Variational inference for large Bayesian vector autoregressions*

Mauro Bernardi† Daniele Bianchi‡ Nicolas Bianco§

Abstract

We propose a novel variational Bayes approach to estimate high-dimensional vector autoregression (VAR) models with hierarchical shrinkage priors. Our approach does not rely on a conventional structural VAR representation of the parameter space for posterior inference. Instead, we elicit hierarchical shrinkage priors directly on the matrix of regression coefficients so that (1) the prior structure directly maps into posterior inference on the reduced-form transition matrix, and (2) posterior estimates are more robust to variables permutation. An extensive simulation study provides evidence that our approach compares favourably against existing linear and non-linear Markov Chain Monte Carlo and variational Bayes methods. We investigate both the statistical and economic value of the forecasts from our variational inference approach within the context of a mean-variance investor allocating her wealth 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. The results hold across different hierarchical shrinkage priors and model dimensions.

Keywords: Bayesian methods, variational inference, hierarchical shrinkage prior, high-dimensional models, vector autoregressions, industry 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, the 2nd Workshop on Dimensionality Reduction and Inference in High-Dimensional Time Series at Maastricht University, and the 2023 Summer Forum Workshop on Macroeconomics and Policy Evaluation at the Barcelona School of Economics for their helpful comments and suggestions. This research has been partially funded by the BERN_BIRD2222_01 - BIRD 2022 grant of the University of Padua. A previous version of this paper was circulating with the title "Variational Bayes inference for large-scale multivariate predictive regressions".

†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 Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain. Email: nicolas.bianco@upf.edu Web: whitenoise8.github.io

1 Introduction

Hierarchical shrinkage priors have been shown to represent an effective regularization technique when estimating large vector autoregression (VAR) models. The use of these priors often relies on a Cholesky decomposition of the residuals covariance matrix so that a large system of equations is reduced to a sequence of univariate regressions. This allows for more efficient computations as priors can be elicited on the structural VAR representation implied by the Cholesky factorization and posterior inference is carried out equation-by-equation.

Such a conventional approach has two important implications for posterior inference: first, priors are not order-invariant, meaning that posterior inference is sensitive to permutations of the endogenous variables for a given prior specification. This is particularly relevant in high dimensions whereby logical orders of the endogenous variables might be unclear or a full search among all possible ordering combinations might be unfeasible (see, e.g., Chan et al., 2021). Second, imposing a shrinkage prior on the structural VAR formulation does not necessarily help to pin down the significance of cross-correlations in the reduced-form VAR formulation. This is especially relevant in forecasting applications whereby the main objective is to accurately identify predictive relationships across variables, rather than to identify structural shocks.

In this paper, we take a different approach towards posterior inference with hierarchical shrinkage priors in large VAR models. Specifically, we propose a novel variational Bayes estimation approach which allows for fast and accurate estimates of the reduced-form regression coefficients without leveraging on a structural VAR representation. This allows us to elicit hierarchical shrinkage priors directly on the matrix of regression coefficients so that (1) the prior structure directly maps into the posterior inference of the reduced-form transition matrix, and (2) posterior estimates are more robust to variables permutation. We also account for the effect of "exogenous" covariates and stochastic volatility in the residuals.

The key feature of our approach is that by abstracting from the linearity constraints implied

by a structural VAR formulation, one can provide a more direct identification of the reduced-form regression parameters. This could have important implications for forecasting within the context of weak predictability whereby the transition matrix and/or the coefficients on exogenous predictors are potentially sparse in nature (see, e.g., Bernardi et al., 2023). The main advantage of our variational inference approach is that an accurate identification of the regression parameters does not translate into a higher computational cost compared to existing Bayesian estimation methods. This is particularly relevant in practice for recursive forecasting implementations with higher frequency data, such as portfolio returns.

We investigate the accuracy of the posterior estimates based on an extensive simulation study for different model dimensions and variables permutation. As benchmarks, we consider a variety of established estimation approaches developed for large Bayesian VAR models, such as the linearized MCMC proposed by Chan and Eisenstat (2018); Cross et al. (2020) and its variational Bayes counterpart proposed by Chan and Yu (2022); Gefang et al. (2023). Both approaches are built upon a structural VAR formulation. In addition, we compare our variational Bayes method against the MCMC approach developed by Gruber and Kastner (2022), which is not constrained by a Cholesky factorization for parameters identification, similar to our approach. We test each estimation method for different hierarchical priors, such as the adaptive-Lasso of Leng et al. (2014), an adaptive version of the Normal-Gamma of Griffin and Brown (2010), and the Horseshoe of Carvalho et al. (2010).

Overall, the simulation results show that our variational inference approach represents the best trade-off between estimation accuracy and computational efficiency. Specifically, posterior inference from our variational Bayes method is as accurate as non-linear MCMC methods (see, e.g., Gruber and Kastner, 2022) but is considerably more efficient. At the same time, our approach is as efficient as conventional MCMC and variational Bayes methods based on a structural VAR formulation, but is considerably more accurate and less sensitive to variables permutation.

Our approach towards posterior inference in large VARs is guided by the principle that a more accurate identification of the reduced-form transition matrix should ultimately lead to better out-of-sample forecasts and financial decision making. To test this assumption, we investigate both the statistical and economic value of the forecasts from our variational Bayes approach within the context of a mean-variance investor who allocates her wealth between an industry portfolio and a risk-free asset based on lagged cross-industry returns and a series of macroeconomic predictors.

Although the model is general and can be applied to any type of financial returns, as far as data are stationary, our focus on different industry portfolios is motivated by a keen interest from researchers (see, e.g., Fama and French, 1997; Hou and Robinson, 2006) and practitioners alike. Indeed, the implications of industry returns predictability are arguably 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 (see, e.g., Rapach and Zhou, 2013), implies that at least some industry portfolio return is predictable.

The main results show that our variational inference approach fares better than competing methods in terms of out-of-sample point and density forecasts. We show that more accurate forecasts translate into larger economic gains as measured by certainty equivalent returns spreads vis-á-vis a naive investor which take investment decisions based on sample estimates of the conditional mean and variance of the returns. This holds across different hierarchical prior specifications. Overall, the empirical results support our view that by a more accurate identification of weak correlations between predictors and portfolio returns, one can significantly improve – both statistically and economically – the out-of-sample performance of large-scale multivariate time-series models.

Our paper connects to a growing literature exploring the use of Bayesian methods to estimate high-dimensional VAR models with shrinkage priors. A non-exhaustive list of works on the

topic contains Chan and Eisenstat (2018); Carriero, Clark, and Marcellino (2019); Huber and Feldkircher (2019); Chan and Yu (2022); Cross, Hou, and Poon (2020); Kastner and Huber (2020); Chan, Koop, and Yu (2021); Chan (2021); Carriero, Chan, Clark, and Marcellino (2022); Gruber and Kastner (2022); Gefang, Koop, and Poon (2023), among others. We contribute to this literature by providing a fast and accurate variational Bayes method which generalize posterior inference of quantities of interest by abstracting from a conventional structural VAR representation.

A second strand of literature we contribute to is related to the predictability of stock returns. 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. Early exceptions are Ferson and Harvey (1991); Ferson and Korajczyk (1995); Ferson and Harvey (1999) and Avramov (2004). We extend this literature by investigating the out-of-sample predictability of industry portfolios through the lens of a novel estimation method for large Bayesian VAR models.

2 Choosing the model parametrization

Let $\mathbf{y}_t = (y_{1,t}, \ldots, y_{d,t})^\intercal \in \mathbb{R}^d$ be a multivariate normal random variable and denote by $\mathbf{x}_t = (1, x_{1,t}, \ldots, x_{p,t})^\intercal \in \mathbb{R}^{(p+1)}$ a vector of covariates at time t. A vector autoregressive model with exogenous covariates and stochastic volatility is defined in compact form as:

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

with $\mathbf{z}_{t-1} = (\mathbf{y}_{t-1}^\intercal, \mathbf{x}_{t-1}^\intercal)^\intercal$ and $\mathbf{\Theta} = (\mathbf{\Phi}, \mathbf{\Gamma})$ consistently partitioned, where $\mathbf{\Phi} \in \mathbb{R}^{d \times d}$ is the transition matrix containing the autoregression coefficients and $\mathbf{\Gamma} \in \mathbb{R}^{d \times (p+1)}$ is the matrix

of regression parameters for the exogenous predictors. Here, $\mathbf{u}_t \in \mathbb{R}^d$ is a sequence of uncorrelated innovation terms such that $\mathbf{u}_{t-k} \perp \mathbf{u}_{t-j}\ \forall k, j$ with $k \neq j$ and $\mathbf{\Omega}_t \in \mathbb{S}^d_{++}$ being a symmetric and positive-definite time-varying precision matrix. A modified Cholesky factorization of $\mathbf{\Omega}_t$ can be conveniently exploited to re-write the model in Eq.(1) with orthogonal innovations (see, e.g., Rothman et al., 2010).

Let $\mathbf{\Omega}_t = \mathbf{L}^\intercal \mathbf{V}_t \mathbf{L}$, where $\mathbf{L} \in \mathbb{R}^{d \times d}$ is unit-lower-triangular and $\mathbf{V}_t \in \mathbb{S}^d_{++}$ is diagonal with time-varying elements $\mathbf{V}_t = \text{Diag}(\nu_{1,t}, \dots, \nu_{d,t})$ (see, e.g., Huber and Feldkircher, 2019; Gefang et al., 2023). By multiplying both sides of Eq.(1) by $\mathbf{L} = \mathbf{I}_d - \mathbf{B}$ one can obtain two alternative re-parametrizations of the same model:

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

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

where $\mathbf{A} = \mathbf{L}\mathbf{\Theta}$ and $\mathbf{B}$ has a strict-lower-triangular structure with elements $\beta_{j,k} = -l_{j,k}$ for $j = 2, \dots, d$ and $k = 1, \dots, j-1$. The key difference is that Eq.(2a) is non-linear in the parameters, while Eq.(2b) is linear. More importantly, Eq.(2b) is known as structural VAR representation, widely used in existing MCMC and variational Bayes estimations methods for high-dimensional VAR models (see, e.g., Chan and Eisenstat, 2018; Chan and Yu, 2022; Gefang et al., 2023). Instead, Eq.(2a) is the reduced-form parametrization at the core of our variational inference approach. This has also been used within the context of MCMC for smaller dimensions (see, e.g., Huber and Feldkircher, 2019; Gruber and Kastner, 2022).

From Eq.(2) one can obtain an equation-by-equation representation in which the j-th component of $\mathbf{y}_t$ becomes:

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

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

for all $j = 1, \dots, d$ and $t = 1, \dots, T$, 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 $j = 1, \dots, d$, 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 $\mathbf{y}_t^j = (y_{1,t}, \dots, y_{j-1,t})^\intercal \in \mathbb{R}^{j-1}$ 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 $\boldsymbol{\Theta}$. We follow Gefang et al. (2023); Chan and Yu (2022) and model the time variation in $\nu_{j,t}^{-1} = \exp(h_{j,t})$ assuming a log-volatility process $h_{j,t} = h_{j,t-1} + e_{j,t}$ with $e_{j,t} \sim \mathsf{N}(0, \psi_j)$, where the initial state $h_{0,j} \sim \mathsf{N}(0, k_0\, \psi_j)$, $k_0 \gg 0$, is unknown.

A discussion on variables permutation. Existing Bayesian approaches for large VAR models often rely on the structural representation in Eq.(2b), and therefore consider the elements in $\mathbf{A}$ as the parameters of interest. This has the key merit of simplifying the implementation of MCMC (see, e.g., Chan and Eisenstat, 2018) and variational Bayes algorithms (see, e.g., Gefang et al., 2023). Under the re-parametrization $\mathbf{A} = \mathbf{L}\boldsymbol{\Theta}$, each element $\vartheta_{i,j}$ – which denotes the (i,j)-entry of $\boldsymbol{\Theta}$ – is 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 $c_{i,j}$ are the (i,j)-entry of $\mathbf{A}$ and $\mathbf{L}^{-1}$, respectively.

This raises two main issues: first, $a_{i,j} = 0$ does not imply $\vartheta_{i,j} = 0$, that is a shrinkage prior on $\mathbf{A}$ does not preserve the structure of $\boldsymbol{\Theta}$. Second, the estimate $\widehat{\boldsymbol{\Theta}} = \widehat{\mathbf{L}}^{-1}\widehat{\mathbf{A}}$ for a given prior is potentially highly sensitive to variables permutation due to its dependence on the Cholesky factorization (see Gruber and Kastner, 2022 for a related discussion). Figure 1 provides a visual representation of this argument by comparing the estimates obtained based on Eq.(2a) vs Eq.(2b), for two different permutations of $\mathbf{y}_t$.

The evidence confirms that the estimates based on the transformation $\widehat{\boldsymbol{\Theta}} = \widehat{\mathbf{L}}^{-1}\widehat{\mathbf{A}}$ clearly diverge from the true $\boldsymbol{\Theta}$. In addition, the posterior estimates are influenced by the variables permutation. Instead, inference based on the representation in Eq.(2a) provides a more accurate identification of $\boldsymbol{\Theta}$ which is also less sensitive to variables permutation. Before taking this intuition to task both in simulation and on actual forecasting, in the next Section

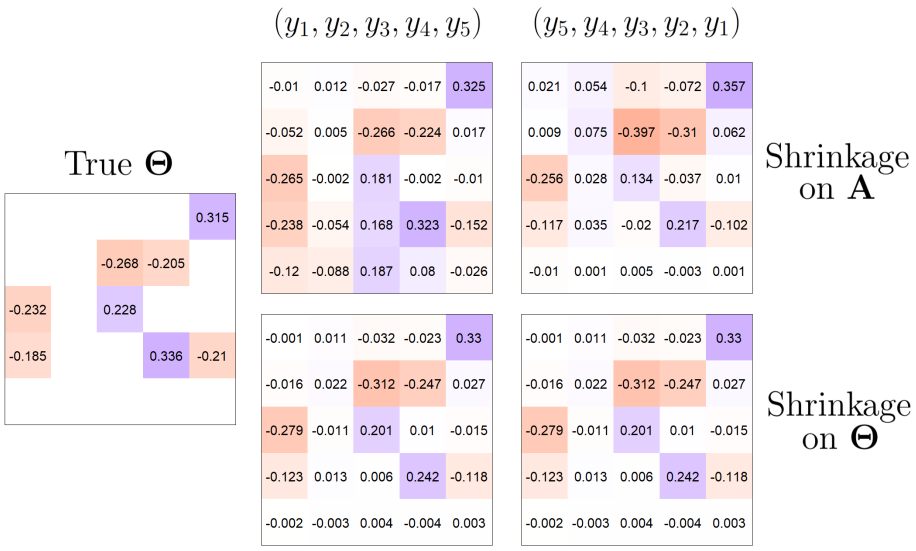


Figure 1: Comparison between the posterior inference for the linear representation $\mathbf{A} = \mathbf{L}\boldsymbol{\Theta}$ (first row) and the original parametrization $\boldsymbol{\Theta}$ (second row), for two different permutations of $\mathbf{y}_t$.

we provide details of our variational Bayes inference approach.

3 Variational Bayes inference

A variational approach to Bayesian inference requires to minimize the Kullback-Leibler (KL) divergence between an approximating density $q(\boldsymbol{\xi})$ and the true posterior density $p(\boldsymbol{\xi}|\mathbf{y})$, where $\boldsymbol{\xi}$ denotes the set of parameters of interest. Ormerod and Wand (2010) show that minimizing the KL divergence can be equivalently stated as the maximization of the "effective lower bound" (ELBO) denoted by $\underline{p}\,(\mathbf{y}; q)$:

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

where $q^*(\boldsymbol{\xi}) \in \mathcal{Q}$ represents the optimal variational density and $\mathcal{Q}$ is a space of density functions. Depending on the assumption on $\mathcal{Q}$, one falls into different variational paradigms. For instance, given a partition of the parameters vector $\boldsymbol{\xi} = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_p\}$, a mean-field variational Bayes (MFVB) approach assumes a factorization of the form $q(\boldsymbol{\xi}) = \prod_{j=1}^{p} q_i(\boldsymbol{\xi}_j)$.

A closed form expression for each optimal variational density $q^*(\boldsymbol{\xi}_j)$ can be defined as:

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

where the expectation is taken with respect to the joint approximating density with the j-th element of the partition removed $q^\star(\boldsymbol{\xi}\setminus\boldsymbol{\xi}_j)$. This allows to implement an efficient iterative algorithm to estimate the optimal density $q^*(\boldsymbol{\xi})$, although some components $q^*(\boldsymbol{\xi}_j)$ may remain too complex to handle and further restrictions are needed. If we assume that $q^*(\boldsymbol{\xi}_j)$ belongs to a pre-specified parametric family of distributions, the MFVB outlined above is sometimes labelled as *semi-parametric* (see Rohde and Wand, 2016).

3.1 Optimal variational densities

We present a factorization of the variational density $q(\boldsymbol{\xi})$ for the model outlined in Eq.(2a). As a benchmark, we consider a non-informative Normal prior for the regression coefficients. 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 $\psi_j \sim \mathsf{InvGa}(a_\psi, b_\psi)$ 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, $\mathsf{InvGa}(\cdot,\cdot)$ denotes the Inverse-Gamma distribution, and $a_\psi > 0$, $b_\psi > 0$, $\tau \gg 0$ and $\upsilon \gg 0$ are the related hyper-parameters. Let $\boldsymbol{\xi} = (\boldsymbol{\vartheta}^\intercal, \mathbf{h}^\intercal, \boldsymbol{\psi}^\intercal, \boldsymbol{\beta}^\intercal)^\intercal$ be the set of parameters of interest, the corresponding variational density can be factorised as $q(\boldsymbol{\xi}) = q(\boldsymbol{\vartheta})q(\mathbf{h})q(\boldsymbol{\psi})q(\boldsymbol{\beta})$, where:

$$q(\boldsymbol{\vartheta}) = \prod_{j=1}^{d} q(\boldsymbol{\vartheta}_j), \quad q(\mathbf{h}) = \prod_{j=1}^{d} q(\mathbf{h}_j), \quad q(\boldsymbol{\psi}) = \prod_{j=1}^{d} q(\psi_j), \quad q(\boldsymbol{\beta}) = \prod_{j=2}^{d} q(\boldsymbol{\beta}_j). \tag{6}$$

For the ease of exposition, in the main text of the paper we summarize the optimal variatonal density for the main parameters of interest $\boldsymbol{\Theta}$, with both a baseline non-informative prior and three alternative hierarchical shrinkage priors. The parameters and the full derivations of the optimal variational densities $q^*(\mathbf{h}_j) \equiv \mathsf{N}_{T+1}(\boldsymbol{\mu}_{q(h_j)}, \boldsymbol{\Sigma}_{q(h_j)})$, $q^*(\psi_j) \equiv \mathsf{InvGa}(a_{q(\psi_j)}, b_{q(\psi_j)})$, and

$q^*(\boldsymbol{\beta}_j) \equiv \mathsf{N}_{j-1}(\boldsymbol{\mu}_{q(\beta_j)}, \boldsymbol{\Sigma}_{q(\beta_j)})$ for $j = 1, \dots, d$, are reported in Proposition B.1.1, B.1.7 and B.1.4 of Appendix B, respectively. Notice these optimal variational densities are invariant across different shrinkage prior specifications for $\boldsymbol{\Theta}$. We leave to Proposition B.1.3 in Appendix B also the derivations for the constant volatility case with $\nu_{j,t} = \nu_j$ and $\nu_j \sim \mathsf{Ga}(a_\nu, b_\nu)$ for $j = 1, \dots, d$, where $\mathsf{Ga}(\cdot, \cdot)$ denotes the gamma distribution, and $a_\nu > 0$, $b_\nu > 0$. For the interested reader, Appendix B also provides the analytical form of the lower bound for each set of parameters.

Proposition 3.1 provides the optimal variational density for the j-th row of $\boldsymbol{\Theta}$ under the baseline Normal prior specification $\vartheta_{j,k} \sim \mathsf{N}(0, \upsilon)$. The proof and analytical derivations are available in Appendix B.1.

Proposition 3.1. *The optimal variational density for* $\boldsymbol{\vartheta}_j$ *is* $q^*(\boldsymbol{\vartheta}_j) \equiv \mathsf{N}_{d+p+1}(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$ *with hyper-parameters:*

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

where $\boldsymbol{\vartheta} = \begin{pmatrix} \boldsymbol{\vartheta}_j \\ \boldsymbol{\vartheta}_{-j} \end{pmatrix}$ *and* $\boldsymbol{\omega}_{j,t}$ *denotes the* j*-th row of* $\boldsymbol{\Omega}_t = \begin{pmatrix} \omega_{j,j,t} & \boldsymbol{\omega}_{j,-j,t} \\ \boldsymbol{\omega}_{-j,j,t} & \boldsymbol{\Omega}_{-j,-j,t} \end{pmatrix}$.

Notice that despite the multivariate model is reduced to a sequence of univariate regressions, the analytical form of the variational mean $\boldsymbol{\mu}_{q(\vartheta_j)}$ in Proposition 3.1 depends on all the other rows through $\boldsymbol{\mu}_{q(\vartheta_{-j})}$. As a result, the variational estimates of $\boldsymbol{\vartheta}_j$ explicitly depend on all of the other $\boldsymbol{\vartheta}_{-j}$. This addresses the issue in the MCMC algorithm of Carriero et al. (2019), which has been highlighted by Bognanni (2022) and corrected by Carriero et al. (2022).

Bayesian adaptive-Lasso. The Bayesian adaptive-Lasso of Leng et al. (2014) extends the original work of Park and Casella (2008) by assuming a different shrinkage for each

regression parameter based on a laplace distribution with an individual 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 normals 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 scaling parameters $\lambda_{j,k}^2$ are not fixed but inferred from the data by assuming a common hyper-prior distribution $\lambda_{j,k}^2 \sim \mathsf{Ga}(h_1, h_2)$, where $h_1, h_2 > 0$.

Let $\boldsymbol{\xi}_{\mathrm{L}} = (\boldsymbol{\xi}^\intercal, \boldsymbol{\upsilon}^\intercal, (\boldsymbol{\lambda}^2)^\intercal))^\intercal$ be the vector $\boldsymbol{\xi}$ augmented with the adaptive-Lasso prior parameters. The 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 density for the j-th row of $\boldsymbol{\Theta}$ under Bayesian adaptive-Lasso prior specification $\vartheta_{j,k}|\upsilon_{j,k} \sim \mathsf{N}(0, \upsilon_{j,k})$, $\upsilon_{j,k}|\lambda_{j,k}^2 \sim \mathsf{Exp}(\lambda_{j,k}^2/2)$, and $\lambda_{j,k}^2 \sim \mathsf{Ga}(h_1, h_2)$. The proof and analytical derivations are available in Appendix B.2.

Proposition 3.2. *The optimal variational density for* $\boldsymbol{\vartheta}_j$ *is* $q^*(\boldsymbol{\vartheta}_j) \equiv \mathsf{N}_{d+p+1}(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$ *with* $\boldsymbol{\Sigma}_{q(\vartheta_j)} = \left(\sum_{t=1}^{T} \boldsymbol{\mu}_{q(\omega_{j,j,t})} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal + \mathrm{Diag}(\boldsymbol{\mu}_{q(1/\upsilon_j)})\right)^{-1}$, *where* $\mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon_j)})$ *is a diagonal matrix with elements* $\boldsymbol{\mu}_{q(1/\upsilon_j)} = (\mu_{q(1/\upsilon_{j,1})}, \mu_{q(1/\upsilon_{j,2})}, \ldots, \mu_{q(1/\upsilon_{j,d+p+1})})$. *The parameters* $\boldsymbol{\mu}_{q(\vartheta_j)}$ *and* $\boldsymbol{\mu}_{q(\omega_{j,j,t})}$ *are as in Proposition 3.1. The optimal variational densities of the scaling parameters are* $q^*(\lambda_{j,k}^2) \equiv \mathsf{Ga}(a_{q(\lambda_{j,k}^2)}, b_{q(\lambda_{j,k}^2)})$ *with* $a_{q(\lambda_{j,k}^2)}, b_{q(\lambda_{j,k}^2)}$ *defined in Eq.(B.20), and* $q^*(1/\upsilon_{j,k}) \equiv \mathsf{IG}(a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})})$ *with* $a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})}$ *defined in Eq.(B.19).*

Adaptive Normal-Gamma. We expand the original Normal-Gamma prior of Griffin and Brown (2010) by assuming that each regression coefficient has a different shrinkage parameter, similar to the adaptive-Lasso. The hierarchical specification 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 $k = 1, \ldots, d+p+1$. Notice 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 a Variance-Gamma distribution. The hyper-parameters η_j and $\lambda_{j,k}$ are not fixed but are inferred from the data by

assuming two common hyper-priors $\lambda_{j,k} \sim \mathsf{Ga}(h_1, h_2)$ and $\eta_j \sim \mathsf{Exp}(h_3)$, where $h_l > 0$ 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 $\boldsymbol{\xi}$ augmented with the parameters of 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 density for the j-th row of $\boldsymbol{\Theta}$ under an adaptive Normal-Gamma specification $\upsilon_{j,k}|\eta_j, \lambda_{j,k} \sim \mathsf{Ga}\left(\eta_j, \eta_j\lambda_{j,k}/2\right)$, $\lambda_{j,k} \sim \mathsf{Ga}(h_1, h_2)$ and $\eta_j \sim \mathsf{Exp}(h_3)$. The proof and analytical derivations are available in Appendix B.3.

Proposition 3.3. *The optimal variational density for $\boldsymbol{\vartheta}_j$ is $q^*(\boldsymbol{\vartheta}_j) \equiv \mathsf{N}_{d+p+1}(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$ with $\boldsymbol{\Sigma}_{q(\vartheta_j)} = \left(\sum_{t=1}^{T} \boldsymbol{\mu}_{q(\omega_{j,j,t})}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal + \text{Diag}(\boldsymbol{\mu}_{q(1/\upsilon_j)})\right)^{-1}$, where $\mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon_j)})$ is a diagonal matrix with elements $\boldsymbol{\mu}_{q(1/\upsilon_j)} = (\mu_{q(1/\upsilon_{j,1})}, \mu_{q(1/\upsilon_{j,2})}, \ldots, \mu_{q(1/\upsilon_{j,d+p+1})})$. The parameters $\boldsymbol{\mu}_{q(\vartheta_j)}$ and $\boldsymbol{\mu}_{q(\omega_{j,j,t})}$ are as in Proposition 3.1. The optimal variational densities of the scaling parameters are $q^*(\lambda_{j,k}) \equiv \mathsf{Ga}(a_{q(\lambda_{j,k})}, b_{q(\lambda_{j,k})})$ with $a_{q(\lambda_{j,k})}, b_{q(\lambda_{j,k})}$ defined in Eq.(B.24), 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 normal distribution with $\zeta_{q(\upsilon_{j,k})}, a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})}$ defined in Eq.(B.23).*

Notice that the optimal density for the parameter η_j is not a known distribution function. Proposition B.3.3 in Appendix B.3 provides an analytical approximation of its moments so that the optimal density can be calculated via numerical integration.

Horseshoe prior. As a third hierarchical shrinkage prior we consider the Horseshoe prior 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 prior that implies an aggressive

shrinkage of weak signals without affecting the strong ones (see, e.g., Polson and Scott, 2011). We follow Wand et al. (2011) and leverage on a scale mixture representation of the half-Cauchy distribution as,

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

where the local and global shrinkage parameters are $\upsilon^2_{j,k}$ 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 $\boldsymbol{\xi}$ augmented with the parameters of 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 density for the j-th row of $\boldsymbol{\Theta}$ under the Horseshoe prior outlined in Eq.(10). The proof and analytical derivations are available in Appendix B.4.

Proposition 3.4. *The optimal variational density for $\boldsymbol{\vartheta}_j$ is $q^*(\boldsymbol{\vartheta}_j) \equiv \mathsf{N}_{d+p+1}(\boldsymbol{\mu}_{q(\vartheta_j)}, \boldsymbol{\Sigma}_{q(\vartheta_j)})$ with $\boldsymbol{\Sigma}_{q(\vartheta_j)} = \left(\sum_{t=1}^{T} \boldsymbol{\mu}_{q(\omega_{j,j,t})}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal + \mu_{q(1/\gamma^2)}\text{Diag}(\boldsymbol{\mu}_{q(1/\upsilon^2_j)})\right)^{-1}$, where $\mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon^2_j)})$ is a diagonal matrix with elements $\boldsymbol{\mu}_{q(1/\upsilon^2_j)} = (\mu_{q(1/\upsilon^2_{j,1})}, \mu_{q(1/\upsilon^2_{j,2})}, \ldots, \mu_{q(1/\upsilon^2_{j,d+p+1})})$. The parameters $\boldsymbol{\mu}_{q(\vartheta_j)}$ and $\boldsymbol{\mu}_{q(\omega_{j,j,t})}$ are as in Proposition 3.1. The optimal variational densities for the global shrinkage is $q^*(\gamma^2) \equiv \mathsf{InvGa}\left(\frac{1}{2}\{d(d+p+1)+1\}, b_{q(\gamma^2)}\right)$ with $b_{q(\gamma^2)}$ defined in Eq.(B.33), and $q^*(\eta) \equiv \mathsf{InvGa}(1, b_{q(\eta)})$ with $b_{q(\eta)}$ defined in Eq.(B.35). The optimal variational densities for the local shrinkage parameters are $q^*(\upsilon^2_{j,k}) \equiv \mathsf{InvGa}(1, b_{q(\upsilon^2_{j,k})})$ and $q^*(\lambda_{j,k}) \equiv \mathsf{InvGa}(1, b_{q(\lambda_{j,k})})$, with $b_{q(\upsilon^2_{j,k})}$ and $b_{q(\lambda_{j,k})}$ defined in Eq.(B.32) and Eq.(B.34), respectively.*

3.2 From shrinkage to sparsity

In addition to computational tractability, shrinking rather than selecting is a defining feature of the hierarchical priors outlined in Section 3.1. That is, posterior estimates of $\boldsymbol{\Theta}$ are non-sparse, and thus can not provide exact differentiation between significant vs non-significant predictors. The latter is particularly relevant since we ultimately want to assess the accuracy of our variational inference approach – versus existing MCMC and variational Bayes algorithms – in identifying the exact structure of $\boldsymbol{\Theta}$.

To address this issue, we build upon Ray and Bhattacharya (2018) and implement a Signal Adaptive Variable Selector (SAVS) algorithm to induce sparsity in $\widehat{\boldsymbol{\Theta}}$, conditional on a given prior. The SAVS is a post-processing algorithm which divides signals and nulls on the basis of the point estimates of the regression coefficients (see, e.g., Hauzenberger, Huber, and Onorante, 2021). Specifically, let $\widehat{\vartheta}_j$ the posterior estimate of ϑ_j and $\mathbf{z}_j$ the associated vector of covariates. 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 we rely on the SAVS post-processing to induce sparsity in the posterior estimates is threefold. First, as highlighted by Ray and Bhattacharya (2018), the SAVS represents an automatic procedure in which the sparsity-inducing property directly depends on the effectiveness of the shrinkage performed on $\widehat{\vartheta}_j$. This refers to the precision of the posterior mean estimates; that is, the more accurate is $\widehat{\vartheta}_j$, the more precise is the identification of the non-zero elements in $\boldsymbol{\Theta}$. Second, the SAVS is "agnostic" with respect to the shrinkage prior or estimation approach adopted, so it represents a natural tool to compare different estimation methods. 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).

In addition to SAVS, we also expand on Hahn and Carvalho (2015) (HC henceforth) and provide a multivariate extension to their least-angle regression which has originally been built for univariate regressions. Appendix D.2 provides the full derivation of our extended HC approach as well as a complete discussion of the drawbacks compared to SAVS. In addition,

for the interested reader, Appendix D provides a direct comparison between the SAVS and our multivariate extension to Hahn and Carvalho (2015) based on simulated data (see also the discussion in Section 4).

3.3 Variational predictive density

Consider the posterior distribution $p(\boldsymbol{\xi}|\mathbf{z}_{1:t})$ given the information set $\mathbf{z}_{1:t} = \{\mathbf{y}_{1:t}, \mathbf{x}_{1:t}\}$ and the conditional likelihood $p(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\xi})$. A standard predictive density takes the 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

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

Although an analytical expression for Eq.(13) is not available, a simulation-based estimator for $q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t})$ can be obtained through Monte Carlo integration by averaging $p(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\xi}^{(i)})$ over the draws $\boldsymbol{\xi}^{(i)} \sim q^*(\boldsymbol{\xi})$, such that $\widehat{q}(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = N^{-1} \sum_{i=1}^{N} p(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\xi}^{(i)})$. Notice that a complete characterization of the optimal variational predictive density entails $q^*(\boldsymbol{\Omega}_t)$ with $\boldsymbol{\Omega}_t = \mathbf{L}^\intercal \mathbf{V}_t \mathbf{L}$. Proposition 3.5 shows that, conditional on $\mathbf{L}$ and $\mathbf{V}_t$, the optimal distribution of $\boldsymbol{\Omega}_t$ can be approximated by a d-dimensional Wishart distribution $\mathsf{Wishart}_d(\delta_t, \mathbf{H}_t)$, where δ_t and $\mathbf{H}_t$ are the degrees of freedom parameter and the scaling matrix, respectively.

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

The complete proof is available in Appendix C.1 and is based on the Expectation Propagation (EP) approach proposed by Minka (2001). In order to implement this approach, there is

no need to know $q^*(\boldsymbol{\Omega}_t)$, but it is sufficient to be able to compute $\mathbb{E}_q(\boldsymbol{\Omega}_t)$. The latter can be reconstructed based on the optimal variational densities of the Cholesky factor $q^*(\boldsymbol{\beta})$ – and therefore for $\mathbf{L}$ – and of $q^*(\mathbf{V}_t)$. The simulation results in Appendix C.1 show that the proposed Wishart distribution provides an accurate approximation of $q^*(\boldsymbol{\Omega}_t)$ for both small and large dimensional models.

Based on Proposition 3.5, we can further simplify Eq.(13) by integrating $\boldsymbol{\Omega}_t$ 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 probability density function of a multivariate Student-t distribution $\mathtt{t}_v(\mathbf{m}, \mathbf{S})$ with mean $\mathbf{m} = \boldsymbol{\Theta}\mathbf{z}_t$, scaling matrix $\mathbf{S} = (v\widehat{\mathbf{H}})^{-1}$, and degrees of freedom parameter $v = \widehat{\delta} - d + 1$. 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.

Notice that the main advantage of the approximation obtained from Proposition 3.5 is to allow for a considerably faster computation of the variational predictive density, compared to using $q^*(\mathbf{L})$ and $q^*(\mathbf{V}_t)$ as stationary distributions to sample $\boldsymbol{\Omega}_t$, similar to an MCMC. This is because the scaling matrix of the Wishart distribution is available in closed form and the computation of degrees of freedom requires only a one-dimensional optimization. In Appendix C.2 we discuss a further simplification that minimizes the KL divergence between the multivariate Student-t and a multivariate Normal distribution.

4 Simulation study

In this section, we report the results of an extensive simulation study designed to compare the properties of our estimation approach against both MCMC and variational Bayes methods

for large VAR models. To begin, we compare our `VB` algorithm against the MCMC approach of Chan and Eisenstat (2018); Cross et al. (2020) and the variational inference framework proposed by Chan and Yu (2022); Gefang et al. (2023). Both these approaches are built upon the structural VAR representation in Eq.(2b). Then, we also compare our `VB` method against the MCMC approach developed by Huber and Feldkircher (2019); Gruber and Kastner (2022) which is based upon a non-linear parametrization as in Eq.(2a), similar to our approach.

For the sake of comparability with Gruber and Kastner (2022); Gefang et al. (2023), which do not consider the presence of exogenous predictors, we consider a standard VAR(1) as data generating process. Consistent with the empirical implementations, we set $T = 360$ and $d = 30, 49$. The choice of d is due to the two alternative industry classifications which are explored in the main empirical analysis. We assume either a moderate – 50% of zeros – or a high – 90% of zeros – level of sparsity in the true matrix $\mathbf{\Theta}$. The latter is generated as follows: we fix to zero $s \cdot d^2$ entries at random, with $s = 0.5, 0.9$ and $d = 30, 49$, while the remaining non-zero coefficients are sampled from a mixture of two normal distributions with means equal to ± 0.08 and standard deviation 0.1. Appendix D provides additional details on the data generating process and additional simulation results for $d = 15$.

4.1 Estimation accuracy

As a measure of point estimation accuracy, we first look at the Frobenius norm $\|\mathbf{\Theta} - \widehat{\mathbf{\Theta}}\|_F$, which measures the difference between the true $\mathbf{\Theta}$ observed at each simulation and its estimate $\widehat{\mathbf{\Theta}}$. In addition, we compare the ability of each estimation method to identify the non-zero elements in the true $\mathbf{\Theta}$ based on the F1 score. The latter can be expressed as a function of counts of true positives (tp), false positives (fp) and false negatives (fn),

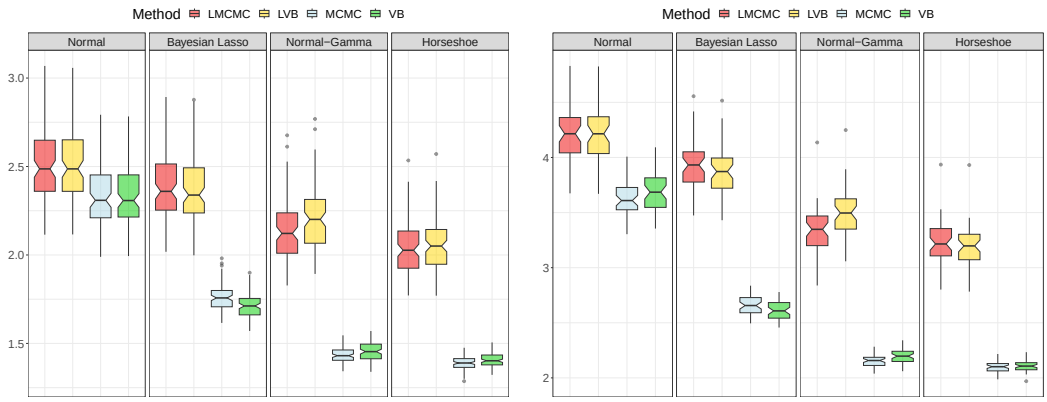
$$\text{F1} = \frac{2tp}{2tp + fp + fn}.$$

The F1 score takes value one if identification is perfect, i.e., no false positives and no false negatives, and zero if there are no true positives. We compute both measures of estimation accuracy on $N = 100$ replications to compare each estimation method and prior specification. The estimates from the MCMC specifications are based on 5,000 posterior simulations, after discarding the first 5,000 as a burn-in sample.

Point estimates. Figure 2 shows the box charts summarizing the Frobenius norm $\|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}\|_F$ across $N = 100$ replications. We label the linearized MCMC and variational methods with `LMCMC` and `LVB`, respectively, with `MCMC` the non-linear method of Gruber and Kastner (2022) and with `VB` our variational inference method, respectively. To increase readability, we separate the results by prior and color-code the four different estimation methods. For instance, for a given sub-plot we report the results for the Normal, adaptive-Lasso, adaptive Normal-Gamma and Horseshoe priors from the left to the right panel. Within each panel, the simulation results for the `LMCMC`, `LVB`, `MCMC` and `VB` estimates are reported in red, yellow, light-blue and green, respectively.

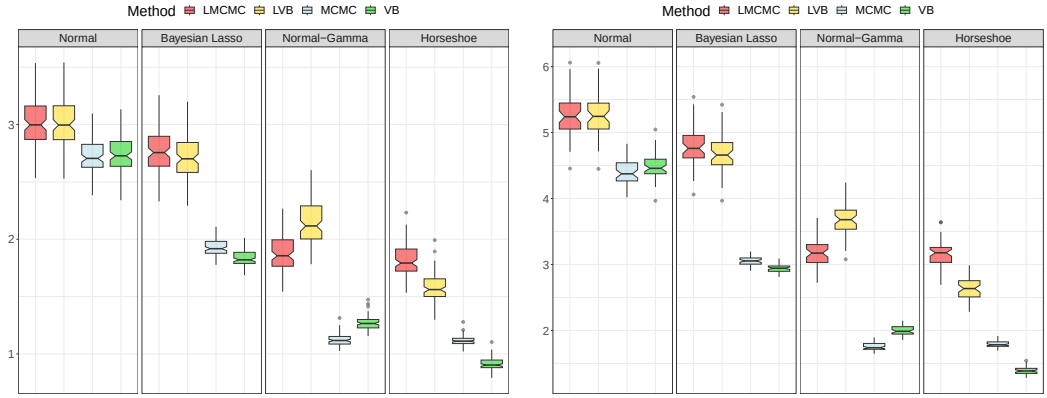
Beginning with the moderate sparsity case (top panels), the simulation results show that `LMCMC` and `LVB` approaches tend to perform equally across different shrinkage priors, with the only exception of the Normal-Gamma prior, in which `LMCMC` slightly outperforms `LVB`. However, the discrepancy between the two structural VAR representation methods tend to increase when sparsity becomes more pervasive (see bottom panels).

Overall, the simulation results support our view that, by eliciting shrinkage priors directly on $\boldsymbol{\Theta}$ – as per the parametrization in Eq.(2a) – the accuracy of the posterior estimates improves. The mean squared errors obtained from `MCMC` and `VB` are lower compared to both `LMCMC` and `LVB`. This holds for all priors and the model dimension. The accuracy with $d = 30$ of the `MCMC` and `VB` is virtually the same. Yet, with $d = 49$ our `VB` produces slightly more accurate estimates than `MCMC` for both the adaptive-Lasso and the Horseshoe prior.



(a) $d = 30$, moderate sparsity

(b) $d = 49$, moderate sparsity



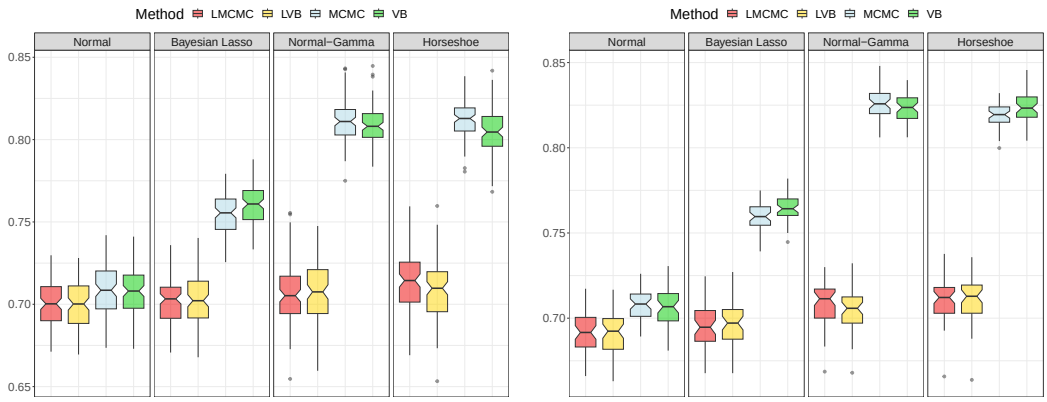
(c) $d = 30$, high sparsity

(d) $d = 49$, high sparsity

Figure 2: Frobenius norm of $\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}$ across $N = 100$ replications, for different shrinkage priors and different inference methods.

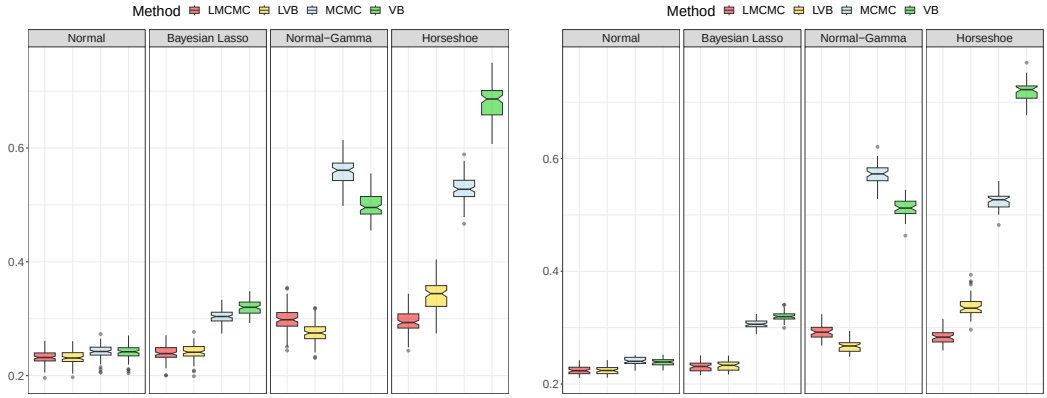
Sparsity identification. Figure 3 shows the box charts of F1 scores across $N = 100$ simulations. The labeling is the same as in Figure 2. Both `LMCMC` and `LVB` produce a rather dismal identification of the non-zero elements in $\boldsymbol{\Theta}$ across prios and model dimensions. This is due to the fact that $\widehat{\boldsymbol{\Theta}} = \widehat{\mathbf{L}}^{-1}\widehat{\mathbf{A}}$ in Eq.(2b), so that a sparse estimate of $\widehat{\mathbf{A}}$ does not map into a sparse estimate of $\widehat{\boldsymbol{\Theta}}$, and therefore produces a lower accuracy in identifying the non-zero coefficients in the true $\boldsymbol{\Theta}$. As the level of sparsity increases, the divergence between $\mathbf{A}$ and $\boldsymbol{\Theta}$ increases.

Consistent with our argument in favor of the parametrization in Eq.(2a), both the `MCMC` and `VB` approaches produce a more accurate identification of the non-zero coefficients in $\boldsymbol{\Theta}$, as



(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 across $N = 100$ replications by looking at the true non-null parameters in $\boldsymbol{\Theta}$ and the non-null parameters estimated based on $\widehat{\boldsymbol{\Theta}}$.

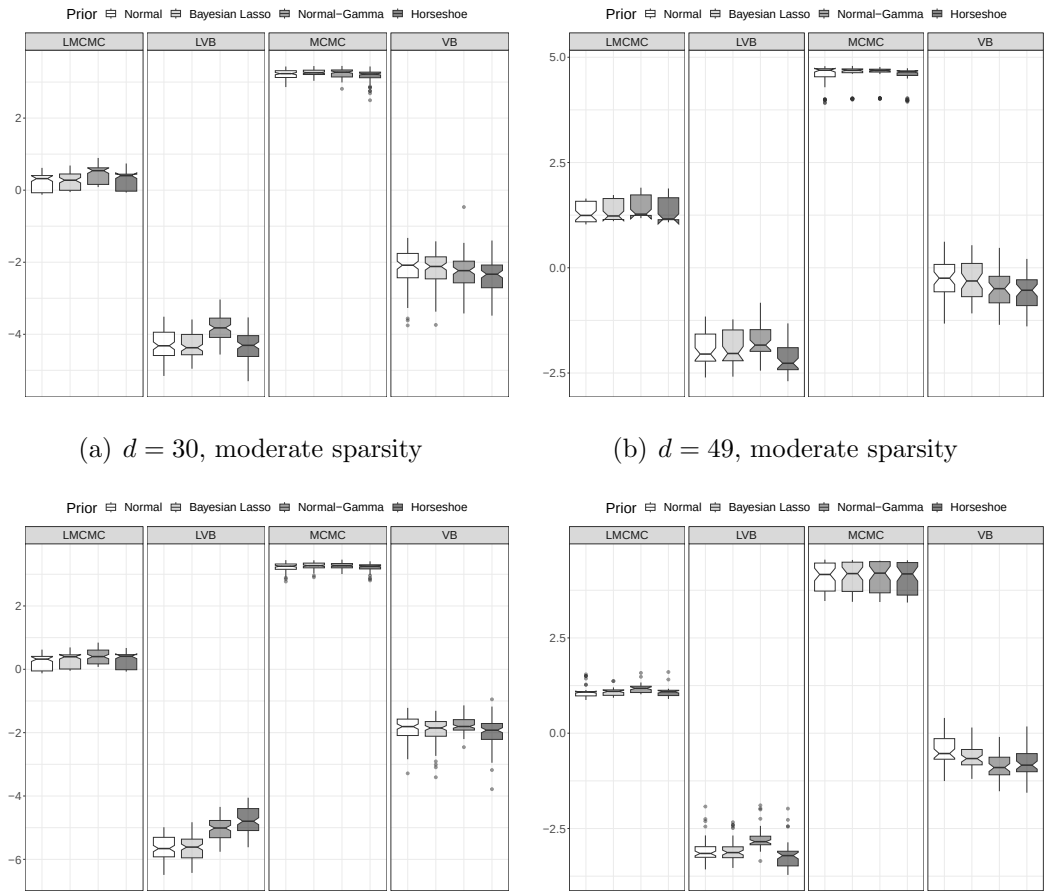
shown by the F1 score. The gap between `LMCMC`, `LVB` versus `MCMC` and `VB` becomes larger for higher levels of sparsity. This result holds across different hierarchical shrinkage priors and for different VAR dimensions. Yet, our `VB` approach turns out to be more accurate than `MCMC` under the adaptive-Lasso and Horseshoe priors for higher levels of sparsity.

As outlined in Section 3, sparsity in the posterior estimates for $\widehat{\boldsymbol{\Theta}}$ for different hierarchical shrinkage priors is induced in the simulation results by using the SAVS algorithm of Ray and Bhattacharya (2018). Appendix D provides additional simulation results obtained by implementing a multivariate version of the post-processing method proposed by Hahn and Carvalho (2015) as an alternative to the SAVS. A full derivation is provided in Appendix

D.2. The F1 scores are largely the same across methods; in fact, the evidence is even more in favour of our `VB`, compared to its `MCMC` counterpart when using the extended Hahn and Carvalho (2015) approach: our `VB` is more accurate than `MCMC` with a Normal-Gamma prior.

Computational efficiency. Chan and Yu (2022) and Gefang et al. (2023) highlight that one of the main advantages of variational Bayes methods is computational efficiency. Figure 4 reports the computational time – expressed in a log-minute scale – required by each estimation approach under different shrinkage priors. To highlight the performance for a given prior, we separate the results by estimation methods and color-code the four different shrinkage priors. For instance, for a given sub-plot, we report the results for the `LMCMC`, `LVB`, `MCMC` and `VB` estimates from left to right panel. Within each panel, the Normal, adaptive-Lasso, adaptive Normal-Gamma, and Horseshoe priors are colored in shades of gray from light (left) to dark (right) grey, respectively. To guarantee a more accurate comparability, we re-coded all competing methods in **Rcpp** and use the same 2.5 GHz Intel Xeon W-2175 with 32GB of RAM for all implementations.

The results highlight that our `VB` approach has a clear computational advantage compared to both linear and non-linear MCMC methods. For instance, for $d = 30$ our `VB` is more than 100 times faster than the `MCMC` of Gruber and Kastner (2022) and more than 10 times faster than the `LMCMC` of Cross et al. (2020), respectively. The gap in favour of our `VB` method compared to both `LMCMC` and `MCMC` increases in larger dimensions; for $d = 49$ the `MCMC` approach takes almost 60 minutes, on average, to generate comparably accurate posterior estimates to our `VB`, which instead takes approximately between 30 to 40 seconds, on average. Such efficiency gap between `VB` and `MCMC` has profound implications for a practical forecasting implementation, especially within the context of recursive predictions with higher frequency data such as stock returns (see Section 5.2). Perhaps not surprisingly, the `LVB` approach of Chan and Yu (2022); Gefang et al. (2023) is highly competitive in terms of computational efficiency. However, being built on a structural VAR formulation, we showed in Figures 2



(a) $d = 30$, moderate sparsity

(b) $d = 49$, moderate sparsity

(c) $d = 30$, high sparsity

(d) $d = 49$, high sparsity

Figure 4: Computational time required by each estimation approach for different hierarchical shrinkage priors. The time is expressed on logarithmic minutes scale.

and 3 that such computational efficiency comes at the cost of a lower estimation accuracy.

Appendix E.1 also provides a broader qualitative discussion on the computational costs of some of the existing MCMC approaches. Specifically, we review some of the results reported in the original papers and show that these largely align with our own findings. In addition, we also discuss some of the limitations of the non-linear `MCMC` for the recursive forecasting implementation (see Section 5.2 for more details).

Robustness to variables permutation. At the outset of the paper, we argue that a conventional structural VAR formulation potentially generates posterior estimates which are

not permutation-invariant. That is, posterior estimates of $\mathbf{\Theta}$ are sensitive to the ordering imposed on the target variables $\mathbf{y}_t$, conditional on a given prior. To highlight this issue, in Appendix D, we report a set of additional simulation results for all estimation methods and shrinkage priors under variables permutation.

The results show that the accuracy of the posterior estimates from both `LMCMC` and `LVB` changes once the variables ordering is reversed (see Figure D.4). This is especially clear for the Normal-Gamma and Horseshoe priors, and when the amount of zero coefficients in $\mathbf{\Theta}$ is more pervasive. On the other hand, the estimation accuracy of both the `MCMC` approach of Gruber and Kastner (2022) and our `VB` method does not substantially deteriorates by arbitrarily changing ordering of the target variables. Overall a substantially higher computational efficiency coupled with a comparable accuracy with complex MCMC, makes our `VB` extremely competitive within the context of recursive forecasts with higher frequency data.

5 A empirical study of industry returns predictability

We investigate both the statistical and economic value of our variational Bayes approach within the context of US industry returns predictability. To expand the scope of the testing framework, 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 size of the cross sections change due to a different industry classification. 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 $t + 1$. The sample periods cover major events, from the great depression to the Covid-19 outbreak.

In addition to cross-industry portfolio returns, we consider a variety of predictors, such as the returns on the market portfolio (`mkt`), and the returns on four alternative long-short investment strategies based on market capitalization (`smb`), book-to-market ratios (`hml`),

operating profitability (rmw) and firm investments (cma) (see Fama and French, 2015). We also consider a set of additional macroeconomic predictors from Goyal and Welch (2008), such as the log price-dividend ratio (pd), the difference between the long term yield on government bonds and the T-bill (term), the BAA-AAA bond yields difference (credit), the monthly log change in the CPI (infl), the aggregate market book-to-market ratio (bm), the net-equity issuing activity (ntis) and the corporate bond returns (corpr).

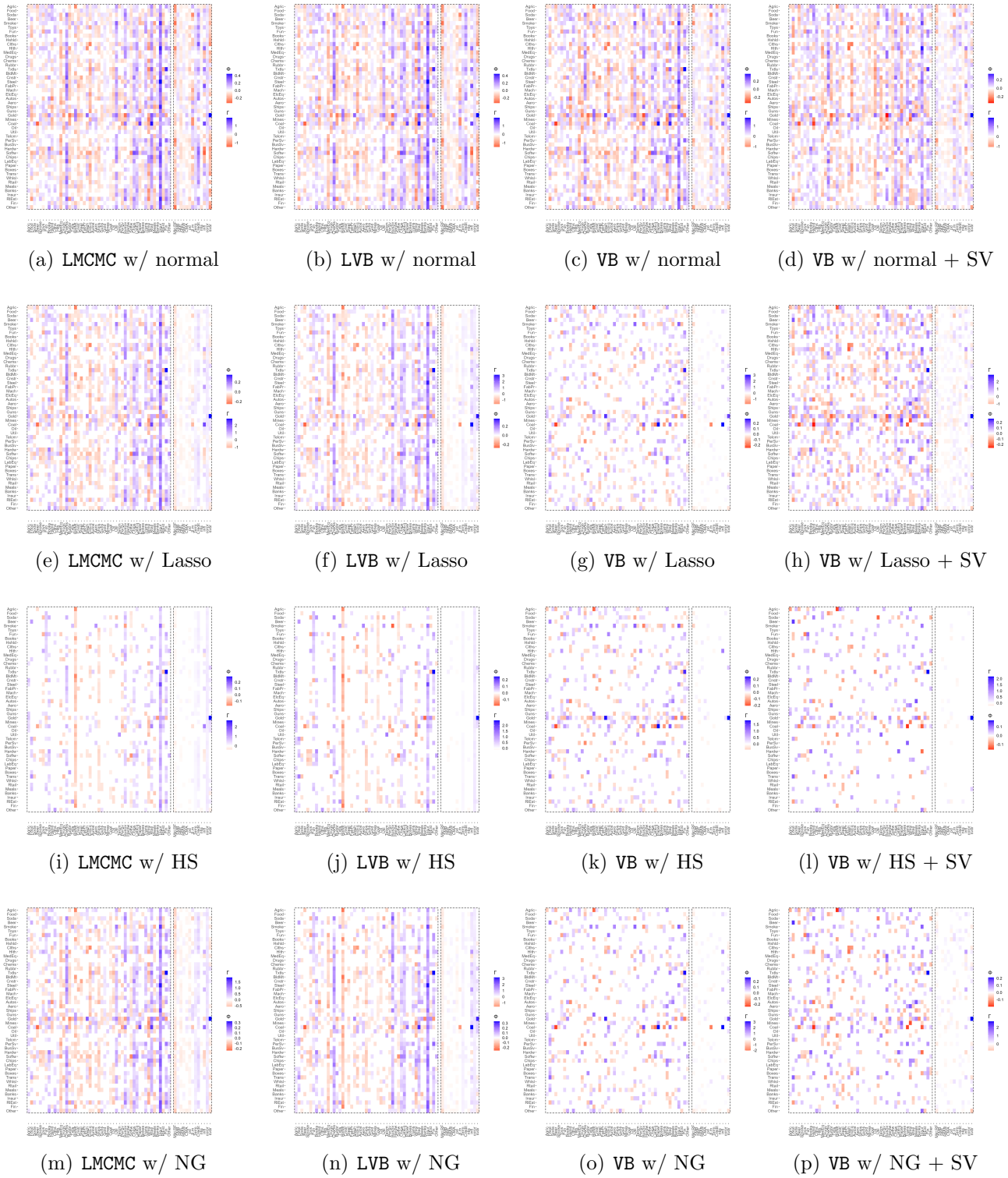
5.1 In-sample estimates of $\mathbf{\Theta}$

In order to highlight some of the main properties of different estimation methods, we first report the in-sample estimates of $\mathbf{\Theta}$ for the $d = 49$ industry case across all priors. Figure 5 compares $\widehat{\mathbf{\Theta}}$ based on the full sample obtained from the LMCMC and the LVB with constant volatility, and our VB with and without stochastic volatility. Appendix E.3 reports the additional in-sample estimates for $d = 30$ industry portfolios.

The in-sample estimates highlight three key results. First, there are visible differences across shrinkage priors. For instance, the Horseshoe tend to shrink parameters more aggressively towards zero so that $\widehat{\mathbf{\Theta}}$ is more sparse compared to, for e.g., the adaptive Normal-Gamma. Second, consistent with Gefang et al. (2023), the estimates of the LMCMC and LVB tend to be closely related. Yet, these in-sample estimates are substantially different compared to our VB approach. This is due to the re-parametrization $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{L}}^{-1}\widehat{\mathbf{A}}$ in Eq.(2b); that is, the estimated $\widehat{\mathbf{A}}$ is not translation-invariant, unlike in our approach. Third, with the exception of the adaptive-Lasso prior, the estimates $\widehat{\mathbf{\Theta}}$ from VB are remarkably stable between constant vs stochastic volatility specifications.

5.2 Out-of-sample forecasting accuracy

Intuitively, different estimates of $\mathbf{\Theta}$ should reflect in different conditional forecasts. To test this intuition we now compare the LMCMC, LVB and the VB estimation approaches with and



(a) LMCMC w/ normal (b) LVB w/ normal (c) VB w/ normal (d) VB w/ normal + SV

(e) LMCMC w/ Lasso (f) LVB w/ Lasso (g) VB w/ Lasso (h) VB w/ Lasso + SV

(i) LMCMC w/ HS (j) LVB w/ HS (k) VB w/ HS (l) VB w/ HS + SV

(m) LMCMC w/ NG (n) LVB w/ NG (o) VB w/ NG (p) VB w/ NG + SV

Figure 5: Variational Bayes estimates of the regression coefficients $\mathbf{\Theta}$ for different estimation methods. We report the estimates for the $d = 49$ industry case obtained for all priors. We report the results for VB with and without stochastic volatility.

without stochastic volatility. For the sake of completeness, we also consider a series of univariate model specifications (`U` henceforth), which corresponds to assuming conditional independence across industry portfolios. We consider a 360 months rolling window period for each model estimation; for instance for the 30-industry classification the out-of-sample period is from July 1957 to May 2020.

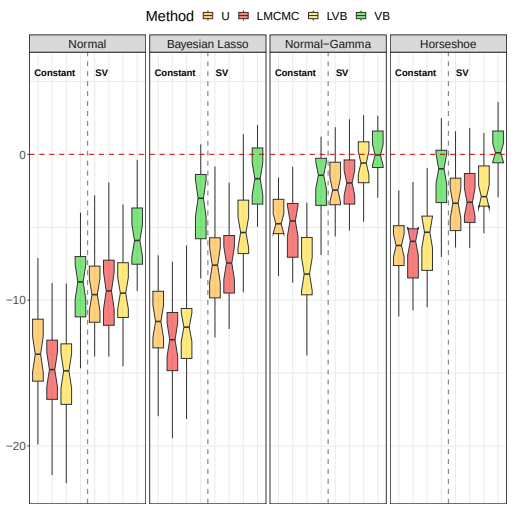
Notice that given the recursive nature of the empirical implementation we do not consider the `MCMC` approach of Gruber and Kastner (2022). This is because the computational cost would make such implementation prohibitive in practice, as discussed in the simulation study based on Figure 4. For instance, on a 2.5 GHz Intel Xeon W-2175 with 32GB of RAM and 14 cores it would take 20 min $\times$ 767 forecasts $\times$ 4 priors $= 61,360$ minutes, or 42 days, to implement the `MCMC` approach for recursive forecasting for the 30 industry portfolios with constant volatility. The computational cost would be even more prohibitive when adding stochastic volatility and/or for the 49 industry portfolios. Appendix E.1 provides an additional discussion on the computational costs of some of the existing MCMC approaches and the key relevance for a higher-frequency forecasting implementation such as ours.

Point forecasts. We begin by inspecting the accuracy of point forecasts for each industry based on the out-of-sample predictive R squared (see, e.g., Goyal and Welch, 2008),

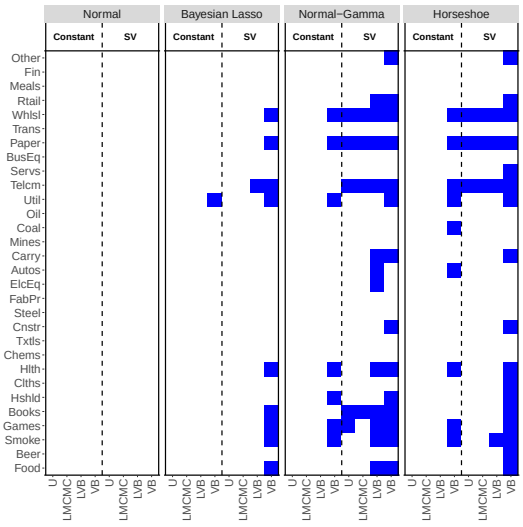
$$R^2_{j,oos}\left(\mathcal{M}_s\right) = 1 - \frac{\sum_{t_0=2}^{T}\left(y_{jt} - \widehat{y}_{jt}\left(\mathcal{M}_s\right)\right)^2}{\sum_{t_0=2}^{T}\left(y_{jt} - \overline{y}_{jt}\right)^2},$$

where t_0 is the date of the first prediction, $\overline{y}_{jt}$ is the naive forecast from the recursive mean – using the same rolling window of observations – and $\widehat{y}_{jt}\left(\mathcal{M}_s\right)$ is the conditional mean returns for industry $j = 1, \ldots, d$ for a given model $\mathcal{M}_s$.

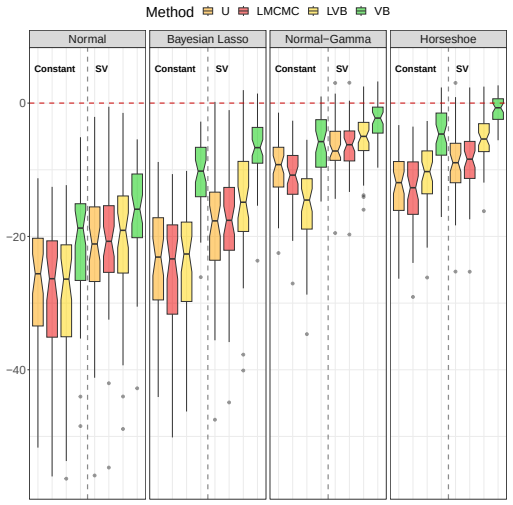
The left panels of Figure 6 show the box charts with the distribution of the $R^2_{j,oos}$ across $j = 1, \ldots, d$ industries. For a given sub-plot the results for the Normal, Bayesian Lasso, Normal-Gamma and Horseshoe priors are reported from the left to the right. Within each



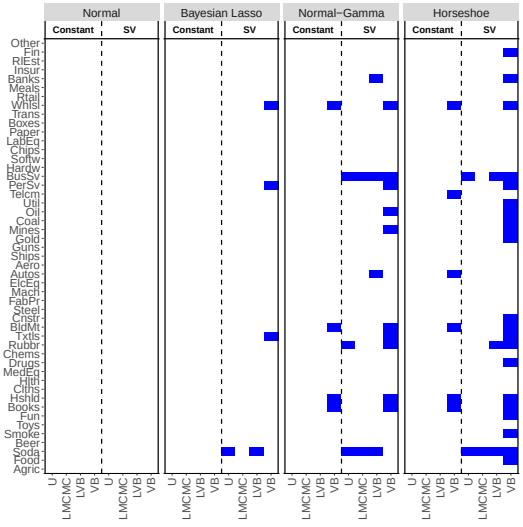
(a) $R_{j,oos}\left(\mathcal{M}_s\right)^2$ across 30 industry portfolios



(b) Portfolios for which $R^2_{j,oos}\left(\mathcal{M}_s\right) > 0$



(c) $R^2_{j,oos}\left(\mathcal{M}_s\right)$ across 49 industry portfolios



(d) Portfolios for which $R^2_{j,oos}\left(\mathcal{M}_s\right) > 0$

Figure 6: Left panels report the $R^2_{j,oos}\left(\mathcal{M}_s\right)$ (in %) across industry portfolios. Right panels report the industries for which a given model can generate $R^2_{j,oos}\left(\mathcal{M}_s\right) > 0$. The top (bottom) panels report the results for 30 (49) industry portfolios.

panel of a sub-plot, the forecasting results for the `U`, `LMCMC`, `LVB`, and `VB` estimates are color coded in orange, red, yellow, and green (from left to right), respectively. The vertical dashed line within each panel separates between constant and stochastic volatility specifications. Based on the same separation across methods and priors, the right panels of Figure 6 report a breakdown of the industries for which the corresponding $R^2_{j,oos}\left(\mathcal{M}_s\right) > 0$.

The out-of-sample $R^2_{j,oos}(\mathcal{M}_s)$ tend to be mostly negative across estimation methods and shrinkage priors. This is consistent with the existing evidence on stock returns predictability: a simple naive forecast based on a rolling sample mean represents a challenging benchmark to beat (see, e.g., Campbell and Thompson, 2007). However, our variational inference approach substantially improves upon univariate regressions, as well as upon the `LMCMC` and `LVB` methods, which are both based on a structural VAR representation.

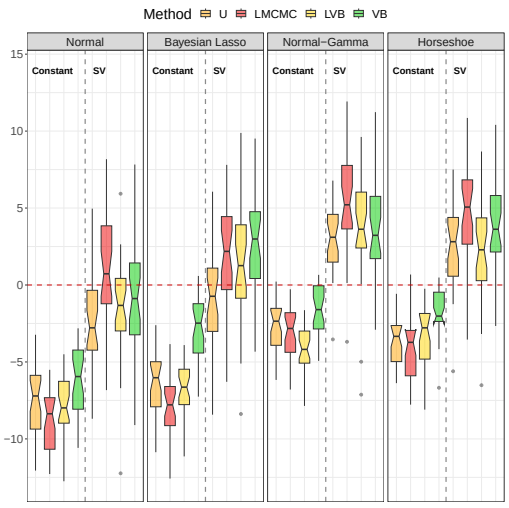
For instance, our `VB` with stochastic volatility generates a positive $R^2_{j,oos}(\mathcal{M}_s)$ for more than half of the 30 industry portfolios based on the adaptive Normal-Gamma and the Horseshoe. This compares to 4 (adaptive Normal-Gamma) and 3 (Horseshoe) positive $R^2_{j,oos}(\mathcal{M}_s)$ obtained from `LMCMC` with stochastic volatility. The gap further increases within the 49-industry classification; our `VB` method is virtually the only approach that can systematically generate positive $R^2_{j,oos}(\mathcal{M}_s)$ across industries. Although concentrated on the Horseshoe prior, the out-performance of our method relative to both `LMCMC` and `VB` holds across different priors.

Density forecasts. We follow Fisher et al. (2020) and assess the accuracy of the density forecasts across priors and estimation methods based on the average log-score (ALS) differential with respect to a "no-predictability" benchmark,

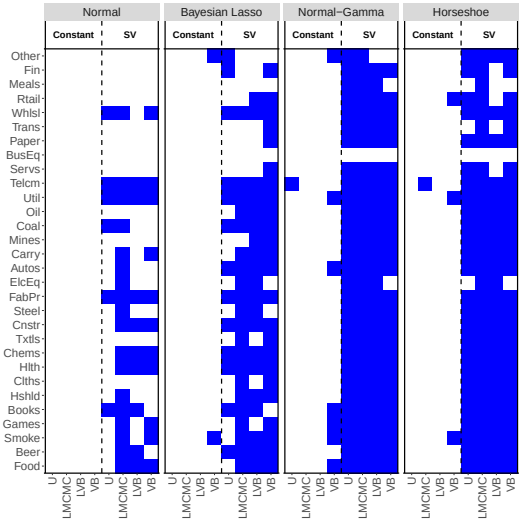
$$\text{ALS}_j(\mathcal{M}_s) = \frac{1}{T-t_0}\sum_{t_0=2}^{T}\left(\ln S_{jt}(\mathcal{M}_s) - \ln \overline{S}_{jt}\right), \tag{15}$$

where $\ln S_{jt}(\mathcal{M}_s)$ denotes the log-score at time t for industry j obtained by evaluating a Normal density with the conditional mean and variance forecast from the model $\mathcal{M}_s$. Consistent with the rationale of $R^2_{j,oos}(\mathcal{M}_s)$, the log-score for the no-predictability benchmark $\ln \overline{S}_{j,t}$ is constructed by evaluating a Normal density based on recursive mean and variance.

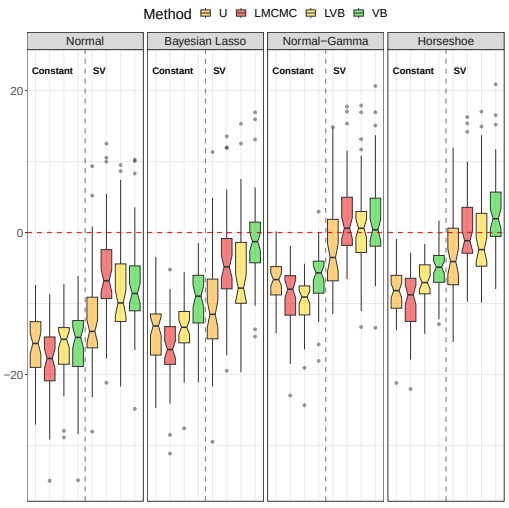
Figure 7 reports the results. The labeling is the same as in Figure 6. Not surprisingly, we find that by adding stochastic volatility the accuracy of density forecasts substantially improves across priors and estimation methods. For instance, our `VB` method with stochastic



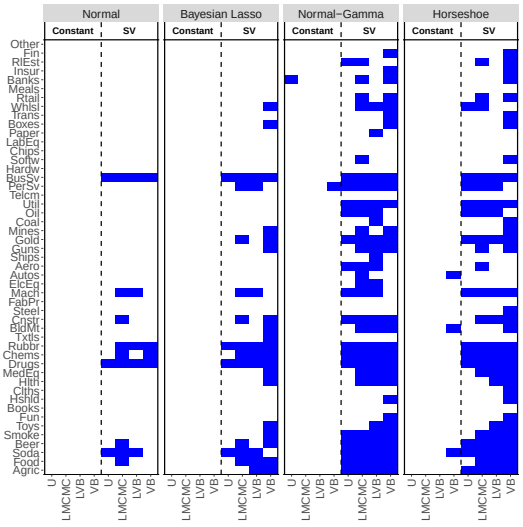
(a) $\text{ALS}_j\left(\mathcal{M}_s\right)$ for the 30 industry portfolios



(b) Portfolios for which $\text{ALS}_j\left(\mathcal{M}_s\right) > 0$



(c) $\text{ALS}_j\left(\mathcal{M}_s\right)$ for the 49 industry portfolios



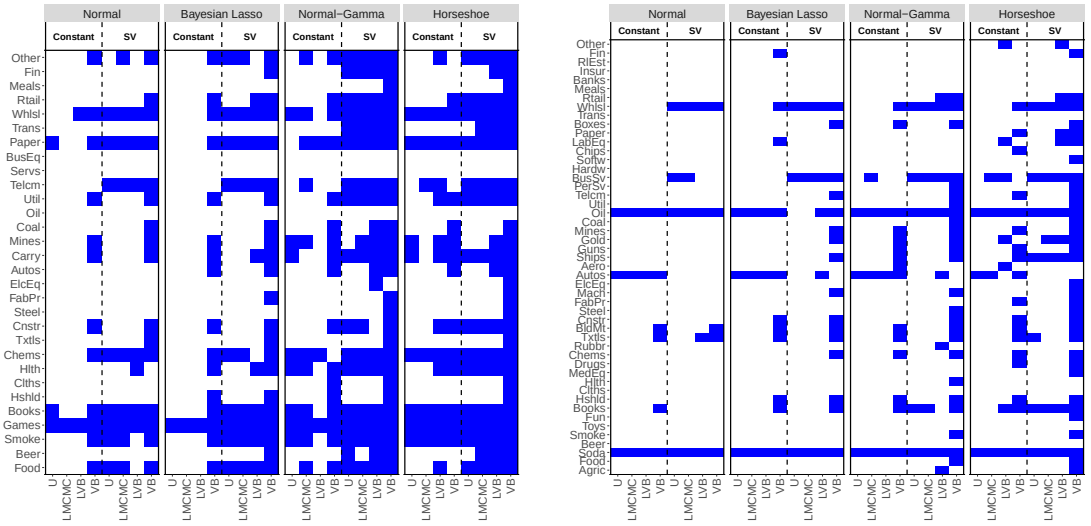
(d) Portfolios for which $\text{ALS}_j\left(\mathcal{M}_s\right) > 0$

Figure 7: Left panels report the log-score differential across industry portfolios. Right panels report the industries for which a given model can generate positive log-score differential. The top (bottom) panels report the results for 30 (49) industry portfolios.

volatility generate positive log-score differentials for almost all of the portfolios for the 30 industry classification and for more than half of the 49 industry portfolios. Interestingly, when it comes to density forecasts rather than modeling expected returns, the Gefang et al. (2023) variational method built on a structural VAR representation performs on par with our `VB` method. This is likely due to stochastic volatility alone, since our `VB` still stands out

within the constant volatility specifications. More generally, our VB approach outperforms the competing estimation methods under all prior specifications.

Returns predictability over the business cycle. Existing literature suggests that expected returns are counter-cyclical and that returns predictability is more concentrated during period of economic contractions vs expansions (see, e.g., Rapach et al., 2010). Thus, we investigate if the forecasting performance of our modeling framework changes over the business cycle. 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 and/or forecasting process. We compute the corresponding $R^2_{j,oos}(\mathcal{M}_s)$ for the recession periods only.



(a) $R^2_{j,oos}(\mathcal{M}_s) > 0$ for 30-industry portfolios (b) $R^2_{j,oos}(\mathcal{M}_s) > 0$ for 49-industry portfolios

Figure 8: The figure reports the industries for which $R^2_{j,oos}(\mathcal{M}_s) > 0$. The left (right) panel report the results for 30 (49) industry portfolios.

Figure 8 reports the industries for which $R^2_{j,oos}(\mathcal{M}_s) > 0$ for both the 30 (left panel) and the 49 (right panel) industry classification. The corresponding cross-sectional distribution of the $R^2_{j,oos}(\mathcal{M}_s)$ and the relative log-scores are reported in Appendix E.3. The labeling of Figure 8 is the same as in Figure 6. By comparing Figure 8 with the results for the full sample, it suggests that the accuracy of the predictions substantially improves across methods

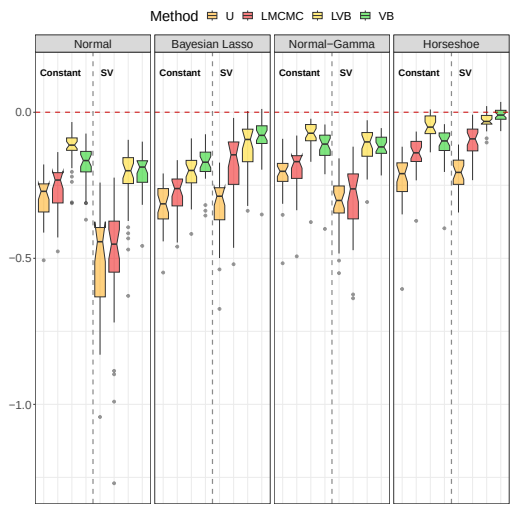
and priors. Nevertheless, our VB method outperforms the naive forecast from the rolling mean for a larger fraction of industry portfolios compared to other methods, in particular when stochastic volatility is considered. The difference between the recession and the full-sample performance persists when considering the 49 industry classification, especially for the adaptive Normal-Gamma and the Horseshoe prior.

5.3 Economic evaluation

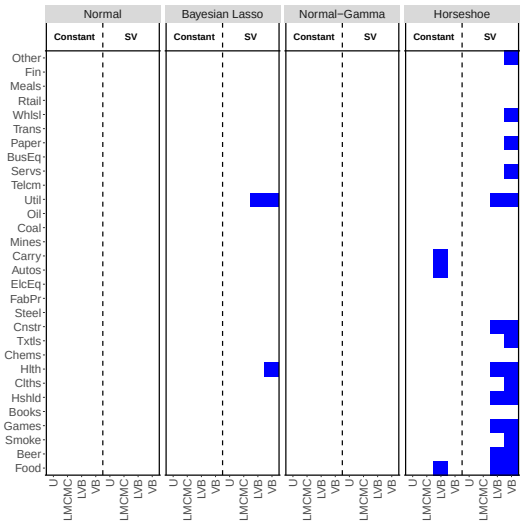
A positive predictive performance does not necessarily translate into economic value. However, in practice an investor is obviously keenly interested in the economic value of returns predictability, perhaps even more than the statistical performance. Hence, it is of paramount importance to evaluate the 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), we consider a representative investor with a single-period horizon and mean-variance preferences who allocates her wealth between an industry portfolio and a risk-free asset. Thus, the investor optimal allocation to stocks for period $t+1$ based on information at time t is given by $w_{jt} = \frac{1}{\gamma}\frac{\widehat{y}_{jt}}{\widehat{\nu}_{jt}^{-1}}$, where $\widehat{y}_{jt}$ represents the returns conditional mean forecast for industry $j = 1, \ldots, d$ and $\widehat{\nu}_{jt}^{-1}$ the corresponding volatility forecast at time t. We also constraint the weights for each of the industry to $-0.5 \leq w_{jt} \leq 1.5$ to prevent extreme short-sales and leverage positions. We assume a risk aversion coefficient of $\gamma = 5$ (see, e.g., Dangl and Halling, 2012).

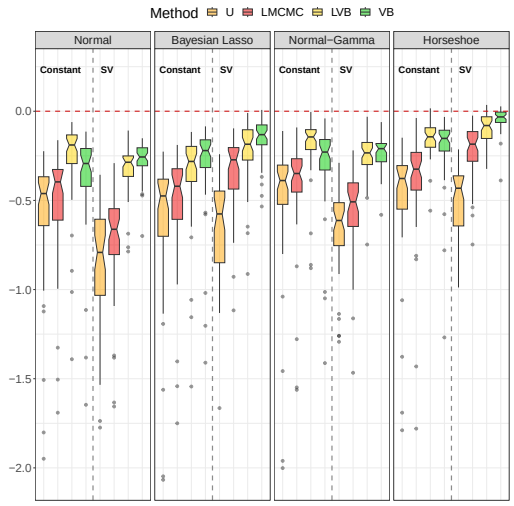
Figure 9 reports the average utility gain – in monthly % – obtained by using a given forecast $\widehat{y}_{jt}$ instead of the recursive sample mean $\overline{y}_{jt}$. The average utility for a given model is calculated as $\widehat{u}_j = \overline{r}_j - 0.5\gamma\overline{\sigma}_j^2$ where $\overline{r}_j$ and $\overline{\sigma}_j^2$ represent the sample mean and variance, respectively, of the portfolio return $r_{jt+1} = w_{jt}y_{jt+1}$ realized over the forecasting period for the industry $j = 1, \ldots, d$ under a given prior specification and estimation method. The utility



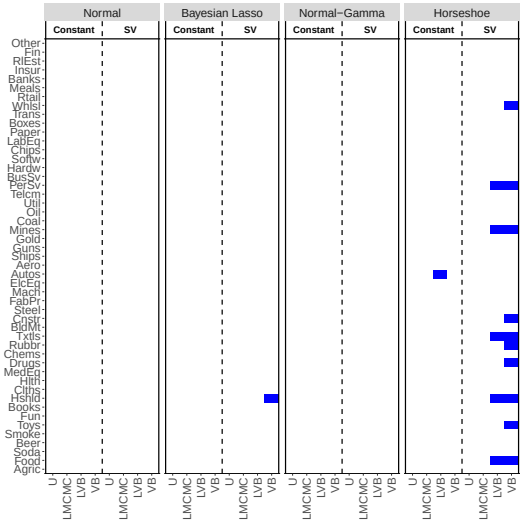
(a) Gain($\mathcal{M}_s$) for 30-industry classification



(b) Gain $(\mathcal{M}_s) > 0$ across 30 industries



(c) Gain($\mathcal{M}_s$) for 49-industry classification



(d) Gain $(\mathcal{M}_s) > 0$ across 49 industries

Figure 9: The left panel reports the cross-sectional distribution of the average utility gain across industry portfolios. The right panel reports the industries for which the utility gain is positive. The top (bottom) panels report the results for the 30-industry (49-industry) classification.

gain is calculated by subtracting the average utility of a given model $\widehat{u}_j$ to the average utility obtained by using the naive forecast from the recursive mean and variance to calculate w_{jt}. A positive value for the utility gain indicates the fee that a risk-averse investor is willing to pay to access the investment strategy implied by $\mathcal{M}_s$.

The economic value of each forecast largely confirms the same evidence offered by the out-

of-sample statistical performance. From a pure economic standpoint, the forecast from a recursive mean are quite challenging to beat: we observe that the average utility gain is mostly negative, with the only exception of those provided by `VB` under an Horseshoe prior specification. Economically, the results show that a representative investor with mean-variance utility is willing to pay, on average, a monthly fee of almost 15 basis points monthly to access the strategy based on our variational inference with stochastic volatility. In addition, the right panels of Figure 9 show that the positive economic value obtained from our `VB` is more broadly spread across industries compared to alternative methods. This holds especially for the 30 industry classification, but also applies to the more granular 49 industry classification.

6 Concluding remarks

We propose a novel variational inference method for large Bayesian vector autoregressions (VAR) with exogenous predictors and stochastic volatility. Differently from most existing estimation methods for high-dimensional VAR models, our approach does not rely on a structural form representation. This allows a fast and accurate identification of the regression coefficients without leveraging on a standard Cholesky-based transformation of the parameter space. We show both in simulation and empirically that our estimation approach outperforms across different prior specifications, both statistically and economically, forecasts from existing benchmark estimation strategies, such as equivalent, non-linear MCMC algorithms (see, e.g., Gruber and Kastner, 2022) linearized MCMC (see, e.g., Cross et al., 2020) and linearized variational inference methods (see, e.g., Gefang et al., 2023).

References

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

M. Bernardi, D. Bianchi, and N. Bianco. Smoothing volatility targeting. *arXiv preprint arXiv:2212.07288*, 2022.

M. Bernardi, D. Bianchi, and N. Bianco. Dynamic variable selection in high-dimensional predictive regressions. *Working Paper*, 2023.

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.

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. Minnesota-type adaptive hierarchical priors for large bayesian vars. *International Journal of Forecasting*, 37(3):1212–1226, 2021.

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.

J. C. Chan, G. Koop, and X. Yu. Large order-invariant bayesian vars with stochastic volatility. *arXiv preprint arXiv:2111.07225*, 2021.

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.

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.

D. Gefang, G. Koop, and A. Poon. Forecasting using variational bayesian inference in large vector autoregressions with hierarchical shrinkage. *International Journal of Forecasting*, 39(1):346–363, 2023.

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

L. Gruber and G. Kastner. Forecasting macroeconomic data with bayesian VARs: Sparse or dense? it depends! *arXiv preprint arXiv:2206.04902*, 2022.

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.

K. Hou and D. T. Robinson. Industry concentration and average stock returns. *The journal of finance*, 61(4):1927–1956, 2006.

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

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.

G. Kastner and F. Huber. Sparse bayesian vector autoregressions in huge dimensions. *Journal of Forecasting*, 39(7):1142–1165, 2020.

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.

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.

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. Rapach and G. Zhou. Forecasting stock returns. In *Handbook of economic forecasting*, volume 2, pages 328–383. Elsevier, 2013.

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.

D. Rohde and M. P. Wand. Semiparametric mean field variational bayes: general principles and numerical issues. *The Journal of Machine Learning Research*, 17(1):5975–6021, 2016.

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.

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.

Supplementary Appendix of:

Variational inference for large Bayesian vector autoregressions

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 variational densities 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 $\boldsymbol{\mu}_{\vartheta_i} = \mathbb{E}(\boldsymbol{\vartheta}_i)$ *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* $tr\big\{\boldsymbol{\Sigma}_{\vartheta_i,\vartheta_j}\mathbf{A}\big\}$*. 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* $\boldsymbol{\Sigma}_\vartheta$*. 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 optimal variational densities

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 vector of log-volatility parameters* $\mathbf{h}_j = (h_{j,0},\ldots,h_{j,T})^\intercal$ *is equal to* $q^*(\mathbf{h}_j) \equiv \mathsf{N}_{T+1}(\boldsymbol{\mu}_{q(h_j)},\boldsymbol{\Sigma}_{q(h_j)})$*, where, for* $j=1,\ldots,d$*, the variational parameters* $(\boldsymbol{\mu}_{q(h_j)},\boldsymbol{\Sigma}_{q(h_j)})$ *are updated as:*

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

$$\boldsymbol{\mu}_{q(h_j)}^{new} = \boldsymbol{\mu}_{q(h_j)}^{new} + \boldsymbol{\Sigma}_{q(h_j)}^{new}\nabla_{\boldsymbol{\mu}_{q(h_j)}} S(\boldsymbol{\mu}_{q(h_j)}^{old},\boldsymbol{\Sigma}_{q(h_j)}^{old}), \tag{B.2}$$

where $\nabla_{\boldsymbol{\mu}} S(\boldsymbol{\mu}^{old},\boldsymbol{\Sigma}^{old})$ *and* $\nabla^2_{\boldsymbol{\mu},\boldsymbol{\mu}} S(\boldsymbol{\mu}^{old},\boldsymbol{\Sigma}^{old})$ *denote the first and second derivative of* $S(\boldsymbol{\mu},\boldsymbol{\Sigma})$ *with respect to* $\boldsymbol{\mu}$ *and evaluated at* $(\boldsymbol{\mu}^{old},\boldsymbol{\Sigma}^{old})$*. The function* S *is the so called non-entropy*

function which is given by $\mathbb{E}_q(\log p(\mathbf{h}_j, \boldsymbol{\xi}_{-h_j}, \mathbf{y}_j))$. *In our scenario, we have that*

$$S(\boldsymbol{\mu}_{q(h_j)}, \boldsymbol{\Sigma}_{q(h_j)}) = -\frac{1}{2}[0, \boldsymbol{\iota}_n^\intercal]\boldsymbol{\mu}_{q(h_j)} - \frac{1}{2}[0, \boldsymbol{\mu}_{q(\boldsymbol{\varepsilon}_j^2)}^\intercal]\mathrm{e}^{-\boldsymbol{\mu}_{q(h_j)}+\frac{1}{2}\boldsymbol{\sigma}^2_{q(\mathbf{h}_j)}} - \frac{1}{2}\mu_{q(1/\psi_j)}\boldsymbol{\mu}_{q(h_j)}\mathbf{Q}\boldsymbol{\mu}_{q(h_j)} - \frac{1}{2}\mu_{q(1/\psi_j)}\mathsf{tr}\{\boldsymbol{\Sigma}_{q(\mathbf{h}_j)}\mathbf{Q}\}, \tag{B.3}$$

where $\boldsymbol{\sigma}^2_{q(h_j)} = \mathsf{diag}(\boldsymbol{\Sigma}_{q(h_j)})$ *is the vector of variances. In addition:*

$$\nabla_{\boldsymbol{\mu}_{q(h_j)}} S(\boldsymbol{\mu}_{q(h_j)}, \boldsymbol{\Sigma}_{q(h_j)}) = -\frac{1}{2}[0, \boldsymbol{\iota}_n^\intercal]^\intercal + \frac{1}{2}[0, \boldsymbol{\mu}_{q(\boldsymbol{\varepsilon}_j^2)}^\intercal]^\intercal \odot \mathrm{e}^{-\boldsymbol{\mu}_{q(h_j)}+\frac{1}{2}\boldsymbol{\sigma}^2_{q(h_j)}} - \mu_{q(1/\psi_j)}\mathbf{Q}\boldsymbol{\mu}_{q(h_j)}, \tag{B.4}$$

$$\nabla^2_{\boldsymbol{\mu}_{q(h_j)}\boldsymbol{\mu}_{q(h_j)}} S(\boldsymbol{\mu}_{q(h_j)}, \boldsymbol{\Sigma}_{q(h_j)}) = -\frac{1}{2}\mathsf{Diag}\left[[0, \boldsymbol{\mu}_{q(\boldsymbol{\varepsilon}_j^2)}^\intercal]^\intercal \odot \mathrm{e}^{-\boldsymbol{\mu}_{q(h_j)}+\frac{1}{2}\boldsymbol{\sigma}^2_{q(h_j)}}\right] - \mu_{q(1/\psi_j)}\mathbf{Q}, \tag{B.5}$$

where $\boldsymbol{\iota}_n$ *is an n-dimensional vector of ones,* $\mu_{q(1/\psi_j)}$ *is the variational mean of* $1/\psi_j$, $\mathbf{Q}$ *is the precision matrix associated to the random walk process with initial state* $h_0 \sim \mathsf{N}(0, k_0\,\psi_j)$, *and* $\odot$ *denotes the Hadamard product. Moreover,* $\boldsymbol{\mu}_{q(\boldsymbol{\varepsilon}_j^2)} = (\mu_{q(\varepsilon^2_{j,1})}, \ldots, \mu_{q(\varepsilon^2_{j,T})})^\intercal$, *with elements* $\mu_{q(\varepsilon^2_{j,t})} = \mathbb{E}_q\left[\varepsilon^2_{j,t}\right]$:

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

where $\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})} = \mathbf{y}_t^j - \boldsymbol{\mu}_{q(\boldsymbol{\Theta}^j)}\mathbf{z}_{t-1}$, *and, for* $i = 1, \ldots, j-1$ *and* $k = 1, \ldots, j-1$, *the elements in the matrix* $\mathbf{K}_{\boldsymbol{\vartheta},t}$ *and in the row vector* $\mathbf{k}_{\boldsymbol{\vartheta},t}$ *are* $[\mathbf{K}_{\vartheta,t}]_{i,k} = tr\left\{Cov(\boldsymbol{\vartheta}_i, \boldsymbol{\vartheta}_k)\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal\right\}$ *and* $[\mathbf{k}_{\vartheta,t}]_i = tr\left\{Cov(\boldsymbol{\vartheta}_i, \boldsymbol{\vartheta}_j)\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal\right\}$ *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:

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

and recall that $h_{j,t} = h_{j,t-1} + e_{j,t}$ with $e_{j,t} \sim \mathsf{N}(0, \psi_j)$ and initial state $h_0 \sim \mathsf{N}(0, k_0\,\psi_j)$. Define $\varepsilon_{j,t} = y_{j,t} - \boldsymbol{\beta}_j\mathbf{r}_{j,t} - \boldsymbol{\vartheta}_j\mathbf{z}_{t-1}$ and $\mathbf{h}_j = (h_{j,0}, \ldots, h_{j,T})^\intercal$. Recall that the random walk can be jointly represented as a Gaussian Markov random field $\mathbf{h}_j \sim \mathsf{N}_{T+1}(0, \psi\mathbf{Q}^{-1})$ with

tri-diagonal precision matrix $\mathbf{Q}^{-1}$. Compute $\log p(\mathbf{h}_j, \boldsymbol{\xi}_{-h_j}, \mathbf{y}_j) \propto \ell_j(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}) + \log p(\mathbf{h}_j)$:

$$\log p(\mathbf{h}_j, \boldsymbol{\xi}_{-h_j}, \mathbf{y}_j) \propto -\frac{1}{2}\sum_{t=1}^{T} h_{j,t} - \frac{1}{2}\sum_{t=1}^{T} \varepsilon_{j,t}^2 \mathrm{e}^{-h_{j,t}} - \frac{1}{2\psi_j}\mathbf{h}_j \mathbf{Q} \mathbf{h}_j.$$

Notice that the latter cannot be recognized as the kernel of a known distribution for $\mathbf{h}_j$, therefore complicating the computations. To overcome this issue we exploit the parametric variational Bayes paradigm and impose a Gaussian approximation $\mathbf{h}_j \sim \mathsf{N}(\boldsymbol{\mu}_{q(h_j)}, \boldsymbol{\Sigma}_{q(h_j)})$ similarly to Bernardi et al. (2022). Then, we follow Rohde and Wand (2016) to implement an iterative updating scheme to derive the optimal values of $(\boldsymbol{\mu}_{q(h_j)}, \boldsymbol{\Sigma}_{q(h_j)})$. To this aim, define the *non-entropy function* S as $\mathbb{E}_q(\log p(\mathbf{h}_j, \boldsymbol{\xi}_{-h_j}, \mathbf{y}_j))$:

$$\begin{aligned} S(\boldsymbol{\mu}_{q(h_j)}, \boldsymbol{\Sigma}_{q(h_j)}) = &-\frac{1}{2}[0, \boldsymbol{\iota}_n^\intercal]\boldsymbol{\mu}_{q(h_j)} - \frac{1}{2}[0, \boldsymbol{\mu}_{q(\boldsymbol{\varepsilon}_j^2)}^\intercal]\mathrm{e}^{-\boldsymbol{\mu}_{q(h_j)} + \frac{1}{2}\boldsymbol{\sigma}^2_{q(\mathbf{h}_j)}} \\ &- \frac{1}{2}\mu_{q(1/\psi_j)}\boldsymbol{\mu}_{q(h_j)}\mathbf{Q}\boldsymbol{\mu}_{q(h_j)} - \frac{1}{2}\mu_{q(1/\psi_j)}\mathsf{tr}\{\boldsymbol{\Sigma}_{q(\mathbf{h}_j)}\mathbf{Q}\}, \end{aligned} \tag{B.6}$$

where we exploit a vector representation of the likelihood term and $\boldsymbol{\sigma}^2_{q(h_j)} = \mathsf{diag}(\boldsymbol{\Sigma}_{q(h_j)})$ is the vector of variances. Moreover each element in the vector $\boldsymbol{\mu}_{q(\boldsymbol{\varepsilon}_j^2)}$, namely $\mu_{q(\varepsilon^2_{j,t})} = \mathbb{E}_q\left[\varepsilon^2_{j,t}\right]$ is given by:

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

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

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

Notice that for the latter derivation we use Results 1 and 2. □

Proposition B.1.2. *The optimal variational density for the vector of time-varying precision parameters* $\boldsymbol{\nu}_j = (\nu_{j,1},\ldots,\nu_{j,T})^\intercal$ *is equal to* $q^*(\boldsymbol{\nu}_j) \equiv \mathsf{logN}_T(-\boldsymbol{\mu}_{q(h_j)}, \boldsymbol{\Sigma}_{q(h_j)})$*, where, for each* $j = 1,\ldots,d$*:*

$$
\begin{aligned}
\mathbb{E}_q[\nu_t] &= \exp\{-\mu_{q(h_{j,t})} + 1/2\sigma_{q(h_{j,t})}^2\}, \\
\mathsf{Var}_q[\nu_t] &= \exp\{-2\mu_{q(h_{j,t})} + \sigma_{q(h_{j,t})}^2\}(\exp\{\sigma_{q(h_{j,t})}^2\} - 1), \\
\mathsf{Cov}_q[\nu_t,\nu_{t+1}] &= \exp\{-\mu_{q(h_{j,t})} - \mu_{q(h_{j,t+1})} + 1/2(\sigma_{q(h_{j,t})}^2 + \sigma_{q(h_{j,t+1})}^2)\}(\exp\{\mathsf{Cov}_q[h_t,h_{t+1}]\} - 1).
\end{aligned} \tag{B.7}
$$

Proof. The proof immediately follows from the fact that $\nu_{j,t} = \mathrm{e}^{-h_{j,t}}$ for $t = 1,\ldots,T$ and the distribution of $\mathbf{h}_j$ is Gaussian, as defined in Proposition B.1.1. □

Proposition B.1.3. *The optimal variational density for the constant precision parameter (homoskedastic modeling)* ν_j *is equal to* $q^*(\nu_j) \equiv \mathsf{Ga}(a_{q(\nu_j)}, b_{q(\nu_j)})$*, where, for* $j = 1,\ldots,d$*:*

$$
a_{q(\nu_j)} = a_\nu + T/2, \quad b_{q(\nu_j)} = b_\nu + \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}_{-\nu_j}\left[\varepsilon_{j,t}^2\right], \tag{B.8}
$$

where $\mathbb{E}_{-\nu_j}\left[\varepsilon_{j,t}^2\right]$ *is defined in Proposition B.1.1.*

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 the computations for $\mathbb{E}_{-\nu_j}\left[\varepsilon_{j,t}^2\right]$ have been previously considered in the Proof of Proposition B.1.1. Take the exponential of the latter equation, and notice that it is the kernel of a gamma random variable $\mathsf{Ga}(a_{q(\nu_j)}, b_{q(\nu_j)})$ as defined in Proposition B.1.3. □

Proposition B.1.4. *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:*

$$\begin{aligned}
\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)} &= \left(\sum_{t=1}^{T}\mu_{q(\nu_{j,t})}\left(\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}^{\intercal} + \mathbf{K}_{\vartheta,t}\right) + 1/\tau\mathbf{I}_{j-1}\right)^{-1}, \\
\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)} &= \boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}\sum_{t=1}^{T}\mu_{q(\nu_{j,t})}\left(\boldsymbol{\mu}_{q(\mathbf{r}_{j,t})}(y_{j,t} - \boldsymbol{\mu}_{q(\vartheta_j)}\mathbf{z}_{t-1})^{\intercal} + \mathbf{k}_{\vartheta,t}\right).
\end{aligned} \tag{B.9}$$

The optimal variational density for the parameter $\boldsymbol{\beta}_j$ *under homoskedastic assumption is obtained by substituting* $\mu_{q(\nu_{j,t})}$ *by* $\mu_{q(\nu_j)}$ *in the latter equations.*

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

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

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

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

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

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

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

Take the exponential and notice that the latter is the kernel of a Gaussian random variable $\mathsf{N}_{j-1}(\boldsymbol{\mu}_{q(\boldsymbol{\beta}_j)}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)})$, as defined in Proposition B.1.4. □

Proposition B.1.5. *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:*

$$\boldsymbol{\Sigma}_{q(\vartheta)} = \left(\sum_{t=1}^{T}(\boldsymbol{\mu}_{q(\boldsymbol{\Omega}_t)} \otimes \mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal) + 1/\upsilon\mathbf{I}_{d(d+p+1)}\right)^{-1}, \quad \boldsymbol{\mu}_{q(\vartheta)} = \boldsymbol{\Sigma}_{q(\vartheta)}\sum_{t=1}^{T}\left(\boldsymbol{\mu}_{q(\boldsymbol{\Omega}_t)} \otimes \mathbf{z}_{t-1}\right)\mathbf{y}_t, \tag{B.10}$$

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

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

The optimal variational density for the parameter $\boldsymbol{\vartheta}$ under homoskedastic assumption is obtained by substituting $\boldsymbol{\mu}_{q(\boldsymbol{\Omega}_t)}$ by $\boldsymbol{\mu}_{q(\boldsymbol{\Omega})} = (\mathbf{I}_d - \boldsymbol{\mu}_{q(\mathbf{B})})^\intercal\boldsymbol{\mu}_{q(\mathbf{V})}(\mathbf{I}_d - \boldsymbol{\mu}_{q(\mathbf{B})}) + \mathbf{C}_\vartheta$ and $\mathbf{C}_\vartheta$ is a constant $d \times d$ 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 $\boldsymbol{\varepsilon}_t \sim \mathsf{N}_d(0, \mathbf{V}_t^{-1})$ and then apply the vectorisation operation on the transposed and get:

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

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

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

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

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

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

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

hence

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

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

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

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.1.5. □

Proposition B.1.6. *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 $j = 1,\dots,d$ of $\boldsymbol{\Theta}$:*

$$\begin{aligned}
\boldsymbol{\Sigma}_{q(\vartheta_j)} &= \left(\sum_{t=1}^{T}\boldsymbol{\mu}_{q(\omega_{j,j,t})}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal + 1/\upsilon\mathbf{I}_{d+p+1}\right)^{-1}, \\
\boldsymbol{\mu}_{q(\vartheta_j)} &= \boldsymbol{\Sigma}_{q(\vartheta_j)}\left(\sum_{t=1}^{T}\left(\boldsymbol{\mu}_{q(\omega_{j,t})}\otimes\mathbf{z}_{t-1}\right)\mathbf{y}_t - \sum_{t=1}^{T}\left(\boldsymbol{\mu}_{q(\omega_{j,-j,t})}\otimes\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal\right)\boldsymbol{\mu}_{q(\vartheta_{-j})}\right).
\end{aligned} \tag{B.11}$$

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. Again, the homoskedastic scenario is recovered with constant elements $\boldsymbol{\mu}_{q(\omega_{j,j})}$, $\boldsymbol{\mu}_{q(\omega_j)}$, and $\boldsymbol{\mu}_{q(\omega_{j,-j})}$.

Proof. Consider the setting as in Proposition B.1.5, define $\boldsymbol{\mu}_{q(\boldsymbol{\Omega}_t)} = \mathbb{E}_{-\vartheta}\left[(\mathbf{I}_d - \mathbf{B})^\intercal\mathbf{V}_t(\mathbf{I}_d - \mathbf{B})\right]$ the expectation of the precision matrix and compute the optimal variational density for the

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

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

Where we used the following partitions:

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

and we denote with $\boldsymbol{\omega}_{j,t}$ the j-th row of $\boldsymbol{\Omega}_t$. 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.1.6. □

Proposition B.1.7. *The optimal variational density for the conditional variance parameter ψ_j is an inverse-gamma distribution $q(\psi_j) \equiv \mathsf{InvGa}(A_{q(\psi_j)}, B_{q(\psi_j)})$, where:*

$$\begin{aligned}
A_{q(\psi_j)} &= A_\psi + \frac{n+1}{2} \\
B_{q(\psi_j)} &= B_\psi + \frac{1}{2}\boldsymbol{\mu}_{q(\mathbf{h}_j)}^\intercal\mathbf{Q}\boldsymbol{\mu}_{q(\mathbf{h}_j)} + \frac{1}{2}\mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\mathbf{h}_j)}\mathbf{Q}\right\},
\end{aligned} \tag{B.12}$$

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

Proof. Recall that a priori $\psi_j \sim \mathsf{InvGa}(A_\psi, B_\psi)$ and compute the optimal variational density as $\log q^*(\psi_j) \propto \mathbb{E}_{-\psi_j}\left[\log p(\mathbf{h}_j|\psi_j) + \log p(\psi_j)\right]$:

$$\begin{aligned}
\log q(\eta^2) &\propto \mathbb{E}_{-\psi_j}\left[-\frac{n+1}{2}\log\psi_j - \frac{1}{2\psi_j}\mathbf{h}_j^\intercal\mathbf{Q}\mathbf{h}_j - (A_\psi+1)\log\psi_j - B_\psi/\psi_j\right] \\
&\propto -\left(A_\psi + \frac{n+1}{2} + 1\right)\log\psi_j - \frac{1}{\psi_j}\left(B_\psi + \frac{1}{2}\mathbb{E}_{h_j}\left[\mathbf{h}_j^\intercal\mathbf{Q}\mathbf{h}_j\right]\right),
\end{aligned}$$

where

$$\mathbb{E}_{h_j}\left[\mathbf{h}_j^\intercal \mathbf{Q}\mathbf{h}_j\right] = \boldsymbol{\mu}_{q(h_j)}^\intercal \mathbf{q}\boldsymbol{\mu}_{q(h_j)} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(h_j)}\mathbf{Q}\right\}.$$

Take the exponential and end up with the kernel of an inverse gamma distribution with parameters as in (B.12). □

In what follows we derive analytically the variational lower bound. Notice that we consider the case of joint approximation $q(\boldsymbol{\vartheta})$, since it represents the more general case, while the lower bound under the further restriction $q(\boldsymbol{\vartheta}) = \prod_{j=1}^{d} q(\boldsymbol{\vartheta}_j)$ can be recovered assuming a block-diagonal structure of $\boldsymbol{\Sigma}_{q(\vartheta)}$ in (B.13) and (B.15).

Proposition B.1.8. *The variational lower bound for the non-sparse homoskedastic multivariate regression model can be derived analytically and it is equal to:*

$$\begin{aligned}
\log \underline{p}(\mathbf{y}; q) = d\left(-\frac{T}{2}\log 2\pi + a_\nu \log b_\nu - \log\Gamma(a_\nu)\right) - \sum_{j=1}^{d}\left(a_{q(\nu_j)}\log b_{q(\nu_j)} - \log\Gamma(a_{q(\nu_j)})\right) \\
- \frac{1}{2}\sum_{j=2}^{d}\sum_{k=1}^{j-1}\left(\log\tau + 1/\tau\mu_{q(\beta_{j,k}^2)}\right) + \frac{1}{2}\sum_{j=2}^{d}\left(\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}| + (j-1)\right) \\
- \frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{d+p+1}\left(\log\upsilon + 1/\upsilon\mu_{q(\vartheta_{j,k}^2)}\right) + \frac{1}{2}\left(\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\vartheta})}| + d(d+p+1)\right).
\end{aligned} \tag{B.13}$$

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:

$$\begin{aligned}
\log \underline{p}(\mathbf{y}; q) &= \sum_{j=1}^{d} \mathbb{E}_q \left[\ell_j(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}) + \log p(\nu_j) - \log q(\nu_j) \right] \\
&\quad + \sum_{j=2}^{d} \mathbb{E}_q \left[\log p(\boldsymbol{\beta}_j) - \log q(\boldsymbol{\beta}_j) \right] + \mathbb{E}_q \left[\log p(\boldsymbol{\vartheta}) - \log q(\boldsymbol{\vartheta}) \right] \\
&= \sum_{j=1}^{d} \Big(\underbrace{\mathbb{E}_q \left[\ell_j(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}) + \log \underline{p}(\mathbf{y}; \nu_j) \right]}_{A} + \sum_{j=2}^{d} \underbrace{\mathbb{E}_q \left[\log \underline{p}(\mathbf{y}; \boldsymbol{\beta}_j) \right]}_{B} + \underbrace{\mathbb{E}_q \left[\log \underline{p}(\mathbf{y}; \boldsymbol{\vartheta}) \right]}_{C},
\end{aligned} \tag{B.14}$$

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.1.3. The

second term to compute is equal to:

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

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.14) and notice that the variational lower bound here computed coincides with the one presented in Proposition B.1.8. □

Proposition B.1.9. *The variational lower bound for the non-sparse multivariate regression model with stochastic volatility can be derived analytically and it is equal to:*

$$
\begin{aligned}
\log \underline{p}(\mathbf{y}; q) = d \left(-\frac{T}{2} \log 2\pi + \frac{T+1}{2} - \frac{1}{2} \log k_0 + a_\psi \log b_\psi - \log \Gamma(a_\psi) \right) \\
+ \frac{1}{2} \sum_{j=1}^{d} \sum_{t=1}^{T} \mu_{q(h_{j,t})} - \frac{1}{2} \sum_{j=1}^{d} \sum_{t=1}^{T} \exp(-\mu_{q(h_{j,t})} + 1/2\sigma_{q(h_{j,t})}^2) \mathbb{E}_q \left[\varepsilon_{j,t}^2 \right] \\
+ \frac{1}{2} \sum_{j=1}^{d} \log |\boldsymbol{\Sigma}_{q(h_j)}| - \sum_{j=1}^{d} \left(a_{q(\psi_j)} \log b_{q(\psi_j)} - \log \Gamma(a_{q(\psi_j)}) \right) \\
- \frac{1}{2} \sum_{j=2}^{d} \sum_{k=1}^{j-1} \left(\log \tau + 1/\tau \mu_{q(\beta_{j,k}^2)} \right) + \frac{1}{2} \sum_{j=2}^{d} \left(\log |\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}| + (j-1) \right) \\
- \frac{1}{2} \sum_{j=1}^{d} \sum_{k=1}^{d+p+1} \left(\log \upsilon + 1/\upsilon \mu_{q(\vartheta_{j,k}^2)} \right) + \frac{1}{2} \left(\log |\boldsymbol{\Sigma}_{q(\boldsymbol{\vartheta})}| + d(d+p+1) \right).
\end{aligned}
\tag{B.15}
$$

Proof. Under the heteroskedastic model specification, we have that

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

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

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

Similarly for the variational density we have:

$$\log q(\boldsymbol{\xi}) = \sum_{j=1}^{d} (\log q(\mathbf{h}_j) + \log q(\psi_j)) + \sum_{j=2}^{d} \log q(\boldsymbol{\beta}_j) + \log q(\boldsymbol{\vartheta}),$$

and the lower bound can be divided into terms referring to each parameter:

$$\begin{aligned}
\log \underline{p}(\mathbf{y}; q) &= \sum_{j=1}^{d} \mathbb{E}_q \left[\ell_j(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}) + \log p(\mathbf{h}_j) - \log q(\mathbf{h}_j) + \log p(\psi_j) - \log q(\psi_j)\right] \\
&\quad + \sum_{j=2}^{d} \mathbb{E}_q \left[\log p(\boldsymbol{\beta}_j) - \log q(\boldsymbol{\beta}_j)\right] + \mathbb{E}_q \left[\log p(\boldsymbol{\vartheta}) - \log q(\boldsymbol{\vartheta})\right] \\
&= \sum_{j=1}^{d} \Big(\underbrace{\mathbb{E}_q \left[\ell_j(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}) + \log \underline{p}(\mathbf{y}; \mathbf{h}_j) + \log \underline{p}(\mathbf{y}; \psi_j)\right]}_{A} \\
&\quad + \sum_{j=2}^{d} \underbrace{\mathbb{E}_q \left[\log \underline{p}(\mathbf{y}; \boldsymbol{\beta}_j)\right]}_{B} + \underbrace{\mathbb{E}_q \left[\log \underline{p}(\mathbf{y}; \boldsymbol{\vartheta})\right]}_{C},
\end{aligned} \tag{B.16}$$

thus our strategy will be to evaluate each piece in the latter separately and then put the results together. The terms B and C are the same computed for the homoskedastic model.

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

$$
\begin{aligned}
A = \mathbb{E}_q \left[-\frac{T}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^{T} h_{j,t} - \frac{1}{2} \sum_{t=1}^{T} \exp(-h_{j,t}) \left(y_{j,t} - \boldsymbol{\beta}_j \mathbf{r}_{j,t} - \boldsymbol{\vartheta}_j \mathbf{z}_{t-1} \right)^2 \right] \\
+ \mathbb{E}_q \left[-\frac{T+1}{2} \log 2\pi - \frac{T+1}{2} \log \psi_j + \frac{1}{2} \underbrace{\log |\mathbf{Q}|}_{= -\log k_0} - \frac{1}{2\psi_j} \mathbf{h}_j^\intercal \mathbf{Q} \mathbf{h}_j \right] \\
- \mathbb{E}_q \left[-\frac{T+1}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}_{q(h_j)}| - \frac{1}{2} \overbrace{(\mathbf{h}_j - \boldsymbol{\mu}_{q(h_j)})^\intercal \boldsymbol{\Sigma}_{q(h_j)}^{-1} (\mathbf{h}_j - \boldsymbol{\mu}_{q(h_j)})}^{\text{See Result 3}} \right] \\
+ \mathbb{E}_q \left[a_\psi \log b_\psi - \log \Gamma(a_\psi) - (a_\psi + 1) \log \psi_j - b_\psi / \psi_j \right] \\
- \mathbb{E}_q \left[a_{q(\psi_j)} \log b_{q(\psi_j)} - \log \Gamma(a_{q(\psi_j)}) - (a_{q(\psi_j)} + 1) \log \psi_j - b_{q(\psi_j)} / \psi_j \right] \\
= -\frac{T}{2} \log 2\pi + \frac{1}{2} \sum_{t=1}^{T} \mu_{q(h_{j,t})} - \frac{1}{2} \sum_{t=1}^{T} \exp(-\mu_{q(h_{j,t})} + 1/2\sigma^2_{q(h_{j,t})}) \mathbb{E}_q \left[\varepsilon_{j,t}^2 \right] \\
- \frac{T+1}{2} \mu_{q(\log \psi_j)} - \frac{1}{2} \log k_0 - \frac{1}{2} \mu_{q(1/\psi_j)} \mathbb{E}_{h_j} \left[\mathbf{h}_j \mathbf{Q} \mathbf{h}_j \right] + \frac{1}{2} \log |\boldsymbol{\Sigma}_{q(h_j)}| + \frac{T+1}{2} \\
+ a_\psi \log b_\psi - \log \Gamma(a_\psi) - (a_\psi + 1) \mu_{q(\log \psi_j)} - \mu_{q(1/\psi_j)} b_\psi \\
- a_{q(\psi_j)} \log b_{q(\psi_j)} + \log \Gamma(a_{q(\psi_j)}) + (a_{q(\psi_j)} + 1) \mu_{q(\log \psi_j)} + \mu_{q(1/\psi_j)} b_{q(\psi_j)} \\
= -\frac{T}{2} \log 2\pi + \frac{1}{2} \sum_{t=1}^{T} \mu_{q(h_{j,t})} - \frac{1}{2} \sum_{t=1}^{T} \exp(-\mu_{q(h_{j,t})} + 1/2\sigma^2_{q(h_{j,t})}) \mathbb{E}_q \left[\varepsilon_{j,t}^2 \right] + \frac{1}{2} \log |\boldsymbol{\Sigma}_{q(h_j)}| \\
+ \frac{T+1}{2} - \frac{1}{2} \log k_0 + a_\psi \log b_\psi - \log \Gamma(a_\psi) - a_{q(\psi_j)} \log b_{q(\psi_j)} + \log \Gamma(a_{q(\psi_j)}),
\end{aligned}
$$

where $\mathbb{E}_q \left[\varepsilon_{j,t}^2 \right]$ is defined in Proposition B.1.1, and to make some simplifications we exploit the definitions of $a_{q(\psi_j)}, b_{q(\psi_j)}$ given in Proposition B.1.7. Put together the terms A, B, C as in (B.16) and notice that the variational lower bound here computed coincides with the one presented in Proposition B.1.9. □

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.

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 $\mathbf{h}_j$, 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

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.8) (homoskedastic);

  Update $q^*(\mathbf{h}_1)$ and therefore $q^*(\boldsymbol{\nu}_1)$ as in (B.1) and (B.7) (heteroskedastic);

  Update $q^*(\psi_1)$ as in (B.12);

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

    Update $q^*(\nu_j)$ as in (B.8) (homoskedastic);

    Update $q^*(\mathbf{h}_j)$ and therefore $q^*(\boldsymbol{\nu}_j)$ as in (B.1) and (B.7) (heteroskedastic);

    Update $q^*(\psi_j)$ as in (B.12);

    Update $q^*(\boldsymbol{\beta}_j)$ as in (B.9);

  end

  Update $q^*(\boldsymbol{\vartheta})$ as in (B.10) or (B.11);

  Compute $\log \underline{p}(\mathbf{y}; q)$ as in (B.13) (homoskedastic) or (B.15) (heteroskedastic);

  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

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:*

$$\boldsymbol{\Sigma}_{q(\vartheta)} = \left(\sum_{t=1}^{T} \boldsymbol{\mu}_{q(\boldsymbol{\Omega}_t)} \otimes \mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal + \mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon)}) \right)^{-1}, \qquad \boldsymbol{\mu}_{q(\vartheta)} = \boldsymbol{\Sigma}_{q(\vartheta)} \sum_{t=1}^{T} \left(\boldsymbol{\mu}_{q(\boldsymbol{\Omega}_t)} \otimes \mathbf{z}_{t-1} \right) \mathbf{y}_t, \tag{B.17}$$

where $\mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon)})$ 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:

$$\begin{aligned} \boldsymbol{\Sigma}_{q(\vartheta_j)} &= \left(\sum_{t=1}^{T} \boldsymbol{\mu}_{q(\omega_{j,j,t})} \mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal + \mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon_j)}) \right)^{-1}, \\ \boldsymbol{\mu}_{q(\vartheta_j)} &= \boldsymbol{\Sigma}_{q(\vartheta_j)} \left(\sum_{t=1}^{T} \left(\boldsymbol{\mu}_{q(\omega_{j,t})} \otimes \mathbf{z}_{t-1} \right) \mathbf{y}_t - \sum_{t=1}^{T} \left(\boldsymbol{\mu}_{q(\omega_{j,-j,t})} \otimes \mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal \right) \boldsymbol{\mu}_{q(\vartheta_{-j})} \right), \end{aligned} \tag{B.18}$$

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

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

$$a_{q(1/\upsilon_{j,k})} = \mu_{q(\vartheta^2_{j,k})}, \quad b_{q(1/\upsilon_{j,k})} = \mu_{q(\lambda^2_{j,k})}. \tag{B.19}$$

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 $\upsilon_{j,k}$:

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

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

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

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:*

$$a_{q(\lambda^2_{j,k})} = h_1 + 1, \quad b_{q(\lambda^2_{j,k})} = \mu_{q(\upsilon_{j,k})}/2 + h_2. \tag{B.20}$$

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 $\log q^*(\lambda^2_{j,k}) \propto \mathbb{E}_{-\lambda^2_{j,k}}\left[\log p(\upsilon_{j,k}) + \log p(\lambda^2_{j,k})\right]$:

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

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:*

$$\begin{aligned}\log \underline{p}(\mathbf{y}; q) = \log \underline{p}^{SV}(\mathbf{y}; \boldsymbol{\beta}, \mathbf{h}, \boldsymbol{\psi}) \quad &\left(or\ \log \underline{p}^{C}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\nu})\ if\ homoskedastic\right) \\ &+ \frac{1}{2}\left(\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\vartheta})}| + d(d+p+1)\right) + \sum_{j=1}^{d}\sum_{k=1}^{d+p+1}\frac{1}{2}\mu_{q(\lambda^2_{j,k})}\mu_{q(\upsilon_{j,k})} \\ &- \sum_{j=1}^{d}\sum_{k=1}^{d+p+1}\left(1/4\log(b_{q(1/\upsilon_{j,k})}/a_{q(1/\upsilon_{j,k})}) - \log K_{1/2}(\sqrt{b_{q(1/\upsilon_{j,k})}a_{q(1/\upsilon_{j,k})}})\right) \\ &+ d(d+p+1)\left(h_1 \log h_2 - \log\Gamma(h_1)\right) - \sum_{j=1}^{d}\sum_{k=1}^{d+p+1}\left(a_{q(\lambda^2_{j,k})}\log b_{q(\lambda^2_{j,k})} - \log\Gamma(a_{q(\lambda^2_{j,k})})\right),\end{aligned} \tag{B.21}$$

where

$$\begin{aligned}\log \underline{p}^{C}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\nu}) &= d\left(-\frac{T}{2}\log 2\pi + a_\nu \log b_\nu - \log\Gamma(a_\nu)\right) - \sum_{j=1}^{d}\left(a_{q(\nu_j)}\log b_{q(\nu_j)} - \log\Gamma(a_{q(\nu_j)})\right) \\ &\quad - \frac{1}{2}\sum_{j=2}^{d}\sum_{k=1}^{j-1}\left(\log\tau + 1/\tau\mu_{q(\beta^2_{j,k})}\right) + \frac{1}{2}\sum_{j=2}^{d}\left(\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}| + (j-1)\right) \\ \log \underline{p}^{SV}(\mathbf{y}; \boldsymbol{\beta}, \mathbf{h}, \boldsymbol{\psi}) &= d\left(-\frac{T}{2}\log 2\pi + \frac{T+1}{2} - \frac{1}{2}\log k_0 + a_\psi \log b_\psi - \log\Gamma(a_\psi)\right) \\ &\quad + \frac{1}{2}\sum_{j=1}^{d}\sum_{t=1}^{T}\mu_{q(h_{j,t})} - \frac{1}{2}\sum_{j=1}^{d}\sum_{t=1}^{T}\exp(-\mu_{q(h_{j,t})} + 1/2\sigma^2_{q(h_{j,t})})\mathbb{E}_q\left[\varepsilon^2_{j,t}\right] \\ &\quad + \frac{1}{2}\sum_{j=1}^{d}\log|\boldsymbol{\Sigma}_{q(h_j)}| - \sum_{j=1}^{d}\left(a_{q(\psi_j)}\log b_{q(\psi_j)} - \log\Gamma(a_{q(\psi_j)})\right) \\ &\quad - \frac{1}{2}\sum_{j=2}^{d}\sum_{k=1}^{j-1}\left(\log\tau + 1/\tau\mu_{q(\beta^2_{j,k})}\right) + \frac{1}{2}\sum_{j=2}^{d}\left(\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}_j)}| + (j-1)\right)\end{aligned}$$

Proof. As we did in (B.14) for Proposition B.1.8, the lower bound can be divided into terms referring to each parameter:

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

where A is equal to (B.14) 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.1.8 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:

$$\begin{aligned} A = \log \underline{p}^{\text{SV}}(\mathbf{y}; \boldsymbol{\beta}, \mathbf{h}, \boldsymbol{\psi}) \quad & \left(\text{or } \log \underline{p}^{\text{C}}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\nu}) \text{ if homoskedastic} \right) \\ & - \frac{1}{2} \sum_{j=1}^{d} \sum_{k=1}^{d+p+1} \left(\mu_{q(\log \upsilon_{j,k})} + \mu_{q(1/\upsilon_{j,k})} \mu_{q(\vartheta^2_{j,k})} \right) + \frac{1}{2} \left(\log |\boldsymbol{\Sigma}_{q(\boldsymbol{\vartheta})}| + d(d+p+1) \right). \end{aligned} \tag{B.22}$$

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

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

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

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

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

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

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

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.

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.8) (homoskedastic);
  Update $q^*(\mathbf{h}_1)$ and therefore $q^*(\boldsymbol{\nu}_1)$ as in (B.1) and (B.7) (heteroskedastic);
  Update $q^*(\psi_1)$ as in (B.12);
  for $j = 2, \dots, d$ **do**
    Update $q^*(\nu_j)$ as in (B.8) (homoskedastic);
    Update $q^*(\mathbf{h}_j)$ and therefore $q^*(\boldsymbol{\nu}_j)$ as in (B.1) and (B.7) (heteroskedastic);
    Update $q^*(\psi_j)$ as in (B.12);
    Update $q^*(\boldsymbol{\beta}_j)$ as in (B.9);
  end
  Update $q^*(\boldsymbol{\vartheta})$ as in (B.17) or (B.18);
  for $j = 1, \dots, d$ **do**
    for $k = 1, \dots, d + p + 1$ **do**
      Update $q^*(\upsilon_{j,k})$, $q^*(\lambda^2_{j,k})$ as in (B.19)-(B.20);
    end
  end
  Compute $\log \underline{p}(\mathbf{y}; q)$ as in (B.21);
  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.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 $\mathbf{h}_j$, ν_j, and for the cholesky factor rows $\boldsymbol{\beta}_j$ 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 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 $k = 1, \ldots, d + p + 1$:*

$$\zeta_{q(\upsilon_{j,k})} = \mu_{q(\eta_j)} - 1/2, \quad a_{q(\upsilon_{j,k})} = \mu_{q(\eta_j)}\mu_{q(\lambda_{j,k})}, \quad b_{q(\upsilon_{j,k})} = \mu_{q(\vartheta^2_{j,k})}. \tag{B.23}$$

Moreover, it is useful to know that

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

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

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

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

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

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

where $\mu_{q(\vartheta^2_{j,k})} = \sigma^2_{q(\vartheta_{j,k})} + \mu^2_{q(\vartheta_{j,k})}$. 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:*

$$a_{q(\lambda_{j,k})} = \mu_{q(\eta_j)} + h_1, \quad b_{q(\lambda_{j,k})} = \frac{\mu_{q(\eta_j)}\mu_{q(\upsilon_{j,k})}}{2} + h_2. \tag{B.24}$$

Moreover, it is useful to know that

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

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

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

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

$$\begin{aligned} \log q^*(\lambda_{j,k}) &\propto \mathbb{E}_{-\lambda_{j,k}}\left[(\eta_j + h_1 - 1)\log \lambda_{j,k} - \lambda_{j,k}\left(\frac{\eta_j \upsilon_{j,k}}{2} + h_2\right)\right] \\ &\propto \left(\mu_{q(\eta_j)} + h_1 - 1\right)\log \lambda_{j,k} - \lambda_{j,k}\left(\frac{\mu_{q(\eta_j)}\mu_{q(\upsilon_{j,k})}}{2} + h_2\right), \end{aligned} \tag{B.25}$$

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:*

$$q^*(\eta_j) = \frac{h(\eta_j)}{c_{\eta_j}} \exp\left\{-\eta_j \sum_{k=1}^{d+p+1}\left(\frac{\mu_{q(\lambda_{j,k})}\mu_{q(\upsilon_{j,k})}}{2} - \mu_{q(\log \lambda_{j,k})} - \mu_{q(\log \upsilon_{j,k})} + \log 2 + h_3\right)\right\}, \tag{B.26}$$

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

$$\begin{aligned}\log q^*(\eta_j) &\propto \mathbb{E}_{-\eta_j}\left[(d+p+1)\left(\eta_j \log \eta_j - \log\Gamma(\eta_j)\right) - \eta_j \sum_{k=1}^{d+p+1}\left(\left(\frac{\lambda_{j,k}\upsilon_{j,k}}{2} - \log\frac{\lambda_{j,k}\upsilon_{j,k}}{2}\right) + h_3\right)\right] \\ &= (d+p+1)\left(\eta_j \log \eta_j - \log\Gamma(\eta_j)\right) \\ &\quad - \eta_j \sum_{k=1}^{d+p+1}\left(\frac{\mu_{q(\lambda_{j,k})}\mu_{q(\upsilon_{j,k})}}{2} - \mathbb{E}_{\upsilon_{j,k}\lambda_{j,k}}\left[\log\frac{\lambda_{j,k}\upsilon_{j,k}}{2}\right] + h_3\right),\end{aligned} \tag{B.27}$$

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.27) 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:*

$$\begin{aligned}\log \underline{p}(\mathbf{y}; q) &= \log \underline{p}^{SV}(\mathbf{y}; \boldsymbol{\beta}, \mathbf{h}, \boldsymbol{\psi}) \quad \left(or\ \log \underline{p}^{C}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\nu})\ if\ homoskedastic\right) \\ &\quad + \frac{1}{2}\left(\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\vartheta})}| + d(d+p+1)\right) - \sum_{j=1}^{d}\sum_{k=1}^{d+p+1} h(\zeta_{q(\upsilon_{j,k})}, a_{q(\upsilon_{j,k})}, b_{q(\upsilon_{j,k})}) \\ &\quad + d(d+p+1)\left(h_1 \log h_2 - \log\Gamma(h_1)\right) - \sum_{j=1}^{d}\sum_{k=1}^{d+p+1}\left(a_{q(\lambda_{j,k})}\log b_{q(\lambda_{j,k})} - \log\Gamma(a_{q(\lambda_{j,k})})\right) \\ &\quad + d\log h_3 + \sum_{j=1}^{d}\log c_{\eta_j} + \sum_{j=1}^{d}\mu_{q(\eta_j)}\sum_{k=1}^{d+p+1}\left(\mu_{q(\lambda_{j,k})}\mu_{q(\upsilon_{j,k})} - \mu_{q(\log\lambda_{j,k})} - \mu_{q(\log\upsilon_{j,k})}\right),\end{aligned} \tag{B.28}$$

where $\log \underline{p}^{SV}(\mathbf{y}; \boldsymbol{\beta}, \mathbf{h}, \boldsymbol{\psi})$ *and* $\log \underline{p}^{C}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\nu})$ *are defined in B.21.*

Proof. As we did in (B.14) for Proposition B.1.8, the lower bound can be divided into terms referring to each parameter:

$$\log \underline{p}(\mathbf{y}; q) = A + \sum_{j=1}^{d}\sum_{k=1}^{d+p+1}\Bigg(\underbrace{\mathbb{E}_q\left[\log \underline{p}(\mathbf{y}; \upsilon_{j,k})\right]}_{B} + \underbrace{\mathbb{E}_q\left[\log \underline{p}(\mathbf{y}; \lambda_{j,k})\right]}_{C} + \underbrace{\mathbb{E}_q\left[\log \underline{p}(\mathbf{y}; \eta_j)\right]}_{D}\Bigg), \tag{B.29}$$

where A is equal to (B.22). 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 $j = 1, \ldots, d$ and $k = 1, \ldots, d + p + 1$, is equal to:

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

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.

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.8) (homoskedastic);
- Update $q^*(\mathbf{h}_1)$ and therefore $q^*(\boldsymbol{\nu}_1)$ as in (B.1) and (B.7) (heteroskedastic);
- Update $q^*(\psi_1)$ as in (B.12);
- **for** $j = 2, \dots, d$ **do**
  - Update $q^*(\nu_j)$ as in (B.8) (homoskedastic);
  - Update $q^*(\mathbf{h}_j)$ and therefore $q^*(\boldsymbol{\nu}_j)$ as in (B.1) and (B.7) (heteroskedastic);
  - Update $q^*(\psi_j)$ as in (B.12);
  - Update $q^*(\boldsymbol{\beta}_j)$ as in (B.9);
- **end**
- Update $q^*(\boldsymbol{\vartheta})$ as in (B.17) or (B.18);
- **for** $j = 1, \dots, d$ **do**
  - **for** $k = 1, \dots, d + p + 1$ **do**
    - Update $q^*(\upsilon_{j,k})$, $q^*(\lambda_{j,k})$ as in (B.23)-(B.24);
  - **end**
  - Update $q^*(\eta_j)$ as in (B.26);
- **end**
- Compute $\log \underline{p}(\mathbf{y}; q)$ as in (B.28);
- 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.4 Horseshoe prior

First of all, notice that the optimal densities for $\mathbf{h}_j$, 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:*

$$\begin{aligned} \boldsymbol{\Sigma}_{q(\vartheta)} &= \left(\sum_{t=1}^{T} \boldsymbol{\mu}_{q(\boldsymbol{\Omega}_t)} \otimes \mathbf{z}_{t-1}\mathbf{z}_{t-1}^\intercal + \boldsymbol{\mu}_{q(1/\boldsymbol{\gamma}^2)} \mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon^2)}) \right)^{-1}, \\ \boldsymbol{\mu}_{q(\vartheta)} &= \boldsymbol{\Sigma}_{q(\vartheta)} \sum_{t=1}^{T} \left(\boldsymbol{\mu}_{q(\boldsymbol{\Omega}_t)} \otimes \mathbf{z}_{t-1} \right) \mathbf{y}_t, \end{aligned} \tag{B.30}$$

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

Under the row-independence assumption, the optimal variational density for the param-

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

$$\begin{aligned}
\boldsymbol{\Sigma}_{q(\vartheta_j)} &= \left(\sum_{t=1}^{T} \boldsymbol{\mu}_{q(\omega_{j,j,t})} \mathbf{z}_{t-1} \mathbf{z}_{t-1}^{\mathsf{T}} + \boldsymbol{\mu}_{q(1/\gamma^2)} \mathsf{Diag}(\boldsymbol{\mu}_{q(1/\upsilon_j^2)}) \right)^{-1}, \\
\boldsymbol{\mu}_{q(\vartheta_j)} &= \boldsymbol{\Sigma}_{q(\vartheta_j)} \left(\sum_{t=1}^{T} \left(\boldsymbol{\mu}_{q(\omega_{j,t})} \otimes \mathbf{z}_{t-1} \right) \mathbf{y}_t - \sum_{t=1}^{T} \left(\boldsymbol{\mu}_{q(\omega_{j,-j,t})} \otimes \mathbf{z}_{t-1} \mathbf{z}_{t-1}^{\mathsf{T}} \right) \boldsymbol{\mu}_{q(\vartheta_{-j})} \right),
\end{aligned} \tag{B.31}$$

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

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* $\upsilon_{j,k}^2$ *is equal to an inverse gamma distribution* $q^*(\upsilon_{j,k}^2) \equiv \mathsf{InvGa}(1, b_{q(\upsilon_{j,k}^2)})$*, where, for* $j = 1, \ldots, d$ *and* $k = 1, \ldots, d+p+1$*:*

$$b_{q(\upsilon_{j,k}^2)} = \mu_{q(1/\lambda_{j,k})} + \frac{1}{2} \mu_{q(\vartheta_{j,k}^2)} \mu_{q(1/\gamma^2)}. \tag{B.32}$$

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

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

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

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

Take the exponential and notice that the latter is the kernel of an inverse gamma random variable $\mathsf{InvGa}(1, b_{q(\upsilon_{j,k}^2)})$, 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:*

$$a_{q(\gamma^2)} = \frac{d(d+p+1)+1}{2}, \quad b_{q(\gamma^2)} = \mu_{q(1/\eta)} + \frac{1}{2} \sum_{j=1}^{d} \sum_{k=1}^{d+p+1} \mu_{q(1/\upsilon_{j,k}^2)} \mu_{q(\vartheta_{j,k}^2)}. \tag{B.33}$$

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

$$b_{q(\lambda_{j,k})} = 1 + \mu_{q(1/\upsilon^2_{j,k})}. \tag{B.34}$$

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

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

Take the exponential and notice that the latter is the kernel of an inverse gamma random variable $\mathsf{InvGa}(1, b_{q(\lambda_{j,k})})$, 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:*

$$b_{q(\eta)} = 1 + \mu_{q(1/\gamma^2)}. \tag{B.35}$$

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:*

$$\begin{aligned}\log \underline{p}(\mathbf{y}; q) = \log \underline{p}^{SV}(\mathbf{y}; \boldsymbol{\beta}, \mathbf{h}, \boldsymbol{\psi}) \quad &\left(\textit{or } \log \underline{p}^{C}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\nu}) \textit{ if homoskedastic}\right) \\ &+ \frac{1}{2}\left(\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\vartheta})}| + d(d+p+1)\right) + \mu_{q(1/\gamma^2)}\left(\mu_{q(1/\eta)} + \sum_{j=1}^{d}\sum_{k=1}^{d+p+1}\mu_{q(\vartheta^2_{j,k})}\mu_{q(1/\upsilon^2_{j,k})}\right) \\ &+ \sum_{j=1}^{d}\sum_{k=1}^{d+p+1}\left(\mu_{q(1/\upsilon^2_{j,k})}\mu_{q(1/\lambda_{j,k})} - \log b_{q(\upsilon^2_{j,k})} - \log b_{q(\lambda_{j,k})} - \log\pi\right) \\ &- a_{q(\gamma^2)}\log b_{q(\gamma^2)} - \log b_{q(\eta)} - \log\pi,\end{aligned} \tag{B.36}$$

where $\log \underline{p}^{SV}(\mathbf{y}; \boldsymbol{\beta}, \mathbf{h}, \boldsymbol{\psi})$ *and* $\log \underline{p}^{C}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\nu})$ *are defined in B.21.*

Proof. As we did in (B.14) for Proposition B.1.8, the lower bound can be divided into terms referring to each parameter:

$$\begin{aligned}\log \underline{p}(\mathbf{y}; q) = A + \underbrace{\mathbb{E}_q\left[\log \underline{p}(\mathbf{y}; \gamma^2)\right]}_{B} + \underbrace{\mathbb{E}_q\left[\log \underline{p}(\mathbf{y}; \eta)\right]}_{C} \\ + \sum_{j=1}^{d}\sum_{k=1}^{d+p+1}\left(\underbrace{\mathbb{E}_q\left[\log \underline{p}(\mathbf{y}; \upsilon^2_{j,k})\right]}_{D} + \underbrace{\mathbb{E}_q\left[\log \underline{p}(\mathbf{y}; \lambda_{j,k})\right]}_{E}\right),\end{aligned} \tag{B.37}$$

where A is similar to (B.14) 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.1.8. Hence, we have that:

$$A = \log \underline{p}^{\mathrm{SV}}(\mathbf{y}; \boldsymbol{\beta}, \mathbf{h}, \boldsymbol{\psi}) \quad \big(\text{or } \log \underline{p}^{\mathrm{C}}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\nu}) \text{ if homoskedastic}\big)$$
$$- \frac{1}{2} \sum_{j=1}^{d} \sum_{k=1}^{d+p+1} \left(\mu_{q(\log \delta^2)} + \mu_{q(\log \upsilon_{j,k}^2)} + \mu_{q(1/\delta^2)} \mu_{q(1/\upsilon_{j,k}^2)} \mu_{q(\vartheta_{j,k}^2)} \right) + \frac{1}{2} \left(\log |\boldsymbol{\Sigma}_{q(\boldsymbol{\vartheta})}| + d(d+p+1) \right). \tag{B.38}$$

Consider now the piece B. We have that:

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

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

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

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.8) (homoskedastic);
  Update $q^*(\mathbf{h}_1)$ and therefore $q^*(\boldsymbol{\nu}_1)$ as in (B.1) and (B.7) (heteroskedastic);
  Update $q^*(\psi_1)$ as in (B.12);
  for $j = 2, \dots, d$ **do**
    Update $q^*(\nu_j)$ as in (B.8) (homoskedastic);
    Update $q^*(\mathbf{h}_j)$ and therefore $q^*(\boldsymbol{\nu}_j)$ as in (B.1) and (B.7) (heteroskedastic);
    Update $q^*(\psi_j)$ as in (B.12);
    Update $q^*(\boldsymbol{\beta}_j)$ as in (B.9);
  end
  Update $q^*(\boldsymbol{\vartheta})$ as in (B.30) or (B.31) ;
  for $j = 1, \dots, d$ **do**
    for $k = 1, \dots, d + p + 1$ **do**
      Update $q^*(\upsilon_{j,k}^2)$, $q^*(\lambda_{j,k})$ as in (B.32)-(B.34);
    end
  end
  Update $q^*(\gamma^2)$, $q^*(\eta)$ as in (B.33)-(B.35);
  Compute $\log \underline{p}(\mathbf{y}; q)$ as in (B.36);
  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

C Variational predictive density

In this section we first discuss the approximation of $q^*(\boldsymbol{\Omega}_t)$. This is instrumental to the derivation of the optimal variational predictive density.

C.1 Inference on the time-varying precision matrix

Proposition 3.5 shows that, conditional on $\mathbf{L}$ and $\mathbf{V}_t$, the optimal distribution of $\boldsymbol{\Omega}_t$ can be approximated by a d-dimensional Wishart distribution $\mathsf{Wishart}_d(\delta_t, \mathbf{H}_t)$, where δ_t and $\mathbf{H}_t$ are the degrees of freedom and the scaling matrix, respectively. The complete proof is based

on the Expectation Propagation (EP) approach proposed by Minka (2001). This has the goal of minimizing the KL divergence between the true and unknown optimal variational distribution $q^*(\mathbf{\Omega}_t)$ and a sub-optimal approximating density $\tilde{q}(\mathbf{\Omega}_t)$. In order to implement this approach, there is no need to know $q^*(\mathbf{\Omega}_t)$, but it is sufficient to be able to compute $\mathbb{E}_q(\mathbf{\Omega}_t)$. The latter can be reconstructed based on the optimal variational densities of the Cholesky factor $q^*(\boldsymbol{\beta})$ – and therefore for $\mathbf{L}$ –, and of $\mathbf{V}_t$.

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

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

$$\psi(\delta_t, \mathbf{H}_t) \propto \frac{d\delta_t}{2}\log 2 + \frac{\delta_t}{2}\log|\mathbf{H}_t| + \log\Gamma_d(\delta_t/2) - \frac{\delta_t}{2}\mathbb{E}_q[\log|\mathbf{\Omega}_t|] + \frac{1}{2}\mathrm{tr}\left\{\mathbf{H}_t^{-1}\mathbb{E}_q[\mathbf{\Omega}_t]\right\}. \quad \text{(C.1)}$$

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

Table 1 compares the sampled distributions with the marginals of the Wishart with $(\widehat{\delta}_t, \widehat{\mathbf{H}}_t)$ in terms of approximation accuracy $\mathcal{ACC} = 100\left\{1 - 0.5\int|\tilde{q}(\omega_t) - q(\omega_t)|\,d\omega_t\right\}\%$, where ω_t is a generic element of $\mathbf{\Omega}_t$.

	$d = 15$		$d = 30$		$d = 50$		$d = 100$	
	$\omega_{j,j,t}$	$\omega_{j,k,t}$	$\omega_{j,j,t}$	$\omega_{j,k,t}$	$\omega_{j,j,t}$	$\omega_{j,k,t}$	$\omega_{j,j,t}$	$\omega_{j,k,t}$
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 Wishart approximation $\tilde{q}(\mathbf{\Omega}_t)$ for dimensions $d = 15, 30, 50, 100$ separately for the diagonal ($\omega_{j,j,t}$) and out-of-diagonal ($\omega_{j,k,t}$) elements of $\mathbf{\Omega}_t$.

The simulation results suggest that our variational inference approach provides an accurate approximation of the optimal distribution of $\boldsymbol{\Omega}_t$ for different dimensions.

C.2 Derivation of the variational predictive density

Recall that the variational predictive posterior can be computed as:

$$q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = \int p(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\xi}) q^*(\boldsymbol{\xi}) d\boldsymbol{\xi} = \int \int p(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\vartheta}, \boldsymbol{\Omega}) q^*(\boldsymbol{\vartheta}) q^*(\boldsymbol{\Omega}_t) d\boldsymbol{\vartheta}\, d\boldsymbol{\Omega}_t, \tag{C.2}$$

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 $\boldsymbol{\Omega}_t$ (as discussed in Section C.1) in the following way:

$$q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = \int q(\boldsymbol{\vartheta}) \underbrace{\left[\int \mathsf{N}_d(\mathbf{y}_{t+1}; \boldsymbol{\Theta}\mathbf{z}_t, \boldsymbol{\Omega}_t^{-1}) \mathsf{Wishart}_d(\boldsymbol{\Omega}_t; \delta_t, \mathbf{H}_t) d\boldsymbol{\Omega}_t \right]}_{A} d\boldsymbol{\vartheta}, \tag{C.3}$$

where

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

$$= \frac{|1 + \frac{1}{v_t}(\mathbf{y}_{t+1} - \boldsymbol{\Theta}\mathbf{z}_t)^\intercal v_t \mathbf{H}_t (\mathbf{y}_{t+1} - \boldsymbol{\Theta}\mathbf{z}_t)|^{-\frac{v_t+d}{2}} \Gamma(\frac{v_t+d}{2})}{\pi^{d/2} v_t^{d/2} |\mathbf{H}_t^{-1}|^{1/2} \Gamma(v_t/2)} = h(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\vartheta}), \tag{C.4}$$

is the density function of a multivariate Student-t distribution with dimension d, $v_t = \delta_t - d + 1$ degrees of freedom, mean vector $\boldsymbol{\Theta}\mathbf{z}_t$ and scaling matrix $\mathbf{S}_t = (v_t \mathbf{H}_t)^{-1}$, i.e. $\mathsf{t}_{v_t}(\boldsymbol{\Theta}\mathbf{z}_t, \mathbf{S}_t)$. Then, the integral in Eq.(C.2) becomes

$$q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = \int h(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\vartheta}) q(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}, \tag{C.5}$$

which requires to simulate only 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 further increase the computational efficiency. To this aim, we propose to approximate the multivariate Student-t in

(C.5) with the closest multivariate normal distribution in terms of KL divergence:

$$\begin{aligned}\mathcal{D}_{KL}(h\|\phi) &\propto -\int \log \phi(\mathbf{y}_{t+1}|\mathbf{m}_t, \mathbf{R}_t^{-1}) h(\mathbf{y}_{t+1}|\mathbf{z}_t, \boldsymbol{\vartheta})\, d\mathbf{y}_{t+1} \\ &= -\mathbb{E}_h(\log \phi(\mathbf{y}_{t+1}|\mathbf{m}_t, \mathbf{R}_t^{-1})) = \psi(\mathbf{m}_t, \mathbf{R}_t),\end{aligned} \tag{C.6}$$

where, in particular,

$$\begin{aligned}\psi(\mathbf{m}_t, \mathbf{R}_t) &\propto \mathbb{E}_h\left(-\frac{1}{2}\log \mathbf{R}_t + \frac{1}{2}(\mathbf{y}_{t+1} - \mathbf{m}_t)^\intercal \mathbf{R}_t (\mathbf{y}_{t+1} - \mathbf{m}_t)\right) \\ &= -\frac{1}{2}\log \mathbf{R}_t + \frac{1}{2}(\boldsymbol{\Theta}\mathbf{z}_t - \mathbf{m}_t)^\intercal \mathbf{R}_t (\boldsymbol{\Theta}\mathbf{z}_t - \mathbf{m}_t) + \frac{v_t}{2(v_t - 2)}\mathrm{tr}\left\{\mathbf{R}_t \mathbf{S}_t\right\},\end{aligned} \tag{C.7}$$

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

$$q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) = \int \phi(\mathbf{y}_{t+1}|\mathbf{m}_t, \mathbf{R}_t^{-1}) q(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}, \tag{C.8}$$

where now $\phi(\mathbf{y}_{t+1}|\boldsymbol{\Theta}\mathbf{z}_t, \mathbf{R}_t^{-1})$ 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.8) can be solved analytically leading to a closed form variational predictive density $q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t})$ which is a multivariate Gaussian distribution with variance matrix $\boldsymbol{\Sigma}_{pred,t}$ and mean vector $\boldsymbol{\mu}_{pred,t}$. Define $\mathbf{Z}_t = (\mathbf{I}_d \otimes \mathbf{z}_t^\intercal)$ and compute the integral above:

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

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

the integral and get:

$$
\begin{aligned}
q(\mathbf{y}_{t+1}|\mathbf{z}_{1:t}) &\propto \exp\left\{-\frac{1}{2}(\mathbf{y}_{t+1}^{\intercal}\mathbf{R}_t\mathbf{y}_{t+1} - \tilde{\boldsymbol{\mu}}_t^{\intercal}\tilde{\boldsymbol{\Sigma}}_t\tilde{\boldsymbol{\mu}}_t)\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\mathbf{y}_{t+1}^{\intercal}\mathbf{R}_t\mathbf{y}_{t+1} - \mathbf{y}_{t+1}^{\intercal}\mathbf{R}_t\mathbf{Z}_t\tilde{\boldsymbol{\Sigma}}_t\mathbf{Z}_t^{\intercal}\mathbf{R}_t\mathbf{y}_{t+1} - 2\mathbf{y}_{t+1}\mathbf{R}_t\mathbf{Z}_t\tilde{\boldsymbol{\Sigma}}_t\boldsymbol{\Sigma}_{q(\vartheta)}^{-1}\boldsymbol{\mu}_{q(\vartheta)})\right\} \\
&= \exp\left\{-\frac{1}{2}(\mathbf{y}_{t+1}^{\intercal}(\mathbf{R}_t - \mathbf{R}_t\mathbf{Z}_t\tilde{\boldsymbol{\Sigma}}_t\mathbf{Z}_t^{\intercal}\mathbf{R}_t)\mathbf{y}_{t+1} - 2\mathbf{y}_{t+1}\mathbf{R}_t\mathbf{Z}_t\tilde{\boldsymbol{\Sigma}}_t\boldsymbol{\Sigma}_{q(\vartheta)}^{-1}\boldsymbol{\mu}_{q(\vartheta)})\right\},
\end{aligned}
\tag{C.10}
$$

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

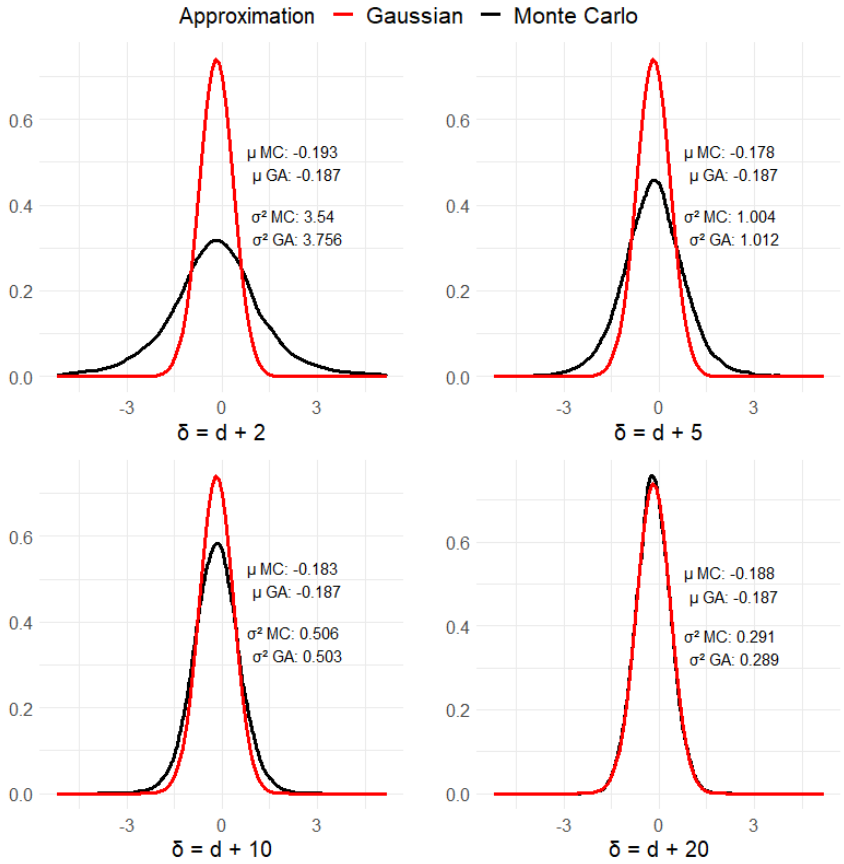


Figure C.10: Second-order approximation of the predictive density.

Figure C.10 shows the approximation of variational predictive posterior with Monte Carlo methods (MC) and via Gaussian approximation (GA) varying the degrees of freedom $\widehat{\delta}_t$ for the distribution of $\boldsymbol{\Omega}_t$. We can see that if $\widehat{\delta}_t \gg d$ the approximation is rather accurate, while the accuracy decreases as $\widehat{\delta}_t$ approaches d. However, even for the case $\widehat{\delta}_t \approx d$, we can still

obtain precise estimates of the first and second moments 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. The true data generating process is an homoskedastic VAR(1):

$$\mathbf{y}_t = \boldsymbol{\Theta}\,\mathbf{y}_{t-1} + \mathbf{u}_t, \qquad \mathbf{u}_t \sim \mathsf{N}_d(\mathbf{0}_d, \boldsymbol{\Omega}^{-1}), \qquad t = 1, \dots, T.$$

The reason why we focus on a VAR(1) data generating process is for direct comparability with the competing estimation methods, such as Gruber and Kastner (2022) and Gefang et al. (2023), which do not consider the presence of exogenous predictors.

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 zeros) and high level of sparsity (90% of zeros). The true matrix $\boldsymbol{\Theta}$ is generated as follows: we fix to zero $s \cdot d^2$ 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. Note 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.

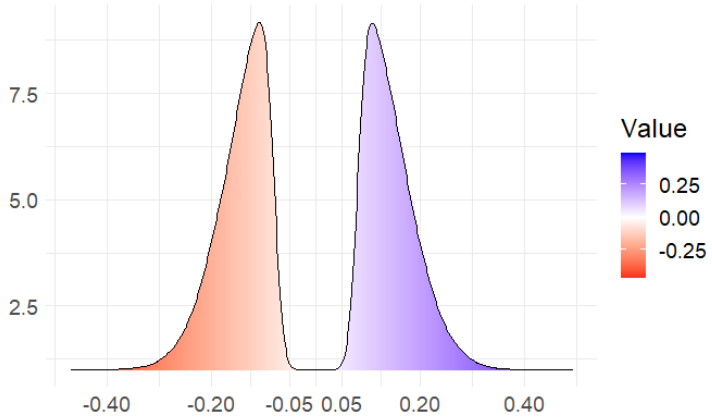
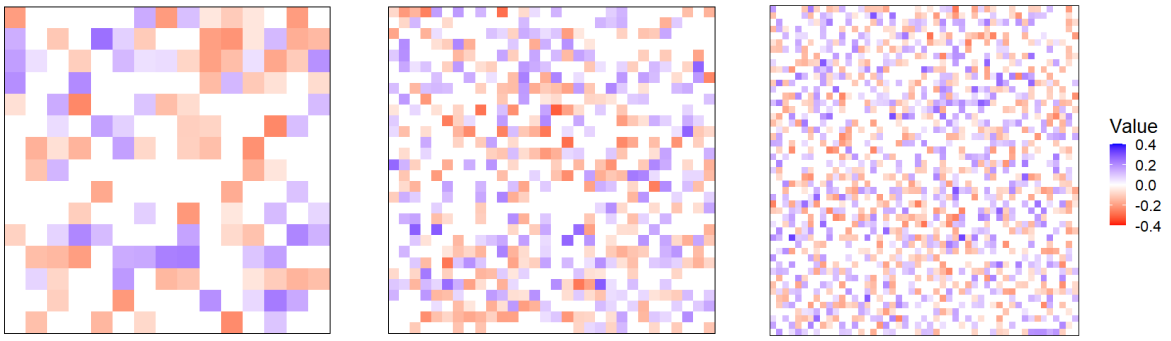
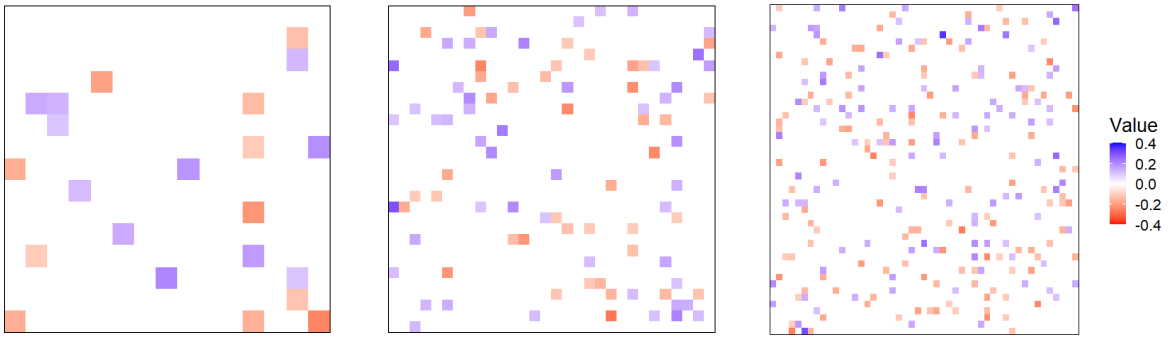


Figure D.1: Distribution of non-zero parameters in the true regression matrix. This figure plots the distribution from which we sample the non-zero entries of the regression matrices used to generate the data for the simulation study.

The variance-covariance matrix $\boldsymbol{\Omega}^{-1}$ coincides with the sample variance covariance matrix



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



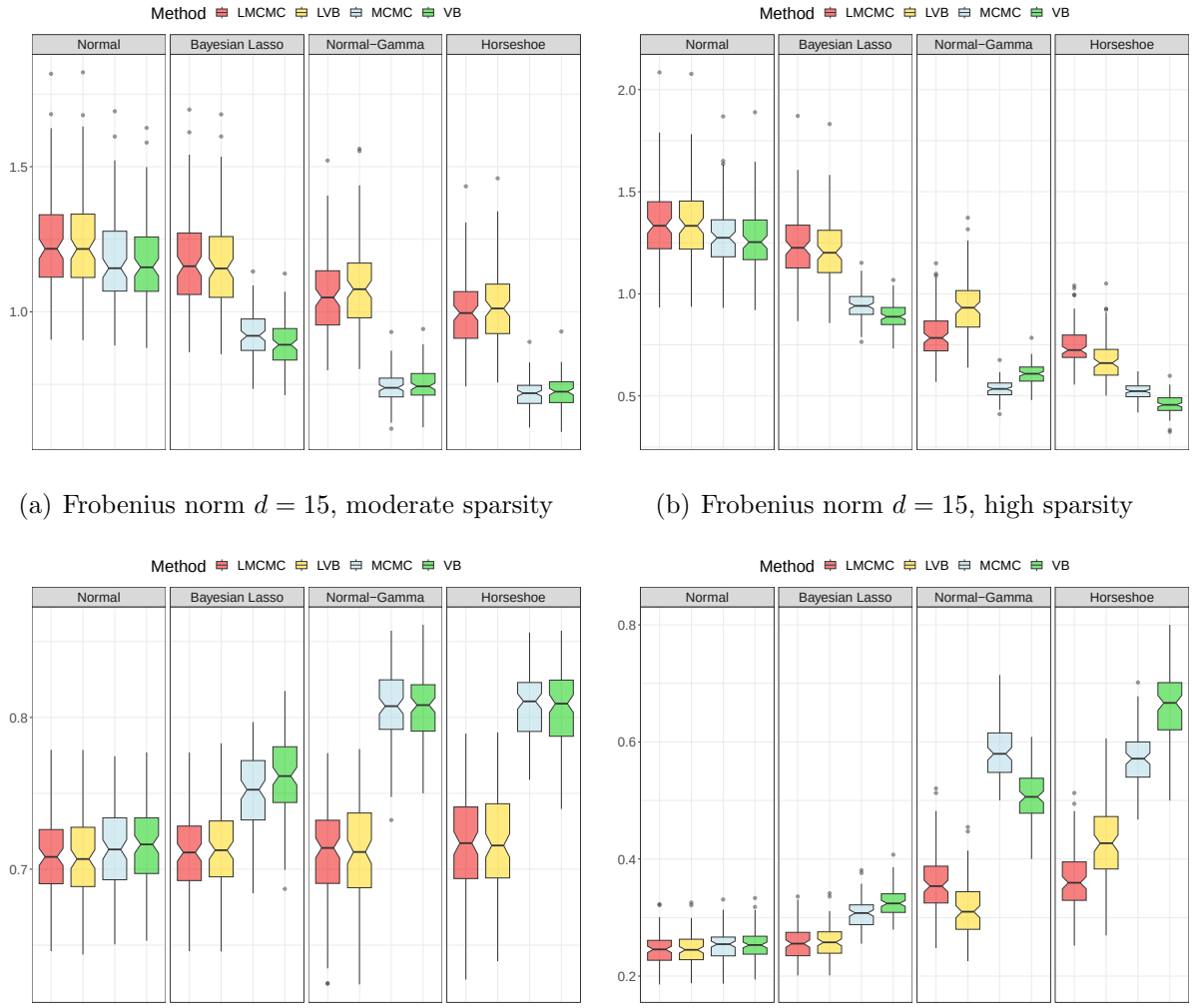
(d) $d = 15$ high sparsity (e) $d = 30$ high sparsity (f) $d = 49$ high sparsity

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).

computed on the real-data used in the empirical application. The initial state $\mathbf{y}_0$ is sampled from the marginal distribution of the VAR(1) defined above, and we consider a burn-in period of $t_{\text{burn}} = 1, \dots, 1000$ before sampling $(\mathbf{y}_1, \dots, \mathbf{y}_T)$ from the VAR(1). 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 $\boldsymbol{\Theta}$ are set to zero.

D.1 Additional simulation results

We complement the results in the main text and show some of the additional results on a smaller model dimension of $d = 15$. Figure D.3 reports the Frobenius norm (top panels) and the F1 score (bottom panels) as in the main text. The labeling and structure of figure is the same as in Figure 2. Similar to the larger VAR cases, our `VB` estimation procedure outperform both MCMC and variational methods based on a structural VAR formulation.



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

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

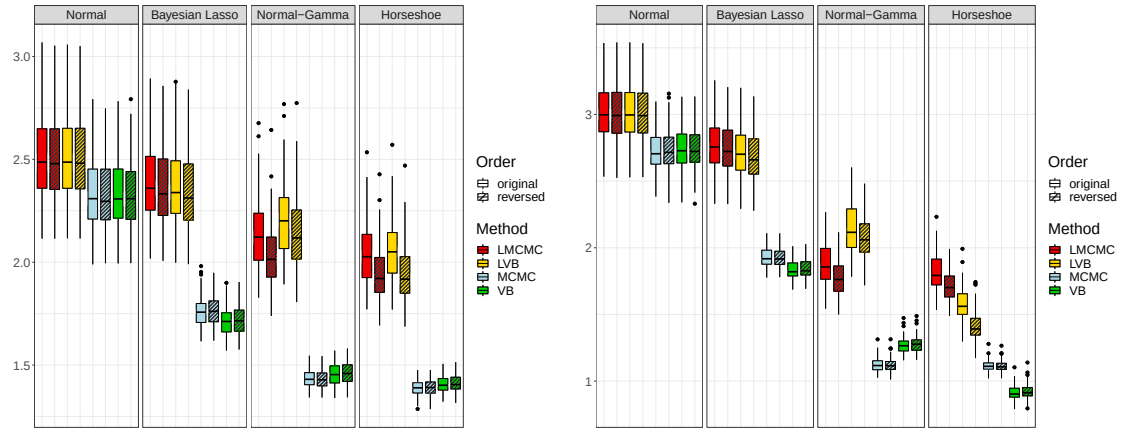
(c) F1 score norm $d = 15$, moderate sparsity

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

Figure D.3: Top panels report the Frobenius norm of $\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}$ for different hierarchical shrinkage priors and estimation methods. 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.

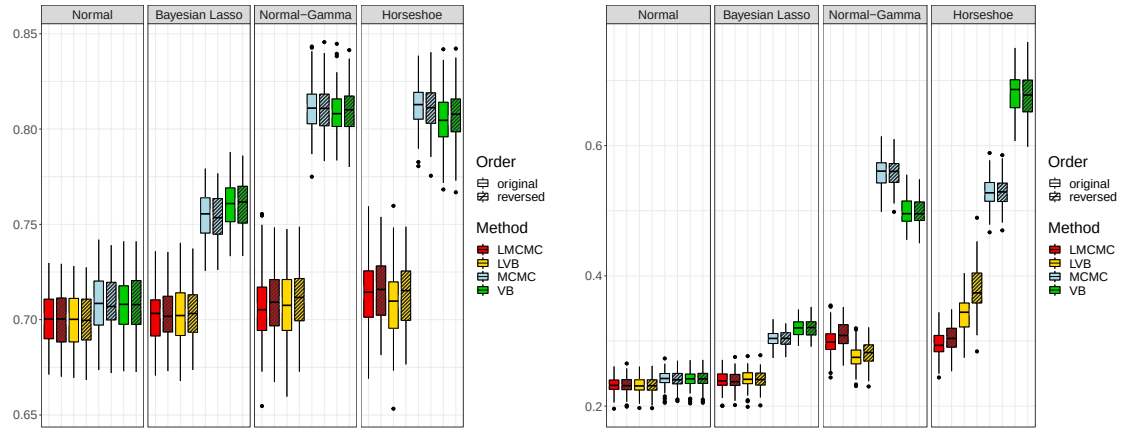
On the other hand, the non-linear MCMC proposed by Gruber and Kastner (2022) turns out to be quite competitive. Nevertheless, our `VB` approach is more accurate for both the adaptive lasso and horseshoe priors, especially when sparsity is more pervasive.

Based on the same simulation setting described above, we now investigate the performance of all estimation methods under variables permutation. Figure D.4 shows the box charts of the Frobenius norms (top panels) and F1 scores (bottom panels) for the $N = 100$ replications for both moderate and high sparsity in the true $\boldsymbol{\Theta}$. For ease of exposition, we only report the case with $d = 30$ predictors. We put in each figure the simulation results



(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 D.4: Top panels report the Frobenius norm of $\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}$ under variables permutation for different 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 $\widehat{\boldsymbol{\Theta}}$. The box charts show the results for $N = 100$ replications, $d = 30$ and different levels of sparsity.

pertaining to the original $\mathbf{y}_t$ (solid) and its reversed order $\mathbf{y}_t^{rev}$ (shaded) next to each other. Colors/labels are the same as in the main simulation study.

The accuracy of the estimates of both `LMCMC` and `LVB` tend to deteriorate when reverting the ordering of the target variables. This is especially clear for the normal-gamma and the horseshoe priors and when the amount of zero coefficients in $\boldsymbol{\Theta}$ is more pervasive. Such performance deterioration is due to the fact that $\boldsymbol{\Theta} = \mathbf{L}^{-1}\mathbf{A}$ from the structural VAR formulation so that the posterior estimate $\widehat{\boldsymbol{\Theta}}$ changes depending on the variables ordering implied by $\mathbf{L}$. The higher the level of sparsity, the larger the disconnect between $\mathbf{A}$ and $\boldsymbol{\Theta}$.

On the other hand, being built on the same non-linear parametrization both the `MCMC` of Gruber and Kastner (2022) and our `VB` approach are substantially less sensitive to variables permutation. This applies across prior specifications, model dimension, and level of sparsity in the true matrix $\mathbf{\Theta}$.

D.2 A multivariate version of Hahn and Carvalho (2015)

The implementation of the sparsity-inducing approach of Hahn and Carvalho (2015) to our multivariate context requires a non-trivial extension. In their original work, the authors assume a linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and uncorrelated Gaussian error terms, $\boldsymbol{\varepsilon} \sim \mathsf{N}_n(0, \sigma^2\mathbf{I}_n)$. Thus, their procedure consists to run the following least-angle regression (LARS) for a grid of tuning parameters λ:

$$\boldsymbol{\beta}_\lambda = \arg\min_{\gamma} \sum_j \frac{\lambda}{|\widehat{\beta}_j|}|\gamma_j| + n^{-1}||\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\gamma}||_2^2, \tag{D.1}$$

where $\widehat{\boldsymbol{\beta}}$ denotes the posterior mean, and, then, to compute, for each λ and each draw $(\boldsymbol{\beta}^{(r)}, \sigma^{2\,(r)})$, the variation-explained for the sparsified linear predictor $\boldsymbol{\beta}_\lambda$:

$$\rho_\lambda^{2\,(r)} = \frac{n^{-1}||\mathbf{X}\boldsymbol{\beta}^{(r)}||^2}{n^{-1}||\mathbf{X}\boldsymbol{\beta}^{(r)}||^2 + \sigma^{2\,(r)} + n^{-1}||\mathbf{X}\boldsymbol{\beta}^{(r)} - \mathbf{X}\boldsymbol{\beta}_\lambda||^2}. \tag{D.2}$$

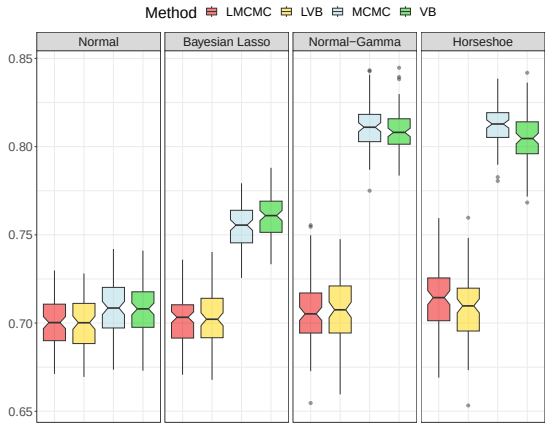
The selection follows a comparison between ρ_λ^2 and $\rho_{\lambda=0}^2$ based on the following heuristic: report the sparsified linear predictor corresponding to the smallest model whose 90% ρ_λ^2 credible interval contains $E(\rho_{\lambda=0}^2)$, that is, select the smallest linear predictor whose variance-explained is not statistically different than the full model.

In our setting, we need to define a suitable formula to compute ρ_λ^2 when $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and the error terms are correlated, i.e. $\boldsymbol{\varepsilon} \sim \mathsf{N}_n(0, \mathbf{\Sigma})$. A natural choice appears to be:

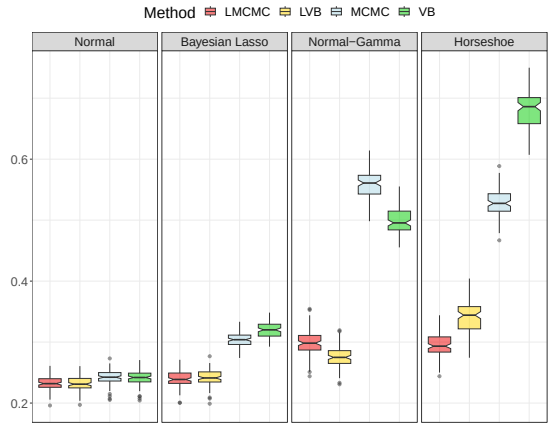
$$\rho_\lambda^{2\,(r)} = \frac{n^{-1}\boldsymbol{\beta}^{\intercal\,(r)}\mathbf{X}^\intercal\mathbf{\Sigma}^{-1\,(r)}\mathbf{X}\boldsymbol{\beta}^{(r)}}{n^{-1}\boldsymbol{\beta}^{\intercal\,(r)}\mathbf{X}^\intercal\mathbf{\Sigma}^{-1\,(r)}\mathbf{X}\boldsymbol{\beta}^{(r)} + 1 + n^{-1}(\mathbf{X}\boldsymbol{\beta}^{(r)} - \mathbf{X}\boldsymbol{\beta}_\lambda)^\intercal\mathbf{\Sigma}^{-1\,(r)}(\mathbf{X}\boldsymbol{\beta}^{(r)} - \mathbf{X}\boldsymbol{\beta}_\lambda)}. \tag{D.3}$$

Notice that, if $\mathbf{\Sigma} = \sigma^2\mathbf{I}_n$ then we obtain the original approach of Hahn and Carvalho (2015).

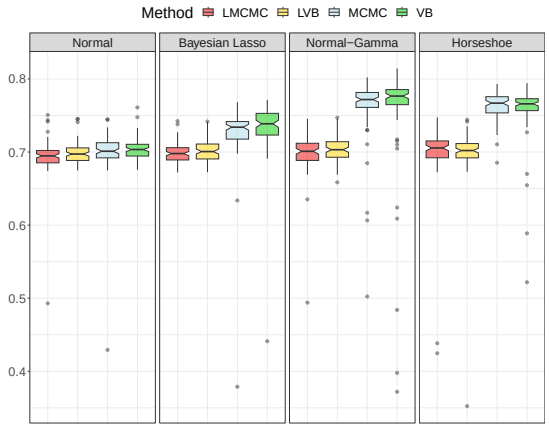
Before discussing some of the additional simulation results, two comments are in order. First, the selection from Hahn and Carvalho (2015) depends on some non-negligible arbitrariness. Specifically, the comparison between ρ_λ^2 and $\rho_{\lambda=0}^2$ is carried out using the selection summary plots (Section 3 of Hahn and Carvalho, 2015). Second, and perhaps more importantly, the post-processing approach based on SAVS is an order of magnitude faster. Indeed,



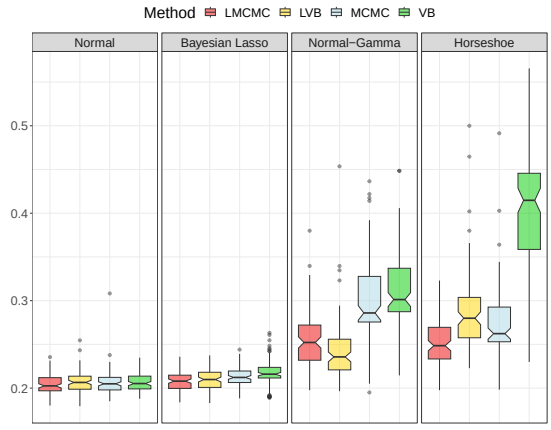
(a) $d = 30$, moderate sparsity, SAVS



(b) $d = 30$, high sparsity, SAVS



(c) $d = 30$, moderate sparsity, HC



(d) $d = 30$, high sparsity, HC

Figure D.5: F1 score computed looking at the true non-null parameters in $\boldsymbol{\Theta}$ and the non-null parameters estimated based on $\widehat{\boldsymbol{\Theta}}$.

the approach of Hahn and Carvalho (2015) requires the evaluation of Eq.(D.3) for each λ and each draws from the posterior. Moreover, λ values are defined over a grid: if the latter is too coarse, then the selection procedure might be inaccurate, while if it is too dense, the computational burden suddenly increases.

According to Ray and Bhattacharya (2018), the latter issue does not affect the SAVS procedures, which indeed does not require tuning parameters and it is computationally fast. To put things into perspective, with $d = 30$, considering $5,000$ draws from the posterior after the burn-in, and a grid of 200 values for λ, the SAVS procedure provides a sparse estimate immediately, while the Hahn and Carvalho (2015) approach takes ≈ 1 minute.

Figure D.5 compares the F1 score based on the same posterior and variational estimates, but with either the SAVS (top panels) or the extended version of Hahn and Carvalho (2015)

as outlined above across different shrinkage priors. For ease of exposition, we report uniquely the results for the $d = 30$ case. The F1 scores across methods remain largely the same, in fact, the results are even more strongly in favor of our `VB` compared to its `MCMC` counterpart when using the extended Hahn and Carvalho (2015) approach. Specifically, our `VB` is more accurate than `MCMC` under the normal-gamma prior.

E Additional empirical considerations

E.1 Computational cost of the recursive forecasts

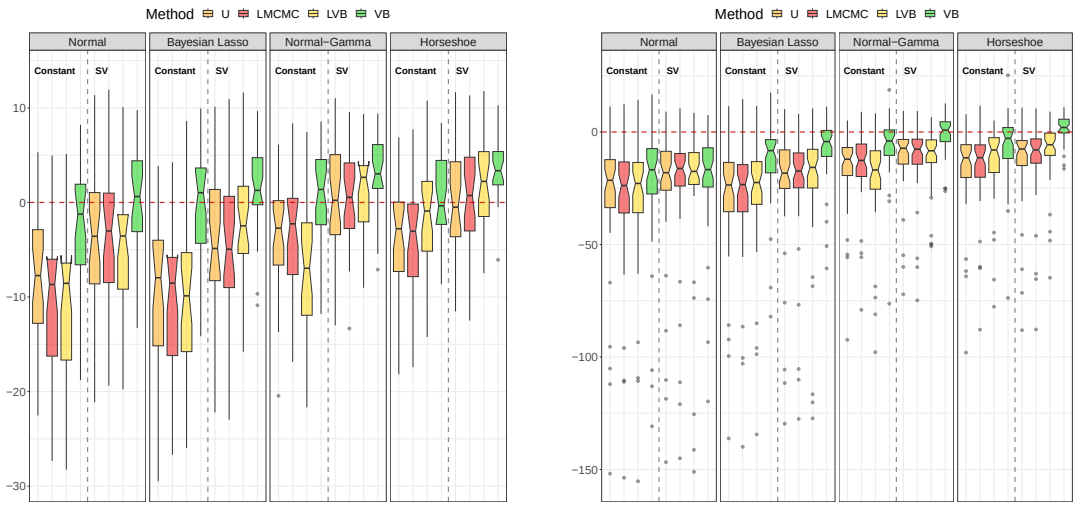
In this section, we discuss more explicitly the qualitative differences in terms of computational efficiency across estimation methods. Starting with Carriero et al. (2019, 2022), they consider $d = 20, 40$ and show that the average computational time to perform 10 draws is 2.5 and 27.3 seconds, respectively, on a 3.5 GHz Intel Core i7 (see Figure 1 in Carriero et al., 2022). This means that for 10,000 draws (as in our case) it takes 41 minutes for $d = 20$ and 7.5 hours for $d = 40$ per monthly forecast. Similarly, on a 2.5 GHz Intel Xeon W-2175 with 32GB of RAM it would take approximately 40 minutes per forecast to implement the `MCMC` approach of Gruber and Kastner (2022) for a $d = 30$ implementation with constant volatility. Huber and Feldkircher (2019), based on a similar non-linear MCMC algorithm for $d = 20$ variables takes around 1.3 hours for 30,000 posterior draws, or 26 minutes for 10,000 draws. These results are all consistent with our own implementations of these methods.

By comparison, our `VB` with stochastic volatility takes less than 3 minutes for each recursive forecast with $d = 30$. This has key implications for practical forecasting use; for instance, a recursive forecast of $d = 30$ industry portfolios for 767 out-of-sample observations based on a constant-volatility specification of Gruber and Kastner (2022) would take $20 \text{ min} \times 767 \text{ forecasts} \times 4 \text{ priors} = 76,700$ minutes, or 42 days to complete. This compares to $10 \text{ sec} \times 767 \text{ forecasts} \times 4 \text{ priors} = 511$ minutes, or almost 9 hours to complete the empirical exercise under a constant-volatility specification with our variational inference approach.

To summarize, a substantially higher computational efficiency coupled with a comparable accuracy with complex MCMC, makes our `VB` extremely competitive within the context of recursive forecasts in higher frequency data.

E.2 Forecasting performance over the business cycle

Figure E.6 reports the $R^2_{j,oos}(\mathcal{M}_s)$ (in %) across 30 (left panel) and 49 (right panel) industry portfolios during recession periods.

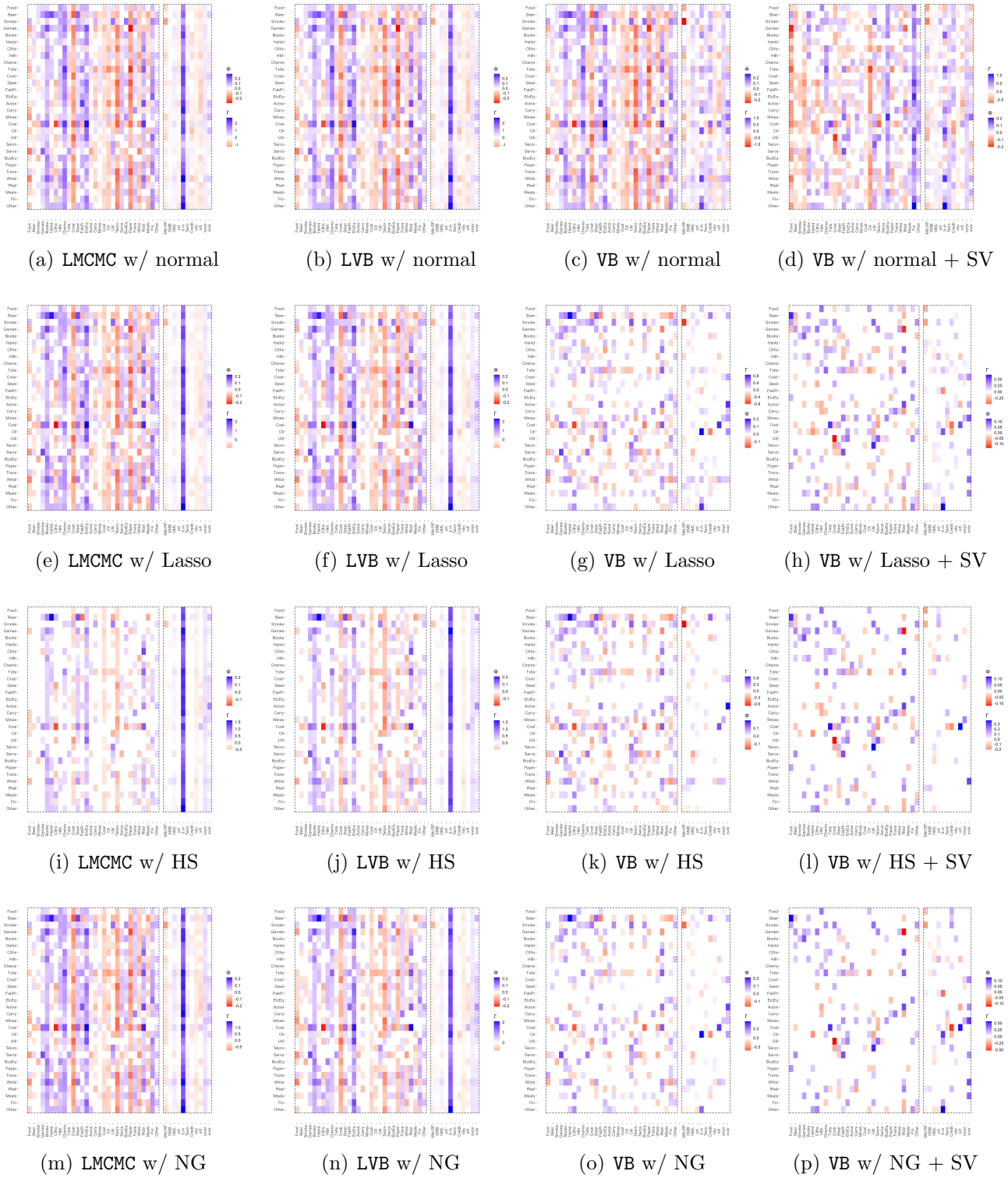


(a) $R_{j,oos}\left(\mathcal{M}_s\right)^2$ across 30 industry portfolios (b) $R^2_{j,oos}\left(\mathcal{M}_s\right)$ across 49 industry portfolios

Figure E.6: This figure reports the $R^2_{j,oos}\left(\mathcal{M}_s\right)$ (in %) across 30 (left panel) and 49 (right panel) industry portfolios.

E.3 Additional in-sample results

Figure E.7 shows the in-sample posterior estimates estimates of the regression coefficients for the $d = 30$ industry case. The in-sample estimates of $\widehat{\mathbf{\Theta}}$ are based on the full sample obtained from the `LMCMC` and the `LVB` with constant volatility, and the `VB` with and without stochastic volatility. Similar to the larger-dimensional setting in the main text, the in-sample estimates highlight three key results. First, and perhaps not surprisingly, there are visible differences across shrinkage priors. For instance, the horseshoe tend to shrinkage parameters more aggressively so that $\widehat{\mathbf{\Theta}}$ is more sparse compared to the normal gamma. Second, the estimates of the `LMCMC` and `LVB` tend to be closely related, consistent with Gefang et al. (2023). Yet, the estimates for the `VB` are substantially different under the same prior. This is due to the fact that $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{L}}^{-1}\widehat{\mathbf{A}}$ in Eq.(2b), so that the estimated $\widehat{\mathbf{A}}$ is not translation-invariant, unlike in our approach. Third, the estimates from `VB` are remarkably stable between constant vs stochastic volatility specifications, with the only exception of the adaptive lasso prior.



(a) LMCMC w/ normal (b) LVB w/ normal (c) VB w/ normal (d) VB w/ normal + SV

(e) LMCMC w/ Lasso (f) LVB w/ Lasso (g) VB w/ Lasso (h) VB w/ Lasso + SV

(i) LMCMC w/ HS (j) LVB w/ HS (k) VB w/ HS (l) VB w/ HS + SV

(m) LMCMC w/ NG (n) LVB w/ NG (o) VB w/ NG (p) VB w/ NG + SV

Figure E.7: Variational Bayes estimates of the regression coefficients $\mathbf{\Theta}$ for different estimation methods. We report the estimates for the $d = 30$ industry case obtained for all priors. We report the results for VB with and without stochastic volatility.