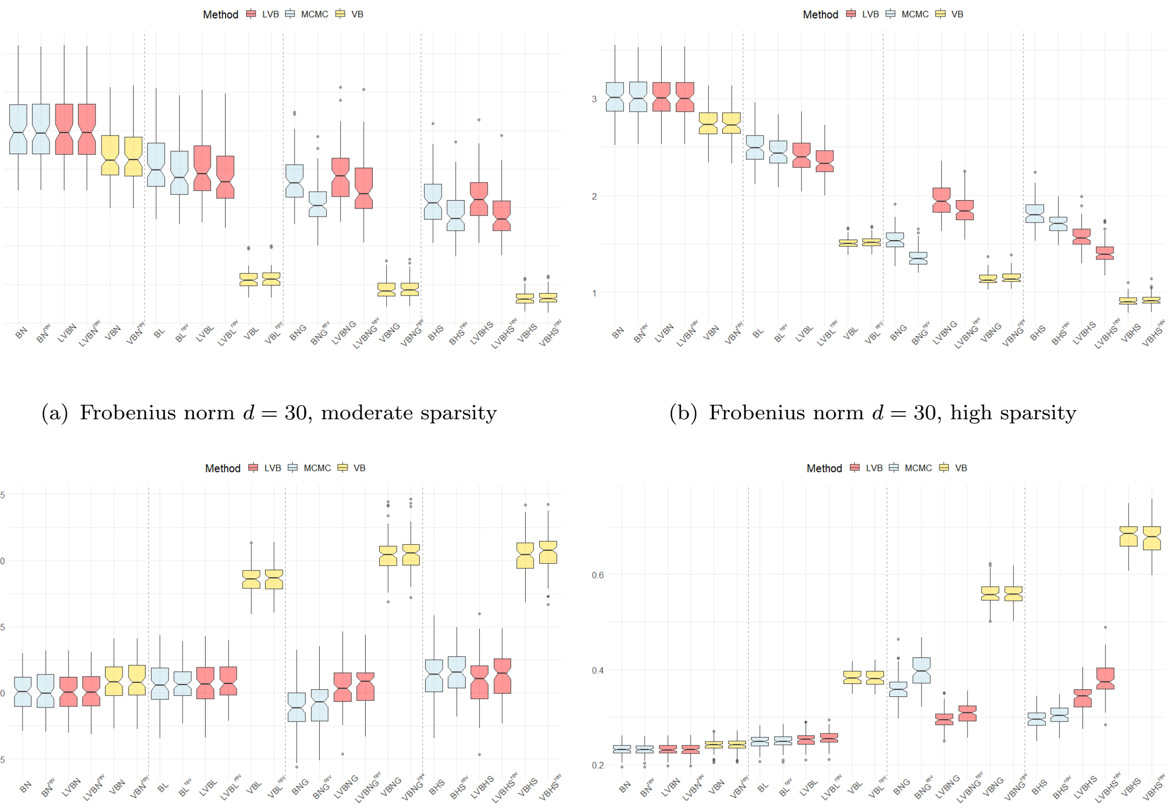
Performance under variables permutation. Based on the same simulation setting described above, we now investigate the performance of all estimation methods under variables permutation. Panel A of Figure 2.5 shows the box charts of the Frobenius norms for the $N = 100$ replications for both moderate and high sparsity in the true $\boldsymbol{\Theta}$. For the ease of exposition we only report the case with $d = 30$ predictors. The case with $d = 49$ allows to draw qualitatively similar conclusions. We put in each figure the simulation results pertaining the original vector $\mathbf{y}_t$ and its reversed order $\mathbf{y}_t^{rev}$ next to each other. Colors and labels are consistent with the initial simulation study.

The accuracy of the estimates from both the MCMC and the linearized variational Bayes approach is affected by the variable order. This is perhaps more evident for the normal-gamma (`BNG`, `LVBNG`) and the Horseshoe (`BHS`, `LVBHS`) priors. The accuracy of the posterior estimates once the ordering of the variables is reversed (labelled with the supscript "rev") is substantially higher, on average for both methods. The impact of the variable order on the posterior estimates from both MCMC and `LVB` is stronger for a highly sparse $\boldsymbol{\Theta}$ (top-right panel). On the other hand, our `VB` method generates consistent posterior estimates across scenarios: its estimation accuracy does not deteriorates or improves depending on an arbitrarily chosen ordering of the target variables.

The bottom panels of Figure 2.5 compares the F1-score under variables permutation across different estimation methods and hierarchical shrinkage priors. Interestingly, when the regression matrix $\boldsymbol{\Theta}$ is moderately sparse, the ordering of the target variables has almost a negligible effect on the ability of MCMC or `LVB` to single out significant predictors. Instead, the effect of ordering on the F1-score increases with the sparsity of the regression coefficient matrix. For instance, when 90% of entries in $\boldsymbol{\Theta}$ are zeros, the identification of significant predictors obtained from MCMC and `LVB` is substantially more accurate under the variables permutation versus the original ordering. This is more visible for the normal-gamma and horseshoe priors. The F1-score confirms the results of the Frobenius norm, meaning that the performance of our `VB` estimation method is permutation-invariant irrespective on how sparse may be the matrix of regression coefficients.

2.5 Industry returns predictability

We now investigate the statistical and economic value of our variational Bayes framework within the context of industry returns predictability in the US. At the end of June of year t each NYSE, AMEX, and NASDAQ stock is assigned to an industry portfolio based on its four-digit SIC code at that time. Thus, the returns on a given value-weighted



(a) Frobenius norm $d = 30$, moderate sparsity

(b) Frobenius norm $d = 30$, high sparsity

(c) F1-score $d = 30$, moderate sparsity

(d) F1-score $d = 30$, high sparsity

FIGURE 2.5: Top panels report the Frobenius norm of $\mathbf{\Theta} - \widehat{\mathbf{\Theta}}$ under variables permutation for different hierarchical shrinkage priors and inference approaches. Bottom panels report the F1-score computed looking at the true non-null parameters in $\mathbf{\Theta}$ and the non-null parameters in the estimated matrix $\widehat{\mathbf{\Theta}}$. The box charts show the results for $N = 100$ replications, $d = 30$ and different levels of sparsity.

portfolio are computed from July of t to June of $t + 1$. We consider two alternative industry aggregations: $d = 30$ industry portfolios from July 1926 to May 2020, and a larger cross section of $d = 49$ industry portfolios from July 1969 to May 2020. The difference of sample length is due to data availability. The sample periods cover major macroeconomic events, from the great depression to the Covid-19 outbreak.

Each stock industry portfolio returns is regressed on lagged cross-industry portfolio returns. In addition, we consider a variety of additional equity risk factors and macroeconomic variables as predictors. For instance, we include in the set of predictors the return on the market portfolio in excess of the risk-free rate (`mkt`), and four alternative long-short investment strategies based on market capitalization (`smb`), book-to-market ratios (`hml`), operating profitability (`rmw`) and investment (`cma`), as proposed by Fama and French (2015). The set of additional macroeconomic predictors is from Goyal and Welch (2008); this includes the log of the aggregate price-dividend ratio (`pd`), the term spread (`term`) (difference between the long term yield on government bonds and the T-bill), the credit spread (`credit`) (the BAA-AAA bond yields difference), the monthly change in inflation (`infl`) measured as the log change in the CPI, the aggregate market book-to-market ratio (`bm`), the net-equity issuing activity (`ntis`) and the long-term corporate bond returns (`corpr`).

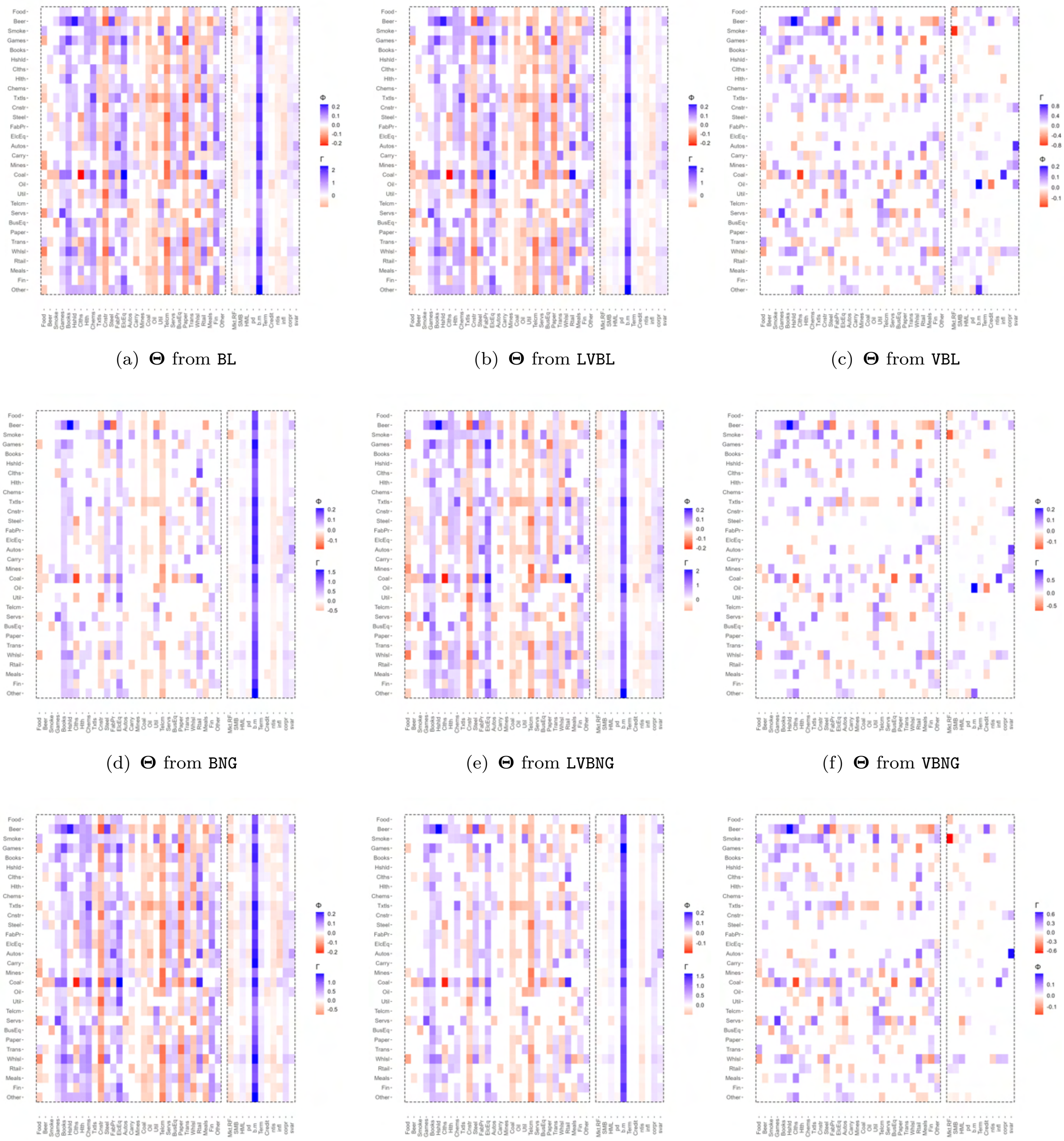
Similar to the simulation study, we add to the labeling of each estimation method the extension `N` for the normal prior, `L` for the Bayesian adaptive lasso, `NG` for the normal-gamma, and `HS` for the horseshoe. For instance, with `BL`, `LVBL` and `VBL` we indicate the MCMC, the `LVB`, and our variational Bayes approach, respectively, under the adaptive lasso prior.

2.5.1 In-sample estimates

Before discussing the out-of-sample forecasting performance, we first report the in-sample posterior estimates of the matrix of regression coefficients $\boldsymbol{\Theta}$. Figure 2.6 shows the estimates. For the sake of brevity we report the results for the $d = 30$ industry case. The posterior estimates highlight three main results. First, the $\widehat{\boldsymbol{\Theta}}$ obtained from the MCMC and the linearised variational Bayes tend to coincide. For instance, the `bm` predictor positive and significant for both methods and across different priors. This is reassuring since, in principle, the `LVB` and the MCMC estimation setting should converge to similar posterior estimates (see, e.g., Gefang *et al.*, 2019; Chan and Yu, 2022).

The second main result from Figure 2.6 is that for both the MCMC and the `LVB` method there are visible differences in the posterior estimates across shrinkage priors. For instance, the $\widehat{\boldsymbol{\Theta}}$ from the `BNG` method is arguably more sparse than the one obtained from the horseshoe prior (`BHS`). Similarly, the regression parameters estimates are more sparse under the `LVBHS` compared to the Bayesian adaptive lasso (`LVBL`). Perhaps more interesting, the third main fact that emerges from Figure 2.6 is that under our variational inference method the estimates of $\boldsymbol{\Theta}$ are (1) more sparse compared to both MCMC and `LVB`, and (2) are remarkably similar across different shinkrage priors.

Section A.5 in the supplementary Appendix shows that the same pattern emerges for the 49 industry portfolios (see Figure A.5). The difference in the posterior estimates for different priors are more marked for the standard MCMC and `LVB` methods, with the normal gamma (horseshoe) producing more sparse estimates within the MCMC (`LVB`) estimation setting. Furthermore, our variational Bayes produces rather stable estimates across priors, yet more sparse compared to both competing estimation methods.



(a) **Θ** from `BL`

(b) **Θ** from `LVBL`

(c) **Θ** from `VBL`

(d) **Θ** from `BNG`

(e) **Θ** from `LVBNG`

(f) **Θ** from `VBNG`

(g) **Θ** from `BHS`

(h) **Θ** from `LVBHS`

(i) **Θ** from `VBHS`

FIGURE 2.6: Posterior estimates of the regression coefficients **Θ** for different estimation methods. We report the estimates for the $d = 30$ industry case obtained from the Bayesian adaptive lasso (top panels), the adaptive normal gamma (middle panels), and the horseshoe (bottom panels).

2.5.2 Out-of-sample forecasting performance

For each industry, we follow Campbell and Thompson (2007); Goyal and Welch (2008) and calculate the out-of-sample predictive R^2 as

$$R^2_{i,oos} = 1 - \frac{\sum_{t_0=2}^{T} \left(y_{it} - \widehat{y}_{it}\left(\mathcal{M}_s\right)\right)^2}{\sum_{t_0=2}^{T} \left(y_{it} - \overline{y}_{it}\right)^2},$$

where t_0 is the date of the first prediction, $\overline{y}_{it}$ is the naive forecast from the recursive mean and $\widehat{y}_{it}\left(\mathcal{M}_s\right)$ is the forecast for a given industry $i = 1, \ldots, d$ from a given shrinkage prior specification $\mathcal{M}_s$. We consider a 360 months rolling window period for the recursive mean and the model estimation, so that for instance for the 30 industry portfolio the out-of-sample forecasting period is from July 1957 to May 2020.

Table 2.2 reports a set of summary statistics for the cross section of industry-specific R^2_{oos}s. In each panel we compare the forecasts obtained from our `VB` estimation versus a standard MCMC and `LVB`, for different shrinkage priors as outlined in Section 2.3. In addition, the first four columns in each panel report the results obtained from univariate models. The latter boil down to assume a diagonal covariance matrix $\mathbf{\Omega}$; that is, the forecasts from the univariate models do not depend on any Cholesky parametrization although ignore potential contemporaneous correlations across industry returns.

Panel A reports the results for the 49 industry classification. For each case we report the mean and median R^2_{oos} across industries as well as a set of percentiles and the min and max values. In addition, we report the percentage of industry portfolios for which a given model can generate positive out-of-sample R^2. Consistent with conventional wisdom, the normal priors tend to overfit in large-dimensional regression models (see, e.g., Korobilis, 2013; Bhattacharya *et al.*, 2015; Hahn and Carvalho, 2015; Griffin and Brown, 2017). This translates in largely negative out-of-sample R^2_{oos}.

Notably, the simple naive forecast based on the rolling mean represents a challenging benchmark to beat for regularised forecasts as well. This is consistent with the existing evidence in returns predictability, such as Campbell and Thompson (2007); Goyal and Welch (2008); Pettenuzzo *et al.* (2014); Bianchi and McAlinn (2020), among others. For instance, none of the univariate models or the multivariate forecasts obtained from MCMC can generate a positive R^2_{oos}. The univariate model with normal gamma prior generates a -0.6% out-of-sample R^2, whereas the `BNG` produces a still negative -1.48% R^2_{oos} under an MCMC estimation procedure.

The performance of the `LVB` method is dismal, with only 4% of the industry portfolio returns that turn out to be predictable under the horseshoe prior. In addition, the magnitude of the predictability is rather low, with the maximum R^2_{oos} equal to 0.99%.

Table 2.2: R^2_{oos} across industries.

This table reports a set of descriptive statistics for the R^2_{oos} expressed in %, across individual industry portfolios. Panel A reports the results for the 49 industry classification, whereas Panel B reports the results for the 30 industry classification. For each case we report the mean and median R^2_{oos} across industries as well as a set of percentiles and the min and max values. In addition, we report the percentage of industry portfolios for which a given model can generate positive out-of-sample R^2, i.e., $\%\left(R^2_{oos} > 0\right)$.

Panel A: R^2_{oos} (%) for 49 industry portfolios

	Univariate				Multivariate											
					MCMC				LVB				VB			
	Normal	BL	NG	HS	Normal	BL	NG	HS	Normal	BL	NG	HS	Normal	BL	NG	HS
Mean	-30.17	-15.13	-5.94	-15.04	-31.59	-17.16	-6.58	-15.30	-31.38	-17.16	-15.80	-6.22	-23.76	-5.09	-4.62	-3.87
Median	-28.23	-14.13	-5.31	-13.97	-28.92	-16.20	-5.96	-14.81	-29.08	-15.41	-14.57	-5.97	-20.84	-5.52	-5.17	-4.08
Percentile																
2.5	-54.67	-26.40	-13.00	-27.89	-60.15	-32.68	-14.32	-29.52	-60.53	-32.56	-29.94	-12.49	-49.95	-11.24	-9.30	-10.22
25	-36.49	-19.70	-7.72	-18.81	-38.58	-21.46	-8.54	-19.33	-38.38	-21.29	-20.06	-8.74	-29.59	-7.98	-7.44	-6.07
75	-22.64	-11.20	-3.70	-11.16	-23.48	-12.75	-4.45	-10.98	-23.42	-13.31	-11.63	-4.08	-17.76	-2.08	-1.62	-1.33
97.5	-13.81	-5.76	-0.89	-5.54	-15.24	-6.92	-1.51	-5.49	-13.14	-5.89	-4.91	0.75	-9.65	1.71	1.49	1.20
Min	-56.74	-31.41	-13.10	-30.55	-60.24	-33.34	-15.81	-32.72	-61.91	-33.86	-31.16	-13.62	-50.91	-12.51	-9.34	-10.31
Max	-12.71	-4.45	-0.60	-4.52	-13.86	-6.60	-1.48	-5.08	-11.96	-5.84	-4.12	0.99	-6.57	2.54	1.68	3.05
$\%\left(R^2_{oos} > 0\right)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.08	0.00	6.12	8.16	8.16

Panel B: R^2_{oos} (%) for 30 industry portfolios

	Univariate				Multivariate											
					MCMC				LVB				VB			
	Normal	BL	NG	HS	Normal	BL	NG	HS	Normal	BL	NG	HS	Normal	BL	NG	HS
Mean	-13.22	-6.38	-1.47	-5.45	-13.84	-6.99	-2.77	-5.29	-36.77	-2.54	-2.53	-3.05	-21.22	-0.59	-0.65	-0.53
Median	-13.71	-6.96	-1.67	-5.91	-14.67	-7.38	-2.97	-5.48	-31.42	-2.80	-2.85	-3.21	-16.46	-0.76	-0.86	-0.77
Percentile																
2.5	-19.87	-10.82	-3.50	-9.63	-20.58	-11.46	-5.78	-9.06	-101.27	-4.18	-4.34	-5.61	-72.42	-2.55	-2.70	-2.73
25	-15.04	-7.84	-2.49	-6.88	-15.58	-8.37	-3.98	-6.56	-36.23	-3.56	-3.43	-4.10	-23.18	-1.44	-1.49	-1.18
75	-11.09	-5.04	-0.47	-4.13	-11.69	-5.77	-1.74	-4.07	-26.65	-1.81	-1.94	-2.06	-12.52	0.48	0.22	0.34
97.5	-6.27	-1.79	1.55	-0.94	-6.66	-2.12	1.07	-0.84	-12.99	0.77	0.85	0.79	-5.24	1.86	1.63	1.53
Min	-20.37	-11.36	-3.57	-10.19	-21.10	-11.78	-5.95	-9.50	-108.95	-4.21	-4.42	-5.77	-79.18	-2.66	-2.82	-2.86
Max	-5.90	-1.56	1.67	-0.65	-6.28	-1.56	1.37	-0.41	-12.35	0.81	0.90	0.97	-4.65	1.92	1.65	1.59
$\%\left(R^2_{oos} > 0\right)$	0.00	0.00	13.33	0.00	0.00	0.00	6.67	0.00	0.00	10.00	10.00	10.00	0.00	33.33	30.00	33.33

On the other hand, the forecasting performance of our VB approach outperform both competing approaches. For instance, more than 8% of the industry portfolios report a $R^2_{oos} > 0$, with a value as high as 3% for monthly returns. The cross section of the R^2_{oos} also provides evidence in favour of our VB method. For instance, the 97.5th percentile of the R^2_{oos} under the LVBL is -6% versus a 1.7% under our VB approach.

Panel B of Table 2.2 reports the performance for the 30 industry classification. The *median* R^2_{oos} is substantially higher across different methods. Athough slightly negative, our VB produces higher R^2_{oos} versus both the MCMC or LVB across all different shrinkage priors. In fact, for about of a third of the 30 industries, the out-of-sample R^2_{oos} from the VB is positive and is as high as 1.9% monthly, which means it outperforms the naive rolling mean forecast.

These results are key since they lie on the fact that unlike the MCMC and LVB approaches, our estimation procedure directly shrinkage the coefficients on the regression matrix $\mathbf{\Theta}$ rather than $\mathbf{A} = \mathbf{L\Theta}$. In addition, we provide solid out-of-sample evidence to the in-sample perspective of Cohen and Frazzini (2008) and Menzly and Ozbas (2010), who find that economic links among certain individual firms and industries could possibly contribute to cross-industry return predictability.

Existing studies, such as Rapach *et al.* (2010), Henkel *et al.* (2011), Dangl and Halling (2012), and Farmer *et al.* (2019) show that the predictability of aggregate stock market returns is primarily concentrated in economic recessions, while it is largely absent during economic expansions. In the next Section we investigate if the performance gap with respect to the naive rolling window forecast decrease during recessions.

Returns predictability during recessions. We now delve further into the analysis of the forecasting performance in recession periods. More precisely, we split the data into recession and expansionary periods using the NBER dates of peaks and troughs. This information is considered *ex-post* and is not used at any time in the estimation of the predictive models. Then, we compute the corresponding R^2_{oos} for the recession periods only. Table 2.3 reports the results for the recession periods using the same structure as in Table 2.2

The predictive ability of all prior specifications, including the normal prior, substantially increases for both the cross section of 49 and 30 industry portfolios. Nevertheless, our VB estimation approach generate the highest *median* R^2_{oos} different shrinkage priors. For instance, the median R^2_{oos} for the VBL is 1.4% against a -17% (19%) obtained from the BL (LVBL) approach. Similarly, the median R^2_{oos} for the VBNG us 1.6% against a still dismal -4% (-16%) obtained from the BNG (LVBNG) method.

Table 2.3: R^2_{oos} across industries during recessions.

This table reports a set of descriptive statistics for the R^2_{oos} calculated during recession periods expressed in %, across individual industry portfolios. Panel A reports the results for the 49 industry classification, whereas Panel B reports the results for the 30 industry classification. For each case we report the mean and median R^2_{oos} across industries as well as a set of percentiles and the min and max values. In addition, we report the percentage of industry portfolios for which a given model can generate positive out-of-sample R^2, i.e., $\%\left(R^2_{oos} > 0\right)$.

Panel A: R^2_{oos} (%) for 49 industry portfolios

	Univariate				Multivariate											
					MCMC				LVB				VB			
	Normal	BL	NG	HS	Normal	BL	NG	HS	Normal	BL	NG	HS	Normal	BL	NG	HS
Mean	-30.20	-19.93	-9.34	-17.67	-31.99	-21.03	-7.40	-16.19	-31.55	-22.19	-19.15	-7.67	-24.07	-0.68	0.05	-0.24
Median	-22.36	-16.31	-5.97	-13.44	-24.02	-17.55	-4.47	-12.96	-24.35	-18.66	-16.80	-6.00	-19.29	1.36	1.60	0.61
Percentile																
2.5	-123.70	-78.74	-38.83	-76.19	-129.91	-81.34	-37.73	-71.35	-142.57	-89.82	-82.66	-40.93	-122.20	-28.24	-29.38	-26.14
25	-35.34	-23.74	-12.55	-21.37	-36.94	-26.26	-10.91	-21.30	-34.56	-26.55	-24.01	-10.86	-28.18	-3.66	-3.35	-2.26
75	-13.44	-8.48	-2.03	-7.18	-14.88	-9.64	-1.58	-6.20	-16.39	-11.73	-7.41	-1.33	-9.03	5.08	6.33	3.93
97.5	6.12	4.34	4.31	5.54	5.64	5.15	6.29	6.62	5.09	2.27	5.17	5.42	13.83	14.23	11.30	9.99
Min	-154.23	-103.91	-55.07	-102.15	-162.39	-102.52	-47.76	-90.60	-159.59	-100.91	-90.23	-47.93	-135.94	-62.53	-48.41	-46.11
Max	10.77	4.69	6.13	6.38	12.23	11.02	6.48	10.59	12.41	2.92	6.68	5.82	15.63	14.73	14.28	11.09
$\%\left(R^2_{oos} > 0\right)$	6.12	6.12	12.24	6.12	6.12	6.12	18.37	8.16	6.12	6.12	8.16	16.33	12.24	55.10	61.22	55.10

Panel B: R^2_{oos} (%) for 30 industry portfolios

	Univariate				Multivariate											
					MCMC				LVB				VB			
	Normal	BL	NG	HS	Normal	BL	NG	HS	Normal	BL	NG	HS	Normal	BL	NG	HS
Mean	-3.68	-0.95	2.89	0.01	-4.13	-1.23	1.22	0.06	-30.92	-0.74	-0.66	-0.40	-15.30	1.31	1.28	1.47
Median	-3.08	-0.93	3.30	0.54	-3.78	-0.76	1.40	0.54	-24.24	-0.56	-0.58	-0.80	-9.28	1.54	1.44	1.18
Percentile																
2.5	-14.90	-10.04	-3.27	-8.65	-15.28	-10.17	-6.21	-8.77	-100.17	-6.08	-5.77	-7.44	-68.71	-3.84	-4.02	-3.29
25	-7.59	-4.12	0.91	-3.05	-7.43	-3.94	-0.72	-2.61	-40.91	-2.84	-2.80	-2.52	-19.44	-1.05	-0.95	-0.76
75	0.95	1.64	4.74	2.77	0.37	1.85	3.17	2.92	-13.12	1.32	1.83	2.06	-4.66	3.40	3.11	3.58
97.5	6.96	6.36	8.61	6.98	6.69	7.49	7.93	7.77	-2.06	4.60	4.56	6.49	5.35	7.22	7.14	6.83
Min	-15.24	-10.57	-3.33	-8.99	-15.45	-10.54	-6.21	-9.08	-101.64	-6.27	-5.93	-7.62	-75.91	-3.85	-4.19	-3.43
Max	7.06	6.55	8.78	7.11	6.83	7.80	8.18	8.08	-1.31	4.89	4.79	6.52	6.43	7.48	7.54	6.83
$\%\left(R^2_{oos} > 0\right)$	30.00	43.33	80.00	53.33	30.00	43.33	63.33	56.67	0.00	40.00	40.00	36.67	6.67	66.67	66.67	73.33

Instead, our VB approach delivers a positive median R^2_{oos} irrespective of the shrinkage prior specification. The performance gap is also clear when we look at the cross section of industries. For instance, more than 50% of the industries have a positive R^2_{oos} when posterior estimates are based on our approach. This is compared to 11% of positive R^2_{oos} across industries, on average across priors, obtained from the MCMC and the LVB method.

The differences in the forecasting performance during recessions narrows when considering the 30 industry classification. For instance, univariate models produce an out-of-sample R^2 which is now comparable to our VB approach. However, this result is limited to the normal gamma shrinkage prior, whereas both the Bayesian adaptive lasso and the horseshoe still under-perform their multivariate counterpart. Interestingly, the MCMC BNG is also quite competitive compared to the VBNG approach.

The results shown in Table 2.3 suggest that the predictive ability of all prior specifications, including the non-sparse normal prior, substantially increases across industries during recessions. Nevertheless, the more accurate estimate of the regression matrix $\mathbf{\Theta}$ obtained through our variational Bayes approach seems to pay off in terms of forecasting accuracy compared to both MCMC and a benchmark linearised variational Bayes method. Perhaps with the only exception of the normal gamma prior for the 30 industry classification, the forecasting performance of our approach is higher for the cross section of industry returns.

2.5.3 Economic significance

It is of paramount importance to evaluate the extent to which apparent gains in predictive accuracy translates into better investment performances. Following existing literature (see, e.g., Goyal and Welch, 2008; Rapach *et al.*, 2010; Dangl and Halling, 2012 and Pettenuzzo *et al.*, 2014), we consider a representative investor with power utility (CRRA) preferences of the form, $\widehat{U}_{t,s} = \widehat{W}_t^{1-\gamma}(\mathcal{M}_s)/(1-\gamma)$, and $\widehat{W}_t(\mathcal{M}_s)$, the wealth generated by the competing model, s, at time, t.

Campbell and Viceira (2004) show that the optimal portfolio allocation based on the conditional forecast can be expressed for the multi-asset case as $\mathrm{w}_t = \gamma^{-1}\mathbf{\Sigma}^{-1}_{t|t-1}[\widehat{\mathbf{y}}_t + \boldsymbol{\sigma}^2_{t|t-1}/2]$, with $\widehat{\mathbf{y}}_t$ the vector of returns forecast at time t, $\boldsymbol{\sigma}^2_{t|t-1}$ a vector containing the diagonal elements of conditional covariance matrix $\mathbf{\Sigma}_{t|t-1}$. Given the optimal weights, we compute the realised returns. Following Fleming *et al.* (2001), we obtain the certainty equivalent differential by subtracting the average utility of each alternative model s, $u_{t,s}$ to the average utility of the historical average forecast, $u_{t,HA}$. A positive value indicates that a representative investor is willing to pay a positive fee to access the investment strategy implied by a given forecasting model.

Notice that this Chapter focuses on the estimation of the regression coefficients under shrinkage priors, and thus modeling time-varying volatility is beyond our scope (see Section 2.2). However, an input to the optimal weights w_t is an estimate of the returns covariance matrix. Consistent with the recursive nature of the forecasts, we consider a simple estimate of $\boldsymbol{\Sigma}_{t|t-1}$ based on the rolling window forecasting errors for each predictive model. We also winsorize the weights for each of the industry to $-1 \leq \mathrm{w}_t \leq 2$ to prevent extreme short-sales and leverage positions. Finally, to make our results directly comparable to other studies we assume a risk aversion $\gamma = 5$ (see, e.g., Johannes *et al.*, 2014).

Table 2.4 shows a set of descriptive statistics summarising the cross section of individual industry certainty equivalent returns expressed in annualised percentage. We report both the individual industry allocation, based on the univariate version of the weights w_t, and the multi-asset case calculated as outline above. Panel A reports the results for the 49 industry classification. For each case we report the mean and median CER across industries as well as a set of percentiles and the min and max values. In addition, we report the percentage of industry portfolios for which a given model can generate positive out-of-sample CERs.

The economic significance confirms the evidence offered by the R^2_{oos}. From a pure economic standpoint, the forecast from a recursive mean are quite challenging to beat, with the mean and median industry CER differentials that are essentially zero. Nevertheless, more than a half of CERs obtained from our variational Bayes approach are positive, compared to a 45%, on average across shrinkage priors, obtained from MCMC and `LVB` methods. Perhaps more importantly, for both the adaptive Bayesian lasso, the adaptive normal-gamma and the horseshoe, our variational Bayes estimation approach produces multi-asset CER which is higher than both the MCMC and the `LVB` approach. Economically, the results show that a representative investor with power utility is willing to pay almost 1% annually to access the strategy based on our variational Bayes estimation.

The results for the cross section of 30 industry portfolios reported in Panel B provide similar evidence. The multi-asset CER obtained from our variational Bayes estimation strategy compares favourably against both MCMC and `LVB` methods, on average across shrinkage priors. In addition, the fraction of positive CERs in the cross section of industry portfolios is higher under our approach, with a certainty equivalent return as high as 0.76% annualised under the forecasts from the horseshoe prior (`VBHS`). This compares

TABLE 2.4: Certainty equivalent returns.

This table reports a set of descriptive statistics for the differential certainty equivalent return (CER) expressed in annualised %. Panel A reports the results for the 49 industry classification, whereas Panel B reports the results for the 30 industry classification. For each case we report the mean and median CER across industries as well as a set of percentiles and the min and max values. In addition, we report the percentage of industry portfolios for which a given model can generate positive out-of-sample CERs.

Panel A: Certainty equivalent for the 49 industry portfolios

	Univariate				Multivariate											
					MCMC				LVB				VB			
	Normal	BL	NG	HS	Normal	BL	NG	HS	Normal	BL	NG	HS	Normal	BL	NG	HS
Mean	-0.02	0.02	0.02	0.01	-0.02	0.00	0.02	0.00	-0.02	-0.01	0.00	-0.01	-0.04	0.00	0.00	-0.02
Median	-0.03	0.04	0.05	0.02	-0.03	-0.01	-0.02	0.01	-0.01	-0.05	-0.05	-0.05	-0.01	0.02	0.05	0.04
Multiasset	0.01	0.15	0.31	0.16	-0.04	0.42	0.67	0.33	0.38	0.39	0.54	0.67	-0.36	0.63	0.26	0.85
Percentile																
2.5	-1.40	-1.20	-0.79	-1.05	-1.63	-1.10	-0.52	-0.94	-1.27	-0.86	-0.87	-0.55	-1.67	-0.83	-0.99	-1.34
25	-0.30	-0.16	-0.07	-0.23	-0.28	-0.24	-0.13	-0.22	-0.39	-0.27	-0.27	-0.16	-0.34	-0.26	-0.20	-0.30
75	0.26	0.38	0.28	0.29	0.27	0.29	0.22	0.22	0.28	0.29	0.25	0.19	0.31	0.26	0.23	0.19
97.5	0.31	0.36	0.44	0.78	0.37	0.77	0.46	0.72	0.56	0.77	0.97	0.55	0.23	0.70	0.81	1.15
Min	-1.61	-1.98	-1.28	-1.39	-1.84	-1.38	-0.66	-1.17	-1.39	-0.91	-1.05	-0.85	-1.74	-1.33	-1.60	-1.46
Max	0.34	0.45	0.46	0.83	1.21	0.90	0.54	0.81	0.76	0.97	1.14	0.67	0.44	0.89	0.95	1.44
% (cer > 0)	33.98	50.14	51.18	52.14	44.90	48.98	41.02	51.10	48.98	48.98	43.06	42.86	46.94	53.06	55.10	55.10

Panel B: Certainty equivalent for the 30 industry portfolios

	Univariate				Multivariate											
					MCMC				LVB				VB			
	Normal	BL	NG	HS	Normal	BL	NG	HS	Normal	BL	NG	HS	Normal	BL	NG	HS
Mean	0.00	0.01	0.00	0.01	0.00	0.01	0.01	0.01	-0.15	-0.01	-0.01	0.00	-0.10	0.00	0.00	0.01
Median	-0.01	-0.01	-0.02	-0.02	-0.02	0.00	0.01	-0.01	-0.12	-0.01	-0.01	-0.02	-0.04	0.01	0.01	0.01
Multiasset	0.13	0.48	0.15	0.44	0.16	0.48	0.32	0.40	-0.20	0.04	0.05	0.08	0.36	0.38	0.38	0.49
Percentile																
2.5	-0.41	-0.25	-0.11	-0.21	-0.38	-0.29	-0.13	-0.21	-0.92	-0.23	-0.24	-0.23	-0.96	-0.25	-0.24	-0.20
25	-0.07	-0.05	-0.06	-0.05	-0.08	-0.08	-0.05	-0.06	-0.30	-0.09	-0.09	-0.08	-0.32	-0.05	-0.06	-0.05
75	0.06	0.05	0.06	0.07	0.05	0.07	0.07	0.06	0.05	0.05	0.07	0.08	0.15	0.06	0.05	0.06
97.5	0.12	0.36	0.17	0.36	0.26	0.34	0.20	0.28	0.26	0.31	0.27	0.21	0.37	0.24	0.23	0.20
Min	-0.48	-0.29	-0.11	-0.22	-0.41	-0.31	-0.14	-0.23	-0.95	-0.24	-0.26	-0.24	-1.09	-0.27	-0.26	-0.22
Max	0.16	0.38	0.19	0.37	0.44	0.35	0.20	0.28	0.34	0.37	0.31	0.59	0.40	0.53	0.44	0.76
% (cer > 0)	40.00	46.67	46.67	40.00	46.67	50.00	50.00	46.67	33.33	50.00	50.00	43.33	46.67	56.67	56.67	56.67

to a 0.59% and a 0.28% obtained from the LVBHS and BHS estimation, respectively.

2.6 Concluding remarks

We are interested in estimating a large-scale multivariate linear regression model and propose a novel variational Bayes estimation algorithm based on a non-linear parametrisation of the regression parameters. This allows a fast and accurate identification of the regression coefficients without leveraging on a standard Cholesky-based transformation of the parameter space. Empirically, we show that our estimation approach substantially outperforms, both statistically and economically, forecasts from state-of-the-art estimation strategies, such as MCMC and linearized variational Bayes methods. This holds across alternative hierarchical shrinkage priors.

Chapter 3

Smooth stochastic volatility

3.1 Introduction

Stochastic volatility (SV, hereafter) models are non-linear hidden Markov models for the dynamic evolution of the conditional variance of an observed random variable (Bernardi *et al.*, 2020). Although the stochastic volatility models were developed in parallel with the ARCH-type models (Engle, 1982; Bollerslev, 1986) they are less popular because of their estimation complexity. Indeed, the latent volatilities enters the observation equation in a non-linear way leading to a likelihood function that depends upon high dimensional integrals. A variety of estimation procedures have been proposed to overcome this difficulty, including the generalized method of moments (GMM) of Melino and Turnbull (1990), the quasi maximum likelihood (QML) approach of Harvey *et al.* (1994) and Ruiz (1994), the efficient method of moments (EMM) of Gallant *et al.* (1997), and MCMC methods Jacquier *et al.* (2002) and Kim *et al.* (1998). In a comparative study of estimation methods, Andersen and Sørensen (1996) showed that MCMC methods are the most efficient estimation techniques for SV models. As concerns the Bayesian paradigm, the analysis of stochastic volatility models has been first proposed by Jacquier *et al.* (2002, 2004). They introduced a single move Gibbs sampler to simulate the log-volatilities one at a time that results in high correlation and bad mixing of the chain. To improve the simulation efficiency, Shephard and Pitt (1997, 2004) propose multi-mover algorithms that leverages the Markovian structure of the model to sample the latent volatilities in blocks. However, their solution to simulate the unknown states, based on a proposal distribution obtained as a second order Taylor expansion of the target, becomes less and less efficient as the dimension of the blocks increases. More recently, and Durbin and Koopman (2000) proposed an importance sampling algorithm that relies on a linear and Gaussian approximation of the non-linear volatility model.

The large amount of data available in financial markets and the need of continuously updating the volatility forecasts as a new observation comes in, has motivated the research on the topic of on-line filtering. In this direction, more recent estimation techniques that joint updates the filtered unknown volatility states and the parameters of the models, as the new information becomes available, have been proposed and named particle filters, (see, e.g., Speekenbrink, 2016, for a systematic review). In their basic and most popular form, particle filter algorithms approximate the sequence of filtering distributions of the volatility process by a set of weighted particles conditional to the parameters that should preventively estimated (see the seminal work of Gordon *et al.* (1993) and the auxiliary particle filter of Pitt and Shephard (1999, 2001); Douc *et al.* (2009)). Various extensions of the basic particle filter to the simulation from the filtering distribution of the states and parameters have been proposed: the works of Liu and West (2001) on the regularized particle filter and Storvik (2002) seems to be the most innovative in this topic. Recently, Andrieu *et al.* (2010); Mendes *et al.* (2020) have introduced a new generation of algorithms that make use of the particle filtering techniques to build a good proposal for the posterior simulation of the volatility states through Markov chain Monte Carlo methods.

Despite recent advances, estimating flexible stochastic volatility models using MCMC methods remains computationally intensive. Variational approximations represent a faster alternative approach to Bayesian inference. Recently Koop and Korobilis (2018) and Gefang *et al.* (2019) adopted variational Bayes methods to estimate stochastic volatility models. Despite the computational advantages, variational methods suffer from a bounded approximation accuracy, as discussed in Chapter 1. The aforementioned works rely on *local* approximations, such as Taylor expansions. These have been shown to perform well in the proximity of the expansion point, but their accuracy decreases rapidly as one moves further away. In contrast, Chan and Yu (2022) propose a *global* approximation of the joint distribution of the (log-)volatility using a multivariate Gaussian approximation. The authors show that their proposal enlarges the family of local approximations and provides a smaller Kullback-Leibler divergence, and therefore higher approximation accuracy.

In this Chapter, we propose a wider family of multivariate Gaussian approximations that generalizes the approach of Chan and Yu (2022) by relaxing some assumptions on the variance-covariance matrix. In particular, the authors does not jointly optimize the mean vector and the variance-covariance matrix of the Gaussian approximation at each variational update, while, for computational simplicity, they fix the latter to match the inverse of the negative Hessian of the log-variational proposal evaluated at its mode.

As pointed out in the reference paper, this might represent a sub-optimal solution. In order to overcome this limitation, we allow the covariance matrix to vary as well, and we tackle the arising computational complexity by leveraging the updating scheme of Rohde and Wand (2016).

Form an empirical perspective, stochastic volatility models have been considered in the financial-econometric literature because of their ability to capture some important stylized facts frequently observed in financial data such as fat-tailed distributions, the slow decay of the autocorrelation function, and the asymmetric response of the volatility to the return shocks (see Ghysels *et al.*, 1996; Shephard and Pitt, 1997, for a review the main properties of the stochastic volatility models). The superior ability of stochastic volatility models to capture stylized facts displayed by the data along with the increasing computer power and the development of efficient estimation techniques gave rise to many financial applications and model extensions of the basic stochastic volatility model. Nowadays, univariate and multivariate stochastic volatility models are effectively applied not only to volatility estimation and forecasting, but also for option pricing, asset allocation and risk management.

A second contribution of this Chapter is motivated by an interesting application is volatility-managed portfolio strategies within the empirical finance literature. Notice that the terms *volatility-targeting* and *volatility-managing* are used interchangeably throughout this Chapter as they carry the same meaning for our purposes. The widespread evidence that volatility tends to cluster over time and negatively correlates with realised returns have motivated the use of volatility targeting to dynamically adjust the notional exposure to a given portfolio. A conventional approach to volatility targeting builds upon the idea that the capital exposure to a given portfolio is levered up (scaled down) based on the inverse of the previous month's realised variance. The theoretical foundation lies in the evolution of the risk-return trade-off over time (see, e.g., Moreira and Muir, 2017). However, volatility management based on realised variance estimates is associated with a dramatic increase in portfolio turnover and significant time-varying leverage. This casts doubt on the usefulness of conventional volatility-managed portfolios, especially for large institutional investors with high *all-in* implementation costs (see, e.g., Patton and Weller, 2020). A simple approach towards cost mitigation is to reduce liquidity demand by slowing down the time-series variation in the factor leverage; this is often achieved by using less erratic estimates of risk, such as the realised volatility instead of the realised variance, or by introducing leverage constraints in the form of a capped notional exposure (see, e.g., Moreira and Muir, 2017; Cederburg *et al.*, 2020; Barroso and Detzel, 2021). While imposing leverage constraints may simplifies an empirical analysis, they do not regularise the often erratic monthly underlying volatility estimates and are typically set arbitrarily, absent sounded economic arguments for their optimal setup. We propose an alternative approach towards slowing down turnover in volatility-managed portfolios which is based on smoothing the predictive density of an otherwise standard stochastic volatility model. In fact, the portfolio turnover directly depends on the measure of volatility estimate used. Our view is that by smoothing monthly volatility forecasts, one can regularise trading turnover and therefore mitigate the effect of transaction costs on volatility-managed portfolios. Such regularisation is achieved in our variational Bayes framework by assuming a general formulation for the mean vector of the Gaussian approximation. This allows for posterior estimates of the latent process to be arbitrary smooth and provides an inference scheme which flexibly encompasses different smoothness assumptions irrespective of the underlying persistence of the latent state.

3.2 Modeling stochastic volatility

Let consider a standard univariate dynamic regression model with stochastic volatility (Taylor, 1994). A general specification is based on a state-space representation of the form:

$$y_t = \mathbf{x}_t^\intercal \beta + \exp(h_t/2)\varepsilon_t, \qquad \varepsilon_t \sim \mathsf{N}(0,1) \tag{3.1}$$

$$h_t = c + \rho(h_{t-1} - c) + u_t, \qquad u_t \sim \mathsf{N}(0,\eta^2), \tag{3.2}$$

where y_t, $\mathbf{x}_t \in \mathbb{R}^p$, $h_t = \log \sigma_t^2$ are, respectively, the observed response variable, a set of covariates, and the log-volatility of the residuals at time t, for $t = 1, 2, \ldots, n$. The error terms ε_t and u_t are mutually independent Gaussian white noise processes. The latent process in (3.2) is a conventional autoregressive process of order one, with unconditional mean c, persistence ρ, and conditional variance η^2. We assume the process to be stationary $|\rho| < 1$, so that the initial state h_0 can be sampled from the marginal distribution, i.e., $h_0 \sim \mathsf{N}\left(c, \frac{\eta^2}{1-\rho^2}\right)$. In what follows we are interested in work with a vector form representation of (3.1)-(3.2). Let $\mathbf{y} = (y_1, \ldots, y_n)^\intercal \in \mathbb{R}^n$ and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{p \times n}$, then the state equation can be written as

$$\mathbf{y} = \mathbf{X}^\intercal \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathsf{N}_n(\mathbf{0}, \mathsf{Diag}\{\exp(\mathbf{h}_1/2)\}) \tag{3.3}$$

where $\mathbf{h}_1 = (h_1, \ldots, h_n)$ is the vector of latent log-volatilities excluding the initial state h_0, and $\mathsf{Diag}(\mathbf{a})$ is the operator that returns a diagonal matrix with diagonal elements equal to the vector $\mathbf{a}$. To achieve a joint representation of the latent process (3.2), we exploit the following remark.

Remark 3.1. The joint distribution of $\mathbf{h} = (h_0, h_1, \dots, h_n)$ can be written using the first-order Markov property $p(\mathbf{h}) = p(h_0)p(h_1|h_0)\dots p(h_n|h_{n-1})$. Then, following Rue and Held (2005), the joint vector $\mathbf{h}$ admits a Gaussian Markov random field (GMRF) representation of order one $\mathbf{h} \sim \mathsf{N}_{n+1}(c\boldsymbol{\iota}_{n+1}, \eta^2\mathbf{Q}^{-1})$ with tridiagonal precision matrix $\mathbf{Q} = \mathbf{Q}(\rho)$ with diagonal elements $q_{1,1} = q_{n+1,n+1} = 1$ and $q_{i,i} = 1 + \rho^2$ for $i = 2, \dots, n$. The off diagonal elements are $q_{i,j} = -\rho$ if $|i - j| = 1$ and 0 elsewhere. Moreover, the inverse of $\mathbf{Q}$, namely $\boldsymbol{\Sigma} = \mathbf{Q}^{-1}$, has elements $\sigma_{i,j} = \rho^{|i-j|}/(1 - \rho^2)$.

To complete the Bayesian model specification we need to define a prior distribution for the parameters that govern the dynamic of the state equation and autoregressive process, i.e., $\boldsymbol{\vartheta} = (\boldsymbol{\beta}^\intercal, c, \rho, \eta^2)^\intercal$. In particular, we assume standard non-informative priors. Let $\boldsymbol{\beta} \sim \mathsf{N}_p(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$ and $c \sim \mathsf{N}(\mu_c, \sigma_c^2)$ be Gaussian distributions, $\eta^2 \sim \mathsf{IGa}(A, B)$ an inverse-gamma, and $\rho \sim \mathsf{U}(-1, 1)$ a continuous uniform between -1 and 1.

3.2.1 Variational inference

In Chapter 1 we discuss different variational Bayes paradigms. In this work, we base the inference on a semi-parametric approach which leverages on both the mean-field factorization assumption and a Gaussian variational approximation. In the following, we consider a factorization of the joint variational density of the latent log-variances $\mathbf{h}$ and the parameters $\boldsymbol{\vartheta} = (\boldsymbol{\beta}, c, \rho, \eta^2)$ of the form:

$$q(\mathbf{h}, \boldsymbol{\vartheta}) = q(\mathbf{h})q(\boldsymbol{\vartheta}) = q(\mathbf{h})q(\boldsymbol{\beta})q(c)q(\rho)q(\eta^2), \tag{3.4}$$

where the variational distribution of the latent states $q(\mathbf{h})$ is a joint distribution over all times, which is introduced in order to preserve the dependence structure. The latter is called global approximation. Recall the closed-form update in (1.7) to compute the optimal variational density for a parameter under the mean-field approach. The following Propositions provide $q^*(\boldsymbol{\beta})$, $q^*(c)$, $q^*(\rho)$, and $q^*(\eta^2)$ and the corresponding proofs are available in Appendix B.1.

Proposition 3.1. *The optimal variational density for the regression parameter vector is* $q(\boldsymbol{\beta}) \equiv \mathsf{N}_p(\boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)})$ *where:*

$$\boldsymbol{\Sigma}_{q(\beta)} = \left(\mathbf{X}\mathbf{H}^{-1}\mathbf{X}^\intercal + \boldsymbol{\Sigma}_\beta^{-1}\right)^{-1} \qquad \boldsymbol{\mu}_{q(\beta)} = \boldsymbol{\Sigma}_{q(\beta)}\left(\mathbf{X}\mathbf{H}^{-1}\mathbf{y} + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta\right), \tag{3.5}$$

and $\mathbf{H}^{-1} = \mathsf{Diag}\left(\mathbb{E}_h\left[\mathrm{e}^{\mathbf{h}_1}\right]\right)$ *is a diagonal matrix with elements that depend on the optimal density for the latent log-volatilities.*

Proof. See proof B.1 in Appendix B.1. □

Proposition 3.2. *The optimal variational density for the unconditional mean of the log-volatility process is* $q(c) \equiv \mathsf{N}(\mu_{q(c)}, \sigma^2_{q(c)})$ *where:*

$$\begin{aligned}\sigma^2_{q(c)} &= (\mu_{q(1/\eta^2)}\boldsymbol{\iota}^\intercal_{n+1}\boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1} + 1/\sigma^2_c)^{-1}\\ \mu_{q(c)} &= \sigma^2_{q(c)}(\mu_{q(1/\eta^2)}\boldsymbol{\iota}^\intercal_{n+1}\boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\mu}_{q(\mathbf{h})} + \mu_c/\sigma^2_c),\end{aligned} \tag{3.6}$$

with

$$\boldsymbol{\mu}_{q(\mathbf{Q})} = \begin{bmatrix} 1 & -\mu_{q(\rho)} & \dots & 0 & 0\\ -\mu_{q(\rho)} & 1+\mu_{q(\rho^2)} & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \dots & 1+\mu_{q(\rho^2)} & -\mu_{q(\rho)}\\ 0 & 0 & \dots & -\mu_{q(\rho)} & 1 \end{bmatrix}.$$

Proof. See proof B.2 in Appendix B.1. □

Proposition 3.3. *The optimal variational density for the autoregressive parameter has the following form:*

$$\log q(\rho) \propto \frac{1}{2}\log(1-\rho^2) - \frac{1}{2}\mu_{q(1/\eta^2)}\left(\rho^2\sum_{t=1}^{n-1}a_t - 2\rho\sum_{t=0}^{n-1}b_t\right), \quad \rho \in (-1,1) \tag{3.7}$$

with

$$\begin{aligned}a_t &= \mathbb{E}_q\left[(h_t-c)^2\right] = (\mu_{q(h_t)} - \mu_{q(c)})^2 + \sigma^2_{q(h_t)} + \sigma^2_{q(c)}\\ b_t &= \mathbb{E}_q\left[(h_t-c)(h_{t+1}-c)\right] = (\mu_{q(h_t)} - \mu_{q(c)})(\mu_{q(h_{t+1})} - \mu_{q(c)}) + \sigma_{q(h_t,h_{t+1})} + \sigma^2_{q(c)},\end{aligned}$$

where $\sigma_{q(h_t,h_{t+1})}$ *denotes the covariance between* h_t *and* h_{t+1} *under the approximating density* q*. Notice that* $\log q(\rho)$ *can be written as:*

$$\log q(\rho) \propto \frac{1}{2}\log(1-\rho^2) - \frac{1}{2}\mu_{q(1/\eta^2)}\left(\sum_{t=1}^{n-1}a_t\right)\left(\rho - \frac{\sum_{t=0}^{n-1}b_t}{\sum_{t=1}^{n-1}a_t}\right)^2, \quad \rho \in (-1,1) \tag{3.8}$$

thus the normalizing constant and the first two moments can be found by Monte Carlo methods by sampling from an univariate Gaussian distribution with mean $\frac{\sum_{t=0}^{n-1}b_t}{\sum_{t=1}^{n-1}a_t}$ *and precision* $\mu_{q(1/\eta^2)}\left(\sum_{t=1}^{n-1}a_t\right)$.

Proof. See proof B.3 in Appendix B.1. □

Proposition 3.4. *The optimal variational density for the variance parameter is an inverse-gamma distribution* $q(\eta^2) \equiv \mathsf{IGa}(A_{q(\eta^2)}, B_{q(\eta^2)})$, *where:*

$$\begin{aligned} A_{q(\eta^2)} &= A + \frac{n+1}{2} \\ B_{q(\eta^2)} &= B + \frac{1}{2}(\boldsymbol{\mu}_{q(\mathbf{h})} - \mu_{q(c)}\boldsymbol{\iota}_{n+1})^\intercal \boldsymbol{\mu}_{q(\mathbf{Q})}(\boldsymbol{\mu}_{q(\mathbf{h})} - \mu_{q(c)}\boldsymbol{\iota}_{n+1}) \\ &\quad + \frac{1}{2}\left(\mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\mathbf{h})}\boldsymbol{\mu}_{q(\mathbf{Q})}\right\} + \sigma^2_{q(c)}\boldsymbol{\iota}_{n+1}^\intercal \boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1}\right), \end{aligned} \tag{3.9}$$

and recall that $\mu_{q(1/\eta^2)} = A_{q(\eta^2)}/B_{q(\eta^2)}$.

Proof. See proof B.4 in Appendix B.1. □

Once the optimal variational densities for the parameters in $\boldsymbol{\vartheta}$ are analytically derived, we focus on the approximating density for the latent process $\mathbf{h}$, where the novelty of our estimation procedure can be well understood when compared to the existing literature (see, e.g., Chan and Yu, 2022). We exploit the representation of $\mathbf{h}$ as GMRF to obtain the approximating density $q(\mathbf{h})$ as $\mathbf{h} \sim \mathsf{N}_{n+1}(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Omega}^{-1}_{q(h)})$ with mean vector $\boldsymbol{\mu}_{q(h)} = \mathbf{W}\mathbf{f}_{q(h)}$ and variance-covariance matrix $\boldsymbol{\Sigma}_{q(h)} = \boldsymbol{\Omega}^{-1}_{q(h)}$, where $\boldsymbol{\Omega}_{q(h)}$ is a tridiagonal matrix. First of all, notice that the choice of $\boldsymbol{\mu}_{q(h)}$ as a linear projection $\mathbf{W}\mathbf{f}_{q(h)}$, with $\mathbf{f}_{q(h)} \in \mathbb{R}^k$ the projection coefficients and $\mathbf{W}$ an $(n+1) \times k$ deterministic matrix, has a direct effect on the posterior estimates of log-volatility. In Section 3.2.2 we discuss in details how different structures of $\mathbf{W}$ leads to different posterior estimates irrespective of the underlying dynamics of the latent state. This is a key feature of our estimation strategy, since it allows to customise the volatility estimates and forecasts without changing the underlying model assumptions. As first approximation proposal, we assume an homoschedastic GMRF specification defined as $\mathbf{h} \sim \mathsf{N}_{n+1}(\mathbf{W}\mathbf{f}, \tau^2\boldsymbol{\Gamma}^{-1})$, where $\boldsymbol{\mu}_{q(h)} = \mathbf{W}\mathbf{f}$ is the mean vector and $\boldsymbol{\Sigma}_{q(h)} = \tau^2\boldsymbol{\Gamma}^{-1}$ is the variance-covariance matrix. More precisely, $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}(\gamma)$ is a tridiagonal precision matrix with diagonal elements $\boldsymbol{\Gamma}_{1,1} = \boldsymbol{\Gamma}_{n+1,n+1} = 1$ and $\boldsymbol{\Gamma}_{i,i} = 1 + \gamma^2$ for $i = 2, \ldots, n$, and off-diagonal elements $\boldsymbol{\Gamma}_{i,j} = -\gamma$ if $|i - j| = 1$ and 0 elsewhere. Under this setting, the density function of the approximate distribution is given by:

$$\log \phi(\mathbf{h}|\mathbf{W}\mathbf{f}, \tau^2\boldsymbol{\Gamma}^{-1}) \propto -\frac{n+1}{2}\log(\tau^2) - \frac{n}{2}\log(1-\gamma^2) - \frac{1}{2\tau^2}(\mathbf{h} - \mathbf{W}\mathbf{f})^\intercal\boldsymbol{\Gamma}(\mathbf{h} - \mathbf{W}\mathbf{f}). \tag{3.10}$$

Let $\boldsymbol{\xi} = (\mathbf{f}, \tau^2, \gamma)$ be the collection of the variational parameters. In order to find the optimal values $\hat{\boldsymbol{\xi}}$, we have to maximize the variational lower bound (ELBO) $\widehat{\boldsymbol{\xi}} =$

$\arg\max_\xi \psi(\mathbf{f}, \tau^2, \gamma)$, where the ELBO is given by:

$$\begin{aligned}\psi(\mathbf{f}, \tau^2, \gamma) &= \mathbb{E}_q(\log p(\mathbf{h}, \mathbf{y})) - \mathbb{E}_q(\log q(\mathbf{h})) \\ &\propto -\frac{1}{2}\boldsymbol{\iota}_n^\intercal \mathbf{W}_1\mathbf{f} - \frac{1}{2}\boldsymbol{\mu}_{q(\mathbf{s})}^\intercal \mathrm{e}^{-\mathbf{W}_1\mathbf{f}+\frac{1}{2}\tau^2\boldsymbol{\iota}_n} \\ &\quad - \frac{1}{2}\mu_{q(1/\eta^2)}(\mathbf{Wf} - \mu_{q(c)}\boldsymbol{\iota}_{n+1})^\intercal \boldsymbol{\mu}_{q(\mathbf{Q})}(\mathbf{Wf} - \mu_{q(c)}\boldsymbol{\iota}_{n+1}) \\ &\quad - \frac{1}{2}\mu_{q(1/\eta^2)}\tau^2 \mathsf{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}_{q(\mathbf{Q})}) \\ &\quad + \frac{n+1}{2}\log(\tau^2) + \frac{n}{2}\log(1-\gamma^2),\end{aligned} \tag{3.11}$$

with $\boldsymbol{\mu}_{q(\mathbf{s})} = (\mu_{q(s_1)}, \ldots, \mu_{q(s_n)})^\intercal$, $\mu_{q(s_t)} = (y_t - \mathbf{x}_t^\intercal \boldsymbol{\mu}_{q(\beta)})^2 + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\beta)}\mathbf{x}_t\mathbf{x}_t^\intercal\right\}$, and $\mathbf{W}_1 \in \mathbb{R}^{n\times k}$ denotes the matrix obtained by deleting the first row of $\mathbf{W}$. The objective function $\psi(\mathbf{f}, \tau^2, \gamma)$ has gradient ∇_ξ and Hessian $\mathcal{H}_\xi$ available in closed form (see Appendix B.2 for a detailed computation) and an iterative Newton-type algorithm can be implemented to fin the optimal variational parameters $(\hat{\mathbf{f}}, \hat{\tau}^2, \hat{\gamma})$. Although the proposed approximation shows a satisfying accuracy within a simulated scenario, it does not represent the optimal approximation choice because of the homoschedastic assumption. Therefore, we relax the latter assumption and also provide a more general approximation which allows for an heteroschedastic Gaussian Markov random field, i.e., $\boldsymbol{\Omega}_{q(h)}$ is still a tridiagonal matrix, but with generic elements $[\boldsymbol{\Omega}_{q(h)}]_{i,j}$. Similarly as before, the optimal parameters $\boldsymbol{\xi} = (\mathbf{f}_{q(h)}, \boldsymbol{\Sigma}_{q(h)})$ of the approximating density $q(\mathbf{h})$ can be found by maximizing the variational lower bound. To solve the optimization, we leverage on the results in Rohde and Wand (2016). The authors provide a closed-form updating scheme for the variational parameters when the approximating density is a multivariate Gaussian. Proposition 3.5 shows the details on the optimal updating scheme for the variational density of the latent volatility states.

Proposition 3.5. *Let* $\mathbf{h} \sim \mathsf{N}_{n+1}(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Omega}_{q(h)}^{-1})$ *be the proposed approximation, with mean vector* $\boldsymbol{\mu}_{q(h)} = \mathbf{Wf}_{q(h)}$ *and variance-covariance matrix* $\boldsymbol{\Sigma}_{q(h)} = \boldsymbol{\Omega}_{q(h)}^{-1}$*, an iterative algorithm can be set as:*

$$\boldsymbol{\Sigma}_{q(h)}^{new} = \left[\nabla^2_{\boldsymbol{\mu}_{q(h)}\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}^{old}, \boldsymbol{\Sigma}_{q(h)}^{old})\right]^{-1}, \tag{3.12}$$

$$\mathbf{f}_{q(h)}^{new} = \mathbf{f}_{q(h)}^{old} + \mathbf{W}^{+}\,\boldsymbol{\Sigma}_{q(h)}^{new}\nabla_{\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}^{old}, \boldsymbol{\Sigma}_{q(h)}^{old}), \tag{3.13}$$

$$\boldsymbol{\mu}_{q(h)}^{new} = \mathbf{Wf}_{q(h)}^{new}, \tag{3.14}$$

where $\mathbf{W}^{+} = (\mathbf{W}^{\intercal}\mathbf{W})^{-1}\mathbf{W}^{\intercal}$ *is the left Moore–Penrose pseudo-inverse of* $\mathbf{W}$. *The function* S *is* $S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}) = \mathbb{E}_q(\log p(\mathbf{h}, \mathbf{y}))$*:*

$$\begin{aligned} S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}) = &-\frac{1}{2}[0, \boldsymbol{\iota}_n^{\intercal}]\boldsymbol{\mu}_{q(h)} - \frac{1}{2}[0, \boldsymbol{\mu}_{q(\mathbf{s})}^{\intercal}]\mathrm{e}^{-\boldsymbol{\mu}_{q(h)}+\frac{1}{2}\boldsymbol{\sigma}^2_{q(\mathbf{h})}} - \frac{1}{2}\mu_{q(1/\eta^2)}\mathsf{tr}(\boldsymbol{\Sigma}_{q(\mathbf{h})}\boldsymbol{\mu}_{q(\mathbf{Q})}) \\ &- \frac{1}{2}\mu_{q(1/\eta^2)}(\boldsymbol{\mu}_{q(h)} - \mu_{q(c)}\boldsymbol{\iota}_{n+1})^{\intercal}\boldsymbol{\mu}_{q(\mathbf{Q})}(\boldsymbol{\mu}_{q(h)} - \mu_{q(c)}\boldsymbol{\iota}_{n+1}), \end{aligned} \tag{3.15}$$

such that its first and second derivative are equal to:

$$\begin{aligned} \nabla_{\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}) = &-\frac{1}{2}[0, \boldsymbol{\iota}_n^{\intercal}]^{\intercal} + \frac{1}{2}[0, \boldsymbol{\mu}_{q(\mathbf{s})}^{\intercal}]^{\intercal} \odot \mathrm{e}^{-\boldsymbol{\mu}_{q(h)}+\frac{1}{2}\mathsf{diag}(\boldsymbol{\Sigma}_{q(\mathbf{h})})} \\ &- \mu_{q(1/\eta^2)}\boldsymbol{\mu}_{q(\mathbf{Q})}(\boldsymbol{\mu}_{q(h)} - \mu_{q(c)}\boldsymbol{\iota}_{n+1}), \end{aligned} \tag{3.16}$$

$$\nabla^2_{\boldsymbol{\mu}_{q(h)}\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}) = -\frac{1}{2}\mathsf{Diag}\left[[0, \boldsymbol{\mu}_{q(\mathbf{s})}^{\intercal}]^{\intercal} \odot \mathrm{e}^{-\boldsymbol{\mu}_{q(h)}+\frac{1}{2}\mathsf{diag}(\boldsymbol{\Sigma}_{q(\mathbf{h})})}\right] - \mu_{q(1/\eta^2)}\boldsymbol{\mu}_{q(\mathbf{Q})}, \tag{3.17}$$

where $\odot$ *denotes the Hadamard product and* $\mathsf{diag}(\mathbf{A})$ *is the operator that returns the diagonal elements in the matrix* $\mathbf{A}$.

Proof. See Appendix B.3. □

The Gaussian variational approximation for $\mathbf{h}$ directly allows for a known distribution function for the variances $\boldsymbol{\sigma}^2 = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_n^2)$.

Remark 3.2. Under the multivariate Gaussian approximation of $q(\mathbf{h})$ with mean vector $\boldsymbol{\mu}_{q(h)}$ and covariance matrix $\boldsymbol{\Sigma}_{q(h)}$, the optimal density of the vector $\boldsymbol{\sigma}^2 = \exp\{\mathbf{h}\}$, namely $q^*(\boldsymbol{\sigma}^2)$, is a multivariate log-normal distribution such that:

$$\begin{aligned} \mathbb{E}_q[\sigma_t^2] &= \exp\{\mu_{q(h_t)} + 1/2\sigma^2_{q(h_t)}\}, \\ \mathsf{Var}_q[\sigma_t^2] &= \exp\{2\mu_{q(h_t)} + \sigma^2_{q(h_t)}\}(\exp\{\sigma^2_{q(h_t)}\} - 1), \\ \mathsf{Cov}_q[\sigma_t^2, \sigma_{t+1}^2] &= \exp\{\mu_{q(h_t)} + \mu_{q(h_{t+1})} + 1/2(\sigma^2_{q(h_t)} + \sigma^2_{q(h_{t+1})})\}(\exp\{\mathsf{Cov}_q[h_t, h_{t+1}]\} - 1). \end{aligned} \tag{3.18}$$

Our approach expands the global approximation method proposed by Chan and Yu (2022) along three main dimensions. First, we relax the assumption that the initial distribution $q(h_0)$ is independent on the trajectory of the latent state $q(\mathbf{h}_1)$, that is, we do not assume $q(\mathbf{h}) = q(h_0)q(\mathbf{h}_1)$. Second, we do not make any assumption on the $\boldsymbol{\Sigma}_{q(h)}$, which is not fixed conditional on $\boldsymbol{\mu}_{q(h)}$, but is estimated jointly with $\boldsymbol{\mu}_{q(h)}$. As highlighted by Chan and Yu (2022), by enlarging the class of distributions the optimal density is expected to be better approximation to $q(\mathbf{h})$. Third, our latent volatility state accommodates a more general AR(1) dynamics, instead of a random walk. While

Algorithm 3.1: Variational Bayes for arbitrary smoothness in stochastic volatility.

```
Initialize: $q^*(\boldsymbol{\vartheta}, \mathbf{h})$, $\mathbf{W}$, $\Delta$
while $\big(\widehat{\Delta} > \Delta\big)$ do
    Update $q^*(\mathbf{h})$ as in Proposition 3.5;
    Update $q^*(c) = \mathsf{N}(\mu_{q(c)}, \sigma^2_{q(c)})$ as in (3.6) with
        $\sigma^2_{q(c)} = (\mu_{q(1/\eta^2)} \boldsymbol{\iota}_{n+1}^\intercal \boldsymbol{\mu}_{q(\mathbf{Q})} \boldsymbol{\iota}_{n+1} + 1/\sigma_c^2)^{-1}$,
        $\mu_{q(c)} = \sigma^2_{q(c)} (\mu_{q(1/\eta^2)} \boldsymbol{\iota}_{n+1}^\intercal \boldsymbol{\mu}_{q(\mathbf{Q})} \boldsymbol{\mu}_{q(\mathbf{h})} + \mu_c/\sigma_c^2)$.
    Update $q(\eta^2) = \mathsf{IGa}(A_{q(\eta^2)}, B_{q(\eta^2)})$ as in (3.9) with
        $A_{q(\eta^2)} = A + \frac{n+1}{2}$,    $B_{q(\eta^2)} = B + \frac{1}{2}(\boldsymbol{\mu}_{q(\mathbf{h})} - \mu_{q(c)}\boldsymbol{\iota}_{n+1})^\intercal \boldsymbol{\mu}_{q(\mathbf{Q})} (\boldsymbol{\mu}_{q(\mathbf{h})} - \mu_{q(c)}\boldsymbol{\iota}_{n+1})$
            $+\frac{1}{2}\left(\mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\mathbf{h})}\boldsymbol{\mu}_{q(\mathbf{Q})}\right\} + \sigma^2_{q(c)}\boldsymbol{\iota}_{n+1}^\intercal \boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1}\right)$.
    Update $q^*(\rho)$ as in (3.8);
    Update $q(\boldsymbol{\beta}) = \mathsf{N}_p(\boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)})$ as in (3.5) with
        $\boldsymbol{\Sigma}_{q(\beta)} = \left(\mathbf{X}^\intercal\mathbf{H}^{-1}\mathbf{X} + \boldsymbol{\Sigma}_\beta^{-1}\right)^{-1}$,    $\boldsymbol{\mu}_{q(\beta)} = \boldsymbol{\Sigma}_{q(\beta)}\left(\mathbf{X}^\intercal\mathbf{H}^{-1}\mathbf{y} + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta\right)$.
    Compute $\widehat{\Delta} = q^*(\boldsymbol{\vartheta}, \mathbf{h})^{(\mathsf{iter})} - q^*(\boldsymbol{\vartheta}, \mathbf{h})^{(\mathsf{iter}-1)}$ ;
end
```

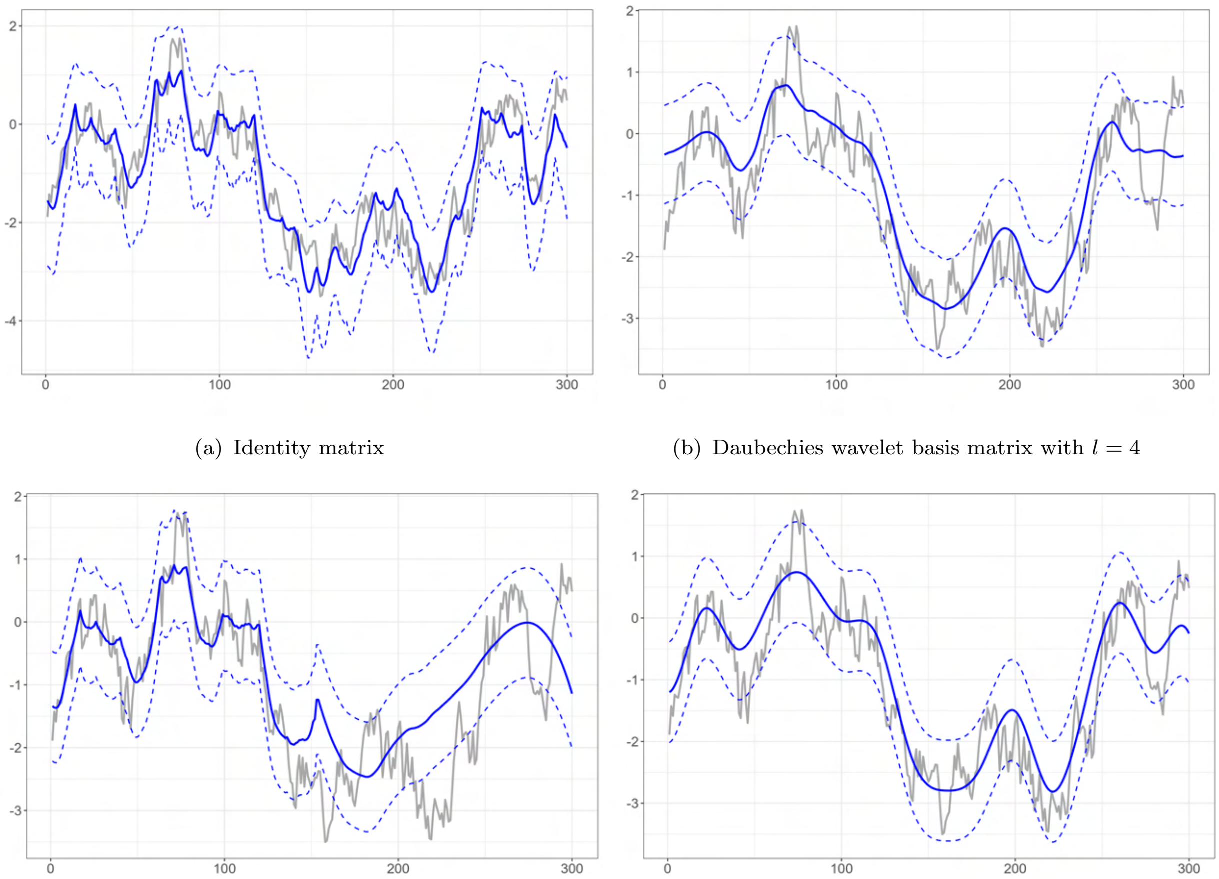
the latter reduces the parameter space, it imposes a strong form of non-stationarity in the log-volatility process. In Section 3.3, we show via a simulation study that all these features have a significant effect on the accuracy of the variational Bayes approximation.

A pseudo-code for the implementation of the proposed iterative estimation procedure is available in Algorithm 3.1, where the convergence is achieved when the variation in the optimal densities update $q^*(\boldsymbol{\vartheta}, \mathbf{h})^{(\mathsf{iter})} - q^*(\boldsymbol{\vartheta}, \mathbf{h})^{(\mathsf{iter}-1)}$ is smaller than a threshold Δ.

3.2.2 Smoothing the volatility estimates

The choice of $\boldsymbol{\mu}_{q(h)}$ as a linear projection $\mathbf{W}\mathbf{f}_{q(h)}$, with $\mathbf{f}_{q(h)} \in \mathbb{R}^k$ being the projection coefficients and $\mathbf{W}$ an $(n+1)\times k$ deterministic matrix, has a direct effect on the posterior estimates of log-volatility. Figure 3.1 shows examples of the shape of $\boldsymbol{\mu}_{q(\mathbf{h})} = \mathbf{W}\mathbf{f}_{q(h)}$ for different choices of $\mathbf{W}$ (solid line), and the corresponding 95% HPD intervals implied by $\boldsymbol{\Sigma}_{q(h)}$ (dashed line). The gray trajectory represents the true simulated value of the log-stochastic volatility. The top-left panel reports the posterior estimates obtained by setting $\mathbf{W} = \mathbf{I}_{n+1}$, with $\mathbf{I}_{n+1}$ an identity matrix of dimension $n+1$. This represents a non-smooth estimate which is akin to the output of a standard MCMC estimation scheme (see, e.g., Hosszejni and Kastner, 2021).

The remaining panels of Figure 3.1 highlight a key feature of our estimation strategy; that is, it allows to customize the volatility forecasts without changing the underlying



(a) Identity matrix

(b) Daubechies wavelet basis matrix with $l = 4$

(c) Identity + Daubechies wavelet basis matrix

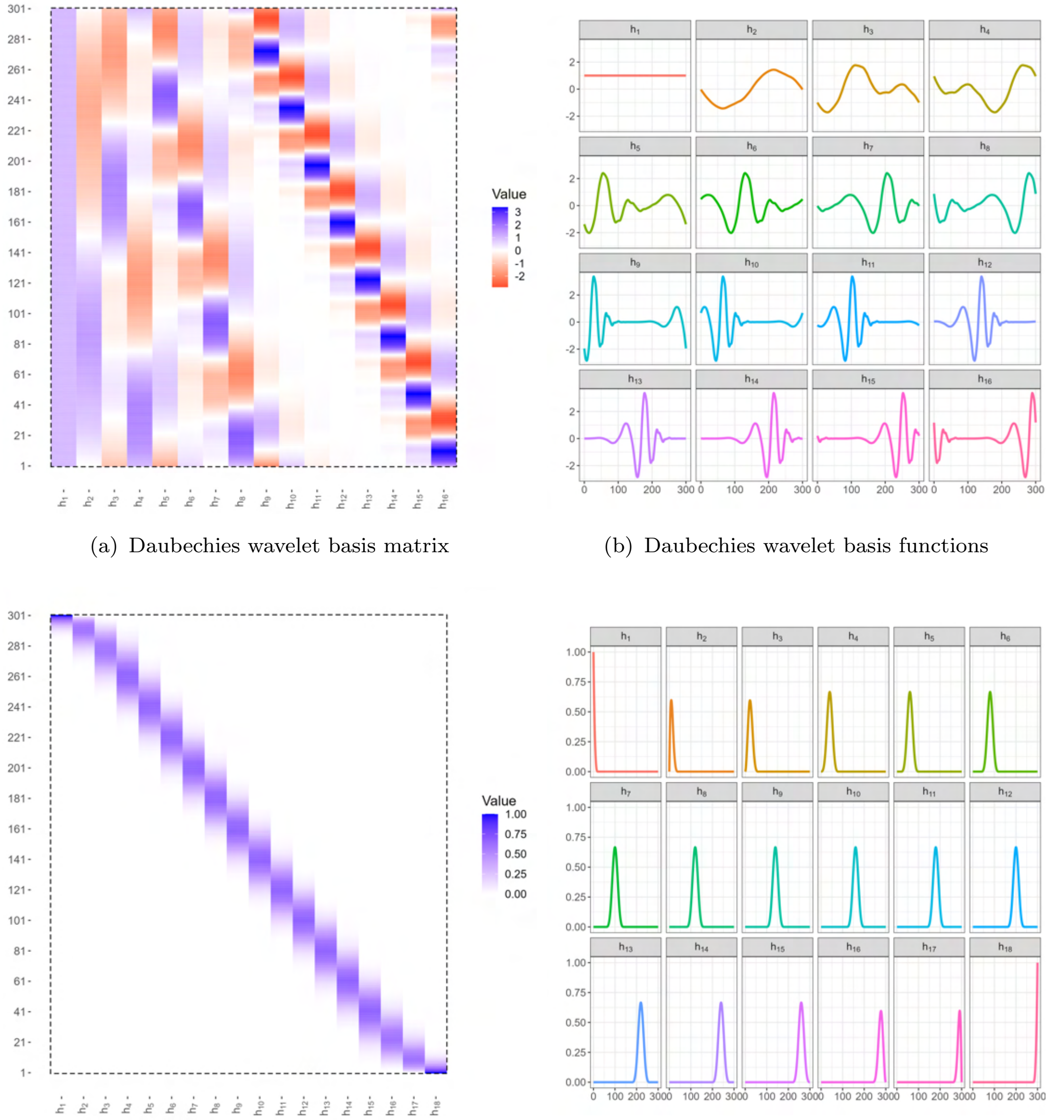
(d) B-spline basis matrix with $kn = 20$ and $dg = 3$

FIGURE 3.1: Shape of the posterior log-volatility estimates for different $\mathbf{W}$.

model assumptions. For instance, the top-right panel shows the posterior estimates of the latent volatility state with $\mathbf{W}$ a matrix of Daubechies wavelet basis functions with a fixed degree of smoothness $l = 4$.

The bottom panels in Figure 3.1 highlight the flexibility of our approach; the left panel shows that more than one smoothing assumption can coexists in the same optimal variational density. For instance, the shape of the posterior estimates assuming $\mathbf{W} = \mathbf{I}_{n+1}$ for the first half of the sample and $\mathbf{W}$ a wavelet basis function with $l = 4$ for the second half of the sample. The bottom-right panel shows that a variety of smoothing functions can be adopted; for instance, the estimates of the latent stochastic volatility can be smoothed based on $\mathbf{W}$ equal to a B-spline basis matrix representing the family of piecewise polynomials with the pre-specified interior knots (kn), degree (dg), and boundary knots.

Figure 3.2 depicts the form of $\mathbf{W}$ when B-spline and Daubechies wavelets are used. The form of $\mathbf{W}$ in case of B-spline basis functions (top) and wavelet basis functions (bottom). Right panels correspond to columns of the matrix $\mathbf{W}$. The B-spline basis



(a) Daubechies wavelet basis matrix

(b) Daubechies wavelet basis functions

(c) B-spline basis matrix with $kn = 20$ and $dg = 3$

(d) B-spline basis functions

FIGURE 3.2: The structure of $\mathbf{W}$ in modeling smoothing volatility.

functions is a sequence of piecewise polynomial functions of a given degree, in this case $dg = 3$. The locations of the pieces are determined by the knots, here we assume $kn = 20$ equally spaced knots. The functions that compose the Daubechies wavelet basis matrix $\mathbf{W}$ are constructed over equally spaced grids on $[0, n]$ of length R, where R is called resolution and it is equal to 2^{l-1}, where l defines the level, and as a result the degree of smoothness. The number of functions at level l is then equal to R and they are defined as dilatation and/or shift of a more general *mother* function. In our case the level is $l = 5$ and therefore the resolution is $R = 16$.

3.2.3 Prediction

Consider the posterior distribution of $p(\mathbf{h},\boldsymbol{\vartheta}|\mathbf{y})$ given the information set available at time t, $\mathbf{y}=\{y_{1:t}\}$, and $p(h_{n+1}|\mathbf{y},\mathbf{h},\boldsymbol{\vartheta})$ the likelihood for the new latent state h_{n+1}. The predictive density then takes the familiar form:

$$p(h_{n+1}|\mathbf{y}) = \int p(h_{n+1}|\mathbf{y},\mathbf{h},\boldsymbol{\vartheta})p(\mathbf{h},\boldsymbol{\vartheta}|\mathbf{y})\, d\mathbf{h}d\boldsymbol{\vartheta}. \tag{3.19}$$

Given a variational density $q(\mathbf{h},\boldsymbol{\vartheta})=q(\mathbf{h})q(\boldsymbol{\vartheta})$ that approximates $p(\mathbf{h},\boldsymbol{\vartheta}|\mathbf{y})$, we follow Gunawan *et al.* (2021) and obtain the variational predictive distribution:

$$\begin{aligned} q(h_{n+1}|\mathbf{y}) &= \int p(h_{n+1}|\mathbf{y},\mathbf{h},\boldsymbol{\vartheta})q(\mathbf{h})q(\boldsymbol{\vartheta})\, d\mathbf{h}d\boldsymbol{\vartheta} \\ &= \int p(h_{n+1}|h_n,\boldsymbol{\vartheta})q(h_n)q(\boldsymbol{\vartheta})\, dh_n d\boldsymbol{\vartheta}, \end{aligned} \tag{3.20}$$

where the second equality follows from Markov property. Recall that our object of interest is the forecast of the variance σ_t^2, rather than the log-volatility h_t for $t=n+1$. Since $h_{n+1}=\log\sigma_{n+1}^2$, the variational predictive density of the conditional variance is readily available as

$$q\left(\sigma_{n+1}^2|\mathbf{y}\right) = \frac{\partial h_{n+1}}{\partial\sigma_{n+1}^2} q\left(h_{n+1}|\mathbf{y}\right)\Big|_{h_{n+1}=\log\sigma_{n+1}^2} = \frac{1}{\sigma_{n+1}^2} q\left(h_{n+1}|\mathbf{y}\right)\Big|_{h_{n+1}=\log\sigma_{n+1}^2}. \tag{3.21}$$

The integral in (3.20) cannot be solved analytically. However, it can be approximated through Monte Carlo integration exploiting the fact that the optimal variational densities $q^*(h_n)$ and $q^*(\boldsymbol{\vartheta})$ are known and we can efficiently sample from them. A simulation-based approximated estimator for the variational predictive distribution of the conditional variance $q(\sigma_{n+1}^2|\mathbf{y})$ is therefore obtained by averaging the density $p(h_{n+1}|h_n^{(i)},\boldsymbol{\vartheta}^{(i)})$ over the draws $h_n^{(i)}\sim q^*(h_n)$ and $\boldsymbol{\vartheta}^{(i)}\sim q^*(\boldsymbol{\vartheta})$, for $i=1,\dots,N$, from the optimal variational density, such that $\widehat{q}(\sigma_{n+1}^2|\mathbf{y})=(N\sigma_{n+1}^2)^{-1}\sum_{i=1}^N p(h_{n+1}|h_n^{(i)},\boldsymbol{\vartheta}^{(i)})$.

3.3 Simulation study and inference properties

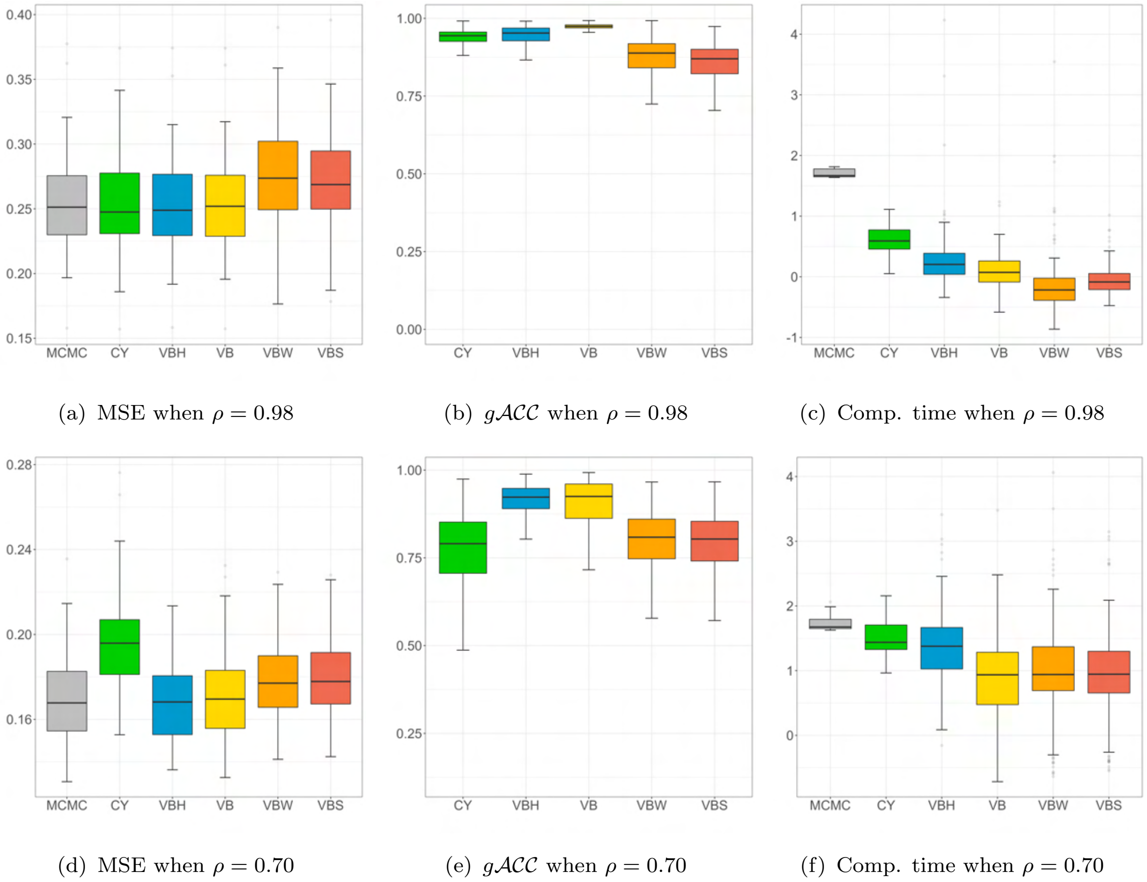
We now perform an extensive simulation study to evaluate the properties of our estimation framework in a controlled setting. We compare our variational Bayes (VB) method against two state-of-the-art Bayesian approaches used within the context of stochastic volatility models, such as MCMC (here we use the R-package stochvol of Hosszejni and Kastner, 2021) and the global variational approximation recently introduced by Chan and Yu (2022) (henceforth CY). Since neither of the benchmark approaches entertain the possibility of arbitrarily smooth predictive densities, the baseline comparison is based on the assumption that $\mathbf{W} = \mathbf{I}_{n+1}$ and the underlying latent state follows an autoregressive dynamics. This gives a cleaner comparison of the accuracy of our variational estimates both in absolute terms and with respect to `MCMC` methods.

We compare each estimation method across $N = 100$ replications and for all different specifications. We consider $n = 600$, consistent with the shortest time series in the empirical application, $c = 0$, $\eta^2 = 0.1$ and both low and high persistence $\rho \in \{0.70, 0.98\}$. Recall that our estimation framework is agnostic on the structure of covariance of the approximating density $\mathbf{\Sigma}_{q(h)}$. However, to better understand the contribution of such generalisation compared to existing methods, we also consider the performance of the homoschedastic approximation (henceforth `VBH`).

Figure 3.3 reports the estimation mean squared error, together with a measure of global accuracy compared to the `MCMC` and computational running time. The mean squared error is measured as $MSE = n^{-1}\sum_{t=1}^{n}(h_t - \hat{h}_t)^2$, where h_t and $\hat{h}$ are the simulated log-variance and its estimate, respectively. The average aggregated accuracy ($g\mathcal{ACC}$) of variatonal Bayes with respect to the `MCMC` approach is calculated aggregating over replicates and time the accuracy measure in (1.4). For the higher-persistence scenario with $\rho = 0.98$ (top panels), the `MCMC`, `CY`, `VB`, and `VBH` provide statistically equivalent performances. The best approximation to the `MCMC` is provided by our `VB` for $\rho = 0.98$.

Interestingly, for the lower-persistent scenario with $\rho = 0.70$ (bottom panels), the `CY` approach shows some difficulty in capturing the full extent of the dynamics of the latent stochastic volatility process. This is also reflected in a generally lower accuracy in approximating the true posterior density $p(\mathbf{h}|\mathbf{y})$ compared to the `MCMC` approach. The lower accuracy of the `CY` approach for $\rho = 0.7$ is due to a more restrictive dynamics of the latent states imposed by their estimation setting. The approximation proposed by Chan and Yu (2022) is based on the computationally convenient assumption that the latent volatility state is a random walk. As a result, it shows a substantially lower accuracy when $\rho \ll 1$.

Although neither the `CY` nor the `MCMC` approach entertain the possibility of smooth volatility forecasts, for a full comparison of the estimation accuracy of our `VB` method we also evaluate the performance of two alternative smoothing approaches, with $\mathbf{W}$ either a B-spline basis matrix with knots equally spaced every 10 time points (henceforth `VBS`), or a Daubechies wavelet basis matrix with $l = 5$ (henceforth `VBW`). The choice of the equally spaced knots in the basis function and the l for the wavelet basis matrix is such that both approaches give a similar degree of smoothness. Notice that both these



(a) MSE when $\rho = 0.98$

(b) $g\mathcal{ACC}$ when $\rho = 0.98$

(c) Comp. time when $\rho = 0.98$

(d) MSE when $\rho = 0.70$

(e) $g\mathcal{ACC}$ when $\rho = 0.70$

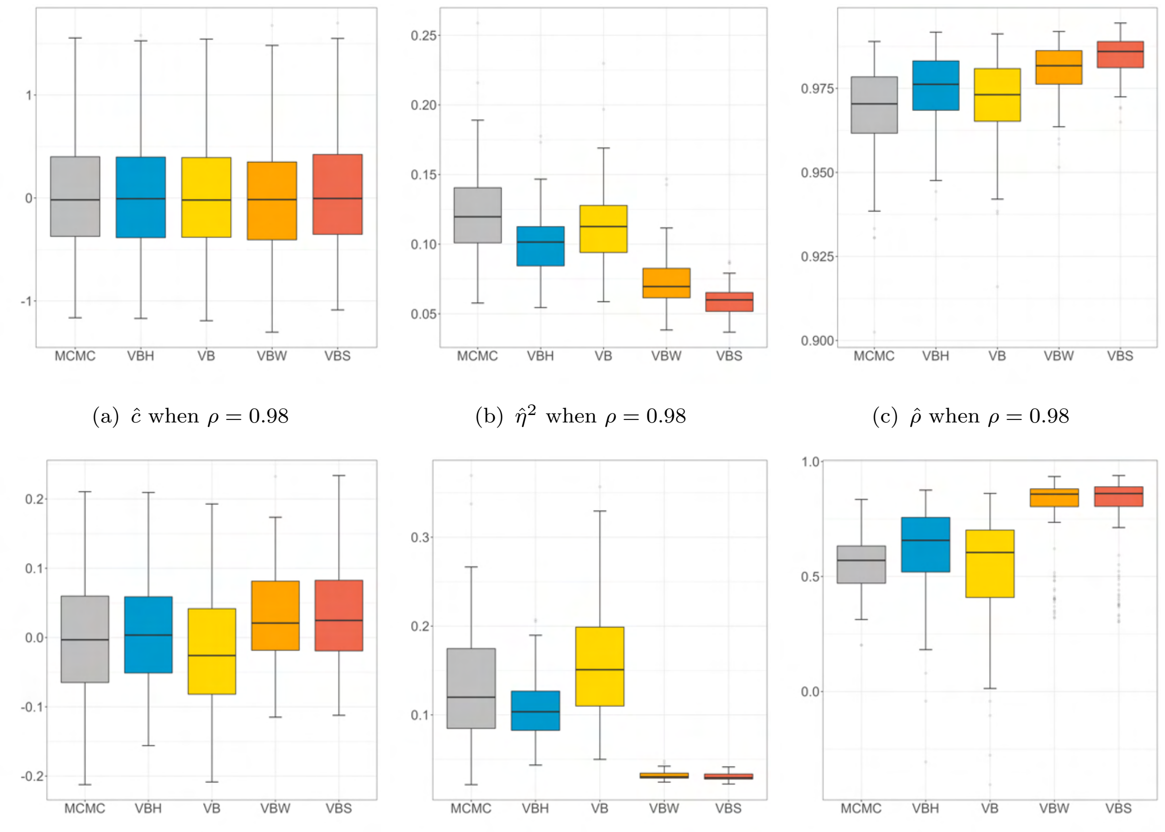
(f) Comp. time when $\rho = 0.70$

FIGURE 3.3: Latent volatility estimates: mean squared error (MSE), global accuracy ($g\mathcal{ACC}$) and computational time across methods. We report the simulation results for both $\rho = 0.98$ (top panels), and $\rho = 0.7$ (bottom panels).

modifications of $\mathbf{W}$ represent an arbitrary intervention on the approximating density $q(\mathbf{h})$. Compared to the baseline `VB`, the smooth approximations have a lower accuracy in the estimate of the underlying AR(1) latent process. Interestingly, similar to `CY` the global accuracy with respect to the `MCMC` deteriorates as the persistence of the latent log-volatility process decreases.

The last column of Figure 3.3 shows that our variational Bayes is less computationally expensive compared to both `MCMC` and `CY` methods. The gain in terms of computational cost holds for both highly persistent latent stochastic volatility (top-right panel) and lower-persistent volatility (bottom-right panel). More generally, our `VB` is almost an order of magnitude faster than `MCMC`, on average. This intuitively represents an advantage when implementing real-time predictions for a large set of equity strategies, as in our main empirical application.

Figure 3.3 suggests that the accuracy of our variational Bayes estimation framework deteriorates when smoothness on the latent state is imposed via the structure in $\mathbf{W}$. We



(a) $\hat{c}$ when $\rho = 0.98$

(b) $\hat{\eta}^2$ when $\rho = 0.98$

(c) $\hat{\rho}$ when $\rho = 0.98$

(d) $\hat{c}$ when $\rho = 0.70$

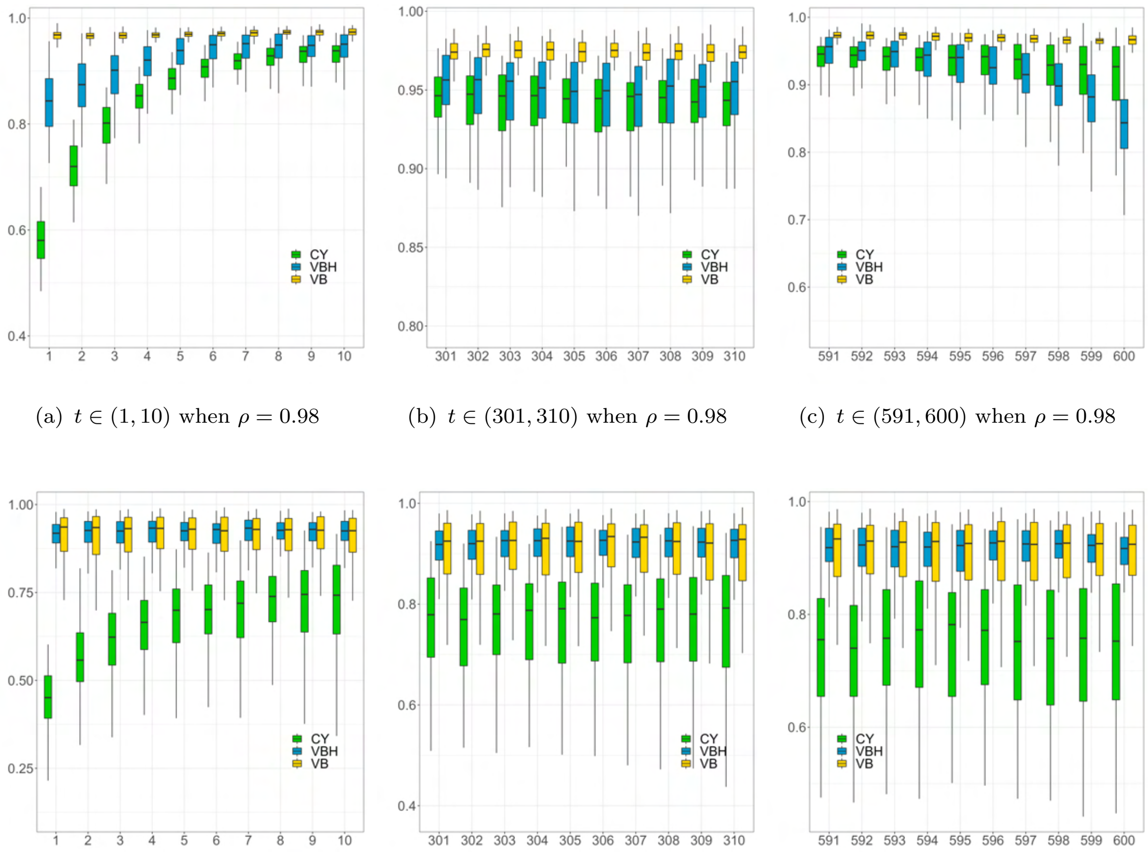
(e) $\hat{\eta}^2$ when $\rho = 0.70$

(f) $\hat{\rho}$ when $\rho = 0.70$

FIGURE 3.4: Estimates for the latent process parameters from different methods. We report the simulation results for both $\rho = 0.98$ (top panels), and $\rho = 0.7$ (bottom panels).

now investigate more in details why that is the case by looking at the posterior estimates of the parameters of interest $\{c, \eta^2, \rho\}$ for difference specifications of $\mathbf{W}$. Figure 3.4 shows that by imposing smoothness in the form of either B-spline or a Daubechies wavelet basis forces the posterior estimates of ρ to be close to one, irrespective of the actual level of persistence in the underlying latent process. Similarly, the estimates of the latent state variance η^2 are smaller for both `VBS` and `VBW` versus `MCMC`'s, and even more so when $\rho = 0.7$. Figure 3.4 confirms the intuition that a lower accuracy of the posterior estimates of the latent state is due to a tight regularization of the parameters implied by smoothing. The effect on the conditional variance estimates is particularly striking.

Beside the possibility of introducing smoothness in the estimates, our variational Bayes approach relax the assumption that the initial distribution $q(h_0)$ is independent on the trajectory of the latent state $q(\mathbf{h}_1)$, that is, we do not assume $q(\mathbf{h}) = q(h_0)q(\mathbf{h}_1)$. Figure 3.5 shows that this generalisation has a non-negligible impact on the posterior



(a) $t \in (1, 10)$ when $\rho = 0.98$
(b) $t \in (301, 310)$ when $\rho = 0.98$
(c) $t \in (591, 600)$ when $\rho = 0.98$

(d) $t \in (1, 10)$ when $\rho = 0.70$
(e) $t \in (301, 310)$ when $\rho = 0.70$
(f) $t \in (591, 600)$ when $\rho = 0.70$

FIGURE 3.5: Accuracy measure at each time. This figure depicts the accuracy of our variational Bayes inference method against the global approximation method proposed by Chan and Yu (2022). The top (bottom) panels report the accuracy when $\rho = 0.98$ ($\rho = 0.7$) in different periods in the timeline.

estimate of the latent state, especially at the beginning on the sample. This is shown by comparing the accuracy ($\mathcal{ACC}$) for different slices of observations. The top (bottom) panels report the accuracy when $\rho = 0.98$ ($\rho = 0.7$). We report the estimation results for $t \in (1, 10)$ in the left panel, $t \in (301, 310)$ in the middle panel, and $t \in (591, 600)$ in the right panel. The simulation results show that our variational Bayes approach maintains an optimal performance over all the timeline. On the other hand, the accuracy of `CY` drops at the beginning of the time series. This is due to the restrictive independence assumption between the initial condition and the rest of the latent state trajectory $q(\mathbf{h}) = q(h_0)q(\mathbf{h}_1)$.

3.4 Smoothing volatility targeting

We now investigate the statistical and economic value of our smooth volatility estimate within the context of volatility targeting across a large set of equity strategies. We first consider the nine equity factors examined by Moreira and Muir (2017). We collect daily and monthly data on the excess returns on the market, and the daily and monthly returns on the size, value, profitability and investment factors as originally proposed by Fama and French (2015), in addition to the profitability and investment factors from Hou *et al.* (2015) and the betting-against-beta factor from Frazzini and Pedersen (2014).

We augment the first group of test portfolios with a second group covering a broader set of trading strategies based on established asset pricing factors. We start with the list of 153 characteristic-managed portfolios, or *factors*, reported in Jensen *et al.* (2022). We then restrict our analysis to value-weighted strategies that can be constructed using the Center for Research in Security Prices (CRSP) monthly and daily stock files, the Compustat Fundamental annual and quarterly files, and the Institutional Broker Estimate (IBES) database. In addition, we exclude a handful of strategies for which there are missing returns. This process identifies 149 value-weighted long-short portfolios for which we collect both daily and monthly returns. For a more detailed description of the portfolio construction we refer to Jensen *et al.* (2022). The combined sample consists of 158 equity trading strategies.

For a given equity trading strategy, let y_t be the buy-and-hold excess portfolio return in month t. We follow Moreira and Muir (2017) and construct the corresponding volatility-managed portfolio return y_t^{σ} as

$$y_t^{\sigma} = \frac{c^*}{\widehat{\sigma}^2_{t|t-1}} y_t, \tag{3.22}$$

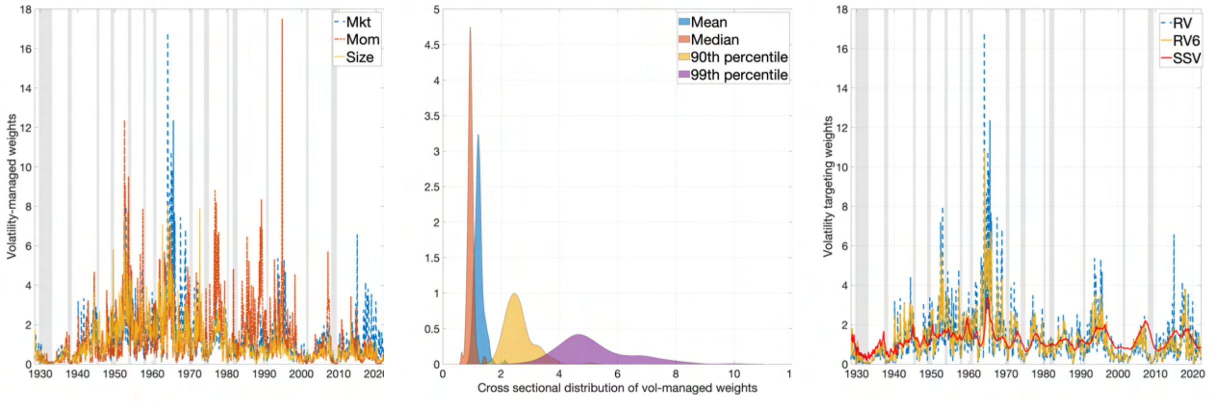
where c^* is a constant chosen such that the unconditional variance of the managed y_t^{σ} and unmanaged y_t portfolios coincide, and $\widehat{\sigma}^2_{t|t-1}$ is the variance forecast of unscaled portfolio returns based on information available up to the previous month $t-1$. The objective of (3.22) is to adjust the capital invested in the original equity strategy based on the inverse of the predicted variance. Effectively, a volatility-managed portfolio is targeting a constant level of volatility, rather than a constant level of notional capital exposure. As such, the dynamics investment position in the underlying portfolio $c^*/\widehat{\sigma}^2_{t|t-1}$ is a measure of (de)leverage required to invest in the volatility-portfolio in month t. Notice that in the standard implementation in (3.23) the scaling parameter c^* is not know by an investor in real time as it requires to observe the full time series of the unscaled returns y_t and the volatility forecasts $\widehat{\sigma}^2_{t|t-1}$.

A benchmark approach to approximate the variance forecast at month t, $\widehat{\sigma}^2_{t|t-1}$ is to use the previous month's realized variance (henceforth `RV`) calculated based on daily portfolio returns (see, e.g., Barroso and Santa-Clara, 2015; Daniel and Moskowitz, 2016; Moreira and Muir, 2017; Cederburg *et al.*, 2020; Barroso and Detzel, 2021),

$$\widehat{\sigma}^2_{t|t-1} = \frac{22}{D_{t-1}} \sum_{j=1}^{D_{t-1}} y^2_{j,t-1}, \tag{3.23}$$

where $y_{j,t-1}$ be the excess returns on a given portfolio in day $j = 1, \dots, D_{t-1}$ for month $t-1$. In addition to the realised variance, we compare our smoothing volatility targeting approach (`SSV`) against a variety of alternative rescaling approaches. The first uses the expected variance from a simple AR(1) rather than realized variance (`RV AR`), which helps to reduce the extremity of the weights. Second, we follow Barroso and Detzel (2021) and consider an alternative six-month window to estimate the longer-term realised variance (`RV6`). Third, we consider both a long-memory model for volatility forecast as proposed by Corsi (2009) (`HAR`), and a standard AR(1) latent stochastic volatility model (`SV`) (see, e.g., Taylor, 1994). Finally, we consider a plain GARCH(1,1) specification (`Garch`), which has been shown to be a challenging benchmark in volatility forecasting (see, Hansen and Lunde, 2005). Notice that, for comparability with the existing literature on volatility-managed portfolios, we assume a constant mean μ in the observation equation (3.1), such that there are no covariates. Throughout the empirical analysis we follow Cederburg *et al.* (2020) and consider both unconditional volatility targeting, whereby c^* is calibrated to match the unconditional volatility of the scaled and unscaled portfolios, as well as real-time volatility targeting, whereby c^*_t is calibrated to match the volatility of the scaled and unscaled portfolios at each month t.

As mentioned in the introduction, volatility management based on realised variance estimates is associated with a dramatic increase in portfolio turnover and significant time-varying leverage. Figure 3.6 shows this case in point. The left panel shows the volatility-managed portfolio allocation based on realised variance estimates for three common portfolios; the market, and the size and momentum factors as originally proposed by Fama and French (1996) and Jegadeesh and Titman (1993), respectively. Simple volatility targeting leads to a tenfold notional exposure compared to the original equity strategy. This excess leverage is pervasive across a broad set of 158 equity trading strategies. For instance, the middle panel in Figure 3.6 shows that volatility targeting based on realised variance leads to a leverage between 1.8 and 4 times for more than 10%, and between 3 to 11 times for at least 1% of the original 158 equity strategies. This makes volatility-managed strategies potentially both risky and costly



(a) Realised variance targeting (b) Leverage distribution (c) Market volatility targeting

FIGURE 3.6: Volatility targeting and portfolio leverage. The left panel reports the rescaling using RV over time for three common factor portfolios. The middle panel reports the cross-sectional distribution of the mean, median, top 10% and top 1% highest leverage weights across all 158 factor portfolios. Right panel shows the weights over time implied by different volatility measures with different smoothness.

to implement, especially when volatility targeting is missed and/or forecasts are not sufficiently accurate (see, e.g., Bongaerts *et al.*, 2020).

3.4.1 A simple statistical appraisal

In this section we provide a statistical appraisal of the performance of our smoothing volatility targeting approach compared to both conventional realised variance measures and benchmark volatility forecasts. This is based on the predictive density of the volatility forecasts obtained for both the non-smooth SV and smooth SSV stochastic volatility models. Recall that real-time volatility targeting for month t takes the form $\omega_t = c_t^*/\widehat{\sigma}^2_{t|t-1}$, $t = 1, \dots, n$. As a result, given the unmanaged factors y_t and the recursively calibrated coefficient c_t^*, for each month we can define the distribution of the volatility-managed returns based on the variational predictive density $q(\sigma_t^2|\mathbf{y})$ with $\mathbf{y}$ collecting the strategy returns up to $t-1$ (see Section 3.2.3 for more details).

Figure 3.7 shows this case in point. The top panels report the distribution of the volatility-managed portfolio returns implied by the non-smooth SV (red area) and smooth SSV (blue area) stochastic volatility models. For the sake of simplicity, we report the volatility-managed returns on the market portfolio over two distinct months, namely October 1995 and March 2009. The returns on the unmanaged portfolio and its scaled version based on previous month's realised variance are indicated as a white and green circle, respectively. By comparing this distribution on a given month with the realised returns on a benchmark strategy for the same month, we can calculate $\mathbb{P}\left(y_t^{\mathcal{M}_1} > y_t^{\mathcal{M}_0}\right)$. Here $y_t^{\mathcal{M}_0}$ represents the returns on the benchmark volatility managing method, for e.g., RV, whereas $y_t^{\mathcal{M}_1}$ the returns on volatility targeting based on either a non-smooth or a smooth stochastic volatility model.

The bottom panel of Figure 3.7 shows an example of the distribution of the returns on a volatility-managed momentum portfolio. Note that the distribution of SSV and SV can be highly time varying and the large negative performance of the unmanaged momentum strategy in March-May 2009 coincides with the so-called *momentum crashes* (see Barroso and Santa-Clara, 2015; Daniel and Moskowitz, 2016; Bianchi *et al.*, 2022).

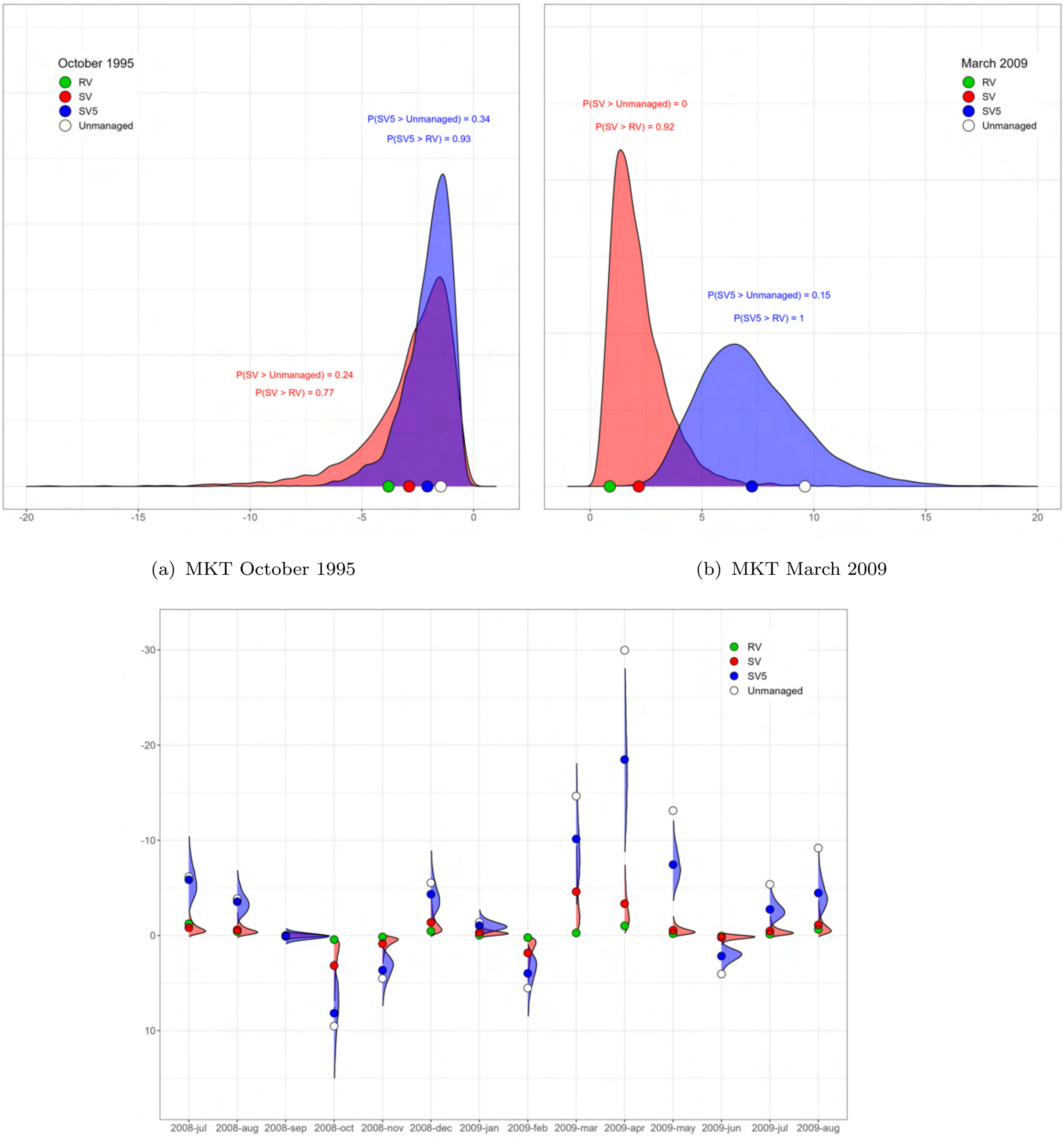
Two interesting facts emerge. First, and perhaps not surprisingly, a non-smooth stochastic volatility model tends to produce relatively similar volatility adjusted returns with few exceptions. In this respect, a standard RV rescaling substantially overperform (underperform) the unmanaged portfolio during periods of large negative (positive) returns. Put it differently, standard volatility targeting helps to mitigate tail risk at the expense of cutting upside opportunities. This is consistent with the abundant empirical evidence that indeed, on average, RV targeting does not systematically outperforms unmanaged portfolios. The second interesting fact pertains our smoothing volatility targeting; the returns on the SSV are closer to the original equity strategy.

We now take to task the intuition highlighted in Figure 3.7 and compare our SSV methodology against all of the competing volatility targeting methods, across all of the 158 equity strategies in our sample. Specifically, we calculate each month two indicator dummies $\mathbb{I}_{i,t}^{+}, \mathbb{I}_{i,t}^{-}$ for each of the $t = 1, \ldots, n$ and each of the $i = 1, \ldots, m$ equity trading strategies,

$$\mathbb{I}_{i,t}^{+} = \begin{cases} 1 & \text{if } \mathbb{P}\left(y_{i,t}^{\texttt{SSV}} > y_{i,t}^{\mathcal{M}_0}\right) > 0.95 \\ 0 & \text{otherwise} \end{cases} \qquad \mathbb{I}_{i,t}^{-} = \begin{cases} 1 & \text{if } \mathbb{P}\left(y_{i,t}^{\texttt{SSV}} < y_{i,t}^{\mathcal{M}_0}\right) > 0.95 \\ 0 & \text{otherwise} \end{cases} \tag{3.24}$$

We can then calculate $p_i^{+} = n^{-1}\sum_{t=1}^{n} \mathbb{I}_{i,t}^{+}$ and $p_i^{-} = n^{-1}\sum_{t=1}^{n} \mathbb{I}_{i,t}^{-}$, with n the sample of observations, for each equity trading strategy. These indicate the frequency over the full sample with which SSV over-performs $\mathcal{M}_0$, i.e., p_i^{+}, or SSV under-performs $\mathcal{M}_0$, i.e., p_i^{-}.

Figure 3.8 reports the difference between p_i^{+} and p_i^{-} for all 158 equity strategies. This indicates the imbalance between out-performance and under-performance of our $y_{i,t}^{\texttt{SSV}}$ compared to a benchmark $y_{i,t}^{\mathcal{M}_0}$. The left panel compares our SSV against the original factor portfolios U and the volatility targeting based on the realised variance RV. The comparison against the unscaled factors confirms the results of Cederburg *et al.* (2020);



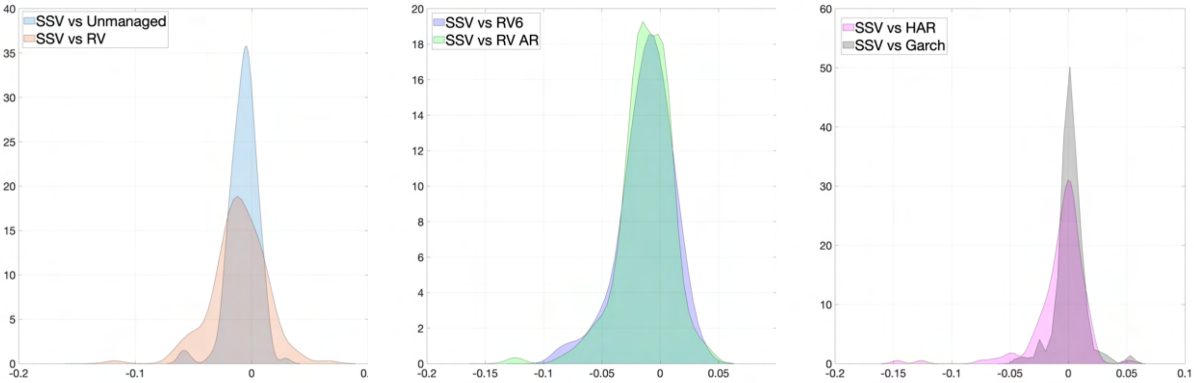
(a) MKT October 1995

(b) MKT March 2009

(c) Momentum factor over 2008/2009

Figure 3.7: Probabilistic assessment of volatility-managed returns. The plot reports the distribution of the volatility-managed portfolio returns implied by the non-smooth `SV` (red area) and smooth `SSV` (blue area) stochastic volatility models. The `RV`-managed an the unmanaged returns are highlighted as green and white circles, respectively.

there is no systematic out-performance of volatility targeting versus unmanaged equity strategies over the sample under investigation. This is reflected in the fact that the difference between p_i^+ and p_i^- is centered around zero for the cross section of equity



(a) $y_t^{\texttt{SSV}}$ vs $y_t^{\texttt{U}}$, $y_t^{\texttt{RV}}$

(b) $y_t^{\texttt{SSV}}$ vs $y_t^{\texttt{RV6}}$, $y_t^{\texttt{RVAR}}$

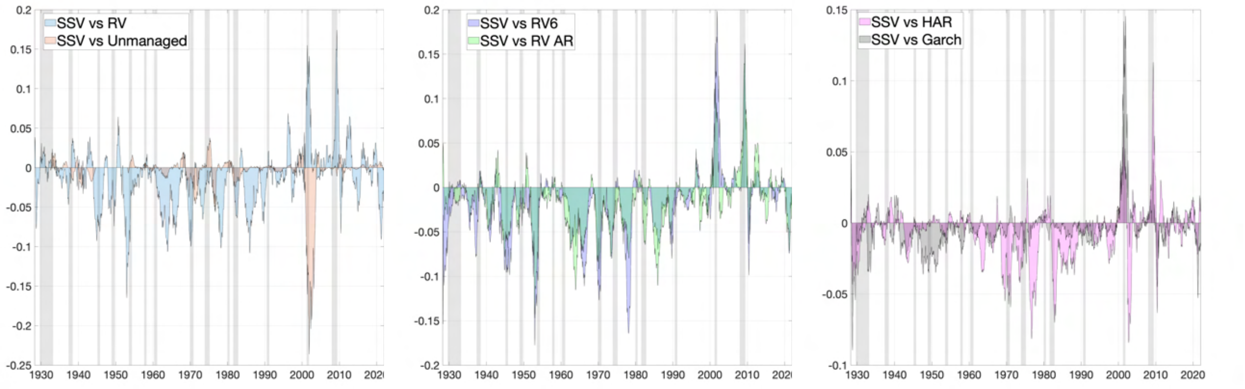
(c) $y_t^{\texttt{SSV}}$ vs $y_t^{\texttt{HAR}}$, $y_t^{\texttt{Garch}}$

FIGURE 3.8: Smoothing vs alternative volatility targeting for the full sample. This figure reports $p_i = p_i^+ - p_i^-$ for the cross section of 158 equity trading strategy. The left panel compares our `SSV` versus `U` and `RV`. The middle panel compares our `SSV` against two alternative smoothing volatility forecasts used in the literature, i.e., `RV6` and `RV AR`. The right panel compares out `SSV` against two popular volatility forecasting methods, such as `HAR` and `Garch`.

strategies. The middle and right panel also confirms that, unconditionally over the full sample, the performance of our `SSV` does not systematically dominate other competing volatility targeting methods. For instance, the spread $p_i = p_i^+ - p_i^-$ is as low as -0.1 and as high as 0.05 when comparing `SSV` vs `RV6`. Similarly, p_i ranges between -0.05 and 0.05 when comparing our `SSV` against the `HAR` or the `Garch` methods.

The results in Figure 3.8 show that the returns on volatility-managed portfolios are statistically equivalent to unscaled factors, at least unconditionally. We now look at a conditional aggregation of the indicators $\mathbb{I}_{i,t}^+$ and $\mathbb{I}_{i,t}^-$. Specifically, we construct a $p_t^+ = m^{-1}\sum_{i=1}^m \mathbb{I}_{i,t}^+$ and $p_t^- = m^{-1}\sum_{i=1}^m \mathbb{I}_{i,t}^-$, with m the number of equity strategies, for month $t = 1, \ldots, n$. Figure 3.9 reports the spread $p_t = p_t^+ - p_t^-$ across the whole sample of observations. The left panel compares the performance of `SSV` versus `RV` and the unmanaged factors `U`. Two interesting facts emerge; first, for the most part of the sample the performance of the `SSV` is subpar compared to the `RV`. This is primarily concentrated in the expansionary periods, whereby volatility is low and the exposure to the original unscaled portfolios is levered up (see, e.g., Figure 3.6).

Second, a smooth volatility targeting substantially improves upon `RV` during the recession in the aftermath of the dot-com bubble and the great financial crisis of 2008/2009. Interestingly, most of the under-performance of `SSV` versus `U` is concentrated during the burst of the dot-com bubble. A possible explanation is that volatility-targeting implies a deleveraging on the original factor, in period in which high volatility



(a) $y_t^{\mathtt{SSV}}$ vs $y_t^{\mathtt{U}}$, $y_t^{\mathtt{RV}}$ (b) $y_t^{\mathtt{SSV}}$ vs $y_t^{\mathtt{RV6}}$, $y_t^{\mathtt{RVAR}}$ (c) $y_t^{\mathtt{SSV}}$ vs $y_t^{\mathtt{HAR}}$, $y_t^{\mathtt{Garch}}$

FIGURE 3.9: Smoothing vs alternative volatility targeting over time. This figure reports the probability $p_t = p_t^+ - p_t^-$ for the sample period. The left panel compares our `SSV` versus `U` and `RV`. The middle panel compares our `SSV` against two alternative smoothing volatility forecasts used in the literature, i.e., `RV6` and `RV AR`. The right panel compares out `SSV` against two popular volatility forecasting methods, such as `HAR` and `Garch`.

did not necessarily correspond to large losses in the original equity factors. The middle and right panel in Figure 3.9 shows that alternative volatility measures to `RV` share a similar pattern compared to our `SSV`; that is, by smoothing volatility forecasts the performance during major recessions improves at the expenses of a subpar performance during economic expansions and/or lower-volatility periods.

3.4.2 Economic evaluation

Most prior studies assess the value of volatility targeting strategies by comparing the Sharpe ratios obtained by scaled factors y_t^σ as in (3.22), with the Sharpe ratios obtained from the original factors y_t (see, e.g., Barroso and Santa-Clara, 2015; Daniel and Moskowitz, 2016; Moreira and Muir, 2017; Bianchi *et al.*, 2022). Moreover, for each equity strategy and volatility-targeting methodology, we estimate the spanning regression on both the scaled and unscaled returns,

$$y_t^\sigma = \alpha + \beta y_t + \epsilon_t. \tag{3.25}$$

The economic implication of $\alpha > 0$ is that volatility scaled portfolios may expand the mean-variance frontier relative to the unscaled portfolios (see, e.g., Gibbons *et al.*, 1989). We test this assumption by comparing the certainty equivalent return (CER) when factoring in moderate levels of notional transaction costs. Specifically, we compare two strategies: (i) a strategy that allocates between a given volatility-managed portfolio and its corresponding original portfolio, and (ii) a strategy constrained to invest only in the original portfolio. The baseline combination correspond to the optimal mean-variance allocation assuming a risk aversion coefficient equal to five. Additional results considering leverage constraints are available in Appendix B.4.

Results without transaction costs. Table 3.1 reports the annualised Sharpe ratio (henceforth SR) and the Sortino ratio, for both unconditional and real-time volatility targeting. For each performance measure, we report both the mean value and the 2.5th, 25th, 50th, 75th, and 97.5th percentiles across all the 158 equity trading strategies. Both the original and the volatility-managed factors yield a positive annualised Sharpe ratio, on average. The risk-adjusted performance is comparable across volatility estimates. For instance, the annualised SR from the `RV` is 0.28 against 0.26 from `SSV`. The dispersion of SRs across equity strategies is also quite comparable across methods. For instance, the 97.5th percentile in the distribution of SRs is 0.69 for the `SSV` against 0.81 from a six-month realised variance `RV6`.

To determine whether the SR from a given volatility-managed portfolio is statistically different from its unmanaged counterpart, we follow Cederburg *et al.* (2020) and implement a block bootstrap approach as proposed by Jobson and Korkie (1981); Ledoit and Wolf (2008). Table 3.1 reports the percentage, out of all 158 equity strategies, of SR differences that are positive or negative, and are statistically significant at the 5% level. The results in Table 3.1 confirms the existing evidence in the literature that volatility-managed portfolios do not systematically outperform their original counterparts (see, e.g., Barroso and Detzel, 2021). For instance, `RV` yields a significantly larger (smaller) SR compared to unmanaged portfolios for 6% (2.5%) of the 158 equity trading strategies considered.

The percentage of higher and significant SRs slightly improves when using our `SSV` method versus both `RV` and all other competing volatility forecasts. Nevertheless, the percentage of significant and positive SRs tend to be quite similar across different volatility targeting estimates. Table 3.1 also reveals that the gross performance across methods is quite comparable when looking at the risk-adjusted returns with a focus on downside risk only. For instance, the average Sortino ratio is 1.44, which is smaller than the 1.77 obtained from the `RV`, but economically fairly close. Again, the Sortino ratios are fairly comparable across scaling methods.

Existing evidence on the performance of volatility-managed portfolios follows from a spanning regression approach. The object of interest is the intercept α, that is a positive

TABLE 3.1: Volatility-managed portfolios and original equity strategies.

This table compares the performance of volatility-managed and original portfolios (U) for the cross section of 158 equity strategies. For a given factor, the volatility-managed factor return in month t is based on a forecast of the conditional variance. For each volatility targeting method we report the mean annualised Sharpe ratio and Sortino ratio, as well as their 2.5th, 25th, 50th, 75th, and 97.5th percentiles in the cross section of equity strategy. In addition, we report the fraction of volatility-managed portfolios that generate a Sharpe ratio which is statistically different from the unscaled strategy (see, Ledoit and Wolf, 2008), and is either positive or negative. The table reports both the performance measure with the scale parameter c^* calibrated over the full sample (unconditional targeting) or at each month t, c^*_t (real time targeting).

	Unconditional targeting								Real time targeting							
	U	RV	RV6	RV AR	HAR	Garch	SV	SSV	U	RV	RV6	RV AR	HAR	Garch	SV	SSV
SR																
Mean	0.24	0.28	0.29	0.29	0.27	0.26	0.26	0.26	0.24	0.27	0.28	0.28	0.27	0.26	0.26	0.26
Percentiles																
2.5	-0.12	-0.20	-0.22	-0.19	-0.20	-0.21	-0.20	-0.20	-0.12	-0.22	-0.23	-0.20	-0.20	-0.22	-0.21	-0.19
25	0.08	0.07	0.06	0.07	0.07	0.03	0.03	0.06	0.08	0.07	0.06	0.08	0.06	0.03	0.02	0.07
50	0.22	0.26	0.27	0.27	0.26	0.25	0.30	0.23	0.22	0.25	0.26	0.26	0.27	0.26	0.28	0.22
75	0.37	0.48	0.48	0.49	0.45	0.43	0.44	0.43	0.37	0.45	0.48	0.46	0.45	0.44	0.43	0.41
97.5	0.63	0.79	0.81	0.80	0.73	0.78	0.79	0.69	0.63	0.75	0.77	0.76	0.74	0.77	0.76	0.68
$p< 0.05$ & $SR> 0$	6.33	7.59	7.59	8.23	8.86	7.59	10.13			5.06	6.96	7.59	8.23	8.86	8.23	11.39
$p< 0.05$ & $SR< 0$	2.53	0.00	1.27	1.90	6.33	5.06	5.06			2.53	0.63	1.27	1.27	4.43	5.70	3.80
Sortino																
Mean	1.44	1.77	1.84	1.79	1.60	1.56	1.61	1.55	1.44	1.74	1.85	1.75	1.61	1.59	1.61	1.51
Percentiles																
2.5	-0.79	-1.06	-1.27	-1.06	-1.20	-1.21	-1.19	-1.12	-0.79	-1.23	-1.39	-1.22	-1.22	-1.23	-1.26	-1.11
25	0.49	0.46	0.44	0.50	0.39	0.17	0.18	0.35	0.49	0.48	0.41	0.47	0.38	0.16	0.13	0.44
50	1.38	1.59	1.66	1.62	1.55	1.67	1.72	1.43	1.38	1.58	1.63	1.61	1.55	1.57	1.67	1.42
75	2.17	2.90	2.95	2.85	2.69	2.63	2.53	2.40	2.17	2.80	2.90	2.81	2.66	2.62	2.54	2.39
97.5	3.50	5.77	5.03	5.47	4.48	4.77	4.64	4.18	3.50	4.84	4.75	4.73	4.55	4.73	4.62	4.09

α implies that a combination of the original unmanaged factor and its volatility-managed counterpart expands the mean-variance frontier compared to investing in the original unscaled portfolio alone (see, e.g., Gibbons *et al.*, 1989). The top panel in Table 3.2 reports the mean alpha (in %) across all the 158 equity strategies obtained from different volatility target methods. Similar to the Sharpe ratios, we report the 2.5th, 25th, 50th, 75th, and 97.5th percentile of the alphas across all rescaled portfolios, in addition to the mean value across equity strategies. Volatility targeting based on realised variance RV achieves the highest average gross α (1.68%), on par with the six-month realised variance RV6. This holds both for the unconditional and the real-time volatility implementation. The fraction of positive and significant gross alphas, at a 5% level, is also higher for the RV and RV6 methods.

Moreira and Muir (2017) link their spanning test results to appraisal ratios and utility gains for investors. Both metrics can be red in the context of mean-variance portfolio choice. The appraisal ratio for a given scaled strategy is $AR = \widehat{\alpha}/\widehat{\sigma}_{\varepsilon}$, where $\widehat{\alpha}$ is the estimated gross alpha from the spanning regression and $\widehat{\sigma}_{\varepsilon}$ the root mean squared error. The squared of the appraisal ratio reflects the extent to which volatility management can be

Table 3.2: Spanning regression results.

This table reports the results from a spanning regression of the form $y_t^\sigma = \alpha + \beta y_t + \epsilon_t$, with y_t^σ the returns on the volatility managed portfolio and y_t^σ its unscaled counterpart. We report the estimated alphas ($\widehat{\alpha}$ in %), the appraisal ratio $AR = \widehat{\alpha}/\widehat{\sigma}_\varepsilon$ and the difference in the certainty equivalent return between and investor that can access both the volatility-managed and the original portfolio, and an investor constrained to invest in the original portfolio only ΔCER. We also report the fraction of volatility-managed alphas that are significant and either positive or negative. The table reports both the performance measure with the scale parameter c^* calibrated over the full sample (unconditional targeting) or at each month t, c_t^* (real time targeting).

	Unconditional targeting							Real-time targeting						
	RV	RV6	RV AR	HAR	Garch	SV	SSV	RV	RV6	RV AR	HAR	Garch	SV	SSV
$\alpha(\%)$														
Mean	1.68	1.68	1.49	0.93	1.20	1.17	0.74	1.78	1.84	1.50	0.98	1.39	0.49	0.34
Percentiles														
2.5	-1.87	-1.77	-1.59	-1.77	-2.51	-2.33	-1.62	-2.93	-1.83	-2.52	-1.45	-2.19	-0.96	-0.97
25	-0.04	-0.10	0.03	-0.13	-0.34	-0.25	-0.32	-0.05	-0.15	0.02	-0.14	-0.29	-0.12	-0.19
50	1.11	1.04	0.92	0.66	0.66	0.69	0.32	1.04	0.99	0.88	0.55	0.60	0.28	0.15
75	2.23	2.23	1.91	1.30	1.80	1.61	1.08	1.98	1.90	1.56	1.26	1.27	0.60	0.56
97.5	7.06	8.03	6.53	5.39	6.49	6.21	3.63	10.78	10.48	9.08	6.38	8.57	2.40	2.12
$p< 0.05$ & $\alpha > 0$	36.08	40.51	34.18	26.58	32.28	31.65	31.01	32.91	34.18	33.54	28.48	32.28	29.75	27.22
$p< 0.05$ & $\alpha < 0$	1.90	2.53	1.90	2.53	8.86	5.70	6.96	1.90	2.53	3.16	2.53	8.23	7.59	9.49
AR														
Mean	0.05	0.05	0.05	0.04	0.03	0.04	0.03	0.04	0.05	0.05	0.04	0.03	0.03	0.03
Percentiles														
2.5	-0.06	-0.06	-0.06	-0.06	-0.09	-0.08	-0.09	-0.06	-0.06	-0.07	-0.06	-0.08	-0.08	-0.09
25	0.00	-0.01	0.00	-0.01	-0.02	-0.02	-0.02	0.00	0.00	0.00	-0.01	-0.02	-0.02	-0.02
50	0.04	0.05	0.05	0.04	0.03	0.04	0.03	0.04	0.05	0.04	0.04	0.04	0.04	0.03
75	0.09	0.09	0.09	0.07	0.08	0.08	0.08	0.08	0.08	0.08	0.07	0.07	0.07	0.07
97.5	0.19	0.19	0.20	0.18	0.18	0.17	0.16	0.16	0.18	0.17	0.19	0.18	0.17	0.16
$\Delta CER(\%)$														
Mean	17.91	18.78	16.37	9.30	13.77	12.55	9.10	14.52	16.77	12.49	6.77	11.99	3.77	4.03
Percentiles														
2.5	-5.93	-4.56	-4.65	-3.82	-7.75	-6.83	-6.10	-23.07	-7.84	-10.55	-9.91	-10.98	-37.33	-34.61
25	0.06	0.52	0.35	0.00	-0.36	-0.03	-0.75	5.02	4.97	3.79	2.75	3.45	0.65	1.10
50	5.69	5.84	5.29	2.85	3.12	3.47	1.83	11.25	10.87	11.26	9.15	9.33	6.44	6.33
75	19.86	17.65	16.26	10.81	12.82	10.34	7.13	22.37	24.68	21.10	16.76	13.78	11.84	11.35
97.5	91.73	65.32	80.30	40.51	49.43	47.05	26.22	75.94	80.12	62.56	38.23	41.86	29.19	22.99

used to increase the slope of the mean-variance frontier (see, Gibbons *et al.*, 1989). The mid panel of Table 3.2 shows the results for both unconditional and real-time volatility targeting. On average, the appraisal ratio from the RV is higher (0.05) compared to our SSV (0.03). The cross-sectional distribution of the ARs is quite symmetric, as the mean and median estimates tend to coincide.

Perhaps more interesting is the fact that the estimates of the $\widehat{\alpha}$ from the spanning regressions can be used to quantify the utility gain from volatility management. This is achieved by comparing the certainty equivalent return (CER) for the investor who has access to both the original and the volatility-managed factor against the investor who is constrained to the original equity strategy only. We follow Cederburg *et al.* (2020); Barroso and Detzel (2021) and define the difference in CER from the unmanaged and the scaled portfolios as

$$\Delta\text{CER} = \frac{\text{SR}\left(z_t^*\right) - \text{SR}\left(y_t\right)}{2\gamma},$$

where $\text{SR}\left(y_t\right)$ is the Sharpe ratio of the unscaled portfolio and $\text{SR}\left(z_t^*\right)$ is the Sharpe ratio of the combined strategy $z_t = x_\sigma \omega_t + x$, with $\omega_t = c^*/\widehat{\sigma}^2_{t|t-1}$. The ex post optimal policy $[x_\sigma,\ x]' = \frac{1}{\gamma}\widehat{\Sigma}^{-1}\widehat{\mu}$ allocates a static weight x_σ to the volatility-managed portfolio and a static x weight on the original factor, based on the sample covariance $\widehat{\Sigma}$ and the sample mean $\widehat{\mu}$ returns of the scaled and unscaled portfolios. This policy is equivalent to dynamically adjust the exposure to the original factor portfolio according to z_t, so that the returns on the combined strategy can be obtained as $z_t^* = z_t\, y_t$. The bottom panel of Table 3.2 reports $\Delta CER(\%)$ for the unconditional and real-time volatility targeting.

We follow Cederburg *et al.* (2020); Wang and Yan (2021) and consider a risk aversion coefficient equal to $\gamma = 5$. The ΔCER confirms that volatility targeting based on realised variance does indeed expands ex post the mean-variance frontier relative to the other volatility targeting methods, when no transaction costs or cost-mitigation strategies are considered. For instance, the ΔCER from the `RV` is 18% versus 9% obtained from our `SSV` smoothing volatility forecast. Interestingly, a slightly smoother estimate of realised volatility, i.e., `RV6`, produces a higher $\Delta CER(\%)$, both unconditionally and in real time.

Turnover and leverage. A standard volatility targeting strategy is built upon scaling the original portfolio returns by $c^*/\widehat{\sigma}^2_{t|t-1}$. The often erratic nature of $\widehat{\sigma}^2_{t|t-1}$ based on realised volatility implies that volatility-managed portfolios are associated with high turnover and significant time-varying leverage ω_t. This is likely to cast doubt on the actual usefulness of volatility targeting portfolios under common liquidity constraints (see Moreira and Muir, 2017; Harvey *et al.*, 2018; Bongaerts *et al.*, 2020; Patton and Weller, 2020; Barroso and Detzel, 2021). Table 3.3 shows the amount of portfolio turnover for different volatility targeting methods. The portfolio turnover is calculated as the average absolute change of the leverage weights $|\Delta w|$ (see Moreira and Muir, 2017). We report the mean turnover as well as the 2.5th, 25th, 50th, 75th, and 97.5th percentile across the 158 equity strategies.

Clearly, our `SSV` method substantially reduces the portfolio turnover compared to all other volatility forecasting methods. For instance, the turnover from the `RV` is 0.65 against a 0.05 from `SSV`, on average across equity strategies. Our `SSV` produces a lower turnover not only on average, but for the full cross section of equity strategies. For instance, the 2.5th (97.5th) percentile is 0.03 (0.06) for the `SSV` against a 0.51 (0.91) from

Table 3.3: Portfolios turnover and leverage dispersion.

This table reports a set of descriptive statistics for the volatility-managed portfolio turnover and leverage. The portfolio turnover is calculated as the average absolute change in monthly volatility-managing weights $|\Delta w|$ (see Moreira and Muir, 2017). The leverage is calculated as $\omega_t = \frac{c^*}{\widehat{\sigma}^2_{t|t-1}}$. The table reports both the performance measure with the scale parameter c^* calibrated over the full sample (unconditional targeting) or at each month t, c^*_t (real time targeting).

	Unconditional targeting							Real time targeting						
	RV	RV6	RV AR	HAR	Garch	SV	SSV	RV	RV6	RV AR	HAR	Garch	SV	SSV
Turnover														
Mean	0.65	0.14	0.48	0.23	0.16	0.21	0.05	69.98	27.22	50.05	22.17	15.66	8.99	2.66
Percentiles														
2.5	0.51	0.11	0.32	0.13	0.05	0.10	0.03	42.08	16.20	29.49	12.82	4.59	4.97	1.36
25	0.57	0.12	0.41	0.20	0.13	0.17	0.04	51.17	19.23	37.26	19.26	10.59	7.64	2.34
50	0.62	0.14	0.45	0.23	0.15	0.20	0.05	59.43	22.04	40.98	21.80	14.09	8.35	2.57
75	0.69	0.16	0.54	0.26	0.19	0.24	0.05	86.49	34.09	64.53	24.94	19.25	10.14	2.92
97.5	0.91	0.22	0.71	0.30	0.29	0.33	0.06	128.35	55.72	98.16	33.43	34.72	14.38	4.21
Average leverage														
Mean	1.24	1.30	1.30	1.23	1.24	1.26	1.22	1.33	1.36	1.34	1.22	1.18	0.56	0.73
Percentiles														
2.5	1.00	1.08	1.07	1.06	1.00	1.04	1.02	0.83	0.89	0.91	0.86	0.76	0.33	0.53
25	1.15	1.20	1.21	1.15	1.15	1.18	1.15	1.00	1.06	1.06	1.01	0.93	0.47	0.67
50	1.22	1.29	1.28	1.22	1.22	1.24	1.20	1.19	1.22	1.19	1.14	1.08	0.56	0.73
75	1.30	1.36	1.35	1.29	1.31	1.33	1.26	1.58	1.63	1.57	1.39	1.38	0.62	0.79
97.5	1.59	1.67	1.65	1.53	1.55	1.56	1.45	2.22	2.21	2.22	1.95	1.93	0.79	0.92
Leverage dispersion														
Mean	1.09	0.92	0.79	0.51	0.72	0.72	0.43	1.21	1.00	0.85	0.48	0.70	0.32	0.27
Percentiles														
2.5	0.71	0.55	0.41	0.29	0.33	0.27	0.22	0.64	0.49	0.38	0.28	0.26	0.13	0.14
25	0.92	0.76	0.62	0.44	0.56	0.56	0.36	0.82	0.68	0.58	0.40	0.48	0.24	0.23
50	1.02	0.87	0.74	0.50	0.66	0.64	0.41	0.97	0.80	0.66	0.46	0.58	0.30	0.26
75	1.22	1.04	0.94	0.55	0.87	0.85	0.49	1.62	1.18	1.10	0.55	0.86	0.37	0.32
97.5	1.71	1.39	1.34	0.80	1.38	1.28	0.70	2.47	2.03	1.82	0.82	1.46	0.61	0.40

RV. Perhaps not unexpectedly, the six-month realised variance implies a lower turnover compared to RV. Nevertheless, our SSV stands out in terms of portfolio stability, both within the context of unconditional or real-time volatility targeting.

The middle panel of Table 3.3 also reports the average leverage implied by volatility targeting, i.e., $\omega_t = c^*/\widehat{\sigma}^2_{t|t-1}$. The real-time implementation of the RV portfolio scaling implies a leverage that is almost twice as large as the one implied by SSV volatility targeting (0.73). Differences across volatility methods are lower for the unconditional targeting. In addition, the bottom panel shows that our smoothing volatility forecasting method significantly reduce liquidity demand, that is increases the stability of ω_t over time. For instance, the variability of leverage from SSV is half (0.43) compared to RV (1.09). The leverage mitigation effect of SSV is even more clear when looking at the real-time implementation; the standard deviation of w_t is 0.27, on average across equity strategies. This compares to 1.21, 1, and 0.85 from the RV, RV6 and RV AR, respectively.

Results with transaction costs. Table 3.3 shows that alternative scaling methods, such as HAR, Garch and RV AR indeed helps to stabilise volatility managing compared to a standard RV. Yet, our smoothing volatility prediction SSV generates by the lowest and most stable liquidity demand across all methods. For each equity factor we now consider the costs of the leverage adjustment associated with volatility targeting. We follow Moreira and Muir (2017); Wang and Yan (2021) and consider two alternative levels of transaction costs of 14 basis points (bps) of the notional value traded to implement volatility targeting (see, e.g., Frazzini *et al.*, 2012) and a more conservative 50 basis points (see, e.g., Wang and Yan, 2021).

Table 3.4 reports the net-of-costs performance statistics for the managed factors. After 14 bps costs, the average SR for RV decreases from 0.23 to 0.17. With a more conservative level of transaction costs, the average SR from RV turns to a negative -0.11 annualised. This is in stark contrast of what we obtain by smoothing the volatility predictions; that is, our SSV generates a remarkable stable SR of 0.25 and 0.23 after 14 and 50 basis points of notional trading costs, respectively. Perhaps more importantly, only 10% of volatility-managed portfolios produce a significantly lower SR compared to the unmanaged counterpart even with conservative 50 bps of trading costs. This is in contrast to RV, for which 79% of Sharpe ratios are significantly lower than the unscaled portfolios. Furthermore, when we consider 50 basis points of transaction costs, the Sortino ratio from SSV is 1.38 versus -0.69 from RV, 0.85 from RV6 and 0.98 from a Garch model, respectively.

Table 3.5 reports the results for the spanning regression $y_t^\sigma = \alpha + \beta y_t + \epsilon_t$, with y_t^σ the returns on the volatility managed portfolio net of transaction costs and y_t^σ its the original equity strategy. The top panels report the estimated alphas ($\widehat{\alpha}$ in %). When considering a conservative notional trading cost of 50 basis points, our SSV volatility forecast generates a positive alpha of 0.46% annualised. This is against a large and negative alpha from the RV, RV AR, HAR, and SV methods. Consistent with Barroso and Detzel (2021), a longer-term six-month estimate of the realised variance RV6 improves the volatility-managed alphas (0.12%). Perhaps more importantly, our SSV method generates a significantly positive alpha for 21% of the equity strategies in our sample, against, for instance, a 3%, 9%, and 14% of the strategies from the RV, RV6 and Garch models, respectively.

The appraisal ratio *AR* reported in the middle panel of Table 3.5 confirms that SSV substantially improves upon realised variance measures RV, especially when a conservative transaction cost is factored in. For instance, with 50 basis points of trading costs

Table 3.4: Volatility-managed portfolios with transaction costs.

This table compares the performance of volatility-managed and original portfolios (U) for the cross section of 158 equity strategies. For a given factor, the volatility-managed factor return in month t is based on a forecast of the conditional variance. For each volatility targeting method we report the mean annualised Sharpe ratio and Sortino ratio, as well as their 2.5th, 25th, 50th, 75th, and 97.5th percentiles in the cross section of equity strategy. In addition, we report the fraction of volatility-managed portfolios that generate a Sharpe ratio which is statistically different from the unscaled strategy (see, Ledoit and Wolf, 2008), and is either positive or negative. The table reports the results for two levels of transaction costs, 14 and 50 basis points of the notional value traded to implement volatility targeting.

	14 basis points								50 basis points							
	U	RV	RV6	RV AR	HAR	Garch	SV	SSV	U	RV	RV6	RV AR	HAR	Garch	SV	SSV
SR																
Mean	0.24	0.17	0.25	0.21	0.23	0.23	0.23	0.25	0.24	-0.11	0.14	0.01	0.13	0.16	0.14	0.23
Percentiles																
2.5	-0.12	-0.32	-0.26	-0.28	-0.26	-0.23	-0.24	-0.20	-0.12	-0.65	-0.39	-0.52	-0.40	-0.31	-0.32	-0.22
25	0.08	-0.03	0.02	0.00	0.02	0.00	-0.01	0.05	0.08	-0.30	-0.09	-0.19	-0.08	-0.06	-0.09	0.03
50	0.22	0.16	0.23	0.20	0.21	0.23	0.26	0.23	0.22	-0.14	0.13	0.00	0.11	0.16	0.16	0.21
75	0.37	0.36	0.43	0.41	0.40	0.40	0.39	0.42	0.37	0.05	0.32	0.17	0.27	0.33	0.30	0.39
97.5	0.63	0.69	0.77	0.72	0.69	0.76	0.76	0.68	0.63	0.48	0.66	0.54	0.59	0.71	0.66	0.66
$p < 0.05$ & $SR > 0$		1.90	4.43	3.80	5.06	6.96	6.96	8.86		0.00	1.27	0.00	1.90	3.80	1.27	6.96
$p < 0.05$ & $SR < 0$		15.19	5.70	10.76	6.96	12.03	12.66	5.70		79.11	27.22	65.82	36.71	27.22	36.08	10.13
Sortino																
Mean	1.44	1.08	1.52	1.30	1.35	1.40	1.40	1.50	1.44	-0.69	0.85	0.04	0.75	0.98	0.86	1.38
Percentiles																
2.5	-0.79	-1.92	-1.55	-1.62	-1.52	-1.32	-1.43	-1.15	-0.79	-4.16	-2.29	-3.05	-2.33	-1.77	-1.91	-1.27
25	0.48	-0.21	0.13	-0.01	0.12	0.03	-0.05	0.32	0.48	-1.82	-0.58	-1.22	-0.50	-0.39	-0.53	0.21
50	1.36	0.91	1.40	1.15	1.27	1.48	1.52	1.37	1.36	-0.91	0.78	0.02	0.68	1.01	1.01	1.25
75	2.16	2.21	2.60	2.30	2.37	2.41	2.30	2.34	2.16	0.32	1.84	1.05	1.62	1.98	1.75	2.21
97.5	3.49	5.14	4.87	5.01	4.17	4.65	4.41	4.14	3.49	3.55	4.32	3.85	3.62	4.43	3.84	4.04

the SSV is the only method that can still generate a positive appraisal ratio. By comparison, the RV, RV6, Garch and RV AR all generate significantly negative ARs. The bottom panels report the difference in the certainty equivalent return between and investor that can access both the volatility-managed and the original portfolio, and an investor constrained to invest in the original portfolio only. The utility gain $\Delta CER(\%)$ is highly in favour of our SSV volatility targeting. For instance, for 14 basis points of transaction costs, the second-best performing strategy is the RV6 rescaling with a ΔCER of 9.56%, annualised, against a 14.5% from our SSV.

3.5 Concluding remarks

In this Chapter we propose a parametric variational Bayes approach to approximate the posterior distribution of a stochastic volatility process. The latter has two main advantages with respect to state-of-the-art approaches. Firstly, it enlarges the family of possible optimal densities, leading to higher approximation accuracy when compared to recent variational Bayes algorithm. Second, it allows arbitrary smooth estimates of the

Table 3.5: Spanning regression results with transaction costs.

This table reports the results from a spanning regression of the form $y_t^\sigma = \alpha + \beta y_t + \epsilon_t$, with y_t^σ the returns on the volatility managed portfolio and y_t^σ its unscaled counterpart. We report the estimated alphas ($\widehat{\alpha}$ in %), the appraisal ratio $AR = \widehat{\alpha}/\widehat{\sigma}_\varepsilon$ and the difference in the certainty equivalent return between and investor that can access both the volatility-managed and the original portfolio, and an investor constrained to invest in the original portfolio only ΔCER. We also report the fraction of volatility-managed alphas that are significant and either positive or negative. The table reports the results for two levels of transaction costs, 14 and 50 basis points of the notional value traded to implement volatility targeting.

	14 basis points							50 basis points						
	`RV`	`RV6`	`RV AR`	`HAR`	`Garch`	`SV`	`SSV`	`RV`	`RV6`	`RV AR`	`HAR`	`Garch`	`SV`	`SSV`
α(%)														
Mean	0.58	1.22	0.68	0.51	0.92	0.82	0.66	-2.23	0.12	-1.39	-0.47	0.23	-0.08	0.46
Percentiles														
2.5	-3.00	-2.50	-2.59	-2.18	-2.73	-2.76	-1.71	-6.30	-3.88	-5.42	-3.11	-3.49	-3.92	-1.92
25	-1.02	-0.46	-0.69	-0.49	-0.62	-0.62	-0.40	-3.65	-1.37	-2.52	-1.42	-1.34	-1.44	-0.62
50	0.13	0.76	0.18	0.26	0.41	0.34	0.25	-2.61	-0.27	-1.71	-0.73	-0.29	-0.46	0.06
75	1.17	1.67	1.04	0.87	1.47	1.25	1.01	-1.66	0.65	-0.92	-0.01	0.86	0.46	0.83
97.5	5.62	6.74	5.39	4.92	6.04	5.66	3.53	2.39	5.16	2.58	3.91	5.01	4.29	3.30
$p < 0.05$ & $\alpha > 0$	11.39	26.58	13.92	15.19	28.48	20.25	28.48	3.16	8.86	4.43	6.33	14.56	8.23	21.52
$p < 0.05$ & $\alpha < 0$	14.56	7.59	12.03	9.49	13.92	13.29	10.13	70.89	23.42	60.13	37.34	23.42	32.28	15.82
AR (%)														
Mean	0.60	3.21	1.22	1.49	2.19	1.80	2.50	-10.23	-1.29	-8.30	-4.31	-1.01	-2.70	1.04
Percentiles														
2.5	-10.51	-8.26	-10.20	-8.60	-9.76	-9.67	-9.72	-25.05	-14.14	-21.95	-16.84	-13.79	-15.14	-11.17
25	-4.24	-1.76	-3.90	-3.39	-2.79	-3.54	-2.75	-15.13	-6.40	-13.33	-8.92	-6.53	-9.02	-4.55
50	0.43	2.85	1.01	1.83	1.83	2.06	2.17	-10.33	-0.99	-8.44	-4.89	-1.88	-2.76	0.52
75	4.82	6.97	5.03	4.70	6.97	6.15	7.55	-5.78	2.82	-4.16	-0.05	4.34	2.27	6.09
97.5	16.35	16.31	16.92	15.75	17.21	16.08	15.43	8.14	12.20	9.53	12.18	13.48	12.50	14.43
ΔCER(%)														
Mean	2.85	9.56	9.05	9.10	6.35	3.57	14.50	-14.50	-0.31	-9.70	-2.26	0.65	-3.75	9.47
Percentiles														
2.5	-17.02	-7.83	-9.22	-6.10	-9.77	-9.42	-6.53	-49.06	-18.03	-31.85	-15.63	-20.97	-21.88	-8.28
25	-3.33	-0.79	-1.94	-1.56	-1.47	-1.68	-0.95	-22.35	-5.21	-15.50	-7.62	-6.03	-7.31	-2.21
50	0.04	3.14	0.07	0.92	1.64	1.13	1.24	-8.72	-0.62	-6.79	-3.25	-0.54	-1.90	0.14
75	5.28	12.63	7.40	4.99	10.24	7.00	6.10	0.45	1.51	-0.91	-0.01	4.30	1.04	4.52
97.5	43.98	59.18	59.85	29.98	46.04	34.18	25.00	19.55	34.38	20.69	21.91	38.48	18.34	22.41

latent process. This is done by forcing the approximation to have a specific form of the mean vector.

Beside methodological results, we highlight the importance of the proposed algorithm in practice. Prior studies found that volatility-managed portfolios that increase leverage when volatility is low produce statistically equivalent economic value compared to the original unscaled factors. We show that such equivalence is primarily due to the extreme leverage implied by volatility targeting. Indeed, volatility-managed portfolios based on standard realised variance tend to have extremely levered exposure to the original factors; such exposure is highly time varying. When factoring in moderate levels of notional transaction costs the benefit of volatility-managing disappears.

The variational Bayes inference with possible smoothness regularises turnover and mitigates the effect of transaction costs on volatility-managed portfolios. Using a large set of 158 equity strategies, we provide evidence that our smoothing volatility targeting approach has economic value when conservative levels of transaction costs are considered. This has important implications for both the risk-adjusted returns and the mean-variance efficiency of volatility-managed portfolios.

Chapter 4

Dynamic sparsity in time-varying parameter regressions

4.1 Introduction

Dynamic linear models are widely used tools in statistical analysis for time series data, since they are highly interpretable and useful for forecasting purposes. However, due to the increasing number of observations (n) and the availability of large set of variables (p), two main problems arise. As a first fact, the assumption of constant parameters on large time windows is unrealistic and yields to poor out-of-sample predictions. In this direction, many authors (see, e.g., Cogley and Sargent, 2001; Primiceri, 2005; West and Harrison, 2006) highlight the importance of consider a time-varying effect of the covariates in the model specification, moving the attention to state-space models. The second issue concerns the possible over-fitting when dealing with large set of predictors. The latter causes unreliable forecasts and complicates the interpretation of the results. For this reason, estimates' regularization and variable selection techniques are essential to perform accurate inference and automatically separate the true signal from noise, therefore basing the entire analysis on a reduced subset of relevant covariates (Varian, 2014). In the Bayesian literature, shrinkage priors (Park and Casella, 2008; Griffin and Brown, 2010; Carvalho *et al.*, 2010) and variable selection methods (Ročková and George, 2014, 2018) are usually considered to push towards zero unimportant coefficients and select the best set of relevant ones. These methods have the effect to control the bias-variance trade off usually leading to improved prediction performances and to provide the best subset of variables on which the practitioners can base their analysis.

A recent stream of literature in Bayesian statistics, considered the combination of time-varying regressions and variable selection approaches. Belmonte *et al.* (2014) and Bitto and Frühwirth-Schnatter (2019) consider a continuous shrinkage prior to force some regression coefficients to be very close to zero for the whole time series. The main limitation of this approach is the assumption that variable's importance is fixed over time, while it is often reasonable to assume that the subset of relevant predictors can change over time. For example, in economics, we can expect that variables have different impact according to the business cycle. In this direction, Kalli and Griffin (2014) proposed a normal-gamma autoregressive process to dynamically shrink towards zero unimportant coefficients. The latter approach falls into the class of dynamic shrinkage processes studied in Kowal *et al.* (2019). Other methods aim to perform model selection rather than shrinkage. Koop and Korobilis (2012), for example, leverage the dynamic model averaging (DMA) method of Raftery *et al.* (2010) to dynamically select a suitable model by choosing at each time the best among the all 2^p possible. It is immediate to see that for moderate p (for example $p > 10$) this approach may be computationally infeasible. More recently, Ročková and McAlinn (2021) and Koop and Korobilis (2020) proposed a dynamic variable selection method that leverages the spike-and-slab type prior of Mitchell and Beauchamp (1988); George and McCulloch (1997); Ishwaran and Rao (2005). In the first aforementioned paper, the authors present the dynamic spike-and-slab process (DSS) and assume a smooth and deterministic trajectory for the dynamic of coefficients' inclusion probabilities, given the past history of the process. In their approach the degree of sparseness is governed by a fixed marginal importance weight which controls for the overall balance between the spike and the slab components of the mixture prior. Instead, Koop and Korobilis (2020) propose a variational Bayes approach for dynamic variable selection (VBDVS), thereby undertaking the problem of variable selection using a similar mixture prior specification while assuming stochastic and independent a-priori inclusion probabilities. Although these methods are intriguing, their main issue lies in the fact that they require a sensible hyper-parameters tuning. In fact, small changes in such parameters produces completely different variables' subset selection.

In this Chapter, we extend the Bernoulli-Gaussian model, previously considered by Ormerod *et al.* (2017) for the linear regression model, to deal with dynamic variable selection. The model specification considers a stochastic process that governs the evolution of the time-varying regression coefficients and possibly for the conditional variance, to account for heteroskedasticity. To increase the flexibility of the model, we further assume an autoregressive process for the a-priori inclusion probability. Unlike existing methods, an important feature of the Bernoulli-Gaussian specification is that it only requires minimal hyperparameter tuning. The inference is carried out within a variational Bayes paradigm which exploits a non-parametric mean-field approach as well as two parametric approximations. Leveraging on the theory developed by Ormerod *et al.* (2017), we further provide a similar results for the time-varying framework here introduced. Specifically, we ensure that the model performs an effective variable selection, therefore achieving sparse estimates. Then, we show that under certain conditions it is possible to reduce the set of covariates directly in real time, while the algorithm is running, leading to a fast procedure able to efficiently deal with large set of variables. The latter results are achieved by inspecting the behavior of the variational updates. As concern the results, we show through and extensive simulation study that our method has excellent performances in separating the true signal from the noise when compared with state-of-the art approaches, while gaining in computational efficiency. We also highlight the main differences that emerge when estimating the underlying model through variational Bayes versus the benchmark true posterior as approximated by MCMC methods. Finally, we show the effectiveness of our model in providing reliable predictions within a real data application in inflation forecasting.

4.2 Model specification and inference

The Bernoulli-Gaussian specification presented within a variational Bayes framework by Ormerod *et al.* (2017) for a static liner regression model represents an interesting approach to variable selection. The time-varying parameter regression model with dynamic variable selection can be seen as a generalization of the aforementioned model specification, and reads as follows:

$$y_t = \mathbf{x}_t^\intercal \tilde{\boldsymbol{\beta}}_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathsf{N}(0, \sigma_t^2), \qquad t = 1, \dots, n, \tag{4.1}$$

where y_t is the response, $\mathbf{x}_t \in \mathbb{R}^p$ is a set of covariates associated to the *sparse* time-varying parameter $\tilde{\boldsymbol{\beta}}_t \in \mathbb{R}^p$. The latter is defined as a product of two components $\tilde{\boldsymbol{\beta}}_t = \boldsymbol{\Gamma}_t \boldsymbol{\beta}_t$, where $\boldsymbol{\Gamma}_t$ is a diagonal matrix of indicator variables with non-zero entries $\gamma_{j,t} \in \{0, 1\}$ and $\boldsymbol{\beta}_t$ is the regression coefficient at time t, and $\sigma_t^2 > 0$ is the unknown time dependent variance of the residuals. The dynamic variable selection comes from the fact that each coefficient $\beta_{j,t} \in \boldsymbol{\beta}_t$, for $j = 1, \dots, p$, can be either included or not in the model depending on the value of the diagonal elements $\gamma_{j,t} \in \{0, 1\}$. The idea behind this model specification is to introduce a stochastic process for both the time evolution of the coefficients and the dynamics of the sparsity over time. Indeed, the assumption of a persistent stochastic process for the inclusion probabilities, i.e., $\mathbb{P}(\gamma_{j,t} = 1)$, represents the main novelty with respect to state-of-the-art techniques. In fact, the latter either assumes independence over time (see Koop and Korobilis, 2020) or defines a deterministic evolution of $\mathbb{P}(\gamma_{j,t} = 1)$, given the information up to $t-1$ (as in Ročková and McAlinn, 2021). To account for a time dependence structure, we assume a random walk specification for the dynamic of $\beta_{j,t}$, i.e.,

$$\beta_{j,t} = \beta_{j,t-1} + v_{j,t}, \qquad v_{j,t} \sim \mathsf{N}(0, \eta_j^2), \qquad j = 1, \dots, p, \tag{4.2}$$

and $\beta_{j,0} \sim \mathsf{N}(0, k_0\eta_j^2)$ defines the initial state. Notice that the latter formulation is equivalent to consider a Gaussian Markov random field (GMRF) for the joint distribution $\boldsymbol{\beta}_j$ as stated in the following Remark 4.1.

Remark 4.1. The joint distribution of $\boldsymbol{\beta}_j = (\beta_{j,0}, \beta_{j,1}, \dots, \beta_{j,n})^\intercal$ can be written leveraging the first-order Markov property $p(\boldsymbol{\beta}_j) = p(\beta_{j,0})p(\beta_{j,1}|\beta_{j,0})\dots p(\beta_{j,n}|\beta_{j,n-1})$, for $j = 1, \dots, p$. Then, following Rue and Held (2005), the joint vector $\boldsymbol{\beta}$ admits a Gaussian Markov random field (GMRF) representation of order one $\boldsymbol{\beta}_j \sim \mathsf{N}_{n+1}(\mathbf{0}, \eta_j^2\mathbf{Q}^{-1})$ with tridiagonal precision matrix $\mathbf{Q}$ with diagonal elements $q_{1,1} = 1 + 1/k_0$, $q_{n+1,n+1} = 1$, and $q_{l,l} = 2$ for $l = 2, \dots, n$. The off diagonal elements are $q_{l,m} = -1$ if $|l - m| = 1$ and 0 elsewhere.

The same dynamic applies to the logarithm of the variance $h_t = \log \sigma_t^2$ to account for possible heteroskedastic nature of the data $\mathbf{h} \sim \mathsf{N}_{n+1}(\mathbf{0}, \nu^2\mathbf{Q}^{-1})$. The homoskedastic alternative can be obtained by setting $\sigma^2 \sim \mathsf{IGa}(A_\sigma, B_\sigma)$. In what follows we detail how we specify the stochastic process that governs the dynamic sparsity of the regression parameters. Specifically, the indicator variables $\gamma_{j,t}$ are assumed to be independent Bernoulli $\gamma_{j,t}|\omega_{j,t} \sim \mathsf{Bern}(p_{j,t})$ given the auxiliary parameters $\omega_{j,t}$, where $\omega_{j,t} = \text{logit}(p_{j,t})$, for $j = 1, \dots, p$. Then, we assume a GMRF specification for the joint vector $\boldsymbol{\omega}_j = (\omega_{j,0}, \omega_{j,1}, \dots, \omega_{j,n})^\intercal \sim \mathsf{N}_{n+1}(\mathbf{0}, \xi_j^2\mathbf{Q}^{-1})$. Under this setting, the marginal distribution for the vector $\boldsymbol{\gamma}_j = (\gamma_{j,1}, \dots, \gamma_{j,n})^\intercal$ is retrieved by integrating out $\boldsymbol{\omega}_j$, as follows:

$$p(\gamma_{j,1}, \dots, \gamma_{j,n}) = \int p(\boldsymbol{\omega}_j) \prod_{t=1}^{n} p(\gamma_{j,t}|\omega_{j,t})\, d\boldsymbol{\omega}_j, \tag{4.3}$$

and it has correlated components.

To complete the Bayesian model specification, we place inverse-gamma priors for the variances parameters $\nu^2 \sim \mathsf{IGa}(A_\nu, B_\nu)$, $\eta_j^2 \sim \mathsf{IGa}(A_\eta, B_\eta)$, and $\xi_j^2 \sim \mathsf{IGa}(A_\xi, B_\xi)$, which represents a common choice in Bayesian analysis.

Approximate Bayesian inference for the parameters of the time-varying Bernoulli-Gaussian model is carried out by the implementation of a semi-parametric variational Bayes algorithm (see, e.g., Wand, 2014). Let $\boldsymbol{\vartheta} = (\mathbf{h}^\intercal, \boldsymbol{\beta}^\intercal, \boldsymbol{\gamma}^\intercal, \boldsymbol{\omega}^\intercal, \nu^2, \boldsymbol{\eta}^{2\intercal}, \boldsymbol{\xi}^{2\intercal})^\intercal$ be the collection of latent variables and model's parameters. Recall that the aim of variational

Bayes (VB) is to find the distribution $q(\boldsymbol{\vartheta}) \in \mathcal{Q}$ that better approximates the posterior $p(\boldsymbol{\vartheta}|\mathbf{y})$ in terms of Kullback-Leibler (Kullback and Leibler, 1951) divergence measure. The choice of the family of distributions over which the optimization problem has to be solved, namely $\mathcal{Q}$, is particularly important, and different choices lead to different approaches. In this Chapter, we consider two main assumptions on $\mathcal{Q}$. First, we propose a convenient mean–field factorisation for the joint variational density. Second, for some components we impose a parametric approximation. The key ingredient for getting the optimal variational densities is the joint distribution of the data, the latent processes, and the parameters $p(\mathbf{y}, \boldsymbol{\vartheta})$ which can be factorized as follows:

$$p(\mathbf{y}, \boldsymbol{\vartheta}) = p(\mathbf{y}|\boldsymbol{\vartheta})p(\mathbf{h})p(\nu^2)\prod_{j=1}^{p} p(\boldsymbol{\beta}_j|\eta_j^2)p(\boldsymbol{\gamma}_j|\boldsymbol{\omega}_j)p(\boldsymbol{\omega}_j|\xi_j^2)p(\eta_j^2)p(\xi_j^2), \tag{4.4}$$

where $p(\boldsymbol{\gamma}_j|\boldsymbol{\omega}_j) = \prod_{t=1}^{n} p(\gamma_{j,t}|\omega_{j,t})$ also factorizes over time. Observe that the full conditional distribution of $\boldsymbol{\omega}_j$, i.e., $p(\boldsymbol{\omega}_j|\text{rest})$, is not recognized as a know distribution and therefore it complicates the computations. Following Polson *et al.* (2013), we exploit the Polya-Gamma representation of $\gamma_{j,t}$:

$$p(\gamma_{j,t}|\omega_{j,t}) = \int_0^{+\infty} p(\gamma_{j,t}|z_{j,t}, \omega_{j,t})p(z_{j,t}|\omega_{j,t})\, dz_{j,t}, \tag{4.5}$$

where $p(z_{j,t})$ is the probability density function of a Polya-Gamma $\mathsf{PG}(1, 0)$ random variable. Leveraging the stochastic representation in (4.5), the augmented version of (4.4) has the advantage of being more tractable. As concerns the first assumption on the variational family $\mathcal{Q}$, i.e., the mean–field factorisation, we propose:

$$q(\boldsymbol{\vartheta}) = q(\mathbf{h})q(\nu^2)\prod_{j=1}^{p} q(\boldsymbol{\beta}_j)q(\boldsymbol{\omega}_j)q(\eta_j^2)q(\xi_j^2)\prod_{t=1}^{n} q(\gamma_{j,t})q(z_{j,t}), \tag{4.6}$$

where a joint distribution for $\mathbf{h}$, $\boldsymbol{\beta}_j$, and $\boldsymbol{\omega}_j$ is required in order to preserve the time dependence and to provide a global approximation for the vector of latent states. The following propositions present the main optimal variational densities, namely $q^*(\boldsymbol{\beta}_j)$ and $q^*(\gamma_{j,t})$. The proofs and the analytical derivation of the remaining optimal densities, are available in Appendix C.1.

Proposition 4.1. *The optimal variational density for the regression parameters is a multivariate Gaussian* $q^*(\boldsymbol{\beta}_j) \equiv \mathsf{N}_{n+1}(\boldsymbol{\mu}_{q(\beta_j)}, \boldsymbol{\Sigma}_{q(\beta_j)})$, *where:*

$$\boldsymbol{\Sigma}_{q(\beta_j)} = (\mathbf{D}_j^{\,2} + \mu_{q(1/\eta_j^2)}\mathbf{Q})^{-1}, \qquad \boldsymbol{\mu}_{q(\beta_j)} = \boldsymbol{\Sigma}_{q(\beta_j)}\mathbf{D}_j\boldsymbol{\mu}_{q(\varepsilon_{-j})}, \tag{4.7}$$

where $\mathbf{D}_j$ *and* $\mathbf{D}_j^2$ *are diagonal matrices with elements* $[\mathbf{D}_j]_t = \mu_{q(1/\sigma_t^2)}\mu_{q(\gamma_{j,t})}x_{j,t}$ *and* $[\mathbf{D}_j]_t^2 = \mu_{q(1/\sigma_t^2)}\mu_{q(\gamma_{j,t})}x_{j,t}^2$*, respectively. Moreover,* $\boldsymbol{\mu}_{q(\varepsilon_{-j})}$ *is the vector of partial residuals with elements* $\mu_{q(\varepsilon_{-j,t})} = y_t - \sum_{k=1,k\neq j}^{p} x_{k,t}\mu_{q(\gamma_{k,t})}\mu_{q(\beta_{k,t})}$.

Proof. See proof C.3 in Appendix C.1. □

Proposition 4.2. *The optimal variational density for the parameters* $\gamma_{j,t}$ *is a Bernoulli random variable* $q^*(\gamma_{j,t}) \equiv \mathsf{Bern}(\text{expit}(\omega_{q(\gamma_{j,t})}))$*, where* $\text{expit}(\cdot)$ *is the inverse of the logit function and* $\omega_{q(\gamma_{j,t})} = \mu_{q(\omega_{j,t})} - \frac{1}{2}\mu_{q(1/\sigma_t^2)}(x_{j,t}^2\mathbb{E}_q[\beta_{j,t}^2] - 2\mu_{q(\beta_{j,t})}x_{j,t}\mu_{q(\varepsilon_{-j,t})})$.

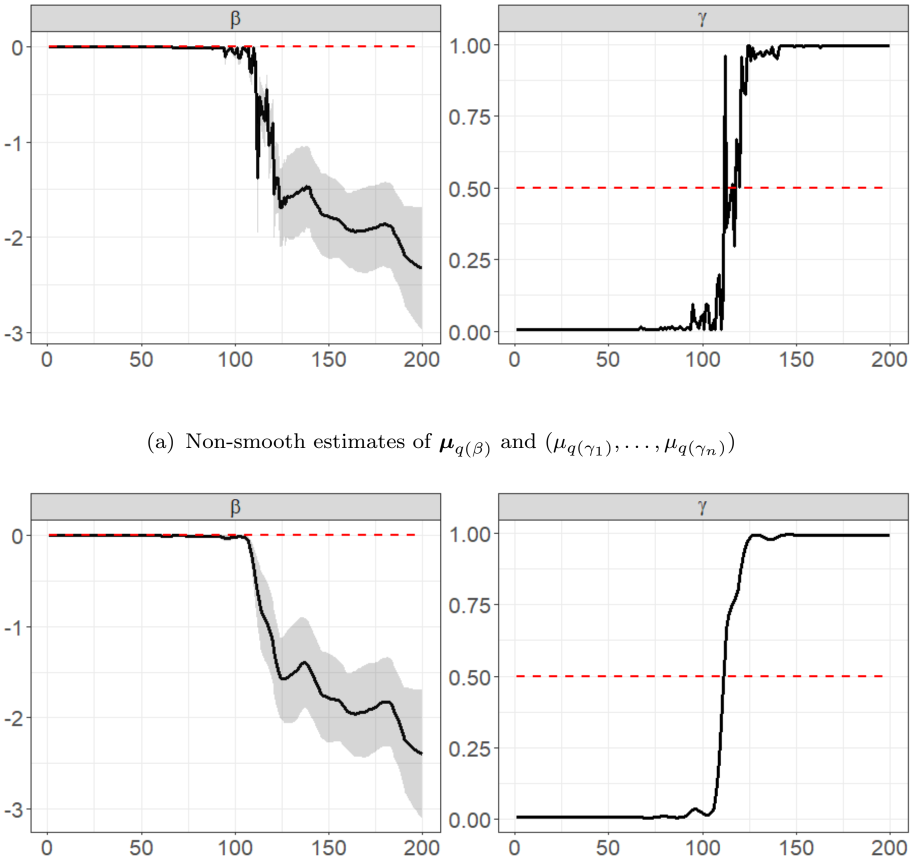
Proof. See proof C.4 in Appendix C.1. □

Beside the mean–field factorisation assumption, we also implement two parametric approximations. The first one imposes a multivariate Gaussian distribution with variational parameters $(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)})$ on $q(\mathbf{h})$. Proposition C.1 in Appendix C.1 provides details on how to find the optimum couple $(\hat{\boldsymbol{\mu}}_{q(h)}, \hat{\boldsymbol{\Sigma}}_{q(h)})$. The second parametric approximation concerns the optimal density of the indicator variables $\gamma_{j,t}$, and it has a practical motivation. In fact, as proved in Proposition 4.2, the optimal variational density $q(\boldsymbol{\gamma}_j) = \prod_{t=1}^n q(\gamma_{j,t})$ is such that each component $q(\gamma_{j,t})$ is $\mathsf{Bern}(\text{expit}(\omega_{q(\gamma_{j,t})}))$ and therefore the whole trajectory over time of the posterior inclusion probabilities can be obtained as the mean vector $\mathbb{E}_q(\boldsymbol{\gamma}_j) = \text{expit}(\boldsymbol{\omega}_{q(\gamma_j)})$. We notice that the latter may not provide a smooth pattern since it is completely data driven, and therefore the variability of the data can cause undesired peaks as shown in Figure 4.1(a). This aspect also impacts on the estimates of the regression coefficient. To this aim, we consider a parametric approximation to force the posterior estimates to follow a smooth trajectory. In particular, we approximate the sequence of densities $\{q(\gamma_{j,t})\}_{t=1}^n$ with the closest sequence $\{\tilde{q}(\gamma_{j,t})\}_{t=1}^n$ in terms of Kullback-Leibler ($\mathcal{KL}$) divergence, such that $\{\tilde{q}(\gamma_{j,t})\}_{t=1}^n$ leads to smooth sequence of posterior inclusion probabilities, which coincides with their expected values. The following Proposition explains this procedure.

Proposition 4.3. *A smooth estimate for the trajectory of the inclusion probabilities can be achieved assuming* $\tilde{q}(\boldsymbol{\gamma}_j) = \prod_{t=1}^n \tilde{q}(\gamma_{j,t})$ *such that* $\tilde{q}(\gamma_{j,t})$ *is* $\mathsf{Bern}(\pi_{j,t})$ *with constrained mean* $\text{logit}(\pi_{j,t}) = \mathbf{w}_t^\intercal\mathbf{f}$*. Therefore,* $\mathbb{E}_{\tilde{q}}(\boldsymbol{\gamma}_j) = \boldsymbol{\pi}_j$ *and* $\text{logit}(\boldsymbol{\pi}_j) = \mathbf{W}\mathbf{f}$*, where* $\mathbf{W}$ *is a* $n\times k$ *B-spline basis matrix. The optimal value of* $\mathbf{f}$ *is the solution of* $\hat{\mathbf{f}} = \arg\max_{\mathbf{f}\in\mathbb{R}^k}\psi(\mathbf{f})$ *where* $\psi(\mathbf{f}) = \sum_{t=1}^n \left[(\omega_{q(\gamma_{j,t})} - \mathbf{w}_t^\intercal\mathbf{f})\text{expit}(\mathbf{w}_t^\intercal\mathbf{f}) + \log(1+\exp(\mathbf{w}_t^\intercal\mathbf{f}))\right]$.

Proof. See proof C.5 in Appendix C.1. □

The implementation of the parametric approximation presented above, leads to better estimates for the regression coefficients and the corresponding inclusion probabilities, and, as a consequence, provides more interpretable results (see Figure 4.1(b)).



(a) Non-smooth estimates of $\boldsymbol{\mu}_{q(\beta)}$ and $(\mu_{q(\gamma_1)}, \dots, \mu_{q(\gamma_n)})$

(b) Smooth estimates of $\boldsymbol{\mu}_{q(\beta)}$ and $(\mu_{q(\gamma_1)}, \dots, \mu_{q(\gamma_n)})$

FIGURE 4.1: Smoothing procedure using parametric variational Bayes.

We presented the derivation of $q(\boldsymbol{\beta}_j)$ and $q(\gamma_{j,t})$ separately, but we remind that our main interest is to provide a distribution for the product-parameter $\tilde{\boldsymbol{\beta}}_j = \boldsymbol{\Gamma}_j \boldsymbol{\beta}_j$. The optimal density is provided in Proposition 4.4.

Proposition 4.4. *Let $q^*(\boldsymbol{\beta}_j)$ and $q^*(\gamma_{j,t})$ be the optimal variational densities presented in Propositions 4.1 and 4.2 (or its smoothed alternative). Define $\tilde{\boldsymbol{\beta}}_j = \boldsymbol{\Gamma}_j \boldsymbol{\beta}_j$, where the matrix $\boldsymbol{\Gamma}_j = \mathsf{Diag}(1, \gamma_{j,1}, \dots, \gamma_{j,n})$ is diagonal. The optimal variational density of $\tilde{\boldsymbol{\beta}}_j$ is given by a mixture of multivariate Gaussian distributions:*

$$q^*(\tilde{\boldsymbol{\beta}}_j) = \sum_{s \in \mathcal{S}} w_s \, \mathsf{N}_{n+1}(\mathbf{D}_s \boldsymbol{\mu}_{q(\beta_j)}, \mathbf{D}_s^{1/2} \boldsymbol{\Sigma}_{q(\beta_j)} \mathbf{D}_s^{1/2}), \tag{4.8}$$

where $\mathcal{S} = \{$sequences of $\{0,1\}$of length $n\}$ with cardinality $|\mathcal{S}| = 2^n$, the diagonal matrix $\mathbf{D}_s = \mathsf{Diag}(1, s_1, \dots, s_n)$, and mixing weights:

$$w_s = \prod_{t=1}^{n} \mu_{q(\gamma_{j,t})}^{s_t} (1 - \mu_{q(\gamma_{j,t})})^{1-s_t}, \tag{4.9}$$

where $\mathbf{s} = (s_1, \dots, s_t, \dots, s_n) \in \mathcal{S}$ *is an element in* $\mathcal{S}$*. Moreover, the mean and variance can be computed analytically:*

$$\boldsymbol{\mu}_{q(\tilde{\beta}_j)} = \boldsymbol{\mu}_{q(\Gamma_j)}\boldsymbol{\mu}_{q(\beta_j)}, \tag{4.10}$$

$$\boldsymbol{\Sigma}_{q(\tilde{\beta}_j)} = (\boldsymbol{\mu}_{q(\gamma_j)}\boldsymbol{\mu}^{\intercal}_{q(\gamma_j)} + \mathbf{W}_{\mu_{q(\gamma_j)}}) \odot \boldsymbol{\Sigma}_{q(\beta_j)} + \mathbf{W}_{\mu_{q(\gamma_j)}} \odot \boldsymbol{\mu}_{q(\beta_j)}\boldsymbol{\mu}^{\intercal}_{q(\beta_j)}, \tag{4.11}$$

where $\mathbf{W}_{\mu_{q(\gamma_j)}}$ *is a diagonal matrix with elements* $(1, \{\mu_{q(\gamma_{j,t})}(1 - \mu_{q(\gamma_{j,t})})\}_{t=1}^{n})$.

Proof. See proof C.6 in Appendix C.1. □

4.3 Properties of the algorithm

In this Section we present some theoretical results and properties of the proposed algorithm. In particular, we focus on the behavior of variational densities' updates from one iteration to the next as the inclusion probabilities tend to zero. Specifically, in Proposition 4.5 we extend the main result of Ormerod *et al.* (2017) to the dynamic variable selection with time-varying coefficients scenario here considered. Some of the definitions and lemmas useful to understand the following propositions, together with the corresponding proofs are postponed to Appendix C.2, for the sake of clarity.

Proposition 4.5. *Assume that the maximum over time of the inclusion probabilities, for a given variable* j*, at the* i*-th iteration of the algorithm is such that* $\max_{t\in\{1,\dots,n\}} \mu^{(i)}_{q(\gamma_{j,t})} = \epsilon$*, and* $\epsilon \ll 1$ *is small enough. Moreover, let* $\boldsymbol{\Sigma}^{(i)}_{q(\omega_j)} - \boldsymbol{\Sigma}^{(i-1)}_{q(\omega_j)}$ *be a positive matrix, which is* $\boldsymbol{\Sigma}^{(i)}_{q(\omega_j)} - \boldsymbol{\Sigma}^{(i-1)}_{q(\omega_j)} \geq 0$*, then:*

1. $\mu^{(i+1)}_{q(\gamma_{j,t})} = \operatorname{expit}\left\{\mu^{(i+1)}_{q(\omega_{j,t})} - \frac{1}{2}\mu^{(i+1)}_{q(1/\sigma^2_t)}x^2_{j,t}\mu^{-1(i+1)}_{q(1/\eta^2_j)}q_{t,t} + O(\epsilon)\right\}$*,* $q_{t,t} = [\mathbf{Q}^{-1}]_{t,t} \geq 0$*;*
2. $\mu^{(i+1)}_{q(\omega_{j,t})} = -1/2\sum_{k=1}^{n} s_{t,k} + O(\epsilon)$*,* $s_{t,k} = [\boldsymbol{\Sigma}_{q(\omega_j)}]_{t,k} \geq 0$*;*
3. $\mu^{(i+1)}_{q(\omega_{j,t})} \leq \mu^{(i)}_{q(\omega_{j,t})}$ *decreases after each iteration.*

Proof. See proof C.12 in Appendix C.2. □

The following lemma is stated in Ormerod *et al.* (2017) and provides expansion results for the expit$(\cdot)$ function.

Lemma 4.1. Let $a \in \mathbb{R}^+$, then, as $a \to +\infty$, the following expansions hold: $\operatorname{expit}(-a) = \exp(-a) + O(\exp(-2a))$ and $\operatorname{expit}(a) = 1 - \exp(-a) + O(\exp(-2a))$.

Leveraging Proposition 4.5 and Lemma 4.1, we can define the following two properties for the proposed semi–parametric variational Bayes algorithm.

Result 4.1. *(Sparsity).* For $\epsilon \ll 1$ sufficiently small, we obtain the following approximation for the update of the inclusion probabilities:

$$\mu_{q(\gamma_{j,t})}^{(i)} \approx \text{expit}\left\{\mu_{q(\omega_{j,t})}^{(i+1)} - 1/2\mu_{q(1/\sigma_t^2)}^{(i+1)} x_{j,t}^2 \left[\mu_{q(1/\eta_j^2)}^{(i+1)}\right]^{-1} q_{t,t}\right\}. \tag{4.12}$$

Moreover, when $\mathrm{M}^{(i)} = \arg\max_{t\in\{1,\dots,n\}} \mu_{q(\omega_{j,t})}^{(i)}$ is small enough, i.e., $\mathrm{M}^{(i)} \ll 0$, after i iterations, the sequence $\{\mu_{q(\gamma_{j,t})}^{(i)}\}_{t=1}^{n}$ is entirely represented as 0 when implemented on a computer.

Result 4.1 shows that, similarly to the static framework discussed in Ormerod *et al.* (2017), this algorithm is able to achieve exact sparsity, i.e., exactly zero inclusion probabilities.

Result 4.2. *(Dimension reduction).* If $\mu_{q(\gamma_{j,t})}^{(i)} \approx 0$, for all t, then all the successive updates $i_k \geq i$ remains numerically $\mu_{q(\gamma_{j,t})}^{(i_k)} \approx 0$ since $\mu_{q(\omega_{j,t})}^{(i_k+1)} \leq \mu_{q(\omega_{j,t})}^{(i_k)}$ and therefore $\mathrm{M}^{(i_k+1)} \leq \mathrm{M}^{(i_k)}$. As a consequence, we can remove the j-th variable from the set of covariates within the algorithm. This procedure reduces the computational cost of the successive iterations.

Result 4.2 provides a real-time dimension reduction strategy embedded into the estimation procedure. The latter is of interest since it ensures that the proposed algorithm is computationally efficient when the dimension of regression coefficient p increases, but the signal $\bar{p} \leq p$ remains constant, where $\bar{p} = \mathsf{card}(\mathcal{J})$ and the set $\mathcal{J} = \{j : \sum_{t=1}^{n} \gamma_{j,t} > 0\}$ collects the indexes of regression coefficients that are included in the model at least for one t.

An efficient iterative algorithm to perform approximate posterior inference within this paradigm is presented in Algorithm 4.1 and the convergence is assessed looking at the variation in the optimal densities updates between two consecutive iterations $q^*(\boldsymbol{\vartheta})^{(\mathsf{iter})} - q^*(\boldsymbol{\vartheta})^{(\mathsf{iter}-1)} < \Delta$.

4.4 Simulation study

In this Section we focus on performances in a controlled simulation-based setting. We divide the discussion into two studies. The first one compares the posterior distributions estimated via the Gibbs sampling algorithm and the variational densities obtained through Algorithm 4.1. The second simulation study serves the purpose of evaluating the proposed method with respect to established state-of-the-art alternative approaches.

Algorithm 4.1: Semi–parametric variational Bayes for time-varying Bernoulli-Gaussian model with dynamic sparsity.

```
Initialize: $q(\boldsymbol{\vartheta})$, $\Delta_{\boldsymbol{\vartheta}}$, $\epsilon$
while $(\widehat{\Delta}_{\boldsymbol{\vartheta}} > \Delta_{\boldsymbol{\vartheta}})$ do
    for $j = 1, \dots, p$ do
        Update $q(\boldsymbol{\beta}_j)$ as in 4.1; and $q(\eta_j)$ as in C.9;
        Update $q(\boldsymbol{\omega}_j)$ as in C.7 and $q(\xi_j)$ as in C.10;
        for $t = 1, \dots, n$ do
            Update $q(z_{j,t})$ as in C.8;
            Update $q(\gamma_{j,t})$ as in 4.2 (non-smooth) or 4.3 (smooth);
        end
    end
    Update $q(\boldsymbol{\sigma})$ as in C.1 (heteroskedastic) or C.2 (homoskedastic);
    Update $q(\nu^2)$ as in C.11;
    if assumptions in 4.5 hold then
        for $j = 1, \dots, p$ do
            if $\max_t\{\mu_{q(\gamma_{j,t})}\} < \epsilon$ then
                Drop the $j$-th variable
            end
        end
    end
    Compute $\widehat{\Delta}_{\boldsymbol{\vartheta}} = q(\boldsymbol{\vartheta})^{(\mathsf{iter})} - q(\boldsymbol{\vartheta})^{(\mathsf{iter}-1)}$ ;
end
```

4.4.1 Comparison with Markov chain Monte Carlo

The evaluation of the variational Bayes approximation compared to MCMC is often a challenging task to undertake. As mentioned in Section 4.2, the data augmentation approach based on the Polya-Gamma representation has the main advantage to lead to a more tractable joint distribution $p(\mathbf{y}, \boldsymbol{\vartheta})$. The latter results in full conditional distributions that are recognized as known density functions. This aspect is crucial to implement an efficient Gibbs sampling scheme to carry out Bayesian inference. In this Section, we compare the posterior distribution for the parameters of the underlying model achieved via variational Bayes and MCMC. To this aim, we use the accuracy measure (1.4) introduced in Chapter 1, where q is the variational density and p denotes the posterior distribution sampled via MCMC. In particular, we retain 20000 draws from the posterior and we discard the first 10000 as a burn-in.

The simulation is set-up as follows. We consider $p = 3$ and $n = 100$, and generate $\{\beta_{1,t}, \beta_{2,t}, \beta_{3,t}\}_{t=1}^{100}$ such that $\beta_{1,t}$ is a time-varying intercept always included in the model, $\beta_{2,t}$ is set equal to zero for all the timeline, while $\beta_{3,t}$ shows a dynamic sparsity pattern. Note that for the scope of this study the small dimension p has a limited impact on

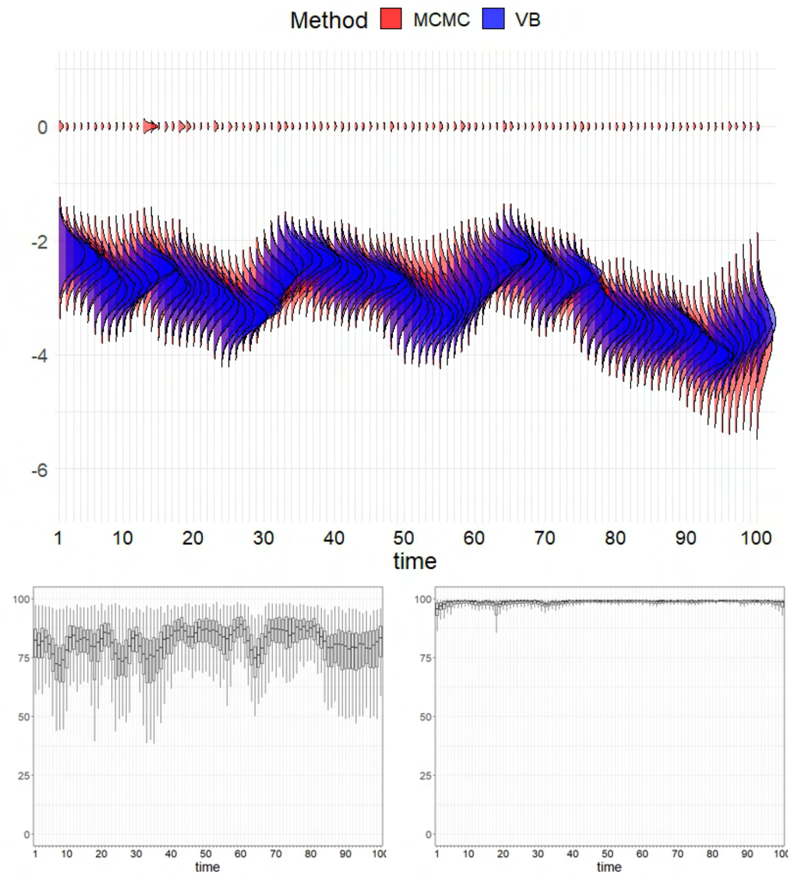


FIGURE 4.2: Comparison with MCMC when $\beta_{1,t}$ is a time-varying intercept with $\gamma_{1,t} = 1$, for all t. Top panel shows the overlapping posterior densities of $\tilde{\beta}_{1,t}$ obtained via VB (blue) and MCMC (red), for one selected replicate. Bottom panel shows the accuracy over time for $\beta_{1,t}$ (left) and $\gamma_{1,t}$ (right).

the validity of the results and it is chosen for convenience to speeding up the MCMC computation. Then, we generate $N = 100$ replicates from $y_t = x_{1,t}\beta_{1,t} + x_{2,t}\beta_{2,t} + x_{3,t}\beta_{3,t} + \varepsilon_t$, with $\varepsilon_t \sim \mathsf{N}(0, 0.25)$, for $t = 1, \ldots, 100$.

As concerns the results, we first focus on the time-varying intercept, which is always included in the true model. Figure 4.2 highlight a good approximation in this scenario. In fact, the posterior distribution obtained via VB and MCMC for $\tilde{\beta}_{1,j}$ overlap enough. Notice that the posterior inclusion probabilities estimated via VB tend to one. This lead to weights in the mixture defining $q^*(\tilde{\boldsymbol{\beta}}_1)$ (see Proposition 4.4) such that $w_s = 1$ if $s = (1, 1, \ldots, 1)$ and $w_s = 0$ otherwise. Hence, we only keep one component of the mixture. This is not the case for MCMC which shows some sampled values of $\gamma_{1,t}$ to be zero. However, the accuracy measure is satisfactory for $\beta_{1,t}$ (it lies around 80%) and close to 100% for $\gamma_{1,t}$.

Then, we consider the opposite case, that is the coefficient which is always equal to zero. Figure 4.3 highlights that VB provides posterior inclusion probabilities very close

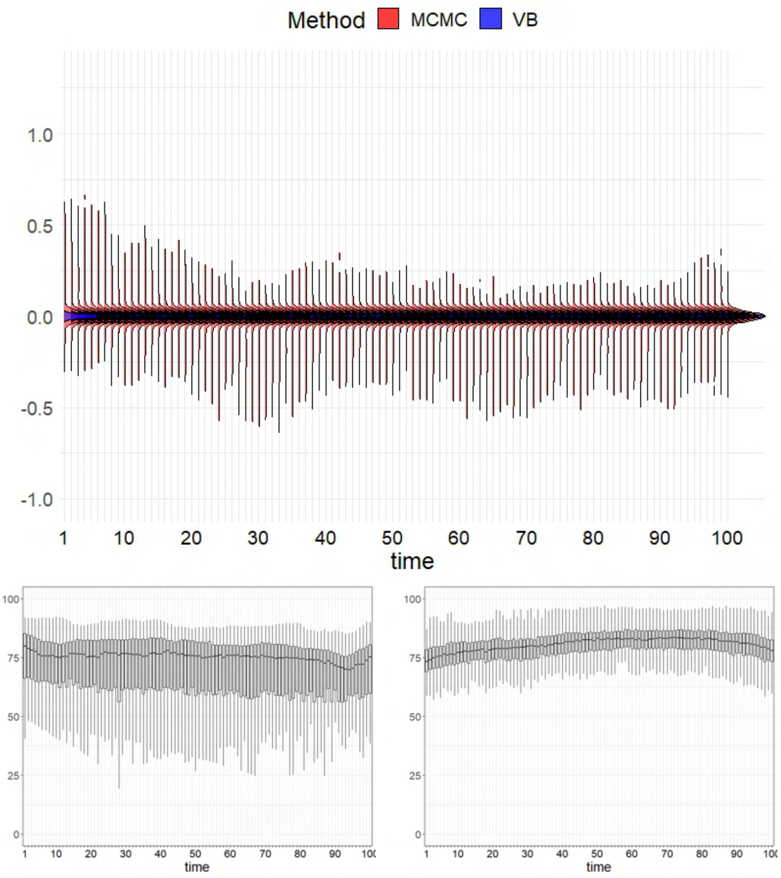


FIGURE 4.3: Comparison with MCMC when $\beta_{3,t}$ is a coefficient constant at zero, i.e., $\gamma_{3,t} = 0$, for all t. Top panel shows the overlapping posterior densities of $\tilde{\beta}_{3,t}$ obtained via VB (blue) and MCMC (red), for one selected replicate. Bottom panel shows the accuracy over time for $\beta_{3,t}$ (left) and $\gamma_{3,t}$ (right).

to zero (recall the exact sparsity property in Result 4.1). This lead to weights in the mixture defining $q^*(\tilde{\boldsymbol{\beta}}_3)$ such that $w_s = 1$ if $s = (0, 0, \dots, 0)$ and $w_s = 0$ otherwise. Hence, we only keep the component of the mixture that identifies a sequence of Dirac at zero $\delta_0(\tilde{\beta}_{3,t})$. This is not the case for MCMC which shows larger variability around zero. Nevertheless, the accuracy measure is around 75% for both $\beta_{3,t}$ and $\gamma_{3,t}$.

The last comment considers the parameter $\beta_{3,t}$, which displays a more interesting dynamic sparsity behavior. Figure 4.4 depicts a good approximation during periods in which the coefficient is included in the model (initial and final part), while the performances deteriorate when $\gamma_{2,t} = 0$ (middle part). The latter aspect is emphasized if we focus on the accuracy measure for $\gamma_{2,t}$ (bottom-right panel). In fact, this scenario resumes the conclusions we carry out for the previous two settings.

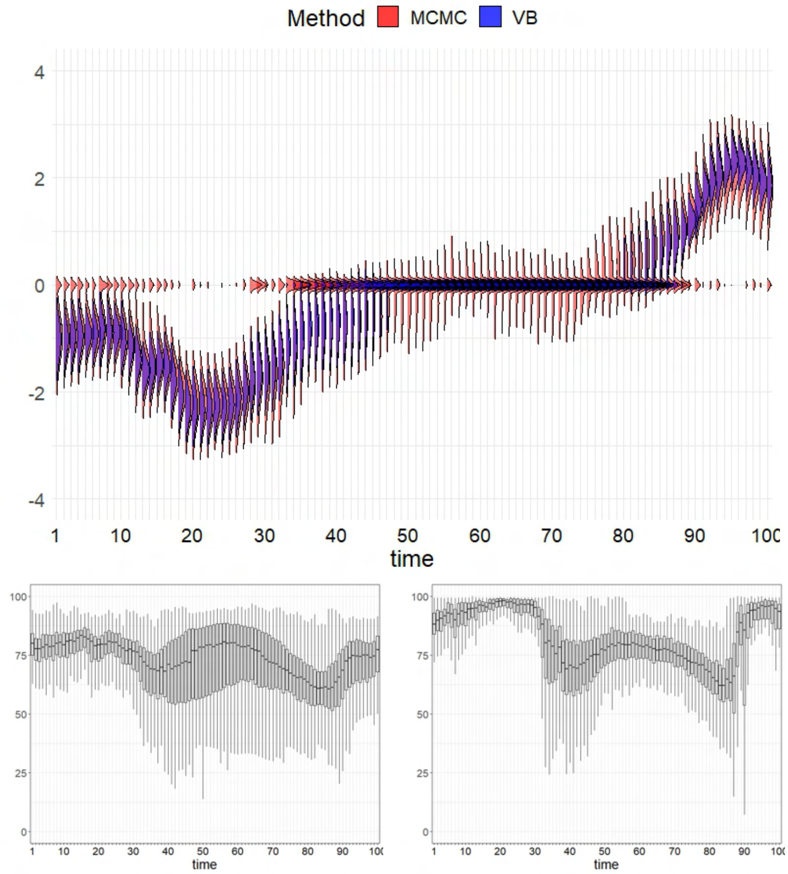


FIGURE 4.4: Comparison with MCMC when $\beta_{2,t}$ is a coefficient that shows dynamic sparsity. Top panel shows the overlapping posterior densities of $\tilde{\beta}_{2,t}$ obtained via VB (blue) and MCMC (red), for one selected replicate. Bottom panel shows the accuracy over time for $\beta_{2,t}$ (left) and $\gamma_{2,t}$ (right).

4.4.2 Comparison with state-of-the-art

In this Section we assess both the selection and estimation accuracy of our approach under different behaviors of the regression coefficients. We consider 100 replicates from the following data generating process

$$y_t = \mathbf{x}_t^\intercal \boldsymbol{\beta}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathsf{N}(0, 0.25), \quad t = 1, \ldots, 200, \tag{4.13}$$

where the entries of $\mathbf{x}_t$ are independently generated from a standard Gaussian distribution. The dimension of the regression parameter $\boldsymbol{\beta}_t$ is equal to $p \in \{50, 100, 200\}$. As before, $\beta_{1,t}$ is a time-varying intercept always included, i.e., $\gamma_{1,t} = 1 \ \forall t$, $\beta_{2:7,t}$ show dynamic sparsity, and $\beta_{8:p,t}$ is set to zero for all t, i.e., $\gamma_{8:p,t} = 0 \ \forall t$. Moreover, we consider different behaviors for the dynamic sparsity in $\beta_{2:7,t}$, which will be discussed later.

We implement different versions of our main algorithm. Hereafter, `BG` denotes the

Bernoulli-Gaussian method accounting for stochastic volatility without smoothness in the inclusion probabilities, BGH is the homoskedastic alternative to BG, and BGS performs smoothing on the posterior inclusion probabilities exploiting the parametric variational approximation. Furthermore, we also consider BG algorithm with fixed variances $\xi_j^2 \in \{0.5, 1\}$ of the latent process $\omega_{j,t}$.

We compare our methods with two recent developments in time-varying regressions with dynamic sparsity. The first one is the dynamic variable selection (DVS) of Koop and Korobilis (2020), while the second matches the dynamic spike-and-slab (DSS) of Ročková and McAlinn (2021) for three different values of the *marginal importance weight* parameter $\Theta \in \{0.1, 0.5, 0.9\}$. Since the literature on dynamic sparsity is not wide, we also compare static variable selection models estimated with recursive rolling windows. The latter is a widely used procedure to mimic a time-varying behavior. Here, we consider two continuous shrinkage priors, i.e., the normal-gamma prior of Griffin and Brown (2010) (BNG) and the horseshoe prior of Carvalho *et al.* (2010) (BHS), and the variable selection method of George and McCulloch (1993) (SSVS) and that of Ročková and George (2014) (EMVS). The estimates shrinked towards zero with BNG and BHS are then sparsified using the signal adaptive variable selector (SAVS) of Ray and Bhattacharya (2018).

We are interested in both point estimation accuracy and signal identification. The first one is assessed by looking at the mean squared error (MSE), while the signal identification is measured via the $F1$-score. As a last, interesting assessment, we compare the methodologies in terms of computational efficiency.

In what follows, we present the simulation results separately for the behavior of the true regression coefficient and we focus on $p = 50, 200$. The results for the case $p = 100$ are reported in Appendix C.3.

Time varying intercept. Figure 4.5 shows an example of simulated time-varying intercept and relative estimates for one selected replicate. Figure 4.6 depicts the results under this first scenario. The F_1-score suggests two main comments. Firstly, all the methods provide good performances (in median) in recognizing a time-varying intercept. As a second fact to mention, DSS seems to be sensible to the choice of the hyperparameter Θ. By looking at the MSE measure, we notice that the dynamics of the coefficient are badly replicated using a rolling window estimation, even if this procedure does not impact on the F_1-score.

Balanced dynamic sparsity with one switch. Figure 4.7 clarifies this behavior. This represents a more intriguing scenario than the previous one. As highlighted in

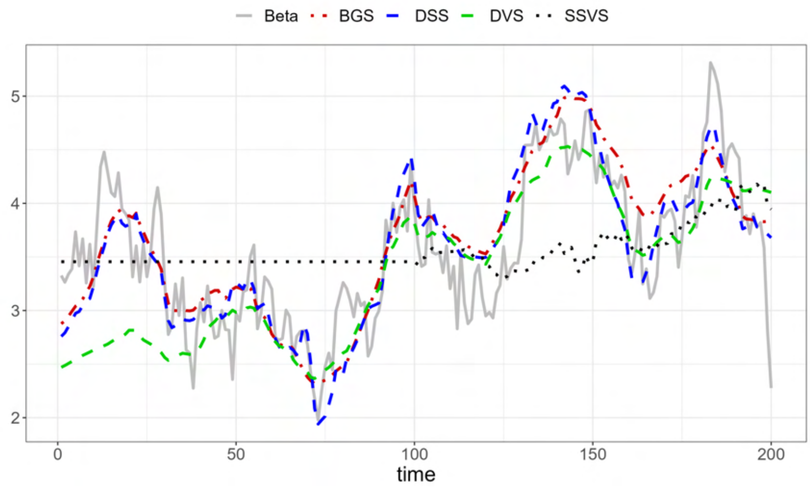


FIGURE 4.5: Example of time-varying intercept $\beta_{1,t}$. It follows an autoregressive process of order one, AR(1) hereafter, with unconditional mean far from zero, persistence $\phi_1 = 0.98$ and conditional variance equal to 0.1.

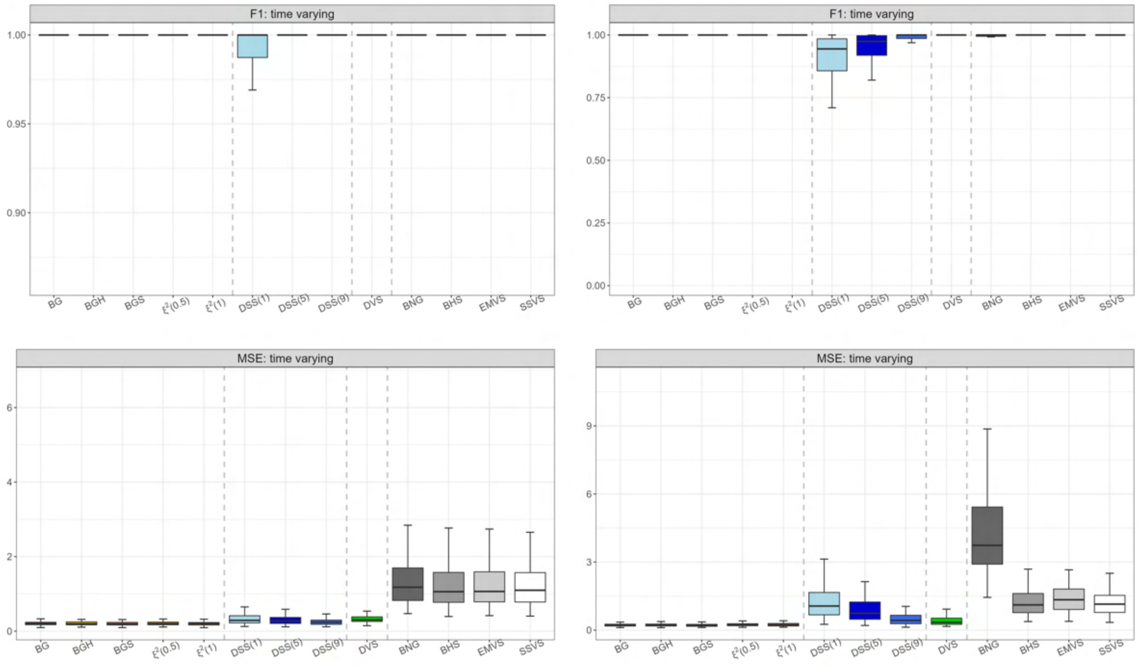


FIGURE 4.6: F_1-score (top panel) and MSE (bottom panel) for the time-varying intercept estimation, when $p = 50$ (left) and $p = 200$ (right).

Figure 4.8, the performances are more heterogeneous. Three comments are in order. First of all, we notice a very similar performance among the different versions of our algorithm. This is a sign of robustness of the dynamic variable selection procedure. Second, the rolling window estimation is not able to properly detect dynamic sparsity, as reflected by the low F_1-score. Third, perhaps more important, we out-perform the state-of-the-art methodologies in dynamically separate the true signal from the noise.

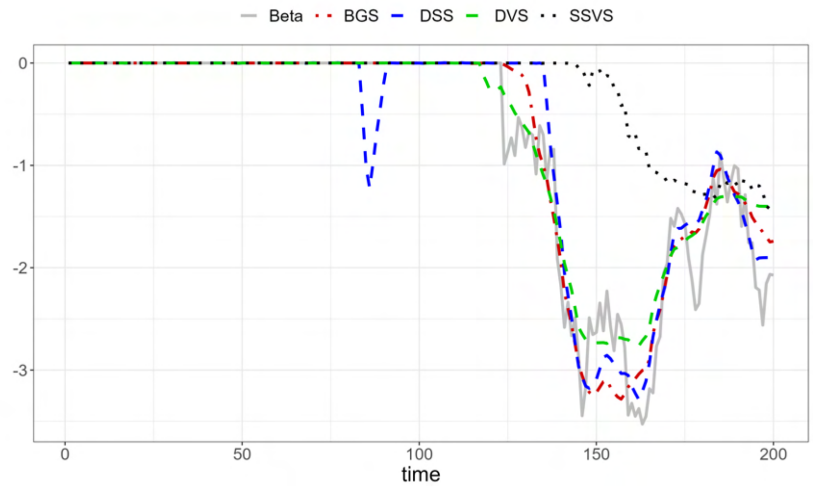


FIGURE 4.7: Example of dynamic sparsity with one switch $\beta_{2,t}$. The parameter is generated as follows. Divide the interval in sub-periods $[1, n] = [1, t_1] \cup [t_1 + 1, t_1 + t_2] \cup ... \cup [t_1 + \ldots + t_n + 1, n]$, where $t_k \sim \mathsf{Pois}(n/2)$, and then alternate periods where $\gamma_{j,t} = 0$ and $\gamma_{j,t} = 1$ starting randomly. For the intervals where $\gamma_{j,t} = 1$ we generate a process as for $\beta_{1,t}$.

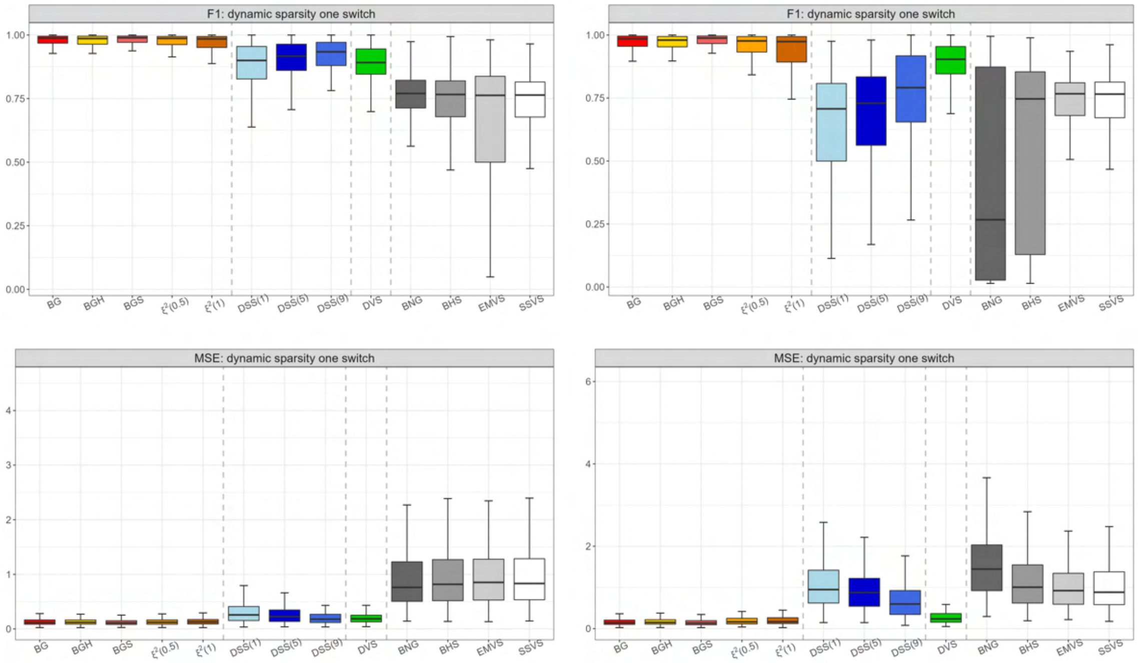


FIGURE 4.8: F_1-score (top panel) and MSE (bottom panel) for the balanced dynamic sparsity setting with one switch, when $p = 50$ (left) and $p = 200$ (right).

As a byproduct, our estimation algorithms also show lower MSE, thus providing more accurate point estimates.

Balanced dynamic sparsity with two switches. Figure 4.9 complicates the previous scenario. This setting represents a more complex scenario to handle. In fact, the

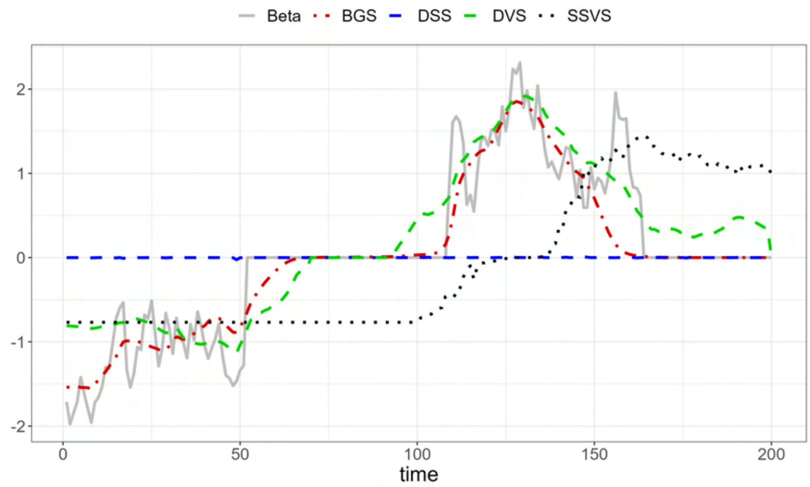


FIGURE 4.9: Example of dynamic sparsity with two switches $\beta_{4,t}$. The parameter is generated as follows. Divide the interval in sub-periods as for $\beta_{2,t}$, but set $t_k \sim \mathsf{Pois}(n/4)$, and then alternate periods where $\gamma_{j,t} = 0$ and $\gamma_{j,t} = 1$ starting randomly. For the intervals where $\gamma_{j,t} = 1$ we generate a process as for $\beta_{1,t}$.

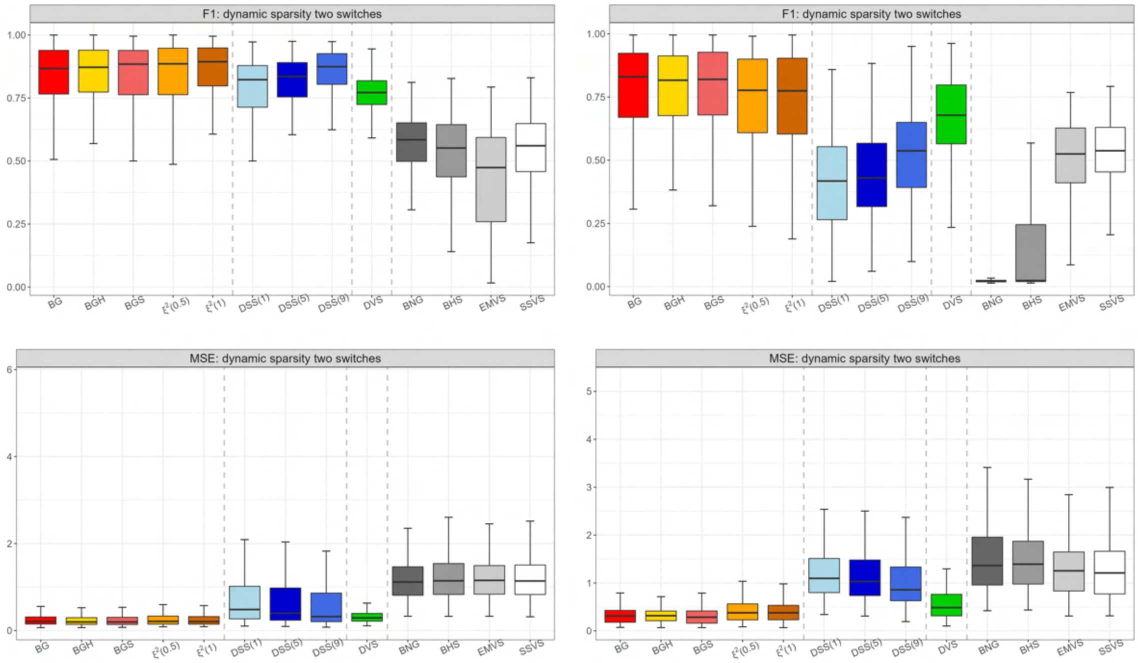


FIGURE 4.10: F_1-score (top panel) and MSE (bottom panel) for the balanced dynamic sparsity setting with two switches, when $p = 50$ (left) and $p = 200$ (right).

F_1-scores in Figure 4.10 are, on median, lower that the previous paragraph. However, in terms of relative comparisons among methods, we can draw the same conclusions.

Low signal dynamic sparsity. Figure 4.11 depicts what we mean with low signal dynamic sparsity. The latter constitutes a very challenging behavior to detect. Indeed, small periods of signal are more complex to extract. Figure 4.12 shows that account for

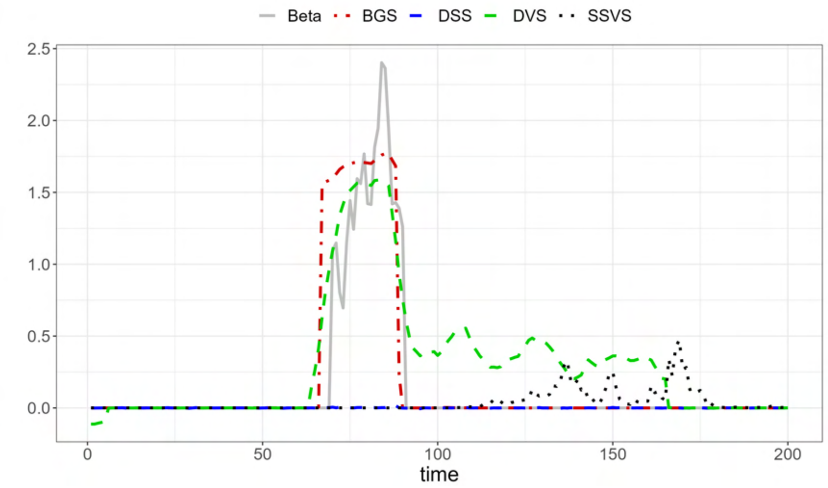


FIGURE 4.11: Example of low signal in dynamic sparsity $\beta_{6,t}$. The parameter is generated as follows. Sample an interval length $\Delta_i \sim \mathsf{Pois}(n/10)$ and place it at random on the timeline, in that period $\gamma_{j,t} = 1$ and we generate a process as for $\beta_{1,t}$.

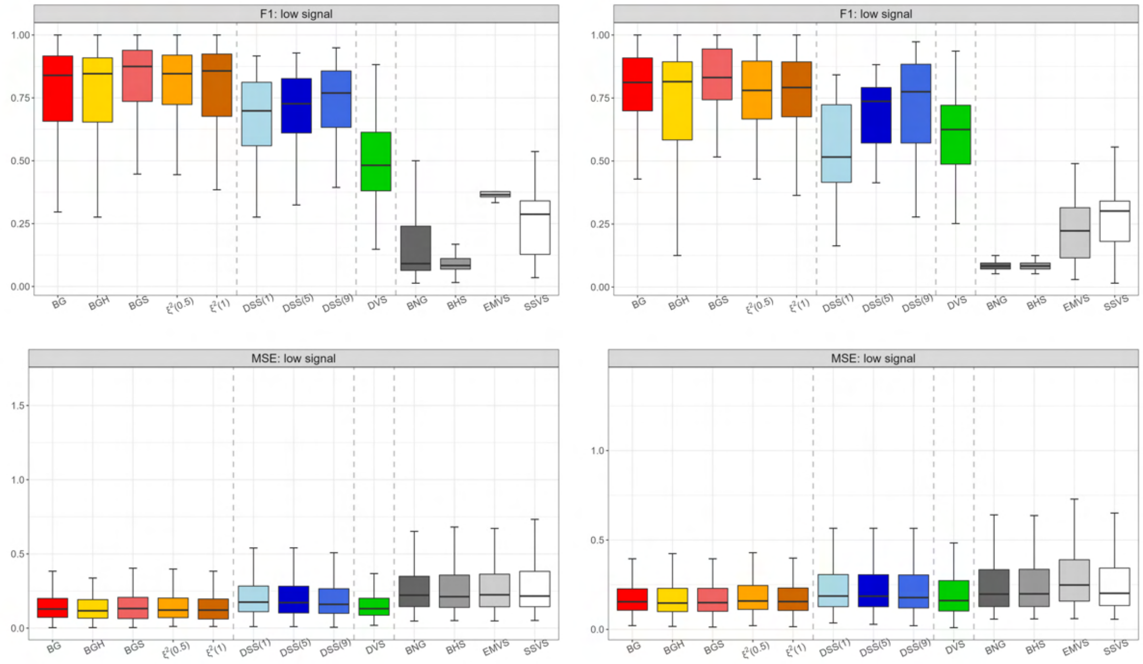


FIGURE 4.12: F_1-score (top panel) and MSE (bottom panel) for the dynamic sparsity setting with low signal, when $p = 50$ (left) and $p = 200$ (right).

a time-varying dynamic in the coefficient is essential to detect short periods of signal since the rolling window procedure returns estimated trajectories that do not respond to sudden variations, and therefore strongly under-performs in terms of F_1-score. A second comment which arises is that, on median, we observe a better performance of our methodologies with respect of both `DSS` and `DVS`.

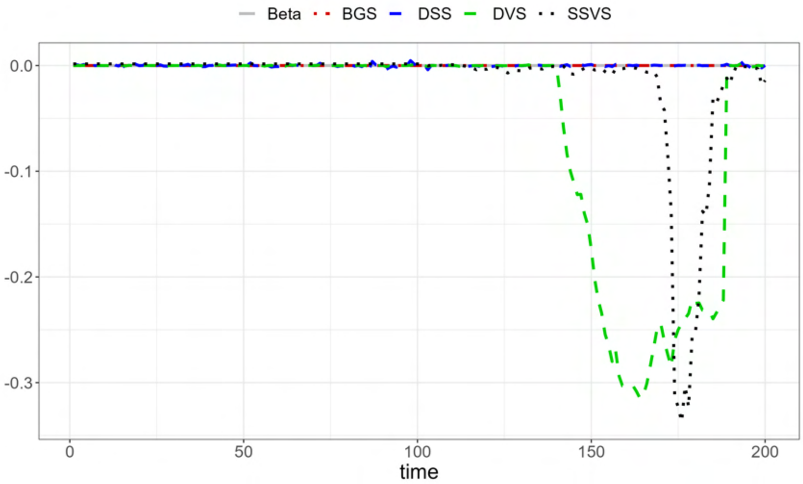


FIGURE 4.13: Example of coefficient constant at zero $\beta_{8,t}$.

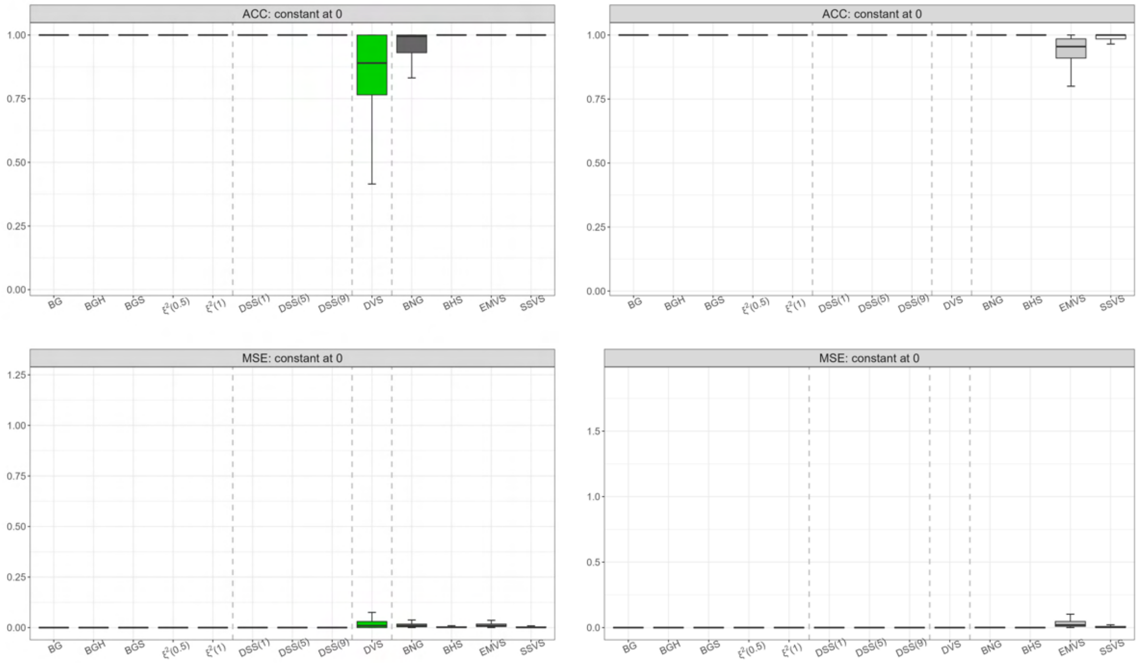


FIGURE 4.14: F_1-score (top panel) and MSE (bottom panel) for the always zero coefficients, when $p = 50$ (left) and $p = 200$ (right).

Constant at zero. Figure 4.13 shows estimation results for one replicate when the coefficient is set always to be zero. Within this setting there's no signal to identify. To this aim we report the total accuracy measure (ACC), computed as % of true zeros recognized, instead of the F_1-score. By looking at Figure 4.14, notice that all the models provide good results when there's only noise to recognize. The only issues appear when we focus on `DVS` for $p = 50$ and some rolling window estimation methods.

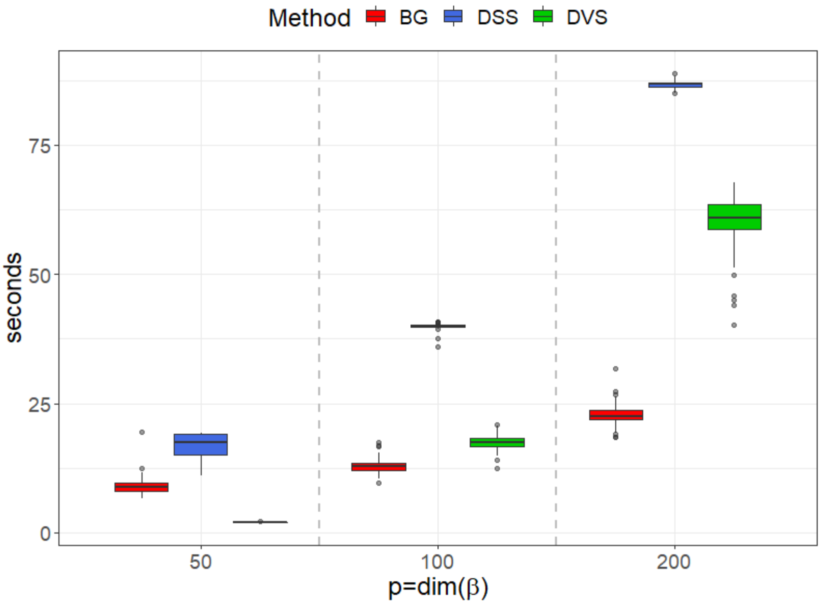


FIGURE 4.15: Computational efficiency of the algorithms computed as running time in second, varying the dimension p.

Computational efficiency. The last simulation result concerns the computational efficiency of the alternative methodologies. To evaluate the latter we track the running time in seconds of the algorithms. Figure 4.15 highlight two main conclusions. First of all, the `DSS` leads to the slowest algorithm, regardless to the dimension of the parameter p. Secondly, and perhaps more interesting, we observe that, when $p = 50$, the `DVS` is faster that `BG`, but when we move towards higher dimensions the situation changes. In fact, for $p = 100$, `BG` slightly over-performs `DVS`, but when $p = 200$, `BG` provides posterior inference 2.70 time faster than `DVS` and 3.85 time faster than `DSS`, on median. The computational efficiency of `BG` comes from the properties of the proposed algorithm presented in Section 4.3. This makes our approach particularly suitable in cases in which n is moderate, p is big and the problem is sparse.

4.5 Inflation forecasting

In this Section we evaluate the performance of Bernoulli-Gaussian model with time-varying sparsity approach to predict the future dynamic of the inflation using a large set of predictors. The latter study has a long history within macroeconomic literature (a non-exaustive list of works includes Stock and Watson, 2010; Koop and Korobilis, 2012; Kalli and Griffin, 2014; Koop and Korobilis, 2020; Ročková and McAlinn, 2021).

We retrieve the macroeconomic data from the FRED-QD database McCracken and Ng (2020). The variables consists in quarterly data spanning the period 3rd quarter

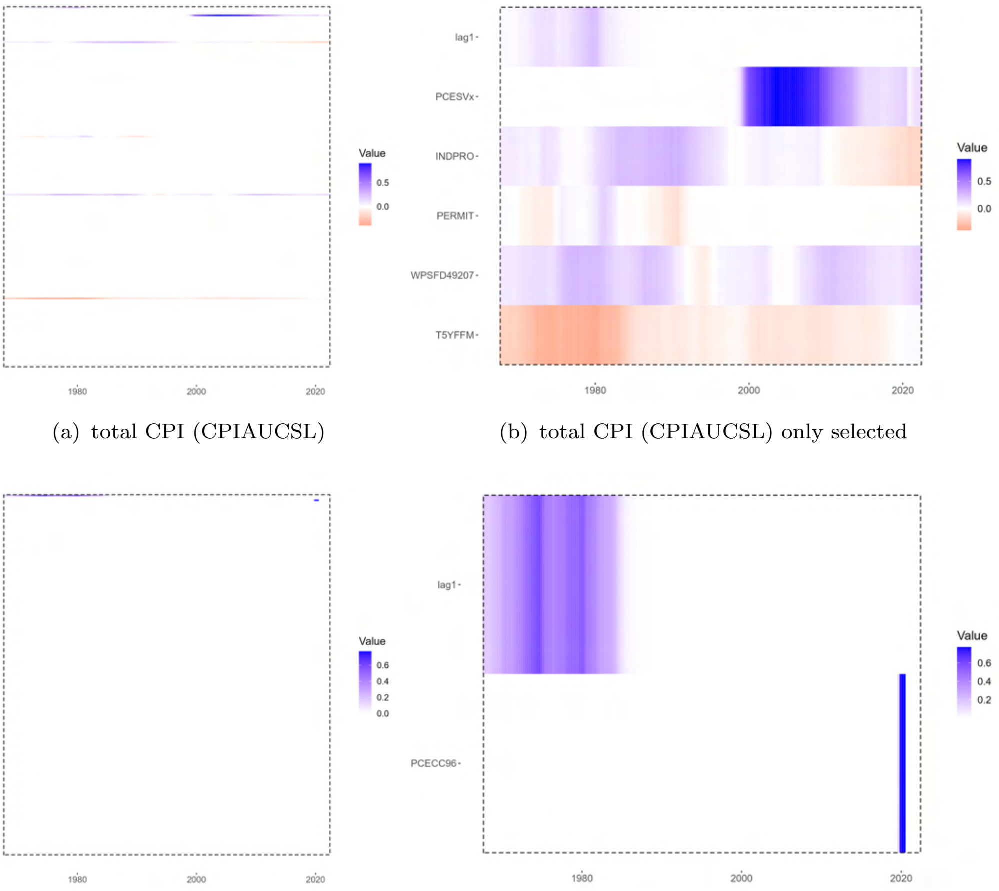
1967 to 2nd quarter 2022, such that the sample includes oil shocks in 1973 and 1979, mild recession in 1990, the dot-com bubble and the great recession in 2007-2009, and the covid-19 pandemic since late 2019. We focus on forecasting four measures of inflation, namely total CPI (CPIAUCSL), core CPI (CPILFESL), GDP deflator (GDPCTPI), and PCE deflator (PCECTPI). The name in parenthesis coincides with the variables' code in the original database. The 229 predictors are transformed according to standard norms in literature and the set also includes first two lags of the response variable.

We divide the following real data application into two parts. The first one focuses on in sample estimation to assess some structural evaluations of the process that drives the inflation. The second part compares three sets of models in terms of both point prediction accuracy and density forecast.

4.5.1 In sample analysis

First of all we focus on an in sample evaluation of the dynamics that drive inflation. Figure 4.16 and Figure 4.17 show the evolution over time (x-axis) of the regression coefficients for the predictors (y-axis). The left panels highlight that the estimates are very sparse and only few parameters enter in the model for at least one time. The behavior of the selected time-varying coefficients is better depicted in Figures 4.18–4.21 together with the trajectories of the posterior inclusion probabilities (bottom panel). Different target measures of inflation are driven by different number of predictors. However, they all show dependence on the first lag in the first part of the sample. Interestingly, some variables such as industrial production (INDPRO), 5-year treasury interest rates (T5YFFM), and producer price index (WPSFD49207) are commonly chosen for different inflation measures and show similar behaviors.

Another interesting fact that arises from the in sample analysis regards the relationship between the signal and volatility. The signal is computed at each time t as the sum of absolute values of the variational mean of the active regression coefficients, i.e., $\sum_{j=1}^{p} |\mu_{q(\tilde{\beta}_{j,t})}|$, and serves as a measure of information available to predict the target variable. Figure 4.22 shows that when the volatility is low, a weak signal is observed meaning that the variability in the target is treated as pure noise. In contrast, when the volatility measure increases, the power of the signal does, meaning there's a part of the variability of the inflation measure that can be explained using the large set of covariates.



(a) total CPI (CPIAUCSL)

(b) total CPI (CPIAUCSL) only selected

(c) core CPI (CPILFESL)

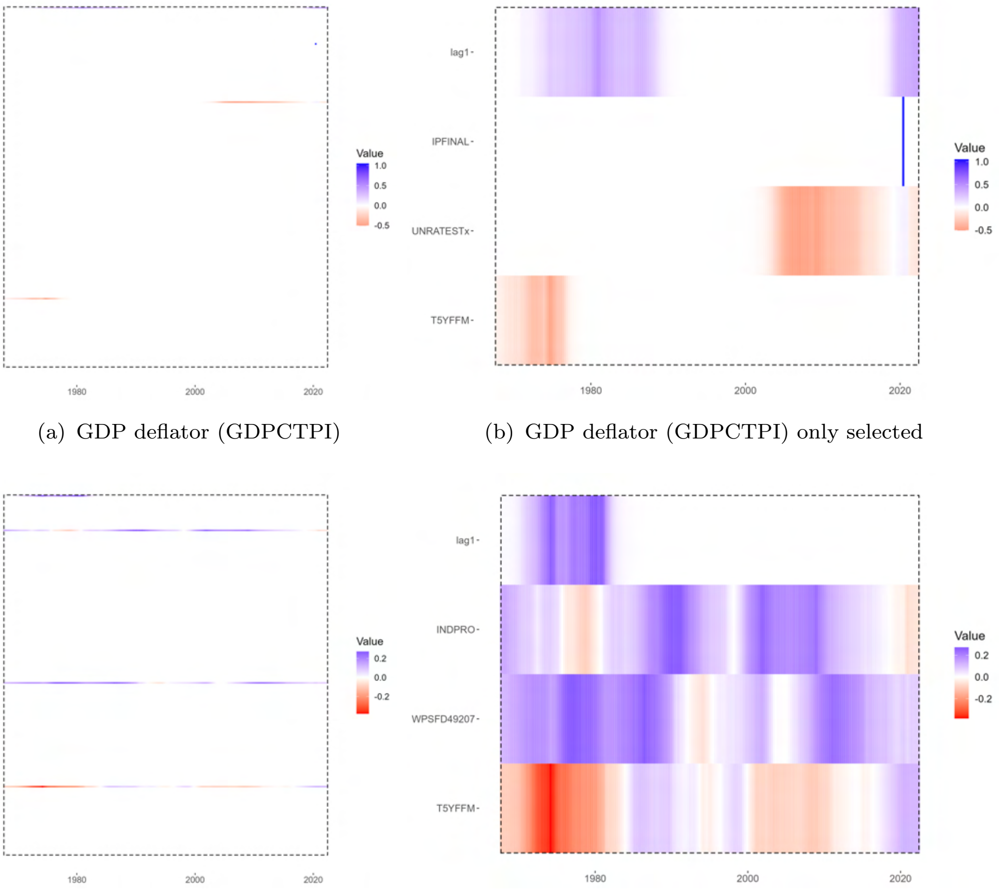
(d) core CPI (CPILFESL) only selected

FIGURE 4.16: In sample estimates for regression coefficients when CPI measures are target. The heatmap can be seen as a matrix collecting the values of the time-varying coefficients for each variable (y-axis) over time (x-axis).

4.5.2 Forecasting performance

In this Section we evaluate the forecasting performances of our methodology compared with some widely used benchmarks and state-of-the-art approaches. We look at both point prediction accuracy, by considering the mean-squared error (MSE), and density forecast performance, assessed in terms of log-predictive density score (LPD). Beside the measure for each target variable, we also provide some aggregated indicators. For the mean-squared error we take into account the mean (Avg) and the weighted mean (W Avg) where the weights depend on the marginal variance of the target variable. As an aggregated measure for density forecast we consider the multivariate predictive density score (Multi) which in our scenario coincides with the sum of the LPD for each target.

We can divide the competing methods into three groups. The first one is made of what we call *benchmark* models and represents tough competitors to beat in macroeconomic forecasting. The latter are the local-level (or random-walk, `RW` hereafter) model,



(a) GDP deflator (GDPCTPI)

(b) GDP deflator (GDPCTPI) only selected

(c) PCE deflator (PCECTPI)

(d) PCE deflator (PCECTPI) only selected

FIGURE 4.17: In sample estimates for regression coefficients when GDP and PCE deflator are target. The heatmap can be seen as a matrix collecting the values of the time-varying coefficients for each variable (y-axis) over time (x-axis).

an autoregressive of order two (`AR2`), and an autoregressive of order two with time-varying parameters (`TVAR`). Notice that these models does not provide any information about the factors that drive inflation since they do not consider the large set of predictors. The second group are composed by static models estimated using a rolling window procedure. A factor model with 5 factors (`F5`) is added to the shrinkage and variable selection methodologies already presented in the simulation study. The last group of models considers recent advances in dynamic variable selection with time varying parameters, i.e., `BGS`, `BGH`, `DSS`, and `DVS`, that have been presented in Section 4.4.2.

Table 4.1 and 4.2 show the results. Here we highlight both the best model (bold) and the best model excluding the benchmarks (red). Both tables suggest that `RW` is a tough benchmark to beat in point prediction, while autoregressive models are difficult to outperform when looking at density forecast. However, our methods, namely `BGS` and `BGH`, provide good performances and are always the best alternative among the models

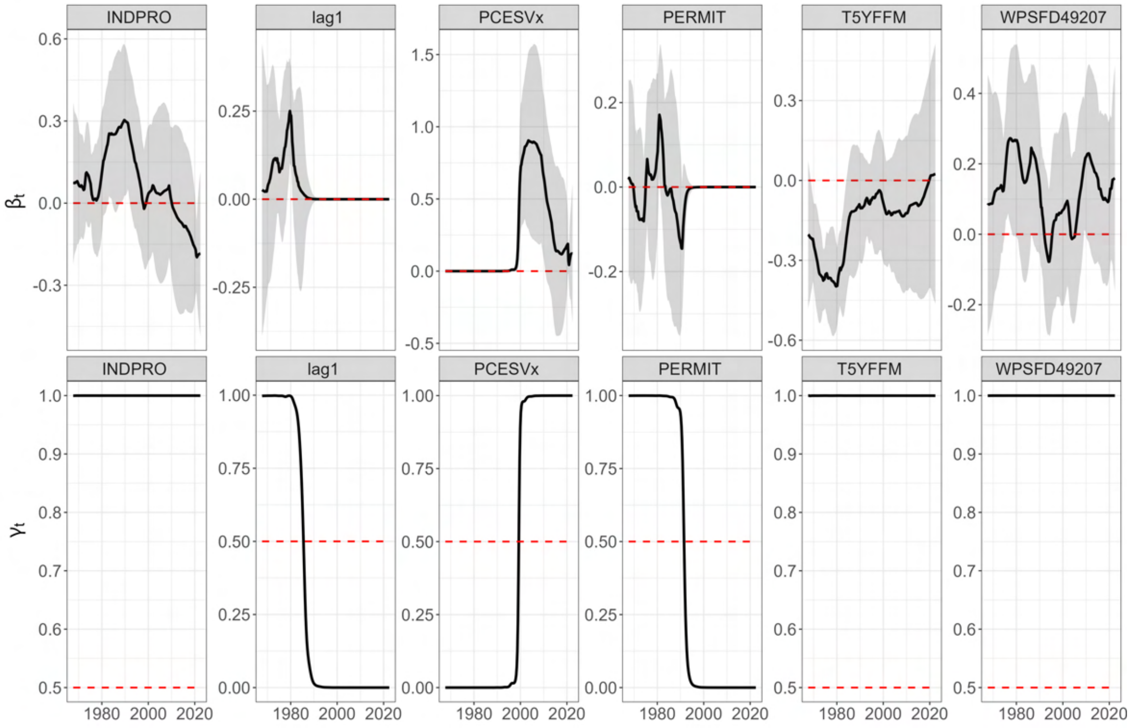


FIGURE 4.18: Time-varying coefficients estimates $\mu_{q(\tilde{\beta}_{j,t})}$ and posterior inclusion probabilities for total CPI (CPIAUCSL).

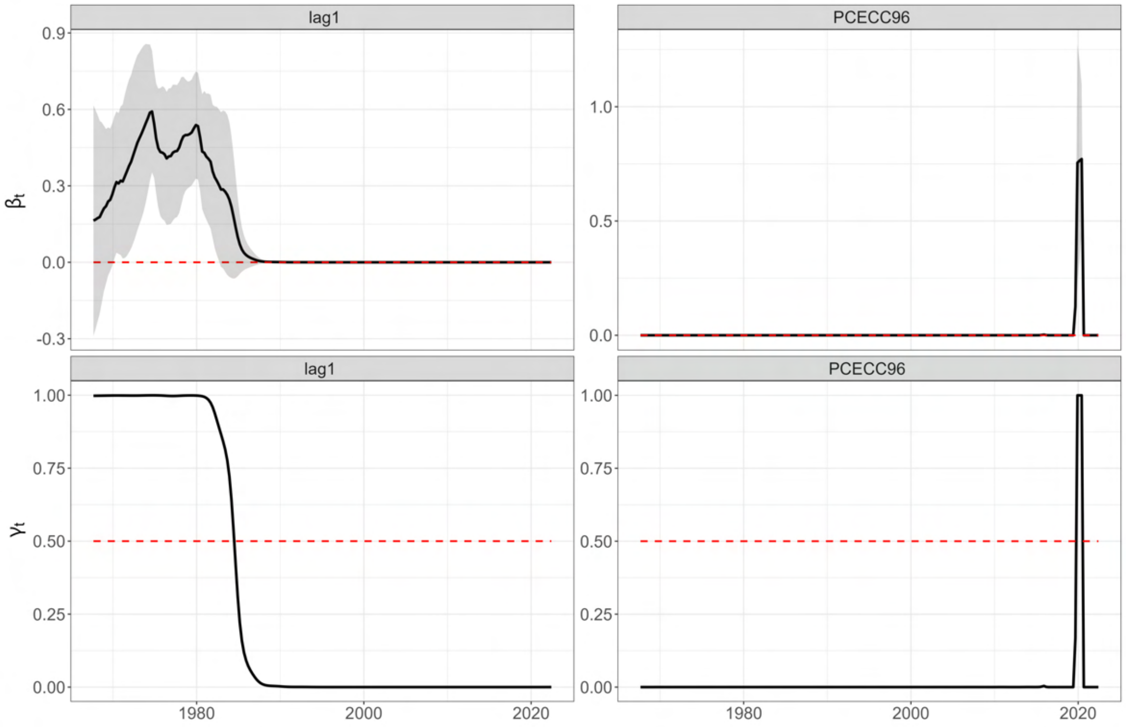


FIGURE 4.19: Time-varying coefficients estimates $\mu_{q(\tilde{\beta}_{j,t})}$ and posterior inclusion probabilities for core CPI (CPILFESL).

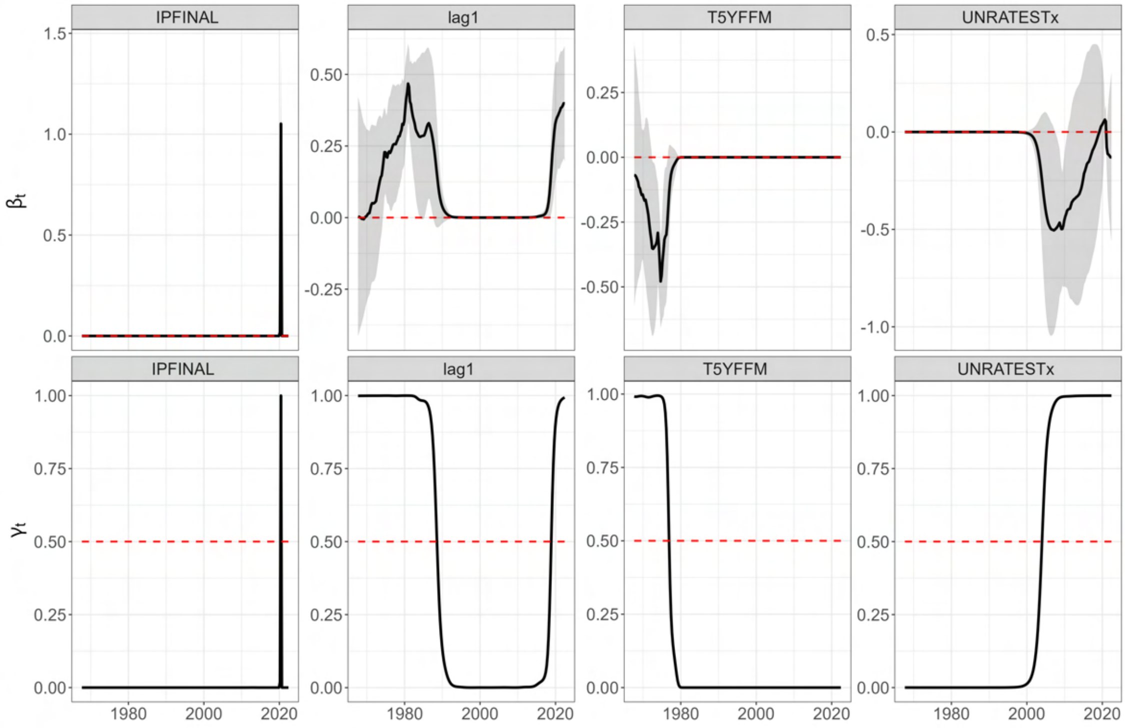


FIGURE 4.20: Time-varying coefficients estimates $\mu_{q(\tilde{\beta}_{j,t})}$ and posterior inclusion probabilities for GDP deflator (GDPCTPI).

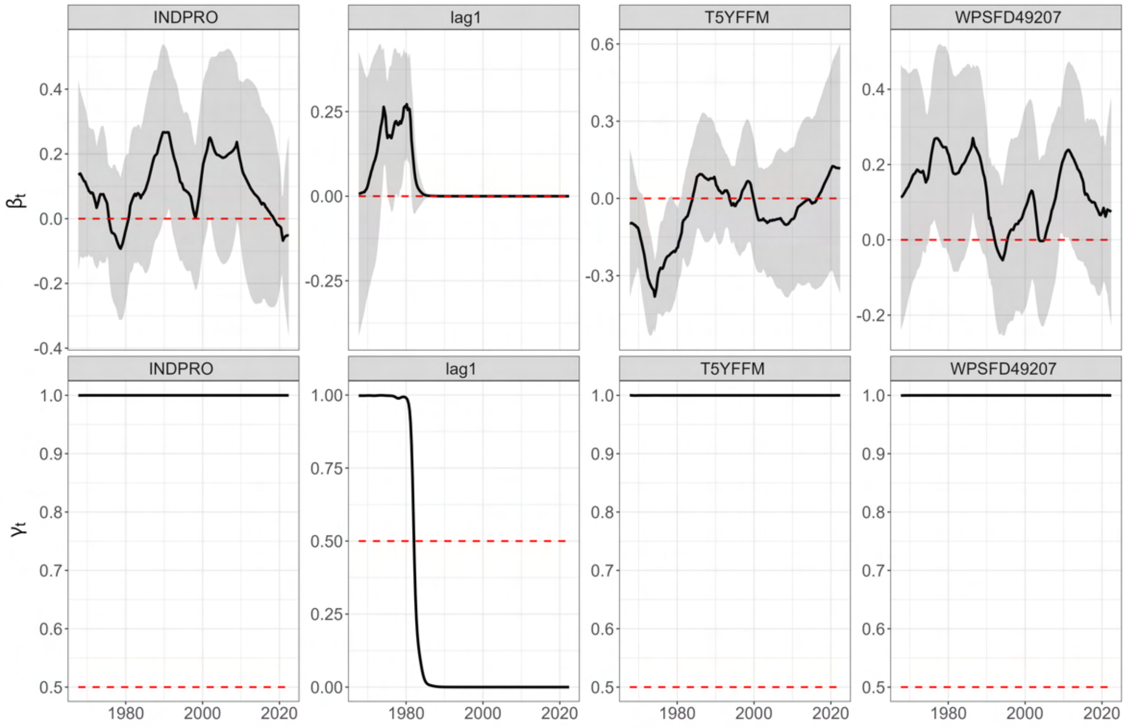
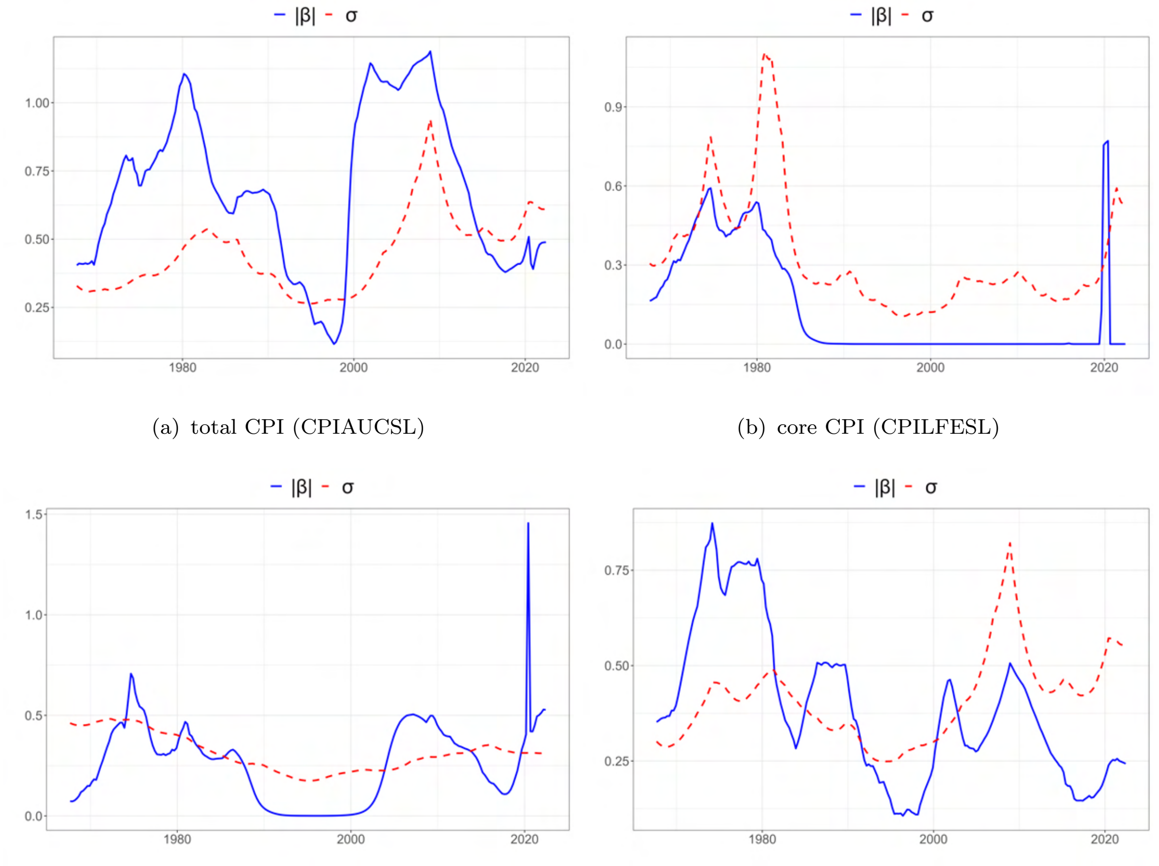


FIGURE 4.21: Time-varying coefficients estimates $\mu_{q(\tilde{\beta}_{j,t})}$ and posterior inclusion probabilities for PCE deflator (PCECTPI).



(a) total CPI (CPIAUCSL)

(b) core CPI (CPILFESL)

(c) GDP deflator (GDPCTPI)

(d) PCE deflator (PCECTPI)

FIGURE 4.22: Signal (blue) computed as $\sum_{j=1}^{p} |\mu_{q(\tilde{\beta}_{j,t})}|$, for $t = 1, \ldots, n$, against the time-variant volatility σ_t.

with many predictors (the only exception is the LPD for CPILFESL).

Although the MSE evaluation metrics indicates better performances for the benchmark, we implement a Diebold-Mariano test (Diebold and Mariano, 1995) to assess weather the performances are statistically equal or not. Figure 4.23 shows that `BGS` and `BGH` provide the same performances as the benchmark, while they are frequently outperforming all the other methodologies (see the first two columns of the plots). In conclusion, if the aim is just providing accurate predictions, the benchmark models and our dynamic Bernoulli-Gaussian specification are equivalent. However, if the final goal is to juxtapose accurate predictions with the understanding of the key drivers of the inflation's dynamic, then the proposed algorithm represents the best alternative so far.

Table 4.1: Point forecast.

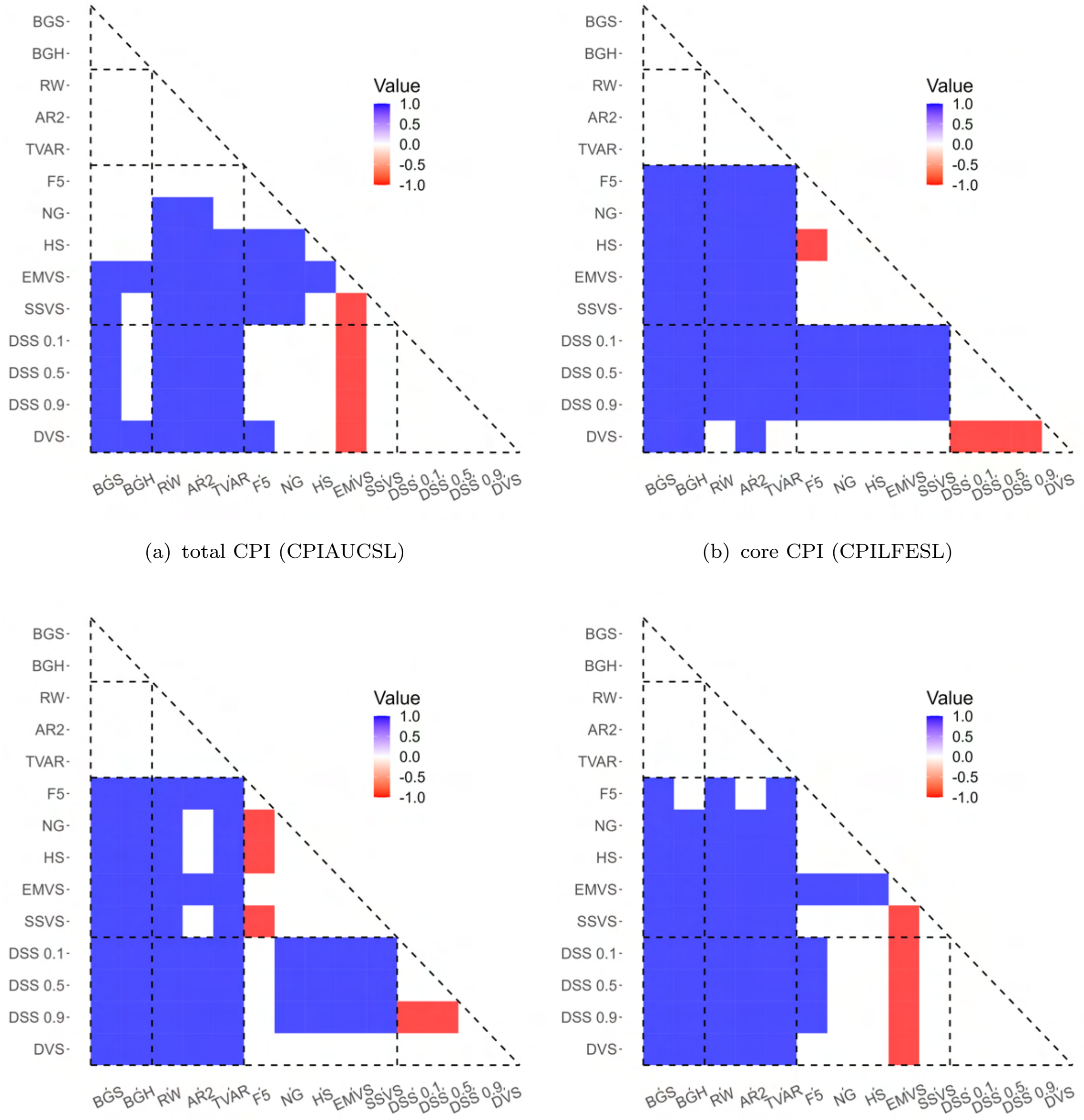
Bold denotes the best model, and red denotes the best model among the ones with large-set of covariates and variable selection. We report the MSE for each target variable together with aggregate measures.

	CPIAUCSL	CPILFESL	GDPCTPI	PCECTPI	Avg	W Avg
RW	**1.745**	**2.382**	1.130	1.071	**1.582**	**1.186**
AR2	1.943	2.582	1.244	0.964	1.683	1.276
TVAR	1.857	2.513	1.194	1.090	1.664	1.250
F5	2.070	2.627	1.727	1.480	1.976	1.469
NG	2.517	3.073	1.357	1.326	2.068	1.567
HS	2.497	3.298	1.360	1.263	2.105	1.601
EMVS	3.389	4.803	1.478	1.417	2.772	2.164
SSVS	2.433	3.422	1.398	1.359	2.153	1.630
BGS	1.772	2.565	1.027	0.989	1.588	1.204
BGH	1.814	2.712	**1.011**	**0.950**	1.622	1.238
DSS 0.1	2.393	2.941	1.952	2.134	2.355	1.742
DSS 0.5	2.394	2.940	1.952	2.132	2.354	1.741
DSS 0.9	2.391	2.939	1.945	2.148	2.356	1.742
DVS	2.855	3.595	1.669	1.330	2.362	1.799

Table 4.2: Density forecast.

Bold denotes the best model, and red denotes the best model among the ones with large-set of covariates and variable selection. We report the LPD for each target variable together with an aggregated measure.

	CPIAUCSL	CPILFESL	GDPCTPI	PCECTPI	Multi
RW	-2.316	-2.576	-2.274	-1.607	-8.773
AR2	-2.064	**-2.327**	**-1.498**	-1.148	**-7.036**
TVAR	-1.994	-2.395	-1.559	**-1.104**	-7.051
F5	-2.200	-2.486	-1.857	-1.558	-8.102
NG	-2.105	-2.348	-1.781	-1.766	-7.999
HS	-2.112	-2.394	-1.866	-1.681	-8.054
EMVS	-5.555	-5.724	-3.235	-2.245	-16.759
SSVS	-2.559	-2.807	-2.519	-2.668	-10.553
BGS	-2.058	-2.427	-1.739	-1.251	-7.474
BGH	**-1.971**	-2.506	-1.740	-1.339	-7.557
DSS 0.1	-4.225	-4.420	-3.960	-3.875	-16.481
DSS 0.5	-4.193	-4.149	-3.900	-4.033	-16.275
DSS 0.9	-4.041	-4.345	-3.902	-3.876	-16.163
DVS	-6.614	-9.006	-4.761	-2.643	-23.024



(a) total CPI (CPIAUCSL)

(b) core CPI (CPILFESL)

(c) GDP deflator (GDPCTPI)

(d) PCE deflator (PCECTPI)

FIGURE 4.23: Diebold-Mariano test for the null hypothesis $\mathcal{H}_0 : MSE^C = MSE^R$ (forecasting equivalence), where MSE^C and MSE^R denote the mean-squared error of the column and row model respectively. If the null is not rejected at level 5% we report 0 (white). If the null is rejected and $MSE^C < MSE^R$, so that the column model provides better forecasts than the row one, we report 1 (blue). If the null is rejected and $MSE^C > MSE^R$ we report -1 (red).

4.6 Concluding remarks

In this Chapter we extend the Bernoulli-Gaussian model in order to deal with dynamic sparsity in time-varying parameters regression. Unlike recent state-of-the-art developments, the proposed model specification assumes that the latent dynamic sparsity is governed by a stochastic process and the weak dependence on fixed hyper-parameters makes this method more robust with respect to more commonly used spike-and-slab type dynamic priors. Moreover, the semi-parametric variational Bayes algorithm presented provides exact sparsity when the coefficient is set to zero. In addition, it is suitable to deal with large set of covariates, thanks to the dimension reduction property. Empirically, we show two results. First of all, the proposed algorithm is sufficiently accurate in approximating the posterior distribution compared to MCMC, especially when the underlying parameter is included in the model. Secondly, we assess the better performances in point estimation, signal identification, and forecasting, when compared to state-of-the-art methods.

Conclusion

In this thesis we develop novel algorithms to perform inference in high-dimensional dynamic models. We consider different notions of dimensionality: *fat* (many variables) and *tall* (long series) data. Together with computational issues, increasing dimension often leads to the over-parametrization problem which causes poor interpretability and prediction performances. As a consequence, regularization of the estimates and variable selection techniques are desirable properties within a large-dimensional setting. The common thread of the Chapters in this thesis is variational Bayes. The latter represents an appealing approach to approximate Bayesian inference since it is proven to be computationally convenient with respect to classical MCMC methods. In this thesis variational Bayes represents the computational solution to tackle high-dimensional problems. As regards the over-parametrization issue, we focus on different solutions across Chapters.

In Chapter 2 we develop a novel algorithm to estimate a multivariate regression model with continuous shrinkage priors. The latter is suitable to deal with the increasing dimension of the cross-sections and size of the covariates' set. We show that the proposed approach permits a better estimation of the regression coefficients and a more accurate identification of the true signal with respect to state-of-the-art algorithms.

Chapter 3 focuses on the estimation of univariate stochastic volatility models. First of all, we provide a more accurate Gaussian approximation with respect to the existing literature. Second, we propose to use the parametric approximation scheme to introduce regularization in the estimates. In particular, we develop a flexible algorithm that can be used to obtain a smooth trajectory for the posterior estimates of the latent volatility process.

In Chapter 4 we extend the static Bernoulli-Gaussian model for variable selection to deal with dynamic sparsity within a variational Bayes approach to approximate Bayesian inference. We highlight that this model specification and the choice of the proposed variational densities provide the optimal solution to dynamically select important covariates in the context of time-varying parameter linear models, so far. The over-performance of our algorithm is emphasized in both a simulated scenario and an application in inflation forecasting.

Future research

The most interesting ideas for a future path of research regards some extensions of the dynamic variable selection described in Chapter 4. The first one that comes to mind is the multivariate case. A second, perhaps more challenging advancement considers the nature of the response variable. To the best of our knowledge, at the moment there's no algorithms suitable to deal with variable selection in time-varying generalized linear models. Our belief is that this should be an intriguing and stimulating topic, as well as useful in practice. In fact, binary and counts time series are interesting in some applications such as fraudolent transactions, cybersecurity (Soundarya and Usha, 2020), companies and countries default analysis (Kristóf, 2021), Covid-19 data, and gene expression data (Bar-Joseph *et al.*, 2012). A last extension of the proposed model, regards the type of dependence. In Chapter 4 we analise the case in which the data are equally spaced over time, but this is not always the case. An example is given by tick-by-tick data in finance (Engle and Sun, 2005). The joint representation of the regression coefficient as a Gaussian Markov random field is of particular convenience in order to easily account for different dependence structures. In fact, by simply changing the precision matrix of the random field, we fall into different scenarios (see Rue and Held, 2005). Irregular spaced time locations can be modeled assuming a joint representation of a continuous random walk process. Moreover, not only time-dependence can be incorporated, but also a *spatial* variable selection can be considered for lattice data or irregular locations in space. The latter has recently gained some attention within a machine learning context (Meyer *et al.*, 2019).

Appendix A

This appendix refers to chapter 2 and it provides the derivation of the optimal densities used in the mean-field variational Bayes algorithms. The derivation concerns the optimal densities for both the normal prior as well as the adaptive Bayesian lasso, the adaptive normal-gamma and the horseshoe. In addition, in this appendix we provide additional simulation and empirical results.

A.1 Auxiliary theoretical results

This Section provides major results that will be repeatedly used in the proofs of the derivation of the optimal densities used in the mean-field variational Bayes algorithms presented in Appendix A.2.

Result A.1. Assume that $\mathbf{y}$ is a n-dimensional row vector, $\mathbf{X}$ a $p \times n$ matrix and $\boldsymbol{\vartheta}$ a p-dimensional row vector of parameters whose distribution is denoted by $q(\boldsymbol{\vartheta})$.
Define $\|\mathbf{y} - \boldsymbol{\vartheta}\mathbf{X}\|_2^2 = (\mathbf{y} - \boldsymbol{\vartheta}\mathbf{X})(\mathbf{y} - \boldsymbol{\vartheta}\mathbf{X})^\intercal$, then it holds:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\vartheta}}\left[\|\mathbf{y} - \boldsymbol{\vartheta}\mathbf{X}\|_2^2\right] &= \mathbf{y}\mathbf{y}^\intercal + \mathbb{E}_{\boldsymbol{\vartheta}}\left[\boldsymbol{\vartheta}\mathbf{X}\mathbf{X}^\intercal\boldsymbol{\vartheta}^\intercal\right] - 2\boldsymbol{\mu}_{q(\vartheta)}\mathbf{X}\mathbf{y}^\intercal \\
&= \mathbf{y}\mathbf{y}^\intercal + \mathsf{tr}\left\{\mathbb{E}_{\boldsymbol{\vartheta}}\left[\boldsymbol{\vartheta}^\intercal\boldsymbol{\vartheta}\right]\mathbf{X}\mathbf{X}^\intercal\right\} - 2\boldsymbol{\mu}_{q(\vartheta)}\mathbf{X}\mathbf{y}^\intercal \\
&= \mathbf{y}\mathbf{y}^\intercal + \boldsymbol{\mu}_{q(\vartheta)}\mathbf{X}\mathbf{X}^\intercal\boldsymbol{\mu}_{q(\vartheta)}^\intercal + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\vartheta)}\mathbf{X}\mathbf{X}^\intercal\right\} - 2\boldsymbol{\mu}_{q(\vartheta)}\mathbf{X}\mathbf{y}^\intercal \\
&= \|\mathbf{y} - \boldsymbol{\mu}_{q(\vartheta)}\mathbf{X}\|_2^2 + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\vartheta)}\mathbf{X}\mathbf{X}^\intercal\right\},
\end{aligned}$$

where $\mathbb{E}_{\boldsymbol{\vartheta}}(f(\boldsymbol{\vartheta}))$ denotes the expectation of the function $f(\boldsymbol{\vartheta}) : \mathbb{R}^p \to \mathbb{R}^k$ with respect to $q(\boldsymbol{\vartheta})$, $\mathsf{tr}(\cdot)$ denotes the trace operator that returns the sum of the diagonal entries of a square matrix, and $\boldsymbol{\mu}_{q(\vartheta)}$ and $\boldsymbol{\Sigma}_{q(\vartheta)}$ denotes the mean and variance-covariance matrix of $\boldsymbol{\vartheta}$.

Result A.2. Let $\boldsymbol{\Theta}$ be a $d \times p$ random matrix with elements $\vartheta_{i,j}$, for $i = 1, \ldots, d$ and $j = 1, \ldots, p$, and let $\mathbf{A}$ be a $p \times p$ matrix. Our interest relies on the computation of the expectation of $\boldsymbol{\Theta}\mathbf{A}\boldsymbol{\Theta}^\intercal$ with respect to the distribution of $\boldsymbol{\Theta}$, where the expectation is taken element-wise. The (i, j)-th entry of $\boldsymbol{\Theta}\mathbf{A}\boldsymbol{\Theta}^\intercal$ is equal to $\boldsymbol{\vartheta}_i\mathbf{A}\boldsymbol{\vartheta}_j^\intercal$, where $\boldsymbol{\vartheta}_i$ and $\boldsymbol{\vartheta}_j$ denote the i-th and j-th row of $\boldsymbol{\Theta}$, respectively.

Therefore, the (i,j)-th entry of $\mathbf{\Theta A \Theta}^\intercal$ is equal to:

$$\mathbb{E}\big(\boldsymbol{\vartheta}_i \mathbf{A} \boldsymbol{\vartheta}_j^\intercal\big) = \mathbb{E}\big(\mathsf{tr}\big\{\boldsymbol{\vartheta}_j^\intercal \boldsymbol{\vartheta}_i \mathbf{A}\big\}\big) = \mathsf{tr}\big\{\mathbb{E}(\boldsymbol{\vartheta}_j^\intercal \boldsymbol{\vartheta}_i \mathbf{A})\big\} = \mathsf{tr}\big\{\mathbb{E}(\boldsymbol{\vartheta}_j^\intercal \boldsymbol{\vartheta}_i)\mathbf{A}\big\}.$$

Denote by $\boldsymbol{\mu}_{\vartheta_i} = \mathbb{E}(\boldsymbol{\vartheta}_i)$ and $\boldsymbol{\Sigma}_{\vartheta_i,\vartheta_j} = \mathsf{Cov}(\boldsymbol{\vartheta}_i, \boldsymbol{\vartheta}_j)$, then the previous expectation reduces to:

$$\mathbb{E}(\boldsymbol{\vartheta}_i \mathbf{A} \boldsymbol{\vartheta}_j^\intercal) = \mathsf{tr}\big\{\big(\boldsymbol{\mu}_{\vartheta_j}^\intercal \boldsymbol{\mu}_{\vartheta_i} + \boldsymbol{\Sigma}_{\vartheta_i,\vartheta_j}\big)\mathbf{A}\big\} = \boldsymbol{\mu}_{\vartheta_i} \mathbf{A} \boldsymbol{\mu}_{\vartheta_j}^\intercal + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{\vartheta_i,\vartheta_j}\mathbf{A}\right\}.$$

In matrix form, $\mathbb{E}(\mathbf{\Theta A \Theta}^\intercal) = \boldsymbol{\mu}_{\mathbf{\Theta}} \mathbf{A} \boldsymbol{\mu}_{\mathbf{\Theta}}^\intercal + \mathbf{K}_{\mathbf{\Theta}}$, where $\boldsymbol{\mu}_{\mathbf{\Theta}}$ is a $d \times p$ matrix with elements $\mu_{\vartheta_{i,j}}$, while $\mathbf{K}_{\mathbf{\Theta}}$ is a $d \times d$ symmetric matrix with elements equal to $\mathsf{tr}\left\{\boldsymbol{\Sigma}_{\vartheta_i,\vartheta_j}\mathbf{A}\right\}$.

Result (A.2) can be further generalized to compute the expectation of $\mathbf{\Theta}_1 \mathbf{A} \mathbf{\Theta}_2^\intercal$ with respect to the joint distribution of $(\mathbf{\Theta}_1, \mathbf{\Theta}_2)$ where $\mathbf{\Theta}_1$ is $d_1 \times p$ and $\mathbf{\Theta}_2$ is $d_2 \times p$.

Result A.3. Let $\boldsymbol{\vartheta}$ be a d-dimensional Gaussian random vector with mean $\boldsymbol{\mu}_\vartheta$ and covariance matrix $\boldsymbol{\Sigma}_\vartheta$. The expectation of the quadratic form $(\boldsymbol{\vartheta} - \boldsymbol{\mu}_\vartheta)^\intercal \boldsymbol{\Sigma}_\vartheta^{-1} (\boldsymbol{\vartheta} - \boldsymbol{\mu}_\vartheta)$ with respect to $\boldsymbol{\vartheta}$ is equal to d. Indeed:

$$\begin{aligned}
\mathbb{E}_\vartheta\left[(\boldsymbol{\vartheta} - \boldsymbol{\mu}_\vartheta)^\intercal \boldsymbol{\Sigma}_\vartheta^{-1} (\boldsymbol{\vartheta} - \boldsymbol{\mu}_\vartheta)\right] &= \mathsf{tr}\left\{\mathbb{E}_\vartheta\left[(\boldsymbol{\vartheta} - \boldsymbol{\mu}_\vartheta)(\boldsymbol{\vartheta} - \boldsymbol{\mu}_\vartheta)^\intercal\right]\boldsymbol{\Sigma}_\vartheta^{-1}\right\} \\
&= \mathsf{tr}\left\{\left[(\boldsymbol{\mu}_\vartheta - \boldsymbol{\mu}_\vartheta)(\boldsymbol{\mu}_\vartheta - \boldsymbol{\mu}_\vartheta)^\intercal + \boldsymbol{\Sigma}_\vartheta\right]\boldsymbol{\Sigma}_\vartheta^{-1}\right\} \\
&= \mathsf{tr}\left\{\boldsymbol{\Sigma}_\vartheta \boldsymbol{\Sigma}_\vartheta^{-1}\right\} \\
&= \mathsf{tr}\left\{\mathbf{I}_d\right\} \\
&= d.
\end{aligned}$$

A.2 Derivation of the variational Bayes algorithms

This appendix explains how to obtain the relevant quantities of the mean-field variational Bayes algorithms described in Section 2.3 for the prior distributions described in Section 2.3.1. We begin by discussing the non-informative prior, then turn to the adaptive Bayesian lasso, the adaptive normal-gamma and conclude with the horseshoe prior.

A.2.1 Normal prior specification

Proposition A.1. *The optimal variational density for the precision parameter ν_j is equal to $q^*(\nu_j) \equiv \mathsf{Ga}(a_{q(\nu_j)}, b_{q(\nu_j)})$, where, for $j = 1, \ldots, d$:*

$$\begin{aligned} a_{q(\nu_j)} &= a_\nu + T/2, \\ b_{q(\nu_j)} &= b_\nu + \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}_{-\nu_j}\left[\varepsilon_{j,t}^2\right], \end{aligned} \tag{A.1}$$

where

$$\begin{aligned} \mathbb{E}_{-\nu_j}\left[\varepsilon_{j,t}^2\right] = &\Big(y_{j,t} - \boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})} - \boldsymbol{\mu}_{q(\vartheta_j)}\mathbf{z}_{t-1}\Big)^2 \\ &+ \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\vartheta_j)}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal\right\} \\ &+ \mathsf{tr}\left\{\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)} + \boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}^\intercal\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}\right)\mathbf{K}_{\vartheta,t}\right\} \\ &+ \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}^\intercal\right\} \\ &- 2\mathbf{k}_{\boldsymbol{\vartheta},t}\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}^\intercal, \end{aligned}$$

where $\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})} = \mathbf{y}_t^j - \boldsymbol{\mu}_{q(\boldsymbol{\Theta}^j)}\mathbf{z}_{t-1}$, and, for $i = 1, \ldots, j-1$ and $k = 1, \ldots, j-1$, the elements in the matrix $\mathbf{K}_{\boldsymbol{\vartheta},t}$ and in the row vector $\mathbf{k}_{\boldsymbol{\vartheta},t}$ are $[\mathbf{K}_{\vartheta,t}]_{i,k} = \mathsf{tr}\left\{\mathsf{Cov}(\boldsymbol{\vartheta}_i, \boldsymbol{\vartheta}_k)\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal\right\}$ and $[\mathbf{k}_{\vartheta,t}]_i = \mathsf{tr}\left\{\mathsf{Cov}(\boldsymbol{\vartheta}_i, \boldsymbol{\vartheta}_j)\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal\right\}$ respectively. Notice that under row-factorization of $\boldsymbol{\Theta}$, we have that $\mathbf{k}_{\boldsymbol{\vartheta},t} = \mathbf{0}_j$.

Proof. Consider the model written for the j-th variable:

$$y_{j,t} = \boldsymbol{\beta}_j\mathbf{r}_{j,t} + \boldsymbol{\vartheta}_j\mathbf{z}_{t-1} + \varepsilon_{j,t}, \quad \varepsilon_{j,t} \sim \mathsf{N}(0, 1/\nu_j),$$

and notice that $\varepsilon_{j,t} = y_{j,t} - \boldsymbol{\beta}_j\mathbf{r}_{j,t} - \boldsymbol{\vartheta}_j\mathbf{z}_{t-1}$. Recall that a priori $\nu_j \sim \mathsf{Ga}(a_\nu, b_\nu)$ and compute $\log q^*(\nu_j) \propto \mathbb{E}_{-\nu_j}\left[\ell_j(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}) + \log p(\nu_j)\right]$:

$$\begin{aligned} \log q^*(\nu_j) &\propto \mathbb{E}_{-\nu_j}\left[\frac{T}{2}\log\nu_j - \frac{\nu_j}{2}\sum_{t=1}^{T}\varepsilon_{j,t}^2 + (a_\nu - 1)\log\nu_j - b_\nu\nu_j\right] \\ &\propto \left(\frac{T}{2} + a_\nu - 1\right)\log\nu_j - \nu_j\left(b_\nu + \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}_{-\nu_j}\left[\varepsilon_{j,t}^2\right]\right), \end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}_{-\nu_j}\left[\varepsilon^2_{j,t}\right] &= \mathbb{E}_{-\nu_j}\left[\left(y_{j,t} - \boldsymbol{\beta}_j \mathbf{r}_{j,t} - \boldsymbol{\vartheta}_j \mathbf{z}_{t-1}\right)^2\right] \\
&= y^2_{j,t} + \mathbb{E}_{\vartheta}\left[\boldsymbol{\vartheta}_j \mathbf{z}_{t-1}\mathbf{z}^{\intercal}_{t-1}\boldsymbol{\vartheta}_j\right] + \overbrace{\mathbb{E}_{\vartheta,\boldsymbol{\beta}_j}\left[\boldsymbol{\beta}_j \mathbf{r}_{j,t}\mathbf{r}^{\intercal}_{j,t}\boldsymbol{\beta}^{\intercal}_j\right]}^{\text{A}} \\
&\quad - 2y_{j,t}\mathbb{E}_{\vartheta}\left[\boldsymbol{\vartheta}_j\right]\mathbf{z}_{t-1} - 2y_{j,t}\mathbb{E}_{\boldsymbol{\beta}_j}\left[\boldsymbol{\beta}_j\right]\mathbb{E}_{\vartheta}\left[\mathbf{r}_{j,t}\right] \\
&\quad + 2\underbrace{\mathbb{E}_{\vartheta}\left[\boldsymbol{\vartheta}_j \mathbf{z}_{t-1}\mathbf{r}^{\intercal}_{j,t}\right]\mathbb{E}_{\boldsymbol{\beta}_j}\left[\boldsymbol{\beta}^{\intercal}_j\right]}_{\text{B}} \\
&= y^2_{j,t} + \boldsymbol{\mu}_{q(\vartheta_j)}\mathbf{z}_{t-1}\mathbf{z}^{\intercal}_{t-1}\boldsymbol{\mu}_{q(\vartheta_j)} + \boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}^{\intercal}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}^{\intercal}_{q(\boldsymbol{\beta}_j)} \\
&\quad - 2y_{j,t}\boldsymbol{\mu}_{q(\vartheta_j)}\mathbf{z}_{t-1} - 2y_{j,t}\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})} \\
&\quad + 2\boldsymbol{\mu}_{q(\vartheta_j)}\mathbf{z}_{t-1}\boldsymbol{\mu}^{\intercal}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}^{\intercal}_{q(\boldsymbol{\beta}_j)} \\
&\quad + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\vartheta_j)}\mathbf{z}_{t-1}\mathbf{z}^{\intercal}_{t-1}\right\} + \mathsf{tr}\left\{\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)} + \boldsymbol{\mu}^{\intercal}_{q(\boldsymbol{\beta}_j)}\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}\right)\mathbf{K}_{\vartheta,t}\right\} \\
&\quad + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}^{\intercal}_{q(\mathbf{r}_{j,t})}\right\} - 2\mathbf{k}_{\vartheta,t}\boldsymbol{\mu}^{\intercal}_{q(\boldsymbol{\beta}_j)} \\
&= \left(y_{j,t} - \boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})} - \boldsymbol{\mu}_{q(\vartheta_j)}\mathbf{z}_{t-1}\right)^2 \\
&\quad + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\vartheta_j)}\mathbf{z}_{t-1}\mathbf{z}^{\intercal}_{t-1}\right\} + \mathsf{tr}\left\{\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)} + \boldsymbol{\mu}^{\intercal}_{q(\boldsymbol{\beta}_j)}\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}\right)\mathbf{K}_{\vartheta,t}\right\} \\
&\quad + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}^{\intercal}_{q(\mathbf{r}_{j,t})}\right\} - 2\mathbf{k}_{\vartheta,t}\boldsymbol{\mu}^{\intercal}_{q(\boldsymbol{\beta}_j)},
\end{aligned}$$

where $\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})} = \mathbf{y}^j_t - \boldsymbol{\mu}_{q(\boldsymbol{\Theta}^j)}\mathbf{z}_{t-1}$.

The computations involving terms A and B are presented in the following equations. Firs of all, define $\boldsymbol{\beta}_j \mathbf{r}_{j,t}\mathbf{r}^{\intercal}_{j,t}\boldsymbol{\beta}^{\intercal}_j = \|\boldsymbol{\beta}_j \mathbf{r}_{j,t}\|^2_2$, then the term A above is equal to:

$$\begin{aligned}
\mathbb{E}_{\vartheta,\boldsymbol{\beta}_j}\left[\|\boldsymbol{\beta}_j \mathbf{r}_{j,t}\|^2_2\right] &= \mathbb{E}_{\boldsymbol{\beta}_j}\Big[\boldsymbol{\beta}_j \overbrace{\mathbb{E}_{\vartheta}\left[\mathbf{r}_{j,t}\mathbf{r}^{\intercal}_{j,t}\right]}^{\text{See Results A.1 and A.2}} \boldsymbol{\beta}^{\intercal}_j\Big] \\
&= \mathbb{E}_{\boldsymbol{\beta}_j}\left[\boldsymbol{\beta}_j\left\{\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}^{\intercal}_{q(\mathbf{r}_{j,t})} + \mathbf{K}_{\vartheta,t}\right\}\boldsymbol{\beta}^{\intercal}_j\right] \\
&= \boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}\left\{\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}^{\intercal}_{q(\mathbf{r}_{j,t})} + \mathbf{K}_{\vartheta,t}\right\}\boldsymbol{\mu}^{\intercal}_{q(\boldsymbol{\beta}_j)} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}\left[\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}^{\intercal}_{q(\mathbf{r}_{j,t})} + \mathbf{K}_{\vartheta,t}\right]\right\} \\
&= \|\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}\|^2_2 + \mathsf{tr}\left\{\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)} + \boldsymbol{\mu}^{\intercal}_{q(\boldsymbol{\beta}_j)}\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}\right)\mathbf{K}_{\vartheta,t}\right\} \\
&\quad + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}^{\intercal}_{q(\mathbf{r}_{j,t})}\right\},
\end{aligned}$$

while the term B is:

$$
\begin{aligned}
\mathbb{E}_{\vartheta}\left[\boldsymbol{\vartheta}_j\mathbf{z}_{t-1}\mathbf{r}_{j,t}^{\intercal}\right]\mathbb{E}_{\boldsymbol{\beta}_j}\left[\boldsymbol{\beta}_j^{\intercal}\right] &= \mathbb{E}_{\vartheta}\left[\boldsymbol{\vartheta}_j\mathbf{z}_{t-1}\mathbf{y}_t^{j\intercal} - \overbrace{\boldsymbol{\vartheta}_j\mathbf{z}_{t-1}\mathbf{z}_{t-1}^{\intercal}\boldsymbol{\Theta}^{j\intercal}}^{\text{See Result A.2}}\right]\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}^{\intercal} \\
&= \left(\boldsymbol{\mu}_{q(\vartheta_j)}\mathbf{z}_{t-1}\mathbf{y}_t^{j\intercal} - \boldsymbol{\mu}_{q(\vartheta_j)}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^{\intercal}\boldsymbol{\mu}_{q(\boldsymbol{\Theta}^j)}^{\intercal} - \mathbf{k}_{\vartheta,t}\right)\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}^{\intercal} \\
&= \boldsymbol{\mu}_{q(\vartheta_j)}\mathbf{z}_{t-1}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}^{\intercal}\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}^{\intercal} - \mathbf{k}_{\vartheta,t}\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}^{\intercal}.
\end{aligned}
$$

Notice that for the latter derivation we use Results A.1 and A.2. To conclude, we obtain:

$$
\log q^*(\nu_j) \propto \left(\frac{T}{2} + a_\nu - 1\right)\log\nu_j - \nu_j\left(b_\nu + \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}_{-\nu_j}\left[\varepsilon_{j,t}^2\right]\right),
$$

then take the exponential and notice that the latter is the kernel of a gamma random variable $\mathsf{Ga}(a_{q(\nu_j)}, b_{q(\nu_j)})$ as defined in Proposition A.1. □

Proposition A.2. *The optimal variational density for the parameter $\boldsymbol{\beta}_j$ for $j = 2, \dots, d$ is equal to $q^*(\boldsymbol{\beta}_j) \equiv \mathsf{N}_{j-1}(\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)})$, where:*

$$
\begin{aligned}
\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)} &= \left(\mu_{q(\nu_j)}\sum_{t=1}^{T}\left(\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}^{\intercal} + \mathbf{K}_{\vartheta,t}\right) + 1/\tau\mathbf{I}_{j-1}\right)^{-1}, \\
\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)} &= \boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}\mu_{q(\nu_j)}\sum_{t=1}^{T}\left(\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}(y_{j,t} - \boldsymbol{\mu}_{q(\vartheta_j)}\mathbf{z}_{t-1})^{\intercal} + \mathbf{k}_{\vartheta,t}\right).
\end{aligned}
\tag{A.2}
$$

Proof. Consider the model written for the j-th variable:

$$
y_{j,t} = \boldsymbol{\beta}_j\mathbf{r}_{j,t} + \boldsymbol{\vartheta}_j\mathbf{z}_{t-1} + \varepsilon_{j,t}, \quad \varepsilon_{j,t} \sim \mathsf{N}(0, 1/\nu_j).
$$

Recall that a priori $\boldsymbol{\beta}_j \sim \mathsf{N}_{j-1}(\mathbf{0}, \tau\mathbf{I}_{j-1})$ and compute the optimal variational density as $\log q^*(\boldsymbol{\beta}_j) \propto \mathbb{E}_{-\boldsymbol{\beta}_j}\left[\ell_j(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}) + \log p(\boldsymbol{\beta}_j)\right]$:

$$
\begin{aligned}
\log q^*(\boldsymbol{\beta}_j) &\propto \mathbb{E}_{-\boldsymbol{\beta}_j}\left[-\frac{\nu_j}{2}\sum_{t=1}^{T}\left(y_{j,t} - \boldsymbol{\vartheta}_j\mathbf{z}_{t-1} - \boldsymbol{\beta}_j\mathbf{r}_{j,t}\right)^2 - \frac{1}{2\tau}\boldsymbol{\beta}_j\boldsymbol{\beta}_j^{\intercal}\right] \\
&\propto \mathbb{E}_{-\boldsymbol{\beta}_j}\left[-\frac{1}{2}\left\{\boldsymbol{\beta}_j\left(\nu_j\sum_{t=1}^{T}\mathbf{r}_{j,t}\mathbf{r}_{j,t}^{\intercal} + 1/\tau\mathbf{I}_{j-1}\right)\boldsymbol{\beta}_j^{\intercal}\right.\right. \\
&\qquad\qquad \left.\left. -2\boldsymbol{\beta}_j\nu_j\sum_{t=1}^{T}\mathbf{r}_{j,t}(y_{j,t} - \boldsymbol{\vartheta}_j\mathbf{z}_{t-1})^{\intercal}\right\}\right],
\end{aligned}
$$

and, applying some results defined is Appendix A.1, we get:

$$
\begin{aligned}
\log q^*(\boldsymbol{\beta}_j) \propto -\frac{1}{2}\Bigg\{ & \boldsymbol{\beta}_j \left(\mu_{q(\nu_j)} \sum_{t=1}^{T} \mathbb{E}_{\vartheta} \overbrace{\left[\mathbf{r}_{j,t}\mathbf{r}_{j,t}^{\intercal}\right]}^{\text{Result A.2}} + 1/\tau \mathbf{I}_{j-1} \right) \boldsymbol{\beta}_j^{\intercal} \\
& - 2\boldsymbol{\beta}_j \mu_{q(\nu_j)} \sum_{t=1}^{T} \mathbb{E}_{\vartheta} \overbrace{\left[\mathbf{r}_{j,t}(y_{j,t} - \boldsymbol{\vartheta}_j \mathbf{z}_{t-1})^{\intercal}\right]}^{\text{Result A.2}} \Bigg\} \\
\propto -\frac{1}{2}\Bigg\{ & \boldsymbol{\beta}_j \left(\mu_{q(\nu_j)} \sum_{t=1}^{T} \left(\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}^{\intercal} + \mathbf{K}_{\vartheta,t}\right) + 1/\tau \mathbf{I}_{j-1} \right) \boldsymbol{\beta}_j^{\intercal} \\
& - 2\boldsymbol{\beta}_j \mu_{q(\nu_j)} \sum_{t=1}^{T} \left(\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}(y_{j,t} - \boldsymbol{\mu}_{q(\boldsymbol{\vartheta}_j)} \mathbf{z}_{t-1})^{\intercal} + \mathbf{k}_{\vartheta,t}\right) \Bigg\}.
\end{aligned}
$$

Take the exponential and notice that the latter is the kernel of a Gaussian random variable $\mathsf{N}_{j-1}(\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)})$, as defined in Proposition A.2. □

Proposition A.3. *The optimal joint variational density for the parameter* $\boldsymbol{\vartheta}$ *is equal to a multivariate Gaussian* $q^*(\boldsymbol{\vartheta}) \equiv \mathsf{N}_{dp}(\boldsymbol{\mu}_{q(\vartheta)}, \boldsymbol{\Sigma}_{q(\vartheta)})$*, where:*

$$
\boldsymbol{\Sigma}_{q(\vartheta)} = \left(\boldsymbol{\mu}_{q(\boldsymbol{\Omega})} \otimes \sum_{t=1}^{T} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^{\intercal} + 1/\upsilon \mathbf{I}_{dp} \right)^{-1}, \qquad \boldsymbol{\mu}_{q(\vartheta)} = \boldsymbol{\Sigma}_{q(\vartheta)} \sum_{t=1}^{T} \left(\boldsymbol{\mu}_{q(\boldsymbol{\Omega})} \otimes \mathbf{z}_{t-1} \right) \mathbf{y}_t, \tag{A.3}
$$

where $\boldsymbol{\mu}_{q(\boldsymbol{\Omega})} = \mathbb{E}_q\left[\boldsymbol{\Omega}\right] = \mathbb{E}_q\left[\mathbf{L}^{\intercal}\mathbf{V}\mathbf{L}\right] = (\mathbf{I}_d - \boldsymbol{\mu}_{q(\mathbf{B})})^{\intercal}\boldsymbol{\mu}_{q(\mathbf{V})}(\mathbf{I}_d - \boldsymbol{\mu}_{q(\mathbf{B})}) + \mathbf{C}_{\vartheta}$ *and* $\mathbf{C}_{\vartheta}$ *is a* $d \times d$ *symmetric matrix whose generic element is given by:*

$$
\left[\mathbf{C}_{\vartheta}\right]_{i,j} = \sum_{k=j+1}^{d} \mathsf{Cov}(\beta_{k,i}, \beta_{k,j}) \mu_{q(\nu_k)}.
$$

Proof. Consider the model written as $\mathbf{L}\mathbf{y}_t = \mathbf{L}\boldsymbol{\Theta}\mathbf{z}_{t-1} + \boldsymbol{\varepsilon}_t$ with $\boldsymbol{\varepsilon}_t \sim \mathsf{N}_d(0, \mathbf{V}^{-1})$ and then apply the vectorisation operation on the transposed and get $\mathbf{L}\mathbf{y}_t = (\mathbf{L} \otimes \mathbf{z}_{t-1}^{\intercal})\boldsymbol{\vartheta} + \boldsymbol{\varepsilon}_t$, where $\boldsymbol{\varepsilon}_t \sim \mathsf{N}_d(0, \mathbf{V}^{-1})$. Recall that a priori $\boldsymbol{\vartheta} \sim \mathsf{N}_{dp}(\mathbf{0}, \upsilon \mathbf{I}_{dp})$. Compute the optimal variational density for the parameter $\boldsymbol{\vartheta}$ as $\log q^*(\boldsymbol{\vartheta}) \propto \mathbb{E}_{-\vartheta}\left[\ell(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}) + \log p(\boldsymbol{\vartheta})\right]$:

$$
\begin{aligned}
\log q^*(\boldsymbol{\vartheta}) & \propto -\frac{1}{2}\mathbb{E}_{-\vartheta}\left[\sum_{t=1}^{T} \left(\mathbf{L}\mathbf{y}_t - (\mathbf{L} \otimes \mathbf{z}_{t-1}^{\intercal})\boldsymbol{\vartheta}\right)^{\intercal} \mathbf{V} \left(\mathbf{L}\mathbf{y}_t - (\mathbf{L} \otimes \mathbf{z}_{t-1}^{\intercal})\boldsymbol{\vartheta}\right) \right] - \frac{1}{2\upsilon}\mathbb{E}_{-\vartheta}\left[\boldsymbol{\vartheta}^{\intercal}\boldsymbol{\vartheta}\right] \\
& \propto -\frac{1}{2}\mathbb{E}_{-\vartheta}\left[\sum_{t=1}^{T} \left(\boldsymbol{\vartheta}^{\intercal}(\mathbf{L}^{\intercal}\mathbf{V}\mathbf{L} \otimes \mathbf{z}_{t-1}\mathbf{z}_{t-1}^{\intercal})\boldsymbol{\vartheta}\right) - 2\sum_{t=1}^{T} \boldsymbol{\vartheta}^{\intercal}\left((\mathbf{L}^{\intercal}\mathbf{V}\mathbf{L} \otimes \mathbf{z}_{t-1})\mathbf{y}_t \right) \right] - \frac{1}{2\upsilon}\boldsymbol{\vartheta}^{\intercal}\boldsymbol{\vartheta} \\
& \propto -\frac{1}{2}\left\{ \boldsymbol{\vartheta}^{\intercal} \left(\boldsymbol{\mu}_{q(\boldsymbol{\Omega})} \otimes \sum_{t=1}^{T} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^{\intercal} + 1/\upsilon \mathbf{I}_{dp} \right) \boldsymbol{\vartheta} - 2\boldsymbol{\vartheta}^{\intercal} \sum_{t=1}^{T} \left(\boldsymbol{\mu}_{q(\boldsymbol{\Omega})} \otimes \mathbf{z}_{t-1} \right) \mathbf{y}_t \right\}.
\end{aligned}
$$

To compute the expectation $\boldsymbol{\mu}_{q(\boldsymbol{\Omega})} = \mathbb{E}_{-\vartheta}\left[(\mathbf{I}_d - \mathbf{B})^\intercal \mathbf{V}(\mathbf{I}_d - \mathbf{B})\right]$ we use the following:

$$\begin{aligned}
\mathbb{E}_{\mathbf{B},\mathbf{V}}\left[(\mathbf{I}_d - \mathbf{B})^\intercal \mathbf{V}(\mathbf{I}_d - \mathbf{B})\right] &= \mathbb{E}_{\mathbf{B},\mathbf{V}}\left[\mathbf{V} - 2\mathbf{B}^\intercal\mathbf{V} - \mathbf{B}^\intercal\mathbf{V}\mathbf{B}\right] \\
&= \boldsymbol{\mu}_{q(\mathbf{V})} - 2\boldsymbol{\mu}_{q(\mathbf{B})}^\intercal \boldsymbol{\mu}_{q(\mathbf{V})} - \mathbb{E}_{\mathbf{B},\mathbf{V}}\left[\mathbf{B}^\intercal\mathbf{V}\mathbf{B}\right] \\
&= \boldsymbol{\mu}_{q(\mathbf{V})} - 2\boldsymbol{\mu}_{q(\mathbf{B})}^\intercal \boldsymbol{\mu}_{q(\mathbf{V})} + \boldsymbol{\mu}_{q(\mathbf{B})}^\intercal \boldsymbol{\mu}_{q(\mathbf{V})}\boldsymbol{\mu}_{q(\mathbf{B})} + \mathbf{C}_\vartheta \\
&= (\mathbf{I}_d - \boldsymbol{\mu}_{q(\mathbf{B})})^\intercal \boldsymbol{\mu}_{q(\mathbf{V})}(\mathbf{I}_d - \boldsymbol{\mu}_{q(\mathbf{B})}) + \mathbf{C}_\vartheta,
\end{aligned}$$

where we exploit the fact that the (i,j)-th element of $\mathbf{B}^\intercal\mathbf{V}\mathbf{B}$ is given by:

$$[\mathbf{B}^\intercal\mathbf{V}\mathbf{B}]_{i,j} = \sum_{k=j+1}^{d} \beta_{k,i}\beta_{k,j}\nu_k, \quad i \leq j \quad \text{and} \quad [\mathbf{B}^\intercal\mathbf{V}\mathbf{B}]_{i,j} = [\mathbf{B}^\intercal\mathbf{V}\mathbf{B}]_{j,i}$$

hence

$$\begin{aligned}
\mathbb{E}_{\mathbf{B},\mathbf{V}}\left[\mathbf{B}^\intercal\mathbf{V}\mathbf{B}\right]_{i,j} &= \mathbb{E}_{\mathbf{B},\mathbf{V}}\left[\sum_{k=j+1}^{d} \beta_{k,i}\beta_{k,j}\nu_k\right] \\
&= \sum_{k=j+1}^{d} \left(\mu_{q(\beta_{k,i})}\mu_{q(\beta_{k,j})} + \mathsf{Cov}(\beta_{k,i},\beta_{k,j})\right)\mu_{q(\nu_k)} \\
&= \sum_{k=j+1}^{d} \mu_{q(\beta_{k,i})}\mu_{q(\beta_{k,j})}\mu_{q(\nu_k)} + \sum_{k=j+1}^{d} \mathsf{Cov}\left(\beta_{k,i},\beta_{k,j}\right)\mu_{q(\nu_k)} \\
&= \left[\boldsymbol{\mu}_{q(\mathbf{B}^\intercal)}\boldsymbol{\mu}_{q(\mathbf{V})}\boldsymbol{\mu}_{q(\mathbf{B})}\right]_{i,j} + \sum_{k=j+1}^{d} \mathsf{Cov}\left(\beta_{k,i},\beta_{k,j}\right))\,\mu_{q(\nu_k)}.
\end{aligned}$$

Thus, each element of $\mathbf{C}_\vartheta$ is given by

$$[\mathbf{C}_\vartheta]_{i,j} = \sum_{k=j+1}^{d} \mathsf{Cov}(\beta_{k,i},\beta_{k,j})\mu_{q(\nu_k)} = [\mathbf{C}_\vartheta]_{j,i}\,.$$

Take the exponential of the $\log q^*(\boldsymbol{\vartheta})$ derived above and notice that it coincides with the kernel of a Gaussian random variable $\mathsf{N}_{dp}(\boldsymbol{\mu}_{q(\vartheta)}, \boldsymbol{\Sigma}_{q(\vartheta)})$, as in Proposition A.3. □

Proposition A.4. *The optimal variational density for the row-parameter $\boldsymbol{\vartheta}_j$ is equal to a multivariate Gaussian $q^*(\boldsymbol{\vartheta}_j) \equiv \mathsf{N}_p(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$, where:*

$$\begin{aligned}
\boldsymbol{\Sigma}_{q(\vartheta_j)} &= \left(\boldsymbol{\mu}_{q(\omega_{j,j})}\sum_{t=1}^{T}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal + 1/\upsilon\mathbf{I}_p\right)^{-1}, \\
\boldsymbol{\mu}_{q(\vartheta_j)} &= \boldsymbol{\Sigma}_{q(\vartheta_j)}\left(\sum_{t=1}^{T}\left(\boldsymbol{\mu}_{q(\omega_j)}\otimes\mathbf{z}_{t-1}\right)\mathbf{y}_t - \left(\boldsymbol{\mu}_{q(\omega_{j,-j})}\otimes\sum_{t=1}^{T}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal\right)\boldsymbol{\mu}_{q(\vartheta_{-j})}\right).
\end{aligned} \tag{A.4}$$

Under this setting the vector $\mathbf{k}_{\vartheta,t}$ computed for $q^*(\nu_j)$ and $q^*(\boldsymbol{\beta}_j)$ is a null vector since the independence among rows of $\boldsymbol{\Theta}$ is assumed.

Proof. Consider the setting as in Proposition A.3, define $\boldsymbol{\mu}_{q(\boldsymbol{\Omega})} = \mathbb{E}_{-\vartheta}\left[(\mathbf{I}_d - \mathbf{B})^\intercal \mathbf{V}(\mathbf{I}_d - \mathbf{B})\right]$ the expectation of the precision matrix and compute the optimal variational density for the parameter $\boldsymbol{\vartheta}_j$ as $\log q^*(\boldsymbol{\vartheta}_j) \propto \mathbb{E}_{-\vartheta_j}\left[\ell(\boldsymbol{\xi};\mathbf{y},\mathbf{x}) + \log p(\boldsymbol{\vartheta}_j)\right]$:

$$\begin{aligned}
\log q^*(\boldsymbol{\vartheta}_j) \propto & -\frac{1}{2}\mathbb{E}_{-\boldsymbol{\vartheta}_j}\left[\boldsymbol{\vartheta}\right]^\intercal \left(\boldsymbol{\mu}_{q(\boldsymbol{\Omega})} \otimes \sum_{t=1}^{T} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal\right) \mathbb{E}_{-\boldsymbol{\vartheta}_j}\left[\boldsymbol{\vartheta}\right] - \frac{1}{2\upsilon}\boldsymbol{\vartheta}_j^\intercal\boldsymbol{\vartheta}_j \\
& + \mathbb{E}_{-\boldsymbol{\vartheta}_j}\left[\boldsymbol{\vartheta}\right]^\intercal \sum_{t=1}^{T}\left(\boldsymbol{\mu}_{q(\boldsymbol{\Omega})} \otimes \mathbf{z}_{t-1}\right)\mathbf{y}_t \\
\propto & -\frac{1}{2}\boldsymbol{\vartheta}_j^\intercal \left(\boldsymbol{\mu}_{q(\omega_{j,j})} \sum_{t=1}^{T} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal\right)\boldsymbol{\vartheta}_j - \frac{1}{2\upsilon}\boldsymbol{\vartheta}_j^\intercal\boldsymbol{\vartheta}_j \\
& + \boldsymbol{\vartheta}_j^\intercal \sum_{t=1}^{T}\left(\boldsymbol{\mu}_{q(\omega_j)} \otimes \mathbf{z}_{t-1}\right)\mathbf{y}_t - \boldsymbol{\vartheta}_j^\intercal\left(\boldsymbol{\mu}_{q(\omega_{j,-j})} \otimes \sum_{t=1}^{T}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal\right)\boldsymbol{\mu}_{q(\vartheta_{-j})}.
\end{aligned}$$

Where we used the following partitions:

$$\boldsymbol{\vartheta} = \begin{pmatrix} \boldsymbol{\vartheta}_j \\ \boldsymbol{\vartheta}_{-j} \end{pmatrix}, \qquad \boldsymbol{\Omega} = \begin{pmatrix} \omega_{j,j} & \boldsymbol{\omega}_{j,-j} \\ \boldsymbol{\omega}_{-j,j} & \boldsymbol{\Omega}_{-j,-j} \end{pmatrix},$$

and we denote with $\boldsymbol{\omega}_j$ the j-th row of $\boldsymbol{\Omega}$. Re-arrange the terms, take the exponential of the $\log q^*(\boldsymbol{\vartheta}_j)$ derived above and notice that it coincides with the kernel of a Gaussian random variable $\mathsf{N}_p(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$, as defined in Proposition A.4. □

Proposition A.5. *The variational lower bound for the non-sparse multivariate regression model can be derived analytically and it is equal to:*

$$\begin{aligned}
\log \underline{p}(\mathbf{y};q) = & \, d\left(-\frac{T}{2}\log 2\pi + a_\nu \log b_\nu - \log\Gamma(a_\nu)\right) - \sum_{j=1}^{d}\left(a_{q(\nu_j)}\log b_{q(\nu_j)} - \log\Gamma(a_{q(\nu_j)})\right) \\
& - \frac{1}{2}\sum_{j=2}^{d}\sum_{k=1}^{j-1}\left(\log\tau + 1/\tau\mu_{q(\beta_{j,k}^2)}\right) + \frac{1}{2}\sum_{j=2}^{d}\left(\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}| + (j-1)\right) \\
& - \frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{p}\left(\log\upsilon + 1/\upsilon\mu_{q(\vartheta_{j,k}^2)}\right) + \frac{1}{2}\left(\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\vartheta})}| + dp\right).
\end{aligned} \tag{A.5}$$

Proof. First of all, notice that the lower bound can be written in terms of expected values with respect to the density q as:

$$\log \underline{p}(\mathbf{y}; q) = \int q(\boldsymbol{\xi}) \log \frac{p(\boldsymbol{\xi}, \mathbf{y})}{q(\boldsymbol{\xi})} \, d\boldsymbol{\xi} = \mathbb{E}_q \left[\log p(\boldsymbol{\xi}, \mathbf{y}) \right] - \mathbb{E}_q \left[\log q(\boldsymbol{\xi}) \right],$$

where $\log p(\boldsymbol{\xi}, \mathbf{y}) = \ell(\boldsymbol{\xi}; \mathbf{y}) + \log p(\boldsymbol{\xi})$. Following our model specification, we have that

$$\log p(\boldsymbol{\xi}, \mathbf{y}) = \sum_{j=1}^{d} \left(\ell_j(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}) + \log p(\nu_j) \right) + \sum_{j=2}^{d} \log p(\boldsymbol{\beta}_j) + \log p(\boldsymbol{\vartheta}),$$

where $\ell_j(\boldsymbol{\vartheta}; \mathbf{y}, \mathbf{x})$ denotes the log-likelihood for the j-th variable:

$$\ell_j(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}) = -\frac{T}{2} \log 2\pi + \frac{T}{2} \log \nu_j - \frac{\nu_j}{2} \sum_{t=1}^{T} \left(y_{j,t} - \boldsymbol{\beta}_j \mathbf{r}_{j,t} - \boldsymbol{\vartheta}_j \mathbf{z}_{t-1} \right)^2.$$

For the variational density we have $\log q(\boldsymbol{\xi}) = \sum_{j=1}^{d} \log q(\nu_j) + \sum_{j=2}^{d} \log q(\boldsymbol{\beta}_j) + \log q(\boldsymbol{\vartheta})$, and the lower bound can be divided into terms referring to each parameter:

$$\begin{aligned} \log \underline{p}(\mathbf{y}; q) &= \sum_{j=1}^{d} \mathbb{E}_q \left[\ell_j(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}) + \log p(\nu_j) - \log q(\nu_j) \right] \\ &\quad + \sum_{j=2}^{d} \mathbb{E}_q \left[\log p(\boldsymbol{\beta}_j) - \log q(\boldsymbol{\beta}_j) \right] + \mathbb{E}_q \left[\log p(\boldsymbol{\vartheta}) - \log q(\boldsymbol{\vartheta}) \right] \\ &= \sum_{j=1}^{d} \big(\underbrace{\mathbb{E}_q \left[\ell_j(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}) + \log \underline{p}(\mathbf{y}; \nu_j) \right]}_{A} + \sum_{j=2}^{d} \underbrace{\mathbb{E}_q \left[\log \underline{p}(\mathbf{y}; \boldsymbol{\beta}_j) \right]}_{B} + \underbrace{\mathbb{E}_q \left[\log \underline{p}(\mathbf{y}; \boldsymbol{\vartheta}) \right]}_{C}, \end{aligned} \tag{A.6}$$

thus our strategy will be to evaluate each piece in the latter separately and then put the results together.

The first part of the lower bound we compute is $A = \ell_j(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}) + \log \underline{p}(\mathbf{y}; \nu_j)$:

$$\begin{aligned} A &= \mathbb{E}_q \left[-\frac{T}{2} \log 2\pi + \frac{T}{2} \log \nu_j - \frac{\nu_j}{2} \sum_{t=1}^{T} \left(y_{j,t} - \boldsymbol{\beta}_j \mathbf{r}_{j,t} - \boldsymbol{\vartheta}_j \mathbf{z}_{t-1} \right)^2 \right] \\ &\quad + \mathbb{E}_q \left[a_\nu \log b_\nu - \log \Gamma(a_\nu) + (a_\nu - 1) \log \nu_j - \nu_j b_\nu \right] \\ &\quad - \mathbb{E}_q \left[a_{q(\nu_j)} \log b_{q(\nu_j)} - \log \Gamma(a_{q(\nu_j)}) + (a_{q(\nu_j)} - 1) \log \nu_j - \nu_j b_{q(\nu_j)} \right] \\ &= -\frac{T}{2} \log 2\pi + \frac{T}{2} \mu_{q(\log \nu_j)} - \frac{\mu_{q(\nu_j)}}{2} \sum_{t=1}^{T} \mathbb{E}_q \left[\varepsilon_{j,t}^2 \right] \\ &\quad + a_\nu \log b_\nu - \log \Gamma(a_\nu) + (a_\nu - 1) \mu_{q(\log \nu_j)} - \mu_{q(\nu_j)} b_\nu \\ &\quad - a_{q(\nu_j)} \log b_{q(\nu_j)} + \log \Gamma(a_{q(\nu_j)}) - (a_{q(\nu_j)} - 1) \mu_{q(\log \nu_j)} + \mu_{q(\nu_j)} b_{q(\nu_j)} \end{aligned}$$

$$= -\frac{T}{2}\log 2\pi + a_\nu \log b_\nu - \log\Gamma(a_\nu) - a_{q(\nu_j)}\log b_{q(\nu_j)} + \log\Gamma(a_{q(\nu_j)}),$$

where we exploit the definitions of $\mathbb{E}_q\left[\varepsilon^2_{j,t}\right], a_{q(\nu_j)}, b_{q(\nu_j)}$ given in Proposition A.1. The second term to compute is equal to:

$$\begin{aligned}
B &= \mathbb{E}_q\left[-\frac{j-1}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{j-1}\log\tau - \frac{1}{2\tau}\sum_{k=1}^{j-1}\beta^2_{j,k}\right] \\
&\quad - \mathbb{E}_q\left[-\frac{j-1}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}| - \frac{1}{2}\overbrace{(\boldsymbol{\beta}_j - \boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)})\boldsymbol{\Sigma}^{-1}_{q(\boldsymbol{\beta}_j)}(\boldsymbol{\beta}_j - \boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)})^\intercal}^{\text{See Result A.3}}\right] \\
&= -\frac{1}{2}\sum_{k=1}^{j-1}\log\tau - \frac{1}{2\tau}\sum_{k=1}^{j-1}\mu_{q(\beta^2_{j,k})} + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}| + \frac{j-1}{2},
\end{aligned}$$

where $\mu_{q(\beta^2_{j,k})} = \mu^2_{q(\beta_{j,k})} + \sigma^2_{q(\beta_{j,k})}$ and $\sigma^2_{q(\beta_{j,k})}$ denotes the k-th element on the diagonal of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}$. To conclude, we compute the last term:

$$\begin{aligned}
C &= \mathbb{E}_q\left[-\frac{dp}{2}\log 2\pi - \frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{p}\log\upsilon - \frac{1}{2\upsilon}\sum_{j=1}^{d}\sum_{k=1}^{p}\vartheta^2_{j,k}\right] \\
&\quad - \mathbb{E}_q\left[-\frac{dp}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\vartheta)}| - \frac{1}{2}\overbrace{(\boldsymbol{\vartheta} - \boldsymbol{\mu}_{q(\vartheta)})^\intercal\boldsymbol{\Sigma}^{-1}_{q(\vartheta)}(\boldsymbol{\vartheta} - \boldsymbol{\mu}_{q(\vartheta)})}^{\text{See Result A.3}}\right] \\
&= -\frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{p}\log\upsilon - \frac{1}{2\upsilon}\sum_{j=1}^{d}\sum_{k=1}^{p}\mu_{q(\vartheta^2_{j,k})} + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\vartheta)}| + \frac{dp}{2}.
\end{aligned}$$

Put together the terms A, B, C as in (A.6) and notice that the variational lower bound here computed coincides with the one presented in Proposition A.5. □

A.2.2 Bayesian adaptive lasso

In order to induce shrinkage towards zero in the estimates of the coefficients $\boldsymbol{\vartheta}$, we assume an adaptive lasso prior. Notice that the optimal densities for the variances ν_j and for the cholesky factor rows $\boldsymbol{\beta}_j$ remain exactly the same computed in Section A.2.1. The changes in the optimal densities $q^*(\boldsymbol{\vartheta})$ consist in the fact that now the prior variances are no more fixed, but random variables themselves.

Proposition A.6. *The joint optimal variational density for the parameter $\boldsymbol{\vartheta}$ is equal to $q^*(\boldsymbol{\vartheta}) \equiv \mathsf{N}_{dp}(\boldsymbol{\mu}_{q(\vartheta)}, \boldsymbol{\Sigma}_{q(\vartheta)})$, where:*

$$\boldsymbol{\Sigma}_{q(\vartheta)} = \left(\boldsymbol{\mu}_{q(\boldsymbol{\Omega})} \otimes \sum_{t=1}^{T} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^{\intercal} + \mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon)})\right)^{-1}, \quad \boldsymbol{\mu}_{q(\vartheta)} = \boldsymbol{\Sigma}_{q(\vartheta)} \sum_{t=1}^{T} \left(\boldsymbol{\mu}_{q(\boldsymbol{\Omega})} \otimes \mathbf{z}_{t-1}\right) \mathbf{y}_t, \tag{A.7}$$

where $\mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon)})$ is a diagonal matrix with elements $\boldsymbol{\mu}_{q(1/\upsilon)} = (\mu_{q(1/\upsilon_{1,1})}, \dots, \mu_{q(1/\upsilon_{d,p})})$.

Under the row-independence assumption, the optimal variational density for the parameter $\boldsymbol{\vartheta}_j$ is equal to $q^*(\boldsymbol{\vartheta}_j) \equiv \mathsf{N}_p(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$, where:

$$\begin{aligned}
\boldsymbol{\Sigma}_{q(\vartheta_j)} &= \left(\boldsymbol{\mu}_{q(\omega_{j,j})} \sum_{t=1}^{T} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^{\intercal} + \mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon_j)})\right)^{-1}, \\
\boldsymbol{\mu}_{q(\vartheta_j)} &= \boldsymbol{\Sigma}_{q(\vartheta_j)} \left(\sum_{t=1}^{T} \left(\boldsymbol{\mu}_{q(\omega_j)} \otimes \mathbf{z}_{t-1}\right) \mathbf{y}_t - \left(\boldsymbol{\mu}_{q(\omega_{j,-j})} \otimes \sum_{t=1}^{T} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^{\intercal}\right) \boldsymbol{\mu}_{q(\vartheta_{-j})}\right),
\end{aligned} \tag{A.8}$$

where $\mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon_j)})$ is a diagonal matrix where $\boldsymbol{\mu}_{q(1/\upsilon_j)} = (\mu_{q(1/\upsilon_{j,1})}, \dots, \mu_{q(1/\upsilon_{j,p})})$.
Hereafter we describe the optimal densities for the parameters used in hierarchical specification of the prior here assumed.

Proposition A.7. *The optimal density for the prior variance $1/\upsilon_{j,k}$ is equal to an inverse Gaussian distribution $q^*(1/\upsilon_{j,k}) \equiv \mathsf{IG}(a_{q(1/\upsilon_{j,k})}, b_{q(1/\upsilon_{j,k})})$, where, for each $j = 1, \dots, d$ and $k = 1, \dots, p$:*

$$a_{q(1/\upsilon_{j,k})} = \mu_{q(\vartheta_{j,k}^2)}, \quad b_{q(1/\upsilon_{j,k})} = \mu_{q(\lambda_{j,k}^2)}. \tag{A.9}$$

Moreover, it is useful to know that

$$\mu_{q(1/\upsilon_{j,k})} = \sqrt{b_{q(1/\upsilon_{j,k})}/a_{q(1/\upsilon_{j,k})}}, \quad \mu_{q(\upsilon_{j,k})} = \sqrt{a_{q(1/\upsilon_{j,k})}/b_{q(1/\upsilon_{j,k})}} + 1/b_{q(1/\upsilon_{j,k})}.$$

Proof. Consider the prior specification which involves the parameter $\upsilon_{j,k}$:

$$\vartheta_{j,k}|\upsilon_{j,k} \sim \mathsf{N}(0, \upsilon_{j,k}), \qquad \upsilon_{j,k}|\lambda_{j,k}^2 \sim \mathsf{Exp}\left(\lambda_{j,k}^2/2\right).$$

Compute the optimal variational density $\log q^*(\upsilon_{j,k}) \propto \mathbb{E}_{-\upsilon_{\mathbf{j},\mathbf{k}}}\left[\log p(\vartheta_{j,k}) + \log p(\upsilon_{j,k})\right]$:

$$\begin{aligned}
\log q^*(\upsilon_{j,k}) &\propto \mathbb{E}_{-\upsilon_{j,k}}\left[-\frac{1}{2}\log \upsilon_{j,k} - \frac{1}{2\upsilon_{j,k}}\vartheta_{j,k}^2 - \upsilon_{j,k}\frac{\lambda_{j,k}^2}{2}\right] \\
&\propto -1/2 \log \upsilon_{j,k} - \frac{1}{2\upsilon_{j,k}}\mu_{q(\vartheta_{j,k}^2)} - \upsilon_{j,k}\frac{\mu_{q(\lambda_{j,k}^2)}}{2},
\end{aligned}$$

and, as a consequence, we obtain:

$$\log q^*(1/\upsilon_{j,k}) \propto -3/2\log(1/\upsilon_{j,k}) - \frac{1}{2}(1/\upsilon_{j,k})\mu_{q(\vartheta^2_{j,k})} - \frac{\mu_{q(\lambda^2_{j,k})}}{2(1/\upsilon_{j,k})}.$$

Take the exponential and notice that the latter is the kernel of an inverse Gaussian random variable $\mathsf{IG}(a_{q(1/\upsilon_{j,k})}, b_{q(1/\upsilon_{j,k})})$, as defined in Proposition A.7. □

Proposition A.8. *The optimal density for the latent parameter $\lambda^2_{j,k}$ for $j = 1, \dots, d$ and $k = 1, \dots, p$ is equal to a $q^*(\lambda^2_{j,k}) \equiv \mathsf{Ga}(a_{q(\lambda^2_{j,k})}, b_{q(\lambda^2_{j,k})})$, where:*

$$a_{q(\lambda^2_{j,k})} = h_1 + 1, \quad b_{q(\lambda^2_{j,k})} = \mu_{q(\upsilon_{j,k})}/2 + h_2. \tag{A.10}$$

Proof. Consider the prior specification which involves the parameter $\lambda^2_{j,k}$:

$$\upsilon_{j,k}|\lambda^2_{j,k} \sim \mathsf{Exp}\left(\lambda^2_{j,k}/2\right), \quad \lambda^2_{j,k} \sim \mathsf{Ga}(h_1, h_2).$$

Compute the optimal variational density as $\log q^*(\lambda^2_{j,k}) \propto \mathbb{E}_{-\boldsymbol{\lambda}^{\mathbf{2}}_{\mathbf{j,k}}}\left[\log p(\upsilon_{j,k}) + \log p(\lambda^2_{j,k})\right]$:

$$\begin{aligned}\log q^*(\lambda^2_{j,k}) &\propto \mathbb{E}_{-\lambda^2_{j,k}}\left[h_1 \log \lambda^2_{j,k} - \lambda^2_{j,k}\left(\upsilon_{j,k}/2 + h_2\right)\right] \\ &\propto h_1 \log \lambda^2_{j,k} - \lambda^2_{j,k}\left(\mu_{q(\upsilon_{j,k})}/2 + h_2\right),\end{aligned}$$

then take the exponential and notice that the latter is the kernel of a gamma random variable $\mathsf{Ga}(a_{q(\lambda^2_{j,k})}, b_{q(\lambda^2_{j,k})})$, as defined in Proposition A.8. □

Proposition A.9. *The variational lower bound for the multivariate regression model with adaptive Bayesian lasso prior can be derived analytically and it is equal to:*

$$\begin{aligned}\log \underline{p}(\mathbf{y}; q) = d&\left(-\frac{T}{2}\log 2\pi + a_\nu \log b_\nu - \log\Gamma(a_\nu)\right) - \sum_{j=1}^{d}\left(a_{q(\nu_j)}\log b_{q(\nu_j)} - \log\Gamma(a_{q(\nu_j)})\right) \\ &- \frac{1}{2}\sum_{j=2}^{d}\sum_{k=1}^{j-1}\left(\log\tau + 1/\tau\mu_{q(\beta^2_{j,k})}\right) + \frac{1}{2}\sum_{j=2}^{d}\left(\log|\mathbf{\Sigma}_{q(\boldsymbol{\beta}_j)}| + (j-1)\right) \\ &+ \frac{1}{2}\left(\log|\mathbf{\Sigma}_{q(\boldsymbol{\vartheta})}| + dp\right) + \sum_{j=1}^{d}\sum_{k=1}^{p}\frac{1}{2}\mu_{q(\lambda^2_{j,k})}\mu_{q(\upsilon_{j,k})} \\ &- \sum_{j=1}^{d}\sum_{k=1}^{p}(1/4\log(b_{q(1/\upsilon_{j,k})}/a_{q(1/\upsilon_{j,k})}) - \log K_{1/2}(\sqrt{b_{q(1/\upsilon_{j,k})}a_{q(1/\upsilon_{j,k})}})) \\ &+ dp\left(h_1\log h_2 - \log\Gamma(h_1)\right) - \sum_{j=1}^{d}\sum_{k=1}^{p}\left(a_{q(\lambda^2_{j,k})}\log b_{q(\lambda^2_{j,k})} - \log\Gamma(a_{q(\lambda^2_{j,k})})\right).\end{aligned} \tag{A.11}$$

Proof. As we did in (A.6) for Proposition A.5, the lower bound can be divided into terms referring to each parameter:

$$\log \underline{p}(\mathbf{y}; q) = A + \sum_{j=1}^{d} \sum_{k=1}^{p} \left(\underbrace{\mathbb{E}_q \left[\log \underline{p}(\mathbf{y}; \upsilon_{j,k}) \right]}_{B} + \underbrace{\mathbb{E}_q \left[\log \underline{p}(\mathbf{y}; \lambda^2_{j,k}) \right]}_{C} \right),$$

where A is equal to (A.6) in the previous non-informative model specification. Our strategy will be to evaluate each piece in the latter separately and then put the results together. Notice that the computations for the piece A are already available from Proposition A.5 and they are equal to the lower bound for the model with the non-informative prior where we still have to take the expectations with respect to the latent parameters $\upsilon_{j,k}$. Thus, we have that:

$$\begin{aligned} A = d \left(-\frac{T}{2} \log 2\pi + a_\nu \log b_\nu - \log \Gamma(a_\nu) \right) - \sum_{j=1}^{d} \left(a_{q(\nu_j)} \log b_{q(\nu_j)} - \log \Gamma(a_{q(\nu_j)}) \right) \\ - \frac{1}{2} \sum_{j=2}^{d} \sum_{k=1}^{j-1} \left(\log \tau + 1/\tau \mu_{q(\beta^2_{j,k})} \right) + \frac{1}{2} \sum_{j=2}^{d} \left(\log |\mathbf{\Sigma}_{q(\boldsymbol{\beta}_j)}| + (j-1) \right) \\ - \frac{1}{2} \sum_{j=1}^{d} \sum_{k=1}^{p} \left(\mu_{q(\log \upsilon_{j,k})} + \mu_{q(1/\upsilon_{j,k})} \mu_{q(\vartheta^2_{j,k})} \right) + \frac{1}{2} \left(\log |\mathbf{\Sigma}_{q(\boldsymbol{\vartheta})}| + dp \right). \end{aligned}$$

Consider now the piece B and recall that, since $q^*(1/\upsilon_{j,k}) \equiv \mathsf{IG}(a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})})$, then its inverse follows $q^*(\upsilon_{j,k}) \equiv \mathsf{GIG}(1/2, b_{q(1/\upsilon_{j,k})}, a_{q(1/\upsilon_{j,k})})$. We have that

$$\begin{aligned} B &= \mathbb{E}_q \left[\log \lambda^2_{j,k} - \log 2 - \upsilon_{j,k} \frac{\lambda^2_{j,k}}{2} \right] \\ &\quad - \mathbb{E}_q \left[h(1/2, b_{q(1/\upsilon_{j,k})}, a_{q(1/\upsilon_{j,k})}) - 1/2 \log \upsilon_{j,k} - \frac{1}{2} \left(b_{q(1/\upsilon_{j,k})} \upsilon_{j,k} + \frac{a_{q(1/\upsilon_{j,k})}}{\upsilon_{j,k}} \right) \right] \\ &= \mu_{q(\log \lambda^2_{j,k})} - \log 2 - h(1/2, b_{q(1/\upsilon_{j,k})}, b_{q(1/\upsilon_{j,k})}) + 1/2 \mu_{q(\log \upsilon_{j,k})} \\ &\quad - \frac{1}{2} \left(\mu_{q(\upsilon_{j,k})} \mu_{q(\lambda^2_{j,k})} - b_{q(1/\upsilon_{j,k})} \mu_{q(\upsilon_{j,k})} - a_{q(1/\upsilon_{j,k})} \mu_{q(1/\upsilon_{j,k})} \right), \end{aligned}$$

where $h(\zeta, a, b)$ denotes the logarithm of the normalizing constant of a GIG distribution, i.e.,

$$h(\zeta, a, b) = \zeta/2 \log(a/b) - \log 2 - \log K_\zeta(\sqrt{ab}).$$

The term involving $\lambda^2_{j,k}$, for $j = 1, \dots, d$ and $k = 1, \dots, p$, is equal to:

$$\begin{aligned}
C &= \mathbb{E}_q \left[h_1 \log h_2 - \log \Gamma(h_1) + (h_1 - 1) \log \lambda^2_{j,k} - \lambda^2_{j,k} h_2 \right] \\
&\quad - \mathbb{E}_q \left[a_{q(\lambda^2_{j,k})} \log b_{q(\lambda^2_{j,k})} - \log \Gamma(a_{q(\lambda^2_{j,k})}) + (a_{q(\lambda^2_{j,k})} - 1) \log \lambda^2_{j,k} - \lambda^2_{j,k} b_{q(\lambda^2_{j,k})} \right] \\
&= h_1 \log h_2 - \log \Gamma(h_1) + (h_1 - 1) \mu_{q(\log \lambda^2_{j,k})} - \mu_{q(\lambda^2_{j,k})} h_2 \\
&\quad - a_{q(\lambda^2_{j,k})} \log b_{q(\lambda^2_{j,k})} + \log \Gamma(a_{q(\lambda^2_{j,k})}) - (a_{q(\lambda^2_{j,k})} - 1) \mu_{q(\log \lambda^2_{j,k})} + \mu_{q(\lambda^2_{j,k})} b_{q(\lambda^2_{j,k})}.
\end{aligned}$$

Group together the terms and exploit the analytical form of the optimal parameters to perform some simplifications. The remaining terms form the lower bound for a multivariate regression model with adaptive lasso prior. □

A.2.3 Adaptive normal-gamma

In order to induce shrinkage towards zero in the estimates of the coefficients, we assume an adaptive normal-gamma prior on $\boldsymbol{\vartheta}$. Notice that the optimal densities for the variances ν_j and for $\boldsymbol{\beta}_j$ remain exactly the same computed in Section A.2.1. The optimal density $q^*(\boldsymbol{\vartheta})$ has the same structure as the one computed in Proposition (A.6) for the lasso prior.

Hereafter we describe the optimal densities for the parameters used in hierarchical specification of the normal-gamma prior.

Proposition A.10. *The optimal density for the prior variance $\upsilon_{j,k}$ is equal to a generalized inverse Gaussian distribution $q^*(\upsilon_{j,k}) \equiv \mathsf{GIG}(\zeta_{q(\upsilon_{j,k})}, a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})})$, where, for $j = 1, \dots, d$ and $k = 1, \dots, p$:*

$$\zeta_{q(\upsilon_{j,k})} = \mu_{q(\eta_j)} - 1/2, \quad a_{q(\upsilon_{j,k})} = \mu_{q(\eta_j)} \mu_{q(\lambda_{j,k})}, \quad b_{q(\upsilon_{j,k})} = \mu_{q(\vartheta^2_{j,k})}. \tag{A.12}$$

Moreover, it is useful to know that

$$\begin{aligned}
\mu_{q(\upsilon_{j,k})} &= \frac{\sqrt{b_{q(\upsilon_{j,k})}} K_{\zeta_{q(\upsilon_{j,k})}+1} \left(\sqrt{a_{q(\upsilon_{j,k})} b_{q(\upsilon_{j,k})}} \right)}{\sqrt{a_{q(\upsilon_{j,k})}} K_{\zeta_{q(\upsilon_{j,k})}} \left(\sqrt{a_{q(\upsilon_{j,k})} b_{q(\upsilon_{j,k})}} \right)}, \\
\mu_{q(1/\upsilon_{j,k})} &= \frac{\sqrt{a_{q(\upsilon_{j,k})}} K_{\zeta_{q(\upsilon_{j,k})}+1} \left(\sqrt{a_{q(\upsilon_{j,k})} b_{q(\upsilon_{j,k})}} \right)}{\sqrt{b_{q(\upsilon_{j,k})}} K_{\zeta_{q(\upsilon_{j,k})}} \left(\sqrt{a_{q(\upsilon_{j,k})} b_{q(\upsilon_{j,k})}} \right)} - \frac{2\zeta_{q(\upsilon_{j,k})}}{b_{q(\upsilon_{j,k})}}, \\
\mu_{q(\log \upsilon_{j,k})} &= \log \frac{\sqrt{b_{q(\upsilon_{j,k})}}}{\sqrt{a_{q(\upsilon_{j,k})}}} + \frac{\partial}{\partial \zeta_{q(\upsilon_{j,k})}} \log K_{\zeta_{q(\upsilon_{j,k})}} \left(\sqrt{a_{q(\upsilon_{j,k})} b_{q(\upsilon_{j,k})}} \right),
\end{aligned}$$

where $K_\zeta(\cdot)$ denotes the modified Bessel function of second kind.

Proof. Consider the prior specification which involves the parameter $\upsilon_{j,k}$:

$$\vartheta_{j,k}|\upsilon_{j,k} \sim \mathsf{N}(0, \upsilon_{j,k}), \qquad \upsilon_{j,k}|\eta_j, \lambda_{j,k} \sim \mathsf{Ga}\left(\eta_j, \frac{\eta_j \lambda_{j,k}}{2}\right).$$

Compute the optimal variational density as $\log q^*(\upsilon_{j,k}) \propto \mathbb{E}_{-\upsilon_{\mathbf{j,k}}}[\log p(\vartheta_{j,k}) + \log p(\upsilon_{j,k})]$:

$$\begin{aligned}\log q^*(\upsilon_{j,k}) &\propto \mathbb{E}_{-\upsilon_{j,k}}\left[-\frac{1}{2}\log \upsilon_{j,k} - \frac{1}{2\upsilon_{j,k}}\beta_{j,k}^2 + (\eta_j - 1)\log \upsilon_{j,k} - \upsilon_{j,k}\frac{\eta_j \lambda_{j,k}}{2}\right] \\ &\propto \left(\mu_{q(\eta_j)} - \frac{1}{2} - 1\right)\log \upsilon_{j,k} - \frac{1}{2\upsilon_{j,k}}\mu_{q(\vartheta_{j,k}^2)} - \upsilon_{j,k}\frac{\mu_{q(\eta_j)}\mu_{q(\lambda_{j,k})}}{2},\end{aligned}$$

where $\mu_{q(\vartheta_{j,k}^2)} = \sigma^2_{q(\vartheta_{j,k})} + \mu^2_{q(\vartheta_{j,k})}$. Take the exponential and notice that the latter is the kernel of a generalized inverse Gaussian random variable $\mathsf{GIG}(\zeta_{q(\upsilon_{j,k})}, a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})})$, as defined in Proposition A.10. □

Proposition A.11. *The optimal density for the latent parameter $\lambda_{j,k}$ for $j = 1, \ldots, d$ and $k = 1, \ldots, p$ is equal to a $q^*(\lambda_{j,k}) \equiv \mathsf{Ga}(a_{q(\lambda_{j,k})}, b_{q(\lambda_{j,k})})$, where:*

$$a_{q(\lambda_{j,k})} = \mu_{q(\eta_j)} + h_1, \quad b_{q(\lambda_{j,k})} = \frac{\mu_{q(\eta_j)}\mu_{q(\upsilon_{j,k})}}{2} + h_2. \tag{A.13}$$

Moreover, it is useful to know that

$$\mu_{q(\lambda_{j,k})} = \frac{a_{q(\lambda_{j,k})}}{b_{q(\lambda_{j,k})}}, \quad \mu_{q(\log \lambda_{j,k})} = -\log b_{q(\lambda_{j,k})} + \frac{\Gamma'(a_{q(\lambda_{j,k})})}{\Gamma(a_{q(\lambda_{j,k})})}.$$

Proof. Consider the prior specification which involves the parameter $\lambda_{j,k}$:

$$\upsilon_{j,k}|\eta_j, \lambda_{j,k} \sim \mathsf{Ga}\left(\eta_j, \frac{\eta_j \lambda_{j,k}}{2}\right), \quad \lambda_{j,k} \sim \mathsf{Ga}(h_1, h_2).$$

Compute the optimal variational density as $\log q^*(\lambda_{j,k}) \propto \mathbb{E}_{-\lambda_{\mathbf{j,k}}}[\log p(\upsilon_{j,k}) + \log p(\lambda_{j,k})]$:

$$\begin{aligned}\log q^*(\lambda_{j,k}) &\propto \mathbb{E}_{-\lambda_{j,k}}\left[(\eta_j + h_1 - 1)\log \lambda_{j,k} - \lambda_{j,k}\left(\frac{\eta_j \upsilon_{j,k}}{2} + h_2\right)\right] \\ &\propto \left(\mu_{q(\eta_j)} + h_1 - 1\right)\log \lambda_{j,k} - \lambda_{j,k}\left(\frac{\mu_{q(\eta_j)}\mu_{q(\upsilon_{j,k})}}{2} + h_2\right),\end{aligned} \tag{A.14}$$

then take the exponential and notice that the latter is the kernel of a gamma random variable $\mathsf{Ga}(a_{q(\lambda_{j,k})}, b_{q(\lambda_{j,k})})$, as defined in Proposition A.11. □

Proposition A.12. *The optimal density for the latent parameter* η_j *for* $j = 1, \dots, d$ *is equal to:*

$$q^*(\eta_j) = \frac{h(\eta_j)}{c_{\eta_j}} \exp\left\{ -\eta_j \sum_{k=1}^{p} \left(\frac{\mu_{q(\lambda_{j,k})}\mu_{q(\upsilon_{j,k})}}{2} - \mu_{q(\log\lambda_{j,k})} - \mu_{q(\log\upsilon_{j,k})} + \log 2 + h_3 \right) \right\}, \tag{A.15}$$

where $\log h(\eta_j) = p(\eta_j \log \eta_j - \log\Gamma(\eta_j))$ *and*

$$c_{\eta_j} = \int_{\mathbb{R}+} h(\eta_j) \exp\left\{ -\eta_j \sum_{k=1}^{p} \left(\frac{\mu_{q(\lambda_{j,k})}\mu_{q(\upsilon_{j,k})}}{2} - \mu_{q(\log\lambda_{j,k})} - \mu_{q(\log\upsilon_{j,k})} + p\log 2 + h_3 \right) \right\} d\eta_j.$$

Then, we have that $\mu_{q(\eta_j)} = \int_{\mathbb{R}+} \eta_j q^*(\eta_j)\, d\eta_j$.

Proof. Consider the prior specification which involves the parameter η_j:

$$\upsilon_{j,k}|\eta_j, \lambda_{j,k} \sim \mathsf{Ga}\left(\eta_j, \frac{\eta_j\lambda_{j,k}}{2}\right), \quad \eta_j \sim \mathsf{Exp}(h_3).$$

Compute the optimal variational density as $\log q^*(\eta_j) \propto \mathbb{E}_{-\eta_\mathrm{j}}\left[\sum_{k=1}^{p} \log p(\upsilon_{j,k}) + \log p(\eta_j)\right]$:

$$\begin{aligned}\log q^*(\eta_j) \propto \mathbb{E}_{-\eta_j}&\left[p\left(\eta_j \log\eta_j - \log\Gamma(\eta_j)\right) - \eta_j \sum_{k=1}^{p} \left(\left(\frac{\lambda_{j,k}\upsilon_{j,k}}{2} - \log\frac{\lambda_{j,k}\upsilon_{j,k}}{2} \right) + h_3 \right) \right] \\ &= p\left(\eta_j \log\eta_j - \log\Gamma(\eta_j)\right) \\ &\quad - \eta_j \sum_{k=1}^{p} \left(\frac{\mu_{q(\lambda_{j,k})}\mu_{q(\upsilon_{j,k})}}{2} - \mathbb{E}_{\upsilon_{j,k}\lambda_{j,k}}\left[\log\frac{\lambda_{j,k}\upsilon_{j,k}}{2}\right] + h_3 \right),\end{aligned} \tag{A.16}$$

which is not the kernel of a know distribution, but since $\mathbb{E}[\log x] \leq \log\mathbb{E}[x] < \mathbb{E}[x]$, it holds that

$$\frac{\mu_{q(\lambda_{j,k})}\mu_{q(\upsilon_{j,k})}}{2} > \mathbb{E}_{\upsilon_{j,k}\lambda_{j,k}}\left[\log\frac{\lambda_{j,k}\upsilon_{j,k}}{2}\right] = \mu_{q(\log\lambda_{j,k})} + \mu_{q(\log\upsilon_{j,k})} - \log 2,$$

hence the exponential of term in (A.16) is integrable and thus we can compute the normalizing constant and its expectation. □

Proposition A.13. *The variational lower bound for the multivariate regression model with adaptive normal-gamma prior can be derived analytically and it is equal to:*

$$\begin{aligned}\log \underline{p}(\mathbf{y};q) = {} & d\left(-\frac{T}{2}\log 2\pi + a_\nu \log b_\nu - \log\Gamma(a_\nu)\right) - \sum_{j=1}^{d}\left(a_{q(\nu_j)}\log b_{q(\nu_j)} - \log\Gamma(a_{q(\nu_j)})\right)\\
& - \frac{1}{2}\sum_{j=2}^{d}\sum_{k=1}^{j-1}\left(\log\tau + 1/\tau\mu_{q(\beta^2_{j,k})}\right) + \frac{1}{2}\sum_{j=2}^{d}\left(\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}| + (j-1)\right)\\
& + \frac{1}{2}\left(\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\vartheta})}| + dp\right) - \sum_{j=1}^{d}\sum_{k=1}^{p} h(\zeta_{q(\upsilon_{j,k})}, a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})}) + d\log h_3\\
& + dp\left(h_1\log h_2 - \log\Gamma(h_1)\right) - \sum_{j=1}^{d}\sum_{k=1}^{p}\left(a_{q(\lambda_{j,k})}\log b_{q(\lambda_{j,k})} - \log\Gamma(a_{q(\lambda_{j,k})})\right)\\
& + \sum_{j=1}^{d}\log c_{\eta_j} + \sum_{j=1}^{d}\mu_{q(\eta_j)}\sum_{k=1}^{p}\left(\mu_{q(\lambda_{j,k})}\mu_{q(\upsilon_{j,k})} - \mu_{q(\log\lambda_{j,k})} - \mu_{q(\log\upsilon_{j,k})}\right).\end{aligned} \tag{A.17}$$

Proof. As we did in (A.6) for Proposition A.5, the lower bound can be divided into terms referring to each parameter:

$$\log \underline{p}(\mathbf{y};q) = A + \sum_{j=1}^{d}\sum_{k=1}^{p}\left(\underbrace{\mathbb{E}_q\left[\log \underline{p}(\mathbf{y};\upsilon_{j,k})\right]}_{B} + \underbrace{\mathbb{E}_q\left[\log \underline{p}(\mathbf{y};\lambda_{j,k})\right]}_{C} + \underbrace{\mathbb{E}_q\left[\log \underline{p}(\mathbf{y};\eta_j)\right]}_{D}\right), \tag{A.18}$$

where A is equal to (A.2.2).

Our strategy will be to evaluate each piece in the latter separately and then put the results together. Consider the piece B:

$$\begin{aligned}B = {} & \mathbb{E}_q\left[\eta_j\log\eta_j + \eta_j\left(\log\lambda_{j,k} - \log 2\right) - \log\Gamma(\eta_j) + (\eta_j - 1)\log\upsilon_{j,k} - \upsilon_{j,k}\frac{\eta_j\lambda_{j,k}}{2}\right]\\
& - \mathbb{E}_q\left[h(\zeta_{q(\upsilon_{j,k})}, a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})}) + (\zeta_{q(\upsilon_{j,k})} - 1)\log\upsilon_{j,k} - \frac{a_{q(\upsilon_{j,k})}\upsilon_{j,k}}{2} - \frac{b_{q(\upsilon_{j,k})}}{2\upsilon_{j,k}}\right]\\
= {} & \mu_{q(\eta_j\log\eta_j)} + \mu_{q(\eta_j)}\left(\mu_{q(\log\lambda_{j,k})} - \log 2\right) - \mu_{q(\log\Gamma(\eta_j))} - h(\zeta_{q(\upsilon_{j,k})}, a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})})\\
& + (\mu_{q(\eta_j)} - 1)\mu_{q(\log\upsilon_{j,k})} - (\zeta_{q(\upsilon_{j,k})} - 1)\mu_{q(\log\upsilon_{j,k})}\\
& - \frac{1}{2}\left(\mu_{q(\upsilon_{j,k})}\mu_{q(\eta_j)}\mu_{q(\lambda_{j,k})} - a_{q(\upsilon_{j,k})}\mu_{q(\upsilon_{j,k})} - b_{q(\upsilon_{j,k})}\mu_{q(1/\upsilon_{j,k})}\right),\end{aligned}$$

where $h(\zeta, a, b)$ denotes the logarithm of the normalizing constant of a GIG distribution, i.e.,

$$h(\zeta, a, b) = \zeta/2\log(a/b) - \log 2 - \log K_\zeta(\sqrt{ab}).$$

The term involving $\lambda_{j,k}$, for $j = 1, \dots, d$ and $k = 1, \dots, p$, is equal to:

$$\begin{aligned} C &= \mathbb{E}_q \left[h_1 \log h_2 - \log \Gamma(h_1) + (h_1 - 1) \log \lambda_{j,k} - \lambda_{j,k} h_2\right] \\ &\quad - \mathbb{E}_q \left[a_{q(\lambda_{j,k})} \log b_{q(\lambda_{j,k})} - \log \Gamma(a_{q(\lambda_{j,k})}) + (a_{q(\lambda_{j,k})} - 1) \log \lambda_{j,k} - \lambda_{j,k} b_{q(\lambda_{j,k})}\right] \\ &= h_1 \log h_2 - \log \Gamma(h_1) + (h_1 - 1)\mu_{q(\log \lambda_{j,k})} - \mu_{q(\lambda_{j,k})} h_2 \\ &\quad - a_{q(\lambda_{j,k})} \log b_{q(\lambda_{j,k})} + \log \Gamma(a_{q(\lambda_{j,k})}) - (a_{q(\lambda_{j,k})} - 1)\mu_{q(\log \lambda_{j,k})} + \mu_{q(\lambda_{j,k})} b_{q(\lambda_{j,k})}, \end{aligned}$$

and, to conclude, compute the term D:

$$\begin{aligned} D &= \mathbb{E}_q \left[\log h_3 - \eta_j h_3\right] \\ &\quad - \mathbb{E}_q \left[\log h(\eta_j) - \log c_{\eta_j} - \eta_j \sum_{k=1}^{p} \left(\frac{\mu_{q(\lambda_{j,k})}\mu_{q(\upsilon_{j,k})}}{2} - \mu_{q(\log \lambda_{j,k})} - \mu_{q(\log \upsilon_{j,k})} + \log 2 + h_3\right)\right] \\ &= \log h_3 - \mu_{q(\eta_j)} h_3 \\ &\quad - \mu_{q(\log h(\eta_j))} + \log c_{\eta_j} + \mu_{q(\eta_j)} \sum_{k=1}^{p} \left(\frac{\mu_{q(\lambda_{j,k})}\mu_{q(\upsilon_{j,k})}}{2} - \mu_{q(\log \lambda_{j,k})} - \mu_{q(\log \upsilon_{j,k})} + \log 2 + h_3\right). \end{aligned}$$

Group together the terms and exploit the analytical form of the optimal parameters to perform some simplifications. The remaining terms form the lower bound for a multivariate regression model with adaptive normal-gamma prior. □

A.2.4 Horseshoe prior

First of all, notice that the optimal densities for the variances ν_j and for the coefficients $\boldsymbol{\beta}_j$ remain the same computed in Section A.2.1. The changes in the optimal densities $q^*(\boldsymbol{\vartheta})$ are stated in the next Proposition.

Proposition A.14. *The joint optimal variational density for the parameter $\boldsymbol{\vartheta}$ is equal to $q^*(\boldsymbol{\vartheta}) \equiv \mathsf{N}_{dp}(\boldsymbol{\mu}_{q(\vartheta)}, \boldsymbol{\Sigma}_{q(\vartheta)})$, where:*

$$\begin{aligned} \boldsymbol{\Sigma}_{q(\vartheta)} &= \left(\boldsymbol{\mu}_{q(\boldsymbol{\Omega})} \otimes \sum_{t=1}^{T} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^{\intercal} + \boldsymbol{\mu}_{q(1/\gamma^2)} \mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon^2)})\right)^{-1}, \\ \boldsymbol{\mu}_{q(\vartheta)} &= \boldsymbol{\Sigma}_{q(\vartheta)} \sum_{t=1}^{T} \left(\boldsymbol{\mu}_{q(\boldsymbol{\Omega})} \otimes \mathbf{z}_{t-1}\right) \mathbf{y}_t, \end{aligned} \tag{A.19}$$

where $\mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon^2)})$ *is a diagonal matrix and* $\boldsymbol{\mu}_{q(1/\upsilon^2)} = (\mu_{q(1/\upsilon^2_{1,1})}, \mu_{q(1/\upsilon^2_{1,2})}, \dots, \mu_{q(1/\upsilon^2_{d,p})})$.

Under the row-independence assumption, the optimal variational density for the parameter $\boldsymbol{\vartheta}_j$ *is equal to* $q^*(\boldsymbol{\vartheta}_j) \equiv \mathsf{N}_p(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$*, where:*

$$\begin{aligned} \boldsymbol{\Sigma}_{q(\vartheta_j)} &= \left(\boldsymbol{\mu}_{q(\omega_{j,j})} \sum_{t=1}^{T} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^{\mathsf{T}} + \boldsymbol{\mu}_{q(1/\gamma^2)} \mathsf{diag}(\boldsymbol{\mu}_{q(1/\upsilon_j^2)}) \right)^{-1}, \\ \boldsymbol{\mu}_{q(\vartheta_j)} &= \boldsymbol{\Sigma}_{q(\vartheta_j)} \left(\sum_{t=1}^{T} \left(\boldsymbol{\mu}_{q(\omega_j)} \otimes \mathbf{z}_{t-1} \right) \mathbf{y}_t - \left(\boldsymbol{\mu}_{q(\omega_{j,-j})} \otimes \sum_{t=1}^{T} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^{\mathsf{T}} \right) \boldsymbol{\mu}_{q(\vartheta_{-j})} \right), \end{aligned} \tag{A.20}$$

where $\mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon_j^2)})$ *is a diagonal matrix and* $\boldsymbol{\mu}_{q(1/\upsilon_j^2)} = (\mu_{q(1/\upsilon_{j,1}^2)}, \mu_{q(1/\upsilon_{j,2}^2)}, \ldots, \mu_{q(1/\upsilon_{j,p}^2)})$.

Hereafter we describe the optimal densities for the parameters used in hierarchical specification of the prior.

Proposition A.15. *The optimal density for the prior local variance* $\upsilon_{j,k}^2$ *is equal to an inverse-gamma distribution* $q^*(\upsilon_{j,k}^2) \equiv \mathsf{IGa}(1, b_{q(\upsilon_{j,k}^2)})$*, where, for* $j = 1, \ldots, d$ *and* $k = 1, \ldots, p$*:*

$$b_{q(\upsilon_{j,k}^2)} = \mu_{q(1/\lambda_{j,k})} + \frac{1}{2}\mu_{q(\vartheta_{j,k}^2)}\mu_{q(1/\gamma^2)}. \tag{A.21}$$

Proof. Consider the prior specification which involves the parameter $\upsilon_{j,k}^2$:

$$\vartheta_{j,k}|\gamma^2, \upsilon_{j,k}^2 \sim \mathsf{N}(0, \gamma^2\upsilon_{j,k}^2), \qquad \upsilon_{j,k}^2|\lambda_{j,k} \sim \mathsf{IGa}\left(1/2, 1/\lambda_{j,k}\right).$$

Compute the optimal variational density $\log q^*(\upsilon_{j,k}^2) \propto \mathbb{E}_{-\upsilon_{\mathbf{j,k}}^2}\left[\log p(\vartheta_{j,k}) + \log p(\upsilon_{j,k}^2)\right]$:

$$\begin{aligned} \log q^*(\upsilon_{j,k}^2) &\propto \mathbb{E}_{-\upsilon_{j,k}^2}\left[-\frac{1}{2}\log\upsilon_{j,k}^2 - \frac{1}{2\gamma^2\upsilon_{j,k}^2}\vartheta_{j,k}^2 - (1/2+1)\log\upsilon_{j,k}^2 - \frac{1}{\upsilon_{j,k}^2\lambda_{j,k}} \right] \\ &\propto -2\log\upsilon_{j,k}^2 - \frac{1}{\upsilon_{j,k}^2}\left(\mu_{q(1/\gamma^2)}\mu_{q(\vartheta_{j,k}^2)}/2 + \mu_{q(1/\lambda_{j,k})} \right). \end{aligned}$$

Take the exponential and notice that the latter is the kernel of an inverse-gamma random variable $\mathsf{IGa}(1, b_{q(\upsilon_{j,k}^2)})$, as defined in Proposition A.15. □

Proposition A.16. *The optimal density for the prior global variance* γ^2 *is equal to an inverse-gamma distribution* $q^*(\gamma^2) \equiv \mathsf{IGa}(a_{q(\gamma^2)}, b_{q(\gamma^2)})$*, where:*

$$a_{q(\gamma^2)} = \frac{dp+1}{2}, \quad b_{q(\gamma^2)} = \mu_{q(1/\eta)} + \frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{p}\mu_{q(1/\upsilon_{j,k}^2)}\mu_{q(\vartheta_{j,k}^2)}. \tag{A.22}$$

Proof. Consider the prior specification which involves the parameter γ^2:

$$\vartheta_{j,k}|\gamma^2, \upsilon^2_{j,k} \sim \mathsf{N}(0, \gamma^2\upsilon^2_{j,k}), \qquad \gamma^2|\eta \sim \mathsf{IGa}\left(1/2, 1/\eta\right).$$

Compute the optimal variational density $\log q^*(\gamma^2) \propto \mathbb{E}_{-\gamma^2}\left[\sum_{j=1}^d \sum_{k=1}^p \log p(\vartheta_{j,k}) + \log p(\gamma^2)\right]$:

$$\begin{aligned}\log q^*(\gamma^2) &\propto \mathbb{E}_{-\gamma^2}\left[-\frac{dp}{2}\log\gamma^2 - \frac{1}{2\gamma^2\upsilon^2_{j,k}}\vartheta^2_{j,k} - (1/2+1)\log\gamma^2 - \frac{1}{\gamma^2\eta}\right] \\ &\propto -\left(\frac{dp+1}{2}+1\right)\log\gamma^2 - \frac{1}{\gamma^2}\left(\sum_{j=1}^d\sum_{k=1}^p \mu_{q(1/\upsilon^2_{j,k})}\mu_{q(\vartheta^2_{j,k})}/2 + \mu_{q(1/\eta)}\right).\end{aligned}$$

Take the exponential and notice that the latter is the kernel of an inverse-gamma random variable $\mathsf{IGa}(a_{q(\gamma^2)}, b_{q(\gamma^2)})$, as defined in Proposition A.16. □

Proposition A.17. *The optimal density for the latent parameter $\lambda_{j,k}$ is equal to an inverse-gamma distribution $q^*(\lambda_{j,k}) \equiv \mathsf{IGa}(1, b_{q(\lambda_{j,k})})$, where, for $j = 1, \ldots, d$ and $k = 1, \ldots, p$:*

$$b_{q(\lambda_{j,k})} = 1 + \mu_{q(1/\upsilon^2_{j,k})}. \tag{A.23}$$

Proof. Consider the prior specification which involves the parameter $\lambda_{j,k}$:

$$\upsilon^2_{j,k}|\lambda_{j,k} \sim \mathsf{IGa}\left(1/2, 1/\lambda_{j,k}\right), \qquad \lambda_{j,k} \sim \mathsf{IGa}\left(1/2, 1\right).$$

Compute the optimal variational density $\log q^*(\lambda_{j,k}) \propto \mathbb{E}_{-\lambda_{\mathbf{j},\mathbf{k}}}\left[\log p(\upsilon^2_{j,k}) + \log p(\lambda_{j,k})\right]$:

$$\begin{aligned}\log q^*(\lambda_{j,k}) &\propto \mathbb{E}_{-\lambda_{j,k}}\left[-\frac{1}{2}\log\lambda_{j,k} - \frac{1}{\upsilon^2_{j,k}\lambda_{j,k}} - (1/2+1)\log\lambda_{j,k} - \frac{1}{\lambda_{j,k}}\right] \\ &\propto -2\log\lambda_{j,k} - \frac{1}{\lambda_{j,k}}\left(1 + \mu_{q(1/\upsilon^2_{j,k})}\right).\end{aligned}$$

Take the exponential and notice that the latter is the kernel of an inverse-gamma random variable $\mathsf{IGa}(1, b_{q(\lambda_{j,k})})$, as defined in Proposition A.17. □

Proposition A.18. *The optimal density for the latent parameter η is equal to an inverse-gamma distribution $q^*(\eta) \equiv \mathsf{IGa}(1, b_{q(\eta)})$, where:*

$$b_{q(\eta)} = 1 + \mu_{q(1/\gamma^2)}. \tag{A.24}$$

Proof. Consider the prior specification which involves the parameter η:

$$\gamma^2|\eta \sim \mathsf{IGa}\left(1/2, 1/\eta\right), \qquad \eta \sim \mathsf{IGa}\left(1/2, 1\right).$$

Compute the optimal variational density $\log q^*(\eta) \propto \mathbb{E}_{-\eta}\left[\log p(\gamma^2) + \log p(\eta)\right]$:

$$\log q^*(\eta) \propto \mathbb{E}_{-\eta}\left[-\frac{1}{2}\log\eta - \frac{1}{\gamma^2\eta} - (1/2+1)\log\eta - \frac{1}{\eta}\right] \propto -2\log\eta - \frac{1}{\eta}\left(1 + \mu_{q(1/\gamma^2)}\right).$$

Take the exponential and notice that the latter is the kernel of an inverse-gamma random variable $\mathsf{IGa}(1, b_{q(\eta)})$, as defined in Proposition A.18. □

Proposition A.19. *The variational lower bound for the multivariate regression model with Horseshoe prior can be derived analytically and it is equal to:*

$$\begin{aligned}
\log \underline{p}(\mathbf{y}; q) = d&\left(-\frac{T}{2}\log 2\pi + a_\nu \log b_\nu - \log\Gamma(a_\nu)\right) - \sum_{j=1}^{d}\left(a_{q(\nu_j)}\log b_{q(\nu_j)} - \log\Gamma(a_{q(\nu_j)})\right) \\
&- \frac{1}{2}\sum_{j=2}^{d}\sum_{k=1}^{j-1}\left(\log\tau + 1/\tau\mu_{q(\beta^2_{j,k})}\right) + \frac{1}{2}\sum_{j=2}^{d}\left(\log|\mathbf{\Sigma}_{q(\boldsymbol{\beta}_j)}| + (j-1)\right) \\
&+ \frac{1}{2}\left(\log|\mathbf{\Sigma}_{q(\boldsymbol{\vartheta})}| + dp\right) + \mu_{q(1/\gamma^2)}\left(\mu_{q(1/\eta)} + \sum_{j=1}^{d}\sum_{k=1}^{p}\mu_{q(\vartheta^2_{j,k})}\mu_{q(1/\upsilon^2_{j,k})}\right) \\
&+ \sum_{j=1}^{d}\sum_{k=1}^{p}\left(\mu_{q(1/\upsilon^2_{j,k})}\mu_{q(1/\lambda_{j,k})} - \log b_{q(\upsilon^2_{j,k})} - \log b_{q(\lambda_{j,k})} - \log\pi\right) \\
&- a_{q(\gamma^2)}\log b_{q(\gamma^2)} - \log b_{q(\eta)} - \log\pi.
\end{aligned} \tag{A.25}$$

Proof. As we did in (A.6) for Proposition A.5, the lower bound can be divided into terms referring to each parameter:

$$\begin{aligned}
\log \underline{p}(\mathbf{y}; q) = A + &\underbrace{\mathbb{E}_q\left[\log \underline{p}(\mathbf{y}; \gamma^2)\right]}_{B} + \underbrace{\mathbb{E}_q\left[\log \underline{p}(\mathbf{y}; \eta)\right]}_{C} \\
&+ \sum_{j=1}^{d}\sum_{k=1}^{p}\left(\underbrace{\mathbb{E}_q\left[\log \underline{p}(\mathbf{y}; \upsilon^2_{j,k})\right]}_{D} + \underbrace{\mathbb{E}_q\left[\log \underline{p}(\mathbf{y}; \lambda_{j,k})\right]}_{E}\right),
\end{aligned} \tag{A.26}$$

where A is similar to (A.6) in the previous non-informative model specification. Our strategy will be to evaluate each piece in the latter separately and then put the results together. Notice that the computations for the piece A are similar to Proposition A.5.

Hence, we have that:

$$\begin{aligned}
A = {} & d\left(-\frac{T}{2}\log 2\pi + a_\nu \log b_\nu - \log\Gamma(a_\nu)\right) - \sum_{j=1}^{d}\big(a_{q(\nu_j)}\log b_{q(\nu_j)} - \log\Gamma(a_{q(\nu_j)})\big) \\
& -\frac{1}{2}\sum_{j=2}^{d}\sum_{k=1}^{j-1}\Big(\log\tau + 1/\tau\mu_{q(\beta^2_{j,k})}\Big) + \frac{1}{2}\sum_{j=2}^{d}\Big(\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}| + (j-1)\Big) \\
& -\frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{p}\Big(\mu_{q(\log\delta^2)} + \mu_{q(\log\upsilon^2_{j,k})} + \mu_{q(1/\delta^2)}\mu_{q(1/\upsilon^2_{j,k})}\mu_{q(\vartheta^2_{j,k})}\Big) \\
& +\frac{1}{2}\left(\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\vartheta})}| + dp\right).
\end{aligned} \tag{A.27}$$

Consider now the piece B. We have that:

$$\begin{aligned}
B = {} & \mathbb{E}_q\left[-\frac{1}{2}\log\eta - \frac{1}{2}\log\pi - (1/2+1)\log\gamma^2 - 1/(\gamma^2\eta)\right] \\
& - \mathbb{E}_q\left[a_{q(\gamma^2)}\log b_{q(\gamma^2)} - \log\Gamma(a_{q(\gamma^2)}) - (a_{q(\gamma^2)}+1)\log\gamma^2 - b_{q(\gamma^2)}/\gamma^2\right] \\
= {} & -\frac{1}{2}\mu_{q(\log\eta)} - \frac{1}{2}\log\pi - (1/2+1)\mu_{q(\log\gamma^2)} - \mu_{q(1/\gamma^2)}\mu_{q(1/\eta)} \\
& - a_{q(\gamma^2)}\log b_{q(\gamma^2)} + \log\Gamma(a_{q(\gamma^2)}) + (a_{q(\gamma^2)}+1)\mu_{q(\log\gamma^2)} + \mu_{q(1/\gamma^2)}b_{q(\gamma^2)},
\end{aligned}$$

while, C reduces to:

$$\begin{aligned}
C = {} & \mathbb{E}_q\left[-\frac{1}{2}\log\pi - (1/2+1)\log\eta - 1/\eta\right] - \mathbb{E}_q\left[\log b_{q(\eta)} - 2\log\eta - b_{q(\eta)}/\eta\right] \\
= {} & -\frac{1}{2}\log\pi - (1/2+1)\mu_{q(\log\eta)} - \mu_{q(1/\eta)} - \log b_{q(\eta)} + 2\mu_{q(\log\eta)} + \mu_{q(1/\eta)}b_{q(\eta)}.
\end{aligned}$$

The remaining terms behave likely B and C. For $j = 1,\dots,d$ and $k = 1,\dots,p$:

$$\begin{aligned}
D = {} & \mathbb{E}_q\left[-\frac{1}{2}\log\lambda_{j,k} - \frac{1}{2}\log\pi - (1/2+1)\log\upsilon^2_{j,k} - 1/(\upsilon^2_{j,k}\lambda_{j,k})\right] \\
& - \mathbb{E}_q\left[\log b_{q(\upsilon^2_{j,k})} - 2\log\upsilon^2_{j,k} - b_{q(\upsilon^2_{j,k})}/\upsilon^2_{j,k}\right] \\
= {} & -\frac{1}{2}\mu_{q(\log\lambda_{j,k})} - \frac{1}{2}\log\pi - (1/2+1)\mu_{q(\log\upsilon^2_{j,k})} - \mu_{q(1/\upsilon^2_{j,k})}\mu_{q(1/\lambda_{j,k})} \\
& - \log b_{q(\upsilon^2_{j,k})} + 2\mu_{q(\log\upsilon^2_{j,k})} + \mu_{q(1/\upsilon^2_{j,k})}b_{q(\upsilon^2_{j,k})},
\end{aligned}$$

and

$$\begin{aligned}
E = {} & \mathbb{E}_q\left[-\frac{1}{2}\log\pi - (1/2+1)\log\lambda_{j,k} - 1/\lambda_{j,k}\right] - \mathbb{E}_q\left[\log b_{q(\lambda_{j,k})} - 2\log\lambda_{j,k} - b_{q(\lambda_{j,k})}/\lambda_{j,k}\right] \\
= {} & -\frac{1}{2}\log\pi - (1/2+1)\mu_{q(\log\lambda_{j,k})} - \mu_{q(1/\lambda_{j,k})} - \log b_{q(\lambda_{j,k})} + 2\mu_{q(\log\lambda_{j,k})} + \mu_{q(1/\lambda_{j,k})}b_{q(\lambda_{j,k})}.
\end{aligned}$$

Group together the terms and exploit the analytical form of the optimal parameters to perform some simplifications. The remaining terms form the lower bound for a multivariate regression model with Horseshoe prior. □

A.3 Variational predictive density

Recall that the variational predictive posterior can be computed as:

$$q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = \int p(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\xi}) q^*(\boldsymbol{\xi}) d\boldsymbol{\xi} = \int\int p(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\vartheta}, \boldsymbol{\Omega}) q(\boldsymbol{\vartheta}) q(\boldsymbol{\Omega}) d\boldsymbol{\vartheta}\, d\boldsymbol{\Omega}, \tag{A.1}$$

which requires only a simulation step according to the first methodology presented in the main paper. If we wish to make the estimation simpler, we can integrate out the precision parameter $\boldsymbol{\Omega}$ (whose optimal variational density is discussed in Section 2.3.2) in the following way:

$$q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = \int q(\boldsymbol{\vartheta}) \underbrace{\left[\int \mathsf{N}_d(\mathbf{y}_{t+1}; \boldsymbol{\Theta}\mathbf{z}_t, \boldsymbol{\Omega}^{-1}) \mathsf{Wishart}_d(\boldsymbol{\Omega}; \delta, \mathbf{H}) d\boldsymbol{\Omega}\right]}_{A} d\boldsymbol{\vartheta}, \tag{A.2}$$

where

$$A = \frac{2^{-d(\delta+1)/2}|\mathbf{H}|^{\delta/2}}{\pi^{d/2}\Gamma_d(\delta/2)} \int \underbrace{|\boldsymbol{\Omega}|^{(\delta-d)/2} \exp\left\{-\frac{1}{2}\mathsf{tr}\left\{\boldsymbol{\Omega}\left(\mathbf{H}^{-1} + (\mathbf{y}_{t+1} - \boldsymbol{\Theta}\mathbf{z}_t)(\mathbf{y}_{t+1} - \boldsymbol{\Theta}\mathbf{z}_t)^\intercal\right)\right\}\right\}}_{\text{Kernel of a } \mathsf{Wishart}_d(\delta+1, (\mathbf{H}^{-1}+(\mathbf{y}_{t+1}-\boldsymbol{\Theta}\mathbf{z}_t)(\mathbf{y}_{t+1}-\boldsymbol{\Theta}\mathbf{z}_t)^\intercal)^{-1})} d\boldsymbol{\Omega}$$

$$= \frac{|1 + \frac{1}{v}(\mathbf{y}_{t+1} - \boldsymbol{\Theta}\mathbf{z}_t)^\intercal v\mathbf{H}(\mathbf{y}_{t+1} - \boldsymbol{\Theta}\mathbf{z}_t)|^{-\frac{v+d}{2}}\Gamma(\frac{v+d}{2})}{\pi^{d/2}v^{d/2}|\mathbf{H}^{-1}|^{1/2}\Gamma(v/2)} = h(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\vartheta}), \tag{A.3}$$

is the density function of a multivariate Student-t distribution with dimension d, $v = \delta - d + 1$ degrees of freedom, mean vector $\boldsymbol{\Theta}\mathbf{z}_t$ and scaling matrix $\mathbf{S} = (v\mathbf{H})^{-1}$, i.e., $\mathsf{t}_v(\boldsymbol{\Theta}\mathbf{z}_t, \mathbf{S})$. Then, the integral in (A.1) becomes

$$q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = \int h(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\vartheta}) q(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}, \tag{A.4}$$

which requires to simulate from the optimal multivariate Gaussian distribution of $\boldsymbol{\vartheta}$ according to the second methodology presented in the main paper.

A second-order approximation can be implemented in order to increase the computational efficiency. To this aim, we propose to approximate the multivariate Student-t in (A.4) with the closest multivariate normal distribution in terms of $\mathcal{KL}$ divergence:

$$\begin{aligned}\mathcal{KL}(h\|\phi) &\propto -\int \log\phi(\mathbf{y}_{t+1}|\mathbf{m},\mathbf{R}^{-1})h(\mathbf{y}_{t+1}|\mathbf{z}_t,\boldsymbol{\vartheta})\,d\mathbf{y}_{t+1}\\ &= -\mathbb{E}_h(\log\phi(\mathbf{y}_{t+1}|\mathbf{m},\mathbf{R}^{-1})) = \psi(\mathbf{m},\mathbf{R}),\end{aligned} \tag{A.5}$$

where, in particular,

$$\begin{aligned}\psi(\mathbf{m},\mathbf{R}) &\propto \mathbb{E}_h\left(-\frac{1}{2}\log\mathbf{R}+\frac{1}{2}(\mathbf{y}_{t+1}-\mathbf{m})^\intercal\mathbf{R}(\mathbf{y}_{t+1}-\mathbf{m})\right)\\ &= -\frac{1}{2}\log\mathbf{R}+\frac{1}{2}(\boldsymbol{\Theta}\mathbf{z}_t-\mathbf{m})^\intercal\mathbf{R}(\boldsymbol{\Theta}\mathbf{z}_t-\mathbf{m})+\frac{v}{2(v-2)}\mathsf{tr}\left\{\mathbf{RS}\right\},\end{aligned} \tag{A.6}$$

which turns out to be minimized when $\mathbf{m}=\boldsymbol{\Theta}\mathbf{z}_t$ and $\mathbf{R}=\frac{v-2}{v}\mathbf{S}^{-1}$. If we substitute the function $h(\cdot)$ with its Gaussian approximation we get

$$q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = \int \phi(\mathbf{y}_{t+1}|\mathbf{m},\mathbf{R}^{-1})q(\boldsymbol{\vartheta})d\boldsymbol{\vartheta}, \tag{A.7}$$

where now $\phi(\mathbf{y}_{t+1}|\boldsymbol{\Theta}\mathbf{z}_t,\mathbf{R}^{-1})$ denotes the density of the multivariate normal distribution that is closest in a $\mathcal{KL}$ sense to the multivariate Student-t $h(\mathbf{y}_{t+1}|\mathbf{z}_t,\boldsymbol{\vartheta})$. The advantage of this procedure is that the integral in (A.7) can be solved analytically leading to a variational predictive density $q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t})$ which is a multivariate Gaussian distribution with variance matrix $\boldsymbol{\Sigma}_{pred}$ and mean vector $\boldsymbol{\mu}_{pred}$. Define $\mathbf{Z}_t=(\mathbf{I}_d\otimes\mathbf{z}_t^\intercal)$ and compute the integral above:

$$\begin{aligned}q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) &\propto \int\exp\left\{-\frac{1}{2}\left[(\mathbf{y}_{t+1}-\mathbf{Z}_t\boldsymbol{\vartheta})^\intercal\mathbf{R}(\mathbf{y}_{t+1}-\mathbf{Z}_t\boldsymbol{\vartheta})+(\boldsymbol{\vartheta}-\boldsymbol{\mu}_{q(\vartheta)})^\intercal\boldsymbol{\Sigma}_{q(\vartheta)}^{-1}(\boldsymbol{\vartheta}-\boldsymbol{\mu}_{q(\vartheta)})\right]\right\}d\boldsymbol{\vartheta}\\ &\propto \exp\left\{-\frac{1}{2}\mathbf{y}_{t+1}^\intercal\mathbf{R}\mathbf{y}_{t+1}\right\}\\ &\qquad\times\int\exp\left\{-\frac{1}{2}\left[\boldsymbol{\vartheta}^\intercal(\boldsymbol{\Sigma}_{q(\vartheta)}^{-1}+\mathbf{Z}_t^\intercal\mathbf{R}\mathbf{Z}_t)\boldsymbol{\vartheta}-2\boldsymbol{\vartheta}^\intercal(\boldsymbol{\Sigma}_{q(\vartheta)}^{-1}\boldsymbol{\mu}_{q(\vartheta)}+\mathbf{Z}_t\mathbf{R}\mathbf{y}_{t+1})\right]\right\}d\boldsymbol{\vartheta},\end{aligned} \tag{A.8}$$

where the term in the integral is the kernel of a multivariate Gaussian random variable with variance matrix $\tilde{\boldsymbol{\Sigma}}=(\boldsymbol{\Sigma}_{q(\vartheta)}^{-1}+\mathbf{Z}_t^\intercal\mathbf{R}\mathbf{Z}_t)^{-1}$ and mean $\tilde{\boldsymbol{\mu}}=\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}_{q(\vartheta)}^{-1}\boldsymbol{\mu}_{q(\vartheta)}+\mathbf{Z}_t\mathbf{R}\mathbf{y}_{t+1})$. Solve the integral and get:

$$\begin{aligned}q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) &\propto \exp\left\{-\frac{1}{2}(\mathbf{y}_{t+1}^\intercal\mathbf{R}\mathbf{y}_{t+1}-\tilde{\boldsymbol{\mu}}^\intercal\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\mu}})\right\}\\ &\propto \exp\left\{-\frac{1}{2}(\mathbf{y}_{t+1}^\intercal\mathbf{R}\mathbf{y}_{t+1}-\mathbf{y}_{t+1}^\intercal\mathbf{R}\mathbf{Z}_t\tilde{\boldsymbol{\Sigma}}\mathbf{Z}_t^\intercal\mathbf{R}\mathbf{y}_{t+1}-2\mathbf{y}_{t+1}\mathbf{R}\mathbf{Z}_t\tilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}_{q(\vartheta)}^{-1}\boldsymbol{\mu}_{q(\vartheta)})\right\}\\ &= \exp\left\{-\frac{1}{2}(\mathbf{y}_{t+1}^\intercal(\mathbf{R}-\mathbf{R}\mathbf{Z}_t\tilde{\boldsymbol{\Sigma}}\mathbf{Z}_t^\intercal\mathbf{R})\mathbf{y}_{t+1}-2\mathbf{y}_{t+1}\mathbf{R}\mathbf{Z}_t\tilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}_{q(\vartheta)}^{-1}\boldsymbol{\mu}_{q(\vartheta)})\right\},\end{aligned} \tag{A.9}$$

which is the kernel of a multivariate Gaussian with variance matrix $\boldsymbol{\Sigma}_{pred} = (\mathbf{R} - \mathbf{R}\mathbf{Z}_t\tilde{\boldsymbol{\Sigma}}\mathbf{Z}_t^\intercal\mathbf{R})^{-1}$ and mean $\boldsymbol{\mu}_{pred} = \boldsymbol{\Sigma}_{pred}\mathbf{R}\mathbf{Z}_t\tilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}_{q(\vartheta)}\boldsymbol{\mu}_{q(\vartheta)}$. To conclude, the second-order Gaussian approximation to the variational predictive posterior is such that $q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) \equiv \mathsf{N}_d(\boldsymbol{\mu}_{pred}, \boldsymbol{\Sigma}_{pred})$.

Figure A.1 shows the approximation of variational predictive posterior with Monte Carlo methods (MC) and via Gaussian approximation (GA) varying the degrees of freedom $\widehat{\delta}$ for the distribution of $\boldsymbol{\Omega}$. We can see that if $\widehat{\delta} \gg d$ the approximation is rather accurate, while the accuracy decreases as $\widehat{\delta}$ approaches d. However, even for the case $\widehat{\delta} \approx d$, we can still obtain rather precise estimates of the first and second moments of the variational predictive posterior.

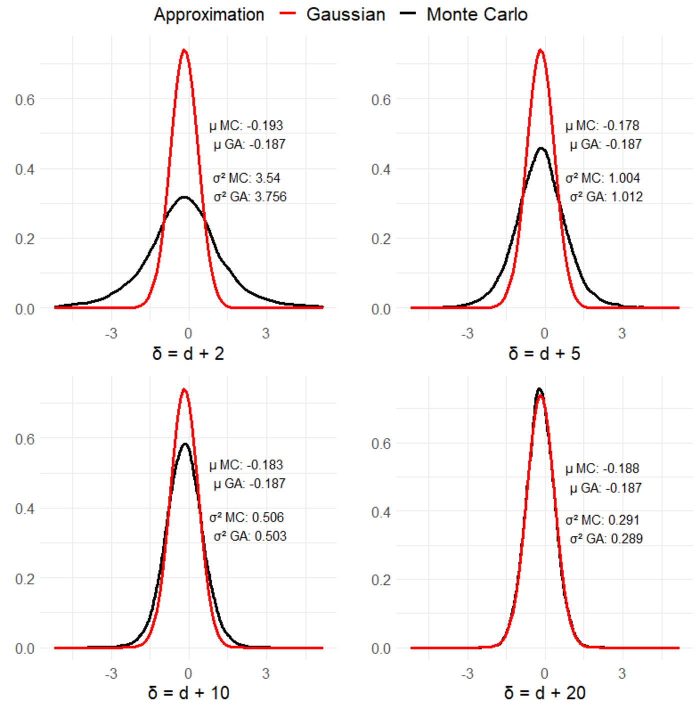


FIGURE A.1: Second-order approximation of the predictive density.

A.4 Simulation details and additional results

In this Section we report additional details and results on the simulation study we highlighted in Section 3.3. We set the length of the time series equal to $T = 360$, corresponding to 30 years of monthly data, the dimension of the multivariate regression model equal to $d = 15, 30, 49$ and we further assume both moderate level of sparsity (50% of true zeros) and high level of sparsity (90% of true zeros). The true matrix $\boldsymbol{\Theta}$ is generated as follows: we fix to zero sd^2 entries at random, where $s = 0.5, 0.9$, while the remaining non zero coefficients are sampled from a mixutre of two Gaussian with means

-0.08 and 0.08, and standard deviation 0.1. Figure A.2 reports the distribution of the non-zero parameters. Notice the draws from the Normal distributions are truncated at -0.05 and 0.05 respectively, to avoid very small values for the non zero parameters.

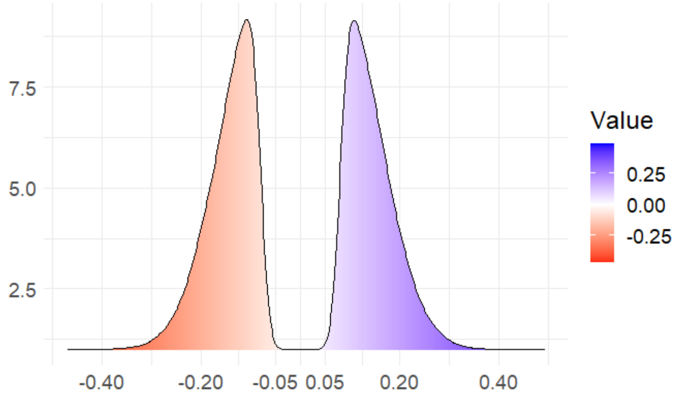
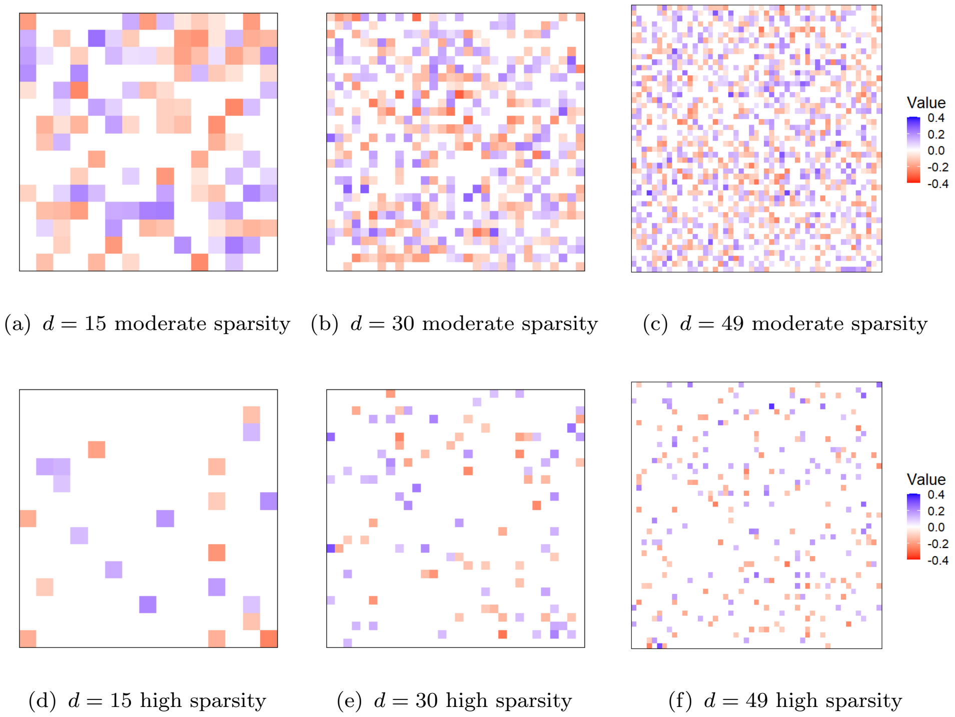


FIGURE A.2: Distribution of non-zero parameters in the regression matrices used to generate the data for the simulation study.

Figure A.3 shows examples of the true regression matrixes for different dimensions $d = 15, 30, 49$ and for two alternative levels of sparsity $s = 0.5, 0.9$, that is 50% and 90% of the entries in the matrix $\mathbf{\Theta}$ are set to zero.

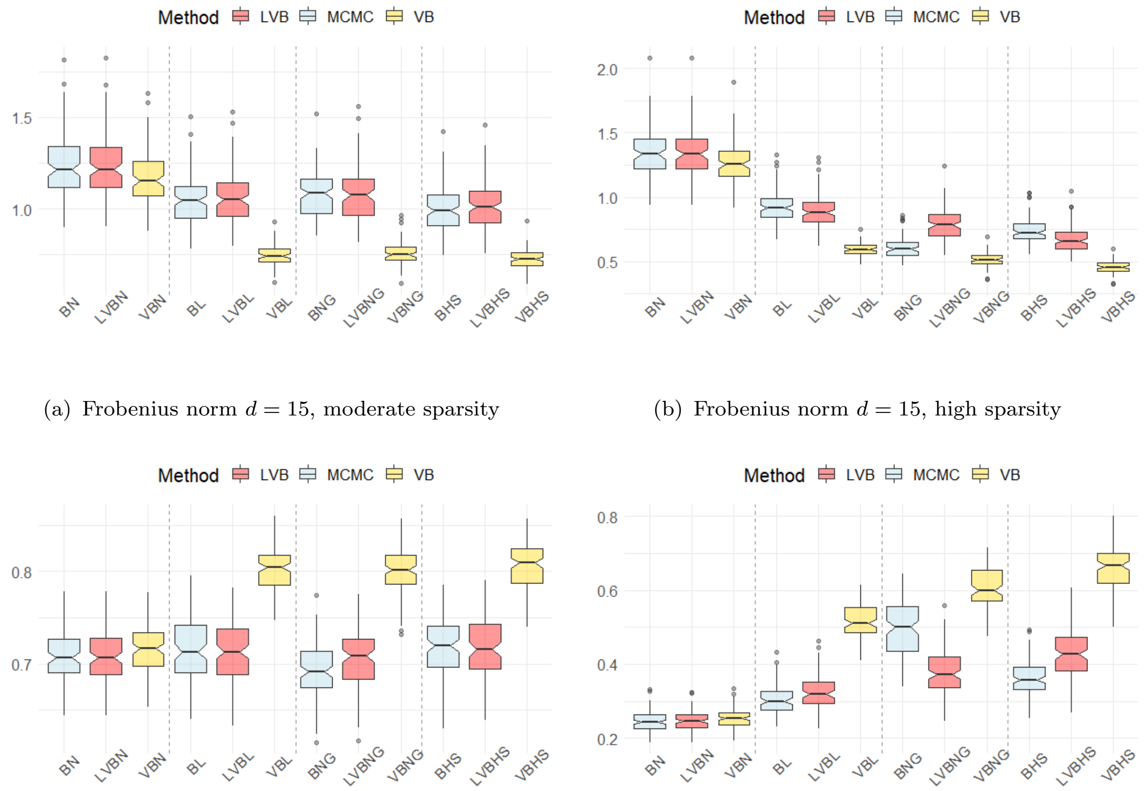


(a) $d = 15$ moderate sparsity (b) $d = 30$ moderate sparsity (c) $d = 49$ moderate sparsity

(d) $d = 15$ high sparsity (e) $d = 30$ high sparsity (f) $d = 49$ high sparsity

FIGURE A.3: Regression matrices for the simulation study. We assume both moderate level of sparsity (top, 50% of zeros) and high level of sparsity (bottom, 90% of zeros).

Additional simulation results. We now show some of the additional results on smaller dimensional simulation cases. Figure A.4 reports the Frobenius norm (top panels) and the F1-score (bottom panels) as in the main text. Similar to the larger-dimensional cases in the main text, our `VB` estimation procedure outperform both MCMC and `LVB` approach. For instance, focusing on the moderate sparsity case (i.e., 50% of zeros in the true matrix), the different priors performs equally given the estimation method, while when the sparsity is high horseshoe tends to perform better that lasso and normal-gamma. Another interesting result is that when the sparsity level is fixed at 50%, methods that work with the reparametrization of the regression matrix provide results similar to the non-informative priors and the difference with our approach is evident.



(a) Frobenius norm $d = 15$, moderate sparsity

(b) Frobenius norm $d = 15$, high sparsity

(c) F1-score norm $d = 15$, moderate sparsity

(d) F1-score norm $d = 15$, high sparsity

FIGURE A.4: Top panels report the Frobenius norm of $\mathbf{\Theta} - \widehat{\mathbf{\Theta}}$ for different hierarchical shrinkage priors and inference approaches. Bottom panels report the F1-score computed looking at the true non-null parameters in $\mathbf{\Theta}$ and the non-null parameters in the estimated matrix $\widehat{\mathbf{\Theta}}$. The box charts show the results for $N = 100$ replications, $d = 15$ and different levels of sparsity.

A.5 Additional empirical results

Computational cost of the recursive forecasts. The faster computation turns out to be key within the context of recursive forecasting. Table A.1 shows the computational time in hours, for each estimation method and across different hierarchical priors. Our approach is significantly faster than an MCMC estimation method and performs virtually on par with a linearized variational Bayes method.

	$d = 30$				$d = 49$			
	Normal	BL	NG	HS	Normal	BL	NG	HS
MCMC	16.9	17.8	160.1	17.8	19.4	20.4	155.9	20.3
LVB	0.1	0.1	100.7	0.2	0.5	0.5	94.6	0.5
VB	1.8	1.5	110.4	1.7	5.4	4.0	113.2	4.4

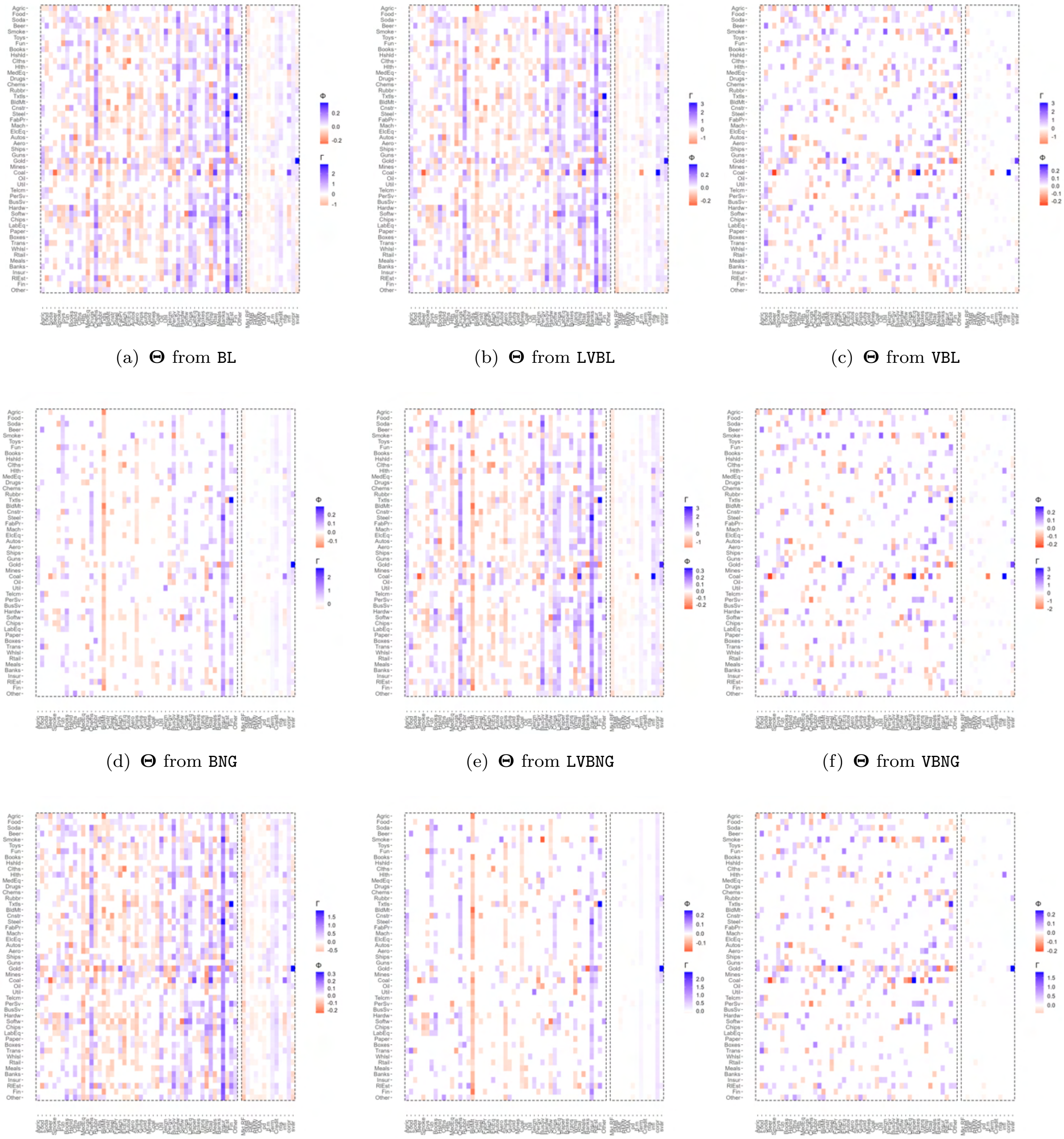
TABLE A.1: Computational time in hours to perform the empirical analysis.

In-sample estimates for 49 industries. Figure A.5 shows the in-sample posterior estimates estimates of the regression coefficients for the $d = 30$ industry case.

The posterior estimates highlight three main results. First, the $\widehat{\mathbf{\Theta}}$ obtained from the MCMC and the linearised variational Bayes tend to coincide. This is reassuring since, in principle, the linearised VB and the MCMC estimation setting should converge to similar posterior estimates (see, e.g., Gefang *et al.*, 2019; Chan and Yu, 2022).

The second main result from Figure A.5 is that for both the MCMC and the linearised VB method there are visible differences in the posterior estimates across shrinkage priors. For instance, the $\widehat{\mathbf{\Theta}}$ from the `BNG` method is arguably more sparse than the one obtained from the horseshoe prior (`BHS`). Similarly, under the linearised variational inference method, the regression parameters estimates are more sparse under the `LVBHS` compared to the Bayesian adaptive lasso (`LVBL`).

Perhaps more interesting, the third main fact that emerges from Figure A.5 is that under our variational inference method the estimates of $\mathbf{\Theta}$ are (1) more sparse compared to both MCMC and linearised VB, and (2) are remarkably similar across different shrinkage priors.



(a) **Θ** from `BL`

(b) **Θ** from `LVBL`

(c) **Θ** from `VBL`

(d) **Θ** from `BNG`

(e) **Θ** from `LVBNG`

(f) **Θ** from `VBNG`

(g) **Θ** from `BHS`

(h) **Θ** from `LVBHS`

(i) **Θ** from `VBHS`

FIGURE A.5: Posterior estimates of the regression coefficients **Θ** for different estimation methods. We report the estimates for the $d = 49$ industry case obtained from the Bayesian adaptive lasso (top panels), the adaptive normal gamma (middle panels), and the horseshoe (bottom panels).

Appendix B

This Appendix refers to Chapter 3. In particular, it provides the complete derivation of the optimal variational density approximations for both the latent stochastic volatility state and the corresponding structural parameters. Moreover, we show some additional empirical results.

B.1 Parameters optimal densities

Proposition B.1. *The optimal variational density for the regression parameter vector is* $q(\boldsymbol{\beta}) \equiv \mathsf{N}_p(\boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)})$ *where:*

$$\boldsymbol{\Sigma}_{q(\beta)} = \left(\mathbf{X}^\intercal \mathbf{H}^{-1}\mathbf{X} + \boldsymbol{\Sigma}_\beta^{-1}\right)^{-1} \qquad \boldsymbol{\mu}_{q(\beta)} = \boldsymbol{\Sigma}_{q(\beta)} \left(\mathbf{X}^\intercal \mathbf{H}^{-1}\mathbf{y} + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta\right), \tag{B.1}$$

where $\mathbf{H}^{-1} = \mathsf{Diag}\left(\mathbb{E}_h\left[\mathrm{e}^{\mathbf{h}_1}\right]\right)$ *is a diagonal matrix with elements that depend on the optimal density for the latent log-volatilities.*

Proof. The logarithm of the full conditional $(\boldsymbol{\beta}|\text{rest})$ is proportional to:

$$\begin{aligned}\log p(\boldsymbol{\beta}|\text{rest}) &\propto -\frac{1}{2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^\intercal \mathsf{diag}\left(\mathrm{e}^{\mathbf{h}_1}\right)\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right) - \frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta\right)^\intercal \boldsymbol{\Sigma}_\beta^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta\right) \\ &\propto -\frac{1}{2}\left(\boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathsf{diag}\left(\mathrm{e}^{\mathbf{h}_1}\right)\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathsf{diag}\left(\mathrm{e}^{\mathbf{h}_1}\right)\mathbf{y}\right) - \frac{1}{2}\left(\boldsymbol{\beta}^\intercal\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\intercal\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta\right).\end{aligned}$$

Compute the optimal variational density as $\log q(\boldsymbol{\beta}) = \mathbb{E}_{-\beta}\left[\log p(\boldsymbol{\beta}|\text{rest})\right]$:

$$\begin{aligned}\log q(\boldsymbol{\beta}) &\propto -\frac{1}{2}\left(\boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathsf{diag}\left(\mathbb{E}_h\left[\mathrm{e}^{\mathbf{h}_1}\right]\right)\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\left(\mathbb{E}_h\left[\mathrm{e}^{\mathbf{h}_1}\right]\right)\mathbf{y}\right) \\ &\quad - \frac{1}{2}\left(\boldsymbol{\beta}^\intercal\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\intercal\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta\right) \\ &= -\frac{1}{2}\left(\boldsymbol{\beta}^\intercal(\mathbf{X}^\intercal\mathbf{H}^{-1}\mathbf{X} + \boldsymbol{\Sigma}_\beta^{-1})\boldsymbol{\beta} - 2\boldsymbol{\beta}^\intercal(\mathbf{X}^\intercal\mathbf{H}^{-1}\mathbf{y} + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta)\right),\end{aligned}$$

where $\mathbf{H}^{-1} = \mathsf{Diag}\left(\mathbb{E}_h\left[\mathrm{e}^{\mathbf{h}_1}\right]\right)$. Take the exponential and end up with the kernel of a multivariate Gaussian distribution with parameters as in (B.1). □

Proposition B.2. *The optimal variational density for the unconditional mean of the log-volatility process is* $q(c) \equiv \mathsf{N}(\mu_{q(c)}, \sigma^2_{q(c)})$ *where:*

$$\begin{aligned}\sigma^2_{q(c)} &= (\mu_{q(1/\eta^2)}\boldsymbol{\iota}_{n+1}^\intercal \boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1} + 1/\sigma^2_c)^{-1} \\ \mu_{q(c)} &= \sigma^2_{q(c)}(\mu_{q(1/\eta^2)}\boldsymbol{\iota}_{n+1}^\intercal \boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\mu}_{q(\mathbf{h})} + \mu_c/\sigma^2_c).\end{aligned} \tag{B.2}$$

where

$$\boldsymbol{\mu}_{q(\mathbf{Q})} = \begin{bmatrix} 1 & -\mu_{q(\rho)} & \dots & 0 & 0 \\ -\mu_{q(\rho)} & 1+\mu_{q(\rho^2)} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1+\mu_{q(\rho^2)} & -\mu_{q(\rho)} \\ 0 & 0 & \dots & -\mu_{q(\rho)} & 1 \end{bmatrix}.$$

Proof. The logarithm of the full conditional $(c|\text{rest})$ is proportional to:

$$\begin{aligned}\log p(c|\text{rest}) &\propto -\frac{1}{2\eta^2}(\mathbf{h} - c\boldsymbol{\iota}_{n+1})^\intercal \mathbf{Q}(\mathbf{h} - c\boldsymbol{\iota}_{n+1}) - \frac{1}{2\sigma^2_c}(c-\mu_c)^2 \\ &\propto -\frac{1}{2\eta^2}(c^2\boldsymbol{\iota}_{n+1}^\intercal \mathbf{Q}\boldsymbol{\iota}_{n+1} - 2c\boldsymbol{\iota}_{n+1}^\intercal \mathbf{Q}\mathbf{h}) - \frac{1}{2\sigma^2_c}(c^2 - 2c\mu_c).\end{aligned}$$

Compute the optimal variational density as $\log q(c) = \mathbb{E}_{-c}\left[\log p(c|\text{rest})\right]$:

$$\begin{aligned}\log q(c) &\propto -\frac{1}{2}\mathbb{E}_{\eta^2}[1/\eta^2](c^2\boldsymbol{\iota}_{n+1}^\intercal \mathbb{E}_\rho[\mathbf{Q}]\boldsymbol{\iota}_{n+1} - 2c\boldsymbol{\iota}_{n+1}^\intercal \mathbb{E}_\rho[\mathbf{Q}]\mathbb{E}_h[\mathbf{h}]) - \frac{1}{2\sigma^2_c}(c^2 - 2c\mu_c) \\ &= -\frac{1}{2}\mu_{q(1/\eta^2)}(c^2\boldsymbol{\iota}_{n+1}^\intercal \boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1} - 2c\boldsymbol{\iota}_{n+1}^\intercal \boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\mu}_{q(h)}) - \frac{1}{2\sigma^2_c}(c^2 - 2c\mu_c) \\ &= -\frac{1}{2}\left(c^2(\mu_{q(1/\eta^2)}\boldsymbol{\iota}_{n+1}^\intercal \boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1} + 1/\sigma^2_c) - 2c(\boldsymbol{\iota}_{n+1}^\intercal \boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\mu}_{q(h)} + \mu_c/\sigma^2_c)\right),\end{aligned}$$

where $\boldsymbol{\mu}_{q(\mathbf{Q})}$ denotes the element-wise expectation of the matrix $\mathbf{Q}$. Take the exponential and end up with the kernel of an univariate Gaussian distribution with parameters as in (B.2). □

Proposition B.3. *The optimal variational density for the autoregressive parameter has the following form:*

$$\log q(\rho) \propto \frac{1}{2}\log(1-\rho^2) - \frac{1}{2}\mu_{q(1/\eta^2)}\left(\rho^2\sum_{t=1}^{n-1} a_t - 2\rho\sum_{t=0}^{n-1} b_t\right), \quad \rho \in (-1, 1) \tag{B.3}$$

with

$$a_t = \mathbb{E}_q\left[(h_t - c)^2\right] = (\mu_{q(h_t)} - \mu_{q(c)})^2 + \sigma^2_{q(h_t)} + \sigma^2_{q(c)} \tag{B.4}$$

$$b_t = \mathbb{E}_q\left[(h_t - c)(h_{t+1} - c)\right] = (\mu_{q(h_t)} - \mu_{q(c)})(\mu_{q(h_{t+1})} - \mu_{q(c)}) + \sigma_{q(h_t,h_{t+1})} + \sigma^2_{q(c)}, \tag{B.5}$$

where $\sigma_{q(h_t,h_{t+1})}$ *denotes the covariance between* h_t *and* h_{t+1} *under the approximating density* q*. Notice that* $\log q(\rho)$ *can be written as:*

$$\log q(\rho) \propto \frac{1}{2}\log(1-\rho^2) - \frac{1}{2}\mu_{q(1/\eta^2)}\left(\sum_{t=1}^{n-1} a_t\right)\left(\rho^2 - \frac{\sum_{t=0}^{n-1} b_t}{\sum_{t=1}^{n-1} a_t}\right)^2, \quad \rho \in (-1,1) \tag{B.6}$$

thus the normalizing constant and the first two moments can be found by Monte Carlo methods by sampling from an univariate Gaussian distribution with mean $\frac{\sum_{t=0}^{n-1} b_t}{\sum_{t=1}^{n-1} a_t}$ *and precision* $\mu_{q(1/\eta^2)}\left(\sum_{t=1}^{n-1} a_t\right)$.

Proof. The logarithm of the full conditional $(\rho|\text{rest})$ is proportional to:

$$\begin{aligned}\log p(\rho|\text{rest}) &\propto \frac{1}{2}\log|\mathbf{Q}| - \frac{1}{2\eta^2}(\mathbf{h} - c\boldsymbol{\iota}_{n+1})^\intercal \mathbf{Q}(\mathbf{h} - c\boldsymbol{\iota}_{n+1}) \\ &\propto \frac{1}{2}\log(1-\rho^2) - \frac{1}{2\eta^2}\left(\rho^2\sum_{t=1}^{n-1}(h_t - c)^2 - 2\rho\sum_{t=0}^{n-1}(h_t - c)(h_{t+1} - c)\right),\end{aligned}$$

for $\rho \in (-1,1)$. Compute the optimal variational density as $\log q(\rho) = \mathbb{E}_{-\rho}\left[\log p(\rho|\text{rest})\right]$:

$$\begin{aligned}\log q(\rho) &\propto \frac{1}{2}\log(1-\rho^2) - \frac{1}{2}\mathbb{E}_q\left[1/\eta^2\right]\left(\rho^2\sum_{t=1}^{n-1}\mathbb{E}_q\left[(h_t - c)^2\right] - 2\rho\sum_{t=0}^{n-1}\mathbb{E}_q\left[(h_t - c)(h_{t+1} - c)\right]\right) \\ &= \frac{1}{2}\log(1-\rho^2) - \frac{1}{2}\mu_{q(1/\eta^2)}\left(\rho^2\sum_{t=1}^{n-1} a_t - 2\rho\sum_{t=0}^{n-1} b_t\right), \quad \rho \in (-1,1),\end{aligned}$$

where a_t and b_t are as in (B.4). Take the exponential and obtain:

$$q(\rho) \propto \sqrt{1-\rho^2}\,\mathbb{I}_{\rho\in(-1,1)}\,\phi\left(\rho; \frac{\sum_{t=0}^{n-1} b_t}{\sum_{t=1}^{n-1} a_t}, \frac{1}{\mu_{q(1/\eta^2)}\sum_{t=1}^{n-1} a_t}\right),$$

where $\phi(x; m, s^2)$ denotes the density function of an univariate gaussian distribution with mean m and variance s^2. □

Proposition B.4. *The optimal variational density for the variance parameter is an inverse-gamma distribution* $q(\eta^2) \equiv \mathsf{IGa}(A_{q(\eta^2)}, B_{q(\eta^2)})$, *where:*

$$\begin{aligned} A_{q(\eta^2)} &= A + \frac{n+1}{2} \\ B_{q(\eta^2)} &= B + \frac{1}{2}(\boldsymbol{\mu}_{q(\mathbf{h})} - \mu_{q(c)}\boldsymbol{\iota}_{n+1})^\intercal \boldsymbol{\mu}_{q(\mathbf{Q})}(\boldsymbol{\mu}_{q(\mathbf{h})} - \mu_{q(c)}\boldsymbol{\iota}_{n+1}) \\ &\quad + \frac{1}{2}\left(\mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\mathbf{h})}\boldsymbol{\mu}_{q(\mathbf{Q})}\right\} + \sigma^2_{q(c)}\boldsymbol{\iota}^\intercal_{n+1}\boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1}\right), \end{aligned} \tag{B.7}$$

and recall that $\mu_{q(1/\eta^2)} = A_{q(\eta^2)}/B_{q(\eta^2)}$.

Proof. The logarithm of the full conditional $(\eta^2|\text{rest})$ is proportional to:

$$\begin{aligned} \log p(\eta^2|\text{rest}) &\propto -\frac{n+1}{2}\log\eta^2 - \frac{1}{2\eta^2}(\mathbf{h} - c\boldsymbol{\iota}_{n+1})^\intercal \mathbf{Q}(\mathbf{h} - c\boldsymbol{\iota}_{n+1}) - (A+1)\log\eta^2 - B/\eta^2 \\ &\propto -\left(A + \frac{n+1}{2} + 1\right)\log\eta^2 - \frac{1}{\eta^2}\left(B + \frac{1}{2}(\mathbf{h} - c\boldsymbol{\iota}_{n+1})^\intercal \mathbf{Q}(\mathbf{h} - c\boldsymbol{\iota}_{n+1})\right). \end{aligned}$$

Compute the optimal variational density as $\log q(\eta^2) = \mathbb{E}_{-\eta^2}\left[\log p(\eta^2|\text{rest})\right]$:

$$\log q(\eta^2) \propto -\left(A + \frac{n+1}{2} + 1\right)\log\eta^2 - \frac{1}{\eta^2}\left(B + \frac{1}{2}\mathbb{E}_{c,\rho,h}\left[(\mathbf{h} - c\boldsymbol{\iota}_{n+1})^\intercal \mathbf{Q}(\mathbf{h} - c\boldsymbol{\iota}_{n+1})\right]\right),$$

where

$$\begin{aligned} \mathbb{E}_{c,\rho,h}\left[(\mathbf{h} - c\boldsymbol{\iota}_{n+1})^\intercal \mathbf{Q}(\mathbf{h} - c\boldsymbol{\iota}_{n+1})\right] &= \mathbb{E}_{c,\rho,h}\left[\mathbf{h}^\intercal\mathbf{Q}\mathbf{h} - 2c\mathbf{h}^\intercal\mathbf{Q}\boldsymbol{\iota}_{n+1} + c^2\boldsymbol{\iota}^\intercal_{n+1}\mathbf{Q}\boldsymbol{\iota}_{n+1}\right] \\ &= \mathbb{E}_h\left[\mathbf{h}^\intercal\boldsymbol{\mu}_{q(\mathbf{Q})}\mathbf{h}\right] + \mathbb{E}_c[c^2]\boldsymbol{\iota}^\intercal_{n+1}\boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1} \\ &\qquad - 2\mu_{q(c)}\boldsymbol{\mu}^\intercal_{q(h)}\boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1} \\ &= \mathsf{tr}\left\{\mathbb{E}_h[\mathbf{h}\mathbf{h}^\intercal]\boldsymbol{\mu}_{q(\mathbf{Q})}\right\} + (\mu^2_{q(c)} + \sigma^2_{q(c)})\boldsymbol{\iota}^\intercal_{n+1}\boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1} \\ &\qquad - 2\mu_{q(c)}\boldsymbol{\mu}^\intercal_{q(h)}\boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1} \\ &= \mathsf{tr}\left\{\left(\boldsymbol{\mu}_{q(h)}\boldsymbol{\mu}^\intercal_{q(h)} + \boldsymbol{\Sigma}_{q(h)}\right)\boldsymbol{\mu}_{q(\mathbf{Q})}\right\} \\ &\qquad + (\mu^2_{q(c)} + \sigma^2_{q(c)})\boldsymbol{\iota}^\intercal_{n+1}\boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1} \\ &\qquad - 2\mu_{q(c)}\boldsymbol{\mu}^\intercal_{q(h)}\boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1} \\ &= \boldsymbol{\mu}^\intercal_{q(h)}\boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\mu}_{q(h)} + \mu^2_{q(c)}\boldsymbol{\iota}^\intercal_{n+1}\boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1} \\ &\qquad - 2\mu_{q(c)}\boldsymbol{\mu}^\intercal_{q(h)}\boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1} \\ &\qquad + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(h)}\boldsymbol{\mu}_{q(\mathbf{Q})}\right\} + \sigma^2_{q(c)}\boldsymbol{\iota}^\intercal_{n+1}\boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1} \\ &= (\boldsymbol{\mu}_{q(h)} - \mu_{q(c)}\boldsymbol{\iota}_{n+1})^\intercal\boldsymbol{\mu}_{q(\mathbf{Q})}(\boldsymbol{\mu}_{q(h)} - \mu_{q(c)}\boldsymbol{\iota}_{n+1}) \\ &\qquad + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(h)}\boldsymbol{\mu}_{q(\mathbf{Q})}\right\} + \sigma^2_{q(c)}\boldsymbol{\iota}^\intercal_{n+1}\boldsymbol{\mu}_{q(\mathbf{Q})}\boldsymbol{\iota}_{n+1}. \end{aligned}$$

Take the exponential and end up with the kernel of an inverse-gamma distribution with parameters as in (B.7). □

B.2 Homoschedastic approximation

First of all, the joint distribution of the latent states and the observations, given the set of covariates is given by:

$$\begin{aligned}\log p(\mathbf{h},\mathbf{y}|\mathbf{X}) &\propto \log p(\mathbf{y}|\mathbf{h}_1,\mathbf{X}) + \log p(\mathbf{h})\\ &= -\frac{1}{2}\boldsymbol{\iota}_n^\intercal \mathbf{h}_1 - \frac{1}{2}\mathbf{s}^\intercal \mathrm{e}^{-\mathbf{h}_1} - \frac{1}{2\eta^2}(\mathbf{h}-c\boldsymbol{\iota}_{n+1})^\intercal \mathbf{Q}(\mathbf{h}-c\boldsymbol{\iota}_{n+1}),\end{aligned} \tag{B.1}$$

where $\mathbf{s}=(s_1,\ldots,s_n)^\intercal$ with $s_t=(y_t-\mathbf{x}_t^\intercal\boldsymbol{\beta})^2$, $\mathbf{h}_1=(h_1,\ldots,h_n)^\intercal$ and $\mathrm{e}^{\mathbf{h}_1}=(\mathrm{e}^{h_1},\ldots,\mathrm{e}^{h_n})^\intercal$. Let the homoschedastic approximation be defined as $\mathbf{h}\sim \mathsf{N}_{n+1}(\mathbf{Wf},\tau^2\boldsymbol{\Gamma}^{-1})$ where $\boldsymbol{\mu}_{q(h)} = \mathbf{Wf}$ is the mean vector and $\boldsymbol{\Sigma}_{q(h)} = \tau^2\boldsymbol{\Gamma}^{-1}$ is the variance-covariance matrix. More precisely, Γ is a tridiagonal precision matrix with diagonal elements $\Gamma_{1,1} = \Gamma_{n+1,n+1} = 1$ and $\Gamma_{i,i} = 1+\gamma^2$ for $i=2,\ldots,n$, and off-diagonal elements $\Gamma_{i,j}=-\gamma$ if $|i-j|=1$ and 0 elsewhere (see Rue and Held, 2005). Under this setting, the density function of the approximate distribution is given by:

$$\log\phi(\mathbf{h}|\mathbf{Wf},\tau^2\boldsymbol{\Gamma}^{-1}) \propto -\frac{n+1}{2}\log(\tau^2) - \frac{n}{2}\log(1-\gamma^2) - \frac{1}{2\tau^2}(\mathbf{h}-\mathbf{Wf})^\intercal\boldsymbol{\Gamma}(\mathbf{h}-\mathbf{Wf}). \tag{B.2}$$

Define the variational lower bound (ELBO) as:

$$\begin{aligned}\psi(\mathbf{f},\tau^2,\gamma) &= \mathbb{E}_q(\log p(\mathbf{h},\mathbf{y})) - \mathbb{E}_q(\log q(\mathbf{h}))\\ &\propto -\frac{1}{2}\boldsymbol{\iota}_n^\intercal \mathbf{W}_1\mathbf{f} - \frac{1}{2}\boldsymbol{\mu}_{q(\mathbf{s})}^\intercal \mathrm{e}^{-\mathbf{W}_1\mathbf{f}+\frac{1}{2}\tau^2\boldsymbol{\iota}_n}\\ &\quad -\frac{1}{2}\mu_{q(1/\eta^2)}(\mathbf{Wf}-\mu_{q(c)}\boldsymbol{\iota}_{n+1})^\intercal\boldsymbol{\mu}_{q(\mathbf{Q})}(\mathbf{Wf}-\mu_{q(c)}\boldsymbol{\iota}_{n+1})\\ &\quad -\frac{1}{2}\mu_{q(1/\eta^2)}\tau^2\mathsf{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}_{q(\mathbf{Q})})\\ &\quad +\frac{n+1}{2}\log(\tau^2)+\frac{n}{2}\log(1-\gamma^2),\end{aligned} \tag{B.3}$$

where $\boldsymbol{\mu}_{q(\mathbf{s})} = (\mu_{q(s_1)},\ldots,\mu_{q(s_n)})^\intercal$ with $\mu_{q(s_t)} = (y_t-\mathbf{x}_t^\intercal\boldsymbol{\mu}_{q(\beta)})^2 + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\beta)}\mathbf{x}_t\mathbf{x}_t^\intercal\right\}$, and $\mathbf{W}_1\in\mathbb{R}^{n\times k}$ denotes the matrix obtained by deleting the first row of $\mathbf{W}$. Moreover

$$\mathsf{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}_{q(\mathbf{Q})}) = 2+(1+\mu_{q(\rho^2)})(n-1)-2n\gamma\mu_{q(\rho)}.$$

Let $\boldsymbol{\xi} = (\mathbf{f}, \tau^2, \gamma)$ be the collection of the optimal parameters, the optimization we have to solve is equal to $\widehat{\boldsymbol{\xi}} = \arg\max_{\xi} \psi(\mathbf{f}, \tau^2, \gamma)$, where the objective function $\psi(\mathbf{f}, \tau^2, \gamma)$ has gradient equal to

$$\nabla_{\xi}\psi(\mathbf{f}, \tau^2, \gamma) = \begin{bmatrix} \nabla_{\mathbf{f}}\psi(\mathbf{f}, \tau^2, \gamma) \\ \nabla_{\tau^2}\psi(\mathbf{f}, \tau^2, \gamma) \\ \nabla_{\gamma}\psi(\mathbf{f}, \tau^2, \gamma) \end{bmatrix},$$

where

$$\begin{aligned}
\nabla_{\mathbf{f}}\psi(\mathbf{f}, \tau^2, \gamma) &= -\frac{1}{2}\mathbf{W}^\intercal[0, \boldsymbol{\iota}_n^\intercal]^\intercal + \frac{1}{2}\mathbf{W}^\intercal\left([0, \boldsymbol{\mu}_{q(\mathbf{s})}^\intercal]^\intercal \odot \mathrm{e}^{-\mathbf{W}\mathbf{f}+\frac{1}{2}\tau^2\boldsymbol{\iota}_{n+1}}\right) \\
&\quad - \mu_{q(1/\eta^2)}\mathbf{W}^\intercal\boldsymbol{\mu}_{q(\mathbf{Q})}(\mathbf{W}\mathbf{f} - \mu_{q(c)}\boldsymbol{\iota}_{n+1}), && \text{(B.4)} \\
\nabla_{\tau^2}\psi(\mathbf{f}, \tau^2, \gamma) &= -\frac{1}{4}(\boldsymbol{\mu}_{q(\mathbf{s})} \odot \boldsymbol{\iota}_n)^\intercal\mathrm{e}^{-\mathbf{W}_1\mathbf{f}+\frac{1}{2}\tau^2\boldsymbol{\iota}_n} \\
&\quad - \frac{1}{2}\mu_{q(1/\eta^2)}(2 + (1 + \mu_{q(\rho^2)})(n-1) - 2n\gamma\mu_{q(\rho)}) + \frac{n+1}{2\tau^2}, && \text{(B.5)} \\
\nabla_{\gamma}\psi(\mathbf{f}, \tau^2, \gamma) &= n\tau^2\mu_{q(1/\eta^2)}\mu_{q(\rho)} - \frac{n\gamma}{1-\gamma^2}, && \text{(B.6)}
\end{aligned}$$

and Hessian equal to:

$$\mathcal{H}_{\xi} = \begin{bmatrix} \nabla^2_{\mathbf{f},\mathbf{f}}\psi(\mathbf{f}, \tau^2, \gamma) & \nabla^2_{\mathbf{f},\tau^2}\psi(\mathbf{f}, \tau^2, \gamma) & \nabla^2_{\mathbf{f},\gamma}\psi(\mathbf{f}, \tau^2, \gamma) \\ \nabla^2_{\mathbf{f},\tau^2}\psi(\mathbf{f}, \tau^2, \gamma) & \nabla^2_{\tau^2,\tau^2}\psi(\mathbf{f}, \tau^2, \gamma) & \nabla^2_{\tau^2,\gamma}\psi(\mathbf{f}, \tau^2, \gamma) \\ \nabla^2_{\mathbf{f},\gamma}\psi(\mathbf{f}, \tau^2, \gamma) & \nabla^2_{\tau^2,\gamma}\psi(\mathbf{f}, \tau^2, \gamma) & \nabla^2_{\gamma,\gamma}\psi(\mathbf{f}, \tau^2, \gamma) \end{bmatrix},$$

with

$$\begin{aligned}
\nabla^2_{\mathbf{f},\mathbf{f}}\psi(\mathbf{f}, \tau^2, \gamma) &= -\frac{1}{2}\mathbf{W}^\intercal\left\{\mathsf{Diag}\left[[0, \boldsymbol{\mu}_{q(\mathbf{s})}^\intercal]^\intercal \odot \mathrm{e}^{-\mathbf{W}\mathbf{f}+\frac{1}{2}\tau^2\boldsymbol{\iota}_{n+1}}\right] + \mu_{q(1/\eta^2)}\boldsymbol{\mu}_{q(\mathbf{Q})}\right\}\mathbf{W} \\
& && \text{(B.7)} \\
\nabla^2_{\tau^2,\tau^2}\psi(\mathbf{f}, \tau^2, \gamma) &= -\frac{1}{8}(\boldsymbol{\mu}_{q(\mathbf{s})} \odot \boldsymbol{\iota}_n)^\intercal\mathrm{e}^{-\mathbf{W}_1\mathbf{f}+\frac{1}{2}\tau^2\boldsymbol{\iota}_n} - \frac{n+1}{2\tau^4} && \text{(B.8)} \\
\nabla^2_{\gamma,\gamma}\psi(\mathbf{f}, \tau^2, \gamma) &= -\frac{n(1+\gamma^2)}{(1-\gamma^2)^2} && \text{(B.9)} \\
\nabla^2_{\mathbf{f},\tau^2}\psi(\mathbf{f}, \tau^2, \gamma) &= \frac{1}{4}\mathbf{W}^\intercal([0, \boldsymbol{\mu}_{q(\mathbf{s})}^\intercal]^\intercal \odot \mathrm{e}^{-\mathbf{W}\mathbf{f}+\frac{1}{2}\tau^2\boldsymbol{\iota}_{n+1}}) && \text{(B.10)} \\
\nabla^2_{\mathbf{f},\gamma}\psi(\mathbf{f}, \tau^2, \gamma) &= \mathbf{0}_k && \text{(B.11)} \\
\nabla^2_{\tau^2,\gamma}\psi(\mathbf{f}, \tau^2, \gamma) &= n\mu_{q(\rho)}\mu_{q(1/\eta^2)} && \text{(B.12)}
\end{aligned}$$

where $\mathbf{a} = \mathsf{diag}(\mathbf{A})$ denotes the operator that returns the vector $\mathbf{a} \in \mathbb{R}^n$ of elements belonging to the main diagonal of the square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, while $\mathbf{A} = \mathsf{Diag}(\mathbf{a})$

denotes the operator that returns a diagonal square matrix $\mathbf{A} \in \mathbb{S}_{+}^{n}$ whose entries consist of the corresponding elements of the vector $\mathbf{a} \in \mathbb{R}^{n}$.

B.3 Heteroschedastic approximation

Let the heteroschedastic approximation be defined as $\mathbf{h} \sim \mathsf{N}_{n+1}(\mathbf{W}\mathbf{f}_{q(h)}, \boldsymbol{\Sigma}_{q(h)})$ where the mean vector is $\boldsymbol{\mu}_{q(h)} = \mathbf{W}\mathbf{f}_{q(h)}$. To find the optimal parameters of the approximating density $(\mathbf{f}_{q(h)}, \boldsymbol{\Sigma}_{q(h)})$, we have to solve the following optimization problem:

$$\widehat{\boldsymbol{\xi}} = \arg\max_{\xi} \psi(\mathbf{f}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}), \tag{B.1}$$

where $\psi(\mathbf{f}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}) = \mathbb{E}_q(\log p(\mathbf{h}, \mathbf{y})) - \mathbb{E}_q(\log q(\mathbf{h}))$ is called variational lower bound (ELBO). To this aim, we can exploit a result provided by Rohde and Wand (2016) valid when the approximating density is a multivariate gaussian distribution. The latter states a closed-form update scheme for the variational parameters:

$$\boldsymbol{\Sigma}^{new} = \left[\nabla^2_{\boldsymbol{\mu},\boldsymbol{\mu}} S(\boldsymbol{\mu}^{old}, \boldsymbol{\Sigma}^{old})\right]^{-1} \tag{B.2}$$

$$\boldsymbol{\mu}^{new} = \boldsymbol{\mu}^{old} + \boldsymbol{\Sigma}^{new} \nabla_{\boldsymbol{\mu}} S(\boldsymbol{\mu}^{old}, \boldsymbol{\Sigma}^{old}), \tag{B.3}$$

where $\nabla_{\boldsymbol{\mu}} S(\boldsymbol{\mu}^{old}, \boldsymbol{\Sigma}^{old})$ and $\nabla^2_{\boldsymbol{\mu},\boldsymbol{\mu}} S(\boldsymbol{\mu}^{old}, \boldsymbol{\Sigma}^{old})$ denote the first and second derivative of $S(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\mu}$ and evaluated at $(\boldsymbol{\mu}^{old}, \boldsymbol{\Sigma}^{old})$. The function S is the so called *non-entropy function* which is given by $\mathbb{E}_q(\log p(\mathbf{h}, \mathbf{y}))$. In our scenario, we have that

$$\begin{aligned} S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}) = &-\frac{1}{2}[0, \boldsymbol{\iota}_n^{\mathsf{T}}]\boldsymbol{\mu}_{q(h)} - \frac{1}{2}[0, \boldsymbol{\mu}_{q(\mathbf{s})}^{\mathsf{T}}]\mathrm{e}^{-\boldsymbol{\mu}_{q(h)} + \frac{1}{2}\boldsymbol{\sigma}^2_{q(\mathbf{h})}} - \frac{1}{2}\mu_{q(1/\eta^2)}\mathsf{tr}(\boldsymbol{\Sigma}_{q(\mathbf{h})}\boldsymbol{\mu}_{q(\mathbf{Q})}) \\ &- \frac{1}{2}\mu_{q(1/\eta^2)}(\boldsymbol{\mu}_{q(h)} - \mu_{q(c)}\boldsymbol{\iota}_{n+1})^{\mathsf{T}}\boldsymbol{\mu}_{q(\mathbf{Q})}(\boldsymbol{\mu}_{q(h)} - \mu_{q(c)}\boldsymbol{\iota}_{n+1}), \end{aligned} \tag{B.4}$$

where $\boldsymbol{\sigma}^2_{q(h)} = \mathsf{diag}(\boldsymbol{\Sigma}_{q(h)})$ is the vector of variances and the diag operator extracts the diagonal vector from the input matrix. Moreover, we obtain:

$$\begin{aligned} \nabla_{\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}) = &-\frac{1}{2}[0, \boldsymbol{\iota}_n^{\mathsf{T}}]^{\mathsf{T}} + \frac{1}{2}[0, \boldsymbol{\mu}_{q(\mathbf{s})}^{\mathsf{T}}]^{\mathsf{T}} \odot \mathrm{e}^{-\boldsymbol{\mu}_{q(h)} + \frac{1}{2}\boldsymbol{\sigma}^2_{q(h)}} \\ &- \mu_{q(1/\eta^2)}\boldsymbol{\mu}_{q(\mathbf{Q})}(\boldsymbol{\mu}_{q(h)} - \mu_{q(c)}\boldsymbol{\iota}_{n+1}), \end{aligned} \tag{B.5}$$

$$\nabla^2_{\boldsymbol{\mu}_{q(h)}\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}) = -\frac{1}{2}\mathsf{Diag}\left[[0, \boldsymbol{\mu}_{q(\mathbf{s})}^{\mathsf{T}}]^{\mathsf{T}} \odot \mathrm{e}^{-\boldsymbol{\mu}_{q(h)} + \frac{1}{2}\boldsymbol{\sigma}^2_{q(h)}}\right] - \mu_{q(1/\eta^2)}\boldsymbol{\mu}_{q(\mathbf{Q})}, \tag{B.6}$$

where $\boldsymbol{\iota}_n$ is an n-dimensional vector of ones, $\mu_{q(1/\eta^2)}$ is the variational mean of $1/\eta^2$, $\boldsymbol{\mu}_{q(\mathbf{Q})}$ is the element-wise variational mean of $\mathbf{Q}$, and $\odot$ denotes the Hadamard product.

Then, since $\boldsymbol{\mu}_{q(h)} = \mathbf{W}\mathbf{f}_{q(h)}$, the updating scheme becomes:

$$\boldsymbol{\Sigma}_{q(h)}^{new} = \left[\nabla^2_{\boldsymbol{\mu}_{q(h)}\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}^{old}, \boldsymbol{\Sigma}_{q(h)}^{old})\right]^{-1}, \tag{B.7}$$

$$\mathbf{f}_{q(h)}^{new} = \mathbf{f}_{q(h)}^{old} + \mathbf{W}^{+}\, \boldsymbol{\Sigma}_{q(h)}^{new} \nabla_{\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}^{old}, \boldsymbol{\Sigma}_{q(h)}^{old}), \tag{B.8}$$

$$\boldsymbol{\mu}_{q(h)}^{new} = \mathbf{W}\mathbf{f}_{q(h)}^{new}, \tag{B.9}$$

with $\mathbf{W}^{+} = (\mathbf{W}^{\intercal}\mathbf{W})^{-1}\mathbf{W}^{\intercal}$ the left Moore–Penrose pseudo-inverse of $\mathbf{W}$.

B.4 Additional empirical results

This Section shows additional empirical results considering transaction costs with leverage constraints.

The results in Tables 3.4–3.5 show that when conservative levels of transaction costs to implement volatility targeting are considered, the performance of standard volatility targeting methods substantially deteriorates. Standard volatility targeting strategies are not designed to mitigate transaction costs. Hence, we next evaluate whether by reducing liquidity demand via capping leverage render volatility targeting still profitable after costs. This approach does not necessarily aim at an optimal allocation from the perspective of a mean-variance investor. Rather, it is a simple, yet effective, risk-management approach that aims to regularise the capital exposure to the original equity trading strategy. We follow Moreira and Muir (2017); Cederburg *et al.* (2020); Barroso and Detzel (2021); Wang and Yan (2021) and consider two different levels of leverage constraint; one that cap the leverage at 1.5 times the original factor, and a second less restrictive that cap leverage at 5 times the exposure to the original factor.

Table B.1 reports the Sharpe and the Sortino ratios considering the same level of transaction costs as in Section 3.4.2, namely 14 and 50 basis points of the notional trading exposure. Panel A shows the results for a 500% leverage constraint. For a conservative 50 basis points transaction costs our SSV produces the highest Sharpe and Sortino ratios among the volatility targeting methods, on average across the 158 equity strategies. For instance, the SSV generates a 0.23 Sharpe ratio on average against a dismal -0.10 annualised Sharpe ratio from the RV. Compared to the unmanaged portfolios, the number of significantly higher SRs is also higher for the SSV case. For instance, none of the re-scaled portfolios with RV has a positive and significant SR differential against 7% of the portfolios re-scaled with SSV.

Panel B shows the results for a more restrictive leverage constraint, which forces the exposure from volatility targeting no more than 1.5 times the original factor portfolio.

Consistent with Moreira and Muir (2017); Barroso and Detzel (2021), a tighter cap does indeed regularise more the performance of volatility targeting across all competing methods. Nevertheless, the performance of our SSV portfolio is quite stable across different levels of leverage constraints. Interestingly, unlike the case without leverage constraints, the RV6 plus leverage cap proves to be a quite competitive benchmark volatility targeting method.

Table B.2 reports the results for the spanning regressions. The top panels report the estimated alphas ($\widehat{\alpha}$ in %). When considering a conservative notional trading cost of 50 basis points, our SSV volatility forecast generates a positive alpha of 0.46% annualised. This is against a large and negative alpha from the RV, RV AR, HAR, and SV methods. Perhaps more importantly, our SSV method generates a significantly positive alpha for 21% of the equity strategies in our sample, against, for instance, 3%, 17%, and 14% from the RV, RV6 and Garch models, respectively.

The appraisal ratio AR reported in the middle panel of Table B.2 confirms that our SSV substantially improves upon standard volatility targeting based on RV, especially when more conservative transaction costs are factored in. For instance, with 50 basis points of trading costs the SSV is the only method that can still generate a positive appraisal ratio together with the RV6 long-term realised variance method. By comparison, the RV, Garch and RV AR all generate significantly negative ARs. The bottom panels report the difference in the certainty equivalent return between and investor that can access both the volatility-managed and the original portfolio, and an investor constrained to invest in the original portfolio only. The utility gain $\Delta CER(\%)$ is highly in favour of our SSV volatility targeting. For instance, for 14 (50) basis points of transaction costs, our SSV method generates a 12% (8%) utility gain. This compares to the 7% from the HAR with 14 basis points and 2.2% from the RV6 with 50 basis points of transaction costs.

Table B.3 reports the spanning regression results with a tighter leverage cap of 1.5. The results are largely in line with Table B.2. That is, the RV6 does indeed represents a challenging benchmark for our SSV method when it comes to the estimated alphas. However, the $\Delta CER(\%)$ from the combination strategy is substantially in favour of our smoothing volatility targeting. For instance, the $\Delta CER(\%)$ from the SSV is 9.52% (13.8%) with 50 (14) basis points of notional transaction costs, against a 4.5% (8/2%) from the RV6 volatility targeting.

Table B.1: Volatility-managed portfolios with leverage constraints.

This table compares the performance of volatility-managed and original portfolios (U) for the cross section of 158 equity strategies. For a given factor, the volatility-managed factor return in month t is based on a forecast of the conditional variance. The volatility-managed weights are capped so that the maximum leverage attainable is 500% (panel A) or 50% (panel B) of the original factor exposure. In addition to our smoothing volatility forecast (SSV), the variance forecasts are from a simple AR(1) fitted on the realised variance (RV AR), an alternative six-month window to estimate the longer-term realised variance (RV6), a long-memory model for volatility forecast as proposed by Corsi (2009) (HAR), a standard AR(1) latent stochastic volatility model (SV), and a plain GARCH(1,1) specification (Garch). For each volatility targeting method we report the mean annualised Sharpe ratio, Sortino ratio and maximum drawdown (in %), as well as their 2.5th, 25th, 50th, 75th, and 97.5th percentiles in the cross section of equity strategy. In addition, we report the fraction of volatility-managed portfolios that generate a Sharpe ratio which is statistically different from the unscaled strategy (see, Ledoit and Wolf, 2008), and is either positive or negative. The table reports the results for two levels of transaction costs, 14 and 50 basis points of the notional value traded to implement volatility targeting.

Panel A: 500% leverage constraint

	14 basis points								50 basis points							
	U	RV	RV6	RV AR	HAR	Garch	SV	SSV	U	RV	RV6	RV AR	HAR	Garch	SV	SSV
SR																
Mean	0.24	0.17	0.27	0.21	0.23	0.23	0.23	0.25	0.24	-0.10	0.21	0.01	0.13	0.16	0.14	0.23
Percentiles																
2.5	-0.12	-0.32	-0.25	-0.28	-0.26	-0.23	-0.24	-0.20	-0.12	-0.66	-0.34	-0.52	-0.40	-0.30	-0.32	-0.22
25	0.08	-0.03	0.05	0.00	0.02	0.00	-0.01	0.05	0.08	-0.29	-0.02	-0.19	-0.08	-0.06	-0.09	0.03
50	0.22	0.15	0.24	0.20	0.21	0.23	0.26	0.23	0.22	-0.11	0.20	0.00	0.11	0.17	0.16	0.21
75	0.37	0.36	0.47	0.41	0.40	0.40	0.39	0.42	0.37	0.06	0.40	0.17	0.27	0.34	0.30	0.39
97.5	0.63	0.73	0.82	0.74	0.70	0.76	0.75	0.68	0.63	0.53	0.75	0.57	0.62	0.71	0.66	0.66
p< 0.05 & SR> 0		1.90	6.33	3.80	3.80	7.59	6.96	8.86		0.00	3.80	0.00	1.27	3.80	1.90	6.96
p< 0.05 & SR< 0		15.19	2.53	12.03	6.33	12.66	12.66	5.70		75.95	12.66	65.82	37.97	28.48	36.08	11.39
Sortino																
Mean	1.44	1.11	1.68	1.31	1.36	1.40	1.39	1.50	1.44	-0.61	1.29	0.05	0.75	0.99	0.86	1.38
Percentiles																
2.5	-0.79	-1.92	-1.40	-1.61	-1.52	-1.33	-1.44	-1.15	-0.79	-4.16	-1.85	-3.05	-2.33	-1.75	-1.93	-1.27
25	0.48	-0.20	0.31	-0.02	0.12	0.03	-0.05	0.32	0.48	-1.78	-0.09	-1.22	-0.47	-0.39	-0.53	0.21
50	1.36	0.88	1.48	1.16	1.27	1.49	1.52	1.37	1.36	-0.77	1.11	0.02	0.68	1.07	1.05	1.25
75	2.16	2.21	2.76	2.30	2.36	2.37	2.30	2.34	2.16	0.36	2.31	1.05	1.59	2.00	1.75	2.21
97.5	3.49	5.22	4.88	5.02	4.31	4.64	4.35	4.14	3.49	3.75	4.54	3.86	3.87	4.42	3.83	4.04

Panel B: 50% leverage constraint

	14 basis points								50 basis points							
	U	RV	RV6	RV AR	HAR	Garch	SV	SSV	U	RV	RV6	RV AR	HAR	Garch	SV	SSV
SR																
Mean	0.24	0.22	0.28	0.24	0.24	0.25	0.25	0.25	0.24	0.04	0.24	0.11	0.16	0.20	0.19	0.24
Percentiles																
2.5	-0.12	-0.30	-0.21	-0.26	-0.24	-0.22	-0.21	-0.19	-0.12	-0.50	-0.28	-0.40	-0.34	-0.26	-0.27	-0.20
25	0.08	0.01	0.07	0.02	0.03	0.02	0.01	0.06	0.08	-0.15	0.03	-0.09	-0.03	-0.03	-0.04	0.05
50	0.22	0.19	0.26	0.20	0.22	0.24	0.24	0.21	0.22	0.04	0.23	0.09	0.14	0.19	0.19	0.20
75	0.37	0.40	0.46	0.41	0.41	0.43	0.42	0.42	0.37	0.23	0.42	0.28	0.33	0.37	0.35	0.41
97.5	0.63	0.74	0.81	0.72	0.70	0.71	0.73	0.68	0.63	0.59	0.77	0.60	0.62	0.67	0.67	0.66
p< 0.05 & SR> 0		1.90	6.33	2.53	3.80	7.59	6.96	4.43		0.63	5.06	1.27	1.90	4.43	4.43	4.43
p< 0.05 & SR< 0		10.13	1.90	5.70	5.70	8.86	8.23	4.43		55.06	4.43	43.67	25.95	20.25	25.32	6.96
Sortino																
Mean	1.44	1.34	1.66	1.42	1.42	1.45	1.44	1.48	1.44	0.28	1.44	0.66	0.97	1.17	1.10	1.40
Percentiles																
2.5	-0.79	-1.67	-1.27	-1.46	-1.35	-1.27	-1.25	-1.07	-0.79	-2.99	-1.55	-2.30	-1.91	-1.48	-1.60	-1.18
25	0.48	0.06	0.41	0.16	0.17	0.14	0.08	0.34	0.48	-0.95	0.18	-0.57	-0.16	-0.15	-0.24	0.26
50	1.36	1.19	1.55	1.21	1.33	1.42	1.46	1.27	1.36	0.28	1.37	0.53	0.84	1.21	1.17	1.21
75	2.16	2.40	2.66	2.49	2.43	2.41	2.34	2.41	2.16	1.46	2.47	1.73	1.95	2.12	1.98	2.31
97.5	3.49	4.73	4.74	4.55	4.19	4.42	4.37	4.13	3.49	4.06	4.54	3.99	3.80	4.21	4.05	4.06

Table B.2: Spanning regression results with 500% leverage constraint.

This table reports the results from a spanning regression of the form $y_t^\sigma = \alpha + \beta y_t + \epsilon_t$, with y_t^σ the returns on the volatility managed portfolio and y_t^σ its unscaled counterpart. The volatility-managed weights are capped so that the maximum leverage attainable is 500% of the original factor exposure. We report the estimated alphas ($\widehat{\alpha}$ in %), the appraisal ratio $AR = \widehat{\alpha}/\widehat{\sigma}_\varepsilon$ and the difference in the certainty equivalent return between and investor that can access both the volatility-managed and the original portfolio, and an investor constrained to invest in the original portfolio only ΔCER. In addition to our smoothing volatility forecast (`SSV`), the variance forecasts are from a simple AR(1) fitted on the realised variance (`RV AR`), an alternative six-month window to estimate the longer-term realised variance (`RV6`), a long-memory model for volatility forecast as proposed by Corsi (2009) (`HAR`), a standard AR(1) latent stochastic volatility model (`SV`), and a plain GARCH(1,1) specification (`Garch`). For each volatility targeting method we report the mean annualised Sharpe ratio, Sortino ratio and maximum drawdown (in %), as well as their 2.5th, 25th, 50th, 75th, and 97.5th percentiles in the cross section of equity strategy. In addition, we report the fraction of volatility-managed alphas that are significant and either positive or negative. The table reports the results for two levels of transaction costs, 14 and 50 basis points of the notional value traded to implement volatility targeting.

	14 basis points							50 basis points						
	RV	RV3	RV AR	HAR	Garch	SV	SV5	RV	RV3	RV AR	HAR	Garch	SV	SV5
α(%)														
Mean	0.56	1.39	0.67	0.54	0.91	0.79	0.66	-2.08	0.78	-1.38	-0.45	0.24	-0.08	0.46
Percentiles														
2.5	-2.92	-2.11	-2.60	-2.16	-2.72	-2.85	-1.71	-5.80	-2.76	-5.36	-3.11	-3.43	-3.89	-1.92
25	-0.97	-0.31	-0.69	-0.49	-0.54	-0.57	-0.40	-3.48	-0.86	-2.50	-1.42	-1.24	-1.44	-0.62
50	0.11	0.84	0.18	0.30	0.40	0.33	0.25	-2.51	0.32	-1.69	-0.71	-0.28	-0.47	0.06
75	1.15	1.92	1.04	0.89	1.47	1.18	1.01	-1.49	1.25	-0.89	-0.01	0.86	0.46	0.83
97.5	5.52	7.57	5.39	4.96	6.05	5.66	3.53	2.51	6.71	2.63	3.91	5.02	4.34	3.30
$p < 0.05$ & $\alpha > 0$	12.03	30.38	13.92	15.82	27.85	20.89	28.48	3.16	17.72	4.43	6.96	13.92	8.86	21.52
$p < 0.05$ & $\alpha < 0$	15.19	3.16	12.03	8.86	13.29	13.92	10.13	70.25	13.92	59.49	36.08	23.42	32.28	15.82
AR(%)														
Mean	0.01	0.04	0.01	0.02	0.02	0.02	0.02	-0.10	0.01	-0.08	-0.04	-0.01	-0.03	0.01
Percentiles														
2.5	-0.11	-0.07	-0.10	-0.09	-0.10	-0.10	-0.10	-0.25	-0.10	-0.22	-0.17	-0.14	-0.15	-0.11
25	-0.04	-0.02	-0.04	-0.03	-0.03	-0.03	-0.03	-0.15	-0.04	-0.13	-0.09	-0.07	-0.09	-0.05
50	0.00	0.04	0.01	0.02	0.02	0.02	0.02	-0.10	0.02	-0.08	-0.05	-0.02	-0.03	0.01
75	0.05	0.08	0.05	0.05	0.07	0.06	0.08	-0.06	0.05	-0.04	0.00	0.04	0.02	0.06
97.5	0.17	0.19	0.17	0.16	0.17	0.16	0.15	0.09	0.17	0.10	0.12	0.13	0.13	0.14
ΔCER(%)														
Mean	1.41	3.52	6.68	7.16	2.97	0.77	11.99	-4.53	2.24	-2.40	1.85	1.67	-1.16	8.26
Percentiles														
2.5	-1.01	-0.17	-0.62	-0.08	-0.49	-0.24	-0.01	-6.80	-0.61	-4.22	-0.35	-0.89	-0.82	-0.01
25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.25	0.00	0.00	0.00	0.00	0.00	0.00
50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
75	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
97.5	18.67	35.18	34.75	28.94	23.41	18.11	16.31	0.96	28.34	4.18	20.08	16.93	9.45	14.89

TABLE B.3: Spanning regression results with 50% leverage constraint.

This table reports the results from a spanning regression of the form $y_t^\sigma = \alpha + \beta y_t + \epsilon_t$, with y_t^σ the returns on the volatility managed portfolio and y_t^σ its unscaled counterpart. The volatility-managed weights are capped so that the maximum leverage attainable is 50% of the original factor exposure. We report the estimated alphas ($\widehat{\alpha}$ in %), the appraisal ratio $AR = \widehat{\alpha}/\widehat{\sigma}_\varepsilon$ and the difference in the certainty equivalent return between and investor that can access both the volatility-managed and the original portfolio, and an investor constrained to invest in the original portfolio only ΔCER. In addition to our smoothing volatility forecast (SSV), the variance forecasts are from a simple AR(1) fitted on the realised variance (RV AR), an alternative six-month window to estimate the longer-term realised variance (RV6), a long-memory model for volatility forecast as proposed by Corsi (2009) (HAR), a standard AR(1) latent stochastic volatility model (SV), and a plain GARCH(1,1) specification (Garch). For each volatility targeting method we report the mean annualised Sharpe ratio, Sortino ratio and maximum drawdown (in %), as well as their 2.5th, 25th, 50th, 75th, and 97.5th percentiles in the cross section of equity strategy. In addition, we report the fraction of volatility-managed alphas that are significant and either positive or negative. The table reports the results for two levels of transaction costs, 14 and 50 basis points of the notional value traded to implement volatility targeting.

	14 basis points							50 basis points						
	RV	RV3	RV AR	HAR	Garch	SV	SV5	RV	RV3	RV AR	HAR	Garch	SV	SV5
α(%)														
Mean	0.47	0.88	0.50	0.48	0.62	0.58	0.44	-0.75	0.61	-0.51	-0.19	0.23	0.10	0.31
Percentiles														
2.5	-1.58	-1.04	-1.44	-1.30	-1.95	-1.90	-1.48	-2.86	-1.34	-2.51	-1.98	-2.29	-2.36	-1.61
25	-0.44	-0.12	-0.42	-0.35	-0.20	-0.31	-0.31	-1.73	-0.41	-1.44	-1.03	-0.72	-0.81	-0.44
50	0.24	0.60	0.26	0.25	0.37	0.32	0.25	-1.00	0.31	-0.77	-0.41	-0.05	-0.14	0.10
75	0.95	1.24	0.94	0.83	1.10	0.93	0.83	-0.24	0.99	-0.07	0.18	0.78	0.48	0.70
97.5	3.34	4.34	3.39	3.57	4.39	4.21	2.82	2.11	4.02	2.35	2.94	3.88	3.62	2.68
$p< 0.05$ & $\alpha > 0$	15.82	28.48	15.82	17.09	25.32	19.62	27.22	5.70	18.99	6.33	8.23	15.82	12.66	20.25
$p< 0.05$ & $\alpha < 0$	10.76	1.90	6.96	5.70	11.39	8.23	8.23	48.10	6.33	41.77	24.68	20.25	24.68	12.66
AR(%)														
Mean	0.02	0.04	0.02	0.02	0.03	0.02	0.02	-0.06	0.02	-0.05	-0.02	0.00	-0.01	0.01
Percentiles														
2.5	-0.10	-0.06	-0.09	-0.07	-0.09	-0.09	-0.09	-0.20	-0.09	-0.18	-0.14	-0.12	-0.13	-0.11
25	-0.03	-0.01	-0.03	-0.02	-0.03	-0.03	-0.03	-0.11	-0.02	-0.09	-0.07	-0.06	-0.06	-0.04
50	0.02	0.04	0.02	0.02	0.02	0.02	0.02	-0.07	0.02	-0.05	-0.03	0.00	-0.01	0.01
75	0.05	0.07	0.05	0.05	0.07	0.06	0.07	-0.01	0.06	0.00	0.01	0.05	0.03	0.06
97.5	0.15	0.19	0.15	0.16	0.17	0.17	0.15	0.10	0.17	0.11	0.13	0.15	0.15	0.14
ΔCER(%)														
Mean	3.52	8.24	9.00	9.20	5.75	3.21	13.84	-9.14	4.96	-5.50	1.07	2.52	-1.26	9.52
Percentiles														
2.5	-7.81	-1.95	-3.83	-4.48	-5.95	-6.19	-4.99	-28.64	-4.98	-19.61	-10.43	-10.47	-12.08	-7.00
25	-0.13	0.00	0.00	0.00	0.00	0.00	0.00	-9.54	-0.01	-6.19	-1.82	-0.12	-0.46	0.00
50	0.00	0.03	0.00	0.00	0.00	0.00	0.00	-0.82	0.00	-0.04	0.00	0.00	0.00	0.00
75	1.34	6.55	1.98	2.55	4.52	3.08	3.25	0.00	2.89	0.00	0.00	1.51	0.02	2.35
97.5	43.98	60.33	65.16	37.37	46.04	32.86	24.00	10.39	46.72	15.88	26.67	38.48	18.34	21.78

Appendix C

This Appendix for Chapter 4 is structured as follows. In Section C.1 we present the computations to derive the optimal variational densities. In Section C.2 we provide useful definitions and the proof of Proposition 4.5. Moreover, we provide additional simulation results in Section C.3.

C.1 Variational densities

Proposition C.1. *The latent log-volatility vector* $\mathbf{h}$ *is approximated by a multivariate Gaussian distribution* $q^*(\mathbf{h}) \equiv \mathsf{N}_{n+1}(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)})$*, where the parameters are updated according to*

$$\boldsymbol{\Sigma}_{q(h)}^{new} = \left[\nabla^2_{\boldsymbol{\mu}_{q(h)},\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}^{old}, \boldsymbol{\Sigma}_{q(h)}^{old})\right]^{-1} \tag{C.1}$$

$$\boldsymbol{\mu}_{q(h)}^{new} = \boldsymbol{\mu}_{q(h)}^{old} + \boldsymbol{\Sigma}_{q(h)}^{new} \nabla_{\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}^{old}, \boldsymbol{\Sigma}_{q(h)}^{old}). \tag{C.2}$$

Define $\boldsymbol{\varepsilon}^2 = \boldsymbol{\varepsilon} \odot \boldsymbol{\varepsilon}$ *with components* $[\boldsymbol{\varepsilon}^2]_t = (y_t - \mathbf{x}_t^\intercal \boldsymbol{\Gamma}_t \boldsymbol{\beta}_t)^2$*, and* $\boldsymbol{\sigma}^2_{q(h)} = \mathsf{diag}(\boldsymbol{\Sigma}_{q(h)})$*. Then,*

$$\nabla_{\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}^{old}, \boldsymbol{\Sigma}_{q(h)}^{old}) = -\frac{\boldsymbol{\iota}_n}{2} + \frac{1}{2}\mathbb{E}_q(\boldsymbol{\varepsilon}^2) \odot \mathrm{e}^{-\boldsymbol{\mu}_{q(h)}^{old} + \boldsymbol{\sigma}^{2\,old}_{q(h)}/2} - \mu_{q(1/\nu^2)} \mathbf{Q} \boldsymbol{\mu}_{q(h)}^{old}, \tag{C.3}$$

and

$$\nabla^2_{\boldsymbol{\mu}_{q(h)},\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}^{old}, \boldsymbol{\Sigma}_{q(h)}^{old}) = -\frac{1}{2}\mathsf{Diag}(\mathbb{E}_q(\boldsymbol{\varepsilon}^2) \odot \mathrm{e}^{-\boldsymbol{\mu}_{q(h)}^{old} + \boldsymbol{\sigma}^{2\,old}_{q(h)}/2}) - \mu_{q(1/\nu^2)} \mathbf{Q}, \tag{C.4}$$

denote the first and second derivative of $S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)})$ *with respect to* $\boldsymbol{\mu}_{q(h)}$ *and evaluated at* $(\boldsymbol{\mu}_{q(h)}^{old}, \boldsymbol{\Sigma}_{q(h)}^{old})$*.*

Proof. The updating scheme follows the algorithm provided in Rohde and Wand (2016) for Gaussian variational approximations. The function S is called *non-entropy function* and it is given by $S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}) = \mathbb{E}_q(\log p(\mathbf{y}, \boldsymbol{\vartheta}))$:

$$\begin{aligned} S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}) = -\frac{\boldsymbol{\iota}_n^\intercal}{2}\boldsymbol{\mu}_{q(h)} - \frac{1}{2}\mathbb{E}_q^\intercal(\boldsymbol{\varepsilon}^2)\mathrm{e}^{-\boldsymbol{\mu}_{q(h)}+\boldsymbol{\sigma}^2_{q(h)}/2} \\ - \frac{1}{2}\mu_{q(1/\nu^2)}\left(\boldsymbol{\mu}_{q(h)}^\intercal \mathbf{Q}\boldsymbol{\mu}_{q(h)} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(h)}\mathbf{Q}\right\}\right), \end{aligned} \tag{C.5}$$

where $\boldsymbol{\varepsilon}^2 = \boldsymbol{\varepsilon} \odot \boldsymbol{\varepsilon}$ with components $[\boldsymbol{\varepsilon}^2]_t = (y_t - \mathbf{x}_t^\intercal \boldsymbol{\Gamma}_t \boldsymbol{\beta}_t)^2$, and $\boldsymbol{\sigma}^2_{q(h)} = \mathsf{diag}(\boldsymbol{\Sigma}_{q(h)})$. Then, the first derivative with respect to the variational mean vector $\boldsymbol{\mu}_{q(h)}$ is given by

$$\nabla_{\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}) = -\frac{\boldsymbol{\iota}_n}{2} + \frac{1}{2}\mathbb{E}_q(\boldsymbol{\varepsilon}^2) \odot \mathrm{e}^{-\boldsymbol{\mu}_{q(h)}+\boldsymbol{\sigma}^2_{q(h)}/2} - \mu_{q(1/\nu^2)}\mathbf{Q}\boldsymbol{\mu}_{q(h)}. \tag{C.6}$$

Moreover, derive $\nabla_{\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)})$ again with respect to $\boldsymbol{\mu}_{q(h)}$:

$$\nabla^2_{\boldsymbol{\mu}_{q(h)},\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}) = -\frac{1}{2}\mathsf{Diag}(\mathbb{E}_q(\boldsymbol{\varepsilon}^2) \odot \mathrm{e}^{-\boldsymbol{\mu}_{q(h)}+\boldsymbol{\sigma}^2_{q(h)}/2}) - \mu_{q(1/\nu^2)}\mathbf{Q}. \tag{C.7}$$

□

Remark C.1. Under the multivariate Gaussian approximation of $q(\mathbf{h})$ with mean vector $\boldsymbol{\mu}_{q(h)}$ and covariance matrix $\boldsymbol{\Sigma}_{q(h)}$, the optimal density of the vector $\boldsymbol{\sigma}^2 = \exp\{\mathbf{h}\}$, namely $q^*(\boldsymbol{\sigma}^2)$, is a multivariate log-normal distribution such that:

$$\begin{aligned} \mathbb{E}_q[\sigma_t^2] &= \exp\{\mu_{q(h_t)} + 1/2\sigma^2_{q(h_t)}\}, \\ \mathsf{Var}_q[\sigma_t^2] &= \exp\{2\mu_{q(h_t)} + \sigma^2_{q(h_t)}\}(\exp\{\sigma^2_{q(h_t)}\} - 1), \\ \mathsf{Cov}_q[\sigma_t^2, \sigma_{t+1}^2] &= \exp\{\mu_{q(h_t)} + \mu_{q(h_{t+1})} + 1/2(\sigma^2_{q(h_t)} + \sigma^2_{q(h_{t+1})})\}(\exp\{\mathsf{Cov}_q[h_t, h_{t+1}]\} - 1). \end{aligned} \tag{C.8}$$

Proposition C.2. *The optimal variational density for the homoskedastic variance σ^2 is an inverse-gamma $q^*(\sigma^2) \equiv \mathsf{IGa}(A_{q(\sigma^2)}, B_{q(\sigma^2)})$ where:*

$$A_{q(\sigma^2)} = A_\sigma + \frac{n}{2}, B_{q(\sigma^2)} = B_\sigma + \frac{1}{2}\mathbb{E}_q\left[\boldsymbol{\varepsilon}^\intercal\boldsymbol{\varepsilon}\right], \tag{C.9}$$

where:

$$\begin{aligned} \mathbb{E}_{-\sigma^2}\left[\boldsymbol{\varepsilon}^\intercal\boldsymbol{\varepsilon}\right] = \mathbf{y}^\intercal\mathbf{y} - 2\left(\sum_{j=1}^p \mathbf{X}_j \boldsymbol{\mu}_{q(\Gamma_j)}\boldsymbol{\mu}_{q(\beta_j)}\right)^\intercal \mathbf{y} + \sum_{j=1}^p \mathsf{tr}\left\{\left(\boldsymbol{\mu}_{q(\beta_j)}\boldsymbol{\mu}_{q(\beta_j)}^\intercal + \boldsymbol{\Sigma}_{q(\beta_j)}\right)\boldsymbol{\mu}_{q(\Gamma_j)}\mathbf{X}_j^2\right\} \\ + \sum_{j=1}^p \boldsymbol{\mu}_{q(\beta_j)}^\intercal \boldsymbol{\mu}_{q(\Gamma_j)}\mathbf{X}_j \sum_{k=1,k\neq j}^p \mathbf{X}_k \boldsymbol{\mu}_{q(\Gamma_k)}\boldsymbol{\mu}_{q(\beta_k)}. \end{aligned}$$

Proof. The full conditional distribution of σ^2 given the rest $p(\sigma^2|\mathbf{y}, \boldsymbol{\vartheta}_{-\sigma^2})$ is proportional to:

$$\begin{aligned} p(\sigma^2|\text{rest}) \propto -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(\mathbf{y} - \sum_{j=1}^{p}\mathbf{X}_j\boldsymbol{\Gamma}_j\boldsymbol{\beta}_j\right)^{\intercal}\left(\mathbf{y} - \sum_{j=1}^{p}\mathbf{X}_j\boldsymbol{\Gamma}_j\boldsymbol{\beta}_j\right) \\ - (A_\sigma + 1)\log\sigma^2 - \frac{B_\sigma}{\sigma^2}, \end{aligned}$$

where $\mathbf{X}_k$, and $\boldsymbol{\Gamma}_k$ are diagonal matrices with elements $x_{k,t}$ and $\gamma_{k,t}$ respectively. Define $\boldsymbol{\varepsilon} = \mathbf{y} - \sum_{j=1}^{p}\mathbf{X}_j\boldsymbol{\Gamma}_j\boldsymbol{\beta}_j$, the optimal variational density is given by:

$$\begin{aligned} \log q(\sigma^2) &\propto \mathbb{E}_{-\sigma^2}[\log p(\sigma^2|\text{rest})] \\ &\propto -(A_\sigma + \frac{n}{2} + 1)\log\sigma^2 - \frac{1}{\sigma^2}\left\{B_\sigma + \frac{1}{2}\mathbb{E}_{-\sigma^2}\left[\boldsymbol{\varepsilon}^{\intercal}\boldsymbol{\varepsilon}\right]\right\}, \end{aligned} \tag{C.10}$$

where:

$$\begin{aligned} \mathbb{E}_{-\sigma^2}\left[\boldsymbol{\varepsilon}^{\intercal}\boldsymbol{\varepsilon}\right] &= \mathbb{E}_{-\sigma^2}\left[\left(\mathbf{y} - \sum_{j=1}^{p}\mathbf{X}_j\boldsymbol{\Gamma}_j\boldsymbol{\beta}_j\right)^{\intercal}\left(\mathbf{y} - \sum_{j=1}^{p}\mathbf{X}_j\boldsymbol{\Gamma}_j\boldsymbol{\beta}_j\right)\right] \\ &= \mathbf{y}^{\intercal}\mathbf{y} - 2\left(\sum_{j=1}^{p}\mathbb{E}_{-\sigma^2}\left[\mathbf{X}_j\boldsymbol{\Gamma}_j\boldsymbol{\beta}_j\right]\right)^{\intercal}\mathbf{y} \\ &\quad + \sum_{j=1}^{p}\mathbb{E}_{-\sigma^2}\left[\boldsymbol{\beta}_j^{\intercal}\boldsymbol{\Gamma}_j\mathbf{X}_j\mathbf{X}_j\boldsymbol{\Gamma}_j\boldsymbol{\beta}_j + \boldsymbol{\beta}_j^{\intercal}\boldsymbol{\Gamma}_j\mathbf{X}_j\sum_{k=1,k\neq j}^{p}\mathbf{X}_k\boldsymbol{\Gamma}_k\boldsymbol{\beta}_k\right] \\ &= \mathbf{y}^{\intercal}\mathbf{y} - 2\left(\sum_{j=1}^{p}\mathbf{X}_j\boldsymbol{\mu}_{q(\Gamma_j)}\boldsymbol{\mu}_{q(\beta_j)}\right)^{\intercal}\mathbf{y} + \sum_{j=1}^{p}\mathsf{tr}\left\{\mathbb{E}_{\beta_j}\left[\boldsymbol{\beta}_j\boldsymbol{\beta}_j^{\intercal}\right]\boldsymbol{\mu}_{q(\Gamma_j)}\mathbf{X}_j^2\right\} \\ &\quad + \sum_{j=1}^{p}\boldsymbol{\mu}_{q(\beta_j)}^{\intercal}\boldsymbol{\mu}_{q(\Gamma_j)}\mathbf{X}_j\sum_{k=1,k\neq j}^{p}\mathbf{X}_k\boldsymbol{\mu}_{q(\Gamma_k)}\boldsymbol{\mu}_{q(\beta_k)} \\ &= \mathbf{y}^{\intercal}\mathbf{y} - 2\left(\sum_{j=1}^{p}\mathbf{X}_j\boldsymbol{\mu}_{q(\Gamma_j)}\boldsymbol{\mu}_{q(\beta_j)}\right)^{\intercal}\mathbf{y} + \sum_{j=1}^{p}\mathsf{tr}\left\{\left(\boldsymbol{\mu}_{q(\beta_j)}\boldsymbol{\mu}_{q(\beta_j)}^{\intercal} + \boldsymbol{\Sigma}_{q(\beta_j)}\right)\boldsymbol{\mu}_{q(\Gamma_j)}\mathbf{X}_j^2\right\} \\ &\quad + \sum_{j=1}^{p}\boldsymbol{\mu}_{q(\beta_j)}^{\intercal}\boldsymbol{\mu}_{q(\Gamma_j)}\mathbf{X}_j\sum_{k=1,k\neq j}^{p}\mathbf{X}_k\boldsymbol{\mu}_{q(\Gamma_k)}\boldsymbol{\mu}_{q(\beta_k)}. \end{aligned}$$

The function in C.10 is recognized to be the kernel of a Inverse-Gamma distribution as in Proposition C.2. □

Proposition C.3. *The optimal variational density for the regression parameters is a multivariate Gaussian* $q(\boldsymbol{\beta}_j) \equiv \mathsf{N}_{n+1}(\boldsymbol{\mu}_{q(\beta_j)}, \boldsymbol{\Sigma}_{q(\beta_j)})$*, where:*

$$\boldsymbol{\Sigma}_{q(\beta_j)} = (\mathbf{D}_j{}^2 + \mu_{q(1/\eta_j^2)}\mathbf{Q})^{-1}, \qquad \boldsymbol{\mu}_{q(\beta_j)} = \boldsymbol{\Sigma}_{q(\beta_j)}\mathbf{D}_j\boldsymbol{\mu}_{q(\varepsilon_{-j})}, \tag{C.11}$$

where $\mathbf{D}_j^m$ *is a diagonal matrix with elements* $[\mathbf{D}_j^m]_t = \mu_{q(1/\sigma_t^2)}\mu_{q(\gamma_{j,t})}x_{j,t}^m$ *and the vector* $\boldsymbol{\mu}_{q(\varepsilon_{-j})} = (0, \mu_{q(\varepsilon_{-j,1})}, \ldots, \mu_{q(\varepsilon_{-j,n})})$ *with* $\mu_{q(\varepsilon_{-j,t})} = y_t - \sum_{k=1,k\neq j}^p x_{k,t}\mu_{q(\gamma_{k,t})}\mu_{q(\beta_{k,t})}$.

Proof. The full conditional distribution of $\boldsymbol{\beta}_j$ given the rest $p(\boldsymbol{\beta}_j|\mathbf{y}, \boldsymbol{\vartheta}_{-\beta_j})$ is proportional to:

$$p(\boldsymbol{\beta}_j|\text{rest}) \propto -\frac{1}{2}\left(\mathbf{y} - \sum_{k=1}^p \mathbf{X}_k\boldsymbol{\Gamma}_k\boldsymbol{\beta}_k\right)^\intercal \mathbf{H}\left(\mathbf{y} - \sum_{k=1}^p \mathbf{X}_k\boldsymbol{\Gamma}_k\boldsymbol{\beta}_k\right) - \frac{1}{2}\mu_{q(1/\eta_j^2)}\boldsymbol{\beta}_j^\intercal\mathbf{Q}\boldsymbol{\beta}_j$$

where $\mathbf{H}$, $\mathbf{X}_k$, and $\boldsymbol{\Gamma}_k$ are diagonal matrices with elements $1/\sigma_t^2$, $x_{k,t}$, and $\gamma_{k,t}$ respectively. Define $\boldsymbol{\varepsilon}_{-j} = \mathbf{y} - \sum_{k=1,k\neq j}^p \mathbf{X}_k\boldsymbol{\Gamma}_k\boldsymbol{\beta}_k$, then

$$\begin{aligned}
p(\boldsymbol{\beta}_j|\text{rest}) &\propto -\frac{1}{2}\left(\boldsymbol{\varepsilon}_{-j} - \mathbf{X}_j\boldsymbol{\Gamma}_j\boldsymbol{\beta}_j\right)^\intercal \mathbf{H}\left(\boldsymbol{\varepsilon}_{-j} - \mathbf{X}_j\boldsymbol{\Gamma}_j\boldsymbol{\beta}_j\right) - \frac{1}{2}\mu_{q(1/\eta_j^2)}\boldsymbol{\beta}_j^\intercal\mathbf{Q}\boldsymbol{\beta}_j \\
&\propto -\frac{1}{2}\left(\boldsymbol{\beta}_j^\intercal\boldsymbol{\Gamma}_j\mathbf{X}_j\mathbf{H}\mathbf{X}_j\boldsymbol{\Gamma}_j\boldsymbol{\beta}_j - 2\boldsymbol{\beta}_j^\intercal\boldsymbol{\Gamma}_j\mathbf{X}_j\mathbf{H}\boldsymbol{\varepsilon}_{-j}\right) - \frac{1}{2}\mu_{q(1/\eta_j^2)}\boldsymbol{\beta}_j^\intercal\mathbf{Q}\boldsymbol{\beta}_j.
\end{aligned}$$

The optimal variational density is given by:

$$\begin{aligned}
\log q(\boldsymbol{\beta}_j) &\propto \mathbb{E}_{-\beta_j}[\log p(\boldsymbol{\beta}_j|\text{rest})] \\
&\propto -\frac{1}{2}\left(\boldsymbol{\beta}_j^\intercal\mathbf{D}_j^2\boldsymbol{\beta}_j - 2\boldsymbol{\beta}_j^\intercal\mathbf{D}_j\boldsymbol{\mu}_{q(\varepsilon_{-j})}\right) - \frac{1}{2}\mu_{q(1/\eta_j^2)}\boldsymbol{\beta}_j^\intercal\mathbf{Q}\boldsymbol{\beta}_j \\
&\propto -\frac{1}{2}\left(\boldsymbol{\beta}_j^\intercal(\mathbf{D}_j^2 + \mu_{q(1/\eta_j^2)}\mathbf{Q})\boldsymbol{\beta}_j - 2\boldsymbol{\beta}_j^\intercal\mathbf{D}_j\boldsymbol{\mu}_{q(\varepsilon_{-j})}\right),
\end{aligned} \tag{C.12}$$

where $\mathbf{D}_j^m$ is a diagonal matrix with elements

$$[\mathbf{D}_j^m]_t = \mathbb{E}_{-\beta_j}[\gamma_{j,t}x_{j,t}^m/\sigma_t^2)] = \mu_{q(1/\sigma_t^2)}\mu_{q(\gamma_{j,t})}x_{j,t}^m,$$

and $\boldsymbol{\mu}_{q(\varepsilon_{-j})} = (0, \mu_{q(\varepsilon_{-j,1})}, \ldots, \mu_{q(\varepsilon_{-j,n})})$ with

$$\mu_{q(\varepsilon_{-j,t})} = \mathbb{E}_{-\beta_j}\left[\mathbf{y} - \sum_{k=1,k\neq j}^p \mathbf{X}_k\boldsymbol{\Gamma}_k\boldsymbol{\beta}_k\right] = y_t - \sum_{k=1,k\neq j}^p x_{k,t}\mu_{q(\gamma_{k,t})}\mu_{q(\beta_{k,t})}.$$

The function in C.12 is recognized to be the kernel of a multivariate Gaussian distribution as in C.3. □

Proposition C.4. *The optimal variational density for the parameters* $\gamma_{j,t}$ *is a Bernoulli random variable* $q(\gamma_{j,t}) \equiv \mathsf{Bern}(\mathrm{e}^{\omega_{q(\gamma_{j,t})}}/(1+\mathrm{e}^{\omega_{q(\gamma_{j,t})}}))$*, where:*

$$\omega_{q(\gamma_{j,t})} = \mu_{q(\omega_{j,t})} - \frac{1}{2}\mu_{q(1/\sigma_t^2)}(x_{j,t}^2\mathbb{E}_q[\beta_{j,t}^2] - 2\mu_{q(\beta_{j,t})}x_{j,t}\mu_{q(\varepsilon_{-j,t})}), \tag{C.13}$$

where $\mathbb{E}_q[\beta_{j,t}^2] = \mu_{q(\beta_{j,t})}^2 + \sigma_{q(\beta_{j,t})}^2$.

Proof. The full conditional distribution of $\gamma_{j,t}$ given the rest $p(\gamma_{j,t}|\mathbf{y}, \boldsymbol{\vartheta}_{-\gamma_{j,t}})$ is proportional to:

$$\begin{aligned}
p(\gamma_{j,t}|\text{rest}) &\propto -\frac{1}{2\sigma_t^2}\left(y_{j,t} - \sum_{k=1}^{p} x_{k,t}\gamma_{k,t}\beta_{k,t}\right)^2 + \gamma_{j,t}\omega_{j,t} \\
&\propto -\frac{1}{2\sigma_t^2}(\gamma_{j,t}^2 x_{j,t}^2\beta_{j,t}^2 - 2\gamma_{j,t}\beta_{j,t}x_{j,t}\mu_{q(\varepsilon_{-j,t})}) + \gamma_{j,t}\omega_{j,t} \\
&\propto \gamma_{j,t}\left\{\omega_{j,t} - \frac{1}{2\sigma_t^2}(x_{j,t}^2\beta_{j,t}^2 - 2\beta_{j,t}x_{j,t}\mu_{q(\varepsilon_{-j,t})})\right\}.
\end{aligned}$$

The optimal variational density is given by:

$$\begin{aligned}
\log q(\gamma_{j,t}) &\propto \mathbb{E}_{-\gamma_{j,t}}\left[\log p(\gamma_{j,t}|\text{rest})\right] \\
&\propto \gamma_{j,t}\{\mu_{q(\omega_{j,t})} - \frac{1}{2}\mu_{q(1/\sigma^2)}(x_{j,t}^2\mathbb{E}_q[\beta_{j,t}^2] - 2\mu_{q(\beta_{j,t})}x_{j,t}\mu_{q(\varepsilon_{-j,t})})\},
\end{aligned} \tag{C.14}$$

where $\mathbb{E}_q[\beta_{j,t}^2] = \mu_{q(\beta_{j,t})}^2 + \sigma_{q(\beta_{j,t})}^2$. The function in C.14 is recognized to be the kernel of a Bernoulli distribution as in C.4. □

Proposition C.5. *A smooth estimate for the trajectory of the inclusion probabilities can be achieved assuming* $\tilde{q}(\boldsymbol{\gamma}_j) = \prod_{t=1}^n \tilde{q}(\gamma_{j,t})$ *such that* $\tilde{q}(\gamma_{j,t})$ *is* $\mathsf{Bern}(\pi_{j,t})$ *with* $\operatorname{logit}(\pi_{j,t}) = \mathbf{w}_t^\intercal\mathbf{f}$*. Therefore,* $\mathbb{E}_{\tilde{q}}(\boldsymbol{\gamma}_j) = \boldsymbol{\pi}_j$ *and* $\operatorname{logit}(\boldsymbol{\pi}_j) = \mathbf{W}\mathbf{f}$*, where* $\mathbf{W}$ *is a* $n \times k$ *b-spline basis matrix.*

(i) The optimal value of $\mathbf{f}$ *is the solution of the following optimization problem:*

$$\hat{\mathbf{f}} = \arg\max_{\mathbf{f}\in\mathbb{R}^k}\psi(\mathbf{f}), \quad \psi(\mathbf{f}) = \sum_{t=1}^{n}\left[(\omega_{q(\gamma_{j,t})} - \mathbf{w}_t^\intercal\mathbf{f})\text{expit}(\mathbf{w}_t^\intercal\mathbf{f}) + \log(1+\exp(\mathbf{w}_t^\intercal\mathbf{f}))\right]. \tag{C.15}$$

(ii) The gradient $\nabla_{\mathbf{f}}\psi(\mathbf{f})$ *of* $\psi(\mathbf{f})$ *is equal to:*

$$\nabla_{\mathbf{f}}\psi(\mathbf{f}) = \sum_{t=1}^{n}\mathbf{w}(\omega_{q(\gamma_{j,t})} - \mathbf{w}_t^\intercal\mathbf{f})\frac{\text{expit}(\mathbf{w}_t^\intercal\mathbf{f})}{1+\exp(\mathbf{w}_t^\intercal\mathbf{f})}. \tag{C.16}$$

Proof. To find the best $\tilde{q}$ that approximates q, minimize the Kullback-Leibler divergence $\mathcal{KL}\left(\tilde{q} \mid\mid q\right)$. This corresponds to maximize $\mathbb{E}_{\tilde{q}}[\log q] - \mathbb{E}_{\tilde{q}}[\log \tilde{q}]$ over the parameters of the approximating density $\tilde{q}$. In our case we obtain:

$$\begin{aligned}\hat{\mathbf{f}} = \arg\max_{\mathbf{f}\in\mathbb{R}^k} \psi(\mathbf{f}) &= \arg\max_{\mathbf{f}\in\mathbb{R}^k} \left\{\mathbb{E}_{\tilde{q}}[\log q(\boldsymbol{\gamma}_j)] - \mathbb{E}_{\tilde{q}}[\log \tilde{q}(\boldsymbol{\gamma})]\right\} \\ &= \arg\max_{\mathbf{f}\in\mathbb{R}^k} \sum_{t=1}^{n} \{\mathbb{E}_{\tilde{q}}[\log q(\gamma_{j,t})] - \mathbb{E}_{\tilde{q}}[\log \tilde{q}(\gamma_{j,t})]\}\end{aligned}$$

and define $\psi_t(\mathbf{f}) = \mathbb{E}_{\tilde{q}}[\log q(\gamma_{j,t})] - \mathbb{E}_{\tilde{q}}[\log \tilde{q}(\gamma_{j,t})]$. The first term is equal to:

$$\mathbb{E}_{\tilde{q}}[\log q(\gamma_{j,t})] = \mathbb{E}_{\tilde{q}}[\gamma_{j,t}\omega_{q(\gamma_{j,t})}] = \omega_{q(\gamma_{j,t})}\text{expit}(\mathbf{w}_t^\intercal \mathbf{f}),$$

while the second one can be written as:

$$\begin{aligned}\mathbb{E}_{\tilde{q}}[\log \tilde{q}(\gamma_{j,t})] &= \mathbb{E}_{\tilde{q}}[\gamma_{j,t}\mathbf{w}_t^\intercal \mathbf{f} - \log(1+\exp(\mathbf{w}_t^\intercal \mathbf{f}))] \\ &= \mathbf{w}_t^\intercal \mathbf{f}\text{expit}(\mathbf{w}_t^\intercal \mathbf{f}) - \log(1+\exp(\mathbf{w}_t^\intercal \mathbf{f})).\end{aligned}$$

Group together and obtain:

$$\psi_t(\mathbf{f}) = (\omega_{q(\gamma_{j,t})} - \mathbf{w}_t^\intercal \mathbf{f})\text{expit}(\mathbf{w}_t^\intercal \mathbf{f}) + \log(1+\exp(\mathbf{w}_t^\intercal \mathbf{f})).$$

which completes the proof of (i). Derive $\psi(\mathbf{f})$ with respect to $\mathbf{f}$:

$$\nabla_{\mathbf{f}}\psi(\mathbf{f}) = \frac{\partial}{\partial \mathbf{f}}\psi(\mathbf{f}) = \sum_{t=1}^{n} \frac{\partial}{\partial \mathbf{f}}\psi_t(\mathbf{f}).$$

Compute the latter and get:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{f}}\psi_t(\mathbf{f}) &= -\mathbf{w}_t\text{expit}(\mathbf{w}_t^\intercal \mathbf{f}) + \mathbf{w}_t(\omega_{q(\gamma_{j,t})} - \mathbf{w}_t^\intercal \mathbf{f})\frac{\text{expit}(\mathbf{w}_t^\intercal \mathbf{f})}{1+\exp(\mathbf{w}_t^\intercal \mathbf{f})} + \mathbf{w}_t\text{expit}(\mathbf{w}_t^\intercal \mathbf{f}) \\ &= \mathbf{w}_t(\omega_{q(\gamma_{j,t})} - \mathbf{w}_t^\intercal \mathbf{f})\frac{\text{expit}(\mathbf{w}_t^\intercal \mathbf{f})}{1+\exp(\mathbf{w}_t^\intercal \mathbf{f})},\end{aligned}$$

which proves (ii). □

Proposition C.6. *Let $q^*(\boldsymbol{\beta}_j)$ and $q^*(\gamma_{j,t})$ be the optimal variational densities presented in Propositions 4.1 and 4.2 (or its smoothed alternative). Define $\tilde{\boldsymbol{\beta}}_j = \boldsymbol{\Gamma}_j\boldsymbol{\beta}_j$, where the matrix $\boldsymbol{\Gamma}_j = \mathsf{Diag}(1, \gamma_{j,1}, \ldots, \gamma_{j,n})$ is diagonal. The optimal variational density of $\tilde{\boldsymbol{\beta}}_j$ is* *given by a mixture of multivariate Gaussian distributions:*

$$q^*(\tilde{\boldsymbol{\beta}}_j) = \sum_{\mathbf{s}\in\mathcal{S}} w_s \, \mathsf{N}_{n+1}(\mathbf{D}_s\boldsymbol{\mu}_{q(\beta_j)}, \mathbf{D}_s^{1/2}\boldsymbol{\Sigma}_{q(\beta_j)}\mathbf{D}_s^{1/2}), \tag{C.17}$$

where $\mathcal{S} = \{$*sequences of* $\{0,1\}$*of length* $n\}$ *with cardinality* $|\mathcal{S}| = 2^n$*, the diagonal matrix* $\mathbf{D}_s = \mathsf{Diag}(1, s_1, \ldots, s_n)$*, and mixing weights:*

$$w_s = \prod_{t=1}^{n} \mu_{q(\gamma_{j,t})}^{s_t}(1-\mu_{q(\gamma_{j,t})})^{1-s_t}, \tag{C.18}$$

where $\mathbf{s} = (s_1, \ldots, s_t, \ldots, s_n) \in \mathcal{S}$ *is an element in* $\mathcal{S}$*. Moreover, the mean and variance can be computed, and they are equal to:*

$$\boldsymbol{\mu}_{q(\tilde{\beta}_j)} = \boldsymbol{\mu}_{q(\Gamma_j)}\boldsymbol{\mu}_{q(\beta_j)}, \tag{C.19}$$

$$\boldsymbol{\Sigma}_{q(\tilde{\beta}_j)} = (\boldsymbol{\mu}_{q(\gamma_j)}\boldsymbol{\mu}_{q(\gamma_j)}^{\mathsf{T}} + \mathbf{W}_{\mu_{q(\gamma_j)}}) \odot \boldsymbol{\Sigma}_{q(\beta_j)} + \mathbf{W}_{\mu_{q(\gamma_j)}} \odot \boldsymbol{\mu}_{q(\beta_j)}\boldsymbol{\mu}_{q(\beta_j)}^{\mathsf{T}}, \tag{C.20}$$

where $\mathbf{W}_{\mu_{q(\gamma_j)}}$ *is a diagonal matrix with elements* $(1, \{\mu_{q(\gamma_{j,t})}(1-\mu_{q(\gamma_{j,t})})\}_{t=1}^{n})$*.*

Proof. Recall that under the variational Bayes assumptions on q, we have that

$$q(\boldsymbol{\beta}_j, \gamma_{j,1}, \ldots, \gamma_{j,n}) = q(\boldsymbol{\beta}_j)\prod_{t=1}^{n} q(\gamma_{j,t}). \tag{C.21}$$

For the sake of simplicity, in what follows we drop the index j and define $\boldsymbol{\gamma} = \mathsf{diag}(\boldsymbol{\Gamma})$ the diagonal elements in $\boldsymbol{\Gamma}$. Consider the following transformation of random variables $(\boldsymbol{\gamma} = \boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}} = \boldsymbol{\Gamma}\boldsymbol{\beta})$, so that $\boldsymbol{\beta} = \boldsymbol{\Gamma}^{-1}\tilde{\boldsymbol{\beta}}$. Hence it follows that:

$$\mathbf{J} = \begin{bmatrix} \nabla_{\gamma}(\gamma_1, \ldots, \gamma_n)^{\mathsf{T}} & \nabla_{\boldsymbol{\beta}}(\gamma_1, \ldots, \gamma_n)^{\mathsf{T}} \\ \nabla_{\gamma}\boldsymbol{\Gamma}^{-1}\tilde{\boldsymbol{\beta}} & \nabla_{\tilde{\boldsymbol{\beta}}}\boldsymbol{\Gamma}^{-1}\tilde{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \nabla_{\gamma}\boldsymbol{\Gamma}^{-1}\tilde{\boldsymbol{\beta}} & \boldsymbol{\Gamma}^{-1} \end{bmatrix}, \tag{C.22}$$

and so $|\mathbf{J}| = |\boldsymbol{\Gamma}^{-1}|$. The joint distribution of $(\tilde{\boldsymbol{\beta}}, \gamma_1, \ldots, \gamma_n)$ can be written as:

$$q(\tilde{\boldsymbol{\beta}}, \gamma_1, \ldots, \gamma_n) = |\boldsymbol{\Gamma}^{-1}|q(\boldsymbol{\Gamma}^{-1}\tilde{\boldsymbol{\beta}})\prod_{t=1}^{n} q(\gamma_{j,t}) = f(\tilde{\boldsymbol{\beta}}|\gamma_1, \ldots, \gamma_n)f(\gamma_1, \ldots, \gamma_n), \tag{C.23}$$

where q are then replaced by the optimal elements q^*. For the conditional distribution in (C.23), we have that:

$$f(\tilde{\boldsymbol{\beta}}|\boldsymbol{\gamma}) = |\boldsymbol{\Gamma}^{-1}|\phi_{n+1}(\boldsymbol{\Gamma}^{-1}\tilde{\boldsymbol{\beta}}|\boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)}), \tag{C.24}$$

where $\phi_{n+1}(\cdot|\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the density function of a multivariate Gaussian. After some computations we have that $f(\tilde{\boldsymbol{\beta}}|\boldsymbol{\gamma}) = \phi_{n+1}(\tilde{\boldsymbol{\beta}}|\boldsymbol{\mu}(\boldsymbol{\gamma}),\boldsymbol{\Sigma}(\boldsymbol{\gamma}))$ with mean vector $\boldsymbol{\mu}(\boldsymbol{\gamma}) = \boldsymbol{\Gamma}\boldsymbol{\mu}_{q(\beta)}$ and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\gamma}) = \boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}_{q(\beta)}\boldsymbol{\Gamma}^{1/2}$. The marginal for $\tilde{\boldsymbol{\beta}}$ can be find:

$$q(\tilde{\boldsymbol{\beta}}) = \sum_{\mathbf{s}\in\mathcal{S}} \phi_{n+1}(\tilde{\boldsymbol{\beta}}|\boldsymbol{\mu}(\boldsymbol{\gamma}=\mathbf{s}),\boldsymbol{\Sigma}(\boldsymbol{\gamma}=\mathbf{s}))\prod_{t=1}^{n} q(\gamma_t = s_t), \tag{C.25}$$

where $\mathcal{S}$ denotes the domain of $\boldsymbol{\gamma} = (1,\gamma_1,\dots,\gamma_n)$, and it is composed by all the possible sequences of $\{0,1\}$ of length n, since the first element is fixed to be 1. The latter set has cardinality $|\mathcal{S}| = 2^n$. The distributional result concerning $\tilde{\boldsymbol{\beta}}$ is therefore proven.

To compute the marginal mean recall that $\mathbb{E}_x(x) = \mathbb{E}_y(\mathbb{E}_x(x|y))$. Hence $\mathbb{E}_q(\tilde{\boldsymbol{\beta}}) = \mathbb{E}_\gamma(\boldsymbol{\Gamma}\boldsymbol{\mu}_{q(\beta)}) = \boldsymbol{\mu}_{q(\Gamma)}\boldsymbol{\mu}_{q(\beta)}$. The marginal variance-covariance matrix is then computed as $\mathsf{Var}_q(\tilde{\boldsymbol{\beta}}) = \mathbb{E}(\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\intercal) - \mathbb{E}(\tilde{\boldsymbol{\beta}})\mathbb{E}(\tilde{\boldsymbol{\beta}})^\intercal$ where

$$\begin{aligned}\mathbb{E}(\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\intercal) &= \mathbb{E}(\boldsymbol{\Gamma}\boldsymbol{\beta}(\boldsymbol{\Gamma}\boldsymbol{\beta})^\intercal) = \mathbb{E}(\boldsymbol{\Gamma}\boldsymbol{\beta}\boldsymbol{\beta}^\intercal\boldsymbol{\Gamma}) = \mathbb{E}(\boldsymbol{\gamma}\boldsymbol{\gamma}^\intercal \odot \boldsymbol{\beta}\boldsymbol{\beta}^\intercal) = \mathbb{E}(\boldsymbol{\gamma}\boldsymbol{\gamma}^\intercal)\odot\mathbb{E}(\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)\\ &= (\boldsymbol{\mu}_{q(\gamma)}\boldsymbol{\mu}_{q(\gamma)}^\intercal + \mathbf{W}_{\mu_{q(\gamma)}})\odot(\boldsymbol{\mu}_{q(\beta)}\boldsymbol{\mu}_{q(\beta)}^\intercal + \boldsymbol{\Sigma}_{q(\beta)}),\end{aligned} \tag{C.26}$$

where $\mathbf{W}_{\mu_{q(\gamma)}}$ is a diagonal matrix with elements $(1,\{\mu_{q(\gamma_t)}(1-\mu_{q(\gamma_t)})\}_{t=1}^n)$. Plug-in the latter in the formula for $\mathsf{Var}_q(\tilde{\boldsymbol{\beta}})$ and recall the analytical form of the mean $\mathbb{E}(\tilde{\boldsymbol{\beta}})$. After some simplification we end up with $\boldsymbol{\Sigma}_{q(\tilde{\beta})} = (\boldsymbol{\mu}_{q(\gamma)}\boldsymbol{\mu}_{q(\gamma)}^\intercal + \mathbf{W}_{\mu_{q(\gamma)}})\odot\boldsymbol{\Sigma}_{q(\beta)} + \mathbf{W}_{\mu_{q(\gamma)}}\odot\boldsymbol{\mu}_{q(\beta)}\boldsymbol{\mu}_{q(\beta)}^\intercal$, which concludes the proof. □

Proposition C.7. *The optimal variational density for the parameter $\boldsymbol{\omega}_j$ is a multivariate Gaussian $q(\boldsymbol{\omega}_j) \equiv \mathsf{N}_{n+1}(\boldsymbol{\mu}_{q(\omega_j)},\boldsymbol{\Sigma}_{q(\omega_j)})$, where:*

$$\boldsymbol{\Sigma}_{q(\omega_j)} = (\mathsf{Diag}(0,\boldsymbol{\mu}_{q(z_j)}) + \mu_{q(1/\xi_j^2)}\mathbf{Q})^{-1}, \qquad \boldsymbol{\mu}_{q(\omega_j)} = \boldsymbol{\Sigma}_{q(\omega_j)}(0,\boldsymbol{\mu}_{q(\bar{\gamma}_j)}^\intercal)^\intercal, \tag{C.27}$$

where $\boldsymbol{\mu}_{q(\bar{\gamma}_j)} = \boldsymbol{\mu}_{q(\gamma_j)} - 1/2\boldsymbol{\iota}_n$.

Proof. The full conditional distribution of $\boldsymbol{\omega}_j$ given the rest $p(\boldsymbol{\omega}_j|\mathbf{y},\boldsymbol{\vartheta}_{-\omega_j})$ is proportional to:

$$p(\boldsymbol{\omega}_j|\text{rest}) \propto \boldsymbol{\omega}_j^\intercal(\boldsymbol{\gamma}_j - 1/2\boldsymbol{\iota}_n) - \frac{1}{2}\boldsymbol{\omega}_j^\intercal\mathsf{Diag}(\mathbf{z}_j)\boldsymbol{\omega}_j - \frac{1}{2\xi_j^2}\boldsymbol{\omega}_j^\intercal\mathbf{Q}\boldsymbol{\omega}_j.$$

The optimal variational density is given by:

$$\begin{aligned}\log q(\boldsymbol{\omega}_j) &\propto \mathbb{E}_{-\omega_j}[\log p(\boldsymbol{\omega}_j|\text{rest})] \\ &\propto \boldsymbol{\omega}_j^\intercal \boldsymbol{\mu}_{q(\bar{\gamma}_j)} - \frac{1}{2}\boldsymbol{\omega}_j^\intercal \mathsf{Diag}(\boldsymbol{\mu}_{q(z_j)})\boldsymbol{\omega}_j - \frac{1}{2}\mu_{q(1/\xi_j^2)}\boldsymbol{\omega}_j^\intercal \mathbf{Q}\boldsymbol{\omega}_j \\ &\propto -\frac{1}{2}\left(\boldsymbol{\omega}_j^\intercal(\mathsf{Diag}(0,\boldsymbol{\mu}_{q(z_j)}) + \mu_{q(1/\xi_j^2)}\mathbf{Q})\boldsymbol{\omega}_j - 2\boldsymbol{\omega}_j^\intercal(0,\boldsymbol{\mu}^\intercal_{q(\bar{\gamma}_j)})^\intercal\right),\end{aligned} \tag{C.28}$$

where $\boldsymbol{\mu}_{q(\bar{\gamma}_j)} = \boldsymbol{\mu}_{q(\gamma_j)} - 1/2\boldsymbol{\iota}_n$. The function in C.28 is recognized to be the kernel of a multivariate Gaussian distribution as in C.7. □

Proposition C.8. *The optimal variational density for the $z_{j,t}$ parameters is a Polya-Gamma $q(z_{j,t}) \equiv \mathsf{PG}(1, \sqrt{\mu_{q(\omega_{j,t}^2)}})$ and define*

$$\mu_{q(z_{j,t})} = \mathbb{E}_q\left[z_{j,t}\right] = \frac{1}{2}\frac{1}{\sqrt{\mu_{q(\omega_{j,t}^2)}}}\tanh\left(\frac{\sqrt{\mu_{q(\omega_{j,t}^2)}}}{2}\right) \tag{C.29}$$

Proof.

$$\log q(z_{j,t}) \propto -z_t\mu_{q(\omega_{j,t}^2)} + \log p(z_{j,t}), \tag{C.30}$$

where $p(z_{j,t})$ is the density function of a $\mathsf{PG}(1,0)$. □

Proposition C.9. *The optimal variational density for the variance parameter η_j^2 is an inverse-gamma distribution $q(\eta_j^2) \equiv \mathsf{IGa}(A_{q(\eta_j^2)}, B_{q(\eta_j^2)})$, where:*

$$A_{q(\eta_j^2)} = A_\eta + \frac{n+1}{2}, \qquad B_{q(\eta_j^2)} = B_\eta + \frac{1}{2}\left(\boldsymbol{\mu}^\intercal_{q(\beta_j)}\mathbf{Q}\boldsymbol{\mu}_{q(\beta_j)} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\beta_j)}\mathbf{Q}\right\}\right). \tag{C.31}$$

Proof. The full conditional distribution of η_j^2 given the rest $p(\eta_j^2|\mathbf{y}, \boldsymbol{\vartheta}_{-\eta_j^2})$ is proportional to:

$$p(\eta_j^2|\text{rest}) \propto -\frac{n+1}{2}\log\eta_j^2 - \frac{1}{2\eta_j^2}\boldsymbol{\beta}_j^\intercal\mathbf{Q}\boldsymbol{\beta}_j - (A_\eta+1)\log\eta_j^2 - \frac{B_\eta}{\eta_j^2}.$$

The optimal variational density is given by:

$$\begin{aligned}\log q(\eta_j^2) &\propto \mathbb{E}_{-\eta_j^2}[\log p(\eta_j^2|\text{rest})] \\ &\propto -\frac{n+1}{2}\log\eta_j^2 - \frac{1}{2\eta_j^2}\mathbb{E}_{-\eta_j^2}\left[\boldsymbol{\beta}_j^\intercal\mathbf{Q}\boldsymbol{\beta}_j\right] - (A_\eta+1)\log\eta_j^2 - \frac{B_\eta}{\eta_j^2} \\ &\propto -(\frac{n}{2} + A_\eta + 1)\log\eta_j^2 - \frac{1}{\eta_j^2}\left(B_\eta + \frac{1}{2}\left(\boldsymbol{\mu}^\intercal_{q(\beta_j)}\mathbf{Q}\boldsymbol{\mu}_{q(\beta_j)} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\beta_j)}\mathbf{Q}\right\}\right)\right).\end{aligned} \tag{C.32}$$

The function in C.32 is recognized to be the kernel of an Inverse-Gaussian distribution as in C.9. □

Proposition C.10. *The optimal variational density for the variance parameter ξ_j^2 is an inverse-gamma distribution $q(\xi_j^2) \equiv \mathsf{IGa}(A_{q(\xi_j^2)}, B_{q(\xi_j^2)})$, where:*

$$A_{q(\xi_j^2)} = A_\xi + \frac{n+1}{2}, \qquad B_{q(\xi_j^2)} = B_\xi + \frac{1}{2}\left(\boldsymbol{\mu}_{q(\omega_j)}^\intercal \mathbf{Q}\boldsymbol{\mu}_{q(\omega_j)} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\omega_j)}\mathbf{Q}\right\}\right). \tag{C.33}$$

Proof. The full conditional distribution of ξ_j^2 given the rest $p(\xi_j^2|\mathbf{y}, \boldsymbol{\vartheta}_{-\xi_j^2})$ is proportional to:

$$p(\xi_j^2|\text{rest}) \propto -\frac{n+1}{2}\log\xi_j^2 - \frac{1}{2\xi_j^2}\boldsymbol{\beta}_j^\intercal\mathbf{Q}\boldsymbol{\beta}_j - (A_\xi+1)\log\xi_j^2 - \frac{B_\xi}{\xi_j^2}.$$

The optimal variational density is given by:

$$\begin{aligned}\log q(\xi_j^2) &\propto \mathbb{E}_{-\xi_j^2}[\log p(\xi_j^2|\text{rest})]\\ &\propto -\frac{n+1}{2}\log\xi_j^2 - \frac{1}{2\xi_j^2}\mathbb{E}_{-\xi_j^2}\left[\boldsymbol{\omega}_j^\intercal\mathbf{Q}\boldsymbol{\omega}_j\right] - (A_\xi+1)\log\xi_j^2 - \frac{B_\xi}{\xi_j^2}\\ &\propto -(\frac{n}{2}+A_\xi+1)\log\xi_j^2 - \frac{1}{\xi_j^2}\left(B_\xi + \frac{1}{2}\left(\boldsymbol{\mu}_{q(\omega_j)}^\intercal\mathbf{Q}\boldsymbol{\mu}_{q(\omega_j)} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\omega_j)}\mathbf{Q}\right\}\right)\right).\end{aligned} \tag{C.34}$$

The function in C.34 is recognized to be the kernel of an Inverse-Gaussian distribution as in C.10. □

Proposition C.11. *The optimal variational density for the variance parameter ν^2 is an inverse-gamma distribution $q(\nu^2) \equiv \mathsf{IGa}(A_{q(\nu^2)}, B_{q(\nu^2)})$, where:*

$$A_{q(\nu^2)} = A_\nu + \frac{n+1}{2}, \qquad B_{q(\nu^2)} = B_\nu + \frac{1}{2}\left(\boldsymbol{\mu}_{q(h)}^\intercal\mathbf{Q}\boldsymbol{\mu}_{q(h)} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(h)}\mathbf{Q}\right\}\right). \tag{C.35}$$

Proof. The full conditional distribution of ν^2 given the rest $p(\nu^2|\mathbf{y}, \boldsymbol{\vartheta}_{-\nu^2})$ is proportional to:

$$p(\nu^2|\text{rest}) \propto -\frac{n+1}{2}\log\nu^2 - \frac{1}{2\nu^2}\mathbf{h}^\intercal\mathbf{Q}\mathbf{h} - (A_\nu+1)\log\nu^2 - \frac{B_\nu}{\nu^2}.$$

The optimal variational density is given by:

$$\begin{aligned}\log q(\nu^2) &\propto \mathbb{E}_{-\nu^2}[\log p(\nu^2|\text{rest})]\\ &\propto -\frac{n+1}{2}\log\nu^2 - \frac{1}{2\nu^2}\mathbb{E}_{-\nu^2}\left[\mathbf{h}^\intercal\mathbf{Q}\mathbf{h}\right] - (A_\nu+1)\log\nu^2 - \frac{B_\nu}{\nu^2}\\ &\propto -(\frac{n}{2}+A_\nu+1)\log\nu^2 - \frac{1}{\nu^2}\left(B_\nu + \frac{1}{2}\left(\boldsymbol{\mu}_{q(h)}^\intercal\mathbf{Q}\boldsymbol{\mu}_{q(h)} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(h)}\mathbf{Q}\right\}\right)\right).\end{aligned} \tag{C.36}$$

The function in C.36 is recognized to be the kernel of an Inverse-Gaussian distribution as in Proposition C.11. □

C.2 Proof of Proposition 4.5

Definition C.1. $\mathbf{A}$ is a Z-matrix if its off-diagonal elements satisfy $a_{i,j} \leq 0$, for $i \neq j$.

Definition C.2. $\mathbf{A}$ is a strictly diagonally dominant (SDD) matrix if, for each i row of $\mathbf{A}$, $|a_{i,i}| > \sum_{j \neq i} |a_{i,j}|$.

Corollary C.1. If a matrix $\mathbf{A}$ is SDD and all its diagonal elements $a_{i,i}$ are positive, then the real parts of its eigenvalues are positive.

Definition C.3. A matrix $\mathbf{A}$ is said to be an M-matrix if it is a strictly diagonally dominant Z-matrix and all its diagonal elements $a_{i,i}$ are positive.

Corollary C.2. If a matrix $\mathbf{A}$ is an M-matrix, then it belongs to inverse-positive matrices, i.e all elements of the inverse are positive $[\mathbf{A}^{-1}]_{i,j} \geq 0$, for all (i,j).

Lemma C.1. The matrix $\mathbf{Q}^{-1}$ is a positive matrix, i.e $[\mathbf{Q}^{-1}]_{i,j} \geq 0$.

Proof. Follows from the tridiagonal form of $\mathbf{Q}$ with $q_{1,1} = 1 + 1/k_0$, and $k_0 < +\infty$. □

Lemma C.2. The matrix $\mathbf{\Sigma}_{q(\omega_j)}$ is a positive matrix, i.e $[\mathbf{\Sigma}_{q(\omega_j)}]_{i,j} \geq 0$.

Proof. Recall that

$$\mathbf{\Sigma}_{q(\omega_j)} = \mathbf{W}^{-1} = \Big(\mathsf{Diag}\left(0, \mathbb{E}_q\left[\mathbf{z}_j\right]\right) + \mu_{q(1/\xi_j^2)}\mathbf{Q}\Big)^{-1}, \tag{C.1}$$

is tridiagonal, where $\mathbb{E}_q\left[z_{j,t}\right] > 0$ and $\mu_{q(1/\xi_j^2)} > 0$. Notice that $\mathbf{W}$ has off-diagonal elements equal to $-\mu_{q(1/\xi_j^2)} < 0$ in the first sub/over-diagonal and 0 elsewhere and therefore it is a Z-matrix. Moreover, $w_{t,t} > 0$ for all t and:

$$w_{1,1} = (1 + k_0^{-1})\mu_{q(1/\xi_j^2)} > \mu_{q(1/\xi_j^2)} = |w_{1,2}| \tag{C.2}$$

$$w_{t,t} = 2\mu_{q(1/\xi_j^2)} + \mathbb{E}_q\left[z_{j,t}\right] > 2\mu_{q(1/\xi_j^2)} = |w_{t,t-1}| + |w_{t,t+1}|, \quad t = 2, \dots, n \tag{C.3}$$

$$w_{n+1,n+1} = \mu_{q(1/\xi_j^2)} + \mathbb{E}_q\left[z_{j,n}\right] > \mu_{q(1/\xi_j^2)} = |w_{n+1,n-1}|, \tag{C.4}$$

thus $\mathbf{W}$ is SDD with positive diagonal elements. Hence, by definition C.3 is an M-matrix and corollary C.2 tells us that its inverse is a positive matrix. □

Proposition C.12. *Assume that the maximum over time of the inclusion probabilities, for a given variable j, at the i-th iteration of the algorithm is such that* $\max_{t \in \{1,\dots,n\}} \mu_{q(\gamma_{j,t})}^{(i)} =$

ϵ, and $\epsilon \ll 1$ is small enough. Moreover, let $\mathbf{\Sigma}^{(i)}_{q(\omega_j)} - \mathbf{\Sigma}^{(i-1)}_{q(\omega_j)}$ be a positive matrix, i.e., $\mathbf{\Sigma}^{(i)}_{q(\omega_j)} - \mathbf{\Sigma}^{(i-1)}_{q(\omega_j)} \geq 0$, then:

1. *$\mu^{(i+1)}_{q(\gamma_{j,t})} = \operatorname{expit}\left\{\mu^{(i+1)}_{q(\omega_{j,t})} - \frac{1}{2}\mu^{(i+1)}_{q(1/\sigma^2_t)} x^2_{j,t}\mu^{-1(i+1)}_{q(1/\eta^2_j)} q_{t,t} + O(\epsilon)\right\}$, $q_{t,t} = [\mathbf{Q}^{-1}]_{t,t} \geq 0$;*
2. *$\mu^{(i+1)}_{q(\omega_{j,t})} = -1/2\sum_{k=1}^{n} s_{t,k} + O(\epsilon)$, $s_{t,k} = [\mathbf{\Sigma}_{q(\omega_j)}]_{t,k} \geq 0$;*
3. *$\mu^{(i+1)}_{q(\omega_{j,t})} \leq \mu^{(i)}_{q(\omega_{j,t})}$ decreases after each iteration.*

Proof. We start proving *1)*. Consider the update for $\mu^{(i+1)}_{q(\gamma_{j,t})}$:

$$\mu^{(i+1)}_{q(\gamma_{j,t})} = \operatorname{expit}\left\{\mu^{(i+1)}_{q(\omega_{j,t})} - 1/2\mu^{(i+1)}_{q(1/\sigma^2_t)}\left(\mathbb{E}^{(i+1)}_q[\beta^2_{j,t}]x^2_{j,t} - 2\mu^{(i+1)}_{q(\beta_{j,t})}x_{j,t}\mathbb{E}^{(i+1)}_q[\varepsilon_{j,t}]\right)\right\}. \tag{C.5}$$

Notice that the vector for all times $\boldsymbol{\mu}^{(i+1)}_{q(\beta_j)}$ has the following formula:

$$\boldsymbol{\mu}^{(i+1)}_{q(\beta_j)} = \left(\mu^{(i+1)}_{q(1/\sigma^2_t)}(\tilde{\mathbf{D}}^{2(i)}_{\gamma_j}) + \mu^{(i)}_{q(1/\eta^2_j)}\mathbf{Q}\right)^{-1}\mu^{(i+1)}_{q(1/\sigma^2_t)}\tilde{\mathbf{D}}^{(i)}_{\gamma_j}\boldsymbol{\mu}^{(i+1)}_{q(\tilde{\varepsilon}_{-j})}, \tag{C.6}$$

where $\tilde{\mathbf{D}}_{\gamma_j} = \mathsf{Diag}((0, \boldsymbol{\mu}_{q(\gamma_j)}) \odot (0, \mathbf{x}_j))$ and $\tilde{\mathbf{D}}^2_{\gamma_j} = \mathsf{Diag}((0, \boldsymbol{\mu}_{q(\gamma_j)}) \odot (0, \mathbf{x}_j \odot \mathbf{x}_j))$. By assumption we can write each $\mu^{(i+1)}_{q(\gamma_{j,t})} = \alpha_t\epsilon$, with $0 < \alpha_t \leq 1$. Now define $\boldsymbol{\alpha}$ the collection of the α_t, and $\mathbf{A}_{\gamma_j} = \mathsf{Diag}((0, \boldsymbol{\alpha}) \odot (0, \mathbf{x}_j))$ and $\mathbf{A}^2_{\gamma_j} = \mathsf{Diag}((0, \boldsymbol{\alpha}) \odot (0, \mathbf{x}_j \odot \mathbf{x}_j))$. (C.6) can be written as

$$\boldsymbol{\mu}^{(i+1)}_{q(\beta_j)} = \left(\mu^{(i+1)}_{q(1/\sigma^2_t)}(\epsilon\mathbf{A}^{2(i)}_{\gamma_j}) + \mu^{(i)}_{q(1/\eta^2_j)}\mathbf{Q}\right)^{-1}\mu^{(i+1)}_{q(1/\sigma^2_t)}\epsilon\mathbf{A}^{(i)}_{\gamma_j}\boldsymbol{\mu}^{(i+1)}_{q(\tilde{\varepsilon}_{-j})}, \tag{C.7}$$

and

$$\lim_{\epsilon\to 0}\frac{\boldsymbol{\mu}^{(i+1)}_{q(\beta_j)}}{\epsilon} < \infty \quad\Longrightarrow\quad \mu^{(i+1)}_{q(\beta_{j,t})} = O(\epsilon). \tag{C.8}$$

Consider now the variance matrix $\mathbf{\Sigma}^{(i+1)}_{q(\beta_j)}$:

$$\mathbf{\Sigma}^{(i+1)}_{q(\beta_j)} = \left(\mu^{(i+1)}_{q(1/\sigma^2_t)}(\epsilon\mathbf{A}^{2(i)}_{\gamma_j}) + \mu^{(i)}_{q(1/\eta^2_j)}\mathbf{Q}\right)^{-1} = f(\epsilon), \tag{C.9}$$

as a scalar to matrix function f with

$$f'(\epsilon) = \left(\mu^{(i+1)}_{q(1/\sigma^2_t)}(\epsilon\mathbf{A}^{2(i)}_{\gamma_j}) + \mu^{(i)}_{q(1/\eta^2_j)}\mathbf{Q}\right)^{-1}\left(\mu^{(i+1)}_{q(1/\sigma^2_t)}(\mathbf{A}^{2(i)}_{\gamma_j})\right)\left(\mu^{(i+1)}_{q(1/\sigma^2_t)}(\epsilon\mathbf{A}^{2(i)}_{\gamma_j}) + \mu^{(i)}_{q(1/\eta^2_j)}\mathbf{Q}\right)^{-1}.$$

Using Taylor expansion in $\epsilon = 0$ we obtain:

$$\mathbf{\Sigma}^{(i+1)}_{q(\beta_j)} = \left(\mu^{(i)}_{q(1/\eta^2_j)}\mathbf{Q}\right)^{-1} + \epsilon\left(\mu^{(i)}_{q(1/\eta^2_j)}\mathbf{Q}\right)^{-1}\left(\mu^{(i+1)}_{q(1/\sigma^2_t)}(\mathbf{A}^{2(i)}_{\gamma_j})\right)\left(\mu^{(i)}_{q(1/\eta^2_j)}\mathbf{Q}\right)^{-1} + \ldots$$

and therefore each diagonal element is $\sigma^{2(i+1)}_{q(\beta_{j,t})} = \left[\mu^{(i)}_{q(1/\eta_j^2)}\right]^{-1} q_{t,t} + O(\epsilon)$ and it follows that

$$\mathbb{E}^{(i+1)}_q[\beta^2_{j,t}] = (\mu^{(i+1)}_{q(\beta_{j,t})})^2 + \sigma^{2(i+1)}_{q(\beta_{j,t})} = \left[\mu^{(i)}_{q(1/\eta_j^2)}\right]^{-1} q_{t,t} + O(\epsilon). \tag{C.10}$$

Put together (C.8) and (C.10) completes the proof. Similarly we prove *2)*. Recall the function to jointly update $\boldsymbol{\mu}^{(i+1)}_{q(\omega_j)}$:

$$\boldsymbol{\mu}^{(i+1)}_{q(\omega_j)} = \boldsymbol{\Sigma}^{(i+1)}_{q(\omega_j)} \left(0, \boldsymbol{\mu}^{(i)\intercal}_{q(\gamma_j)} - 1/2\boldsymbol{\iota}^\intercal_n\right)^\intercal, \tag{C.11}$$

then the update of the t-th component is:

$$\begin{aligned}\mu^{(i+1)}_{q(\omega_{j,t})} &= \mathbf{s}^\intercal_t \left(0, \boldsymbol{\mu}^{(i)\intercal}_{q(\gamma_j)} - 1/2\boldsymbol{\iota}^\intercal_n\right)^\intercal \\ &= -1/2\mathbf{s}^\intercal_t\,(0, \boldsymbol{\iota}^\intercal_n)^\intercal + \mathbf{s}^\intercal_t \left(0, \boldsymbol{\mu}^{(i)\intercal}_{q(\gamma_j)}\right)^\intercal \\ &= -1/2\sum_{k=1}^n s_{t,k} + \sum_{k=1}^n s_{t,k}\mu^{(i)}_{q(\gamma_{j,k})},\end{aligned} \tag{C.12}$$

where $\mathbf{s}_t$ denotes the t-th column in $\boldsymbol{\Sigma}^{(i+1)}_{q(\omega_j)}$. Notice that, since $\mu^{(i)}_{q(\gamma_{j,k})} \leq \epsilon$, for all k, we can write $\mu^{(i)}_{q(\gamma_{j,k})} = \alpha_k\epsilon$, where $0 < \alpha_k \leq 1$. Plug-in the latter in (C.12) and get

$$\mu^{(i+1)}_{q(\omega_{j,t})} = -1/2\sum_{k=1}^n s_{t,k} + \epsilon\sum_{k=1}^n \alpha_k s_{t,k} = -1/2\sum_{k=1}^n s_{t,k} + O(\epsilon). \tag{C.13}$$

To prove the last statement *3)*, assume that we observe $\boldsymbol{\Sigma}^{(i)}_{q(\omega_j)} - \boldsymbol{\Sigma}^{(i-1)}_{q(\omega_j)}$ positive matrix. Then we have that, for ϵ small:

$$|\boldsymbol{\mu}^{(i)}_{q(\omega_j)}| = \frac{1}{2}\boldsymbol{\Sigma}^{(i)}_{q(\omega_j)}\,(0, \boldsymbol{\iota}^\intercal_n)^\intercal \geq \frac{1}{2}\boldsymbol{\Sigma}^{(i-1)}_{q(\omega_j)}\,(0, \boldsymbol{\iota}^\intercal_n)^\intercal = |\boldsymbol{\mu}^{(i-1)}_{q(\omega_j)}|, \tag{C.14}$$

and therefore:

$$\mathbb{E}^{(i)}_q(\boldsymbol{\omega}_j\boldsymbol{\omega}^\intercal_j) = \boldsymbol{\mu}^{(i)}_{q(\omega_j)}(\boldsymbol{\mu}^{(i)}_{q(\omega_j)})^\intercal + \boldsymbol{\Sigma}^{(i)}_{q(\omega_j)} \geq \boldsymbol{\mu}^{(i-1)}_{q(\omega_j)}(\boldsymbol{\mu}^{(i-1)}_{q(\omega_j)})^\intercal + \boldsymbol{\Sigma}^{(i-1)}_{q(\omega_j)} = \mathbb{E}^{(i-1)}_q(\boldsymbol{\omega}_j\boldsymbol{\omega}^\intercal_j), \tag{C.15}$$

which means that $\mathbb{E}^{(i)}_q(\boldsymbol{\omega}_j\boldsymbol{\omega}^\intercal_j) - \mathbb{E}^{(i-1)}_q(\boldsymbol{\omega}_j\boldsymbol{\omega}^\intercal_j)$ is a positive matrix. Consider now the update for the variable $z_{j,t}$:

$$\mathbb{E}^{(i)}_q\left[z_{j,t}\right] = \frac{1}{2}\frac{1}{\sqrt{\mathbb{E}^{(i)}_q(\omega^2_{j,t})}}\tanh(\frac{\sqrt{\mathbb{E}^{(i)}_q(\omega^2_{j,t})}}{2}) \leq \mathbb{E}^{(i-1)}_q\left[z_{j,t}\right], \tag{C.16}$$

since it is decreasing in $\mathbb{E}_q^{(i)}(\omega_{j,t}^2)$, for all t. And similarly for ξ_j^2:

$$\mu^{(i)}_{q(1/\xi_j^2)} = \frac{A_\xi + \frac{n+1}{2}}{B_\xi + \frac{1}{2}\mathsf{tr}\left\{\mathbb{E}_q^{(i)}(\boldsymbol{\omega}_j\boldsymbol{\omega}_j^\intercal)\mathbf{Q}\right\}} \leq \mu^{(i-1)}_{q(1/\xi_j^2)}, \tag{C.17}$$

since it is decreasing in $\mathbb{E}_q^{(i)}(\boldsymbol{\omega}_j\boldsymbol{\omega}_j^\intercal)$ and $\mathbb{E}_q^{(i)}(\boldsymbol{\omega}_j\boldsymbol{\omega}_j^\intercal) - \mathbb{E}_q^{(i-1)}(\boldsymbol{\omega}_j\boldsymbol{\omega}_j^\intercal)$ is a positive matrix. The next update of $\boldsymbol{\Sigma}_{q(\omega_j)}$ is equal to:

$$\boldsymbol{\Sigma}^{(i+1)}_{q(\omega_j)} = \left(\mathsf{Diag}\left(0, \mathbb{E}_q^{(i)}\left[\mathbf{z}_j\right]\right) + \mu^{(i)}_{q(1/\xi_j^2)}\mathbf{Q}\right)^{-1}, \tag{C.18}$$

which increases as both $\mathbb{E}_q^{(i)}\left[\mathbf{z}_j\right]$ and $\mu^{(i)}_{q(1/\xi_j^2)}$ decreases. Hence also $\boldsymbol{\Sigma}^{(i+1)}_{q(\omega_j)} - \boldsymbol{\Sigma}^{(i)}_{q(\omega_j)}$ is a positive matrix and therefore, for ϵ small:

$$|\boldsymbol{\mu}^{(i+1)}_{q(\omega_j)}| \geq |\boldsymbol{\mu}^{(i)}_{q(\omega_j)}|, \tag{C.19}$$

and from statement *2)* we have that $\boldsymbol{\mu}^{(i+1)}_{q(\omega_j)} \leq \boldsymbol{\mu}^{(i)}_{q(\omega_j)}$. Set $i = i + 1$ and repeat the procedure from (C.14). Observe that $\boldsymbol{\mu}_{q(\omega_j)}$ decreases after each iteration until convergence. □

C.3 Additional simulation results

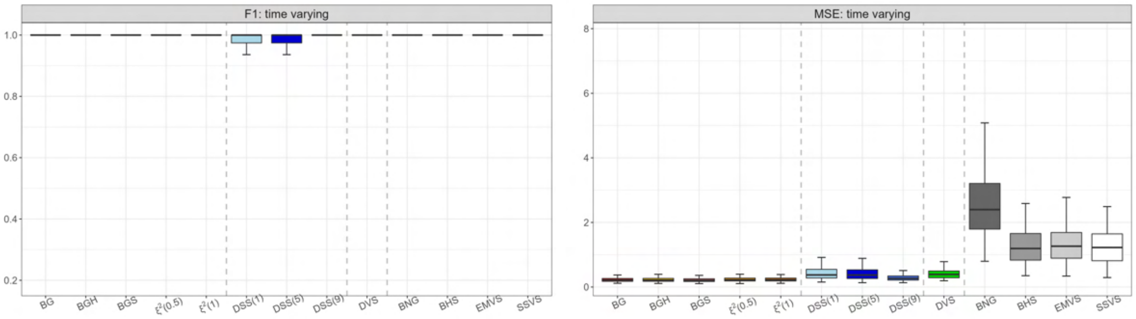


FIGURE C.1: F_1-score (left panel) and MSE (right panel) for the time-varying intercept estimation, when $p = 100$.

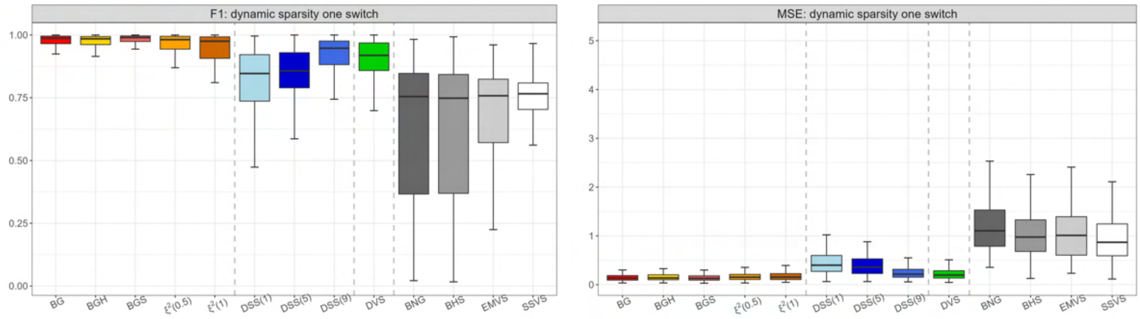


FIGURE C.2: F_1-score (left panel) and MSE (right panel) for the balanced dynamic sparsity setting with one switch, when $p = 100$.

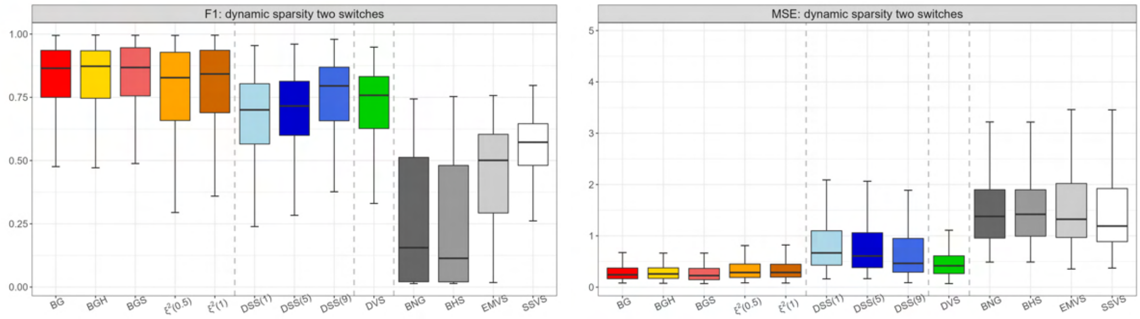


FIGURE C.3: F_1-score (left panel) and MSE (right panel) for the balanced dynamic sparsity setting with two switches, when $p = 100$.

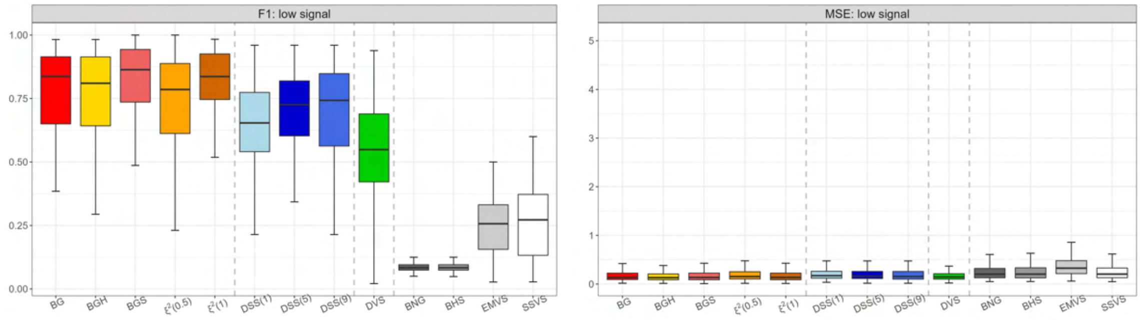


FIGURE C.4: F_1-score (left panel) and MSE (right panel) for the dynamic sparsity setting with low signal, when $p = 100$.

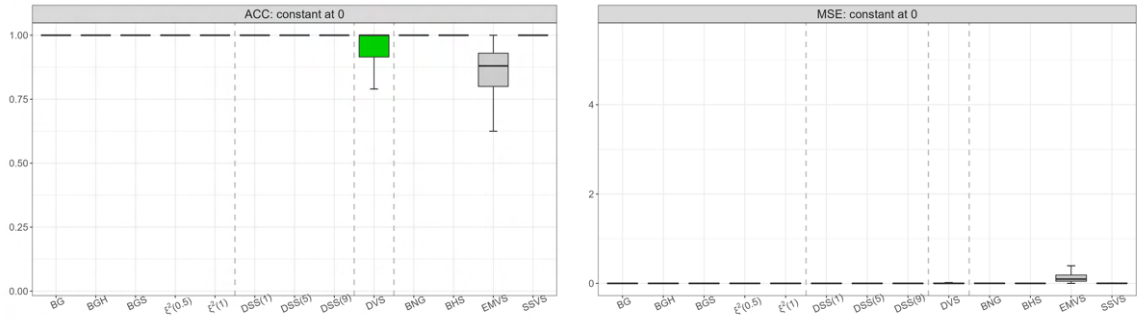


FIGURE C.5: F_1-score (left panel) and MSE (right panel) for the always zero coefficients, when $p = 100$.

Bibliography

Aguayo-Orozco, A., Bois, F. Y., Brunak, S. and Taboureau, O. (2018) Analysis of time-series gene expression data to explore mechanisms of chemical-induced hepatic steatosis toxicity. *Frontiers in Genetics* **9**.

Andersen, T. G. and Sørensen, B. E. (1996) GMM estimation of a stochastic volatility model: A Monte Carlo study. *Journal of Business & Economic Statistics* **14**(3), 328–352.

Andrieu, C., Doucet, A. and Holenstein, R. (2010) Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(3), 269–342.

Avramov, D. (2004) Stock return predictability and asset pricing models. *Review of Financial Studies* **17**(3), 699–738.

Bar-Joseph, Z., Gitter, A. and Simon, I. (2012) Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics* **13**(8), 552–564.

Barroso, P. and Detzel, A. (2021) Do limits to arbitrage explain the benefits of volatility-managed portfolios? *Journal of Financial Economics* **140**(3), 744–767.

Barroso, P. and Santa-Clara, P. (2015) Momentum has its moments. *Journal of Financial Economics* **116**(1), 111–120.

Belmonte, M. A., Koop, G. and Korobilis, D. (2014) Hierarchical shrinkage in time-varying parameter models. *Journal of Forecasting* **33**(1), 80–94.

Bernardi, M., Bonaccolto, G., Caporin, M. and Costola, M. (2020) Volatility forecasting in a data rich environment. *Macroeconomic Forecasting in the Era of Big Data* pp. 127–160.

Bhattacharya, A., Chakraborty, A. and Mallick, B. K. (2016) Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika* **103**(4), 985–991.

Bhattacharya, A., Pati, D., Pillai, N. S. and Dunson, D. B. (2015) Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association* **110**(512), 1479–1490.

Bianchi, D., De Polis, A. and Petrella, I. (2022) Taming momentum crashes. *Working paper* .

Bianchi, D. and McAlinn, K. (2020) Divide and conquer: Financial ratios and industry returns predictability. *Available at SSRN* .

Bitto, A. and Frühwirth-Schnatter, S. (2019) Achieving shrinkage in a time-varying parameter model framework. *J. Econometrics* **210**(1), 75–97.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017) Variational inference: a review for statisticians. *J. Amer. Statist. Assoc.* **112**(518), 859–877.

Bognanni, M. (2022) Comment on "Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors". *Journal of Econometrics* **227**(2), 498–505.

Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **31**(3), 307–327.

Bongaerts, D., Kang, X. and van Dijk, M. (2020) Conditional volatility targeting. *Financial Analysts Journal* **76**(4), 54–71.

Campbell, J. Y. and Thompson, S. B. (2007) Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies* **21**(4), 1509–1531.

Campbell, J. Y. and Viceira, L. M. (2004) Long-horizon mean-variance analysis: A user guide. *Manuscript, Harvard University, Cambridge, MA* .

Carriero, A., Chan, J., Clark, T. E. and Marcellino, M. (2022) Corrigendum to "Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors"[j. econometrics 212 (1)(2019) 137–154]. *Journal of Econometrics* **227**(2), 506–512.

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*, volume 5, pp. 73–80.

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

Casella, G. and George, E. I. (1992) Explaining the Gibbs sampler. *The American Statistician* **46**(3), 167–174.

Cederburg, S., O'Doherty, M. S., Wang, F. and Yan, X. S. (2020) On the performance of volatility-managed portfolios. *Journal of Financial Economics* **138**(1), 95–117.

Chan, J. C. and Eisenstat, E. (2018) Bayesian model comparison for time-varying parameter vars with stochastic volatility. *Journal of Applied Econometrics* **33**(4), 509–532.

Chan, J. C. and Yu, X. (2022) Fast and accurate variational inference for large Bayesian vars with stochastic volatility. *Journal of Economic Dynamics and Control* **143**, 104505.

Clark, T. E. (2011) Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility. *Journal of Business & Economic Statistics* **29**(3), 327–341.

Cogley, T. and Sargent, T. J. (2001) Evolving post-world war II US inflation dynamics. *NBER macroeconomics annual* **16**, 331–373.

Cohen, L. and Frazzini, A. (2008) Economic links and predictable returns. *The Journal of Finance* **63**(4), 1977–2011.

Corsi, F. (2009) A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* **7**(2), 174–196.

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* .

Dangl, T. and Halling, M. (2012) Predictive regressions with time-varying coefficients. *Journal of Financial Economics* **106**(1), 157–181.

Daniel, K. and Moskowitz, T. J. (2016) Momentum crashes. *Journal of Financial Economics* **122**(2), 221–247.

Diebold, F. X. and Mariano, R. S. (1995) Comparing predictive accuracy. *Journal of Business & economic statistics* **20**, 134–144.

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J. and Blei, D. (2017) Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, pp. 2732–2741.

Douc, R., Moulines, E. and Olsson, J. (2009) Optimality of the auxiliary particle filter. *Probab. Math. Statist.* **29**(1), 1–28.

Duane, S., Kennedy, A. D., Pendleton, B. J. and Roweth, D. (1987) Hybrid Monte Carlo. *Physics letters B* **195**(2), 216–222.

Durbin, J. and Koopman, S. J. (2000) Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **62**(1), 3–56.

Elezaj, O. and Tole, D. (2018) Big data: Potential, challenges, and implications in official statistics. *CBU International Conference Proceedings* **6**, 95.

Engle, R. F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**(4), 987–1007.

Engle, R. F. and Sun, Z. (2005) Forecasting volatility using tick by tick data .

Fama, E. F. and French, K. R. (1996) Multifactor explanations of asset pricing anomalies. *The journal of finance* **51**(1), 55–84.

Fama, E. F. and French, K. R. (1997) Industry costs of equity. *Journal of financial economics* **43**(2), 153–193.

Fama, E. F. and French, K. R. (2015) A five-factor asset pricing model. *Journal of financial economics* **116**(1), 1–22.

Farmer, L., Schmidt, L. and Timmermann, A. (2019) Pockets of predictability. *Available at SSRN 3152386* .

Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Barekatain, M., Novikov, A., Ruiz, F., Schrittwieser, J., Swirszcz, G., Silver, D., Hassabis, D. and Kohli, P. (2022) Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature* **610**, 47–53.

Ferson, W. E. and Harvey, C. R. (1991) The variation of economic risk premiums. *Journal of political economy* **99**(2), 385–415.

Ferson, W. E. and Harvey, C. R. (1999) Conditioning variables and the cross section of stock returns. *The Journal of Finance* **54**(4), 1325–1360.

Ferson, W. E. and Korajczyk, R. A. (1995) Do arbitrage pricing models explain the predictability of stock returns? *Journal of Business* pp. 309–349.

Fleming, J., Kirby, C. and Ostdiek, B. (2001) The economic value of volatility timing. *Journal of Finance* **56**(1), 329–352.

Frazzini, A., Israel, R. and Moskowitz, T. J. (2012) Trading costs of asset pricing anomalies. *Fama-Miller Working Paper, Chicago Booth Research Paper* (14-05).

Frazzini, A. and Pedersen, L. H. (2014) Betting against beta. *Journal of financial economics* **111**(1), 1–25.

Fulton, C. and Hubrich, K. (2021) Forecasting US inflation in real time. Finance and Economics Discussion Series 2021-014, Board of Governors of the Federal Reserve System (U.S.).

Gallant, A., Hsieh, D. and Tauchen, G. (1997) Estimation of stochastic volatility models with diagnostics. *Journal of Econometrics* **81**(1), 159–192.

Gefang, D., Koop, G. and Poon, A. (2019) Variational Bayesian inference in large Vector Autoregressions with hierarchical shrinkage. *CAMA Working Paper* (2019-08).

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.

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* pp. 339–373.

Ghysels, E., Harvey, A. C. and Renault, E. (1996) 5 stochastic volatility. In *Statistical Methods in Finance*, volume 14 of *Handbook of Statistics*, pp. 119–191. Elsevier.

Gibbons, M. R., Ross, S. A. and Shanken, J. (1989) A test of the efficiency of a given portfolio. *Econometrica: Journal of the Econometric Society* pp. 1121–1152.

Gordon, N., Salmond, D. and Smith, A. (1993) Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)* **140**, 107–113(6).

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

Griffin, J. and Brown, P. (2017) Hierarchical shrinkage priors for regression models. *Bayesian Anal.* **12**(1), 135–159.

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

Gunawan, D., Kohn, R. and Nott, D. (2021) Variational Bayes approximation of factor stochastic volatility models. *International Journal of Forecasting* **37**(4), 1355–1375.

Hahn, P. R. and Carvalho, C. M. (2015) Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association* **110**(509), 435–448.

Hansen, P. R. and Lunde, A. (2005) A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of applied econometrics* **20**(7), 873–889.

Harvey, A., Ruiz, E. and Shephard, N. (1994) Multivariate stochastic variance models. *The Review of Economic Studies* **61**(2), 247–264.

Harvey, C. R., Hoyle, E., Korgaonkar, R., Rattray, S., Sargaison, M. and Van Hemert, O. (2018) The impact of volatility targeting. *The Journal of Portfolio Management* **45**(1), 14–33.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications .

Hauzenberger, N., Huber, F. and Onorante, L. (2021) Combining shrinkage and sparsity in conjugate vector autoregressive models. *Journal of Applied Econometrics* **36**(3), 304–327.

Henkel, S. J., Martin, J. S. and Nardari, F. (2011) Time-varying short-horizon predictability. *Journal of financial economics* **99**(3), 560–580.

Hinton, G. and Van Camp, D. (1993) Keeping neural networks simple by minimizing the description length of the weights. In *in Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory.*

Hosszejni, D. and Kastner, G. (2021) Modeling univariate and multivariate stochastic volatility in R with stochvol and factorstochvol. *Journal of Statistical Software* **100**(12), 1–34.

Hou, K., Xue, C. and Zhang, L. (2015) Digesting anomalies: An investment approach. *The Review of Financial Studies* **28**(3), 650–705.

Huber, F., Koop, G. and Onorante, L. (2021) Inducing sparsity and shrinkage in time-varying parameter models. *Journal of Business & Economic Statistics* **39**(3), 669–683.

Ishwaran, H. and Rao, J. S. (2005) Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics* **33**(2), 730–773.

Jacquier, E., Polson, N. G. and Rossi, P. E. (2002) Bayesian analysis of stochastic volatility models. Volume 20, pp. 69–87. Twentieth anniversary commemorative issue.

Jacquier, E., Polson, N. G. and Rossi, P. E. (2004) Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *J. Econometrics* **122**(1), 185–212.

Jegadeesh, N. and Titman, S. (1993) Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance* **48**(1), 65–91.

Jensen, T. I., Kelly, B. T. and Pedersen, L. H. (2022) Is there a replication crisis in finance? *Journal of Finance* (Forthcoming).

Jobson, J. D. and Korkie, B. M. (1981) Performance hypothesis testing with the sharpe and treynor measures. *Journal of Finance* pp. 889–908.

Johannes, M., Korteweg, A. and Polson, N. (2014) Sequential learning, predictability, and optimal portfolio returns. *Journal of Finance* **69**(2), 611–644.

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.

Kalli, M. and Griffin, J. (2014) Time-varying sparsity in dynamic regression models. *Journal of Econometrics* **178**(2), 779–793.

Kim, S., Shephard, N. and Chib, S. (1998) Stochastic volatility: Likelihood inference and comparison with ARCH models. *The Review of Economic Studies* **65**(3), 361–393.

Koop, G. and Korobilis, D. (2012) Forecasting inflation using dynamic model averaging. *International Economic Review* **53**(3), 867–886.

Koop, G. and Korobilis, D. (2018) Variational Bayes inference in high-dimensional time-varying parameter models .

Koop, G. and Korobilis, D. (2020) Bayesian dynamic variable selection in high dimensions. *International Economic Review* .

Korobilis, D. (2013) Hierarchical shrinkage priors for dynamic regressions with many predictors. *International Journal of Forecasting* **29**(1), 43–59.

Kowal, D. R., Matteson, D. S. and Ruppert, D. (2019) Dynamic shrinkage processes. *Journal of the Royal Statistical Society Series B* **81**(4), 781–804.

Kristóf, T. (2021) Sovereign default forecasting in the era of the covid-19 crisis. *Journal of Risk and Financial Management* **14**(10), 494.

Kullback, S. and Leibler, R. A. (1951) On information and sufficiency. *The annals of mathematical statistics* **22**(1), 79–86.

Ledoit, O. and Wolf, M. (2008) Robust performance hypothesis testing with the sharpe ratio. *Journal of Empirical Finance* **15**(5), 850–859.

Leng, C., Tran, M. N. and Nott, D. (2014) Bayesian adaptive lasso. *Annals of the Institute of Statistical Mathematics* **66**(2), 221–244.

Lewellen, J., Nagel, S. and Shanken, J. (2010) A skeptical appraisal of asset pricing tests. *Journal of Financial economics* **96**(2), 175–194.

Liu, J. and West, M. (2001) *Combined parameter and state estimation in simulation-based filtering*, pp. 197–223. New York, NY: Springer New York.

Luts, J. and Wand, M. P. (2015) Variational inference for count response semiparametric regression. *Bayesian Analysis* **10**(4), 991 – 1023.

McCracken, M. and Ng, S. (2020) Fred-qd: A quarterly database for macroeconomic research. Technical report, National Bureau of Economic Research.

Melino, A. and Turnbull, S. M. (1990) Pricing foreign currency options with stochastic volatility. *Journal of Econometrics* **45**(1), 239–265.

Mendes, E. F., Carter, C. K., Gunawan, D. and Kohn, R. (2020) A flexible particle Markov chain Monte Carlo method. *Stat. Comput.* **30**(4), 783–798.

Menictas, M. and Wand, M. P. (2015) Variational inference for heteroscedastic semiparametric regression. *Australian & New Zealand Journal of Statistics* **57**(1), 119–138.

Menzly, L. and Ozbas, O. (2010) Market segmentation and cross-predictability of returns. *The Journal of Finance* **65**(4), 1555–1580.

Meyer, H., Reudenbach, C., Wöllauer, S. and Nauss, T. (2019) Importance of spatial predictor variable selection in machine learning applications–moving from data reproduction to spatial prediction. *Ecological Modelling* **411**, 108815.

Minka, T. *et al.* (2005) Divergence measures and message passing. Technical report, Technical report, Microsoft Research.

Minka, T. P. (2001) Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 362–369.

Mitchell, T. J. and Beauchamp, J. J. (1988) Bayesian variable selection in linear regression. *Journal of the american statistical association* **83**(404), 1023–1032.

Moreira, A. and Muir, T. (2017) Volatility-managed portfolios. *The Journal of Finance* **72**(4), 1611–1644.

Mudelsee, M. (2019) Trend analysis of climate time series: A review of methods. *Earth-Science Reviews* **190**, 310–322.

Ormerod, J. T. and Wand, M. P. (2010) Explaining variational approximations. *Amer. Statist.* **64**(2), 140–153.

Ormerod, J. T. and Wand, M. P. (2012) Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics* **21**(1), 2–17.

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

Parisi, G. (1988) *Statistical field theory.* Volume 66 of *Frontiers in Physics.* Benjamin/Cummings Publishing Co., Inc., Advanced Book Program, Reading, MA. With a foreword by David Pines.

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

Patton, A. J. and Weller, B. M. (2020) What you see is not what you get: The costs of trading market anomalies. *Journal of Financial Economics* **137**(2), 515–549.

Pettenuzzo, D., Timmermann, A. and Valkanov, R. (2014) Forecasting stock returns under economic constraints. *Journal of Financial Economics* **114**(3), 517–553.

Pitt, M. K. and Shephard, N. (1999) Filtering via simulation: auxiliary particle filters. *J. Amer. Statist. Assoc.* **94**(446), 590–599.

Pitt, M. K. and Shephard, N. (2001) Auxiliary variable based particle filters. In *Sequential Monte Carlo methods in practice*, Stat. Eng. Inf. Sci., pp. 273–293. Springer, New York.

Polson, N. G. and Scott, J. G. (2011) Shrink globally, act locally: sparse Bayesian regularization and prediction. In *Bayesian statistics 9*, pp. 501–538. Oxford Univ. Press, Oxford.

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.

Pourahmadi, M. and Noorbaloochi, S. (2016) Multivariate time series analysis of neuroscience data: some challenges and opportunities. *Current Opinion in Neurobiology* **37**, 12–15. Neurobiology of cognitive behavior.

Primiceri, G. E. (2005) Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies* **72**(3), 821–852.

Raftery, A. E., Kárný, M. and Ettler, P. (2010) Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics* **52**(1), 52–66. PMID: 20607102.

Rapach, D. E., Strauss, J. K. and Zhou, G. (2010) Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies* **23**(2), 821–862.

Ray, P. and Bhattacharya, A. (2018) Signal adaptive variable selector for the horseshoe prior. *arXiv: Methodology* .

Robert, C. P. and Casella, G. (2004) *Monte Carlo statistical methods.* Springer.

Rohde, D. and Wand, M. P. (2016) Semiparametric mean field variational Bayes: General principles and numerical issues. *Journal of Machine Learning Research* **17**(172), 1–47.

Rothman, A. J., Levina, E. and Zhu, J. (2010) A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika* **97**(3), 539–550.

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. *J. Amer. Statist. Assoc.* **113**(521), 431–444.

Ročková, V. and McAlinn, K. (2021) Dynamic variable selection with spike-and-slab process priors. *Bayesian Analysis* **16**(1), 233 – 269.

Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications.* Volume 104 of *Monographs on Statistics and Applied Probability.* London: Chapman & Hall.

Ruiz, E. (1994) Quasi-maximum likelihood estimation of stochastic volatility models. *Journal of Econometrics* **63**(1), 289–306.

Rustagi, J. S. (1976) *Variational methods in statistics.* Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London. Mathematics in Science and Engineering, Vol. 121.

Sakurai, J. J. (1994) *Modern quantum mechanics; rev. ed.* Reading, MA: Addison-Wesley.

Sezer, O. B., Gudelek, M. U. and Ozbayoglu, A. M. (2020) Financial time series forecasting with deep learning : A systematic literature review: 2005–2019. *Applied Soft Computing* **90**, 106181.

Shephard, N. and Pitt, M. K. (1997) Likelihood analysis of non-Gaussian measurement time series. *Biometrika* **84**(3), 653–667.

Shephard, N. and Pitt, M. K. (2004) Erratum: “Likelihood analysis of non-Gaussian measurement time series” [Biometrika **84** (1997), no. 3, 653–667; mr1603940]. *Biometrika* **91**(1), 249–250.

Smith, S. C. and Timmermann, A. (2021) Break risk. *The Review of Financial Studies* **34**(4), 2045–2100.

Soundarya, C. and Usha, S. (2020) Analyzing and predicting cyber hacking with time series models. *International Journal of Research in Engineering, Science and Management* **3**(7), 1–8.

Speekenbrink, M. (2016) A tutorial on particle filters. *Journal of Mathematical Psychology* **73**, 140–152.

Stock, J. H. and Watson, M. W. (2010) Modeling inflation after the crisis. Technical report, National Bureau of Economic Research.

Storvik, G. (2002) Particle filters for state-space models with the presence of unknown static parameters. *IEEE Transactions on Signal Processing* **50**(2), 281–289.

Tang, X., Ghosh, M., Xu, X. and Ghosh, P. (2018) Bayesian variable selection and estimation based on global-local shrinkage priors. *Sankhya A* **80**(2), 215–246.

Taylor, S. J. (1994) Modeling stochastic volatility: A review and comparative study. *Mathematical finance* **4**(2), 183–204.

Varian, H. R. (2014) Big data: New tricks for econometrics. *Journal of Economic Perspectives* **28**(2), 3–28.

Wand, M. P. (2014) Fully simplified multivariate normal updates in non-conjugate variational message passing. *Journal of Machine Learning Research* .

Wand, M. P., Ormerod, J. T., Padoan, S. A. and Frühwirth, R. (2011) Mean field variational Bayes for elaborate distributions. *Bayesian Analysis* **6**(4), 847–900.

Wang, F. and Yan, X. S. (2021) Downside risk and the performance of volatility-managed portfolios. *Journal of Banking & Finance* **131**, 106198.

Wang, Y. and Blei, D. M. (2018) Frequentist consistency of variational Bayes. *Journal of the American Statistical Association* pp. 1–15.

West, M. and Harrison, J. (2006) *Bayesian forecasting and dynamic models.* Springer Science & Business Media.

Xu, M. and Duan, L. L. (2020) Bayesian inference with the L1-ball prior: Solving combinatorial problems with exact zeros.

Zhang, C., Bütepage, J., Kjellström, H. and Mandt, S. (2019) Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence* **41**(8), 2008–2026.

Zhang, F. and Gao, C. (2020) Convergence rates of variational posterior distributions. *The Annals of Statistics* **48**(4), 2180–2207.

Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**(476), 1418–1429.

Nicolas Bianco

Dept. of Statistical Sciences, University of Padova
Via Cesare Battisti, 241–243,
35121, Padova, Italy

Email: nicolas.bianco@phd.unipd.it
Website: whitenoise8.github.io
Phone: 0039 340-165-2200

Research interests

Bayesian inference, computational methods, dynamic sparsity, high-dimensional statistics, sparse methods, time series analysis, time-varying parameter, variational approximations

Current position

- **PhD student in Statistics** Padova, Italy
  Dept. of Statistical Sciences, University of Padova *Oct 2019 - Dec 2022*
  Supervisor: Mauro Bernardi; Co-supervisior: Daniele Bianchi
  Expected Thesis defense: April-May 2023

Education

- **Master degree in Statistical Sciences** Padova, Italy
  Dept. of Statistical Sciences, University of Padova *Oct 2017 - Sep 2019*
  Final grade: 110/110 cum laude
  Thesis title: Variational inference in sparse high-dimensional models
  Supervisor: Mauro Bernardi; Co-supervisior: Daniele Bianchi
- **Erasmus+** Lisbon, Portugal
  ISEG – Instituto Superior de Economia e Gestão *Feb 2017 - Jul 2017*
- **Bachelor degree in Statistics for Economics and Business** Padova, Italy
  Dept. of Statistical Sciences, University of Padova *Oct 2014 - Sep 2017*
  Final grade: 110/110
  Thesis title: Non-negative constrained penalised matching quantiles estimation
  Supervisor: Mauro Bernardi

Awards and Grants

- **Best poster award** Padova, Italy
  Conference: Statistical Methods and Models for Complex Data (prize of 1500 €) *Sep 2022*
  Evaluation committee: Frühwirth-Schnatter, Sylvia; Gijbels, Irène; Sartori, Nicola
- **Students travel award 2022** ISBA, Montreal, Canada
  Economic support (500 $) to participate at the conference *Jun 2022*
- **Prize Oliviero Lessi** SIS – Italian Statistical Society, Rome, Italy
  Best Master Thesis in mathematical Statistics (prize of 500 €) *Jul 2020*
- **Stats Under the Stars – V edition** Bocconi University, Milan, Italy
  Best overall project *Jun 2019*
- **Hackathon on speech recognition** Unox S.p.A, Padua, Italy
  First place position in the competition *May 2018*

Research

Articles

1. Bianco N., Bernardi M., Bianchi D. (2022). Smoothing volatility targeting. *arXiv preprint arXiv:2212.07288*
2. Bianco N., Bernardi M., Bianchi D. (2022). Sparse multivariate modeling for stock returns predictability. *arXiv preprint arXiv:2202.12644*
3. Bianco N., Bernardi M., Bianchi D. (2022). Bernoulli-Gaussian model for dynamic sparsity in time varying parameter regression. *Proceedings of the 36th International Workshop on Statistical Modeling, Trieste*
4. Bianco N., Bernardi M., Bianchi D. (2021). Variational inference for sparse vector autoregression. *Proceedings of the 35th International Workshop on Statistical Modeling, Bilbao, 19–24*

Ongoing research

1. Variational Inference for dynamic sparsity in time varying parameter models (with Bernardi M. and Bianchi D.)
2. A Bayesian approach to dynamic quantiles network (with Roverato A., Bernardi M. and Castiglione C.)

Conferences

12th European Seminar on Bayesian Econometrics, Salzburg (contributed); 24th International Conference on Computational Statistics, Bologna (invited); 36th International Workshop on Statistical Modeling, Trieste (contributed); World meeting of the International Society for Bayesian Analysis, Montreal (invited); 14th International Conference of the ERCIM WG on Computational and Methodological Statistics (invited, online); 35th International Workshop on Statistical Modeling (contributed, online); World meeting of the International Society for Bayesian Analysis (contributed, online); 13th International Conference of the ERCIM WG on Computational and Methodological Statistics (contributed, online).

Visiting periods

- **Queen Mary university of London** London, UK
  School of Economics and Finance *Oct 2022 – Dec 2022*

Teaching experience

- **Academic tutor** Padova, Italy
  Dept. of Statistical Sciences, University of Padova *Oct 2018 – Sep 2019*
  Courses: Calculus (bachelor degree), Statistics (advanced course, master degree)

Service at the department

- **PhD students representative** Padova, Italy
  Dept. of Statistical Sciences, University of Padova *Oct 2019 - Dec 2022*
  Member of department council and PhD program council
- **Senior year career guidance** Padova, Italy
  Dept. of Statistical Sciences, University of Padova *2017 - 2022*
- **Tutor for scientific degrees national program (PNLS)** Padova, Italy
  Dept. of Statistical Sciences, University of Padova *Oct 2017 - Sep 2018*

Other

- **Technologies**: R/Rstudio (advanced), C++ (advanced), LaTex (advanced), Python (intermediate), Matlab (basic), Julia (basic)
- **Languages**: Italian (native); English (fluent); Portuguese (basic)