Variational Bayes inference for large-scale multivariate predictive regressions*

Mauro Bernardi† Daniele Bianchi‡ Nicolas Bianco†

First draft: January 2021. This draft: November 28, 2022

Abstract

We propose a novel variational Bayes approach to estimate large-scale multivariate linear predictive regressions. Differently from conventional Bayesian inference methods, our approach does not rely on a Cholesky-based transformation of the parameters space. This allows to elicit hierarchical shrinkage priors directly on the matrix of regression coefficients. An extensive simulation study provides evidence that our approach produces more accurate estimates under different sparsity assumptions. We test empirically both the statistical and economic performance 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. The results show that more accurate estimates translate into substantial statistical and economic out-of-sample gains compared to existing Bayesian estimation methods. Both the simulation and empirical results hold across different hierarchical shrinkage priors and model dimensions.

Keywords: Bayesian methods, mean-field approximation, variational Bayes, hierarchical shrinkage prior, multivariate regressions, returns predictability.

JEL codes: C11, C32, C55, C53, G11

*We are thankful to Andrea Carriero and seminar participants at the 2021 Virtual NBER-NSF SBIES, the 2021 European Summer Meeting of the Econometric Society and the 2nd Workshop on Dimensionality Reduction and Inference in High-Dimensional Time Series at Maastricht University for their helpful comments and suggestions. A previous version of this paper was circulating with the title "Sparse multivariate modeling for stock returns predictability".

†Department of Statistical Sciences, University of Padova, Italy. Email: mauro.bernardi@unipd.it

‡School of Economics and Finance, Queen Mary University of London, United Kingdom. Email: d.bianchi@qmul.ac.uk Web: whitesphd.com

†Department of Statistical Sciences, University of Padova, Italy. Email: nicolas.bianco@phd.unipd.it Web: whitenoise8.github.io

1 Introduction

Within a Bayesian linear regression context, parameters regularization is often based on continuous shrinkage priors.[1] 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 based on the Cholesky factor. While this greatly simplifies posterior inference – especially for high-dimensional models –, it potentially prevents to recover the original structure of the regression coefficients.

In this paper, we take a different approach towards the identification of the regression coefficients in multivariate predictive 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 hierarchical shrinkage priors directly on the matrix of regression coefficients, differently from benchmark Bayesian inference methods.

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,

[1] 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), and Griffin and Brown (2017), among others.

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 variables permutations. 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 paper 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 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 $\mathbf{x}_t = (1, x_{1,t}, \dots, x_{p,t})^\intercal \in \mathbb{R}^{(p+1)}$ be a vector of exogenous covariates observed at time t. A multivariate predictive regression model with constant volatility is defined as follows:

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

with $\mathbf{\Theta} = (\mathbf{\Gamma}, \mathbf{\Phi})$ where $\mathbf{\Gamma} \in \mathbb{R}^{d\times(p+1)}$ is the matrix of regression parameters for the exogenous predictors and $\mathbf{\Phi} \in \mathbb{R}^{d\times d}$ is the transition matrix containing the autoregressive parameters, $\mathbf{z}_{t-1} = (\mathbf{x}_{t-1}^\intercal, \mathbf{y}_{t-1}^\intercal)^\intercal$. 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 $k, j = \pm 1, \pm 2, \dots$ and covariance matrix equal to $\mathbf{\Omega}^{-1}$, with $\mathbf{\Omega} \in \mathbb{S}_{++}^d$ being a symmetric and positive definite precision matrix.

The modified Cholesky factorization of the precision matrix $\mathbf{\Omega}$ can be conveniently exploited to re-write the model in Eq.(1) with orthogonal innovations, (see, e.g. Rothman et al., 2010). Let $\mathbf{\Omega} = \mathbf{L}^\intercal\mathbf{V}\mathbf{L}$, 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 (1) by $\mathbf{L} = \mathbf{I}_d - \mathbf{B}$. After some simple algebra one can obtain two alternative

parametrizations of the same model:

$$\mathbf{y}_t = \mathbf{B}(\mathbf{y}_t - \boldsymbol{\Theta}\mathbf{z}_{t-1}) + \boldsymbol{\Theta}\mathbf{z}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathsf{N}_d(\mathbf{0}_d, \mathbf{V}^{-1}), \tag{2a}$$

$$\mathbf{y}_t = \mathbf{B}\mathbf{y}_t - \mathbf{A}\mathbf{z}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathsf{N}_d(\mathbf{0}_d, \mathbf{V}^{-1}), \tag{2b}$$

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 $j = 2, \ldots, d$ and $k = 1, \ldots, j-1$. The key difference is that Eq.(2a) shows non-linearity in the parameters, while Eq.(2b) is linear. More importantly, Eq.(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 Eq.(2a) is the parametrization at the core of our variational Bayes approach. From Eq.(2) one can obtain an equation-by-equation representation in which the j-th component of $\mathbf{y}_t$ becomes:

$$y_{j,t} = \boldsymbol{\beta}_j \mathbf{r}_{j,t} + \boldsymbol{\vartheta}_j \mathbf{z}_{t-1} + \varepsilon_{j,t}, \qquad \varepsilon_{j,t} \sim \mathsf{N}(0, 1/\nu_j), \tag{3a}$$

$$y_{j,t} = \boldsymbol{\beta}_j \mathbf{y}_t^j + \mathbf{a}_j \mathbf{z}_{t-1} + \varepsilon_{j,t}, \qquad \varepsilon_{j,t} \sim \mathsf{N}(0, 1/\nu_j), \tag{3b}$$

for all $j = 1, \ldots, d$ and $t = 2, 3, \ldots$, 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 $\mathbf{r}_{j,t} = \mathbf{y}_t^j - \boldsymbol{\Theta}^j \mathbf{z}_{t-1}$ 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 $\boldsymbol{\Theta}^j \in \mathbb{R}^{(j-1)\times d}$ is the sub-matrix containing the first $j-1$ rows of $\boldsymbol{\Theta}$.

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 $\boldsymbol{\Theta}$, we consider for simplicity that each of the elements in $\boldsymbol{\nu}$ 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 Eq.(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 $\mathbf{A} = \mathbf{L}\boldsymbol{\Theta}$, each element $\vartheta_{i,j}$, which denotes the (i,j)-entry of $\boldsymbol{\Theta}$, 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 $\boldsymbol{\Theta}$; and *ii)* the estimate $\boldsymbol{\Theta} = \mathbf{L}^{-1}\mathbf{A}$ is not permutation invariant, which is a direct consequence of the Cholesky factorization. Figure 1 provides a numerical representation of this argument. We compare the posterior estimates obtained from a `LVB` method based on Eq.(2a) versus our `VB` approach based on Eq.(2b), for two different permutations of $\mathbf{y}_t$.

The evidence confirms that the estimates based on the transformation $\boldsymbol{\Theta} = \mathbf{L}^{-1}\mathbf{A}$ do not match with the true $\boldsymbol{\Theta}$. In addition, the estimates are influenced by the variables permutation. Instead, our `VB` approach provides a more accurate, permutation invariant, identification of $\boldsymbol{\Theta}$. 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.

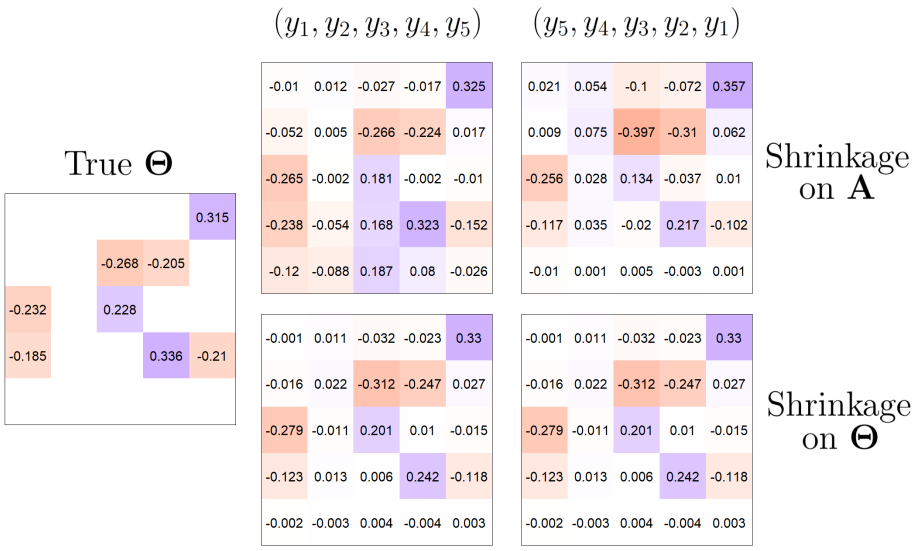


Figure 1: Comparison between the posterior inference for the linear representation $\mathbf{A} = \mathbf{L}\boldsymbol{\Theta}$ (first row) and the original parametrization $\boldsymbol{\Theta}$ (second row), for two different permutations of $\mathbf{y}_t$.

3 Variational Bayes inference

A variational Bayes approach to inference requires to minimize the Kullback-Leibler (KL) divergence between an approximating density $q(\boldsymbol{\psi})$ and the true posterior density $p(\boldsymbol{\psi}|\mathbf{y})$, (see, e.g. Ormerod and Wand, 2010; Blei et al., 2017). The KL divergence cannot be directly minimized with respect to $\boldsymbol{\psi}$ because it involves the expectation with respect to the unknown true posterior distribution. Ormerod and Wand (2010) show that the problem of minimizing KL can be equivalently stated as the maximization of the variational lower bound (ELBO) denoted by $\underline{p}\,(\mathbf{y}; q)$:

$$q^*(\boldsymbol{\psi}) = \arg \max_{q(\boldsymbol{\psi}) \in \mathcal{Q}} \log \underline{p}\,(\mathbf{y}; q)\,, \quad \underline{p}\,(\mathbf{y}; q) = \int q(\boldsymbol{\psi}) \log \left\{ \frac{p(\mathbf{y}, \boldsymbol{\psi})}{q(\boldsymbol{\psi})} \right\} d\boldsymbol{\psi}, \tag{4}$$

where $q^*(\boldsymbol{\psi}) \in \mathcal{Q}$ represents the optimal variational density and $\mathcal{Q}$ is a space of functions. Depending on the assumption on the space $\mathcal{Q}$, we fall into different variational paradigms. The mean-field variational Bayes (MFVB) approach only assumes a non-parametric restriction for the variational density, i.e. $q(\boldsymbol{\psi}) = \prod_{i=1}^{p} q_i(\boldsymbol{\psi}_i)$ for a partition $\{\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p\}$ of the parameter vector $\boldsymbol{\psi}$. Under the MFVB restriction, a closed form expression for the optimal

variational density of each component $q(\boldsymbol{\psi}_j)$ is defined as:

$$q^*(\boldsymbol{\psi}_j) \propto \exp\left\{\mathbb{E}_{q^\star(\boldsymbol{\psi}\setminus\boldsymbol{\psi}_j)}\Big[\log p(\mathbf{y},\boldsymbol{\psi})\Big]\right\}, \qquad q^\star(\boldsymbol{\psi}\setminus\boldsymbol{\psi}_j) = \prod_{\substack{i=1\\ i\neq j}}^{p} q_i(\boldsymbol{\psi}_i), \tag{5}$$

where the expectation is taken with respect to the joint approximating density with the j-th element of the partition removed $q^\star(\boldsymbol{\psi}\setminus\boldsymbol{\psi}_j)$. This allows to implement a relatively easy iterative algorithm to estimate the optimal density $q^*(\boldsymbol{\psi})$.

Equation (5) shows that the factorization of $q(\boldsymbol{\psi})$ 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 $\boldsymbol{\Theta}$. In the main text we will summarize the optimal approximating density $q^\star$, show how to perform approximate inference on $\boldsymbol{\Omega}$, and illustrate how to make predictions within this framework. For the interested reader, in Appendix B we provide the full set of derivations of the optimal variational densities together with the analytical form of the lower bound and the correspondent iterative algorithms.

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 $\boldsymbol{\Theta}$, let $\vartheta_{j,k} \sim \mathsf{N}(0, \upsilon)$, for $j = 1,\dots,d$ and $k = 1,\dots,d+p+1$. In addition, let $\nu_j \sim \mathsf{Ga}(a_\nu, b_\nu)$ for $j = 1,\dots,d$, and $\beta_{j,k} \sim \mathsf{N}(0,\tau)$, for $j = 2,\dots,d$ and $k = 1,\dots,j-1$. 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:

$$q(\boldsymbol{\xi}) = q(\boldsymbol{\nu})q(\boldsymbol{\beta})q(\boldsymbol{\vartheta}), \quad q(\boldsymbol{\nu}) = \prod_{j=1}^{d} q(\nu_j), \quad q(\boldsymbol{\beta}) = \prod_{j=2}^{d} q(\boldsymbol{\beta}_j), \quad q(\boldsymbol{\vartheta}) = \prod_{j=1}^{d} q(\boldsymbol{\vartheta}_j). \tag{6}$$

Proposition 3.1 provides the optimal variational density for this default normal prior specification. The proof and analytical derivations are available in Appendix B.1.

Proposition 3.1. *The optimal variational densities for* ν_j *and* $\boldsymbol{\beta}_j$ *are* $q^*(\nu_j) \equiv \mathsf{Ga}(a_{q(\nu_j)}, b_{q(\nu_j)})$ *and* $q^*(\boldsymbol{\beta}_j) \equiv \mathsf{N}_{j-1}(\boldsymbol{\mu}_{q(\beta_j)}, \boldsymbol{\Sigma}_{q(\beta_j)})$. *The hyper-parameters* $a_{q(\nu_j)}, b_{q(\nu_j)}$ *and* $\boldsymbol{\mu}_{q(\beta_j)}, \boldsymbol{\Sigma}_{q(\beta_j)}$ *are defined in Eq.(B.1) and Eq.(B.2). The optimal variational density for the* j*-th row of the matrix* $\boldsymbol{\Theta}$ *is a multivariate gaussian* $q^*(\boldsymbol{\vartheta}_j) \equiv \mathsf{N}_{d+p+1}(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$ *with hyper-parameters:*

$$
\begin{aligned}
\boldsymbol{\Sigma}_{q(\vartheta_j)} &= \left(\boldsymbol{\mu}_{q(\omega_{j,j})} \sum_{t=1}^{T} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^{\intercal} + 1/\upsilon \mathbf{I}_{d+p+1} \right)^{-1}, \\
\boldsymbol{\mu}_{q(\vartheta_j)} &= \boldsymbol{\Sigma}_{q(\vartheta_j)} \left(\sum_{t=1}^{T} \left(\boldsymbol{\mu}_{q(\omega_j)} \otimes \mathbf{z}_{t-1} \right) \mathbf{y}_t - \left(\boldsymbol{\mu}_{q(\omega_{j,-j})} \otimes \sum_{t=1}^{T} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^{\intercal} \right) \boldsymbol{\mu}_{q(\vartheta_{-j})} \right),
\end{aligned}
\tag{7}
$$

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

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

Note from Proposition 3.1 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).

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 $\vartheta_{j,k}|\lambda_{j,k} \sim \mathsf{Lap}(\lambda_{j,k})$, for $j = 1, \ldots, d$ and $k = 1, \ldots, d+p+1$. The latter can be represented as a scale mixture of Gaussians with an exponential mixing density, $\vartheta_{j,k}|\upsilon_{j,k} \sim \mathsf{N}(0, \upsilon_{j,k})$, $\upsilon_{j,k}|\lambda_{j,k}^2 \sim \mathsf{Exp}(\lambda_{j,k}^2/2)$.

The choice of the scaling parameters $\lambda_{j,k}^2$ 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 hyper-prior distribution $\lambda_{j,k}^2 \sim \mathsf{Ga}(h_1, h_2)$, 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. The joint distribution $q(\boldsymbol{\xi}_{\mathrm{L}})$ can be factorised as:

$$q(\boldsymbol{\xi}_{\mathrm{L}}) = q(\boldsymbol{\xi})q(\boldsymbol{\upsilon}, \boldsymbol{\lambda}^2), \qquad q(\boldsymbol{\upsilon}, \boldsymbol{\lambda}^2) = \prod_{j=1}^{d} \prod_{k=1}^{d+p+1} q(\upsilon_{j,k})q(\lambda_{j,k}^2), \tag{8}$$

Proposition 3.2 provides the optimal variational densities for the Bayesian adaptive lasso prior. The proof and analytical derivation of the parameters are available in Appendix B.2.

Proposition 3.2. *The optimal variational densities for ν_j and $\boldsymbol{\beta}_j$ are the same as in Proposition 3.1. The distribution $q^*(\boldsymbol{\vartheta}_j)$ is a multivariate gaussian as in Proposition 3.1 with covariance matrix $\boldsymbol{\Sigma}_{q(\vartheta_j)} = \left(\boldsymbol{\mu}_{q(\omega_{j,j})} \sum_{t=1}^{T} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal + diag(\boldsymbol{\mu}_{q(1/\upsilon)})\right)^{-1}$. For the scaling parameters we have that $q^*(\lambda_{j,k}^2) \equiv \mathsf{Ga}(a_{q(\lambda_{j,k}^2)}, b_{q(\lambda_{j,k}^2)})$ with $a_{q(\lambda_{j,k}^2)}, b_{q(\lambda_{j,k}^2)}$ defined in Eq.(B.10), and $q^*(1/\upsilon_{j,k}) \equiv \mathsf{IG}(a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})})$ is an inverse-gaussian distribution with the parameters $a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})}$ defined in Eq.(B.9).*

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 $j = 1, \ldots, d$ and $k = 1, \ldots, d + p + 1$. 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:

$$q(\boldsymbol{\xi}_{\text{NG}}) = q(\boldsymbol{\xi})q(\boldsymbol{\upsilon}, \boldsymbol{\lambda}, \boldsymbol{\eta}), \qquad q(\boldsymbol{\upsilon}, \boldsymbol{\lambda}, \boldsymbol{\eta}) = \prod_{j=1}^{d} q(\eta_j) \prod_{k=1}^{d+p+1} q(\upsilon_{j,k})q(\lambda_{j,k}). \tag{9}$$

Proposition 3.3 provides the optimal variational densities for the adaptive normal-gamma prior. The proof and analytical derivation of the parameters are available in Appendix B.3.

Proposition 3.3. *The optimal variational densities for ν_j and $\boldsymbol{\beta}_j$ are the same as in Proposition 3.1. The distribution of $q^*(\boldsymbol{\vartheta}_j)$ is a multivariate gaussian as in Proposition 3.1 with covariance matrix $\boldsymbol{\Sigma}_{q(\vartheta_j)} = \left(\boldsymbol{\mu}_{q(\omega_{j,j})} \sum_{t=1}^{T} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal + diag(\boldsymbol{\mu}_{q(1/\upsilon)})\right)^{-1}$. 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 $a_{q(\lambda_{j,k})}, b_{q(\lambda_{j,k})}$ defined in Eq.(B.13), 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 parameters $\zeta_{q(\upsilon_{j,k})}, a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})}$ defined in Eq.(B.12).*

Note that 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 (B.15).

Horseshoe prior

Finally we consider an horseshoe prior originally as proposed by Carvalho et al. (2009, 2010). This is based on the hierarchical specification $\vartheta_{j,k}|\upsilon_{j,k}^2, \gamma^2 \sim \mathsf{N}(0, \gamma^2\upsilon_{j,k}^2)$, $\gamma \sim \mathsf{C}^+(0, 1)$, $\upsilon_{j,k} \sim \mathsf{C}^+(0, 1)$, where $\mathsf{C}^+(0, 1)$ denotes the standard half-cauchy distribution with probability density function equal to $f(x) = 2/\{\pi(1 + x^2)\}\mathbb{1}_{(0,\infty)}(x)$. 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:

$$\begin{aligned} \vartheta_{j,k}|\upsilon^2_{j,k},\gamma^2 \sim \mathsf{N}(0,\gamma^2\upsilon^2_{j,k}), \quad \gamma^2|\eta \sim \mathsf{InvGa}(1/2,1/\eta), \quad \upsilon^2_{j,k}|\lambda_{j,k} \sim \mathsf{InvGa}(1/2,1/\lambda_{j,k}), \\ \eta \sim \mathsf{InvGa}(1/2,1), \qquad \lambda_{j,k} \sim \mathsf{InvGa}(1/2,1), \end{aligned} \tag{10}$$

where $\mathsf{InvGa}(\cdot,\cdot)$ denotes the inverse gamma distribution, and the local and overall shrinkage are determined by $\upsilon^2_{j,k}$ 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:

$$q(\boldsymbol{\xi}_{\text{HS}}) = q(\boldsymbol{\xi})q(\boldsymbol{\upsilon}^2,\gamma^2,\boldsymbol{\lambda},\eta), \qquad q(\boldsymbol{\upsilon}^2,\gamma^2,\boldsymbol{\lambda},\eta) = q(\gamma^2)q(\eta)\prod_{j=1}^{d}\prod_{k=1}^{d+p+1} q(\upsilon^2_{j,k})q(\lambda_{j,k}). \tag{11}$$

Proposition 3.4 provides the optimal variational densities for the horseshoe. The proof and analytical derivation of the parameters are available in Appendix B.4.

Proposition 3.4. *The optimal variational densities for ν_j and $\boldsymbol{\beta}_j$ are the same as in Proposition 3.1. The distribution family of $q^*(\boldsymbol{\vartheta}_j)$ is a multivariate gaussian as in Proposition 3.1 having variance matrix $\boldsymbol{\Sigma}_{q(\vartheta_j)} = \left(\boldsymbol{\mu}_{q(\omega_{j,j})}\sum_{t=1}^{T}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal + \mu_{q(1/\gamma^2)}diag(\boldsymbol{\mu}_{q(1/\upsilon^2)})\right)^{-1}$. For the global shrinkage parameters we have that $q^*(\gamma^2) \equiv \mathsf{InvGa}(\frac{1}{2}\{d(d+p+1)+1\}, b_{q(\gamma^2)})$ with $b_{q(\gamma^2)}$ defined in Eq.(B.22), and $q^*(\eta) \equiv \mathsf{InvGa}(1, b_{q(\eta)})$ where $b_{q(\eta)}$ is defined in Eq.(B.24). For the local shrinkage parameters we have that $q^*(\upsilon^2_{j,k}) \equiv \mathsf{InvGa}(1, b_{q(\upsilon^2_{j,k})})$ and $q^*(\lambda_{j,k}) \equiv \mathsf{InvGa}(1, b_{q(\lambda_{j,k})})$, where $b_{q(\upsilon^2_{j,k})}$ and $b_{q(\lambda_{j,k})}$ are defined in Eq.(B.21) and Eq.(B.23) respectively.*

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 (1) we still need an approximating density for the precision matrix $\boldsymbol{\Omega} = \mathbf{L}^\intercal \mathbf{V} \mathbf{L}$.

Proposition 3.5 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 complete proof is available in Appendix B.5. To simplify, the proof is based on the Expectation Propagation (EP) variational procedure of Minka (2001), which has the goal of minimizing the KL divergence between the true and unknown distribution $p(\boldsymbol{\Omega})$ and the approximating density $q(\boldsymbol{\Omega})$.

Proposition 3.5. *The approximate distribution q of $\boldsymbol{\Omega}$ is $\mathsf{Wishart}_d(\widehat{\delta}, \widehat{\mathbf{H}})$, where the scaling matrix is given by $\widehat{\mathbf{H}} = \widehat{\delta}^{-1} \mathbb{E}_p[\boldsymbol{\Omega}]$ and $\widehat{\delta}$ can be obtained numerically as the solution of a convex optimization problem.*

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 $q(\boldsymbol{\Omega})$ exploiting the modified Cholesky decomposition. Table 1 compares the sampled distributions with the marginals of the Wishart with parameters $(\widehat{\delta}, \widehat{\mathbf{H}})$ in terms of approximation accuracy $\mathcal{ACC} = 100 \left\{1 - 0.5 \int |q(\omega) - p(\omega)| \, d\omega\right\}\%$, where ω is a generic element of $\boldsymbol{\Omega}$.

	$d = 15$		$d = 30$		$d = 50$		$d = 100$	
	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 1: Accuracy (%) of the whishart approximation $q(\boldsymbol{\Omega})$ for dimensions $d = 15, 30, 50, 100$ separately for the diagonal (ω_{jj}) and out-of-diagonal (ω_{jk}) elements of $\boldsymbol{\Omega}$.

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

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 paper, 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 $\boldsymbol{\Theta}$, 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.

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,

$$p(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = \int p(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\xi}) p(\boldsymbol{\xi}|\mathbf{z}_{1:t}) d\boldsymbol{\xi}. \tag{12}$$

Given an optimal variational density $q^*(\boldsymbol{\xi})$ that approximates $p(\boldsymbol{\xi}|\mathbf{z}_{1:t})$, we follow Gunawan et al. (2020) and obtain the variational predictive distribution

$$q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = \int p(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\xi}) q^*(\boldsymbol{\xi}) d\boldsymbol{\xi} = \int \int p(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\vartheta}, \boldsymbol{\Omega}) q(\boldsymbol{\vartheta}) q(\boldsymbol{\Omega}) d\boldsymbol{\vartheta}\, d\boldsymbol{\Omega}. \tag{13}$$

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 Eq.(13) by integrating $\boldsymbol{\Omega}$ such that:

$$q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = \int h(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\vartheta}) q(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}, \tag{14}$$

where $h(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\vartheta})$ denotes the density function of a multivariate Student-t distribution $\mathbf{t}_v(\mathbf{m}, \mathbf{S})$ 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 C.

4 Simulation study

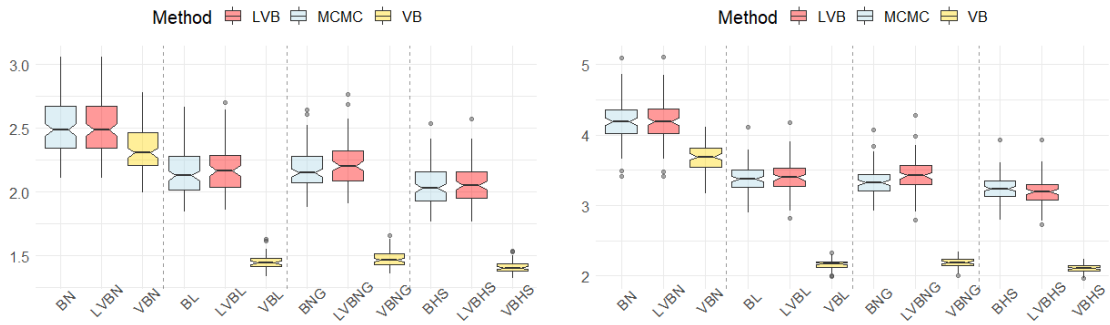
We now perform an extensive simulation study to compare the properties of our `VB` approach against standard MCMC and `LVB` approaches. Consistent with the empirical application, we set the length of the time series $T = 360$ and the cross-sectional dimension of the data generating process $d = 30, 49$.[2] 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 $\mathbf{\Theta}$ 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 D 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 3.1. Recall that while our approach is based on the parametrization in Eq.(2a), both the competing MCMC and linearized variational Bayes approach are built upon the linearized system of equations implied by Eq.(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).

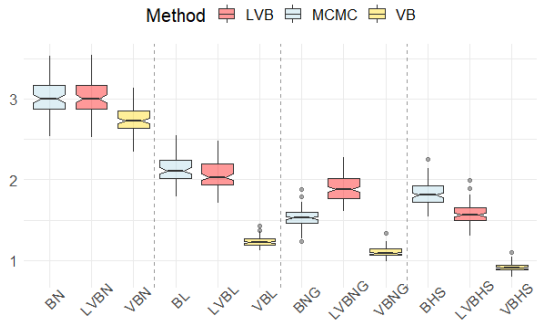
As a measure of point estimation accuracy of the regression matrix $\mathbf{\Theta}$, we first look at the Frobenius norm, denoted by $\|\cdot\|_F$. This measure the difference between the true matrix $\mathbf{\Theta}$, which is observed at each simulation, and the corresponding estimate $\widehat{\mathbf{\Theta}}$. Figure 2 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.

[2] Additional results with $d = 15$ are reported in Appendix D.

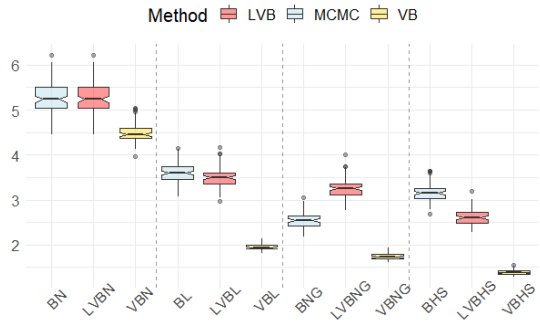


(a) $d = 30$, moderate sparsity

(b) $d = 49$, moderate sparsity



(c) $d = 30$, high sparsity

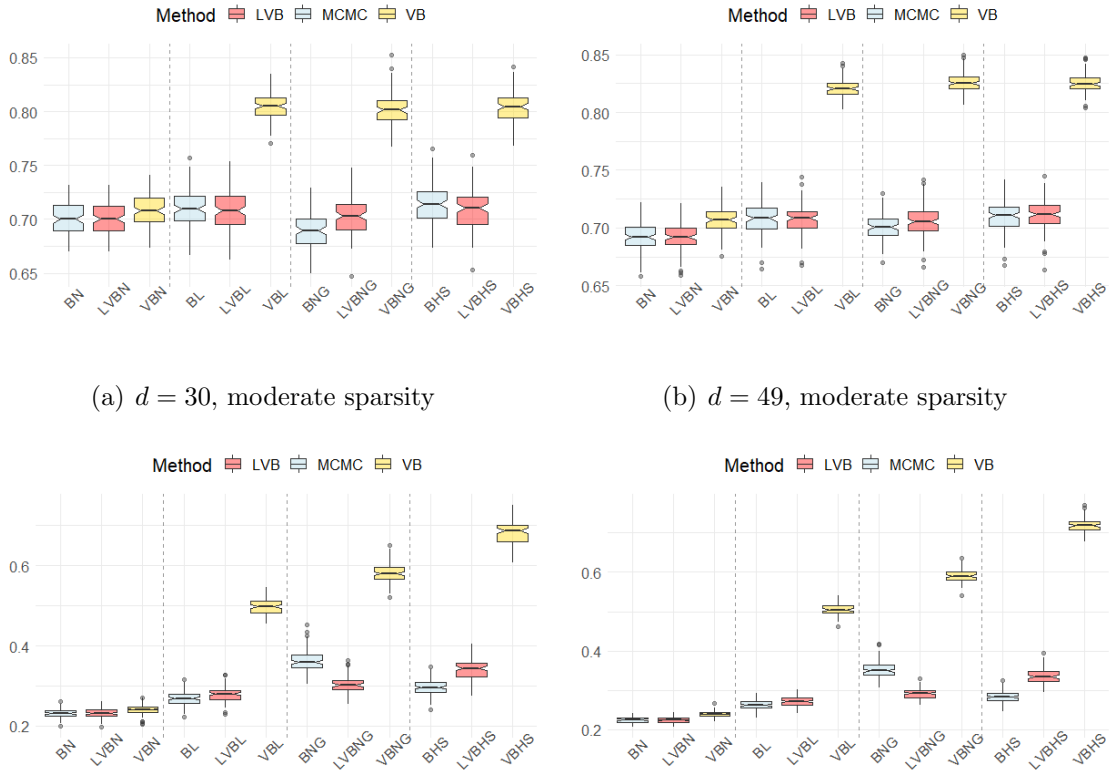


(d) $d = 49$, high sparsity

Figure 2: Frobenius norm of $\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}$ for different hierarchical shrinkage priors and different inference approaches.

Beginning with the moderate sparsity case (i.e. 50% of zeros in $\boldsymbol{\Theta}$), 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 $d = 49$, and is reassuring, considering both approaches are built upon the same Cholesky-based parametrization (see Eq.2a). 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 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 3 for more details).



(a) $d = 30$, moderate sparsity

(b) $d = 49$, moderate sparsity

(c) $d = 30$, high sparsity

(d) $d = 49$, high sparsity

Figure 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.

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 $\boldsymbol{\Theta} = \mathbf{L}^{-1}\mathbf{A}$ from the estimated $\widehat{\mathbf{A}}$ leads to a dense $\widehat{\boldsymbol{\Theta}}$, and therefore a lower identification accuracy. This is not the case for our variational Bayes approach which directly shrinks the regression matrix $\boldsymbol{\Theta}$, 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 D show that the higher accuracy of our framework is preserved in smaller-dimensional settings (i.e. $d = 15$). In addition, compared to the MCMC approach, Appendix D shows that our `VB` estimation scheme is computationally faster.

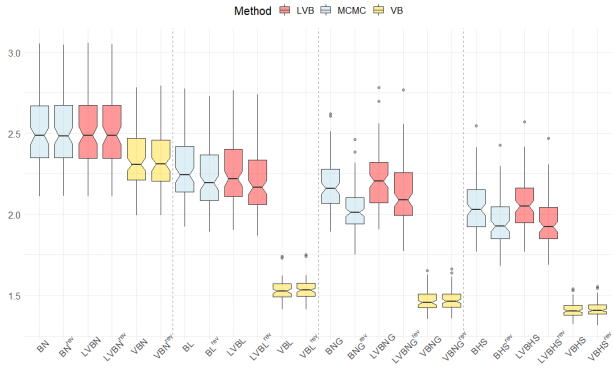
4.1 Performance under variables permutation

Based on the same simulation setting described above, we now investigate the performance of all estimation methods under variables permutation. Panel A of Figure 4 shows the box charts of the Frobenius norms for the $N = 100$ replications for both moderate and high sparsity in the true $\mathbf{\Theta}$.[3] 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.

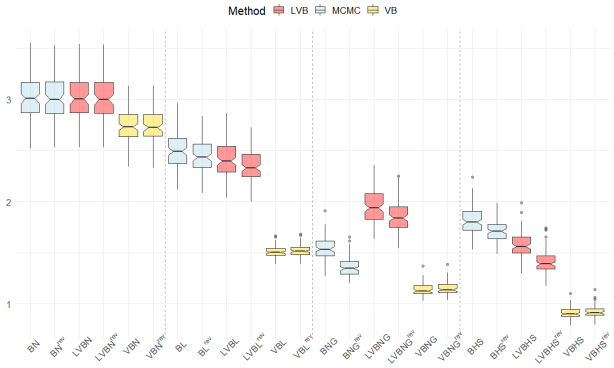
The accuracy of the estimates from both the MCMC and the linearized variational Bayes approach is affected by the variables ordering. This is perhaps more evident for the normal-gamma (`BNG`, `LVBNG`) and the Horseshoe (`BHS`, `LVBHS`) priors. The accuracy of the posterior estimates once the ordering of the variables is reversed (labelled with the supscript "rev") is substantially higher, on average for both methods. The impact of the variables ordering on the posterior estimates from both MCMC and `LVB` is stronger for a highly sparse $\mathbf{\Theta}$ (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 4 compares the F1-score under variables permutation across different estimation methods and hierarchical shrinkage priors. Interestingly, when the regression matrix $\mathbf{\Theta}$ is moderately sparse, the ordering of the target variables has almost a

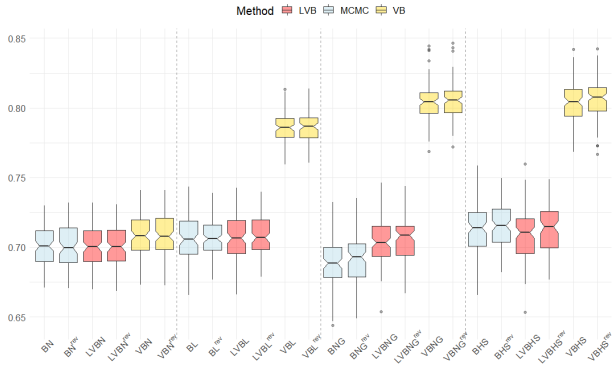
[3]For the ease of exposition we only report the case with $d = 30$ predictors. The case with $d = 49$ allows to draw qualitatively similar conclusions. The results are available upon request.



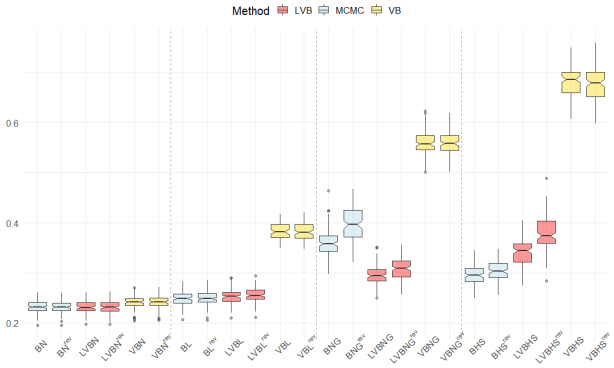
(a) Frobenius norm $d = 30$, moderate sparsity



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



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



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

Figure 4: Top panels report the Frobenius norm of $\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}$ 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 $\boldsymbol{\Theta}$ and the non-null parameters in the estimated matrix $\widehat{\boldsymbol{\Theta}}$. The box charts show the results for $N = 100$ replications, $d = 30$ and different levels of sparsity.

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 $\mathbf{\Theta}$ 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 norma, meaning that the performance of our `VB` estimation method is permutation-invariant irrespective on how sparse may be the matrix of regression coefficients.

5 A empirical study of 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 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

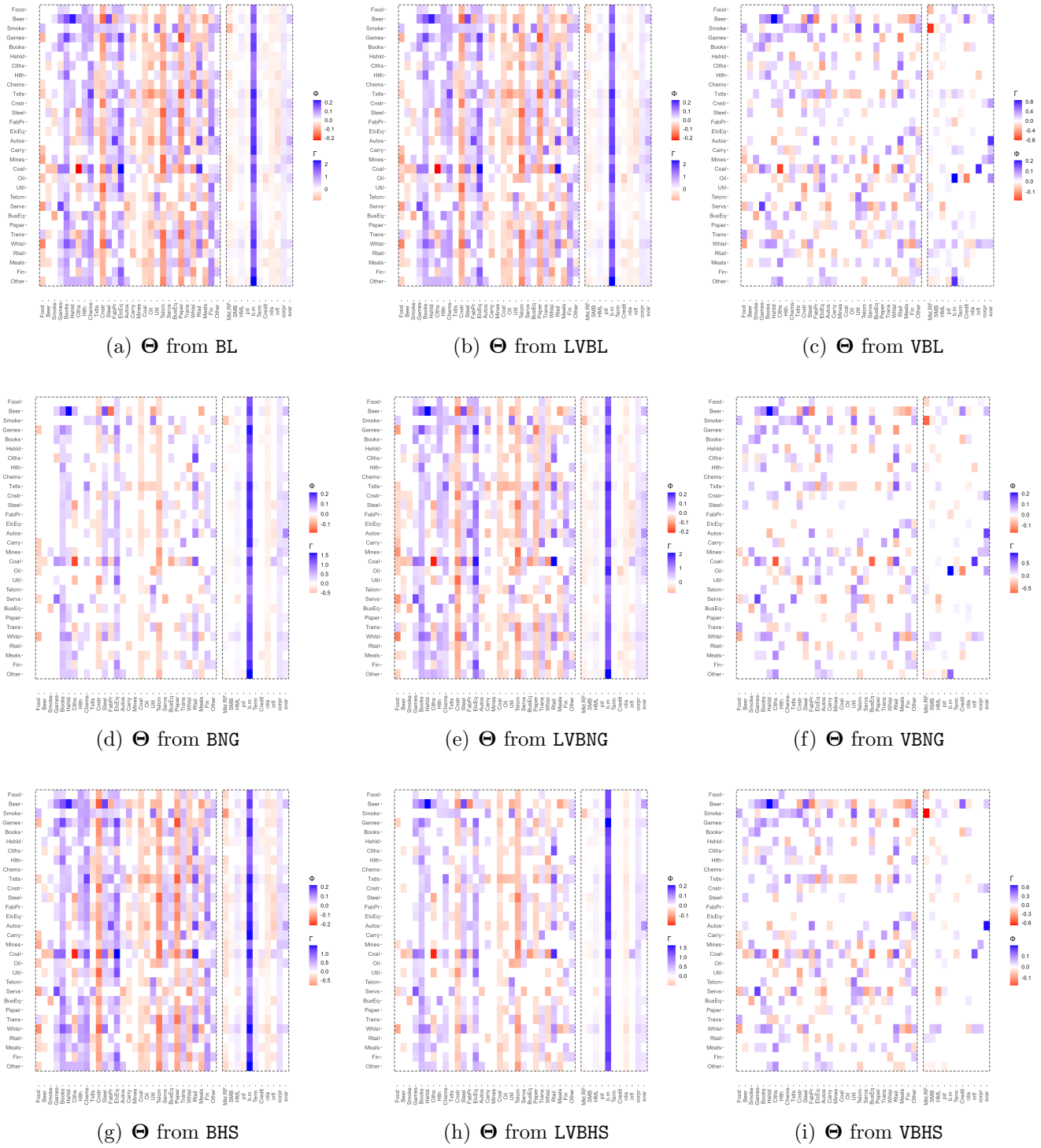
the aggregate price-dividend ratio (`pd`), the term spread (`term`) (difference between the long term yield on government bonds and the T-bill), the credit spread (`credit`) (the BAA-AAA bond yields difference), the monthly change in inflation (`infl`) measured as the log change in the CPI, the aggregate 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.

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 $\mathbf{\Theta}$. Figure 5 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 $\widehat{\mathbf{\Theta}}$ 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 5 is that for both the MCMC and the `LVB` method there are visible differences in the posterior estimates across shrinkage priors. For instance, the $\widehat{\mathbf{\Theta}}$ 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 5 is that under our variational inference method the estimates



(a) $\boldsymbol{\Theta}$ from BL (b) $\boldsymbol{\Theta}$ from LVBL (c) $\boldsymbol{\Theta}$ from VBL

(d) $\boldsymbol{\Theta}$ from BNG (e) $\boldsymbol{\Theta}$ from LVBNG (f) $\boldsymbol{\Theta}$ from VBNG

(g) $\boldsymbol{\Theta}$ from BHS (h) $\boldsymbol{\Theta}$ from LVBHS (i) $\boldsymbol{\Theta}$ from VBHS

Figure 5: Posterior estimates of the regression coefficients $\boldsymbol{\Theta}$ 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).

of $\boldsymbol{\Theta}$ are (1) more sparse compared to both MCMC and `LVB`, and (2) are remarkably similar across different shinkrage priors.

Section E in the supplementary Appendix shows that the same pattern emerges for the 49 industry portfolios (see Figure E.7). 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.

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 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 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 Ω; that is, the forecasts from the univariate models do not depend on any Cholesky parametrization although ignore potential contemporaneous correlations across industry returns.

Table 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., Prob $\left(R^2_{oos} > 0\right)$.

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
Prob $\left(R^2_{oos} > 0\right)$	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
Prob $\left(R^2_{oos} > 0\right)$	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

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, i.e., Prob $(R^2_{oos} > 0)$. Consistent with conventional wisdom that 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%. 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 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 $\mathbf{A} = \mathbf{L}\mathbf{\Theta}$. 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.

5.2.1 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 3 reports the results for the recession periods using the same structure as in Table 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%

Table 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., Prob $\left(R^2_{oos} > 0\right)$.

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
Prob $\left(R^2_{oos} > 0\right)$	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
Prob $\left(R^2_{oos} > 0\right)$	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

(-16%) obtained from the BNG (LVBNG) method.

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 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.

Section E in the supplementary Appendix reports some additional forecasting results based on two aggregations of the mean squared forecasting errors along the lines of Christoffersen and Diebold (1998) and Fisher et al. (2020). More specifically, we computed the aggregated $R^2_{oos,W}$ based on the weighted multivariate mean squared forecast error of model $\mathcal{M}_s$ and the no-predictability benchmark. Figure E.5 and E.6 shows that our variational Bayes method substantially outperforms both MCMC and LVB across priors.

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 $\mathrm{w}_t = \gamma^{-1}\boldsymbol{\Sigma}_{t|t-1}^{-1}[\widehat{\mathbf{y}}_t + \boldsymbol{\sigma}_{t|t-1}^2/2]$, with $\widehat{\mathbf{y}}_t$ the vector of returns forecast at time t, $\boldsymbol{\sigma}_{t|t-1}^2$ a vector containing the diagonal elements of conditional covariance matrix $\boldsymbol{\Sigma}_{t|t-1}$. 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 our paper focuses on the estimation of the regression coefficients under shrinkage priors, and thus modeling time-varying volatility is beyond our scope (see Section 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 4 shows a set of descriptive statistics summarising the cross section of individual industry certainty equivalent returns expressed in annualised percentage. We report both

Table 4: **Certainty equivalent returns**

This table reports a set of descriptive statistics for the differential certainty equivalent return expressed in annualised %. 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. 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, i.e., Prob (cer > 0).

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
Prob (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
Prob (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

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, i.e., Prob (cer > 0).

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

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.

References

D. Avramov. Stock return predictability and asset pricing models. *Review of Financial Studies*, 17(3):699–738, 2004.

A. Bhattacharya, D. Pati, N. S. Pillai, and D. B. Dunson. Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490, 2015.

A. Bhattacharya, A. Chakraborty, and B. K. Mallick. Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103(4):985–991, 2016.

D. Bianchi and K. McAlinn. Divide and conquer: Financial ratios and industry returns predictability. *Available at SSRN*, 2020.

A. Bitto and S. Frühwirth-Schnatter. Achieving shrinkage in a time-varying parameter model framework. *J. Econometrics*, 210(1):75–97, 2019.

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

M. Bognanni. Comment on "large bayesian vector autoregressions with stochastic volatility and non-conjugate priors". *Journal of Econometrics*, 227(2):498–505, 2022.

J. Y. Campbell and S. B. Thompson. Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4):1509–1531, 2007.

J. Y. Campbell and L. M. Viceira. Long-horizon mean-variance analysis: A user guide. *Manuscript, Harvard University, Cambridge, MA*, 2004.

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

A. Carriero, J. Chan, T. E. Clark, and M. Marcellino. 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, 2022.

C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, volume 5, pages 73–80, 16–18 Apr 2009.

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

J. C. Chan and E. Eisenstat. Bayesian model comparison for time-varying parameter vars with stochastic volatility. *Journal of Applied Econometrics*, 33(4):509–532, 2018.

J. C. Chan and X. Yu. Fast and accurate variational inference for large bayesian vars with stochastic volatility. *Journal of Economic Dynamics and Control*, 143:104505, 2022.

P. F. Christoffersen and F. X. Diebold. Cointegration and long-horizon forecasting. *Journal of Business & Economic Statistics*, 16(4):450–456, 1998.

T. E. Clark. Real-time density forecasts from bayesian vector autoregressions with stochastic volatility. *Journal of Business & Economic Statistics*, 29(3):327–341, 2011.

L. Cohen and A. Frazzini. Economic links and predictable returns. *The Journal of Finance*, 63(4):1977–2011, 2008.

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

T. Dangl and M. Halling. Predictive regressions with time-varying coefficients. *Journal of Financial Economics*, 106(1):157–181, 2012.

E. F. Fama and K. R. French. Industry costs of equity. *Journal of financial economics*, 43 (2):153–193, 1997.

E. F. Fama and K. R. French. A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22, 2015.

L. Farmer, L. Schmidt, and A. Timmermann. Pockets of predictability. *Available at SSRN 3152386*, 2019.

W. E. Ferson and C. R. Harvey. The variation of economic risk premiums. *Journal of political economy*, 99(2):385–415, 1991.

W. E. Ferson and C. R. Harvey. Conditioning variables and the cross section of stock returns. *The Journal of Finance*, 54(4):1325–1360, 1999.

W. E. Ferson and R. A. Korajczyk. Do arbitrage pricing models explain the predictability of stock returns? *Journal of Business*, pages 309–349, 1995.

J. D. Fisher, D. Pettenuzzo, C. M. Carvalho, et al. Optimal asset allocation with multivariate bayesian dynamic linear models. *Annals of Applied Statistics*, 14(1):299–338, 2020.

J. Fleming, C. Kirby, and B. Ostdiek. The Economic Value of Volatility Timing. *Journal of Finance*, 56(1):329–352, 2001.

D. Gefang, G. Koop, and A. Poon. Variational Bayesian inference in large Vector Autoregressions with hierarchical shrinkage. *CAMA Working Paper*, (2019-08), Jan. 2019.

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

J. Griffin and P. Brown. Hierarchical shrinkage priors for regression models. *Bayesian Analysis*, 12(1):135–159, 2017.

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

D. Gunawan, R. Kohn, and D. Nott. Variational Approximation of Factor Stochastic Volatility Models. *arXiv e-prints*, art. arXiv:2010.06738, Oct. 2020.

P. R. Hahn and C. M. Carvalho. Decoupling shrinkage and selection in bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448, 2015.

N. Hauzenberger, F. Huber, and L. Onorante. Combining shrinkage and sparsity in conjugate vector autoregressive models. *Journal of Applied Econometrics*, 36(3):304–327, 2021.

S. J. Henkel, J. S. Martin, and F. Nardari. Time-varying short-horizon predictability. *Journal of financial economics*, 99(3):560–580, 2011.

F. Huber, G. Koop, and L. Onorante. Inducing sparsity and shrinkage in time-varying parameter models. *Journal of Business & Economic Statistics*, 39(3):669–683, 2021.

M. Johannes, A. Korteweg, and N. Polson. Sequential Learning, Predictability, and Optimal Portfolio Returns. *Journal of Finance*, 69(2):611–644, 2014.

D. Korobilis. Hierarchical shrinkage priors for dynamic regressions with many predictors. *International Journal of Forecasting*, 29(1):43–59, 2013.

C. Leng, M. N. Tran, and D. Nott. Bayesian adaptive Lasso. *Annals of the Institute of Statistical Mathematics*, 66(2):221–244, sep 2014.

J. Lewellen, S. Nagel, and J. Shanken. A skeptical appraisal of asset pricing tests. *Journal of Financial economics*, 96(2):175–194, 2010.

L. Menzly and O. Ozbas. Market segmentation and cross-predictability of returns. *The Journal of Finance*, 65(4):1555–1580, 2010.

T. P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 362–369, 2001.

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

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

D. Pettenuzzo, A. Timmermann, and R. Valkanov. Forecasting stock returns under economic constraints. *Journal of Financial Economics*, 114(3):517–553, 2014.

N. G. Polson and J. G. Scott. Shrink globally, act locally: sparse Bayesian regularization and prediction. In *Bayesian statistics 9*, pages 501–538. Oxford Univ. Press, Oxford, 2011.

D. E. Rapach, J. K. Strauss, and G. Zhou. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, 23 (2):821–862, 2010.

P. Ray and A. Bhattacharya. Signal adaptive variable selector for the horseshoe prior. *arXiv: Methodology*, 10 2018.

A. J. Rothman, E. Levina, and J. Zhu. A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika*, 97(3):539–550, 2010.

S. C. Smith and A. Timmermann. Break risk. *The Review of Financial Studies*, 34(4): 2045–2100, 2021.

X. Tang, M. Ghosh, X. Xu, and P. Ghosh. Bayesian variable selection and estimation based on global-local shrinkage priors. *Sankhya A*, 80(2):215–246, 2018.

M. P. Wand, J. T. Ormerod, S. A. Padoan, and R. Frührwirth. Mean field variational bayes for elaborate distributions. *Bayesian Analysis*, 6(4):847–900, 2011. ISSN 19360975.

H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476): 1418–1429, 2006.

Supplementary Appendix of:

Variational Bayes inference for large-scale multivariate predictive regressions

This appendix provide 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 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 B.

Result 1. *Assume that* $\mathbf{y}$ *is a* n*-dimensional vector,* $\mathbf{X}$ *a* $p \times n$ *matrix and* $\boldsymbol{\vartheta}$ *a* p*-dimensional 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:*

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

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* $q(\boldsymbol{\vartheta})$*,* $tr(\cdot)$ *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 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* $\boldsymbol{\Theta}\mathbf{A}\boldsymbol{\Theta}^\intercal$ *is equal to:*

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

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

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

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

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

Result 3. *Let* $\boldsymbol{\vartheta}$ *be a* d*-dimesnional gaussian random vector with mean vector* $\boldsymbol{\mu}_\vartheta$ *and variance-covariance matrix* $\boldsymbol{\Sigma}_\vartheta$*. Then 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:*

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

B 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 3 for the prior distributions described in Section 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.

B.1 Normal prior specification

Proposition B.1.1. *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* $j = 1,\dots,d$*:*

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

where

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

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* $\left[\mathbf{K}_{\vartheta,t}\right]_{i,k} = tr\left\{Cov(\boldsymbol{\vartheta}_i, \boldsymbol{\vartheta}_k)\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal\right\}$ *and* $\left[\mathbf{k}_{\vartheta,t}\right]_i = tr\left\{Cov(\boldsymbol{\vartheta}_i, \boldsymbol{\vartheta}_j)\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal\right\}$ *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_{j,t}^2\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_{j,t}^2 + \mathbb{E}_{\vartheta}\left[\boldsymbol{\vartheta}_j\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal\boldsymbol{\vartheta}_j\right] + \overbrace{\mathbb{E}_{\vartheta,\boldsymbol{\beta}_j}\left[\boldsymbol{\beta}_j\mathbf{r}_{j,t}\mathbf{r}_{j,t}^\intercal\boldsymbol{\beta}_j^\intercal\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}_{j,t}^\intercal\right]\mathbb{E}_{\boldsymbol{\beta}_j}\left[\boldsymbol{\beta}_j^\intercal\right]}_{\text{B}} \\
&= y_{j,t}^2 + \boldsymbol{\mu}_{q(\vartheta_j)}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal\boldsymbol{\mu}_{q(\vartheta_j)} + \boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}^\intercal\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}^\intercal \\
&\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}_{q(\mathbf{r}_{j,t})}^\intercal\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}^\intercal \\
&\quad + \operatorname{tr}\left\{\boldsymbol{\Sigma}_{q(\vartheta_j)}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal\right\} + \operatorname{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\} \\
&\quad + \operatorname{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}_{\vartheta,t}\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}^\intercal \\
&= \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 + \operatorname{tr}\left\{\boldsymbol{\Sigma}_{q(\vartheta_j)}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal\right\} + \operatorname{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\} \\
&\quad + \operatorname{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}_{\vartheta,t}\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}^\intercal,
\end{aligned}
$$

where $\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})} = \mathbf{y}_t^j - \boldsymbol{\mu}_{q(\boldsymbol{\Theta}^j)}\mathbf{z}_{t-1}$. The computations involving terms A and B are presented in the following equations. Firs of all, define $\boldsymbol{\beta}_j\mathbf{r}_{j,t}\mathbf{r}_{j,t}^\intercal\boldsymbol{\beta}_j^\intercal = \|\boldsymbol{\beta}_j\mathbf{r}_{j,t}\|_2^2$, then the term A above is equal to:

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

while the term B is:

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

Notice that for the latter derivation we use Results 1 and 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 B.1.1. □

Proposition B.1.2. *The optimal variational density for the parameter $\boldsymbol{\beta}_j$ for $j = 2, \ldots, d$ 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:*

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

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 $\log q^*(\boldsymbol{\beta}_j) \propto \mathbb{E}_{-\boldsymbol{\beta}_j}\left[\ell_j(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}) + \log p(\boldsymbol{\beta}_j)\right]$:

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

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

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

$$\propto -\frac{1}{2}\Bigg\{\boldsymbol{\beta}_j\Bigg(\mu_{q(\nu_j)}\sum_{t=1}^{T}\left(\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}^{\intercal}+\mathbf{K}_{\vartheta,t}\right)+1/\tau\mathbf{I}_{j-1}\Bigg)\boldsymbol{\beta}_j^{\intercal}$$

$$-2\boldsymbol{\beta}_j\mu_{q(\nu_j)}\sum_{t=1}^{T}\left(\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}(y_{j,t}-\boldsymbol{\mu}_{q(\vartheta_j)}\mathbf{z}_{t-1})^{\intercal}+\mathbf{k}_{\vartheta,t}\right)\Bigg\}.$$

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 B.1.2. □

Proposition B.1.3. *The optimal variational density for the parameter $\boldsymbol{\vartheta}$ is equal to a multivariate gaussian $q^*(\boldsymbol{\vartheta})\equiv\mathsf{N}_{d(d+p+1)}(\boldsymbol{\mu}_{q(\vartheta)},\boldsymbol{\Sigma}_{q(\vartheta)})$, where:*

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

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:

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

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:

$$\mathbf{L}\mathbf{y}_t=(\mathbf{L}\otimes\mathbf{z}_{t-1}^{\intercal})\boldsymbol{\vartheta}+\boldsymbol{\varepsilon}_t,\quad \boldsymbol{\varepsilon}_t\sim\mathsf{N}_d(0,\mathbf{V}^{-1}).$$

Recall that a priori $\boldsymbol{\vartheta}\sim\mathsf{N}_{d(d+p+1)}(\mathbf{0},\upsilon\mathbf{I}_{d(d+p+1)})$. Compute the optimal variational density

for the parameter $\boldsymbol{\vartheta}$ as $\log q^*(\boldsymbol{\vartheta}) \propto \mathbb{E}_{-\vartheta}\left[\ell(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}) + \log p(\boldsymbol{\vartheta})\right]$:

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

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:

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

hence

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

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

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

Take the exponential of the $\log q^*(\boldsymbol{\vartheta})$ derived above and notice that it coincides with the kernel of a gaussian random variable $\mathsf{N}_{d(d+p+1)}(\boldsymbol{\mu}_{q(\vartheta)}, \boldsymbol{\Sigma}_{q(\vartheta)})$, as defined in Proposition B.1.3. □

Proposition B.1.4. *The optimal variational density for the parameter $\boldsymbol{\vartheta}_j$ is equal to a multivariate gaussian $q^*(\boldsymbol{\vartheta}_j) \equiv \mathsf{N}_{d+p+1}(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$, where, for each row $j = 1, \ldots, d$ of $\boldsymbol{\Theta}$:*

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

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 B.1.3, 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 $\mathsf{N}_{d+p+1}(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$, as defined in Proposition B.1.4. □

Proposition B.1.5. *The variational lower bound for the non-sparse multivariate regression model can be derived analytically and it is equal to:*

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

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:

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

Similarly 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:

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

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 $A = \ell_j(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}) + \log \underline{p}(\mathbf{y}; \nu_j)$:

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

where we exploit the definitions of $\mathbb{E}_q \left[\varepsilon_{j,t}^2\right], a_{q(\nu_j)}, b_{q(\nu_j)}$ given in Proposition B.1.1. The second term to compute is equal to:

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

where $\mu_{q(\beta_{j,k}^2)} = \mu_{q(\beta_{j,k})}^2 + \sigma_{q(\beta_{j,k})}^2$ and $\sigma_{q(\beta_{j,k})}^2$ denotes the k-th element on the diagonal of

$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}$. To conclude, we compute the last term:

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

Put together the terms A, B, C as in (B.6) and notice that the variational lower bound here computed coincides with the one presented in Proposition B.1.5. □

The moments of the optimal variational densities are updated at each iteration of the Algorithm 1 and the convergence is assessed by checking the variation both in the lower bound and the parameters.

Algorithm 1: MFVB with non-informative prior.

Initialize: $q^*(\boldsymbol{\xi})$, Δ_ξ, Δ_{ELBO}
while $\left(\widehat{\Delta}_{ELBO} > \Delta_{ELBO}\right) \vee \left(\widehat{\Delta}_\xi > \Delta_\xi\right)$ **do**
  Update $q^*(\nu_1)$ as in (B.1);
  for $j = 2, \ldots, d$ **do**
    Update $q^*(\nu_j)$ and $q^*(\boldsymbol{\beta}_j)$ as in (B.1)-(B.2);
  end
  Update $q^*(\boldsymbol{\vartheta})$ as in (B.3) or (B.4);
  Compute $\log \underline{p}(\mathbf{y}; q)$ as in (B.5);
  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

B.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 B.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 B.2.1. *The joint optimal variational density for the parameter $\boldsymbol{\vartheta}$ is equal to*

$q^*(\boldsymbol{\vartheta}) \equiv \mathsf{N}_{d(d+p+1)}(\boldsymbol{\mu}_{q(\vartheta)}, \boldsymbol{\Sigma}_{q(\vartheta)})$, *where:*

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

where diag$(\boldsymbol{\mu}_{q(1/\upsilon)})$ *is a diagonal matrix where* $\boldsymbol{\mu}_{q(1/\upsilon)} = (\mu_{q(1/\upsilon_{1,1})}, \mu_{q(1/\upsilon_{1,2})}, \ldots, \mu_{q(1/\upsilon_{d,d+p+1})})$.

Under the row-independence assumption, the optimal variational density for the parameter $\boldsymbol{\vartheta}_j$ *is equal to* $q^*(\boldsymbol{\vartheta}_j) \equiv \mathsf{N}_{d+p+1}(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$, *where:*

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

where diag$(\boldsymbol{\mu}_{q(1/\upsilon_j)})$ *is a diagonal matrix where* $\boldsymbol{\mu}_{q(1/\upsilon_j)} = (\mu_{q(1/\upsilon_{j,1})}, \mu_{q(1/\upsilon_{j,2})}, \ldots, \mu_{q(1/\upsilon_{j,d+p+1})})$. Hereafter we describe the optimal densities for the parameters used in hierarchical specification of the prior here assumed.

Proposition B.2.2. *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* $j = 1, \ldots, d$ *and* $k = 1, \ldots, d+p+1$*:*

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

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 B.2.2. □

Proposition B.2.3. *The optimal density for the latent parameter $\lambda^2_{j,k}$ for $j = 1, \ldots, d$ and $k = 1, \ldots, d + p + 1$ 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:*

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

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 B.2.3. □

Proposition B.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:

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

Proof. As we did in (B.6) for Proposition B.1.5, the lower bound can be divided into terms referring to each parameter:

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

where A is equal to (B.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 B.1.5 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:

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

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

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

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 $\lambda_{j,k}^2$, for $j = 1, \ldots, d$ and $k = 1, \ldots, d + p + 1$, is equal to:

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

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. □

The moments of the optimal variational densities are updated at each iteration of the Algorithm 2 and the convergence is assessed by checking the variation both in the lower bound and the parameters.

B.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 B.1. The optimal density $q^*(\boldsymbol{\vartheta})$ has the same structure as the one computed in Proposition (B.2.1) for the lasso prior.
Hereafter we describe the optimal densities for the parameters used in hierarchical specification of the normal-gamma prior.

Proposition B.3.1. *The optimal density for the prior variance $\upsilon_{j,k}$ is equal to a generalized*

Algorithm 2: MFVB with Bayesian adaptive lasso prior.

Initialize: $q^*(\boldsymbol{\xi})$, Δ_ξ, Δ_{ELBO}
while $\left(\widehat{\Delta}_{ELBO} > \Delta_{ELBO}\right) \vee \left(\widehat{\Delta}_\xi > \Delta_\xi\right)$ **do**
  Update $q^*(\nu_1)$ as in (B.1);
  for $j = 2, \ldots, d$ **do**
    Update $q^*(\nu_j)$ and $q^*(\boldsymbol{\beta}_j)$ as in (B.1)-(B.2);
  end
  Update $q^*(\boldsymbol{\vartheta})$ as in (B.7) or (B.8);
  for $j = 1, \ldots, d$ **do**
    for $k = 1, \ldots, d + p + 1$ **do**
      Update $q^*(\upsilon_{j,k})$, $q^*(\lambda^2_{j,k})$ as in (B.9)-(B.10);
    end
  end
  Compute $\log \underline{p}\,(\mathbf{y}; q)$ as in (B.11);
  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

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* $j = 1, \ldots, d$ *and* $k = 1, \ldots, d + p + 1$*:*

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

Moreover, it is useful to know that

$$\mu_{q(\upsilon_{j,k})} = \frac{\sqrt{b_{q(\upsilon_{j,k})}} K_{\zeta_{q(\upsilon_{j,k})}+1}\left(\sqrt{a_{q(\upsilon_{j,k})} b_{q(\upsilon_{j,k})}}\right)}{\sqrt{a_{q(\upsilon_{j,k})}} K_{\zeta_{q(\upsilon_{j,k})}}\left(\sqrt{a_{q(\upsilon_{j,k})} b_{q(\upsilon_{j,k})}}\right)},$$

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

$$\mu_{q(\log \upsilon_{j,k})} = \log \frac{\sqrt{b_{q(\upsilon_{j,k})}}}{\sqrt{a_{q(\upsilon_{j,k})}}} + \frac{\partial}{\partial \zeta_{q(\upsilon_{j,k})}} \log K_{\zeta_{q(\upsilon_{j,k})}}\left(\sqrt{a_{q(\upsilon_{j,k})} b_{q(\upsilon_{j,k})}}\right),$$

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 $\log q^*(\upsilon_{j,k}) \propto \mathbb{E}_{-\upsilon_{j,k}}[\log p(\vartheta_{j,k}) + \log p(\upsilon_{j,k})]$:

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

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 B.3.1. □

Proposition B.3.2. *The optimal density for the latent parameter $\lambda_{j,k}$ for $j = 1, \ldots, d$ and $k = 1, \ldots, d + p + 1$ is equal to a $q^*(\lambda_{j,k}) \equiv \mathsf{Ga}(a_{q(\lambda_{j,k})}, b_{q(\lambda_{j,k})})$, where:*

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

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 $\log q^*(\lambda_{j,k}) \propto \mathbb{E}_{-\lambda_{j,k}}[\log p(\upsilon_{j,k}) + \log p(\lambda_{j,k})]$:

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

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 B.3.2. □

Proposition B.3.3. *The optimal density for the latent parameter η_j for $j = 1, \ldots, d$ is equal to:*

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

where $\log h(\eta_j) = (d+p+1)(\eta_j \log \eta_j - \log \Gamma(\eta_j))$ *and*

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

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 $\log q^*(\eta_j) \propto \mathbb{E}_{-\eta_j}\left[\sum_{k=1}^{d+p+1} \log p(\upsilon_{j,k}) + \log p(\eta_j) \right]$:

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

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 (B.16) is integrable and thus we can compute the normalizing constant and its expectation. □

Proposition B.3.4. *The variational lower bound for the multivariate regression model with*

adaptive normal-gamma prior can be derived analytically and it is equal to:

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

Proof. As we did in (B.6) for Proposition B.1.5, the lower bound can be divided into terms referring to each parameter:

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

where A is equal to (B.2). Our strategy will be to evaluate each piece in the latter separately and then put the results together. Consider the piece B:

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

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 $k = 1, \dots, d + p + 1$, 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:

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

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. □

The moments of the optimal variational densities are updated at each iteration of the Algorithm 3 and the convergence is assessed by checking the variation both in the lower bound and the parameters.

B.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 B.1. The changes in the optimal densities $q^*(\boldsymbol{\vartheta})$ are stated in the next proposition.

Proposition B.4.1. *The joint optimal variational density for the parameter $\boldsymbol{\vartheta}$ is equal to $q^*(\boldsymbol{\vartheta}) \equiv \mathsf{N}_{d(d+p+1)}(\boldsymbol{\mu}_{q(\vartheta)}, \boldsymbol{\Sigma}_{q(\vartheta)})$, where:*

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

Algorithm 3: MFVB with adaptive normal-gamma prior.

Initialize: $q^*(\boldsymbol{\xi})$, Δ_ξ, Δ_{ELBO}

while $\left(\widehat{\Delta}_{ELBO} > \Delta_{ELBO}\right) \vee \left(\widehat{\Delta}_\xi > \Delta_\xi\right)$ **do**

  Update $q^*(\nu_1)$ as in (B.1);

  for $j = 2, \ldots, d$ **do**

    Update $q^*(\nu_j)$ and $q^*(\boldsymbol{\beta}_j)$ as in (B.1)-(B.2);

  end

  Update $q^*(\boldsymbol{\vartheta})$ as in (B.7) or (B.8);

  for $j = 1, \ldots, d$ **do**

    for $k = 1, \ldots, d + p + 1$ **do**

      Update $q^*(\upsilon_{j,k})$, $q^*(\lambda_{j,k})$ as in (B.12)-(B.13);

    end

    Update $q^*(\eta_j)$ as in (B.15);

  end

  Compute $\log \underline{p}\,(\mathbf{y}; q)$ as in (B.17);

  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

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

Under the row-independence assumption, the optimal variational density for the parameter $\boldsymbol{\vartheta}_j$ *is equal to* $q^*(\boldsymbol{\vartheta}_j) \equiv \mathsf{N}_{d+p+1}(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$, *where:*

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

where $\text{diag}(\boldsymbol{\mu}_{q(1/\upsilon_j^2)})$ *is a diagonal matrix and* $\boldsymbol{\mu}_{q(1/\upsilon_j^2)} = (\mu_{q(1/\upsilon^2_{j,1})}, \mu_{q(1/\upsilon^2_{j,2})}, \ldots, \mu_{q(1/\upsilon^2_{j,d+p+1})})$.

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

Proposition B.4.2. *The optimal density for the prior local variance* $\upsilon^2_{j,k}$ *is equal to an inverse gamma distribution* $q^*(\upsilon^2_{j,k}) \equiv \mathsf{InvGa}(1, b_{q(\upsilon^2_{j,k})})$, *where, for* $j = 1, \ldots, d$ *and* $k = 1, \ldots, d + p + 1$*:*

$$b_{q(\upsilon^2_{j,k})} = \mu_{q(1/\lambda_{j,k})} + \frac{1}{2}\mu_{q(\vartheta^2_{j,k})}\mu_{q(1/\gamma^2)}. \tag{B.21}$$

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

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

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

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

Take the exponential and notice that the latter is the kernel of an inverse gamma random variable $\mathsf{InvGa}(1, b_{q(\upsilon^2_{j,k})})$, as defined in Proposition B.4.2. □

Proposition B.4.3. *The optimal density for the prior global variance γ^2 is equal to an inverse gamma distribution $q^*(\gamma^2) \equiv \mathsf{InvGa}(a_{q(\gamma^2)}, b_{q(\gamma^2)})$, where:*

$$a_{q(\gamma^2)} = \frac{d(d+p+1)+1}{2}, \quad b_{q(\gamma^2)} = \mu_{q(1/\eta)} + \frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{d+p+1}\mu_{q(1/\upsilon^2_{j,k})}\mu_{q(\vartheta^2_{j,k})}. \tag{B.22}$$

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

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

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

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

Take the exponential and notice that the latter is the kernel of an inverse gamma random variable $\mathsf{InvGa}(a_{q(\gamma^2)}, b_{q(\gamma^2)})$, as defined in Proposition B.4.3. □

Proposition B.4.4. *The optimal density for the latent parameter $\lambda_{j,k}$ is equal to an inverse gamma distribution $q^*(\lambda_{j,k}) \equiv \mathsf{InvGa}(1, b_{q(\lambda_{j,k})})$, where, for $j = 1, \ldots, d$ and $k = 1, \ldots, d+$*

$p+1$:

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

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

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

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

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

Take the exponential and notice that the latter is the kernel of an inverse gamma random variable $\mathsf{InvGa}(1, b_{q(\lambda_{j,k})})$, as defined in Proposition B.4.4. □

Proposition B.4.5. *The optimal density for the latent parameter η is equal to an inverse gamma distribution $q^*(\eta) \equiv \mathsf{InvGa}(1, b_{q(\eta)})$, where:*

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

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

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

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

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

Take the exponential and notice that the latter is the kernel of an inverse gamma random variable $\mathsf{InvGa}(1, b_{q(\eta)})$, as defined in Proposition B.4.5. □

Proposition B.4.6. *The variational lower bound for the multivariate regression model with*

Horseshoe prior can be derived analytically and it is equal to:

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

Proof. As we did in (B.6) for Proposition B.1.5, the lower bound can be divided into terms referring to each parameter:

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

where A is similar to (B.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 B.1.5. Hence, we have that:

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

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. In particular, for $j = 1,\dots,d$ and $k = 1,\dots,d+p+1$:

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

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. □

The moments of the optimal variational densities are updated at each iteration of the Algorithm 4 and the convergence is assessed by checking the variation both in the lower bound and the parameters.

Algorithm 4: MFVB with Horseshoe prior.

Initialize: $q^*(\boldsymbol{\xi})$, Δ_ξ, Δ_{ELBO}

while $\left(\widehat{\Delta}_{ELBO} > \Delta_{ELBO}\right) \vee \left(\widehat{\Delta}_\xi > \Delta_\xi\right)$ **do**

  Update $q^*(\nu_1)$ as in (B.1);

  for $j = 2, \ldots, d$ **do**

    Update $q^*(\nu_j)$ and $q^*(\boldsymbol{\beta}_j)$ as in (B.1)-(B.2);

  end

  Update $q^*(\boldsymbol{\vartheta})$ as in (B.19) or (B.20) ;

  for $j = 1, \ldots, d$ **do**

    for $k = 1, \ldots, d + p + 1$ **do**

      Update $q^*(\upsilon^2_{j,k})$, $q^*(\lambda_{j,k})$ as in (B.21)-(B.23);

    end

  end

  Update $q^*(\gamma^2)$, $q^*(\eta)$ as in (B.22)-(B.24);

  Compute $\log \underline{p}\left(\mathbf{y}; q\right)$ as in (B.25);

  Compute $\widehat{\Delta}_{\text{ELBO}} = \log \underline{p}\left(\mathbf{y}; q\right)^{(\text{iter})} - \log \underline{p}\left(\mathbf{y}; q\right)^{(\text{iter}-1)}$;

  Compute $\widehat{\Delta}_\xi = q^*(\boldsymbol{\xi})^{(\text{iter})} - q^*(\boldsymbol{\xi})^{(\text{iter}-1)}$;

end

B.5 Inference on the precision matrix

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

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

$$\psi(\delta, \mathbf{H}) \propto \frac{d\delta}{2}\log 2 + \frac{\delta}{2}\log|\mathbf{H}| + \log\Gamma_d(\delta/2) - \frac{\delta}{2}\mathbb{E}_p\left[\log|\boldsymbol{\Omega}|\right] + \frac{1}{2}\text{tr}\left\{\mathbf{H}^{-1}\mathbb{E}_p\left[\boldsymbol{\Omega}\right]\right\}. \tag{B.28}$$

Notice that $\mathbb{E}_p\left[\log|\boldsymbol{\Omega}|\right] = \mathbb{E}_{q(V)}\left[\log|\mathbf{V}|\right] = \sum_{j=1}^d \mu_{q(\log\nu_j)}$ and $\mathbb{E}_p\left[\boldsymbol{\Omega}\right] = \mathbb{E}_{q(L),q(V)}\left[\mathbf{L}^\intercal\mathbf{V}\mathbf{L}\right]$ are available as byproduct of the mean-field Variational Bayes algorithm. Differentiating (B.28) with respect to the scaling matrix $\mathbf{H}$, solving $\partial\psi(\delta, \mathbf{H})/\partial\mathbf{H} = 0$ provides $\widehat{\mathbf{H}}_\delta = \delta^{-1}\mathbb{E}_p\left[\boldsymbol{\Omega}\right]$ 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. □

C Variational predictive density

Recall that the variational predictive posterior can be computed as:

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

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 B.5) in the following way:

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

where

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

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

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 (C.1) becomes

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

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 (C.4) with the closest multivariate normal distribution in terms of KL divergence:

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

where, in particular,

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

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

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

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 *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 (C.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:

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

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:

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

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 $\boldsymbol{\mu}_{pred}=\boldsymbol{\Sigma}_{pred}\mathbf{R}\mathbf{Z}_t\tilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}_{q(\vartheta)}^{-1}\boldsymbol{\mu}_{q(\vartheta)}$. 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 C.6 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 $\mathbf{\Omega}$. 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.

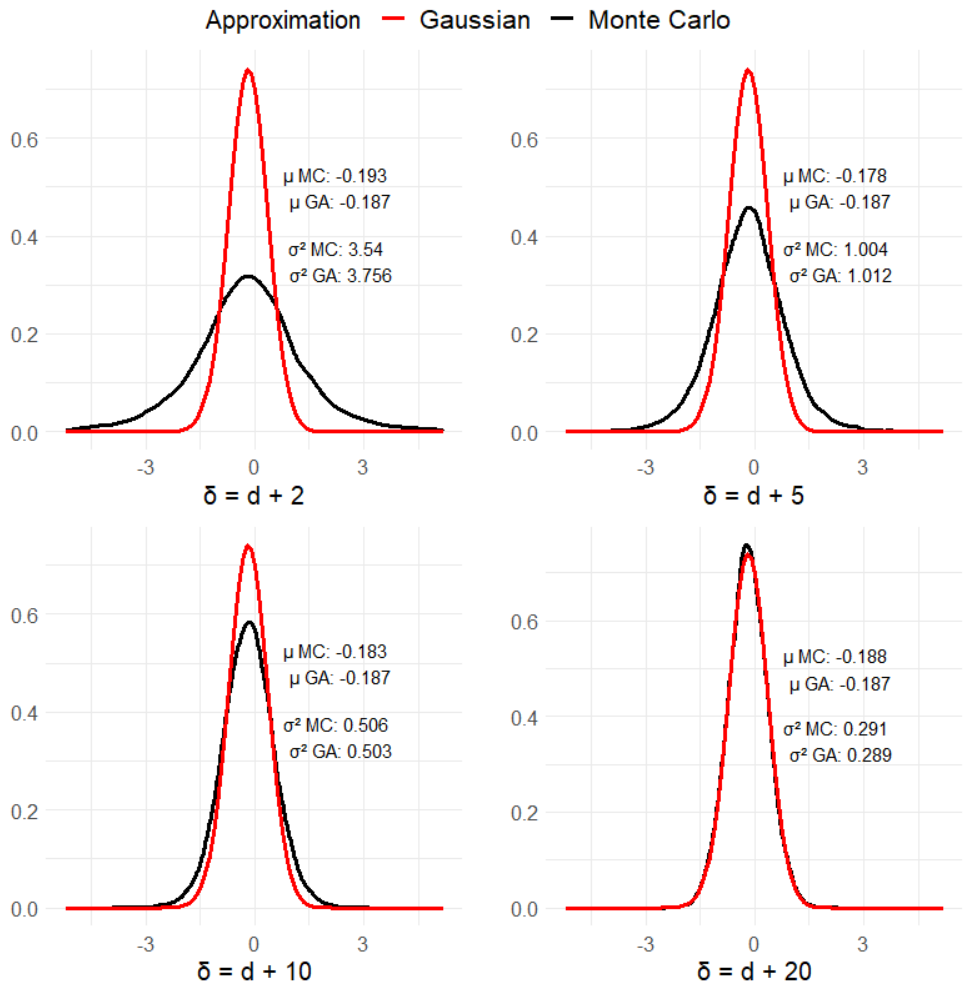


Figure C.6: Second-order approximation of the predictive density.

D Simulation details and additional results

In this section we report additional details and results on the simulation study we highlighted in Section 4. 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 $\mathbf{\Theta}$ 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 D.1 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.

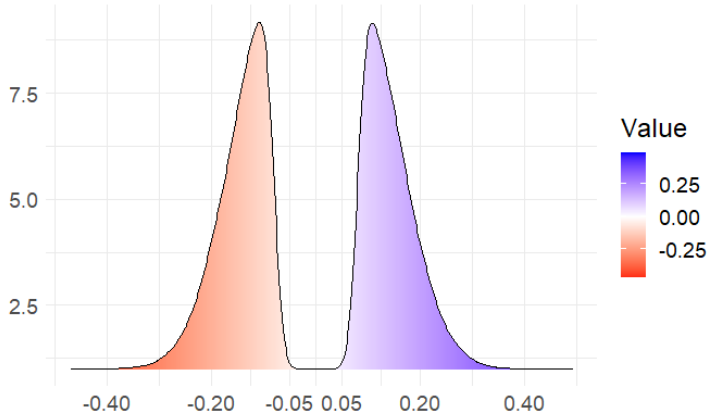
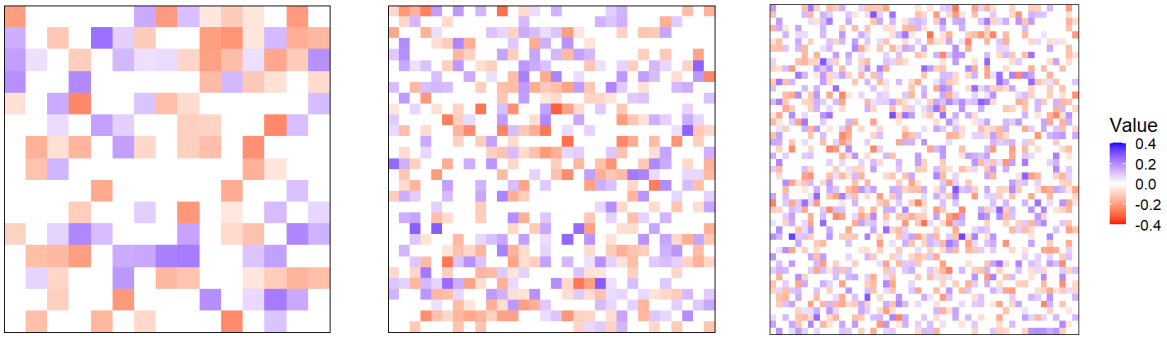


Figure D.1: Distribution of non-zero parameters in the true regression matrix. This figure plots the distribution from which we sample the non-zero entries of the regression matrices used to generate the data for the simulation study.

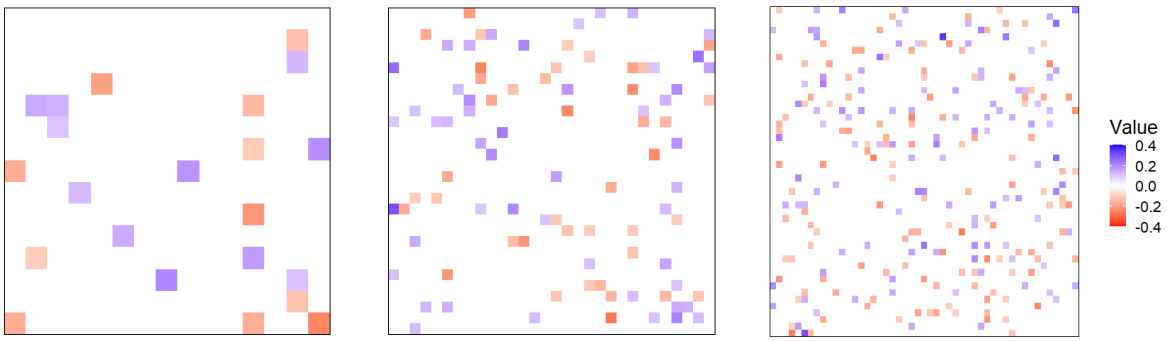
Figure D.2 shows examples of the true regression matrixes for different dimensions $d = 15, 30, 49$ and for two alternative levels of sparsity $s = 0.5, 0.9$, that is 50% and 90% of the entries in the matrix $\mathbf{\Theta}$ are set to zero.

D.1 Additional simulation results

We now show some of the additional results on smaller dimensional simulation cases. Figure D.3 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.

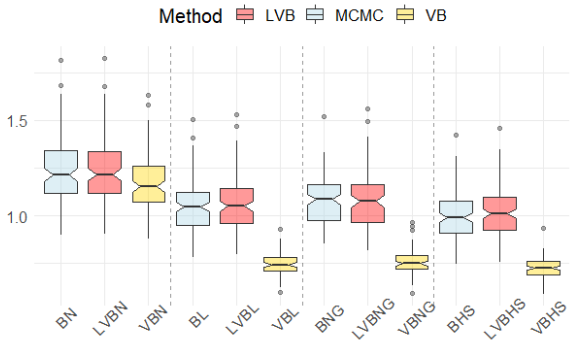


(a) $d = 15$ moderate sparsity (b) $d = 30$ moderate sparsity (c) $d = 49$ moderate sparsity

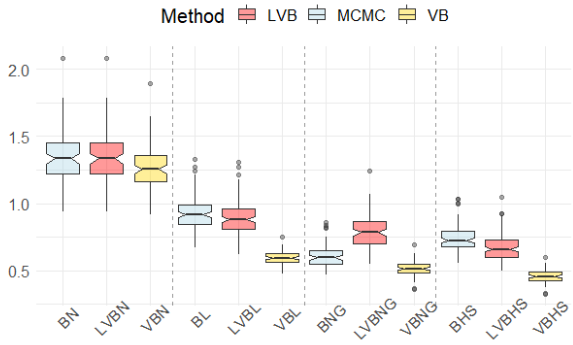


(d) $d = 15$ high sparsity (e) $d = 30$ high sparsity (f) $d = 49$ high sparsity

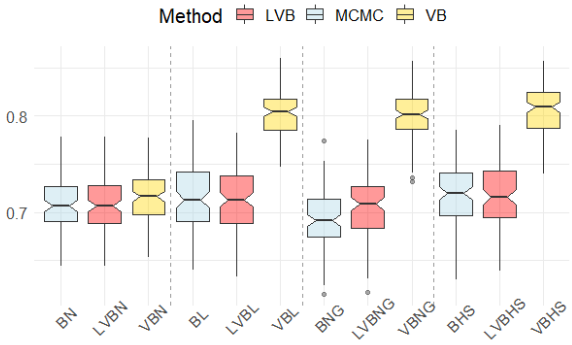
Figure D.2: True regression matrices for the simulation study. This figure plots the regression matrices used in the simulation study. We assume both moderate level of sparsity (top panels, 50% of true zeros) and high level of sparsity (bottom panels, 90% of true zeros).



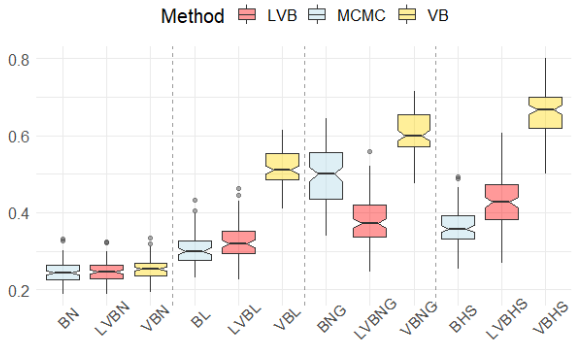
(a) Frobenius norm $d = 15$, moderate sparsity



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



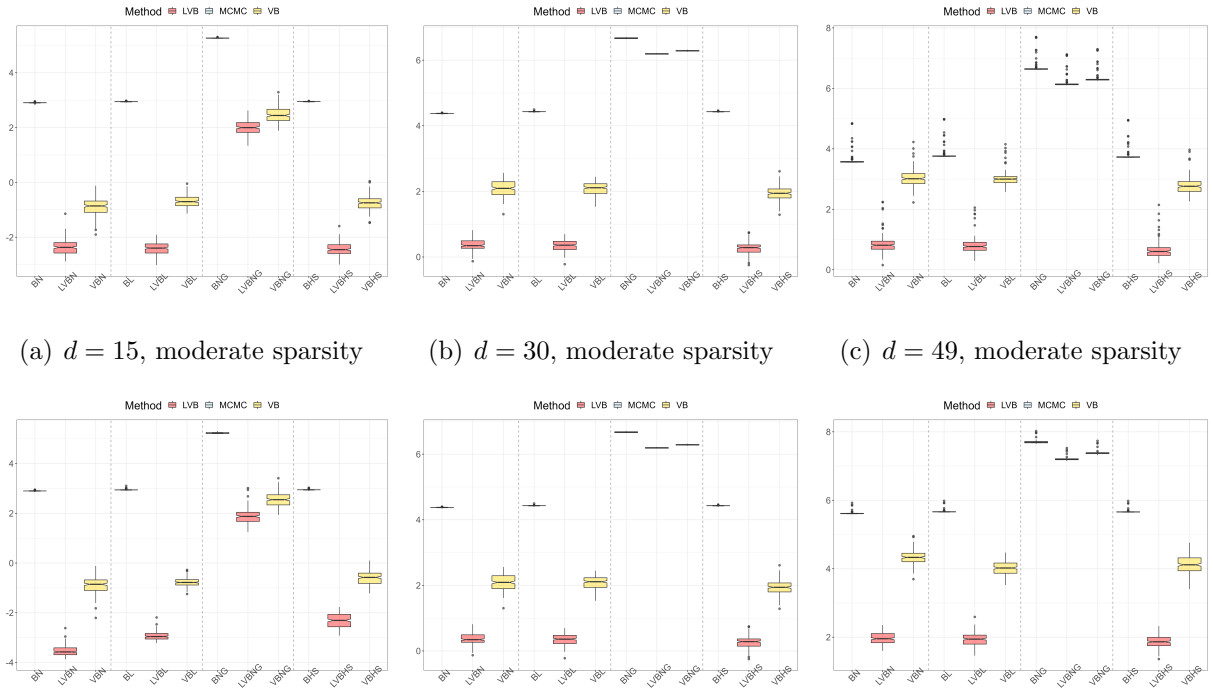
(c) F1-score norm $d = 15$, moderate sparsity



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

Figure D.3: Top panels report the Frobenius norm of $\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}$ 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 $\boldsymbol{\Theta}$ and the non-null parameters in the estimated matrix $\widehat{\boldsymbol{\Theta}}$. The box charts show the results for $N = 100$ replications, $d = 15$ and different levels of sparsity.

Another advantage of the variational methods is the computational time (see Figure D.4). For example, in dimension $d = 49$, the `VB` algortihms 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 step in the `MCMC` approach which translates in a numerical integration in the `VB` algorithm.



(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 D.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.

E Additional empirical results

E.1 Computational cost of the recursive forecasts

The faster computation turns out to be key within the context of recursive forecasting. Table 5 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

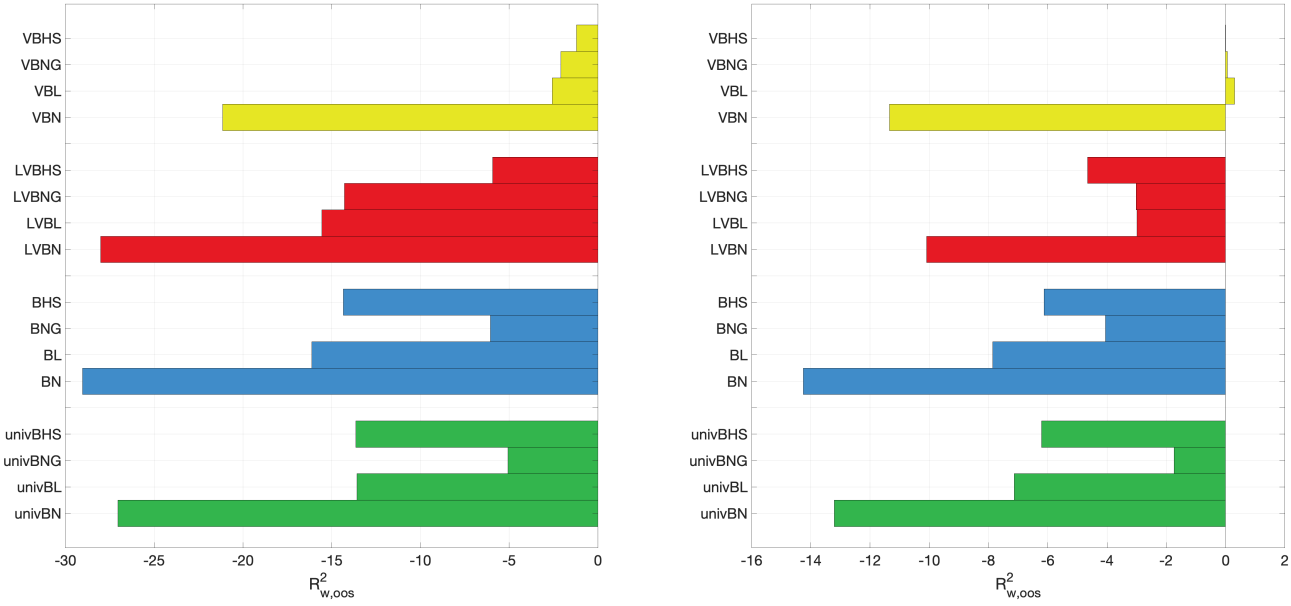
Table 5: Computational time in hours to perform the empirical analysis.

E.2 Aggregate out-of-sample $R^2_{W,oos}$

We aggregate the individual forecasting errors for each industry portfolios as follows,

$$R^2_{W,oos} = 1 - \frac{\sum_{t_0=2}^{T} \sum_{i=1}^{n} \mathrm{w}_{it} \widehat{e}^2_{it}}{\sum_{t_0=2}^{T} \sum_{i=1}^{n} \mathrm{w}_{it} \overline{e}^2_{it}},$$

where $\widehat{e}_{it} = y_{it} - \widehat{y}_{it}(\mathcal{M}_s)$ the forecasting error from the $\mathcal{M}_s$ model and $\overline{e}_{it} = y_{it} - \overline{y}_{it}$ the forecasting error associated to the naive forecast $\overline{y}_{it}$. Figure E.5 represents the results for an aggregation where the weights w_{it} represent the inverse of the average market capitalization (size) of a firm within a given industry.



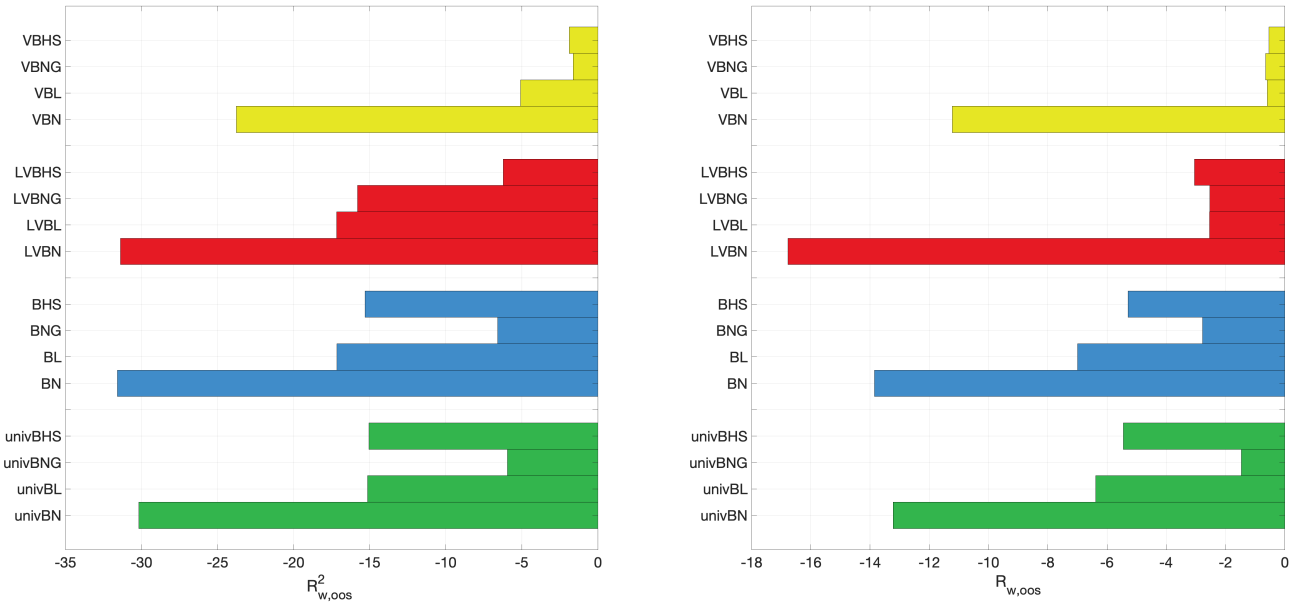
(a) $R^2_{W,oos}$ across 49 industries

(b) $R^2_{W,oos}$ across 30 industries

Figure E.5: Weighted out-of-sample R^2_{oos}. This figure plots the out-of-sample $R^2_{W,oos}$ where the aggregation is based on the average market capitalization of firms within a given industry. The left (right) panel shows the results for the cross section of 49 (30) industry portfolios.

By weighting the forecasting errors by market capitalization there will be less (more) penalty for the forecast errors of a small (large) asset, which carries less (more) relevance in the dynamics of the market portfolio, but also less reward when accurate.

The naive rolling mean forecast represents a highly challenging benchmark to beat, once individual industry forecasts are aggregated. However, our `VB` approach outperforms both the MCMC and the `LVB` approaches irrespective of the hierarchical shrinkage prior specification. This is more evident for the 30 industry classification, where the $R^2_{W,oos}$ for the `VB` approach is slightly positive, whereas the `LVB`, which ranks second, is as low as -3%. At a more general level, a sparse multivariate regression model outperforms a model with a normal prior, as indicated by a large and negative weighted $R^2_{W,oos}$ obtained by the latter.



(a) $R^2_{W,oos}$ across 49 industries

(b) $R^2_{W,oos}$ across 30 industries

Figure E.6: Weighted out-of-sample R^2_{oos}. This figure plots the out-of-sample $R^2_{W,oos}$ where the aggregation is based on the average market capitalization of firms within a given industry. The left (right) panel shows the results for the cross section of 49 (30) industry portfolios.

Figure E.6 reports the results for a different aggregation scheme where the wights w_{it} in $R^2_{W,oos}$ represents the inverse of the sample volatility of the asset. In this respect, there will be less (more) penalty for the forecast errors of an asset with large (small) returns volatility, but also more reward when the forecast is accurate.

Similar to the market cap aggregation, the naive rolling mean forecast turns out to be a highly challenging benchmark to beat. However, our `VB` approach outperforms both the

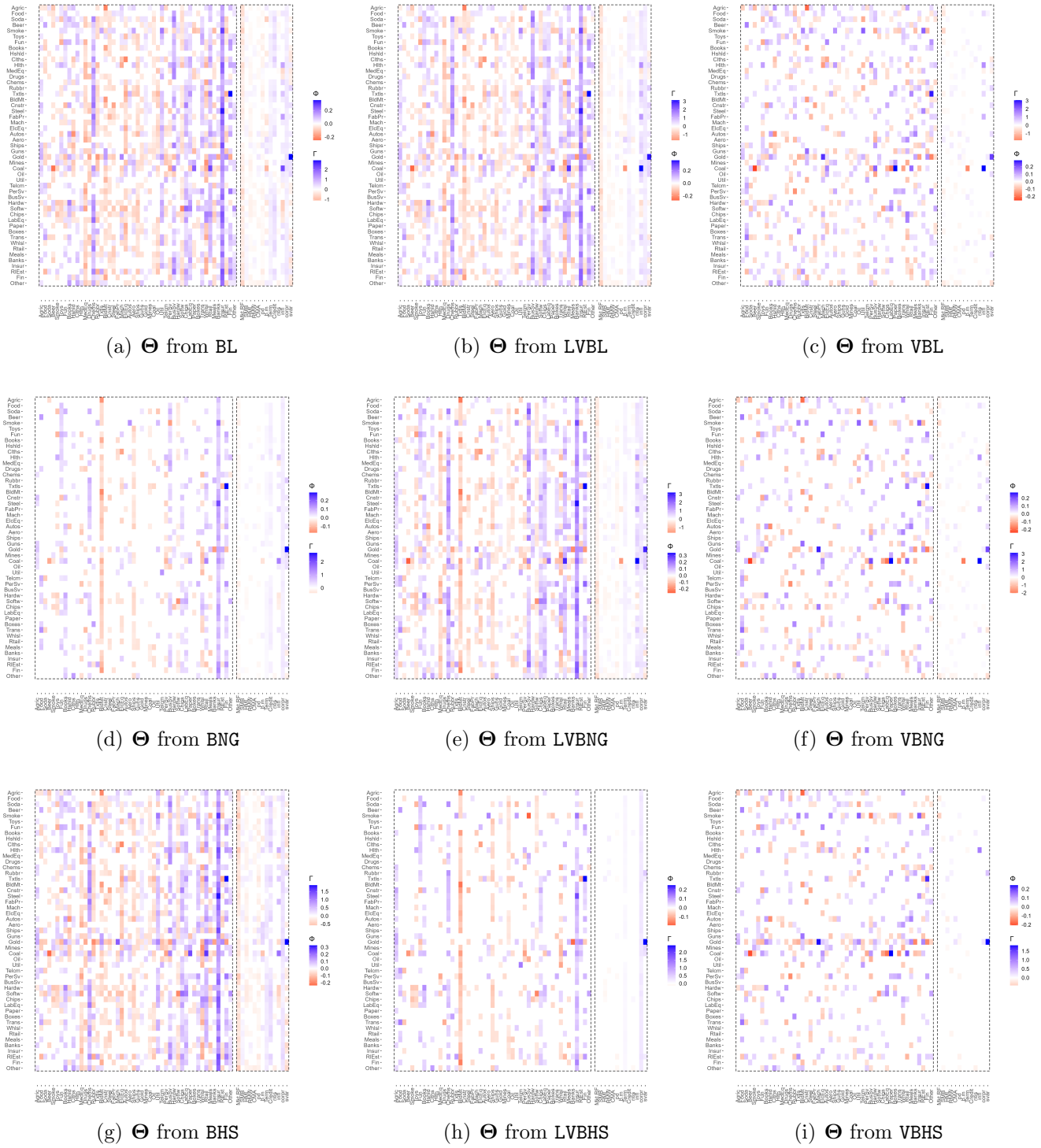
MCMC and the LVB estimation approaches, irrespective of the shrinkage prior specification. This is more evident for the 30 industry classification, where the $R^2_{W,oos}$ for the VB approach is essentially zero, whereas the LVB, which ranks second, has $R^2_{W,oos}$ as low as -3%.

E.3 In-sample estimates for 49 industries

Figure E.7 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 E.7 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 E.7 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 shinkrage priors.



(a) $\boldsymbol{\Theta}$ from BL (b) $\boldsymbol{\Theta}$ from LVBL (c) $\boldsymbol{\Theta}$ from VBL

(d) $\boldsymbol{\Theta}$ from BNG (e) $\boldsymbol{\Theta}$ from LVBNG (f) $\boldsymbol{\Theta}$ from VBNG

(g) $\boldsymbol{\Theta}$ from BHS (h) $\boldsymbol{\Theta}$ from LVBHS (i) $\boldsymbol{\Theta}$ from VBHS

Figure E.7: Posterior estimates of the regression coefficients $\boldsymbol{\Theta}$ 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).