Adaptive Hierarchical Priors for High-Dimensional Vector

Davide Pettenuzzo, Economics Department, Brandeis University

Dimitris Korobilis, Business School, University of Essex

Working Paper Series

2017| 115

Adaptive Hierarchical Priors for High-Dimensional Vector Autoregressions*

Dimitris Korobilis
University of Essex†

Davide Pettenuzzo
Brandeis University‡

September 14, 2017

Abstract

This paper proposes a scalable and simulation-free estimation algorithm for vector autoregressions (VARs) that allows fast approximate calculation of marginal posterior distributions. We apply the algorithm to derive analytical expressions for popular Bayesian shrinkage priors that admit a hierarchical representation and which would typically require computationally intensive posterior simulation methods. The proposed algorithm is modular, parallelizable, and scales linearly with the number of predictors, allowing fast and efficient estimation of large Bayesian VARs. The benefits of our approach are explored using three quantitative exercises. First, a Monte Carlo experiment illustrates the accuracy and computational gains of the proposed estimation algorithm and priors. Second, a forecasting exercise involving VARs estimated on macroeconomic data demonstrates the ability of hierarchical shrinkage priors to find useful parsimonious representations. Finally, we show that our approach can be used successfully for structural analysis and can replicate important features of structural shocks predicted by economic theory.

Keywords: Bayesian VARs, Mixture prior, Large datasets, Macroeconomic forecasting

JEL Classifications: C11, C13, C32, C53

*We would like to thank Linda Bui, Andrea Carriero, Roberto Casarin, Sid Chib, Todd Clark, Sylvia Früwirth-Schnatter, Anna Galvao, Domenico Giannone, Jim Griffin, Maria Kalli, Gary Koop, Robert McCulloch, Ivan Petrella, Simon Price, Giorgio Primiceri, Francesco Ravazzolo, Giovanni Ricco, Barbara Rossi, Mark Steel, Robert Taylor, Allan Timmermann, and Frank Windmeijer. We would also like to acknowledge for their helpful comments participants at the University of Pennsylvania "Big Data in Predictive Dynamic Econometric Modeling" conference, the 23^{rd} Computing in Economics and Finance conference, the 2017 NBER Seminar on Bayesian Inference in Econometrics and Statistics, the 2017 Sheffield Macroeconomic Workshop, as well as seminar participants at the National Bank of Poland, the European Central Bank, Brandeis University, and the Universities of Bristol, Kent, and Warwick.

†Essex Business School, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom. d.korobilis@essex.ac.uk

‡Brandeis University, Sachar International Center, 415 South St, Waltham, MA. dpettenu@brandeis.edu

1 Introduction

There is ample evidence that exploiting large information sets can be beneficial for macroeconomic forecasting and structural analysis. While the early literature has established this fact in univariate applications (Stock and Watson, 2002), a more recent literature applies this concept to multivariate vector autoregressions (Banbura et al., 2010). Not surprisingly, a large body of this literature relies on Bayesian methods, exploiting subjective prior information to help with the often very low signal-to-noise ratio in the data. As an example, the early literature on vector autoregressions (Doan et al., 1984; Litterman, 1979) realized the benefits of using priors to achieve regularized and reliable estimators, and this led to the so-called Minnesota (or Littermann) prior. There is also no shortage of work taking advantage of recent advances in Bayesian computation to examine more complex prior structures. For example, Del Negro and Schorfheide (2004) and Ingram and Whiteman (1994) specify priors from general equilibrium models; Andersson and Karlsson (2008), George et al. (2008), Koop and Potter (2004) , Korobilis (2013), Stock and Watson (2006) and Strachan and Dijk (2013) focus on model averaging and selection priors; De Mol et al. (2008), Gefang (2014), Giannone et al. (2015) and Huber and Feldkircher (forthcoming) examine the properties of shrinkage priors with a hierarchical structure.

One of the main features distinguishing all these recent approaches from the earlier literature is the ability to rely on prior distributions whose moments can be directly informed from the data. More specifically, hyperparameters controlling how informative a prior distribution is (and that would typically be a subjective choice) can now be integrated in the estimation phase, simply by eliciting appropriate prior distributions on them. Such hierarchical priors have excellent shrinkage properties and can approximate a wide class of penalized regression estimators. Notable examples are the double-exponential prior that leads to the least absolute shrinkage and selection operator (LASSO) estimator (Tibshirani, 1994) and the Spike-and-Slab prior, which is connected to the generalized ridge estimator (Ishwaran and Rao, 2005). In the machine learning literature, hierarchical priors are referred to as "sparse Bayesian learning" or "adaptive sparseness" priors, due to the fact that the informativeness of the prior is learned from the data; see Tipping (2001) and Figueiredo (2003). At the same time, estimators based

on hierarchical priors are also highly correlated with estimators based on principal component shrinkage; see De Mol et al. (2008).

Notwithstanding their excellent properties and empirical success, the vast majority of existing applications featuring hierarchical priors have been severely limited due to their reliance on computationally intensive Markov Chain Monte Carlo (MCMC) methods. For example, Huber and Feldkircher (forthcoming) consider VARs with five variables, while George et al. (2008) work with seven-variable models. In high-dimensions, when the VAR parameters proliferate at a polynomial rate, such simulation-based methods become computationally cumbersome, if not infeasible. A notable exception is Giannone et al. (2015) who, in order to estimate systems with more than 20 equations, rely on the natural conjugate prior to obtain posterior estimates for the degree of informativeness of their prior. However, their approach is restricted by the fact that the natural conjugate prior treats each VAR equation symmetrically, and imposes that the prior covariances of the coefficients in any two equations must be proportional to one another. This means that if one wants to impose money neutrality in the VAR by shrinking to zero the coefficient of money in the equation for GDP, then the symmetry of the natural conjugate prior requires that the effect of money be removed in all other VAR equations in the system, even if ones believes that money could still be a useful predictor in some of the other equations, for example, that of inflation. Additionally, the natural conjugate prior only allows the estimation of a single prior hyperparameter that is common to all VAR parameters, a situation that may be quite restrictive in the presence of thousands of VAR parameters. In constrast, typical hierarchical priors such as the Bayesian LASSO (Park and Casella, 2008), allow each individual scalar coefficient of a regression model to have its own individual variance, thus leading to the notion of "adaptive shrinkage"; see also Tipping (2001) and Huber and Feldkircher (forthcoming).

In this paper we develop a new estimation algorithm for VARs under the proposed class of hierarchical shrinkage priors. Unlike the previous methods available in the literature, the proposed approach is simulation-free and can be used with models of high dimensions. Following van den Boom et al. (2015a), the estimation relies on a simple transformation ("rotation") of the VAR, into a form that is observationally equivalent but which permits us to obtain analytical expressions for the marginal posteriors of all VAR parameters. In addition to the analytical

nature of the estimation algorithm, another main advantage of our method is that it can be easily parallelized to take advantage of the processing capabilities of modern PCs and GPUs. In particular, suppose that the interest is on estimating a k-dimensional vector of parameters $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)'$. In this case, our algorithm can be easily parallelized into k separate tasks, each one leading to an analytical expression for the marginal posterior of β_i, $i = 1, ..., k$. We show, using a Monte Carlo exercise, that our algorithm is as accurate as the comparable simulation-based methods, but at a fraction of the computing time. At the same time, the simulation-free nature of the algorithm means that there are no "convergence" or other similar numerical issues.

These features of our estimation algorithm bring us to the remaining contributions of the paper. We first show how to use the new estimation algorithm to derive analytical posteriors and adaptive shrinkage for three popular cases of hierarchical priors: Normal-Jeffreys (Hobert and Casella, 1996); Spike-and-Slab (Mitchell and Beauchamp, 1988); and Normal-Gamma (Griffin and Brown, 2010). We then focus on two empirical exercises inspired by the recent literature on high-dimensional VARs. Our first application is a macroeconomic forecasting exercise using large-dimensional VARs. Banbura et al. (2010), Carriero et al. (2012), Carriero et al. (2016) and Koop et al. (2017) provide strong evidence that high-dimensional Bayesian VARs can consistently outperform smaller models. We show that when combined with the three hierarchical priors we focus on, our algorithm outperforms all competing methods in terms of forecast accuracy. Our second exercise involves using our algorithm to estimate impulse response functions from an identified BVAR. In particular, we simulate artificial time-series data from a large-scale DSGE model and show that our methods can be used to obtain very accurate impulse response functions.

The remainder of the paper is organized as follows. Section 2 describes in detail the estimation procedure we rely on to obtain analytical posteriors for the regression parameters in the presence of non-conjugate priors. Next, Section 3 examines the properties of three popular cases of hierarchical shrinkage priors and provides analytical derivations for the marginal posteriors of the coefficients of interest. Section 4 extends the methods described in Section 2 and Section 3 to the VAR case. After that, Section 5 describes the Monte Carlo exercise we use to test the accuracy of the estimation algorithm and the properties of the implied adaptive shrinkage. Section 6 is devoted to the macroeconomic forecasting application, while Section 7

focuses on extending our algorithm to estimate impulse response functions using artificial data obtained from a large-scale DSGE model. Finally, Section 8 offers some concluding remarks.

2 A new Bayesian estimation methodology

To illustrate how our estimation procedure works in a regression context, consider a simple univariate linear regression model of the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{v}, \tag{1}$$

where $\boldsymbol{y} = (y_1, ..., y_T)'$ is a $T \times 1$ vector featuring our dependent variable, $\boldsymbol{X} = (\boldsymbol{X}'_1, ..., \boldsymbol{X}'_T)'$ is a $T \times k$ matrix involving T observations on k regressors, $\boldsymbol{\beta}$ is the corresponding $k \times 1$ vector of regression coefficients, and $\boldsymbol{v} = (v_1, ..., v_T)' \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_T\right)$. When k is large, estimation of the high-dimensional posterior distribution $p\left(\boldsymbol{\beta}|\boldsymbol{y}\right)$ involves very costly operations (e.g. inversion of the high-dimensional matrix $\boldsymbol{X}$), and quickly becomes computationally demanding or even infeasible.

We now introduce an alternative approach to evaluate the marginal posteriors $\{p\left(\beta_j|\boldsymbol{y}\right)\}_{j=1}^{k}$, without the need to compute a number of high-dimensional integrals over the joint posterior distribution $p\left(\boldsymbol{\beta}|\boldsymbol{y}\right)$. We then proceed by approximating the full posterior $p\left(\boldsymbol{\beta}|\boldsymbol{y}\right)$ using the product of all k marginal posteriors.[1] Put simply, our approach works by transforming a complex and often intractable k-dimensional posterior evaluation problem into the product of k independent (and much simpler) estimation steps. We follow van den Boom et al. (2015a,b) and, one at a time, for each of the k columns in $\boldsymbol{X}$, define the following rotation,

$$y_j^* = \boldsymbol{q}_j'\boldsymbol{y}, \qquad \widetilde{\boldsymbol{y}}_j = \boldsymbol{W}_j'\boldsymbol{y}, \tag{2}$$

where $\boldsymbol{q}_j = \boldsymbol{X}_j / \left\|\boldsymbol{X}_j\right\|$ is a $T \times 1$ unit vector in the direction of j-th column of $\boldsymbol{X}$ and $\boldsymbol{W}_j$ is an arbitrarily chosen $T \times T-1$ matrix, subject to the constraint $\boldsymbol{W}_j\boldsymbol{W}_j' = \boldsymbol{I}_T - \boldsymbol{q}_j\boldsymbol{q}_j'$. Note that since the $T \times T$ orthogonal matrix $\boldsymbol{Q}_j = \left[\boldsymbol{q}_j|\boldsymbol{W}_j\right]$ is of full rank, the suggested rotation provides a one-to-one mapping between the original data $\boldsymbol{y}$ and the rotated data $\left(y_j^*, \widetilde{\boldsymbol{y}}_j'\right)'$. We

[1] This assumption implies posterior independence among coefficients, that is, $p\left(\boldsymbol{\beta}|\boldsymbol{y}\right) \equiv \prod_j p\left(\beta_j|\boldsymbol{y}\right)$. While such independence assumption can be very helpful for prediction, in Section 7 we show how, in the context of structural VAR inference, to modify our procedure in order to obtain the exact joint posterior.

show in Appendix A.1 that if we multiply both sides of (1) by $\boldsymbol{Q}_j$, after rearranging we obtain the following observationally equivalent regressions

$$\begin{aligned} y_j^* = &\quad \|\boldsymbol{X}_j\|\,\beta_j + \boldsymbol{X}_{(-j)}^*\boldsymbol{\beta}_{(-j)} + v_j^*, \\ \widetilde{\boldsymbol{y}}_j = &\qquad\qquad \widetilde{\boldsymbol{X}}_{(-j)}\boldsymbol{\beta}_{(-j)} + \widetilde{\boldsymbol{v}}_j, \end{aligned} \tag{3}$$

where $\boldsymbol{X}_{(-j)}^* = \boldsymbol{q}_j'\boldsymbol{X}_{(-j)}$ is a $1 \times (k-1)$ vector, $v_j^* = \boldsymbol{q}_j'\boldsymbol{v}$ is a scalar, $\widetilde{\boldsymbol{X}}_{(-j)} = \boldsymbol{W}_j'\boldsymbol{X}_{(-j)}$ is a $(T-1) \times (k-1)$ matrix, $\widetilde{\boldsymbol{v}}_{\boldsymbol{j}} = \boldsymbol{W}_j'\boldsymbol{v}$ is a $(T-1) \times 1$ vector, and $\boldsymbol{X}_{(-j)} = \boldsymbol{X} \setminus \boldsymbol{X}_j$ denotes the $k-1$ columns of $\boldsymbol{X}$ after its j-th column has been removed. Similarly, $\boldsymbol{\beta}_{(-j)} = \boldsymbol{\beta} \setminus \beta_j$ denotes the $k-1$ elements of $\boldsymbol{\beta}$ after its j-th element has been removed. It also follows that the joint likelihood of the rotated data $\left(y_j^*, \widetilde{\boldsymbol{y}}_j'\right)'$ can we represented as

$$\begin{bmatrix} y_j^* \\ \widetilde{\boldsymbol{y}}_j \end{bmatrix} \Bigg| \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}\left(\begin{bmatrix} \|\boldsymbol{X}_j\| \\ 0 \end{bmatrix} \beta_j + \begin{bmatrix} \boldsymbol{X}_{(-j)}^* \\ \widetilde{\boldsymbol{X}}_{(-j)} \end{bmatrix} \boldsymbol{\beta}_{(-j)}, \sigma^2 \boldsymbol{I}_T \right), \tag{4}$$

where, due to the orthogonality of $\boldsymbol{Q}_j = \left[\boldsymbol{q}_j | \boldsymbol{W}_j\right]$, the variance of the rotated data is still σ^2. Most importantly, the rescaled regression in (3) separates the scalar y_j^*, which depends on β_j, from the remaining $T-1$ observations $\widetilde{\boldsymbol{y}}_j$, which are conditionally independent of the effect of β_j. At the same time, the form of the rescaled likelihood in (4) implies that y_j^* and $\widetilde{\boldsymbol{y}}_j$ do not share covariance terms, which ultimately means that we can treat (3) as two conditionally separable regression models. Combined, these last two equations provide insights on how to devise a simple two-step OLS procedure to estimate β_j. First, regress $\widetilde{\boldsymbol{y}}_j$ on $\widetilde{\boldsymbol{X}}_{(-j)}$ to obtain estimates for $\boldsymbol{\beta}_{(-j)}$ and σ^2, namely $\widehat{\boldsymbol{\beta}}_{(-j)}$ and $\widehat{\sigma}^2$. Next, condition on the regression variance $\widehat{\sigma}^2$ and regress $\left(y_j^* - \boldsymbol{X}_{(-j)}^*\widehat{\boldsymbol{\beta}}_{(-j)}\right)$ on $\|\boldsymbol{X}_j\|$ to obtain an estimate for β_j. Note that the estimates that we obtain from this two-step procedure are numerically identical to the OLS estimates we would recover if working with the original regression model in (1).

We now exploit the form of the likelihood in (4), along with Bayes Theorem, to derive the following expression for the marginal posterior distribution $p\left(\beta_j | \boldsymbol{y}\right)$

$$\begin{aligned} p\left(\beta_j | \boldsymbol{y}\right) &= p\left(\beta_j | y_j^*, \widetilde{\boldsymbol{y}}_j\right) \\ &= \frac{p\left(\beta_j, y_j^* | \widetilde{\boldsymbol{y}}_j\right)}{p\left(y_j^* | \widetilde{\boldsymbol{y}}_j\right)} \\ &\propto p\left(y_j^* | \beta_j, \widetilde{\boldsymbol{y}}_j\right) p\left(\beta_j\right), \end{aligned} \tag{5}$$

where we have used the fact that $p\left(y_j^* | \widetilde{\boldsymbol{y}}_j\right)$ does not involve β_j, meaning it is simply a normalizing constant that can be removed, and also the result that $\widetilde{\boldsymbol{y}}_j$ does not convey any information about

β_j, i.e. $p\left(\beta_j|\widetilde{\boldsymbol{y}}_j\right) \equiv p\left(\beta_j\right)$. Equation (5) shows that thanks to the rotation in (2), the marginal posterior distribution of β_j is now proportional to the rotated conditional likelihood $p\left(y_j^*|\beta_j, \widetilde{\boldsymbol{y}}_j\right)$ and the prior $p\left(\beta_j\right)$.[2] While we postpone our discussion on the prior distribution until the next section, it is of immediate interest to derive an expression for $p\left(y_j^*|\beta_j, \widetilde{\boldsymbol{y}}_j\right)$, and this is where we now turn our attention.

Note that, from a Bayesian standpoint, this conditional likelihood function can be interpreted as the predictive distribution of the "out-of-sample" data y_j^* given the "in-sample" data $\widetilde{\boldsymbol{y}}_j$, after the parameters $\boldsymbol{\beta}_{(-j)}$ and σ^2 have been integrated out. Using standard results for Bayesian predictive analysis (Koop, 2003), we show in Appendix A.2 that under a natural conjugate prior for $(\boldsymbol{\beta}_{(-j)}, \sigma^2)$ it follows that[3]

$$\begin{aligned} p\left(y_j^*|\beta_j, \widetilde{\boldsymbol{y}}_j\right) &= \|\boldsymbol{X}_j\|\, \beta_j + t_{2\overline{d}}\left(\overline{\mu}_j, \overline{\tau}_j^2\right) \\ &\approx \|\boldsymbol{X}_j\|\, \beta_j + \mathcal{N}\left(\overline{\mu}_j, \overline{\tau}_j^2\right), \end{aligned} \tag{6}$$

where

$$\overline{\mu}_j = \boldsymbol{X}_{(-j)}^* \overline{\boldsymbol{\beta}}_{(-j)}, \tag{7}$$

and

$$\overline{\tau}_j^2 = \frac{\overline{\psi}_{(-j)}}{\overline{d}}\left(1 + \boldsymbol{X}_{(-j)}^* \overline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(-j)}} \boldsymbol{X}_{(-j)}^{*\prime}\right). \tag{8}$$

The exact formulas for the posterior moments $\overline{\boldsymbol{\beta}}_{(-j)}$, $\overline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(-j)}}$, $\overline{\psi}_{(-j)}$, and $\overline{d}$ are standard to derive, and are also provided in Appendix A.2. Two key remarks are in order. First, note that in equation (6) we have chosen to approximate a Student-t predictive distribution using a Normal distribution. An immediate question is how good an approximation this will be. Note that if σ^2 is known, then the formulas are exact. In other words, the rotated likelihood $p\left(y_j^*|\beta_j, \widetilde{\boldsymbol{y}}_j\right)$ is indeed normal with the moments specified above. When σ^2 is unknown then the approximation can still be quite accurate, and the accuracy will increases with the sample size.[4] Second, jointly

[2] One implicit assumption we will rely on throughout is that the elements of $\boldsymbol{\beta}$ need to be a-priori independent, that is, $p\left(\boldsymbol{\beta}\right) = \prod_{j=1}^{k} p\left(\beta_j\right)$. This is a standard assumption in Bayesian analysis using hierarchical or other priors (e.g. Minnesota prior), since it is generally quite hard to objectively specify prior beliefs on the coefficients' cross-correlations.

[3] While there are many alternative prior choices available for $(\boldsymbol{\beta}_{(-j)}, \sigma^2)$, we have chosen to rely on the natural conjugate prior because, among other things, it leads to proper posteriors for the regression parameters even when the number of parameters $(k-1)$ is larger than the total number of observations $(T-1)$, and at the same time leads to a closed-form expression for the conditional likelihood $p\left(y_j^*|\beta_j, \widetilde{\boldsymbol{y}}_j\right)$.

[4] This is related to the fact that a Student-t distribution with a sufficient number of degrees of freedom - typically 100 or more - converges to a Normal distribution.

equations (5) and (6) imply that it is now possible to compute the marginal posterior for β_j by solving a scalar linear regression model with normal data and known variance, $\overline{\tau}_j^2$. Most importantly, the fact that the variance of this regression is known and fixed implies that we can derive analytically the marginal posterior for β_j even for priors that would normally require time-consuming simulation methods. This is a key result that we exploit in Section 3 to compute simulation-free marginal posteriors for a host of hierarchical shrinkage priors.

The estimation steps resulting from the above analysis are summarized in Algorithm 1. While exact expressions depend on the choice of prior distribution, $p(\beta_j)$, here we give an example of how our algorithm would work with a generic prior.

Algorithm 1 Scalable posterior estimation algorithm

for $j = 1$ **to** k

Step 1: Prepare rotation matrices

- Compute $\boldsymbol{q}_j = \boldsymbol{X}_j / \|\boldsymbol{X}_j\|$
- Generate each element of $\boldsymbol{W}_j$ from $\mathcal{N}(0,1)$
- Create $\boldsymbol{Q}_j = [\boldsymbol{q}_j | \boldsymbol{W}_j]$, using QR decomposition to ensure orthogonality

Step 2: Apply rotation

- Compute rotated data y_j^* and $\widetilde{\boldsymbol{y}}_j$, $\boldsymbol{X}_{(-j)}^*$ and $\widetilde{\boldsymbol{X}}_{(-j)}$

Step 3: Estimate auxiliary regression

- Regress $\widetilde{\boldsymbol{y}}_j$ on $\widetilde{\boldsymbol{X}}_{(-j)}$, obtain moments of $p\left(\boldsymbol{\beta}_{(-j)} \middle| \sigma^2, \widetilde{\boldsymbol{y}}_j\right)$ and $p\left(\sigma^2 \middle| \widetilde{\boldsymbol{y}}_j\right)$ analytically
- Derive moments of rotated likelihood, $\overline{\mu}_j$ and $\overline{\tau}_j^2$, analytically

Step 4: Estimate parameter of interest

- Given $\overline{\mu}_j$ and $\overline{\tau}_j^2$, regress $\left(y_j^* - \overline{\mu}_j\right)$ on $\|\boldsymbol{X}_j\|$
- Obtain moments of $p(\beta_j | \boldsymbol{y})$ analytically

end for

3 Hierarchical shrinkage priors

We now turn our focus to the prior for β_j $(j = 1, ..., k)$ in (5). While the estimation algorithm that we have just described can be applied to a variety of priors (provided that the elements of $\boldsymbol{\beta}$ are a-priori independent), we focus here on the following class of adaptive hierarchical priors

for β_j,[5]

$$\begin{aligned} \beta_j |\, \lambda_j^2 &\sim \mathcal{N}\left(0, \lambda_j^2 \underline{V}_{\beta_j}\right), \\ \lambda_j^2 &\sim G, \end{aligned} \tag{9}$$

where $\underline{V}_{\beta_j}$ denotes the part of the prior scale parameter chosen by the researcher, while λ_j^2 (or its square root, λ_j, depending on the specification) is a random variable with its own prior distribution, G.[6] Two observations are in order. First, the hierarchical form of the prior shows that conditional on the idiosyncratic scale parameter λ_j^2, the j-the regression coefficient β_j has a normal prior distribution. Combined with the approximation in (6), this is the key element that will allow us to derive the posterior of β_j without resorting to simulation methods. Second, while the conditional prior for β_j is normal, the marginal prior of β_j, $p(\beta_j) = \int \mathcal{N}\left(0, \lambda_j^2 \underline{V}_{\beta_j}\right) dG\left(\lambda_j^2\right)$ ought not to be and, depending on the choice of G, can result in very different shapes, with possibly a large mass around zero and much heavier tails than a bell-shaped Normal prior, two features that will impose shrinkage in the regression model.

Within the class of adaptive hierarchical priors, we focus on three special cases for G, which in turn lead to three well-known Bayesian shrinkage estimators.

3.1 Normal-Jeffreys

The first choice of prior for λ_j^2 is a Jeffreys prior, i.e. $p\left(\lambda_j^2\right) \propto 1/\lambda_j^2$, which is fully uninformative about λ_j^2. Notice that this particular choice of prior for λ_j^2 leads to an improper marginal prior for β_j, i.e. $p(\beta_j) \propto |\beta_j|^{-1}$, a prior that is sharply peaked at zero and is similar to the popular Laplace prior, and therefore favors sparsity in the regression model (see for example Tipping, 2001; Figueiredo, 2003).

Thanks to the approximation in (6) and the conditional normality of the prior, it is straightforward to derive the marginal likelihood for y_j^* analytically. This takes the form

$$\begin{aligned} p\left(y_j^* |\, \lambda_j^2, \widetilde{\boldsymbol{y}}_j\right) &= \int p\left(y_j^* |\, \beta_j, \widetilde{\boldsymbol{y}}_j\right) p\left(\beta_j |\, \lambda_j^2\right) d\beta_j \\ &= \mathcal{N}\left(y_j^* |\, \overline{\mu}_j, \|\boldsymbol{X}_j\|^2 \lambda_j^2 \underline{V}_{\beta_j} + \overline{\tau}_j^2\right), \end{aligned} \tag{10}$$

[5] The assumption that the prior mean of β_j is zero is without loss of generality. All the results that follow can be trivially updated to allow for a non-zero prior mean.

[6] Alternatively, we could also refer to λ_j^2 as the local variance component. See for example Polson and Scott (2010).

where $\mathcal{N}(z|a,b)$ denotes the probability of a random variable z evaluated at a Normal distribution with mean a and variance b. Next, similar to the analysis of Giannone et al. (2015), we can choose the optimal shrinkage intensity λ_j^2 in (9) by maximizing (10), i.e.

$$\widehat{\lambda}_j^2 = \arg\max_{\lambda_j^2} \; p\left(y_j^* \mid \lambda_j^2, \widetilde{\boldsymbol{y}}_j\right). \tag{11}$$

We show in Appendix A.3 that the posterior estimate of λ_j^2 that maximizes the marginal likelihood takes the form

$$\widehat{\lambda}_j^2 = \max\left[0, \frac{\left(y_j^* - \overline{\mu}_j\right)^2 - \overline{\tau}_j^2}{\|\boldsymbol{X}_j\|^2 \underline{V}_{\beta_j}}\right]. \tag{12}$$

Finally, plugging the optimal shrinkage intensity $\widehat{\lambda}_j^2$ into (9) leads to the marginal posterior

$$p\left(\beta_j|\widehat{\lambda}_j^2, \boldsymbol{y}\right) \sim \mathcal{N}\left(\overline{\beta}_j, \overline{V}_{\beta_j}\right), \tag{13}$$

where both $\overline{\beta}_j$ and $\overline{V}_{\beta_j}$ depend on $\widehat{\lambda}_j^2$, and are given by

$$\overline{V}_{\beta_j} = \frac{\overline{\tau}_j^2 \widehat{\lambda}_j^2 \underline{V}_{\beta_j}}{\|\boldsymbol{X}_j\|^2 \widehat{\lambda}_j^2 \underline{V}_{\beta_j} + \overline{\tau}_j^2}, \qquad \overline{\beta}_j = \frac{\|\boldsymbol{X}_j\| \widehat{\lambda}_j^2 \underline{V}_{\beta_j} \left(y_j^* - \overline{\mu}_j\right)}{\|\boldsymbol{X}_j\|^2 \widehat{\lambda}_j^2 \underline{V}_{\beta_j} + \overline{\tau}_j^2}. \tag{14}$$

Notice, to conclude, that both the maximization in (12) and the prior moments in (14) only include scalar operations, so they are trivial to compute $\forall j \in [1, k]$.

3.2 Normal-Gamma

The second prior specification we consider within the class of hierarchical priors in (9) is the popular class of Normal-Gamma priors, studied in Griffin and Brown (2010) and extended to the VAR case by Huber and Feldkircher (forthcoming). This prior assumes that $\lambda_j^2 \sim \mathcal{G}(\underline{c}_1, \underline{c}_2)$, where $\underline{c}_1$ and $\underline{c}_2$ denote the shape and scale of the Gamma distribution $\mathcal{G}$. To see the effect of the hyperparameters $\underline{c}_1$ and $\underline{c}_2$ on the shape of the marginal prior for β_j, the bottom panels of Figure 1 plot the marginal distribution of β_j for two different choices of $\underline{c}_1$ and $\underline{c}_2$. As a benchmark to compare against, the top left panel of the figure plots the empirical distribution of the non-hierarchical version of (9), where $\lambda_j^2 = 1$ is non-stochastic and $\underline{V}_{\beta_j} = 10$.[7] The bottom left panel plots the marginal prior of β_j when G is the Gamma density and the hyperparameters

[7] For a large prior variance this can be considered a locally uninformative prior, while for small values of $\underline{V}_{\beta_j}$ it results in the ridge estimator.

are set to $\underline{c}_1 = 1$ and $\underline{c}_2 = 2$. As it can be seen from this panel, this choice of hyperparameters generates a marginal prior for β_j that, compared to the benchmark bell-shaped Normal prior in the top left panel of the figure, shrinks towards zero at a much faster rate. Next, the bottom right panel of the figure considers the case where $\underline{c}_1 = 0.1, \underline{c}_2 = 2$. This choice leads to a much more intense shrinkage, with a clear spike around zero and tails that are significantly heavier than a Normal density.[8]

We can proceed in an analogous manner as in the Normal-Jeffreys case, and choose the optimal shrinkage intensity by maximizing the posterior of λ_j^2,

$$\widehat{\lambda}_j^2 = \arg\max_{\lambda_j^2} \; p\left(y_j^* \mid \lambda_j^2, \widetilde{\boldsymbol{y}}_j\right) p\left(\lambda_j^2\right), \tag{15}$$

which, after taking logs, leads to the following maximization

$$\widehat{\lambda}_j^2 = \arg\max_{\lambda_j^2} \left\{ -\frac{1}{2}\ln\left(\overline{\tau}_j^2 + \|\boldsymbol{X}_j\|^2 \lambda_j^2 \underline{V}_{\beta_j}\right) - \frac{1}{2}\frac{\left(y_j^* - \overline{\mu}_j\right)^2}{\tau_j^2 + \|\boldsymbol{X}_j\|^2 \lambda_j^2 \underline{V}_{\beta_j}} + (\underline{c}_1 - 1)\ln\lambda_j^2 - \underline{c}_2 \lambda_j^2 \right\}. \tag{16}$$

Once again, this is a straightforward maximization over scalar quantities, hence trivial to compute. Finally, plugging the optimal shrinkage intensity $\widehat{\lambda}_j^2$ into (9) leads to a marginal posterior for β_j with moments as in (14).

3.3 Spike-and-Slab

The third specification we consider for our hierarchical prior is the popular Spike-and-Slab prior. While it is possible to cast this prior in the hierarchical form of (9) (see for example Griffin and Brown, 2010, p. 175), we follow the literature and write this prior as an explicit mixture of distributions

$$\begin{aligned} \beta_j|\lambda_j &\sim (1-\lambda_j)\,\delta_0 + \lambda_j \mathcal{N}\left(0, \underline{V}_{\beta_j}\right), \\ \lambda_j &\sim Bernoulli\left(\underline{\pi}_0\right), \end{aligned} \tag{17}$$

[8] Notice that the Normal-Jeffreys prior is not plotted in this figure because it is an improper prior for λ_j^2, and leads to a marginal prior for β_j that does not integrate to one (and, thus, cannot be represented graphically). However, following Tipping (2001) we can think of the Normal-Jeffreys prior as a special case of a Normal-Inverse Gamma (IG) mixture, with $\lambda_j^2 \sim \mathcal{IG}\left(\underline{\alpha}_1, \underline{\alpha}_2\right)$ where $\underline{\alpha}_1, \underline{\alpha}_2 \to 0$. The Normal-IG mixture is the typical representation of the Student-t distribution, which is more peaked at zero compared to the Normal distribution. Therefore, the shrinkage induced by a Normal-Jeffreys can be broadly thought of as the limit of a Student-t prior with very large (infinite in practice) variance.

where δ_0 is the Dirac delta function at zero, while λ_j is now a Bernoulli random variable with mean $\underline{\pi}_0$ which, in turn, denotes the prior proportion of non-zero regressors in the model. As noted by Griffin and Brown (2010), the Spike-and-Slab and Normal-Gamma priors can lead to very similar forms of shrinkage. It is in fact possible to elicit the prior hyperparameters $\underline{c}_1$ and $\underline{c}_2$ of the Normal-Gamma prior and the prior inclusion probability $\underline{\pi}_0$ of the Spike-and-Slab prior in a way to similarly constrain most of the variation in the priors to a small set of regressors. Figure 1 makes this point explicitly, where in the top right panel we show the marginal prior of β_j for the Spike-and-Slab case and $\underline{\pi}_0 = 0.5$ (as with the other three panels, we set $\underline{V}_{\beta_j} = 10$). As it can be seen, the Spike-and-Slab prior with $\underline{\pi}_0 = 0.5$ leads to a marginal prior for β_j that behaves very much like the Normal-Gamma case when $\underline{c}_1 = 0.1$ and $\underline{c}_2 = 2$ (bottom right panel), placing a considerable mass at zero and featuring very heavy tails.

It follows that the posterior of λ_j is of the same form, that is $\lambda_j | \boldsymbol{y} \sim Bernoulli\left(\widehat{\pi}_j\right)$, where

$$\widehat{\pi}_j = p\left(\lambda_j = 1 | \boldsymbol{y}\right) = \frac{p\left(y_j^* \middle| \ \lambda_j = 1, \widetilde{\boldsymbol{y}}_j\right) p\left(\lambda_j = 1\right)}{p\left(y_j^* \middle| \lambda_j = 0, \widetilde{\boldsymbol{y}}_j\right) p\left(\lambda_j = 0\right) + p\left(y_j^* \middle| \lambda_j = 1, \widetilde{\boldsymbol{y}}_j\right) p\left(\lambda_j = 1\right)} \tag{18}$$

where $\widehat{\pi}_j$ is the posterior probability of inclusion (PIP) of predictor j in the regression model (not to be confused with a "p-value" or "significance level"). We show in Appendix A.4 that $\widehat{\pi}_j$ simplifies to

$$\widehat{\pi}_j = \frac{\mathcal{N}\left(y_j^* \middle| \overline{\mu}_j, \overline{\tau}_j^2 + \|\boldsymbol{X}_j\|^2 \underline{V}_{\beta_j}\right) \underline{\pi}_0}{\mathcal{N}\left(y_j^* \middle| \overline{\mu}_j, \overline{\tau}_j^2\right)\left(1 - \underline{\pi}_0\right) + \mathcal{N}\left(y_j^* \middle| \overline{\mu}_j, \overline{\tau}_j^2 + \|\boldsymbol{X}_j\|^2 \underline{V}_{\beta_j}\right) \underline{\pi}_0} \tag{19}$$

Finally, in this case the marginal posterior of β_j is equal to

$$\begin{aligned} p\left(\beta_j | \boldsymbol{y}\right) &= \int p\left(\beta_j | \lambda_j, \boldsymbol{y}\right) p\left(\lambda_j | \boldsymbol{y}\right) d\lambda_j \\ &= p\left(\lambda_j = 0 | \boldsymbol{y}\right) p\left(\beta_j | \lambda_j = 0, \boldsymbol{y}\right) + p\left(\lambda_j = 1 | \boldsymbol{y}\right) p\left(\beta_j | \lambda_j = 1, \boldsymbol{y}\right) \\ &= \left(1 - \widehat{\pi}_j\right)\delta_0 + \widehat{\pi}_j \mathcal{N}\left(\overline{\beta}_j, \overline{V}_{\beta_j}\right) \end{aligned} \tag{20}$$

where $\overline{\beta}_j$ and $\overline{V}_{\beta_j}$ are again given by (14) in the special case when $\widehat{\lambda}_j = 1$.

4 Application to BVAR estimation

Up to this point, we have focused our exposition on a univariate regression model. Consider now the following n-dimensional VAR(p) model,

$$\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \ldots + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t, \quad t = 1, ..., T, \tag{21}$$

where $\boldsymbol{y}_t$ is an $n \times 1$ vector of time series of interest, $\boldsymbol{c}$ is an $n \times 1$ vector of intercepts, $\boldsymbol{A}_1, ..., \boldsymbol{A}_p$ are $n \times n$ matrices of coefficients on the lagged dependent variables, and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega})$, with $\boldsymbol{\Omega}$ an $n \times n$ covariance matrix. We next rewrite the original VAR model in (21) in a recursive form, which allows to estimate the VAR coefficients $\{\boldsymbol{c}, \boldsymbol{a}\}$ and the elements of the covariance matrix $\boldsymbol{\Omega}$ one equation at a time. This, in turns, allows us to readily apply the estimation method we presented in Section 2 to the VAR, by iterating recursively through a collection of univariate regressions.[9]

From a computational perspective, there are at least two ways one can re-write the reduced-form VAR in (21) as a recursive system. For example, Koop et al. (2017) rely on a recursive structural VAR representation. Here we use an alternative recursive form that is due to Carriero et al. (2016). We show in Appendix A.5 that the i-th equation of the VAR ($i = 1, ..., n$) can be written as

$$y_{i,t} = c_i + \boldsymbol{a}_{i,\cdot} \boldsymbol{Z}_t + \gamma_{i,1}\sigma_1 u_{1,t} + ... + \gamma_{i,i-1}\sigma_{i-1} u_{i-1,t} + \sigma_i u_{i,t}, \tag{22}$$

where c_i is the scalar intercept, $\boldsymbol{Z}_t = \left[\boldsymbol{y}'_{t-1}, ..., \boldsymbol{y}'_{t-p}\right]'$ is a $np \times 1$ vector containing all p lags of $\boldsymbol{y}_t$, $\boldsymbol{a}_{i,\cdot} = [a_{i,1}, ..., a_{i,np}]$ denotes the corresponding vector of coefficients, $u_{1,t}, ..., u_{i-1,t}$ and $\sigma_1, ..., \sigma_{i-1}$ are the VAR structural residuals and standard deviations from all the previous $i-1$ equations, and $\gamma_{i,1}, ..., \gamma_{i,i-1}$ their associated coefficients. Next, let $\boldsymbol{X}_{i,t} = (\boldsymbol{Z}'_t, \sigma_1 u_{1,t}, ..., \sigma_{i-1} u_{i-1,t})$ and rewrite (22) as

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_i + \boldsymbol{v}_i, \tag{23}$$

where $\boldsymbol{y}_i = (y_{i,t}, ..., y_{i,T})'$, $\boldsymbol{X}_i = \left(\boldsymbol{X}'_{i,1}, ..., \boldsymbol{X}'_{i,T}\right)'$, $\boldsymbol{\beta}_i = (c_i, \boldsymbol{a}_{i,\cdot}, \gamma_{i,1}, ..., \gamma_{i,i-1})'$, and $\boldsymbol{v}_i = (\sigma_i u_{i,1}, ..., \sigma_i u_{i,T})'$. With the i-th equation of the VAR now in the same form as (1), we can straightforwardly apply the algorithm in Section 2 tot the VAR, one equation at a time. Next, we can modify the hierarchical prior in (9) to work with the VAR i-th equation by re-writing it as follows

$$\begin{aligned} \beta_{ij} |\, \lambda_{ij}^2 &\sim \mathcal{N}\left(0, \lambda_{ij}^2 \underline{V}_{\beta_{ij}}\right), \\ \lambda_{ij}^2 &\sim G, \end{aligned} \tag{24}$$

[9] Following standard results in multivariate models, one can factorize the covariance matrix $\boldsymbol{\Omega}$ into a diagonal matrix of variance terms and a lower triangular matrix of covariance terms. This factorization allows the covariance terms to be treated as contemporaneous right-hand side predictors in each equation of the VAR and, because of the imposed recursive ordering, allows to estimate the VAR equation-by-equation; see Hausman (1983) for an early discussion of this approach.

where $j = 1, ..., k_i$ indexes the elements of $\boldsymbol{\beta}_i$, and $k_i = np + i$ denotes the total number of coefficients in the i-th equation of the VAR. One final point worth mentioning is that an added benefit of the procedure in (23)-(24) is that we can now apply our hierarchical shrinkage priors also to the coefficients $\gamma_{i,1}, ..., \gamma_{i,i-1}$, thus, explicitly providing shrinkage to the contemporaneous covariance elements in the VAR.

5 Monte Carlo analysis

In this section we evaluate numerically the new approach using simulated data. The purpose of this exercise is manifold. First, we want to assess the numerical precision of the new estimation method. We have already argued that if we apply OLS to the two-stage rotated regression in (3), we will obtain coefficients estimates that are identical to those we would obtain from the original regression problem in (1). However, it is important to evaluate whether the new estimation algorithm works well under a wide variety of Bayesian priors that will lead to biased penalized estimators. Second, we want to establish whether the three hierarchical priors introduced in Section 3 have good shrinkage properties when applied to a VAR setting and a finite amount of data. While the properties of such priors have been thoroughly examined and discussed in the literature, it is important to assess how the approximations we have introduced affect their performance. Finally, we want to obtain a measure of how well the proposed method fares against popular methods in recovering the true VAR coefficients.

5.1 Setup of Monte Carlo experiment

In order to investigate the importance of shrinkage as a function of the VAR size, we consider VARs of three different dimensions, that is, small ($n = 3$), medium ($n = 7$), and large ($n = 20$). For each VAR dimension, we generate 1,000 datasets with $T = 150$ observations each. In all three cases, we set the number of lags to $p = 2$. The data generating process is that of a sparse VAR, where we allow the sparsity pattern to be random. We first model the persistence of each variable in the VAR by setting the first own lag coefficient to be in the range $[0.4, 0.6]$, i.e.

$$\boldsymbol{A}_1 = diag\left(\rho_1, \rho_2, ..., \rho_n\right), \tag{25}$$

where $\rho_i \sim \mathcal{U}(0.4, 0.6)$, $i = 1, ..., n$. The coefficients on the subsequent own lags, $(A_l)_{i,i}$ are then generated according to the rule that $(A_l)_{i,i} = (A_1)_{i,i} / l^2$ $(l = 2, ..., p)$, implying a geometric decay

in their magnitudes, with the more distant lags having a lesser impact.[10] As for the coefficients on the other lags, we set them according to the following rule:

$$(A_l)_{i,j} = \begin{cases} \mathcal{N}\left(0, \sigma_A^2\right) & \text{with prob } \xi_A \\ 0 & \text{with prob } (1-\xi_A) \end{cases} \qquad l = 1, ..., p, \qquad i \neq j, \tag{26}$$

where $\xi_A \in (0,1)$ is the probability of obtaining a non-zero coefficient. We set $\sigma_A^2 = 0.1$ and calibrate the inclusion probability according the the VAR size by setting $\xi_A = 1/(n-1)$. This means, for example, that in a seven-variable VAR only 1/6 of the coefficients are expected to be non-zero. Next, we decompose the covariance matrix $\mathbf{\Omega}$ as $\mathbf{\Omega} = \mathbf{\Phi}\mathbf{\Phi}'$ where

$$\mathbf{\Phi} = \begin{bmatrix} 1 & 0 & ... & 0 \\ \varphi_{2,1} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \varphi_{n,1} & ... & \varphi_{n,n-1} & 1 \end{bmatrix}, \tag{27}$$

and generate the element of $\mathbf{\Phi}$ according to the following rule

$$\varphi_{i,j} = \begin{cases} \mathcal{U}(0,1) & \text{with prob } \xi_{\mathbf{\Phi}} \\ 0 & \text{with prob } 1-\xi_{\mathbf{\Phi}} \end{cases} \qquad i > j. \tag{28}$$

where we set $\xi_{\mathbf{\Phi}} = 0.5$.

Along with our proposed algorithm and the three priors described in Section 3 (**Normal-Jeffreys**; **Normal-Gamma**; **Spike-and-Slab**), we consider the following three competing estimation methods: OLS (**VAR**); hierarchical Minnesota shrinkage as in Giannone et al. (2015) (**BVAR-GLP**); stochastic search for VAR restrictions algorithm of George et al. (2008) (**SSVS**). The BVAR-GLP approach relies on Minnesota-type moments, so due to the fact that the generated VARs are all stationary we set the prior mean on the first own lag coefficient to 0.9. For all the remaining coefficients, we set the prior mean to zero (see Kadiyala and Karlsson, 1997, for a discussion of these choices). For consistency, we use the same prior means in all the other Bayesian approaches, including ours (that is, we modify the hierarchical prior in (24) to allow for a non-zero mean, which we denote with $\underline{\beta}_{ij}$). The remaining settings for the BVAR-GLP algorithm are the default ones suggested by the authors. As for the SSVS algorithm, we follow George et al. (2008) and set (using the authors' notation) the prior

[10] The relatively low value of ρ_i and the decay in the own lag coefficients is done to guarantee that all variables in the VAR are stationary. In practice, in all cases we examine the roots of the generated VAR coefficients and discard all simulated DGPs producing non-stationarity variables.

inclusion probabilities to $p_i = q_{ij} = 0.5$, and the prior variances to $R = R_j = I$, $\tau_0 = \kappa_0 = 0.1$ and $\tau_1 = \kappa_1 = 1$. As for the remaining details of our approach, we set the prior variance in (24) to $\underline{V}_{\beta_j} = 10$. Also, in the Spike-and-Slab case we set the prior inclusion probability for all predictors to $\underline{\pi}_0 = 0.5$, while in the Normal-Gamma case we set $\underline{c}_1 = 0.1$ and $\underline{c}_2 = 2$.[11]

5.2 Results

We begin by drawing attention to the estimated shrinkage intensity implied by our approach under the three different priors we considered. The top panels of Figure 2 and Figure 3 plot the empirical distribution of the average shrinkage intensity $\overline{\lambda}$ over the 1,000 Monte Carlo iterations for the three VAR sizes and for the Normal-Jeffreys and Normal-Gamma cases, respectively. In both figures, $\overline{\lambda} = \frac{1}{K}\sum_{i=1}^{n}\sum_{j=1}^{k_i}\widehat{\lambda}_{ij}$, where $K = \sum_{i=1}^{n} k_i$ denotes the total number of VAR coefficients, including the covariance terms in $\mathbf{\Phi}$. As one may expect, both in the case of the Normal-Jeffreys and the Normal-Gamma prior, the average shrinkage intensity becomes smaller as the VAR size increases, implying that more shrinkage is imposed in higher dimensions. This is a desirable feature of shrinkage estimation in VARs, and in line with previous findings in the literature; see Banbura et al. (2010) and their relevant discussion. This result is particularly clear in the case of the Normal-Gamma prior, where the empirical distribution of $\overline{\lambda}$ becomes more concentrated and informative as the VAR size increases.

A notable feature of our procedure is that it yields individualized shrinkage hyperparameters for each VAR coefficient, including the elements of the covariance matrix $\mathbf{\Phi}$. It would then be conceivable to expect that the VAR parameters which are equal to zero in the DGP should be accompanied by, on average, lower $\widehat{\lambda}_{ij}$'s. In order to verify this claim, the bottom panels of Figure 2 and Figure 3 plot the empirical distributions of the average shrinkage intensity $\overline{\lambda}$, after the individual $\widehat{\lambda}_{ij}$'s have been grouped according to whether the underlying VAR coefficients are equal to zero or not in the DGP. As expected, for both priors we find that the average shrinkage intensity of the zero VAR parameters (red bars) is significantly on the left of the average shrinkage intensity corresponding to non-zero VAR coefficients (blue bars). Notably, Figure 3 shows that for a large number Monte Carlo iterations, the average shrinkage intensity

[11] In all cases, intercepts are left unrestricted using a diffuse prior. Note also that for both the SSVS algorithm and our estimation algorithms, we allow for shrinkage estimation of the sparse covariance terms $\varphi_{i,j}$.

associated with the zero VAR coefficients is exactly zero, meaning that the hierarchical Gamma prior is capable of accurately flagging the irrelevant coefficients, shrinking all of them to zero. This result is more pronounced for the $n = 3$ and $n = 7$ VAR sizes, implying that for the larger $n = 20$ case, different values of the hyperparameters $\underline{c}_1, \underline{c}_2$ may be needed to achieve a similar result.

Figure 4 plots the distribution of the average posterior inclusion probabilities (PIPs) for the Spike-and-Slab prior, $\overline{\pi} = \frac{1}{K}\sum_{i=1}^{n}\sum_{j=1}^{k_i}\widehat{\pi}_{ij}$. In this case, due to the fact that there is a well-established alternative MCMC algorithm for VARs that relies on this prior, we contrast the results of our Spike-and-Slab hierarchical prior with those from the SSVS approach of George et al. (2008). In particular, the top panels of the figure plot the empirical distributions of $\overline{\pi}$ estimated with the SSVS algorithm, while the bottom panels plot the empirical distribution of $\overline{\pi}$ estimated using our algorithm and the Spike-and-Slab hierarchical prior. Once again, we separately plot the average PIPs corresponding to VAR parameters that are equal to zero (different from zero) in the DGP. As it can be seen from inspecting the figure, both algorithms are quite accurate at flagging which VAR parameters should be zero (or not), with the empirical distributions of the average PIPs from the zero VAR coefficients on the left of the corresponding non-zero coefficients' empirical distributions. Nevertheless, our algorithm performs visibly much better than the SSVS, with the estimated distributions being closer to zero and one (in the case of the SSVS algorithm, both distributions are close to 0.5 implying a decreased ability to determine if a VAR parameter is zero or not).

We next look at the effectiveness of the various methods in recovering the parameters of the true data generating process. To this end, for each of the approaches considered in this section, we compute the Mean Absolute Deviation (MAD), defined as

$$MAD^{(r,s)} = \frac{1}{K}\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left|\beta_{ij}^{(r)} - \widehat{\beta}_{ij}^{(r,s)}\right|, \tag{29}$$

where s denotes the method used (VAR, BVAR-GLP, SSVS, Normal-Jeffreys, Normal-Gamma, Spike-and-Slab), $r = 1, ..., 1,000$ keeps track of the MC simulations, K denotes the total number of lag coefficients in the VAR, $\beta_{ij}^{(r)}$ is the true VAR coefficient from the r-th simulation, and $\widehat{\beta}_{ij}^{(r,s)}$ denotes the (posterior mean of the) corresponding estimate according to method s. Figure 5 shows the quartiles and median of the MAD statistic over all 1,000 Monte Carlo iterations, by

means of box plots. For $n = 3$ the various shrinkage methods do not appear to improve much compared to OLS in recovering the true VAR parameters. However, as the VAR size increases, OLS begins to work less well. On the other hand, our estimation algorithm combined with the three hierarchical priors we introduced in Section 3 seems capable of accurately recovering the true VAR parameters, performing better than SSVS and as well or better than the BVAR-GLP method.

We conclude this section with a look at the computational loads of the various approaches considered thus far. For each of the Bayesian methods considered in this section, Table 1 reports the CPU time in seconds required to complete a single Monte Carlo iteration for the three VAR sizes, where lower values imply faster estimation. As it can be seen from the table, our approach is significantly faster than all the Bayesian alternatives considered in this section. In particular, thanks to the scalability of our posterior estimation algorithm, the CPU time of our proposed approach does not grow with the VAR size, a feature that makes it a particularly appealing tool to use with large dimensional VARs. We exploit this property of our approach in the next section.[12]

6 Macroeconomic forecasting

Combined with the speed-up in CPU times implied by our proposed simulation-free algorithm, the excellent properties of the hierarchical priors we introduced in Section 3 make them a very natural choice for a large dimensional VAR application. In this section, we investigate this claim empirically.

6.1 Data, models, and prior settings

We collect 40 quarterly variables for the US spanning the period 1959Q1 to 2015Q4.[13] The data, which are obtained from the Federal Reserve Economic Data (FRED) and are available

[12] We should note that in this comparison the various competing methods were tuned to be as fast as possible so, in practice, their quoted times in Table 1 could be higher than showed here. In particular, we have relied on the analytical version of the BVAR-GLP algorithm (their simulation-based version provided quantitatively similar results). As for the SSVS algorithm, we have set the total number of MCMC iterations (including burn-in) to 1,100, which we found to be sufficient with the simulated data but possibly on the low side with real-data applications.

[13] For the variables which are originally observed at the monthly frequency, we transform them into quarterly series by computing the average of their monthly values within each quarter.

at https://fred.stlouisfed.org, cover a wide range of key macroeconomic variables that applied economists monitor regularly, such as different measures of output, prices, interest and exchange rates, and stock market performance. We provide a full list of the data and their transformations in order to achieve stationarity in Appendix B. Out of the 40 series, we further distinguish seven "variables of interest", that is, key variables of interest which we will inspect very closely in order to evaluate how well the different models perform. These variables are: real gross domestic product (GDP), GDP deflator (GDPDEFL), and federal funds rate (FEDFUNDS), total employment (PAYEMS), unemployment rate (UNRATE), consumer prices (CPIAUCSL), and the 10-year rate on government securities (GS10).

We estimate VARs of three different sizes: medium (only the seven variables of interest), large (variables in medium plus an additional 13), and X-large (variables in large plus an additional 20), that is, we consider seven, 20 and 40-variable VARs. All VARs have a maximum of $p = 5$ lags. For each model size, we estimate a range of different models. In addition to the three hierarchical priors estimated using our simulation-free method, which we denote as N-J (Normal-Jeffreys), SNS (Spike-and-Slab), and N-G (Normal-Gamma), we consider five established methods for dealing with VARs of possibly large dimensions. The first two methods are based on the Minnesota prior with optimal tuning of its shrinkage, one allows for Bayesian variable selection and model averaging, and two methods rely on factor shrinkage. In particular, we denote as BVAR-BGR the model of Banbura et al. (2010) who optimize the Minnesota shrinkage hyperparameter using a grid, while we denote as BVAR-GLP the model of Giannone et al. (2015) who introduce a hierarchical prior on the same Minnesota shrinkage hyperparameter and derive its posterior update formula.[14] As a representative of simulation-based hierarchical shrinkage, we consider the stochastic restrictions search algorithm of George et al. (2008), which we denote as SSVS. This algorithm is based on a mixture shrinkage prior, similar to the Spike-and-Slab prior we introduced in Section 3. Finally, we use a dynamic factor model (denoted DFM), and a factor augmented VAR (denoted FAVAR); see Stock and Watson (2002) and Bernanke et al. (2005).

[14]Both the BVAR-BGR and BVAR-GLP approaches approximate inference using a natural conjugate prior which, as explained in the Introduction, has the disadvantage of symmetry across VAR equations, but the big advantage of leading to analytical expressions for the posterior moments of the VAR coefficients.

For the sake of comparability, whenever possible, we use the same exact prior settings. In particular, all Bayesian VAR models (including our three hierarchical prior and the SSVS) feature the same Minnesota-based prior moments, which we write as

$$\underline{\beta}_{ij} = \begin{cases} 0.9 & \text{if own first lag} \\ 0 & \text{otherwise} \end{cases}, \qquad \underline{V}_{\beta_{ij}} = \begin{cases} \frac{1}{l_{ij}^2} & \text{if own lags} \\ \frac{\psi \times \widehat{\sigma}_i^2}{l_{ij}^2 \times \widehat{\sigma}_k^2} & \text{otherwise} \end{cases}, \tag{30}$$

where $i = 1, ..., n$, $j = 2, ..., np + 1$, $\widehat{\sigma}_i^2$ ($\widehat{\sigma}_k^2$) is the OLS estimate of the variance of an AR(p) model on y_{it} (y_{kt}), $l_{ij} = \lfloor j/i \rfloor$ is the lag-length associated with the coefficient β_{ij} in the VAR, and ψ is a hyperparameter that allows coefficients of variable k showing up in VAR equation i ($i \neq k$) to shrink differently than own coefficients (k denotes the variable that the β_{ij} coefficient belongs to, i.e. $k = j - n\,(l_{ij} - 1)$).[15,16] Next, note that both in the BVAR-BGR and BVAR-GLP models the shrinkage intensity is the same across all VAR coefficients i.e., using the notation in (24), $\lambda_{ij}^2 = \lambda^2$, so these two priors do not allow for adaptive shrinkage. In addition, in the BVAR-BGR case we follow Banbura et al. (2010) and use a wide grid of possible λ^2 values, while in the BVAR-GLP case the choice of the optimal shrinkage intensity is fully automatic.[17] In contrast, the SSVS prior of George et al. (2008) and our three hierarchical priors, N-J, SNS and N-G, do allow separate shrinkage intensities λ_{ij}^2. The other shrinkage hyperparameter ψ is set in all models to be a function of the VAR size, with $\psi = 0.001$ for the medium and large VARs, and $\psi = 0.0001$ for the X-large VAR (note that the BVAR-BGR and BVAR-GLP models require $\psi = 1$). The remaining prior settings for the SSVS SNS, N-J, and N-G priors are: $\pi_0 = 0.1$, that is, our prior expectation is that only 10% of VAR coefficients are non-zero; $\underline{c}_1 = 0.1$, and $\underline{c}_1 = 2$. As for the prior hyperparameters specific to the SSVS we also set, using notation from George et al. (2008), $\tau_0 = \kappa_0 = 0.001$ and $\tau_1 = \kappa_1 = 10$. Finally, the DFM and FAVAR are estimated using principal components of the factors and a non-informative prior. We use the Bayesian information criterion (BIC) to select the optimal number of factors (minimum allowed is 1 and maximum is $\lfloor \sqrt{n} \rfloor$, with n the VAR size) and the optimal number of lags (ranging from one to five).

[15] We denote with $\lfloor x \rfloor$ the floor of x, i.e. the largest integer less than or equal to x.

[16] Both the intercepts and the elements of $\mathbf{\Phi}$ are left unrestricted with flat and uninformative priors, i.e. $\underline{\beta}_{ij} = 0$ and $\underline{V}_{\beta_{ij}} = 10$, $i = 1, ..., n$, $j = 1, np + 2, ..., k_i$.

[17] The BVAR-GLP approach allows alternative prior variants, such as the sum-of-coefficients prior. We have estimated a number of these variants and, with the exception of the sum-of-coefficients prior, by and large the results do not change significantly. As expected with the stationary data we use, the sum-of-coefficients prior does not work particularly well.

6.2 Measuring predictive accuracy

We use the first twenty five years of data, 1959:Q3–1984:Q4, to obtain initial parameter estimates for all models, which are then used to predict outcomes from 1985:Q1 ($h = 1$) to 1985:Q4 ($h = 4$). The next period, we include data for 1985:Q1 in the estimation sample, and use the resulting estimates to predict the outcomes from 1985:Q2 to 1986:Q1. We proceed recursively in this fashion until 2015:Q4, thus generating a time series of point and density forecasts for each forecast horizon h, with $h = 1, ..., 4$.[18]

Next, for each of the seven key variables listed above we summarize the precision of the h-step-ahead point forecasts for model i, relative to that from a benchmark VAR(p^*), by means of the ratio of MSFEs:

$$MSFE_{ijh} = \frac{\sum_{\tau=\underline{t}}^{\bar{t}-h} e_{i,j,\tau+h}^2}{\sum_{\tau=\underline{t}}^{\bar{t}-h} e_{bcmk,j,\tau+h}^2}, \tag{31}$$

where the benchmark VAR(p^*) has flat prior and is estimated using OLS, p^* denotes the largest lag length that can be estimated in a VAR with OLS and the data at hand, $\underline{t}$ and $\bar{t}$ denote the start and end of the out-of-sample period, and $e_{i,j,\tau+h}^2$ and $e_{bcmk,j,\tau+h}^2$ are the squared forecast errors of variable j at time τ and forecast horizon h associated with model i ($i \in$ {DFM,FAVAR,BVAR-BGR,BVAR-GLP,SSVS,N-J,SNS,N-G}) and the VAR(p^*) model, respectively.[19] The point forecasts used to compute the forecast errors are obtained by averaging over the draws from the various models' h-step-ahead predictive densities. Values of $MSFE_{ijh}$ below one suggest that model i produces more accurate point forecasts than the VAR(p^*) benchmark for variable j and forecast horizon h.

We also assess the accuracy of the point forecasts of the various methods using the multivariate loss function of Christoffersen and Diebold (1998). Specifically, we compute the ratio between the multivariate weighted mean squared forecast error (WMSFE) of model i and the WMSFE of the benchmark VAR(p^*) model as follows:

$$WMSFE_{ih} = \frac{\sum_{\tau=\underline{t}}^{\bar{t}-h} we_{i,\tau+h}}{\sum_{\tau=\underline{t}}^{\bar{t}-h} we_{bcmk,\tau+h}}, \tag{32}$$

where $we_{i,\tau+h} = \left(e'_{i,\tau+h} \times W \times e_{i,\tau+h}\right)$ and $we_{bcmk,\tau+h} = \left(e'_{bcmk,\tau+h} \times W \times e_{bcmk,\tau+h}\right)$ are time

[18] Note that when $h > 1$, point forecasts are iterated and predictive simulation is used to produce the predictive densities.

[19] That is, $p = 5$ for the medium VAR, $p = 2$ for the large VAR, and $p = 1$ for the X-large VAR.

$\tau + h$ weighted forecast errors of model i and the benchmark model, $e_{i,\tau+h}$ and $e_{bcmk,\tau+h}$ are either the (3×1) or the (7×1) vector of forecast errors for the key series we focus on, and W is either a (3×3) or a (7×7) matrix of weights. Following Carriero et al. (2011), we set the matrix W to be a diagonal matrix featuring on the diagonal the inverse of the variances of the series to be forecast.

As for the quality of the density forecasts, we follow Geweke and Amisano (2010) and compute the average log predictive likelihood differential between model i and the VAR(p^*) benchmark,

$$ALPL_{ijh} = \frac{1}{\bar{t} - \underline{t} - h + 1} \sum_{\tau=\underline{t}}^{\bar{t}-h} \left(LPL_{i,j,\tau+h} - LPL_{bcmk,j,\tau+h} \right), \tag{33}$$

where $LPL_{i,j,\tau+h}$ ($LPL_{bcmk,j,\tau+h}$) denotes model i's (benchmark's) log predictive score of variable j, computed at time $\tau + h$, i.e., the log of the h-step-ahead predictive density evaluated at the outcome. Positive values of $ALPL_{ijh}$ indicate that for variable j and forecast horizon h on average model i produces more accurate density forecasts than the benchmark model.

Finally, in order to test the statistical significance of differences in point and density forecasts, we consider pairwise tests of equal predictive accuracy (henceforth, EPA; Diebold and Mariano, 1995; West, 1996) in terms of MSFE, WMSFE, and ALPL. All EPA tests we conduct are based on a two sided test with the null hypothesis being the VAR(p^*) benchmark. We use standard normal critical values. Based on simulation evidence in Clark and McCracken (2013), when computing the variance estimator which enters the test statistic we rely on serial correlation robust standard errors, and incorporate the finite sample correction due to Harvey et al. (1997). In the tables, we use ***, ** and * to denote results which are significant at the 1%, 5% and 10% levels, respectively, in favor of the model listed at the top of each column.

6.3 Forecasting results

We now present results on the short-term forecasting performance of the various methods described above, based on the model sizes and forecast metrics outlined in the previous subsections. Table 2 shows relative $WMSFE$ statistics using all seven series of interest (right panel), as well as a smaller subset comprising three key variables (left panel), namely GDP, GDPDEFL, and FEDFUNDS. We find that, for both sets of variables and across all three

VAR dimensions, the hierarchical Spike-and-Slab (SNS) and Normal-Gamma (N-G) priors clearly dominate all other methods. As we saw in Figure 1, these two priors are very closely related. It is therefore not surprising to see that the forecasts implied by both priors are highly correlated, as indicated by their WMSFE values. Interestingly, the forecasts implied by the hierarchical Normal-Jeffreys (N-J) prior appear to be somewhat less accurate, even though in many cases they end only very slightly below the N-G and SNS ones. Overall, for all three hierarchical priors the forecast gains are quite substantial compared to the benchmark VAR model, exceeding 50% improvements in terms of $WMSFE$ for a number of horizons. Gains relative to alternative methods are also quite large, in several instances achieving improvements of 20% or more over the two BVAR methods, and 10% or more over the two factor-based methods. Another indication that the adaptive shrinkage imposed by the proposed hierarchical priors works well is that forecast performance improves as the VAR size increases (this pattern is particularly clear at horizons $h = 1, 2$). Finally, as we move from the medium to the large and to the X-large VARs, for most of the alternative BVAR and factor methods, the $WMSFEs$ tend to become larger implying larger estimation error resulting from over-parametrization and their reduced efficiency in shrinking irrelevant coefficients. All in all, the point-forecast accuracy of the three hierarchical priors proposed in this paper is excellent.

Table 3, Table 4 and Table 5 expand the analysis of the point forecast performance to each of the seven series of interest, separately for each of the three VAR sizes. As it was the case with the $WMSFE$ metrics, in the vast majority of cases the three hierarchical priors dominate, with improvements for some of the individual variables that are at times quite substantial. For example, for the FEDFUNDS series the improvements reach 70% (relative to the benchmark VAR). Also, in a few cases the SSVS method outperforms all the alternatives, especially in the case of the CPIAUCSL and GDPDEFL variables. Finally, it appears that in the large and X-large cases the DFM and FAVAR methods lead the way for some variables, especially at $h = 3$ and at $h = 4$, but overall their MSFEs are not significantly different from the benchmark, as measured by the Diebold-Mariano statistic.

Table 6, Table 7, and Table 8 shed lights on the quality of the density forecasts, by reporting averages of log predictive likelihoods ($ALPLs$) for all three VAR sizes, separately for each of

the seven variables of interest and the four forecast horizons. Results appear more mixed in this case, with no single method emerging as a clear winner. In particular, the three hierarchical priors appear to work particularly well for the PAYEMS and GS10 series, with improvements over the benchmark VAR method that in the latter case are often statistically significant. On the other hand, the BVAR-GLP method performs best with the GDPDEFL series, while the BVAR-BGR approach tends to work particularly well with the FEDFUNDS series. Overall, the fact that there is no clear winner when looking at the accuracy of the density forecasts should not come as a surprise, as all methods considered, including the benchmark VAR, do not differ in their treatment of the diagonal elements of $\boldsymbol{\Omega}$ in (21). Still, it is reassuring to find that the improvements in point forecasts we observed in Table 3 through Table 5 for the three hierarchical priors are not accompanied by any obvious deterioration in the quality of the density forecasts.

7 Structural VARs and impulse response analysis

The excellent forecast performance of our methodology is in line with a vast literature in statistics that praises the use of hierarchical priors for providing successful regularized estimation. As explained in Section 2, we have paired such priors with a fast approximate procedure that provides as output a joint parameter posterior $p(\boldsymbol{\beta}|\, \boldsymbol{y})$ under the assumption that all the elements of the vector $\boldsymbol{\beta}$ are a-posteriori uncorrelated. This approximation appears to be quite satisfactory in the high-dimensional forecasting application we have considered, where the final outcome of interest is simply a set of predictions for some economic variables of interest.

In addition to forecasting, VARs are also used regularly to identify structural shocks and assess the transmission mechanisms of the macro-economy through impulse response analysis and historical decompositions. In these cases, the assumption of a-posteriori independence may hinder the ability of the economist to provide reliable policy recommendations. In this section, we present a simple modification of our algorithm that is better suited for structural analysis.

In order to demonstrate this procedure, we follow papers such as Giannone et al. (2015) and generate 500 artificial datasets of $T = 216$ quarters from a large-scale dynamic stochastic general equilibrium (DSGE) model. The model we use is an extension of Görtz and Tsoukalas (2017) and

Görtz et al. (2016), and focuses on sectoral total factor productivity (TFP) shocks and financial frictions.[20] Among all possible sectoral and aggregate variables that this model generates, we focus only on the aggregate ones, to stay consistent with the bulk of the news shock literature.[21] In particular, we follow Barsky and Sims (2011) and use TFP, real GDP, consumption, and hours as our four variables of interest; in addition, to better identify news shocks, we include three additional series from the DSGE model, namely inflation, interest rate spread (the difference between long-term and short-term interest rates), and equity prices.[22] Finally, as the news shocks are not directly observed in a VAR setting, we rely on the identification scheme of Forni et al. (2014) to extract them.[23]

For each of the 500 datasets, we use the artificial data on the seven variables listed above to estimate a VAR with flat priors and a hierarchical prior BVAR. In particular, to estimate the latter model we rely on a simple two-stage procedure. In the first step of this procedure, we use the estimation algorithm described in Section 4 along with the hierarchical Spike-and-Slab prior to obtain posterior inclusion probabilities $\widehat{\pi}_{ij}$ for each of the VAR coefficients. Next, in the second step, we insert the restrictions implied by the posterior inclusion probabilities in a BVAR, which is estimated using an independent Normal-Wishart prior.[24]

Figure 6 plots the DSGE theoretical impulse responses to a productivity news shock, along with the average across the 500 replications of the median impulse responses for the flat prior (VAR) and our hierarchical prior (BVAR).[25] In general, both the VAR and BVAR models seem

[20] More specifically, we generate the artificial data using the default parameter settings that Görtz et al. (2016) use when financial frictions are present.

[21] See Beaudry and Portier (2013) for an excellent review of empirical studies on news and business cycles.

[22] Forni and Gambetti (2014) have shown that many of the smaller VARs considered in this literature are non-fundamental, meaning that they will not recover news shocks correctly. On the other hand, Beaudry et al. (2015) have argued that even non-fundamental VARs can correctly recover the responses of TFP to news shock. Regardless of this, larger information sets are still needed in order to identify correctly the remaining responses of interest to policy-makers, namely, output, consumption and hours.

[23] The identification scheme of Forni et al. (2014) relies on a combination of long and short-run restrictions on TFP. The alternative identification schemes proposed in Barsky and Sims (2011) and Francis et al. (2014) produce identical results.

[24] In particular, we start from (23) and rewrite the VAR in (21) in its SUR form. Using notation from Koop and Korobilis (2010), we rewrite the reduced-form VAR in (21) as $\boldsymbol{Y} = \widetilde{\boldsymbol{X}}\boldsymbol{B} + \boldsymbol{V}$ where $\boldsymbol{Y} = (\boldsymbol{y}_1', ..., \boldsymbol{y}_n')'$ and $\boldsymbol{V} = (\boldsymbol{v}_1', ..., \boldsymbol{v}_n')'$ are $Tn \times 1$ vectors, while $\widetilde{\boldsymbol{X}}$ is a $Tn \times K$ block-diagonal matrix obtained by stacking together the $T \times k_i$ matrices $\widetilde{\boldsymbol{X}}_1$,...,$\widetilde{\boldsymbol{X}}_n$ that incorporate the constraints implied by the estimated PIPs in (19). The elements in the generic matrix $\widetilde{\boldsymbol{X}}_i$ $(i = 1, ..., n)$, in turn, are computed by multiplying each row of $\boldsymbol{X}_i$ by $\widehat{\boldsymbol{\pi}}_i$, the $k_i \times 1$ vector of PIPs estimated from the VAR's i-th equation, i.e. $\widetilde{\boldsymbol{X}}_{i,t} = \boldsymbol{X}_{i,t} \circ \widehat{\boldsymbol{\pi}}_i'$, where $\circ$ denotes the Hadamard product, and $t = 1, ..., T$.

[25] Interestingly, the shape of the responses of output and consumption have a distinct double-hump shape. This

to capture fairly well the responses of output, consumption and hours. On the other hand, news shock in the DSGE model are anticipated 12 quarters ahead, therefore the response of TFP is zero for the first 12 periods. Such feature is generally harder to capture with a VAR or BVAR. Nevertheless, the empirical responses of TFP of both models are still quite reasonable, and in line with the VAR estimates reported elsewhere (Barsky and Sims, 2011). Next, Figure 7 provides a more accurate assessment of the differences in the estimated impulse responses. For each of the 500 replications, we compute the difference between the theoretical DSGE response and the estimated VAR and BVAR median responses, across the seven variables and the 40 horizons. Then, for each variable and horizon, we take the average of the squared errors across replications (MSE). Figure 7 plots the ratio between the MSE of the VAR with flat priors and the MSE of the hierarchical BVAR. As it can be seen from the figure, for the vast majority of periods the MSE ratios are higher than one, implying that the two-step BVAR procedure based on the hierarchical Spike-and-Slab prior generates more accurate responses than the flat prior VAR.

8 Conclusions

We have introduced a novel methodology for estimating BVARs which features a number of desirable properties, including scalability, flexibility, and computational efficiency. We exploited the flexibility of this novel approach to study empirically the benefits of a wide class of hierarchical shrinkage priors that lead to individualized adaptive shrinkage on the VAR coefficients. Thanks to the estimation method we introduced, we are able to derive analytical expressions for the marginal posteriors implied by three popular cases of hierarchical priors, namely Normal-Jeffreys, Spike-and-Slab, and Normal-Gamma. Our approach works extremely well with BVARs of both medium and large dimensions, delivering analytical approximations to the marginal posterior distributions of the BVAR coefficients that are very accurate. In addition, our proposed algorithm for posterior inference is multiple times faster than existing Bayesian VAR methods that rely on simulation methods. We implement a thorough Monte Carlo analysis to quantify the benefits of our approach, and find that it can recover very

is the direct consequence of working with a model with financial intermediation; see Figure 10 of Görtz et al. (2016) for more details.

accurately the underlying VAR coefficients. We also demonstrate, using an extensive forecasting application, the benefits of our adaptive shrinkage procedure in preventing over-fitting of large VARs and providing excellent forecasting performance. Finally, we demonstrate using a simulated numerical example with artificial data extracted from a large structural macroeconomic model, that our algorithm can be useful also in recovering structural impulse responses.

References

Andersson, M. K. and S. Karlsson (2008): "Bayesian forecast combination for VAR models," in *Bayesian Econometrics (Advances in Econometrics, vol. 23)*, ed. by S. Chib, W. Griffiths, G. Koop, and D. Terrell, Emerald Group, 501–524.

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

Barsky, R. B. and E. R. Sims (2011): "News shocks and business cycles," *Journal of Monetary Economics*, 58, 273 – 289.

Beaudry, P., P. Fve, A. Guay, and F. Portier (2015): "When is Nonfundamentalness in VARs a Real Problem? An Application to News Shocks," Working Paper 21466, National Bureau of Economic Research.

Beaudry, P. and F. Portier (2013): "News Driven Business Cycles: Insights and Challenges," Working Paper 19411, National Bureau of Economic Research.

Bernanke, B. S., J. Boivin, and P. Eliasz (2005): "Measuring the effects of monetary policy: A factor-augmented vector autoregressive (favar) approach," *The Quarterly Journal of Economics*, 120, 387–422.

Carriero, A., T. Clark, and M. Marcellino (2016): "Large Vector Autoregressions with stochastic volatility and flexible priors," Working paper 16-17, Federal Reserve Bank of Cleveland.

Carriero, A., G. Kapetanios, and M. Marcellino (2011): "Forecasting large datasets with Bayesian reduced rank multivariate models," *Journal of Applied Econometrics*, 26, 735–761.

——— (2012): "Forecasting government bond yields with large Bayesian vector autoregressions," *Journal of Banking and Finance*, 36, 2026 – 2047.

Christoffersen, P. F. and F. X. Diebold (1998): "Cointegration and Long-Horizon Forecasting," *Journal of Business & Economic Statistics*, 16, 450–458.

Clark, T. and M. McCracken (2013): "Advances in Forecast Evaluation," in *Handbook of Economic Forecasting*, ed. by A. Timmermann and G. Elliott, Elsevier, vol. 2, chap. 20, 1107–1201.

De Mol, C., D. Giannone, and L. Reichlin (2008): "Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?" *Journal of Econometrics*, 146, 318–328.

Del Negro, M. and F. Schorfheide (2004): "Priors from General Equilibrium Models for VARS," *International Economic Review*, 45, 643–673.

Diebold, F. X. and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263.

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

Figueiredo, M. A. T. (2003): "Adaptive sparseness for supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 1150–1159.

Forni, M. and L. Gambetti (2014): "Sufficient information in structural VARs," *Journal of Monetary Economics*, 66, 124–136.

Forni, M., L. Gambetti, and L. Sala (2014): "No News in Business Cycles," *The Economic Journal*, 124, 1168–1191.

Francis, N., M. T. Owyang, J. E. Roush, and R. DiCecio (2014): "A Flexible Finite-Horizon Alternative to Long-Run Restrictions with an Application to Technology Shocks," *The Review of Economics and Statistics*, 96, 638–647.

Gefang, D. (2014): "Bayesian doubly adaptive elastic-net Lasso for VAR shrinkage," *International Journal of Forecasting*, 30, 1–11.

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

Geweke, J. and G. Amisano (2010): "Comparing and evaluating Bayesian predictive distributions of asset returns," *International Journal of Forecasting*, 26, 216 – 230.

Giannone, D., G. E. Primiceri, and M. Lenza (2015): "Prior Selection for Vector Autoregressions," *Review of Economics and Statistics*, 97, 436–451.

Görtz, C. and J. D. Tsoukalas (2017): "News and Financial Intermediation in Aggregate Fluctuations," *The Review of Economics and Statistics*, 99, 514–530.

Görtz, C., J. D. Tsoukalas, and F. Zanetti (2016): "News Shocks under Financial Frictions," .

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

Harvey, D., S. Leybourne, and P. Newbold (1997): "Testing the equality of prediction mean squared errors," *International Journal of Forecasting*, 13, 281 – 291.

Hausman, J. A. (1983): "Chapter 7 Specification and estimation of simultaneous equation models," *Handbook of Econometrics*, 1, 391 – 448.

Hobert, J. P. and G. Casella (1996): "The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models," *Journal of the American Statistical Association*, 91, 1461–1473.

Huber, F. and M. Feldkircher (forthcoming): "Adaptive Shrinkage in Bayesian Vector Autoregressive Models," *Journal of Business and Economic Statistics*, 1 – 33.

Ingram, B. F. and C. H. Whiteman (1994): "Supplanting the 'Minnesota' prior: Forecasting macroeconomic time series using real business cycle model priors," *Journal of Monetary Economics*, 34, 497–510.

Ishwaran, H. and J. S. Rao (2005): "Spike and slab variable selection: Frequentist and Bayesian strategies," *Ann. Statist.*, 33, 730–773.

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

Koop, G. (2003): *Bayesian Econometrics*, John Wiley & Sons, Ltd.

Koop, G. and D. Korobilis (2010): "Bayesian Multivariate Time Series Methods for Empirical Macroeconomics," *Foundations and Trends in Econometrics*, 3, 267–358.

Koop, G., D. Korobilis, and D. Pettenuzzo (2017): "Bayesian Compressed Vector Autoregressions," Working paper.

Koop, G. and S. Potter (2004): "Forecasting in dynamic factor models using Bayesian model averaging," *Econometrics Journal*, 7, 550–565.

Korobilis, D. (2013): "VAR Forecasting Using Bayesian Variable Selection," *Journal of Applied Econometrics*, 28, 204–230.

Litterman, R. B. (1979): "Techniques of forecasting using vector autoregressions," Working Papers 115, Federal Reserve Bank of Minneapolis.

Mitchell, T. J. and J. J. Beauchamp (1988): "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1032.

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

Polson, N. and J. Scott (2010): "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction," in *Bayesian Statistics*, Oxford University Press, vol. 9, 1–24.

Primiceri, G. E. (2005): "Time Varying Structural Vector Autoregressions and Monetary Policy," *The Review of Economic Studies*, 72, 821–852.

Smith, M. and R. Kohn (2002): "Parsimonious Covariance Matrix Estimation for Longitudinal Data," *Journal of the American Statistical Association*, 97, 1141–1153.

Stock, J. H. and M. W. Watson (2002): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business & Economic Statistics*, 20, 147–162.

——— (2006): "Forecasting with Many Predictors," Elsevier, vol. 1 of *Handbook of Economic Forecasting*, 515 – 554.

Strachan, R. W. and H. K. V. Dijk (2013): "Evidence On Features Of A Dsge Business Cycle Model From Bayesian Model Averaging," *International Economic Review*, 54, 385–402.

Tibshirani, R. (1994): "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Tipping, M. E. (2001): "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Machine Learing Research*, 1, 211–244.

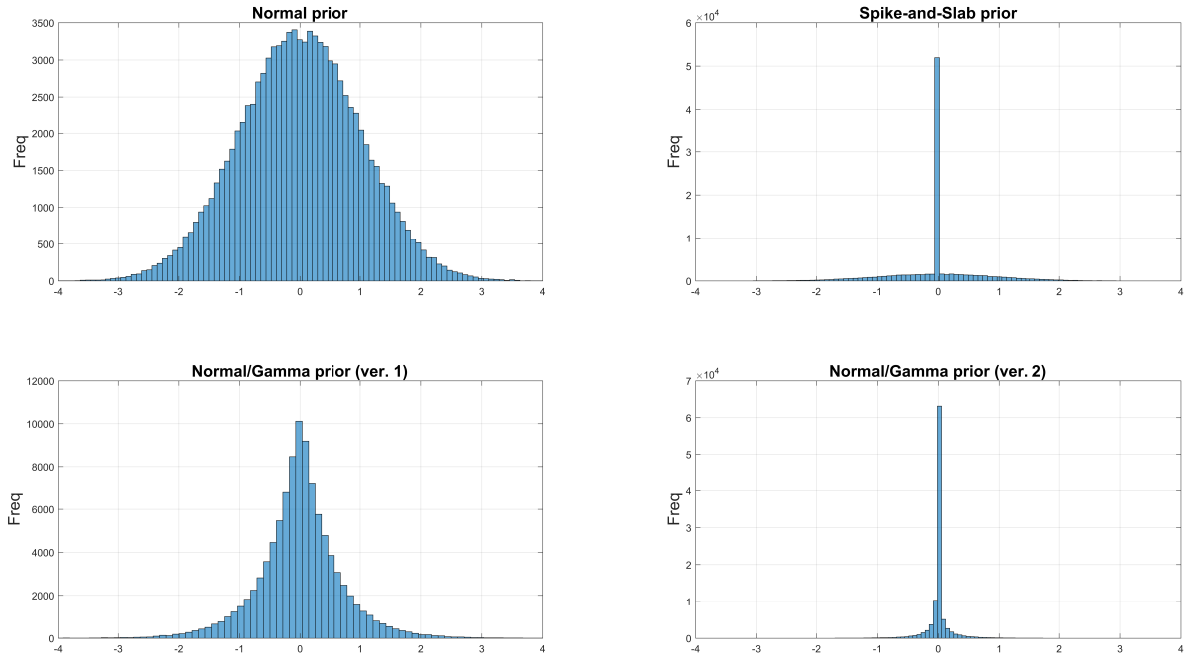
van den Boom, W., G. Reeves, and D. B. Dunson (2015a): "Quantifying Uncertainty in Variable Selection with Arbitrary Matrices," in *Proceedings of the IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), Cancun, Mexico*.

——— (2015b): "Scalable Approximations of Marginal Posteriors in Variable Selection," Working paper, Duke University Department of Statistical Science. Available at arXiv:1506.06629.

West, K. D. (1996): "Asymptotic Inference about Predictive Ability," *Econometrica*, 64, pp. 1067–1084.

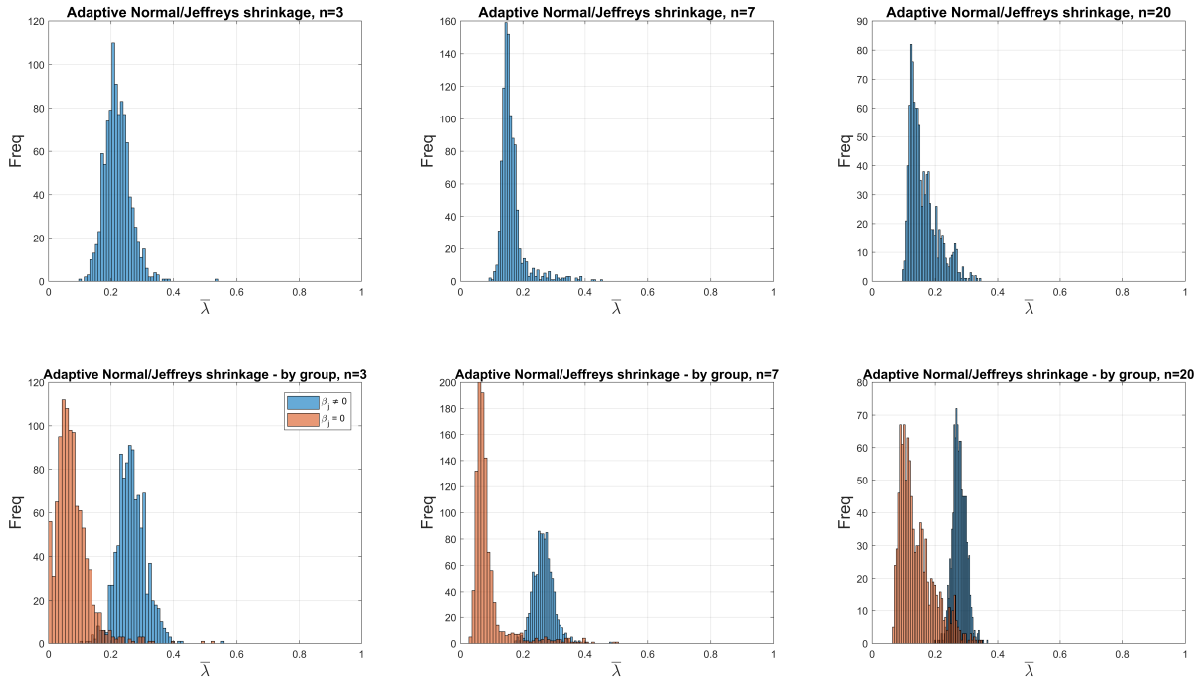
Figures and Tables

Figure 1. Histograms of hierarhical priors



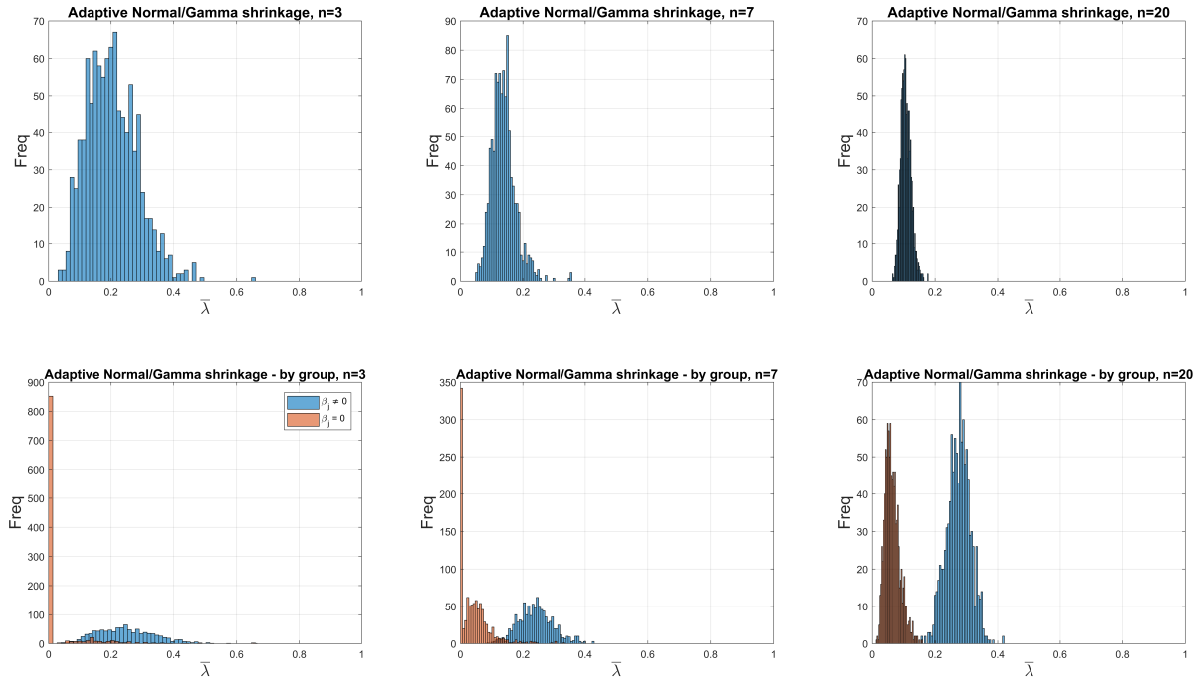
Top left panel: an example of a Normal prior for β_j in one dimension, where $\beta_j \sim \mathcal{N}\left(0, \underline{V}_{\beta_j}\right)$, and $\underline{V}_{\beta_j} = 10$. Top right panel: an example of a Spike-and-Slab prior for β_j in one dimension, where $\beta_j \sim (1-\lambda_j)\,\delta_0 + \lambda_j \mathcal{N}\left(0, \underline{V}_{\beta_j}\right)$, $\lambda_j \sim Bernoulli\,(\underline{\pi}_0)$, and $\underline{\pi}_0 = 0.5$. Bottom panels: two examples of a hierarchical Normal/Gamma prior for β_j in one dimension, where the hyperparameter λ_j^2 has been integrated out, i.e. $p(\beta_j) = \int p\left(\beta_j|\lambda_j^2\right) p\left(\lambda_j^2\right) d\lambda_j^2$, with $\beta_j|\lambda_j^2 \sim \mathcal{N}\left(0, \lambda_j^2 \underline{V}_{\beta_j}\right)$ and $\lambda_j^2 \sim \mathcal{G}\left(\underline{c}_1, \underline{c}_2\right)$. In the bottom left panel, we set $\underline{c}_1 = 1$ $\underline{c}_2 = 2$, while in the bottom right panel we have $\underline{c}_1 = 0.1$ $\underline{c}_2 = 2$.

Figure 2. Monte Carlo simulation - Shrinkage intensity, Normal/Jeffreys prior



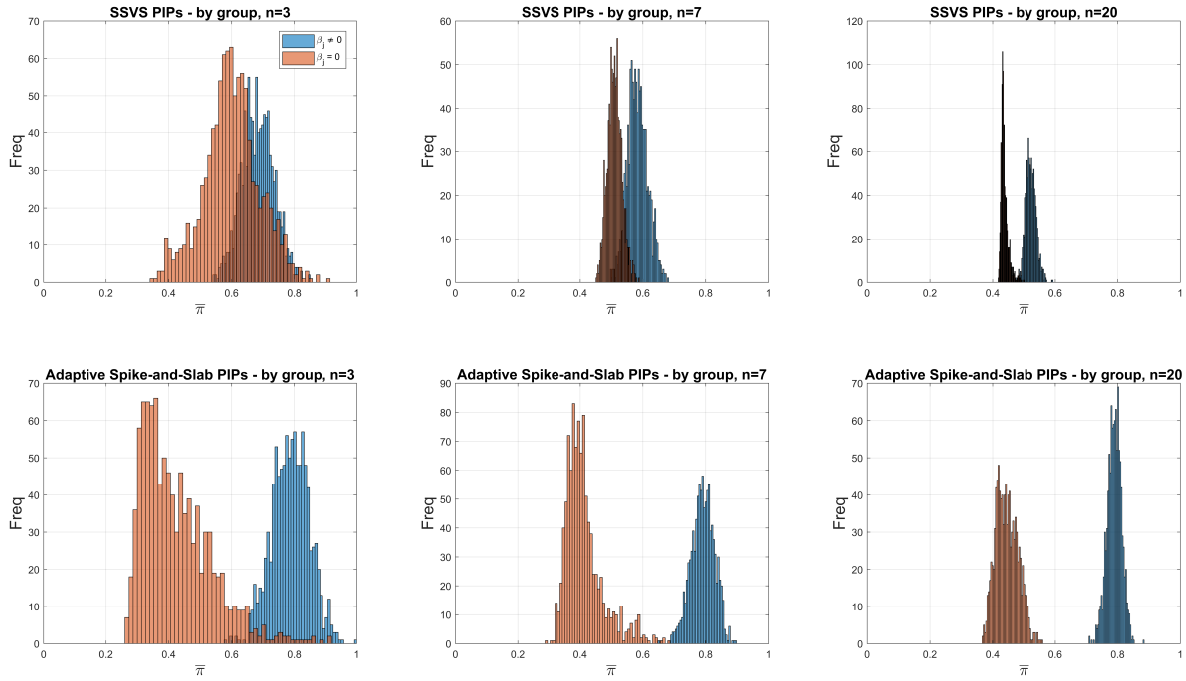
The top three panels plot the empirical distribution of the average estimated shrinkage intensity $\overline{\lambda} = \frac{1}{K}\sum_{i=1}^{n}\sum_{j=1}^{k_i}\widehat{\lambda}_{ij}$ for a small ($n = 3$), medium ($n = 7$), and large ($n = 20$) VAR(p), averaged over all VAR coefficients. $K = \sum_{i=1}^{n} k_i$ denotes the total number of VAR coefficients, including the covariance terms in $\mathbf{\Phi}$, and $k_i = np + i$. Results are based on our adaptive shrinkage procedure and the Normal/Jeffreys prior. The bottom three panels plot the average shrinkage intensity estimated by our adaptive procedure, broken down according to whether the corresponding VAR coefficients in the simulated data are equal to zero (red bars) or not (blue bars). All empirical distributions are obtained by simulating $1,000$ VAR(p) of sample size $T = 150$ and lag length $p = 2$. See Section 5 for additional details on the design of the Monte Carlo simulation.

Figure 3. Monte Carlo simulation - Shrinkage intensity, Normal/Gamma prior



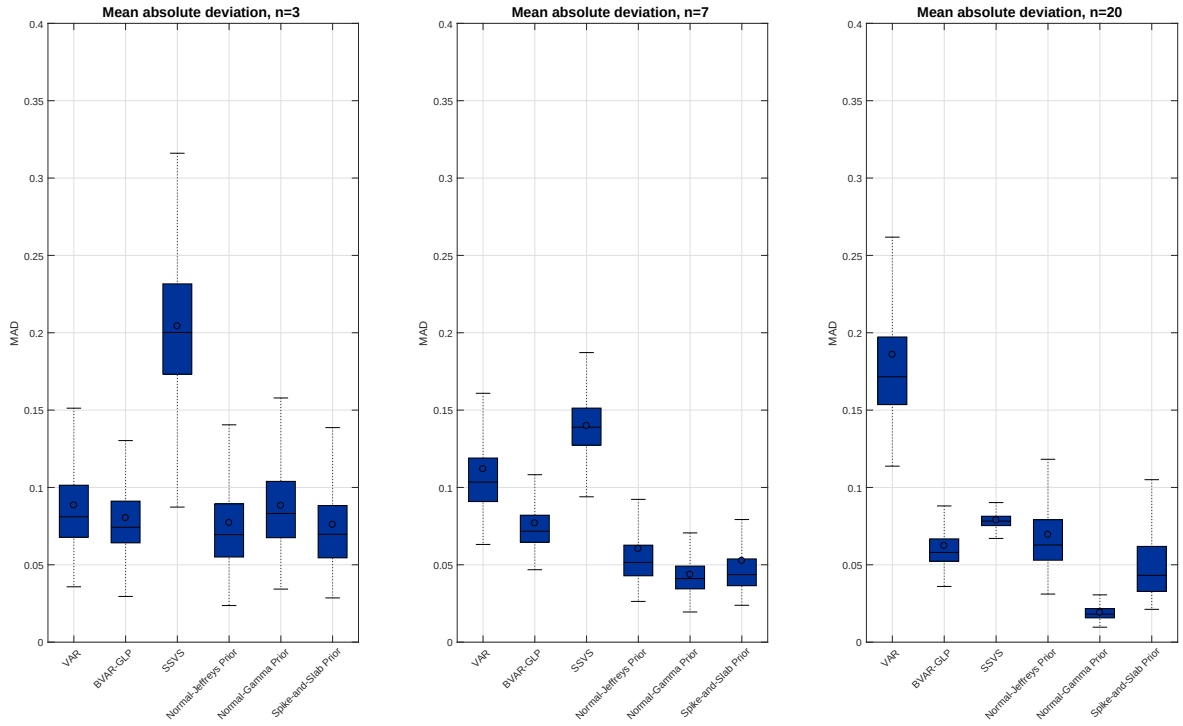
The top three panels plot the empirical distribution of the average estimated shrinkage intensity $\overline{\lambda} = \frac{1}{K}\sum_{i=1}^{n}\sum_{j=1}^{k_i}\widehat{\lambda}_{ij}$ for a small ($n = 3$), medium ($n = 7$), and large ($n = 20$) VAR(p), averaged over all VAR coefficients. $K = \sum_{i=1}^{n} k_i$ denotes the total number of VAR coefficients, including the covariance terms in $\mathbf{\Phi}$, and $k_i = np + i$. Results are based on our adaptive shrinkage procedure and the Normal/Gamma prior. The bottom three panels plot the average shrinkage intensity estimated by our adaptive procedure, broken down according to whether the corresponding VAR coefficients in the simulated data are equal to zero (red bars) or not (blue bars). All empirical distributions are obtained by simulating $1,000$ VAR(p) of sample size $T = 150$ and lag length $p = 2$. See Section 5 for additional details on the design of the Monte Carlo simulation.

Figure 4. Monte Carlo simulation - Posterior Inclusion Probabilities (PIPs)



The top three panels of this figure plot the empirical distribution of the average posterior inclusion probability (PIP) obtained using the George et al. (2008) SSVS approach for a small ($n = 3$), medium ($n = 7$), and large ($n = 20$) VAR(p), and broken down according to whether the corresponding VAR coefficients in the simulated data are equal to zero (red bars) or not (blue bars). The bottom three panels plot the analogous empirical distributions of the averaged PIPs estimated using our adaptive shrinkage procedure with the Spike-and-Slab prior. All empirical distributions are obtained by simulating $1,000$ VAR(p) of sample size $T = 150$ and lag length $p = 2$. See Section 5 for additional details on the design of the Monte Carlo simulation.

Figure 5. Monte Carlo simulation - Mean Absolute Deviations

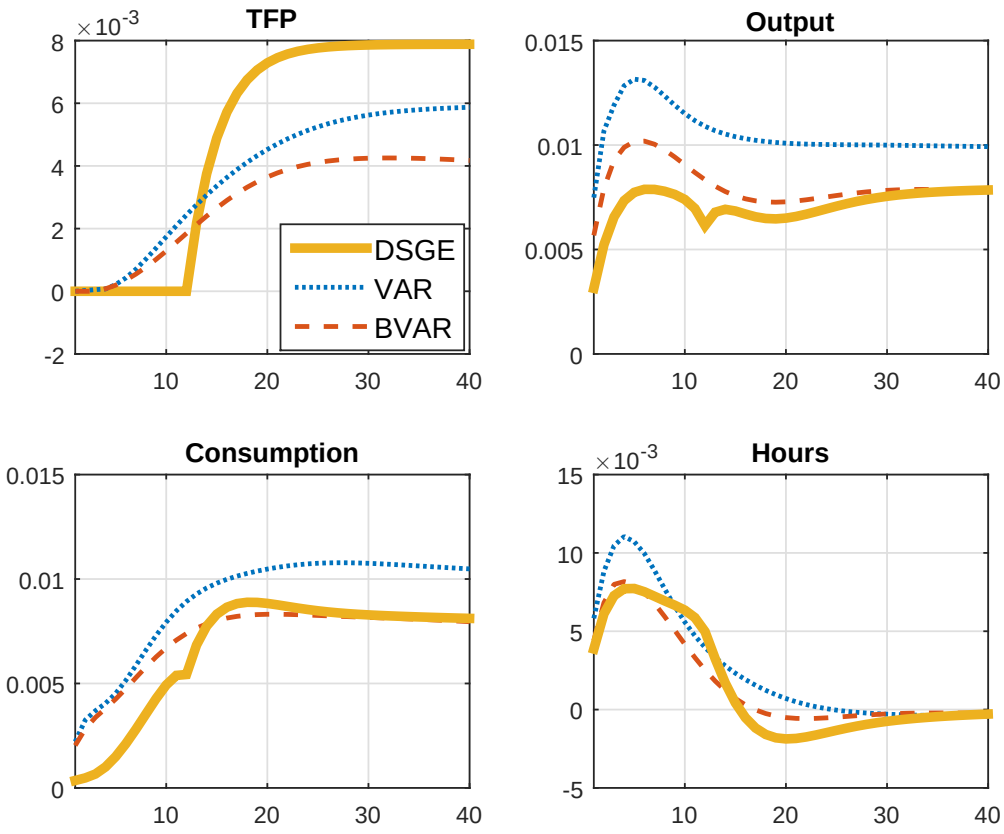


This figure reports box plots for the empirical distributions of the Mean Absolute Deviations (MAD), obtained from estimating a VAR(p) with OLS, a BVAR using the Giannone et al. (2015) (BVAR-GLP), the George et al. (2008) SSVS approach, and our adaptive shrinkage procedure with Normal/Jeffreys, Normal/Gamma, and Spike-and-Slab priors. These empirical distributions are obtained by simulating $1,000$ VAR(p) of sample size $T = 150$ and lag length $p = 2$. For each of the approaches listed and each of the 1,000 simulations we compute the Mean Absolute Deviation (MAD), defined as

$$MAD^{(r,s)} = \frac{1}{K} \sum_{i=1}^{n} \sum_{j=1}^{k_i} \left| \beta_{ij}^{(r)} - \widehat{\beta}_{ij}^{(r,s)} \right|,$$

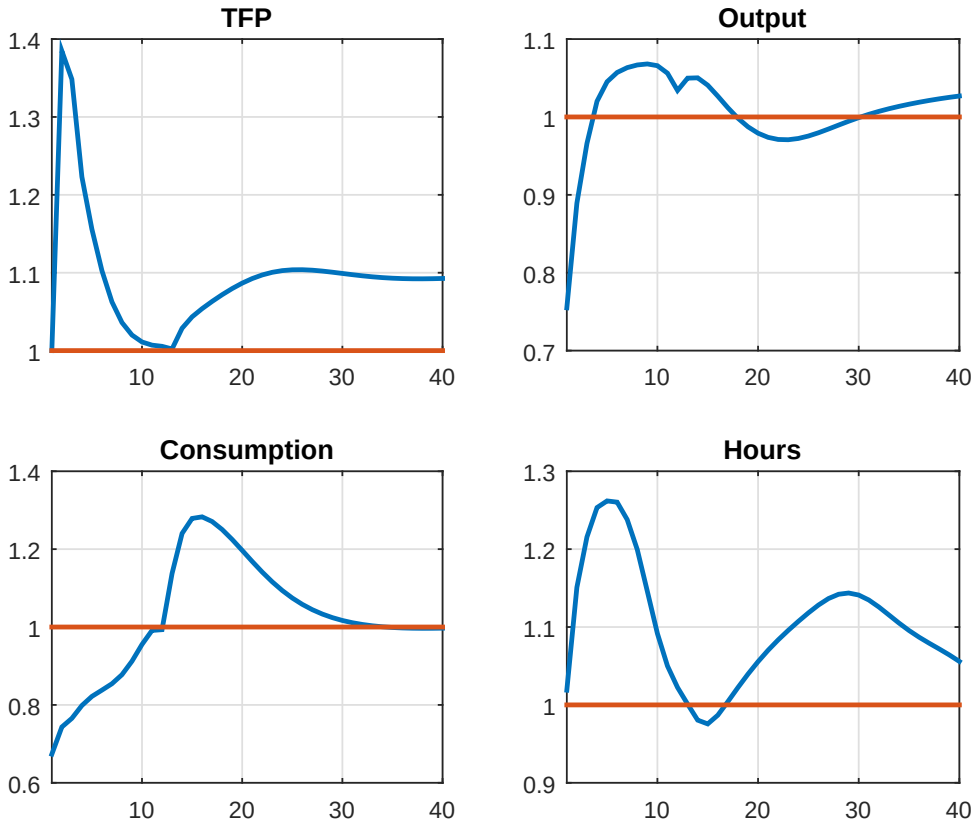
where s denotes the method used, $r = 1, ..., 1,000$ keeps track of the MC simulations, $K = \sum_{i=1}^{n} k_i$ denotes the total number of lag coefficients in the VAR, $\beta_{ij}^{(r)}$ is the true DGP coefficient from the r-th simulation, and $\widehat{\beta}_{ij}^{(r,s)}$ denotes the (posterior mean of the) corresponding estimate according to method s. Results are reported separately for small ($n = 3$), medium ($n = 7$), and large ($n = 20$) VARs.

Figure 6. Impulse responses on simulated data



This figure reports the impulse responses to a productivity news shock in the DSGE model used to generate the data (solid line), and the median across Monte Carlo replications of the BVAR (dashed line) and the VAR (dotted line) impulse responses.

Figure 7. Ratio of MSE: VAR versus BVAR



This figure reports the ratio of the MSE of the VAR over the MSE of the BVAR. Values larger than one indicate that the MSE of the VAR is larger than that of the BVAR.

Table 1. Computing time (seconds) per Monte Carlo iteration

CPU time (in seconds)

Method	*VAR size*		
	Small	*Medium*	*Large*
BVAR-GLP	0.043	0.094	0.949
SSVS	0.469	2.624	52.247
Adaptive Normal/Jeffreys prior	0.022	0.024	0.023
Adaptive Spike-and-Slab prior	0.020	0.024	0.023
Adaptive Normal/Gamma prior	0.063	0.074	0.071

This table reports the average CPU time required to complete one iteration of the Monte Carlo simulation. In each iteration of the Monte Carlo, we estimate a VAR(p) of sample size $T = 150$ and lag length $p = 2$ with a BVAR using the Giannone et al. (2015) method (BVAR-GLP), the George et al. (2008) SSVS approach, and our adaptive shrinkage procedure with Normal/Jeffreys, Spike-and-Slab, and Normal/Gamma priors. The reference machine is a 64 bit Windows 7-based PC with a 3.4 Ghz Quad-Core Intel i7-3770 CPU with 16GB DDR3 RAM and running MATLAB R2017a. See Section 5 for additional details on the design of the Monte Carlo simulation.

Table 2. Out-of-sample point forecast performance: Multivariate results

	DFM	*FAVAR*	*BVAR-BGR*	*BVAR-GLP*	*SSVS*	*N-J*	*SNS*	*N-G*	*DFM*	*FAVAR*	*BVAR-BGR*	*BVAR-GLP*	*SSVS*	*N-J*	*SNS*	*N-G*
	Medium VAR															
	3 series								*7 series*							
h=1	0.693***		0.738***	0.730***	0.714*	0.528***	0.520***	**0.495*** **	0.757***		0.780***	0.773***	0.784*	0.639***	0.616***	**0.586*** **
h=2	0.704***		0.713***	0.701***	0.649***	0.620***	**0.560*** **	0.571***	0.764***		0.772***	0.761***	0.777**	0.724***	**0.644*** **	0.652***
h=3	0.750***		0.757***	0.763***	0.740***	0.709***	**0.652*** **	0.677***	0.788***		0.802***	0.798***	0.798**	0.792***	**0.701*** **	0.706***
h=4	0.712***		0.743***	0.745***	0.715***	0.681***	**0.645*** **	0.660***	0.763***		0.810***	0.808***	0.775**	0.784**	0.713***	**0.713*** **
	Large VAR															
	3 series								*7 series*							
h=1	0.646***	0.580***	0.703***	0.673***	0.692**	0.464***	0.459***	**0.445*** **	0.765***	0.666***	0.751***	0.713***	0.806*	0.574***	0.558***	**0.540*** **
h=2	0.656***	0.557***	0.760***	0.778***	0.586***	0.627***	**0.504*** **	0.505***	0.702***	0.607***	0.782***	0.793***	0.654***	0.686***	0.570***	**0.562*** **
h=3	0.744**	0.678**	0.792**	0.807**	0.718**	0.776	**0.618*** **	0.633**	0.785**	0.725***	0.851**	0.869*	0.780**	0.863	0.702***	**0.690*** **
h=4	0.830	0.810*	0.878*	0.878*	0.829	0.918	**0.743** **	0.760**	0.844*	0.827**	0.926	0.941	0.853*	0.990	0.814***	**0.791*** **
	X-large VAR															
	3 series								*7 series*							
h=1	0.631***	0.575***	0.661***	0.731***	0.538***	0.466***	0.463***	**0.441*** **	0.739***	0.642***	0.727***	0.746***	0.647***	0.569***	0.567***	**0.546*** **
h=2	0.807**	0.716***	0.925	0.992	0.731**	0.677**	**0.608*** **	0.621***	0.806**	0.713***	0.867**	0.902	0.733**	0.696***	**0.637*** **	0.648***
h=3	0.859**	0.792***	0.860**	0.907	0.806**	0.758***	**0.693*** **	0.718***	0.866**	0.783***	0.877**	0.911	0.813***	0.806***	**0.735*** **	0.746***
h=4	0.935	0.897	0.923	0.918	0.892	0.851*	**0.788** **	0.816*	0.906	0.856**	0.943	0.958	0.866*	0.874*	**0.801*** **	0.811***

This table reports the ratio between the multivariate weighted mean squared forecast error (WMSFE) of model i and the WMSFE of the benchmark VAR(p^*) model, computed as

$$WMSFE_{ih} = \frac{\sum_{\tau=\underline{t}}^{\bar{t}-h} we_{i,\tau+h}}{\sum_{\tau=\underline{t}}^{\bar{t}-h} we_{bcmk,\tau+h}},$$

where p^* is the largest lag length that can be estimated in a VAR with flat priors and the data at hand, $we_{i,\tau+h} = \left(e'_{i,\tau+h} \times W \times e_{i,\tau+h}\right)$ and $we_{bcmk,\tau+h} = \left(e'_{bcmk,\tau+h} \times W \times e_{bcmk,\tau+h}\right)$ denote the weighted forecast errors of model i and the benchmark model at time $\tau + h$, $e_{i,\tau+h}$ and $e_{bcmk,\tau+h}$ are the $(N \times 1)$ vector of forecast errors, and W is an $(N \times N)$ matrix of weights. The left panels are based on $N = 3$, and focus on the following three series {FEDFUNDS, GDP, GDPDEFL}. The right panels focus on $N = 7$ and the following series {PAYEMS, CPIAUCSL,FEDFUNDS, GDP, UNRATE, GDPDEFL, GS10}. We set the matrix W to be a diagonal matrix featuring on the diagonal the inverse of the variances of the series to be forecast. $\underline{t}$ and $\bar{t}$ denote the start and end of the out-of-sample period, $i \in$ {DFM, FAVAR, BVAR-BGR, BVAR-GLP, SSVS, N-J, SNS, N-G}, and $h = 1, ..., 4$. All forecasts are generated out-of-sample using recursive estimates of the models, with the out of sample period starting in 1985:Q1 and ending in 2015:Q4. Bold numbers indicate the lowest WMSFE across all models for any given VAR size - forecast horizon pair. * significance at the 10% level; ** significance at the 5% level; *** significance at the 1% level.

† The factor-augmented VAR (FAVAR) of medium size only has the seven variables of interest observed but no additional variables to extract factors from. Therefore, the FAVAR estimated on the medium size is equivalent to the VAR estimated with OLS, and for that reason we do not report its results.

Table 3. Out-of-sample point forecast performance, Medium VAR

Variable	*DFM*	*BVAR-BGR*	*BVAR-GLP*	*SSVS*	*N-J*	*SNS*	*N-G*
			$h = 1$				
PAYEMS	0.882	0.746**	0.738**	1.025	0.706*	0.539**	**0.510****
CPIAUCSL	0.983	0.975	0.970	1.056	0.961	0.966	**0.945**
FEDFUNDS	0.573***	0.687***	0.675***	0.320**	0.294***	0.328**	**0.273****
GDP	0.842	0.761***	0.767***	1.297	0.841	**0.724****	0.764**
UNRATE	0.786	0.796***	0.829**	1.002	0.855	0.673*	**0.627***
GDPDEFL	0.835*	0.844*	0.834**	1.071	0.778*	0.783*	**0.764****
GS10	0.790**	0.810**	0.781**	**0.676*****	0.694***	0.728**	0.683***
			$h = 2$				
PAYEMS	0.670	0.735**	0.721**	0.964	0.733	0.517***	**0.485*****
CPIAUCSL	1.004	0.964	0.969	**0.963**	0.984	0.974	1.034
FEDFUNDS	0.576***	0.606***	0.588***	0.371***	0.391***	**0.360*****	0.374***
GDP	0.768**	0.742***	0.768***	1.021	0.862	**0.711*****	0.744**
UNRATE	0.797	0.799**	0.809*	1.048	0.905	0.702**	**0.665****
GDPDEFL	0.917	0.924	0.875*	0.800*	0.830*	0.822*	**0.799***
GS10	0.873	0.879	0.843**	0.820*	0.812*	**0.790****	0.810*
			$h = 3$				
PAYEMS	0.692	0.753**	0.744*	0.895	0.796	0.576***	**0.542*****
CPIAUCSL	1.043	0.976	0.965	**0.929**	1.003	0.966	0.951
FEDFUNDS	0.594***	0.643***	0.661***	0.516***	0.512***	**0.496*****	0.524***
GDP	0.824*	0.775***	0.797***	0.928	0.822*	**0.703*****	0.741**
UNRATE	0.765*	0.807**	0.799*	0.766	0.871	0.681**	**0.659****
GDPDEFL	0.908	0.934	0.890	**0.829**	0.875	0.850	0.846
GS10	0.856*	0.885	0.856**	0.852	0.861	**0.831****	0.836**
			$h = 4$				
PAYEMS	0.695*	0.777*	0.772*	0.789	0.819	0.617**	**0.587****
CPIAUCSL	1.016	**1.002**	1.002	1.009	1.041	1.033	1.020
FEDFUNDS	0.481***	0.610***	0.601***	0.462***	**0.446*****	0.452***	0.455***
GDP	0.915	0.804***	0.847**	0.962	0.865	**0.761*****	0.803**
UNRATE	0.736	0.844	0.846	0.694*	0.839	0.703*	**0.675****
GDPDEFL	0.862**	0.923	0.883**	**0.842****	0.874**	0.860**	0.857**
GS10	0.909	0.956	0.934	0.937	0.924	**0.904**	0.917

This table reports the ratio between the MSFE of model i and the MSFE of the benchmark VAR(p) for the medium size VAR, computed as

$$MSFE_{ijh} = \frac{\sum_{\tau=\underline{t}}^{\bar{t}-h} e^2_{i,j,\tau+h}}{\sum_{\tau=\underline{t}}^{\bar{t}-h} e^2_{bcmk,j,\tau+h}},$$

where $p = 5$, $e^2_{i,j,\tau+h}$ and $e^2_{bcmk,j,\tau+h}$ are the squared forecast errors of variable j at time τ and forecast horizon h generated by model i and the VAR(p) model, respectively. $\underline{t}$ and $\bar{t}$ denote the start and end of the out-of-sample period, $i \in$ {DFM, BVAR-BGR, BVAR-GLP, SSVS, N-J, SNS, N-G}, $j \in$ {PAYEMS, CPIAUCSL,FEDFUNDS, GDP, UNRATE, GDPDEFL, GS10}, and $h = 1, ..., 4$. All forecasts are generated out-of-sample using recursive estimates of the models, with the out of sample period starting in 1985:Q1 and ending in 2015:Q4. Bold numbers indicate the lowest MSFE across all models for a given variable-forecast horizon pair. * significance at the 10% level; ** significance at the 5% level; *** significance at the 1% level.

Table 4. Out-of-sample point forecast performance, Large VAR

Variable	*DFM*	*FAVAR*	*BVAR-BGR*	*BVAR-GLP*	*SSVS*	*N-J*	*SNS*	*N-G*
				$h = 1$				
PAYEMS	1.225	0.587***	0.661**	0.643***	0.870	0.525**	**0.454****	0.458***
CPIAUCSL	1.286	1.095	1.016	**0.955**	1.169	1.060	0.999	0.986
FEDFUNDS	0.473***	0.468***	0.675**	0.671**	0.315***	0.325***	0.307***	**0.271*****
GDP	1.026	0.686	0.720*	0.661*	1.317	**0.592****	0.593*	0.621*
UNRATE	0.770**	0.744**	0.730***	0.798**	1.464	0.689*	0.592**	**0.584*****
GDPDEFL	**0.607*****	0.714**	0.747***	0.691***	0.832	0.632***	0.652***	0.639***
GS10	0.775**	0.721**	0.818**	0.700***	0.652***	0.662***	0.687***	**0.639*****
				$h = 2$				
PAYEMS	0.890	0.625***	0.879	0.875	0.859	0.726*	0.555**	**0.491*****
CPIAUCSL	0.818*	0.777**	0.766**	0.787***	**0.745****	0.823*	0.757**	0.753**
FEDFUNDS	0.422***	**0.304*****	0.659***	0.722**	0.332***	0.472***	0.325***	0.320***
GDP	1.078	0.856	0.930	0.955	0.991	0.826	**0.694***	0.700
UNRATE	0.700***	0.664***	0.844**	0.874	0.815	0.843	0.673**	**0.625*****
GDPDEFL	0.698***	0.784**	0.793*	0.699***	0.691***	0.746**	**0.689*****	0.700***
GS10	0.688***	0.616***	0.781**	0.753**	0.616***	0.671***	**0.609*****	0.620***
				$h = 3$				
PAYEMS	0.938	0.754**	1.005	1.023	0.942	1.010	0.744	**0.634*****
CPIAUCSL	0.807	0.782*	0.824**	0.842**	**0.780***	0.863**	0.807**	0.797*
FEDFUNDS	0.509**	0.454**	0.641**	0.689*	0.463**	0.631	**0.445****	0.461**
GDP	1.016	0.901	0.943	0.983	1.021	0.919	**0.743***	0.768
UNRATE	**0.812**	0.826	0.998	1.017	0.948	1.192	0.879	0.813
GDPDEFL	0.780**	0.764**	0.849*	0.768**	**0.744****	0.836*	0.754**	0.752**
GS10	0.783***	**0.753*****	0.864	0.891*	0.759***	0.822**	0.763**	0.758***
				$h = 4$				
PAYEMS	0.860	0.768**	0.996	1.067	0.873	1.065	0.824	**0.697****
CPIAUCSL	**0.868***	0.876	0.885	0.878*	0.900	0.917	0.899	0.895
FEDFUNDS	0.520**	0.541**	0.700**	0.657***	0.520**	0.710	**0.511****	0.526**
GDP	1.178	1.077	1.036	1.122	1.168	1.107	**0.905**	0.945
UNRATE	**0.806***	0.844	1.025	1.089	0.849	1.268	0.920	0.821*
GDPDEFL	**0.928**	0.947	0.997	0.952	0.938	1.051	0.973	0.956
GS10	0.896	0.895	0.975	0.955	0.888*	0.966	0.891	**0.886***

This table reports the ratio between the MSFE of model i and the MSFE of the benchmark VAR(p) for the large size VAR, computed as

$$MSFE_{ijh} = \frac{\sum_{\tau=\underline{t}}^{\bar{t}-h} e^2_{i,j,\tau+h}}{\sum_{\tau=\underline{t}}^{\bar{t}-h} e^2_{bcmk,j,\tau+h}},$$

where $p = 2$, $e^2_{i,j,\tau+h}$ and $e^2_{bcmk,j,\tau+h}$ are the squared forecast errors of variable j at time τ and forecast horizon h generated by model i and the VAR(p) model, respectively. $\underline{t}$ and $\bar{t}$ denote the start and end of the out-of-sample period, $i \in$ {DFM, FAVAR, BVAR-BGR, BVAR-GLP, SSVS, N-J, SNS, N-G}, $j \in$ {PAYEMS, CPIAUCSL,FEDFUNDS, GDP, UNRATE, GDPDEFL, GS10}, and $h = 1, ..., 4$. All forecasts are generated out-of-sample using recursive estimates of the models, with the out of sample period starting in 1985:Q1 and ending in 2015:Q4. Bold numbers indicate the lowest MSFE across all models for a given variable-forecast horizon pair. * significance at the 10% level; ** significance at the 5% level; *** significance at the 1% level.

Table 5. Out-of-sample point forecast performance, X-large VAR

Variable	*DFM*	*FAVAR*	*BVAR-BGR*	*BVAR-GLP*	*SSVS*	*N-J*	*SNS*	*N-G*
				$h = 1$				
PAYEMS	1.092	0.560***	0.612***	0.555***	0.741	0.516***	**0.484*** **	0.516***
CPIAUCSL	1.317	1.091	0.994	0.954	1.245	**0.946**	0.980	0.975
FEDFUNDS	0.449***	0.476***	0.655***	0.612***	0.317***	0.324***	0.344**	**0.280*** **
GDP	1.022	0.791*	0.750**	0.760*	1.016	0.738*	**0.656*** **	0.705**
UNRATE	0.638**	0.576***	0.633***	0.736*	0.636**	0.636**	0.581***	**0.576*** **
GDPDEFL	0.627**	0.570***	0.601***	0.907	0.531***	**0.488*** **	0.511***	0.504***
GS10	0.766**	0.690***	0.905	0.736***	0.687**	0.721**	0.739*	**0.685** **
				$h = 2$				
PAYEMS	0.962	0.604***	0.792*	0.730***	0.791	0.621***	**0.524*** **	0.541***
CPIAUCSL	0.963	0.961	0.884**	0.924	0.909	0.897	**0.879**	0.911
FEDFUNDS	0.584***	0.508***	0.971	1.106	0.469**	0.511**	0.452***	**0.443*** **
GDP	1.241	0.964	0.975	0.857	1.163	0.895	**0.752* **	0.800
UNRATE	0.760*	**0.656*** **	0.773**	0.839	0.775	0.814	0.700**	0.685**
GDPDEFL	0.706***	0.794*	0.793*	0.950	**0.699*** **	0.718**	0.709***	0.722***
GS10	0.669***	0.638***	0.782**	0.750***	**0.584*** **	0.598***	0.591***	0.594***
				$h = 3$				
PAYEMS	0.984	0.728***	0.893	0.861	0.825	0.777*	**0.642*** **	0.648***
CPIAUCSL	0.939	0.913	0.959	1.022	**0.902**	0.975	0.930	0.933
FEDFUNDS	0.604***	0.581***	0.762*	0.807*	0.563***	**0.535*** **	0.540***	0.547***
GDP	1.122	0.961	0.922	0.893	1.028	0.905	**0.748* **	0.799
UNRATE	0.818	**0.705*** **	0.869	0.923	0.817	0.919	0.776*	0.758**
GDPDEFL	0.885	0.893	0.937	1.111	**0.877**	0.919	0.881	0.896
GS10	0.781*	0.760***	0.865	0.869*	0.760**	0.772**	0.763**	**0.760** **
				$h = 4$				
PAYEMS	0.921	0.775**	0.980	0.999	0.801*	0.828	0.694**	**0.685** **
CPIAUCSL	**0.975**	0.981	0.992	1.002	0.989	1.017	1.007	0.990
FEDFUNDS	0.642**	0.625***	0.803	0.724***	0.616**	**0.599*** **	0.606**	0.625**
GDP	1.204	1.120	0.969	0.999	1.112	1.001	**0.840**	0.907
UNRATE	0.797**	**0.731*** **	0.972	1.055	0.781**	0.907	0.758***	0.760***
GDPDEFL	**1.009**	1.011	1.068	1.141	1.023	1.058	1.037	1.017
GS10	0.854	**0.836**	0.903	0.917	0.837	0.858	0.840	0.836

This table reports the ratio between the MSFE of model i and the MSFE of the benchmark VAR(p) for the X-large size VAR, computed as

$$MSFE_{ijh} = \frac{\sum_{\tau=\underline{t}}^{\bar{t}-h} e^2_{i,j,\tau+h}}{\sum_{\tau=\underline{t}}^{\bar{t}-h} e^2_{bcmk,j,\tau+h}},$$

where $p = 1$, $e^2_{i,j,\tau+h}$ and $e^2_{bcmk,j,\tau+h}$ are the squared forecast errors of variable j at time τ and forecast horizon h generated by model i and the VAR(p) model, respectively. $\underline{t}$ and $\bar{t}$ denote the start and end of the out-of-sample period, $i \in$ {DFM, FAVAR, BVAR-BGR, BVAR-GLP, SSVS, N-J, SNS, N-G}, $j \in$ {PAYEMS, CPIAUCSL,FEDFUNDS, GDP, UNRATE, GDPDEFL, GS10}, and $h = 1, ..., 4$. All forecasts are generated out-of-sample using recursive estimates of the models, with the out of sample period starting in 1985:Q1 and ending in 2015:Q4. Bold numbers indicate the lowest MSFE across all models for a given variable-forecast horizon pair. * significance at the 10% level; ** significance at the 5% level; *** significance at the 1% level.

Table 6. Out-of-sample density forecast performance, Medium VAR

Variable	*DFM*	*BVAR-BGR*	*BVAR-GLP*	*SSVS*	*N-J*	*SNS*	*N-G*
			$h = 1$				
PAYEMS	0.354	0.474	0.447	0.315	0.529	**0.560**	0.560
CPIAUCSL	1.675	1.394*	1.061*	**2.798**	2.072	2.277*	2.576*
FEDFUNDS	0.482	0.371	0.408	0.459	**0.588**	0.502	0.492
GDP	0.069	0.126**	0.081*	-0.087	0.032	**0.149**	0.139
UNRATE	0.848	0.399	0.487	0.659	0.522	0.823	**0.942**
GDPDEFL	0.002	0.032	**0.044**	-0.108	0.022	-0.002	-0.001
GS10	0.235*	0.244*	0.244**	0.274**	0.308**	0.282**	**0.311****
			$h = 2$				
PAYEMS	0.251	0.311*	0.367*	0.055	0.321	**0.414***	0.410*
CPIAUCSL	0.815	0.256	0.877	1.577	0.481	0.891	**1.765**
FEDFUNDS	0.048	0.132***	0.127***	0.055	**0.149***	0.101	0.059
GDP	0.073	**0.181*****	0.105*	0.023	-0.037	0.113	0.141*
UNRATE	0.115	0.043	0.224*	-0.076	-0.202	**0.225****	0.201*
GDPDEFL	-0.032	0.006	0.024	-0.006	**0.027**	0.021	0.011
GS10	0.083	0.081	0.099*	0.098	0.125*	**0.126***	0.111
			$h = 3$				
PAYEMS	0.195	0.112	0.208**	-0.351	0.117	0.338*	**0.344**
CPIAUCSL	1.347	0.969	0.583	**2.127**	0.294	0.941	1.212
FEDFUNDS	-0.027	**0.064*****	0.042**	-0.032	0.061**	0.027	-0.029
GDP	0.151	0.077	0.154**	0.035	0.016	**0.212*****	0.173**
UNRATE	0.520	0.366	0.414	0.479	-0.059	0.280**	**0.591**
GDPDEFL	-0.029	-0.007	**0.011**	-0.030	0.001	0.009	-0.021
GS10	0.058	0.053	0.067	0.051	0.075	**0.083**	0.082*
			$h = 4$				
PAYEMS	0.435	0.176*	0.194*	-0.185	0.246	0.441	**0.461**
CPIAUCSL	**1.209**	0.470	0.447	1.042	0.484	0.755	0.968
FEDFUNDS	0.024	0.091***	0.070**	0.019	**0.122*****	0.060	0.028
GDP	0.000	**0.166****	-0.007	0.056	0.046	0.099*	0.101**
UNRATE	0.466	0.067	0.037	**0.510**	-0.249	0.155	0.297*
GDPDEFL	0.004	0.011	0.015	-0.024	**0.021**	0.010	-0.022
GS10	0.017	0.028	0.041	0.018	**0.058**	0.047	0.043

This table reports the average log predictive likelihood (ALPL) differential between model i and the benchmark VAR(p) for the medium VAR, computed as

$$ALPL_{ijh} = \frac{1}{\bar{t} - \underline{t} - h + 1} \sum_{\tau=\underline{t}}^{\bar{t}-h} \left(LPL_{i,j,\tau+h} - LPL_{bcmk,j,\tau+h} \right),$$

where $p = 5$, while $LPL_{i,j,\tau+h}$ and $LPL_{bcmk,j,\tau+h}$ are the log predictive likelihoods of variable j at time τ and forecast horizon h generated by model i and the VAR(p), respectively. $\underline{t}$ and $\bar{t}$ denote the start and end of the out-of-sample period, $i \in$ {DFM, BVAR-BGR, BVAR-GLP, SSVS, N-J, SNS, N-G}, $j \in$ {PAYEMS, CPIAUCSL,FEDFUNDS, GDP, UNRATE, GDPDEFL, GS10}, and $h = 1, ..., 4$. All forecasts are generated out-of-sample using recursive estimates of the models, with the out of sample period starting in 1985:Q1 and ending in 2015:Q4. Bold numbers indicate the highest ALPL across all models for a given variable-forecast horizon pair. * significance at the 10% level; ** significance at the 5% level; *** significance at the 1% level.

Table 7. Out-of-sample density forecast performance, Large VAR

Variable	*DFM*	*FAVAR*	*BVAR-BGR*	*BVAR-GLP*	*SSVS*	*N-J*	*SNS*	*N-G*
				$h = 1$				
PAYEMS	0.174	0.494	0.472	0.346**	0.321	0.598	**0.602**	0.557
CPIAUCSL	1.566	0.775	-1.318	-0.014	**2.323**	0.933	0.936	1.211*
FEDFUNDS	0.252	0.248	0.082	-0.047	0.187	**0.348***	0.267	0.255
GDP	0.061	0.159	0.230*	**0.282***	-0.102	0.282*	0.244	0.192
UNRATE	0.507	0.582	0.260***	0.196*	0.160	0.670	**0.707**	0.676
GDPDEFL	0.136	0.124	0.178**	**0.200****	0.051	0.188*	0.140	0.140
GS10	0.256**	0.286**	0.222**	0.312***	0.301**	**0.338*****	0.311***	0.324***
				$h = 2$				
PAYEMS	0.052	0.214**	0.030	-0.160	0.049	0.194	**0.279****	0.275**
CPIAUCSL	**2.072**	1.928	0.068	0.028	1.831	1.336	1.368	2.021
FEDFUNDS	0.111*	0.137*	**0.174*****	0.119*	0.085	0.154***	0.163**	0.108
GDP	-0.055	0.037	0.010	-0.091	-0.070	0.096	**0.147**	0.100
UNRATE	0.175	0.221*	-0.134	-0.139	0.093	0.167	0.276**	**0.315****
GDPDEFL	0.044	0.038	0.084	**0.137*****	0.053	0.074	0.074*	0.045
GS10	0.332*	0.372*	0.276	0.295	0.365*	0.346*	**0.397***	0.394*
				$h = 3$				
PAYEMS	-0.035	0.038	-0.362	-0.394	-0.234	-0.001	0.120	**0.121**
CPIAUCSL	0.752	0.777	-1.464	-0.866	0.790	0.223	-0.065	**0.900**
FEDFUNDS	0.046	0.046	**0.171*****	0.135***	0.017	0.071	0.081*	0.031
GDP	0.170	0.216	0.057	0.092	0.148	0.254	**0.340**	0.291
UNRATE	0.006	-0.063	-0.366	-0.263	-0.124	-0.023	-0.101	**0.049**
GDPDEFL	0.054	0.065*	0.087*	**0.119*****	0.060	0.076**	0.070*	0.059
GS10	0.123**	0.141**	0.115**	0.088*	0.137**	0.123**	0.143**	**0.150****
				$h = 4$				
PAYEMS	-0.026	0.059	-0.697	-0.864	-0.336	-0.023	0.087	**0.105**
CPIAUCSL	0.663	0.333	-0.753	-0.873	**0.681**	0.549	0.380	0.437
FEDFUNDS	0.042	0.040	**0.178*****	0.154***	0.015	0.062	0.086**	0.032
GDP	-0.110	-0.008	-0.375	-0.117	-0.042	-0.048	**0.115**	0.058
UNRATE	0.016	**0.092**	-0.703	-0.439	-0.042	-0.066	0.072	0.043
GDPDEFL	0.004	0.000	0.050	**0.069*****	-0.009	0.004	0.010	-0.018
GS10	0.035	0.039	0.024	0.045*	0.037	0.035	0.054**	**0.059****

This table reports the average log predictive likelihood (ALPL) differential between model i and the benchmark VAR(p) for the large VAR, computed as

$$ALPL_{ijh} = \frac{1}{\bar{t} - \underline{t} - h + 1} \sum_{\tau=\underline{t}}^{\bar{t}-h} \left(LPL_{i,j,\tau+h} - LPL_{bcmk,j,\tau+h} \right),$$

where $p = 2$, while $LPL_{i,j,\tau+h}$ and $LPL_{bcmk,j,\tau+h}$ are the log predictive likelihoods of variable j at time τ and forecast horizon h generated by model i and the VAR(p), respectively. $\underline{t}$ and $\bar{t}$ denote the start and end of the out-of-sample period, $i \in$ {DFM, FAVAR, BVAR-BGR, BVAR-GLP, SSVS, N-J, SNS, N-G}, $j \in$ {PAYEMS, CPIAUCSL,FEDFUNDS, GDP, UNRATE, GDPDEFL, GS10}, and $h = 1, ..., 4$. All forecasts are generated out-of-sample using recursive estimates of the models, with the out of sample period starting in 1985:Q1 and ending in 2015:Q4. Bold numbers indicate the highest ALPL across all models for a given variable-forecast horizon pair. * significance at the 10% level; ** significance at the 5% level; *** significance at the 1% level.

Table 8. Out-of-sample density forecast performance, X-large VAR

Variable	*DFM*	*FAVAR*	*BVAR-BGR*	*BVAR-GLP*	*SSVS*	*N-J*	*SNS*	*N-G*
				$h = 1$				
PAYEMS	-0.031	0.310**	0.213	0.284**	0.113	**0.351*** **	0.298**	0.244**
CPIAUCSL	0.867*	0.541	-1.426	-1.663	**1.821**	0.718*	1.170**	1.045**
FEDFUNDS	0.351	0.367	-0.036	0.115	0.296	**0.398**	0.333	0.340
GDP	0.113	0.189	**0.302*** **	0.250*	0.071	0.244**	0.256**	0.246**
UNRATE	0.528*	**0.645** **	0.477	0.452	0.542*	0.542*	0.528	0.563*
GDPDEFL	0.134*	0.196***	**0.247*** **	0.033	0.191**	0.240***	0.209***	0.208***
GS10	0.172**	**0.208*** **	0.016	0.169**	0.190**	0.195**	0.182**	0.200**
				$h = 2$				
PAYEMS	0.116	0.327*	-0.046	0.212**	0.097	**0.350***	0.325*	0.286
CPIAUCSL	-0.161	-0.080	-1.554	-2.276	**0.669**	-0.901	0.408	-0.088
FEDFUNDS	0.029	0.046	0.045	-0.308	0.001	**0.073**	0.038	0.014
GDP	-0.090	0.022	-0.030	-0.078	-0.129	0.071	**0.121***	-0.028
UNRATE	0.198	**0.271***	-0.213	-0.020	0.075	0.072	0.223***	0.211**
GDPDEFL	0.059	0.047	**0.106***	-0.084	0.071*	0.082**	0.080**	0.058
GS10	0.276***	0.284***	0.150	0.170**	0.307***	0.328***	0.318***	**0.343*** **
				$h = 3$				
PAYEMS	0.001	0.062	-0.399	-0.522	0.036	0.077	**0.147***	0.139
CPIAUCSL	0.202	0.996	-1.582	-2.471	**1.284**	-0.386	0.103	0.506
FEDFUNDS	0.034	0.031	**0.206*** **	0.134***	0.015	0.095***	0.059**	0.029
GDP	0.052	0.119	-0.013	0.029	0.109	0.146	**0.228***	0.211
UNRATE	**0.010**	-0.050	-0.234	-0.359	-0.010	-0.295	-0.082	-0.194
GDPDEFL	-0.004	0.002	**0.078***	-0.194	0.015	0.016	0.021	0.005
GS10	0.079	0.095**	0.068	0.045	0.093**	0.104**	0.110**	**0.115** **
				$h = 4$				
PAYEMS	-0.113	0.008	-0.610	-0.858	-0.051	0.047	**0.138**	0.056
CPIAUCSL	0.048	-0.247	-1.798	-2.590	**0.813**	-0.717	-0.643	-0.071
FEDFUNDS	0.038	0.047*	**0.220*** **	0.162***	0.025	0.111***	0.071**	0.046
GDP	0.040	0.138	0.018	0.049	0.106	0.130	**0.270**	0.120
UNRATE	-0.210	-0.060	-0.794	-0.722	**0.005**	-0.551	-0.231	-0.220
GDPDEFL	-0.019	-0.019	**0.049***	-0.218	-0.016	0.008	-0.011	-0.020
GS10	0.052	0.048	0.067*	0.039	0.047	**0.072**	0.071	0.072

This table reports the average log predictive likelihood (ALPL) differential between model i and the benchmark VAR(p) for the X-large VAR, computed as

$$ALPL_{ijh} = \frac{1}{\bar{t} - \underline{t} - h + 1} \sum_{\tau=\underline{t}}^{\bar{t}-h} \left(LPL_{i,j,\tau+h} - LPL_{bcmk,j,\tau+h} \right),$$

where $p = 1$, while $LPL_{i,j,\tau+h}$ and $LPL_{bcmk,j,\tau+h}$ are the log predictive likelihoods of variable j at time τ and forecast horizon h generated by model i and the VAR(p), respectively. $\underline{t}$ and $\bar{t}$ denote the start and end of the out-of-sample period, $i \in$ {DFM, FAVAR, BVAR-BGR, BVAR-GLP, SSVS, N-J, SNS, N-G}, $j \in$ {PAYEMS, CPIAUCSL,FEDFUNDS, GDP, UNRATE, GDPDEFL, GS10}, and $h = 1, ..., 4$. All forecasts are generated out-of-sample using recursive estimates of the models, with the out of sample period starting in 1985:Q1 and ending in 2015:Q4. Bold numbers indicate the highest ALPL across all models for a given variable-forecast horizon pair. * significance at the 10% level; ** significance at the 5% level; *** significance at the 1% level.

Appendix A Technical appendix

In this section, we provide detailed derivations and proofs of all the main results in the paper.

A.1 Derivation of the rotated regression and rotated likelihood

We begin by providing details on the derivation of the rotated regression in equation (3) and the joint likelihood of the rotated data in (4). Start with the simple univariate linear regression model in (1), which for convenience we report here

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{v}, \tag{A.1}$$

Next, introduce the $T \times T$ full-rank rotation matrix $\boldsymbol{Q}_j = \left[\boldsymbol{q}_j | \boldsymbol{W}_j\right]$ where $\boldsymbol{q}_j = \boldsymbol{X}_j / \left\|\boldsymbol{X}_j\right\|$ and $\boldsymbol{W}_j$ is an arbitrarily chosen $T \times (T-1)$ matrix subject to the constraint $\boldsymbol{W}_j \boldsymbol{W}_j' = \boldsymbol{I}_T - \boldsymbol{q}_j \boldsymbol{q}_j'$. Next, rewrite (A.1) as

$$\boldsymbol{y} = \boldsymbol{X}_j \beta_j + \boldsymbol{X}_{(-j)} \boldsymbol{\beta}_{(-j)} + \boldsymbol{v} \tag{A.2}$$

where $\boldsymbol{X}_{(-j)} = \boldsymbol{X} \setminus \boldsymbol{X}_j$ and $\boldsymbol{\beta}_{(-j)} = \boldsymbol{\beta} \setminus \beta_j$. Proceed by pre-multiplying both LHS and RHS of (A.2) by $\boldsymbol{Q}_j'$, to obtain

$$\boldsymbol{Q}_j' \boldsymbol{y} = \boldsymbol{Q}_j' \boldsymbol{X}_j \beta_j + \boldsymbol{Q}_j' \boldsymbol{X}_{(-j)} \boldsymbol{\beta}_{(-j)} + \boldsymbol{Q}_j' \boldsymbol{v}, \tag{A.3}$$

or, using the fact that $\boldsymbol{Q}_j = \left[\boldsymbol{q}_j | \boldsymbol{W}_j\right]$,

$$\begin{bmatrix} \boldsymbol{q}_j' \\ \boldsymbol{W}_j' \end{bmatrix} \boldsymbol{y} = \begin{bmatrix} \boldsymbol{q}_j' \\ \boldsymbol{W}_j' \end{bmatrix} \boldsymbol{X}_j \beta_j + \begin{bmatrix} \boldsymbol{q}_j' \\ \boldsymbol{W}_j' \end{bmatrix} \boldsymbol{X}_{(-j)} \boldsymbol{\beta}_{(-j)} + \begin{bmatrix} \boldsymbol{q}_j' \\ \boldsymbol{W}_j' \end{bmatrix} \boldsymbol{v}. \tag{A.4}$$

Now using the definition of $\boldsymbol{q}_j$ and the formulas for y_j^* and $\widetilde{\boldsymbol{y}}_j$ in (2), we have that

$$\begin{bmatrix} y_j^* \\ \widetilde{\boldsymbol{y}}_j \end{bmatrix} = \begin{bmatrix} \left(\boldsymbol{X}_j' \boldsymbol{X}_j / \left\|\boldsymbol{X}_j\right\|\right) \\ \boldsymbol{W}_j' \boldsymbol{q}_j \left\|\boldsymbol{X}_j\right\| \end{bmatrix} \beta_j + \begin{bmatrix} \boldsymbol{q}_j' \boldsymbol{X}_{(-j)} \\ \boldsymbol{W}_j' \boldsymbol{X}_{(-j)} \end{bmatrix} \boldsymbol{\beta}_{(-j)} + \begin{bmatrix} \boldsymbol{q}_j' \boldsymbol{v} \\ \boldsymbol{W}_j' \boldsymbol{v} \end{bmatrix}, \tag{A.5}$$

Further simplifications lead to (3), i.e.

$$\begin{bmatrix} y_j^* \\ \widetilde{\boldsymbol{y}}_j \end{bmatrix} = \begin{bmatrix} \left\|\boldsymbol{X}_j\right\| \beta_j \\ \boldsymbol{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{X}_{(-j)}^* \boldsymbol{\beta}_{(-j)} \\ \widetilde{\boldsymbol{X}}_{(-j)} \boldsymbol{\beta}_{(-j)} \end{bmatrix} + \begin{bmatrix} v_j^* \\ \widetilde{\boldsymbol{v}}_j \end{bmatrix}, \tag{A.6}$$

where we have exploited the following two results:

1. $\left(\boldsymbol{X}_j' \boldsymbol{X}_j / \left\|\boldsymbol{X}_j\right\|\right) = \left\|\boldsymbol{X}_j\right\|$. This is due to the fact that $\boldsymbol{X}_j' \boldsymbol{X}_j = \left\|\boldsymbol{X}_j\right\|^2$;

2. By definition, $\boldsymbol{W}_j'$ and $\boldsymbol{q}_j$ are orthogonal. They all are columns of the orthogonal matrix $\boldsymbol{Q}_j$, so by construction $\boldsymbol{W}_j'\boldsymbol{q}_j = \boldsymbol{0}$.

Next, to go from (3) to (4), note that $E\left(\boldsymbol{Q}_j'\boldsymbol{v}\right) = \boldsymbol{0}$ while $var\left(\boldsymbol{Q}_j'\boldsymbol{v}\right) = \sigma^2\boldsymbol{Q}_j\boldsymbol{Q}_j' = \sigma^2\boldsymbol{I}_T$ which, combined with (A.6), leads to the rotated likelihood in equation (4). ■

A.2 Derivation of the rotated conditional likelihood

In this subsection, we provide details on the results in equations (6), (7), and (8). Start by focusing on the top row of (4), and note that the conditional density $p\left(y_j^*|\boldsymbol{\beta},\sigma^2\right)$ can be decomposed as follows

$$y_j^* = \|\boldsymbol{X}_j\|\,\beta_j + y_j^+ \tag{A.7}$$

where

$$y_j^+|\boldsymbol{\beta}_{(-j)},\sigma^2 \sim \mathcal{N}\left(\boldsymbol{X}_{(-j)}^*\boldsymbol{\beta}_{(-j)},\sigma^2\right) \tag{A.8}$$

Notice that the newly defined $p\left(y_j^+|\boldsymbol{\beta}_{(-j)},\sigma^2\right)$ can be interpreted as essentially the predictive distribution associated with the auxiliary regression that is defined in the second row of (4). This leads to the following result,

$$\begin{aligned} p\left(y_j^*|\beta_j,\widetilde{\boldsymbol{y}}_j\right) &= \|\boldsymbol{X}_j\|\,\beta_j + p\left(y_j^+|\widetilde{\boldsymbol{y}}_j\right) \\ &= \|\boldsymbol{X}_j\|\,\beta_j + \int\int p\left(y_j^+|\boldsymbol{\beta}_{(-j)},\sigma^2,\widetilde{\boldsymbol{y}}_j\right) p\left(\boldsymbol{\beta}_{(-j)},\sigma^2|\widetilde{\boldsymbol{y}}_j\right) d\boldsymbol{\beta}_{(-j)}d\sigma^2 \end{aligned} \tag{A.9}$$

The key to solving (A.9) is to compute the integral in the second row of the equation, which in turn will depend on the prior distribution adopted for $p\left(\boldsymbol{\beta}_{(-j)},\sigma^2\right)$. As we discussed in Section 2, for computational tractability we chose to rely on the natural conjugate prior,

$$\begin{aligned} \boldsymbol{\beta}_{(-j)}|\sigma^2 &\sim \mathcal{N}\left(\underline{\boldsymbol{\beta}}_{(-j)},\sigma^2\underline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(-j)}}\right) \\ \sigma^2 &\sim \mathcal{IG}\left(\underline{\psi},\underline{d}\right) \end{aligned} \tag{A.10}$$

It is straightforward to show that the posterior distribution $p\left(\boldsymbol{\beta}_{(-j)},\sigma^2|\widetilde{\boldsymbol{y}}_j\right)$ also belongs to the Normal-Inverse-Gamma (NIG) family, and is given by

$$\begin{aligned} \boldsymbol{\beta}_{(-j)}|\sigma^2,\widetilde{\boldsymbol{y}}_j &\sim \mathcal{N}\left(\overline{\boldsymbol{\beta}}_{(-j)},\sigma^2\overline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(-j)}}\right) \\ \sigma^2|\widetilde{\boldsymbol{y}}_j &\sim \mathcal{IG}\left(\overline{\psi}_{(-j)},\overline{d}\right) \end{aligned} \tag{A.11}$$

where $\overline{d} = \underline{d} + (T-1)/2$,

$$\overline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(-j)}} = \left(\underline{\boldsymbol{V}}^{-1}_{\boldsymbol{\beta}_{(-j)}} + \widetilde{\boldsymbol{X}}'_{(-j)}\widetilde{\boldsymbol{X}}_{(-j)}\right)^{-1}, \tag{A.12}$$

$$\overline{\boldsymbol{\beta}}_{(-j)} = \overline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(-j)}}\left(\underline{\boldsymbol{V}}^{-1}_{\boldsymbol{\beta}_{(-j)}}\underline{\boldsymbol{\beta}}_{(-j)} + \widetilde{\boldsymbol{X}}'_{(-j)}\widetilde{\boldsymbol{y}}_j\right), \tag{A.13}$$

and

$$\overline{\psi}_{(-j)} = \underline{\psi} + \frac{1}{2}\left(\widetilde{\boldsymbol{y}}'_j\widetilde{\boldsymbol{y}}_j + \underline{\boldsymbol{\beta}}'_{(-j)}\underline{\boldsymbol{V}}^{-1}_{\boldsymbol{\beta}_{(-j)}}\underline{\boldsymbol{\beta}}_{(-j)} - \overline{\boldsymbol{\beta}}'_{(-j)}\overline{\boldsymbol{V}}^{-1}_{\boldsymbol{\beta}_{(-j)}}\overline{\boldsymbol{\beta}}_{(-j)}\right). \tag{A.14}$$

Armed with an analytical expression for the posterior $p\left(\boldsymbol{\beta}_{(-j)}, \sigma^2|\widetilde{\boldsymbol{y}}_j\right)$, we are now ready to derive the rotated conditional likelihood:

$$\begin{aligned}
p\left(y^*_j|\beta_j, \widetilde{\boldsymbol{y}}_j\right) &= \|\boldsymbol{X}_j\|\beta_j + \int\int p\left(y^+_j|\boldsymbol{\beta}_{(-j)}, \sigma^2, \widetilde{\boldsymbol{y}}_j\right) p\left(\boldsymbol{\beta}_{(-j)}, \sigma^2|\widetilde{\boldsymbol{y}}_j\right) d\boldsymbol{\beta}_{(-j)}d\sigma^2 \\
&= \|\boldsymbol{X}_j\|\beta_j + \int\int \mathcal{N}\left(\boldsymbol{X}^*_{(-j)}\boldsymbol{\beta}_{(-j)}, \sigma^2\right) \times \\
&\times \mathcal{N}\left(\overline{\boldsymbol{\beta}}_{(-j)}, \sigma^2\overline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(-j)}}\right)\mathcal{IG}\left(\overline{\psi}_{(-j)}, \overline{d}\right) d\boldsymbol{\beta}_{(-j)}d\sigma^2 \\
&= \|\boldsymbol{X}_j\|\beta_j + t_{2\overline{d}}\left(\overline{\mu}_j, \overline{\tau}^2_j\right) \\
&\approx \|\boldsymbol{X}_j\|\beta_j + \mathcal{N}\left(\overline{\mu}_j, \overline{\tau}^2_j\right)
\end{aligned} \tag{A.15}$$

where

$$\overline{\mu}_j = \boldsymbol{X}^*_{(-j)}\overline{\boldsymbol{\beta}}_{(-j)} \tag{A.16}$$

and

$$\overline{\tau}^2_j = \frac{\overline{\psi}_{(-j)}}{\overline{d}}\left(1 + \boldsymbol{X}^*_{(-j)}\overline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(-j)}}\boldsymbol{X}^{*\prime}_{(-j)}\right). \tag{A.17}$$

This concludes the derivations of equations (6), (7), and (8). ■

A.3 Calculation of optimal shrinkage intensity under a Normal-Jeffreys prior

Start with the approximation in (6), which here we slightly rearrange to be

$$\left(y^*_j - \overline{\mu}_j\right)|\beta_j, \widetilde{\boldsymbol{y}}_j \sim \mathcal{N}\left(\|\boldsymbol{X}_j\|\beta_j, \overline{\tau}^2_j\right),$$

and write the Normal-Jeffreys prior as in (9)

$$\beta_j|\lambda^2_j \sim \mathcal{N}\left(0, \lambda^2_j\underline{V}_{\beta_j}\right) \tag{A.18}$$

Next, the marginal likelihood $p\left(y_j^* - \overline{\mu}_j \middle| \lambda_j^2, \widetilde{\boldsymbol{y}}_j\right)$ is given by

$$\begin{aligned} p\left(y_j^* - \overline{\mu}_j \middle| \lambda_j^2, \widetilde{\boldsymbol{y}}_j\right) &= \int p\left(y_j^* - \overline{\mu}_j \middle| \beta_j, \widetilde{\boldsymbol{y}}_j\right) p\left(\beta_j \middle| \lambda_j^2\right) d\beta_j \\ &= \mathcal{N}\left(y_j^* - \overline{\mu}_j \middle| \left\|\boldsymbol{X}_j\right\|^2 \lambda_j^2 \underline{V}_{\beta_j} + \overline{\tau}_j^2\right). \end{aligned} \tag{A.19}$$

or, more explicitly,

$$p\left(y_j^* - \overline{\mu}_j | \lambda_j^2, \widetilde{\boldsymbol{y}}_j\right) = \frac{1}{\sqrt{2\pi}\sqrt{\overline{\tau}_j^2 + \left\|\boldsymbol{X}_j\right\|^2 \lambda_j^2 \underline{V}_{\beta_j}}} \times \exp\left(-\frac{\left(y_j^* - \overline{\mu}_j\right)^2}{2\left(\overline{\tau}_j^2 + \left\|\boldsymbol{X}_j\right\|^2 \lambda_j^2 \underline{V}_{\beta_j}\right)}\right)$$

To find the λ_j^2 that maximizes $p\left(y_j^* - \overline{\mu}_j\right) | \lambda_j^2, \widetilde{\boldsymbol{y}}_j$, take the log and only focus on the terms that involve λ_j^2:

$$\ln p\left(y_j^* - \overline{\mu}_j | \lambda_j^2, \widetilde{\boldsymbol{y}}_j\right) \propto -\frac{1}{2} \ln\left(\overline{\tau}_j^2 + \left\|\boldsymbol{X}_j\right\|^2 \lambda_j^2 \underline{V}_{\beta_j}\right) - \frac{1}{2} \frac{\left(y_j^* - \overline{\mu}_j\right)^2}{\left(\overline{\tau}_j^2 + \left\|\boldsymbol{X}_j\right\|^2 \lambda_j^2 \underline{V}_{\beta_j}\right)}$$

Now taking the derivative with respect to λ_i^2 and setting it to zero

$$\frac{\partial \ln p\left(y_j^* - \overline{\mu}_j | \lambda_j^2, \widetilde{\boldsymbol{y}}_j\right)}{\partial \lambda_j^2} = -\frac{1}{2} \frac{\left\|\boldsymbol{X}_j\right\|^2 \underline{V}_{\beta_j}}{\left(\overline{\tau}_j^2 + \left\|\boldsymbol{X}_j\right\|^2 \lambda_j^2 \underline{V}_{\beta_j}\right)} + \frac{1}{2} \frac{\left(y_j^* - \overline{\mu}_j\right)^2 \left\|\boldsymbol{X}_j\right\|^2 \underline{V}_{\beta_j}}{\left(\overline{\tau}_j^2 + \left\|\boldsymbol{X}_j\right\|^2 \lambda_j^2 \underline{V}_{\beta_j}\right)^2} = 0$$

leads to the solution in (12),

$$\widehat{\lambda}_j^2 = \max\left[0, \frac{\left(y_j^* - \overline{\mu}_j\right)^2 - \overline{\tau}_j^2}{\left\|\boldsymbol{X}_j\right\|^2 \underline{V}_{\beta_j}}\right]. \tag{A.20}$$

■

A.4 Derivation of posterior probability of inclusion under a Spike-and-Slab prior

Start with (18), which for convenience we rewrite here as

$$\widehat{\pi}_j = p\left(\lambda_j = 1 \middle| \, y_j^*, \widetilde{\boldsymbol{y}}_j\right) = \frac{p\left(y_j^* \middle| \, \lambda_j = 1, \widetilde{\boldsymbol{y}}_j\right) p\left(\lambda_j = 1\right)}{p\left(y_j^* \middle| \, \lambda_j = 0, \widetilde{\boldsymbol{y}}_j\right) p\left(\lambda_j = 0\right) + p\left(y_j^* \middle| \, \lambda_j = 1, \widetilde{\boldsymbol{y}}_j\right) p\left(\lambda_j = 1\right)} \tag{A.21}$$

Next, notice that $= p(\lambda_j = 1) = \underline{\pi}_0$ and $p(\lambda_j = 0) = 1 - \underline{\pi}_0$. Furthermore, the approximation in (6) along with the independence between β_j and $\widetilde{\boldsymbol{y}}_j$ imply that

$$\begin{aligned} p\left(y_j^* \middle|\ \lambda_j = 1, \widetilde{\boldsymbol{y}}_j\right) &\approx \int p\left(y_j^* \middle|\ \beta_j, \lambda_j = 1, \widetilde{\boldsymbol{y}}_j\right) p\left(\beta_j \middle|\, \lambda_j = 1, \widetilde{\boldsymbol{y}}_j\right) d\beta_j \\ &\approx \int p\left(y_j^* \middle|\ \beta_j, \lambda_j = 1, \widetilde{\boldsymbol{y}}_j\right) p\left(\beta_j \middle|\, \lambda_j = 1\right) d\beta_j \\ &\sim \mathcal{N}\left(y_j^* \middle|\, \overline{\mu}_j, \overline{\tau}_j^2 + \|\boldsymbol{X}_j\|^2 \underline{V}_{\beta_j}\right) \end{aligned} \tag{A.22}$$

while, similarly,

$$\begin{aligned} p\left(y_j^* \middle|\ \lambda_j = 0, \widetilde{\boldsymbol{y}}_j\right) &\approx \int p\left(y_j^* \middle|\ \beta_j, \lambda_j = 0, \widetilde{\boldsymbol{y}}_j\right) p\left(\beta_j \middle|\, \lambda_j = 0, \widetilde{\boldsymbol{y}}_j\right) d\beta_j \\ &\approx \int p\left(y_j^* \middle|\ \beta_j, \lambda_j = 0, \widetilde{\boldsymbol{y}}_j\right) p\left(\beta_j \middle|\, \lambda_j = 0\right) d\beta_j \\ &\sim \mathcal{N}\left(y_j^* \middle|\, \overline{\mu}_j, \overline{\tau}_j^2\right) \end{aligned} \tag{A.23}$$

Plugging (A.22) and (A.23) into (A.21) leads to (19). ■

A.5 Triangularization of the VAR

Start from the n-dimensional VAR(p) model in (21), which for convenience we rewrite here

$$\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \ldots + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t, \quad t = 1, ..., T, \tag{A.24}$$

where $\boldsymbol{y}_t$ is an $n \times 1$ vector of time series of interest, $\boldsymbol{c}$ is an $n \times 1$ vector of intercepts, $\boldsymbol{A}_1, ..., \boldsymbol{A}_p$ are $n \times n$ matrices of coefficients on the lagged dependent variables, and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega})$, with $\boldsymbol{\Omega}$ an $n \times n$ covariance matrix. Next, following Carriero et al. (2016), decompose the VAR covariance matrix $\boldsymbol{\Omega}$ in (A.24) as $\boldsymbol{\Omega} = \boldsymbol{\Gamma}^{-1} \boldsymbol{\Sigma} \left(\boldsymbol{\Gamma}^{-1}\right)'$, where

$$\boldsymbol{\Gamma}^{-1} = \begin{bmatrix} 1 & 0 & ... & 0 & 0 \\ \gamma_{2,1} & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 \\ \gamma_{n-1,1} & ... & \gamma_{n-1,n-2} & 1 & 0 \\ \gamma_{n,1} & ... & \gamma_{n,n-2} & \gamma_{n,n-1} & 1 \end{bmatrix}, \tag{A.25}$$

and $\boldsymbol{\Sigma} = diag\left(\sigma_1^2, ..., \sigma_n^2\right)$. Under this decomposition the residuals of the original VAR(p) in (A.24) can be written using the identity $\boldsymbol{\varepsilon}_t = \boldsymbol{\Gamma}^{-1} \boldsymbol{\Sigma}^{1/2} \boldsymbol{u}_t$, with $\boldsymbol{u}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, which implies that the i-th row of this identity is

$$\varepsilon_{i,t} = \gamma_{i,1} \sigma_1 u_{1,t} + ... + \gamma_{i,i-1} \sigma_{i-1} u_{i-1,t} + \sigma_i u_{i,t}. \tag{A.26}$$

As a result, the VAR(p) in equation (A.24) admits the following triangular structure,

$$\begin{aligned} y_{1,t} &= c_1 + \boldsymbol{a}_{1,\cdot}\boldsymbol{Z}_t + \sigma_1 u_{1t}, \\ y_{2,t} &= c_2 + \boldsymbol{a}_{2,\cdot}\boldsymbol{Z}_t + \gamma_{2,1}\sigma_1 u_{1,t} + \sigma_2 u_{2,t}, \\ &\vdots \\ y_{n,t} &= c_n + \boldsymbol{a}_{n,\cdot}\boldsymbol{Z}_t + \gamma_{n,1}\sigma_1 u_{1,t} + ... + \gamma_{n,n-1}\sigma_{n-1} u_{n-1,t} + \sigma_n u_{n,t}, \end{aligned} \tag{A.27}$$

where $\boldsymbol{a}_{i,\cdot} = [a_{i,1}, ..., a_{i,np}]$ denotes the vector of coefficients in the i-th VAR equation, and $\boldsymbol{Z}_t = \left[\boldsymbol{y}'_{t-1}, ..., \boldsymbol{y}'_{t-p}\right]'$. As noted by Carriero et al. (2016), the re-parametrization of the VAR(p) in (A.27) allows for estimation of the system recursively, equation-by-equation.[26] For example, consider the generic equation i, which we rewrite as

$$y_{i,t} = c_i + \boldsymbol{a}_{i,\cdot}\boldsymbol{Z}_t + \gamma_{i,1}\sigma_1 u_{1,t} + ... + \gamma_{i,i-1}\sigma_{i-1} u_{i-1,t} + \sigma_i u_{i,t}, \tag{A.28}$$

Provided that all previous $i-1$ equations have been already estimated, all terms on the right hand side of (A.28) involving the previous equation error terms can be replaced by their estimated counterparts. As a result, the full posterior for the VAR parameters $\left\{\boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\Gamma}^{-1}, \boldsymbol{\Sigma}\right\}$ can now be obtained recursively, one equation at a time.

[26] It is worth pointing out an important feature that affects all models that rely on the triangularization in (A.27). If the priors for $\boldsymbol{\Gamma}^{-1}$ and $\boldsymbol{\Sigma}$ are elicited separately, the implied prior for $\boldsymbol{\Omega}$ will change when the ordering of the equations in the VAR changes. As a result, different orderings of the variables in the VAR will lead to different prior specifications for $\boldsymbol{\Omega}$ and potentially different joint posteriors of the BVAR parameters $\{\boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\Omega}\}$. As noted by Primiceri (2005), this problem will likely be less severe in the case as it is here in which the elements of the covariance matrix in $\boldsymbol{\Gamma}^{-1}$ do not vary with time, because the likelihood will quickly dominate the prior as the sample size increases. On this point, see also the estimation algorithms of Smith and Kohn (2002) and George et al. (2008) and discussions therein.

Appendix B Data and transformations

Table B.1. List of series

Series id	*Tcode*[†]	*Medium*	*Large*	*X-large*	*FRED*	*Description*
1	5		X	X	RPI	Real Personal Income
2	5		X	X	W875RX1	RPI ex. Transfers
3	5		X	X	DPCERA3M086SBEA	Real PCE
4	5		X	X	CMRMTSPLx	Real M&T Sales
5	5		X	X	RETAILx	Retail and Food Services Sales
6	5			X	INDPRO	IP Index
7	2			X	HWI	Help-Wanted Index for US
8	2			X	HWIURATIO	Help Wanted to Unemployed ratio
9	5			X	CLF16OV	Civilian Labor Force
10	2	X	X	X	UNRATE	Civilian Unemployment Rate
11	5	X	X	X	PAYEMS	All Employees: Total nonfarm
12	5			X	CES0600000007	Hours: Goods-Producing
13	5			X	M1SL	M1 Money Stock
14	5			X	M2SL	M2 Money Stock
15	5			X	M2REAL	Real M2 Money Stock
16	5		X	X	BUSLOANS	Commercial and Industrial Loans
17	5		X	X	NONREVSL	Total Nonrevolving Credit
18	2		X	X	CONSPI	Credit to PI ratio
19	5			X	S&P 500	S&P 500
20	5			X	S&P: indust	S&P Industrial
21	2			X	S&P div yield	S&P Divident yield
22	5			X	S&P PE ratio	S&P Price/Earnings ratio
23	2	X	X	X	FEDFUNDS	Effective Federal Funds Rate
24	2		X	X	CP3M	3-Month AA Comm. Paper Rate
25	2			X	TB3MS	3-Month T-bill
26	2			X	TB6MS	6-Month T-bill
27	2			X	GS1	1-Year T-bond
28	2			X	GS5	5-Year T-bond
29	2	X	X	X	GS10	10-Year T-bond
30	2			X	AAA	Aaa Corporate Bond Yield
31	2			X	BAA	Baa Corporate Bond Yield
32	5		X	X	EXSZUS	Switzerland / U.S. FX Rate
33	5		X	X	EXJPUS	Japan / U.S. FX Rate
34	5		X	X	EXUSUK	U.S. / U.K. FX Rate
35	5		X	X	EXCAUS	Canada / U.S. FX Rate
36	5			X	OILPRICEx	Crude Oil Prices: WTI
37	6	X	X	X	CPIAUCSL	CPI: All Items
38	5			X	INVEST	Securities in Bank Credit
39	5	X	X	X	GDP	Real Gross Domestic Product
40	6	X	X	X	GDPDEFL	GDP deflator

[†] Transformation code. These stand for: 2 - first differences; 5 - first differences of logarithms; 6 - second differences of logarithms