Macroeconomic Forecasting with Large Bayesian VARs: Global-local Priors and the Illusion of Sparsity

Jamie L. Cross [1,2] Chenghan Hou[3*] Aubrey Poon[2,4]

[1]BI Norwegian Business School, Centre of Applied Macroeconomics and Commodity Prices (CAMP)

[2]Centre for Applied Macroeconomic Analysis (CAMA)

[3]Hunan University, Center for Finance and Management Studies (CEFMS)

[4]University of Strathclyde

August 3, 2019

Abstract

A class of global-local hierarchical shrinkage priors for estimating large Bayesian vector autoregressions (BVARs) has recently been proposed. We question whether three such priors: Dirichlet-Laplace, Horseshoe and Normal-Gamma, can systematically improve upon the forecast accuracy of two commonly used benchmarks: hierarchical Minnesota prior and stochastic search variable selection prior (SSVS), when predicting key macroeconomic variables. Using small and large data sets, both point and density forecasts suggest that the answer is no. Instead, our results indicate that a hierarchical Minnesota prior remains a solid practical choice when forecasting macroeconomic variables. In light of existing optimality results, a possible explanation for our finding is that macroeconomic data is not sparse, but instead dense.

Keywords: Bayesian VAR, Forecasting, Shrinkage Prior, Stochastic Volatility.

*Corresponding author. We thank Gary Koop, Dimitris Korobilis and participants of the 13th RCEA Bayesian Workshop for their comments on the paper.

1 Introduction

A key finding of 21st century macroeconometrics is that big data sets can be beneficial for forecasting macroeconomic variables (see, e.g. Stock and Watson (2002a,c); De Mol et al. (2008); Bańbura et al. (2010); Koop and Korobilis (2013); Huber and Feldkircher (2017); Chan (2018)). Following Bańbura et al. (2010), the benchmark model is a large BVAR with Minnesota prior. However, a well-known weakness of this prior is that its estimates are sensitive to the hyperparameter values (Giannone et al., 2015). This led to the development of hierarchical priors that introduce priors on the hyperparameters—so-called *hyperpriors*—and integrate them out in a Bayesian manner. For instance, Giannone et al. (2015) show that placing a hyperprior on the key regularization hyperparameter of the conjugate Minnesota prior, can generate more accurate macroeconomic forecasts than conventional choices. Another popular approach is the stochastic search variable selection (SSVS) prior (George et al., 2008). This prior specifies that each coefficient follows a two-state Gaussian mixture in which one distribution is relatively tight, while the other is more disperse. Set in this manner, the SSVS prior groups coefficients into those that are strongly regularized and those that are not. Despite their popularity, these hierarchical priors impose a trade-off in terms of flexibility and complexity. For instance, the hierarchical conjugate Minnesota prior possesses a Kronecker structure which induces symmetric regularization across the BVAR equations.

With this potential limitation in mind, the main contribution of this paper is to assess whether a class of recently developed *global-local hierarchical shrinkage priors* (hereon referred to as GL priors), can improve upon the forecast accuracy of these two popular benchmarks. As the name suggests, GL priors introduce both global and local variance components into their prior distribution. The *global shrinkage component* places substantial probability mass near the prior mean, while the *local shrinkage component* induces fat-tailed behavior in the marginal prior distribution. Set in this manner, GL priors heavily regularize the model parameters while retaining enough flexibility to prevent possible weak signals from being forced to zero. This feature is important in the presence of sparsity, where such priors attain optimal rates of posterior contraction (van der Pas et al., 2016). Despite this result, it remains to be seen whether this property is useful for macroeconomic forecasting, where the presence of sparsity is less clear (Giannone et al., 2018).

To investigate this idea, we consider three GL priors that have been recently introduced to the large BVAR literature: Dirichlet-Laplace (Kastner and Huber, 2017), Horseshoe (Follett and Yu, 2017) and Normal-Gamma (Huber and Feldkircher, 2017). While each of these priors has been shown to provide reasonable forecasts of macroeconomic variables against simple benchmarks, their relative forecast accuracy against both each other and hierarchical Minnesota and SSVS benchmarks remains unexplored. Specifically, Kastner and Huber (2017) show that a large BVAR with Dirichlet-Laplace prior and factor stochastic volatility (FSV) has both superior in- and out-of-sample properties when compared to a pure FSV model with conditional mean equal to zero; Follett and Yu (2017) show how that a large BVAR with a Horseshoe prior can provide more accurate macroeconomic forecasts than a range of alternative shrinkage priors: Student's-t, Laplace, Ridge and discrete mixture; and Huber and Feldkircher (2017) show that the Normal-Gamma prior provides competitive macroeconomic forecasts compared to a simple Minnesota prior benchmark. Our paper therefore builds on this emerging literature by comparing both the relative forecast performance of these priors both against each other and conventional benchmarks. In particular, we directly extend the set of models in Huber and Feldkircher (2017) to include two additional GL priors: the Dirichlet-Laplace and Horseshoe. Moreover, in the spirit of Giannone et al. (2015), we adopt a hierarchical approach when estimating the hyperparameters associated with each prior. As we will later show, this addition avoids the well-known pitfalls associated with user-specified values of the hyperparameters; i.e., they may lead to too little regularization, or too much, and consequently generate misleading forecast results.

Our forecasting exercise consists of comparing both the point and density forecast performance of the five BVARs mentioned above on a variant of the well known Stock and Watson (2011) dataset of quarterly US macroeconomic variables provided by McCracken and Ng (2016). Specifically, we use both small and large data sets over the evaluation period 1979Q1-2018Q2 to forecast three key macroeconomic indicators: real GDP growth, GDP deflator measured inflation, and the effective federal funds rate (FFR). Since our sample contains the entire zero-lower-bound period, we use the Wu and Xia (2016) shadow rate during this part of the sample. Also, since stochastic volatility is well known to improve the forecast accuracy of these variables (Clark and Ravazzolo, 2015; Carriero et al., 2016, 2019), we show how a stochastic volatility framework of Cogley and Sargent (2005) can be easily implemented under each prior specification.

Using a range of metrics for forecast accuracy, we find that the simple hierarchical Minnesota prior often provides more accurate forecasts of inflation, the Normal-Gamma and Horseshoe prior respectively provide similarly accurate forecasts of real GDP and the FFR. That being said, the improvements offered by the Minnesota prior are often statistically significant, while the GL priors never provide such improvements upon the Minnesota prior. This finding is important because it suggests that the documented concerns of this prior in sparse settings do not lead to subpar forecast accuracy, at least for macroeconomic forecasting. In light of existing optimality results for GL priors in sparse settings (van der Pas et al., 2016), one possible explanation for this result is that macroeconomic data is not sparse, but instead dense. In fact, this idea is in line with a recent paper by Giannone et al. (2018), who comment on the *illusion of sparsity* in economic and financial modeling.

To test the validity of this claim, we conduct an extensive simulation study based on both dense and sparse data generating processes (DGPs). The dense DGP allows for a broad set of explanatory variables to be important for prediction, although their individual impact is small. In contrast, the sparse DGP features a small set of explanatory variables, each with high predictive power. Indeed, our results suggest that while the GL priors are superior in recovering the true model parameters in a sparse environment, a suitable hierarchical Minnesota prior can attain competitive performance in dense environments. Of course, such an exercise can not be taken as a formal proof of the density of macroeconomic data; however, it is suggestive of such a phenomenon. Since such a process requires the development of sophisticated statistical tools, we leave this open as an important avenue of future research.

The rest of the paper is structured as follows. We present the modeling framework used in the empirical study in Section 2, discuss the priors and associated posteriors in Section 3, and present the main forecasting results in Section 4. In Section 5, we examine the theoretical properties of these priors through a simulation exercise. We present some additional analysis and robustness checks in Section 6 and conclude in Section 7.

2 The Bayesian VAR with Stochastic Volatility (BVAR-SV)

The BVAR-SV model has a state-space representation with the observation equation

$$\mathbf{A}_0\mathbf{y}_t = \mathbf{B}_1\mathbf{y}_{t-1} + \ldots + \mathbf{B}_p\mathbf{y}_{t-p} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_t\right), \tag{2.1}$$

where $\mathbf{y}_t$ is an $n \times 1$ vector of endogenous variables, the $\mathbf{B}_i$'s, $i = 1, \ldots, p$ are the $n \times n$ coefficients matrices, $\mathbf{A}_0$ is an $n \times n$ lower triangular matrix with ones on the diagonal, $\boldsymbol{\epsilon}_t$ is an $n \times 1$ vector of iid residuals, $\mathcal{N}(\cdot, \cdot)$ denotes the Gaussian distribution and $\boldsymbol{\Sigma}_t = \text{diag}(\exp(h_{1,t}), \ldots, \exp(h_{n,t}))$.[1] For specification purposes it is useful to rewrite (2.1) as

$$\mathbf{y}_t = \tilde{\mathbf{X}}_t\boldsymbol{\beta} + \mathbf{W}_t\boldsymbol{\gamma} + \boldsymbol{\epsilon}_t, \tag{2.2}$$

where $\tilde{\mathbf{X}}_t = \text{I}_n \otimes [1, \mathbf{y}_{t-1}^{'}, \ldots, \mathbf{y}_{t-p}^{'}]$, $\boldsymbol{\beta} = \text{vec}([\mathbf{B}_1, \ldots, \mathbf{B}_p]^{'})$ is a $k_\beta \times 1$ vector and $\boldsymbol{\gamma}$ is a $k_\gamma \times 1$ vector of the contemporaneous coefficients in $\mathbf{A}_0$ (collected by rows) and $\mathbf{W}_t$ is a $n \times k_\gamma$ matrix of associated elements in $-\mathbf{y}_t$. It is easy to check that $k_\beta = n^2p$ and $k_\gamma = n(n-1)/2$.

Following Cogley and Sargent (2005), we assume that the state equation for the latent log-volatilities $\mathbf{h}_t = (h_{1,t}, \ldots, h_{n,t})'$ follow the subsequent independent random walk process

$$\mathbf{h}_t = \mathbf{h}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}), \tag{2.3}$$

where $\boldsymbol{\eta}_t$ is an $n \times 1$ vector of iid residuals, $\boldsymbol{\Omega} = \text{diag}\left(\sigma_{h_1}^2, \ldots, \sigma_{h_n}^2\right)$ and the initial condition $\mathbf{h}_0$ is treated as a parameter to be estimated.

The model specification is completed by specifying prior distributions for the lagged and contemporaneous coefficients ($\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$), the state variances ($\boldsymbol{\Omega}$) and the initial state ($\mathbf{h}_0$). We consider a range of hierarchical shrinkage priors on the coefficients and defer details to the next section. As for the state variances and initial condition, we specify independent prior distributions as follows

$$\mathbf{h}_0 \sim \mathcal{N}\left(\mathbf{a}_\mathbf{h}, \mathbf{V}_\mathbf{h}\right), \;\; \sigma_{h_i}^2 \sim \mathcal{IG}(\nu_{h_i}, S_{h_i}), \;\; \text{for} \quad i = 1, \ldots n, \tag{2.4}$$

[1] Since we demean each of the series prior to estimation, we have excluded any intercept from (2.1).

where $\mathcal{IG}(\cdot,\cdot)$ denotes the inverse Gamma distribution. In the forecasting exercise and simulation study we set $\mathbf{a_h} = \mathbf{0}$, $\mathbf{V_h} = 10\times\mathbf{I}_n$, $\nu_{h_i} = 10$ and $S_{h_i} = 0.1^2\,(\nu_{h_i} - 1)$ across each model specification. The conditional posteriors for the stochastic volatilities in this framework have been documented in Chan and Eisenstat (2018). To sample the volatilities, we make use of the efficient precision sampling algorithm in Chan and Hsiao (2014). We will not reproduce the details but note that the algorithm combines both the auxiliary mixture sampler of Kim et al. (1998) with the precision sampling methods for linear Gaussian state-space models developed in Chan and Jeliazkov (2009). As we are considering a large dateset, a computational difficulty arises from the high dimension of the lagged coefficients. Specifically, when n is large, inverting the $n^2p \times n^2p$ covariance matrix of the conditional posterior distribution of the lagged coefficients is extremely computationally demanding. To overcome this issue, we follow Carriero et al. (2019) and sample from the conditional posterior distributions equation-by-equation. Finally, in line with key papers in the literature, we specify a lag length of $p = 4$ in all model variants (Koop, 2013; Chan, 2018).

3 Priors

A major difficulty in estimating large BVARs is that the number of coefficients increases rapidly with the number of endogenous variables. Without appropriate regularization, parameter uncertainty would make any estimates and associated analysis unreliable. In this section we show how this overparameterization problem can be overcome by using the prior on the lagged coefficients as a regularization tool. A similar setup can also be applied to the contemporaneous coefficients.

To facilitate the discussion, note that a generic Gaussian prior on the BVAR coefficients takes the form

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V}), \tag{3.1}$$

where $\boldsymbol{\mu}$ and $\mathbf{V}$ are researcher specified hyperparameters. When the variables are stationary, regularization can be achieved by setting $\boldsymbol{\mu} = \mathbf{0}$.[2] The prior variance matrix $\mathbf{V}$ is assumed to be diagonal, i.e. $\mathbf{V} = \text{diag}\left(v_1, \ldots, v_{k_{\boldsymbol{\beta}}}\right)$. Various shrinkage priors differentiate themselves through their treatment of $\mathbf{V}$. In this paper we focus on two traditional priors: hierarchical Minnesota and SSVS, and three GL priors: Dirichlet-Laplace, Horseshoe and Normal-Gamma.

[2] In our empirical study we ensure that all variables in the BVAR system are stationary.

3.1 Minnesota Prior

Many versions of the Minnesota prior have been used in practice (see, e.g., Karlsson (2013) for a comprehensive discussion). Here we follow Koop et al. (2010) and specify the following variant from Litterman (1979, 1980). For expository purposes, note that the diagonal elements of the prior variance matrix can be written as $\left(v_1, \ldots, v_{k_{\boldsymbol{\beta}}}\right) = \text{vec}\left((\mathbf{V}_1, \ldots, \mathbf{V}_p)'\right)$. Specifically, the (i,j)-th element of $\mathbf{V}_r$, $\mathbf{V}_r^{ij}$, denotes the variance of the (i,j)-th element of the VAR coefficient matrix $\mathbf{B}_r$, $r = 1, \ldots, p$, i.e.

$$\mathbf{V}_r^{ij} = \begin{cases} \frac{\pi_1^2}{r^2} & \text{for coefficients on own lag } r \text{ for } r = 1, \ldots, p, \\ \frac{\pi_1^2 \pi_2 \sigma_j}{r^2 \sigma_i} & \text{for coefficients on lag } r \text{ of variable } j \neq i, \text{ for } r = 1, \ldots, p, \end{cases} \tag{3.2}$$

where π_1 and π_2 are hyperparameters, and σ_l is the standard deviation from an AR(4) model for the variable l, $l = 1, \ldots, n$. The hyperparameter π_1 controls the overall tightness of the marginal distributions around zero. It also governs the relative importance of the prior compared to information contained in the data. The choice of this hyperparameter consequently has a large effect on the overall degree of parameter shrinkage (Chan et al., 2018). In a similar vein, π_2 governs the relative importance of cross-lag coefficients. If $\pi_2 = 1$, then both types of lags are a priori equally important. Conversely, setting $\pi_2 < 1$ implies that cross-lags are less important than own-lags, and vice-versa. Next, the term $1/r^2$ is the rate at which prior variance decreases with the lag length. This incorporates the idea that recent lags are more important than distant ones. Following Koop et al. (2010) we set $r = 2$.

Instead of choosing values for the hyperparameters, we specify two hyperpriors. Specifically, we set Uniform priors of the form $\pi_1 \sim \mathcal{U}(1/k_{\boldsymbol{\beta}}, 1)$ and $\pi_2 \sim \mathcal{U}(0.5, 1)$. The boundaries are chosen to cover conventional parameter ranges that are used in the literature, while the distribution ensures that we are uninformative about which choice is best. The trade-off with this extra flexibility is that the posterior distributions for π_1 and π_2 are non-standard. We consequently sample each of them with a random walk Metropolis-Hastings step.

3.2 SSVS Prior

The SSVS prior was developed by George et al. (2008) as an early hierarchical shrinkage prior that aims to endogenously determine the degree of coefficient shrinkage. The main idea is to partition the coefficients into two groups. The first group of coefficients are strongly regularized towards zero, while those in the second group are not. More formally, each of the coefficients are assumed to follow a two-component Gaussian mixture of the form

$$\beta_j|\pi_j \sim \pi_j\mathcal{N}(0, c_j v_j) + (1-\pi_j)\mathcal{N}(0, v_j), \tag{3.3}$$

where

$$\begin{aligned}\mathbb{P}(\pi_j = 1) &= q_j,\\ \mathbb{P}(\pi_j = 0) &= 1 - q_j,\end{aligned} \tag{3.4}$$

in which c_j is a scale parameter that we set to be 0.001, which implies that the variance of the Gaussian distribution obtained with $\pi_j = 1$ is strictly less than the one with $\pi_j = 0$. The mixture probability q_j, and the variance v_j are assumed to be the parameters to be estimated. Since q_j is a probability, we specify an uninformative prior $q_j \sim \mathcal{U}(0,1)$. Finally, we set $v_j \sim \mathcal{IG}(1,1)$. This distribution ensures that the support of v_j is the positive segment of the real line, while the shape and scale parameters are chosen to set both the prior mean and variance to one.

3.3 Global-Local Hierarchical Shrinkage Priors

In addition to the two conventional priors, we consider three new GL priors: Dirichlet-Laplace, Horseshoe and Normal-Gamma. First introduced by Polson and Scott (2010), these priors can be viewed as setting

$$\beta_j \sim \mathcal{N}(0, \tau^2\psi_j^2), \tag{3.5}$$

$$\tau^2 \sim \mathcal{X}_\tau, \tag{3.6}$$

$$\psi_j^2 \sim \mathcal{X}_\psi, \tag{3.7}$$

where τ^2 is the *global variance component*, each ψ_j^2 is a *local variance component* and both $\mathcal{X}_\tau$ and $\mathcal{X}_\psi$ denote arbitrary distributions. By specifying the global and local variance components separately, GL priors can shrink the BVAR coefficients towards zero, while allowing the data to speak through the maintenance of idiosyncratic tail behavior in the marginal distributions (after integrating out τ^2). Moreover, the Dirichlet-Laplace, Horseshoe and Normal-Gamma priors can be viewed as GL shrinkage priors that differ based on the choices of $\mathcal{X}_\tau$ and $\mathcal{X}_\psi$. We now consider each of these specifications in turn.

3.3.1 Dirichlet-Laplace Prior

The Dirichlet-Laplace (DL) prior was first introduced by Bhattacharya et al. (2015) as a regularization prior for large linear regressions, and was recently extended to the BVAR framework in Kastner and Huber (2017). The DL prior can be represented as the following mixture

$$\beta_j \sim \mathcal{N}(0, \psi_j \tau^2 \phi_j^2), \tag{3.8}$$

$$\psi_j \sim \mathcal{L}(\frac{1}{2}), \tag{3.9}$$

$$\tau \sim \mathcal{G}(\alpha_\beta k_\beta, \frac{1}{2}), \tag{3.10}$$

$$\boldsymbol{\phi} \sim \mathcal{D}(\alpha_\beta, \ldots, \alpha_\beta), \tag{3.11}$$

where $\mathcal{L}(\cdot)$, $\mathcal{G}(\cdot)$ and $\mathcal{D}(\cdot, \ldots, \cdot)$ respectively denote the Laplace, Gamma and Dirichlet distributions, ψ_j are *local shrinkage parameters*, and the Dirichlet-Laplace prior on the convolution $\tau^2\phi_j^2$ replaces the single *global shrinkage parameter* in (3.6) by a vector of scales $\left(\tau\phi_1, \ldots, \tau\phi_{k_{\boldsymbol{\beta}}}\right)$, in which $\boldsymbol{\phi} = \left(\phi_1, \ldots, \phi_{k_\beta}\right)$ is bounded to the $(k_\beta - 1)$ simplex $\mathcal{S} = \{\boldsymbol{\phi} : \phi_j \geq 0, \sum_{j=1}^{k_{\boldsymbol{\beta}}} \phi j = 1\}$.

Aside from having the advantages pertaining to GL priors that we discussed previously, an additional advantage of the DL prior over the traditional Minnesota and SSVS priors is that it requires less a priori input from the researcher. Specifically, the above equations show that the researcher only has to specify a single hyperparameter: α_β, compared to three in the Minnesota and SSVS priors. Bhattacharya et al. (2015) show that varying α_β has two effects. First, smaller values of α_β generate a higher degree of shrinkage within the coefficients and vice versa. Intuition for this can be obtained through the equations of the global shrinkage convolution. Note that lower values of

α_β imply that greater prior probability mass is placed on small values of τ through (3.10) and ϕ_j through (3.11). Both of these effects reduce the variance in (3.8) and thus push β_j towards zero. Second, since it translates into lower values of τ, smaller values of α_β also generate thicker tails on the marginal prior of β_j (after integrating out ϕ_j). Taken together, this implies that lowering α_β has the effect of increasing the amount of global parameter shrinkage, while allowing for thicker tails in the (local) marginal distributions. This latter feature is important as variables with strong predictive power will be prevented from being incorrectly zeroed out.

This raises an important question: how should researchers choose α_β? In the regression context, Bhattacharya et al. (2015) show that setting $\alpha_\beta = k_\beta^{-(1+\delta)}$, for small $\delta > 0$, results in an optimal *minimax rate* of posterior contraction. In practice however, they find that such values may result in the over shrinkage of relatively weak yet significant signals and generate numerical errors when k_β is large. To avoid this issue, they suggest softening the mass at zero by adopting a default value of $\alpha_\beta = 0.5$. While in some sense arbitrary, they show that this choice leads to similar in-sample fit relative to $\alpha_\beta = k_\beta^{-1}$ in over more than 100 fictitious samples. This lead Kastner and Huber (2017) to specify $\alpha_\beta = k_\beta^{-1}$ in their main BVAR analysis and $\alpha_\beta = 0.5$ as robustness. Here we take a more agnostic approach and specify a third layer uniform hyperprior over the range $[k_\beta^{-1}, 0.5]$. This approach has the advantage of allowing the data to select the value of α_β and thus avoids the possibility that our selection impacts the results.

It is straight forward to show that the posteriors of ψ_j and τ are

$$\psi_j|\beta_j, \phi_j, \tau \sim \mathcal{IG}\left(\frac{\phi_j \tau}{|\beta_j|}, 1\right), \quad \text{for } j = 1, \ldots, k, \tag{3.12}$$

and

$$\tau|\beta_j, \phi_j, \psi_j \sim \mathcal{GIG}\left(k_\beta(\alpha_\beta - 1), 1, 2\sum_{j=1}^{K} \frac{|\beta_j|}{\phi_j}\right), \tag{3.13}$$

where $\mathcal{GIG}(\cdot, \cdot, \cdot)$ denotes the generalized Inverse-Gaussian distribution. Since one at a time updates of a Dirichlet vector are computationally intensive, we follow Bhattacharya et al. (2015, p.1482) and first obtain a draw from

$$R_j|\beta_j \sim \mathcal{GIG}\left(\alpha_\beta - 1, 1, 2|\beta_j|\right), \quad \text{for } j = 1, \ldots, k. \tag{3.14}$$

and then set

$$\phi_j = \frac{R_j}{\sum_{j=1}^{k} R_j}. \tag{3.15}$$

Draws from the $\mathcal{GIG}$ distribution can be obtained with the algorithm in Devroye (2014). Finally, since α_β is of one dimensional, it can be shown that the log of the full conditional distribution for α_β is given

$$\log p(\alpha_\beta|\tau, \boldsymbol{\phi}) = -k_\beta \log \Gamma(\alpha_\beta) + \alpha_\beta \left(k_\beta \log \tau - k_\beta \log 2 + \sum_{j=1}^{k_\beta} \phi_j \right) + c$$

where c is a constant independent of α_β. The posterior draws of α_β can be obtained by using a Griddy-Gibbs algorithm on the interval $[k_\beta^{-1}, 0.5]$.[3]

3.3.2 Horseshoe Prior

The Horseshoe (HS) prior was first introduced by Carvalho et al. (2009, 2010) in the machine learning literature as a method for parameter shrinkage in the presence of sparsity, and was recently adapted to the BVAR framework in Follett and Yu (2017). The general HS prior takes the form

$$\beta_j|\lambda_j^\beta, \tau^\beta \sim \mathcal{N}(0, \lambda_j^\beta \tau^\beta), \tag{3.16}$$

$$\lambda_j^\beta \sim \mathcal{C}^+(0, 1), \tag{3.17}$$

$$\tau^\beta \sim \mathcal{C}^+(0, 1), \tag{3.18}$$

where $\mathcal{C}^+(\cdot, \cdot)$ denotes the half-Cauchy distribution, λ_j^β is the local shrinkage parameter and τ^β is the global shrinkage parameter. Carvalho et al. (2009, 2010) show that specifying half-Cauchy distributions over the global and local hyperparameters allows for optimal rates of shrinkage near zero, while having sufficiently thick tails to preserve strong signals. Another interesting feature of the HS prior is that it has no hyperparameters. It therefore avoids the aforementioned issues relating to user specification of these parameters.

[3] Details of the Griddy-Gibbs algorithm can be found in Ritter and Tanner (1992).

Despite its theoretical advantages, the HS prior results in non-standard conditional posterior distributions for the VAR coefficients, making standard Gibbs sampling infeasible. Rather than resorting to a Metropolis-within-Gibbs sampler as in Follett and Yu (2017), we exploit the scale mixture representation of the half-Cauchy distribution to get known conditional posterior distributions for each of the coefficients (Makalic and Schmidt, 2016).[4] In particular, we consider the following inverse-Gamma representation of HS prior

$$\lambda_j^\beta | \nu_j^\beta \sim \mathcal{IG}(\frac{1}{2}, \frac{1}{\nu_j^\beta}), \tag{3.19}$$

$$\tau^\beta | \xi_\beta \sim \mathcal{IG}(\frac{1}{2}, \frac{1}{\xi_\beta}), \tag{3.20}$$

$$\nu_j^\beta \sim \mathcal{IG}(\frac{1}{2}, 1), \tag{3.21}$$

$$\xi_\beta \sim \mathcal{IG}(\frac{1}{2}, 1), \tag{3.22}$$

where $\left(\nu_1, \ldots, \nu_{k\boldsymbol{\beta}}\right)$ and ξ are independent auxiliary random variables. Note that this hierarchical representation retains the advantage of not having to specify any hyperparameters.

Since the Gaussian and inverse-Gamma distributions are conjugate distributions, it is straight forward to show that the posteriors of these parameters are

$$\lambda_j^\beta | \beta_j, \tau^\beta, \nu_j, \xi_\beta \sim \mathcal{IG}(1, \frac{1}{\nu_j^\beta} + \frac{\beta_j^2}{2\tau^\beta}), \tag{3.23}$$

$$\tau^\beta | \beta_j, \lambda_j^\beta, \nu_j, \xi_\beta \sim \mathcal{IG}(\frac{k+1}{2}, \frac{1}{\xi_\beta} + \frac{1}{2}\sum_{j=1}^{k} \frac{\beta_j^2}{\lambda_j^\beta}), \tag{3.24}$$

$$\nu_j^\beta | \beta_j, \lambda_j^\beta, \tau^\beta, \xi_\beta \sim \mathcal{IG}(1, 1 + \frac{1}{\lambda_j^\beta}), \tag{3.25}$$

$$\xi_\beta | \beta_j, \lambda_j^\beta, \tau^\beta, \nu_j \sim \mathcal{IG}(1, 1 + \frac{1}{\tau^\beta}). \tag{3.26}$$

Finally, since all of the conditional posteriors are inverse-Gamma distributions, we can sample from them directly without the need for a Metropolis-Hastings step.

[4] The scalar mixture representation stems from the fact that if X and ω are random variables such that $X^2|\omega \sim \mathcal{IG}\left(\frac{1}{2}, \frac{1}{\omega}\right)$ and $\omega \sim \mathcal{IG}\left(\frac{1}{2}, \frac{1}{a^2}\right)$, where a is a constant, then $X \sim \mathcal{C}^+(0, a)$.

3.3.3 Normal-Gamma Prior

The Normal-Gamma (NG) prior was first proposed by Griffin et al. (2010) in the regression context, and recently extended to a BVAR in Huber and Feldkircher (2017). The general form is

$$\beta_j|\widetilde{\psi}_j, \lambda^2 \sim \mathcal{N}(0, \frac{2}{\lambda^2}\widetilde{\psi}_j), \tag{3.27}$$

$$\widetilde{\psi}_j \sim \mathcal{G}(\vartheta_\psi^\beta, \vartheta_\psi^\beta), \tag{3.28}$$

$$\lambda^2 \sim \mathcal{G}(c_{0,}, c_1), \tag{3.29}$$

where $\widetilde{\psi}_j$ is the local shrinkage parameter and λ^2 is the global shrinkage parameter. Griffin et al. (2010) show that the variance of the marginal density of β_j, after integrating out $\widetilde{\psi}_j$, is negatively related to the global shrinkage parameter, and that the hyperparameter of the local shrinkage parameter: ϑ_ψ^β, controls the degree of kurtosis. For instance, decreasing ϑ_ψ^β places more probability mass close to zero, but at the same time generates thicker tails. Thus, a tight prior given by a large value of λ^2 and small values of ϑ_ψ^β will avoid the incorrect zeroing out of local signals. In light of this fact, we set $c_0 = k_{\boldsymbol{\beta}}^{-1}$ and $c_1 = 1$. This implies that the prior mean in (3.29) is $k_{\boldsymbol{\beta}}^{-1}$ and the degree of global shrinkage on the VAR coefficients is able to vary with the model size.

As for ϑ_ψ^β, we follow Huber and Feldkircher (2017) and impose an exponentially distributed prior centered on unity, i.e $\vartheta_\psi^\beta \sim \mathcal{E}(1)$. The motivation for this hyperprior stems from the fact that setting $\vartheta_\psi = 1$ results in the double exponential (Laplace) prior in Park and Casella (2008).[5] In fact, the original motivation for the NG prior more broadly stemmed from addressing a key weakness of the double exponential prior. Specifically, the full posterior distribution does not contract at the same speed as its mode (Castillo et al., 2015). The consequence of this result is that the full posterior distribution does not provide an accurate measure of estimation uncertainty, a central idea of Bayesian inference and (density) forecasting. This issue is overcome by introducing the local shrinkage parameter $\widetilde{\psi}_j$. In short, this variable ensures that the resulting posterior distribution converges at the same rate as the mode. It consequently provides superior uncertainty quantification compared to the double-exponential prior while retaining its regularization properties.

[5]Since the maximum a posteriori probability (MAP) estimate of this distribution corresponds to the frequentest least absolute shrinkage and selection operator (LASSO), this prior is also commonly called the Bayesian LASSO.

To derive the posteriors it is first useful to note that (3.27) - (3.29) can be equivalently written as

$$\beta_j|\psi_j \sim \mathcal{N}(0, \psi_j), \tag{3.30}$$

$$\psi_j|\lambda^2, \vartheta_\psi^\beta \sim \mathcal{G}(\vartheta_\psi^\beta, \frac{\vartheta_\psi^\beta \lambda^2}{2}), \tag{3.31}$$

$$\lambda^2 \sim \mathcal{G}(c_{0,}, c_1), \tag{3.32}$$

where $\psi_j = \frac{2}{\lambda^2}\widetilde{\psi}_j$. Given this representation it can be easily shown that the (conditional) posteriors are

$$\psi_j|\lambda^2, \vartheta_\psi^\beta \sim \mathcal{GIG}\left(\vartheta_\psi^\beta - \frac{1}{2}, \vartheta_\psi^\beta \lambda^2, \beta_j^2\right), \tag{3.33}$$

$$\lambda^2|\vartheta_\psi^\beta \sim \mathcal{G}\left(c_0 + \vartheta_\psi^\beta k_{\boldsymbol{\beta}}, c_1 + \frac{\vartheta_\psi^\beta}{2}\sum_{j=1}^{k_{\boldsymbol{\beta}}} \psi_j\right). \tag{3.34}$$

The conditional posterior of ϑ_ψ^β is non-standard and thus requires a Metropolis-Hastings step. Following Huber and Feldkircher (2017), we use a random walk sampler in which candidate draws, denoted $\log(\vartheta_\psi^{\beta c})$, is obtained from $\mathcal{N}\left(\ln\left(\vartheta_\psi^\beta\right), \omega^2\right)$, where ω^2 is a tuning parameter and we set $\omega^2 = 0.01^2$ in our empirical study. The acceptance probability for $\vartheta_\psi^{\beta c}$ is given by

$$\min\left\{1, \left(\frac{\vartheta_\psi^{\beta c}}{\vartheta_\psi^\beta}\right)\left(\frac{\Gamma(\vartheta_\psi^\beta)}{\Gamma(\vartheta_\psi^{\beta c})}\right)^{k_{\boldsymbol{\beta}}} \frac{\left(\vartheta_\psi^{\beta c}\lambda^2/2\right)^{k_{\boldsymbol{\beta}}\vartheta_\psi^{\beta c}}}{\left(\vartheta_\psi^\beta \lambda^2/2\right)^{k_{\boldsymbol{\beta}}\vartheta_\psi^\beta}}\left(\prod_{i=1}^{k_{\boldsymbol{\beta}}}\psi_i\right)^{\vartheta_\psi^{\beta c}-\vartheta_\psi^\beta} \exp\left((\vartheta_\psi^\beta - \vartheta_\psi^{\beta c})(\frac{\lambda^2}{2}\sum_{i=1}^{k_{\boldsymbol{\beta}}}\psi_i + 1)\right)\right\}, \tag{3.35}$$

where $\Gamma(\cdot)$ denotes the Gamma function.

4 Forecasting

4.1 Data

Following Giannone et al. (2015); Huber and Feldkircher (2017) and Kastner and Huber (2017), we forecast three commonly used US macroeconomic indicators: real GDP growth, GDP Deflator based inflation and the Federal Funds Rate (FFR), using both small ($n = 3$) and large ($n = 21$)

BVAR-SV models. As in Huber and Feldkircher (2017) and Kastner and Huber (2017) the data is a variant of the well-known Stock and Watson (2011) dataset for the US economy, provided by McCracken and Ng (2016) and maintained by the Federal Reserve Bank of St. Louis. The sample period is from 1960Q1 to 2018Q2. Since this includes the entire duration of the zero lower bound period of interest rates in the US economy, we replace the conventional FFR with the shadow rate provided by Wu and Xia (2016). Following McCracken and Ng (2016) we transform the data to be approximately stationary. A complete list of the variables (in the order that they appear in the BVAR-SV model) and associated transformations is provided in Table 1. Finally, following convention we standardize the stationary series prior to estimation (Follett and Yu, 2017; Huber and Feldkircher, 2017; Kastner and Huber, 2017).

Table 1: Description of variables used in the forecasting exercise.

Name	FRED mnemonic	Transformation
Real Gross Domestic Product	GDPC1	1
Gross Domestic Product: Chain-type Price Index	GDPCTPI	1
Federal Funds Rate (Shadow rate)	FEDFUNDS	2
Real Personal Consumption Expenditures	PCECC96	1
Real Gross Private Domestic Investment	GPDIC1	1
Real private fixed investment: Residential	PRFIx	1
Industrial Production Index	INDPRO	1
Capacity Utilization: Manufacturing	CUMFNS	1
All Employees: Service-Providing Industries	SRVPRD	1
Civilian Employment	CE16OV	1
Average Weekly Hours of Production and Nonsupervisory Employees: Manufacturing	AWHMAN	2
Personal Consumption Expenditures	PCECTPI	2
Gross Private Domestic Investment: Chain-type Price Index	GPDICTPI	1
Consumer Price Index for All Urban Consumers: All Items	CPIAUCSL	1
Real Average Hourly Earnings of Production and Nonsupervisory Employees: Construction	CES2000000008x	1
1-Year Treasury Constant Maturity Rate	GS1	2
10-Year Treasury Constant Maturity Rate	GS10	2
Real M2 Money Stock	M2REALx	1
U.S. / U.K. Foreign Exchange Rate	EXUSUKx	1
University of Michigan: Consumer Sentiment	UMCSENTx	2
S&Ps Common Stock Price Index: Composite	S&P 500	1

Note: Transformation codes are as follows: 1 - log differences, 2 - raw data. Variables are listed in the order that they appear in the BVAR-SV model.

4.2 Forecast Method and Metrics

We conduct a pseudo out-of-sample forecasting exercise in which we compare both point and density forecasts provided by the five BVAR-SV models presented in Section 3.3. In each exercise, we divide the data into three sub-samples. The first part is the *initialization period*, which consists

the first $p = 4$ observations, i.e. 1960Q1-1960Q4. The second part is the *estimation period*, which consists of the eighteen years of quarterly data, i.e. 1961Q1-1978Q4. The third part is the *evaluation period*, which contains the remaining observations that are used to assess the forecast performance of the model, i.e. 1979Q1-2018Q2.

To demonstrate how the forecasts are conducted, let $\mathbf{y}_{1:t}$ denote the data from the *estimation period* and $\hat{\mathbf{y}}_{t+k}$ represent the vector of k-steps-ahead forecasts with k=1,2 and 4. Density forecasts are obtained by the predictive density: $f\left(\mathbf{y}_{t+k}|\mathbf{y}_{1:t}\right)$, and point forecasts are taken to be the mean of this density: $\hat{\mathbf{y}}_{t+k} = \mathbb{E}\left[\mathbf{y}_{t+k}|\mathbf{y}_{1:t}\right]$. The forecasts are then created with predictive simulation. For concreteness, suppose we want to produce a 4-step ahead forecast of inflation from 2010Q1. Then, given the MCMC draws up to 2010Q1 along with the relevant transition equations, we simulate the future states up to 2010Q4. The conditional expectation of this equation is then taken to be the point forecast of $\mathbf{y}_{2011Q1}$, and the observed data is used to evaluate implied density to produce a density forecast. The exercise is then repeated using data up to the end of the evaluation period.

Point forecast accuracy is assessed with two commonly used metrics: the average mean absolute forecast error (MAFE) and the average root mean square forecast error (RMSFE). These metrics are defined as

$$\text{MAFE} = \frac{1}{T - T_0 - k + 1} \sum_{t=1}^{T-T_0-k+1} |\mathbf{y}_{T_0+t+k-1} - \hat{\mathbf{y}}_{T_0+t+k-1}|, \tag{4.1}$$

$$\text{RMSFE} = \frac{1}{T - T_0 - k + 1} \sum_{t=1}^{T-T_0-k+1} \sqrt{(\mathbf{y}_{T_0+t+k-1} - \hat{\mathbf{y}}_{T_0+t+k-1})^2}, \tag{4.2}$$

where $\hat{\mathbf{y}}_t = \mathbb{E}(\mathbf{y}_t|\mathbf{y}_{1:t-1})$. Since a smaller forecast error corresponds to both a smaller MSFE and MAFE, for interpretation purposes relatively smaller values indicate superior forecast performance.

Density forecast accuracy is assessed with two commonly used metrics: the average log predictive likelihood (LPL) and the average continuous rank probability score (CRPS). These metrics are defined as

$$\text{LPL} = \frac{1}{T - T_0 - k + 1} \sum_{t=1}^{T-T_0-k+1} \log p(\hat{\mathbf{y}}_{T_0+t+k-1} = \mathbf{y}_{T_0+t+k-1}|\mathbf{y}_{1:T_0+t}), \tag{4.3}$$

$$\text{CRPS} = \frac{1}{T - T_0 - k + 1} \sum_{t=1}^{T-T_0-k+1} |\mathbf{y}_{T_0+t+k-1} - \hat{\mathbf{y}}_{A,T_0+t+k-1}| - \frac{1}{2}|\hat{\mathbf{y}}_{A,T_0+t+k-1} - \hat{\mathbf{y}}_{B,T_0+t+k-1}|, \tag{4.4}$$

where $\hat{\mathbf{y}}_{A,t}$ and $\hat{\mathbf{y}}_{B,t}$ are independently drawn from the joint distribution of $\mathbf{y}_t$. For interpretation purposes, note that a larger LPL value implies a better density forecast performance, while a smaller CRPS value implies better performance.

The choice of these metrics is twofold. First, the RMSFE and LPL have been widely used in existing studies and thus allows us to directly compare results. Second, the CRPS generalizes the MAFE to the case of probabilistic forecasts. These two metrics therefore provide a direct way of comparing point and density forecasts in a consistent manner (Gneiting and Raftery, 2007).

4.3 Forecasting Results

Before discussing the forecast results, we briefly assess the in-sample fit of the priors over the evaluation-period. For that purpose, we provide the joint log predictive likelihood (JLPL) across all three variables of interest in Table 2. Theoretically, the JLPL is as an approximation to the marginal likelihood in which we have conditioned on the initial estimation period, and therefore provides a direct measure of in-sample fit (Geweke, 2001). The results show that the hierarchical Minnesota prior provides the best fit compared to all other priors in both small and large models. One exception is the Horseshoe prior which performs similarly in the large dataset. In contrast to Huber and Feldkircher (2017), we find that the hierarchical Minnesota prior outperforms the Normal-Gamma prior. Our result therefore highlights the importance of adopting a hierarchical framework as specified in Section 2 as compared to their simpler specification. Consistent with existing studies, we also find that the large BVAR specification tends to provide better in-sample fit compared to the smaller BVAR (Bańbura et al., 2010; Koop et al., 2010; Huber and Feldkircher, 2017; Chan, 2018). Two exceptions are the Dirichlet-Laplace and SSVS priors, which do worse as the model size increases. While the former result is in line with existing studies (e.g. Koop et al. (2010)), the latter result has not been documented elsewhere in the literature.

Table 2: Joint log predictive likelihood (JLPL) for small ($n = 3$) and large $n = 21$ BVAR-SV models.

Small	
BVARSV-DL	-0.85
BVARSV-HS	-0.86
BVARSV-M	-0.78
BVARSV-NG	-0.86
BVARSV-SSVS	-0.86
Large	
BVARSV-DL	-0.92
BVARSV-HS	-0.78
BVARSV-M	-0.78
BVARSV-NG	-0.83
BVARSV-SSVS	-0.99

Having shown that the hierarchical Minnesota prior provides competitive in-sample fit, we now compare the out-of-sample forecast performance of the priors. The point forecast results are presented in Tables 3 and 4. Two facts immediately stand out. First, both metrics agree that the large model with Minnesota prior provides the best forecasts of inflation and that the small model with Minnesota prior provides the best interest rate forecasts. The second result worth noting is that despite the relatively good forecast performance of the Minnesota prior, the forecast accuracy is extremely close across all priors. In fact, in many cases the Minnesota prior only improves upon the alternatives at the second decimal place, resulting in a non-statistically significant difference in many cases. One exception is the MAFE inflation for which the Minnesota prior often provides statistically significant improvements upon the GL priors at short horizons. We defer a possible explanation for this to the end of the section.

Table 3: RMSFE of small ($n = 3$) and large ($n = 21$) models for real GDP growth, GDP Deflator measured inflation and the (shadow) Federal Funds Rate.

	GDP Growth			**GDP Deflator**			**Federal Funds Rate**		
Small	k = 1	k = 2	k = 4	k = 1	k = 2	k = 4	k = 1	k = 2	k = 4
BVARSV-DL	0.87	0.89	0.95	0.39	0.98	0.61	0.22	0.35	0.49
BVARSV-HS	0.86	0.88	0.94	0.39	0.49	0.61	0.22	0.34	0.48
BVARSV-M	0.81	0.81	0.86	0.36	0.41	0.52	**0.22**	**0.34**	**0.47**
BVARSV-NG	0.86	0.88	0.95	0.39	0.49	0.61	0.22	0.34	0.48
BVARSV-SSVS	0.87*	0.90*	0.96	0.39	0.49	0.61	0.22	0.34	0.48
Large	k = 1	k = 2	k = 4	k = 1	k = 2	k = 4	k = 1	k = 2	k = 4
BVARSV-DL	0.77	0.86	0.93	0.39	0.47	0.60	0.23	0.37	0.55
BVARSV-HS	0.75	0.82	0.85	0.38	0.41	0.53	0.23	0.34	0.49
BVARSV-M	0.74	0.81	0.85	**0.34**	**0.40**	**0.50**	0.22	0.35	0.49
BVARSV-NG	**0.73**	**0.81**	**0.84**	0.36	0.42	0.53	0.23	0.34	0.49
BVARSV-SSVS	0.80	0.87	0.95	0.40*	0.48*	0.62	0.23	0.38	0.56

Note: Best model is indicated in Bold. The superscripts ***, ** and * indicate rejection of equal forecast accuracy relative to the BVAR-SV with hierarchical Minnesota prior at significance level 0.01, 0.05 and 0.1, respectively, when using an asymptotic test in Diebold and Mariano (1995).

Table 4: MAFE of small ($n = 3$) and large ($n = 21$) models for real GDP growth, GDP Deflator measured inflation and the (shadow) Federal Funds Rate.

	GDP Growth			**GDP Deflator**			**Federal Funds Rate**		
Small	k = 1	k = 2	k = 4	k = 1	k = 2	k = 4	k = 1	k = 2	k = 4
BVARSV-DL	0.61	0.63	0.65*	0.29**	0.34**	0.40	0.11	0.21	0.35**
BVARSV-HS	0.60	0.62	0.64*	0.29**	0.34**	0.41	**0.11**	0.21	0.34
BVARSV-M	0.58	**0.57**	0.61	0.27	0.31	0.38	0.11	**0.21**	**0.34**
BVARSV-NG	0.61	0.62	0.65*	0.29*	0.34*	0.40	0.11	0.21	0.35**
BVARSV-SSVS	0.62	0.63	0.65*	0.30**	0.34**	0.40	0.11	0.21	0.34*
Large	k = 1	k = 2	k = 4	k = 1	k = 2	k = 4	k = 1	k = 2	k = 4
BVARSV-DL	0.58	0.64	0.67*	0.29**	0.35**	0.40	0.12	0.23	0.38
BVARSV-HS	0.56	0.58	**0.59**	0.29	0.32**	0.39	0.11	0.21**	0.34
BVARSV-M	**0.55**	0.58	0.61	**0.27**	**0.30**	**0.37**	0.12	0.22	0.34
BVARSV-NG	0.55	0.59	0.61	0.28	0.33	0.39	0.13	0.22	0.34
BVARSV-SSVS	0.60*	0.65	0.69**	0.30**	0.35**	0.42*	0.13	0.25	0.39

Note: Best model is indicated in Bold. The superscripts ***, ** and * indicate rejection of equal forecast accuracy relative to the BVAR-SV with hierarchical Minnesota prior at significance level 0.01, 0.05 and 0.1, respectively, when using an asymptotic test in Diebold and Mariano (1995).

The associated density forecast results are presented across Tables 5 and 6. Consistent with the point forecast results, both metrics agree that the large BVAR with Minnesota prior provides the best inflation forecasts when compared to the alternative models. When using the small dataset,

the metrics agree that the Minnesota prior often provides statistically significant improvements over the remaining priors in the small models. Interestingly, when using the large model, the LPL suggests that the Normal-Gamma and Horseshoe priors provide superior accuracy for real GDP growth and the FFR respectively, however these improvements are not statistically significant at the 10% level or lower. Nonetheless, the result that the Normal-Gamma prior provides good one-step-ahead point and density forecasts for real GDP (as measured by the RMSFE and LPL) is in line with Huber and Feldkircher (2017). In contrast, the CRPS suggests that the Minnesota prior provides superior forecasts to the alternatives, and these improvements are often statistically significant. This result is consistent with the point forecast evaluations provided by the MAFE. As discussed previously, one advantage of the CRPS is that it reduces to the mean absolute error (MAE) if the forecast is deterministic. In practice, this makes it possible to compare an ensemble forecast with a deterministic forecast of the same variable in a consistent fashion. Hence, our results show the importance of using a range of metrics when evaluating density forecasts in particular.

Table 5: LPL of small ($n = 3$) and large ($n = 21$) models for real GDP growth, GDP Deflator measured inflation and the (shadow) Federal Funds Rate.

	GDP Growth			**GDP Deflator**			**Federal Funds Rate**		
Small	k = 1	k = 2	k = 4	k = 1	k = 2	k = 4	k = 1	k = 2	k = 4
BVARSV-DL	−1.16*	−1.21**	−1.29*	−0.36	−0.51	−0.68	0.65	−0.02	−0.65
BVARSV-HS	−1.16**	−1.20**	−1.28*	−0.36	−0.51	−0.68	0.64	−0.01	−0.63
BVARSV-M	−1.12	−1.15	−1.22	−0.33	−0.46	−0.65	0.64	−0.01	−0.61
BVARSV-NG	−1.16**	−1.20**	−1.28*	−0.36	−0.51	−0.67	0.64	−0.01	−0.63
BVARSV-SSVS	−1.16**	−1.21**	−1.30*	−0.370	−0.51	−0.67	0.64	−0.01	−0.63*
Large	k = 1	k = 2	k = 4	k = 1	k = 2	k = 4	k = 1	k = 2	k = 4
BVARSV-DL	−1.10	−1.20	−1.31	−0.40	−0.54	−0.73	0.57	−0.03	−0.58
BVARSV-HS	−1.06	−1.14	−1.22	−0.38	−0.48	−0.69*	**0.65**	**0.02**	−0.58
BVARSV-M	−1.06	−1.14	**−1.22**	**−0.29**	**−0.41**	**−0.60**	0.57	−0.04	**-0.58**
BVARSV-NG	**−1.05**	**−1.14**	−1.22	−0.34	−0.47	−0.66	0.56	−0.04	−0.59*
BVARSV-SSVS	−1.11	−1.21	−1.31	−0.41	−0.55**	−0.70	0.53	−0.08	−0.60

Note: Best model is indicated in Bold. The superscripts ***, ** and * indicate rejection of equal forecast accuracy relative to the BVAR-SV with hierarchical Minnesota prior at significance level 0.01, 0.05 and 0.1, respectively, when using an asymptotic test in Diebold and Mariano (1995).

Table 6: CRPS of small ($n = 3$) and large ($n = 21$) models for real GDP growth, GDP Deflator measured inflation and the (shadow) Federal Funds Rate.

	GDP Growth			**GDP Deflator**			**Federal Funds Rate**		
Small	k = 1	k = 2	k = 4	k = 1	k = 2	k = 4	k = 1	k = 2	k = 4
BVARSV-DL	0.47*	0.48**	0.51*	0.22**	0.26**	0.31	0.09*	0.17*	0.27**
BVARSV-HS	0.46	0.47*	0.50*	0.21*	0.26*	0.31	0.09	0.16	0.27*
BVARSV-M	0.44	**0.44**	0.47	0.20	0.23	0.28	**0.09**	**0.16**	**0.26**
BVARSV-NG	0.47*	0.48*	0.50*	0.22*	0.26*	0.31	0.09*	0.17*	0.27*
BVARSV-SSVS	0.47*	0.48*	0.51*	0.22*	0.26*	0.31	0.09**	0.17*	0.27*
Large	k = 1	k = 2	k = 4	k = 1	k = 2	k = 4	k = 1	k = 2	k = 4
BVARSV-DL	0.51**	0.55***	0.59*	0.26***	0.31***	0.38*	0.18	0.21	0.34
BVARSV-HS	0.42	0.45	0.47	0.22**	0.25**	0.31**	0.10	0.17	0.27
BVARSV-M	**0.41**	0.45	**0.47**	**0.20**	**0.23**	**0.28**	0.11	0.18	0.28
BVARSV-NG	0.42	0.46	0.47	0.22**	0.26**	0.31	0.11	0.18	0.29
BVARSV-SSVS	0.52**	0.55*	0.59*	0.26***	0,31**	0,38*	0.12**	0.22	0.35

Note: Best model is indicated in Bold. The superscripts ***, ** and * indicate rejection of equal forecast accuracy relative to the BVAR-SV with hierarchical Minnesota prior at significance level 0.01, 0.05 and 0.1, respectively, when using an asymptotic test in Diebold and Mariano (1995).

Our forecasting exercise also corroborates some results that have been previously documented elsewhere in the literature. For instance, in the case of real GDP and inflation, the forecasting performance of most prior specifications increases with the model size, implying that adding information aids in forecast accuracy (Bańbura et al., 2010; Koop and Korobilis, 2013; Huber and Feldkircher, 2017). In contrast, our result that the small BVAR often outperforms the large specification when forecasting interest rates reflects the idea that the Federal Reserve largely focuses on inflation and output growth when making their interest rate decisions. Finally, consistent with Koop et al. (2010), we find that the performance of the SSVS prior often deteriorates with model size. This is especially evident when generating density forecasts.

Taken together, our results suggests that the hierarchical Minnesota prior provides a solid practical choice when forecasting macroeconomic variables. This is important because it suggests that the documented concerns of this prior in sparse settings do not lead to subpar forecast accuracy, at least for macroeconomic forecasting. In light of existing optimality results for GL priors in sparse settings (van der Pas et al., 2016), one possible explanation for this result is that macroeconomic data is not sparse, but instead dense. In fact, this idea is in line with a recent paper by Giannone et al. (2018), who comment on the *illusion of sparsity* in economic and financial modeling. We investigate this possibility in the next section.

5 Simulation Study

To test the validity of the claim that the Minnesota prior can perform at least as well as the GL priors in dense environments, we compare each priors ability detect non-zero signals in both sparse and dense environments. The dense DGP allows for a large set of explanatory variables to be important for prediction, although their individual impact is small. In contrast, the sparse DGP features a smaller set of important explanatory variables, each with high predictive power.

Our simulation study is based on twenty fictitious samples of stationary data: ten dense and ten sparse, each with $T = 300$ observations from a large ($n = 20$) homoscedastic VAR. The sample size and number of variables are chosen to reflect our forecasting exercise. In each case we set the number of lags to $p = 2$. Consistent with conventional economic ideas, we distinguish between the importance of own-lags, cross-lags and intercepts. In the first instance, the own-lag coefficients in each equation are assumed to have a moderate degree of persistence that decays as the lag length increases. In particular, we set

$$b_{1,ii} \sim \mathcal{N}\left(0, \sigma_o^2\right), \quad i = 1, \ldots, n \tag{5.1}$$

$$b_{r,ii} = \frac{b_{1,ii}}{2}, \quad r = 1, \ldots, p \tag{5.2}$$

where $b_{r,ij}$ is the (i, j) entry of the coefficient matrix $\mathbf{B}_r$ and $\mathcal{N}(\cdot, \cdot)$ is a univariate Gaussian distribution. Equation (5.2) incorporates the notion that recent lags contain more information than distant ones. Next, the cross-lags are produced via the specification

$$b_{r,ij} = \begin{cases} \mathcal{N}(0, \sigma_c^2) & \text{with probability } p_c, \quad r = 1, \ldots, p, \quad i \neq j \\ 0 & \text{with probability } 1 - p_c, \end{cases} \tag{5.3}$$

where $p_c \in (0, 1)$ is the probability of a non-zero coefficient. Finally, for simplicity we set each of the intercepts equal to zero. This choice is without loss of generality since the intercept in any model can be forced to zero by simply subtracting the mean of the data. To ensure the resulting simulated data are stationary we discard any non-stationary draws.

Note that the inclusion probability on the cross-lags p_c controls the number of non-zero coefficients, while the variance terms σ_o^2 and σ_c^2 control the possible strength of the signals. We can therefore

set these values to simulate sparse and dense DGPs. In the sparse setting, we set the inclusion probability equal to 0.01, the variance of the own lags to 0.6 and the variance of cross-lags to 0.3. This ensures that, on average, one percent of the cross-product terms are not equal to zero, while the resulting non-zero terms have relatively strong signals. In the dense setting, we set the inclusion probability equal to 0.8 and the variance of the own- and cross-lags to 0.2 and 0.15 respectively. This ensures that, on average, 80 percent of the cross-product terms are not equal to zero, while the resulting non-zero terms have relatively weak signals. In both cases, setting $\sigma_0^2 > \sigma_c^2$ allows for the idea that own-lags are more important than cross-lags.

Next, we similarly generate elements in the vector of contemporaneous coefficients $\boldsymbol{\gamma}$ according to the rule

$$\gamma_i = \begin{cases} \mathcal{N}(0, \sigma_\gamma) & \text{with probability } p_\gamma, \quad i = 1, \ldots, k_{\boldsymbol{\gamma}} \\ 0 & \text{with probability } 1 - p_\gamma, \end{cases} \tag{5.4}$$

where $p_\gamma \in (0, 1)$ is the probability of a non-zero element. Consistent with the BVAR coefficients, in practice we set $p_\gamma = 0.01$ and $\sigma_\gamma = 0.3$ in the sparse setting and $p_\gamma = 0.8$ and $\sigma_\gamma = 0.15$ in the dense environment. Finally, for simplicity we assume that each equation in the BVAR has an independent and identically distributed Gaussian white noise disturbance term with unit variance.

To measure the models ability to detect the non-zero signals in the sparse and dense DGPs, we compare the *mean absolute deviation* (MAD) of the posterior mean induced by each prior against the true parameters in the underlying data generating process (DGP). The MAD is defined as

$$\text{MAD}^{(r,m)} = \frac{1}{k_{\boldsymbol{\theta}}} \sum_{i=1}^{n} \sum_{j=1}^{k_{\boldsymbol{\theta},i}} |\theta_{i,j}^{(r)} - \hat{\theta}_{i,j}^{(r,m)}|, \tag{5.5}$$

where r is the MC simulation, m is the model associated with a given prior, $\theta_{i,j}$ is the true value of the parameter used to generate the DGP and $\hat{\theta}_{i,j}$ is the posterior mean.

5.1 Results for Simulation Study

Figures 1 and 2 respectively contain boxplots of the MAD statistics for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ across each of the sparse and dense DGPs. In each figure the box represents second and third quartiles, while the vertical line inside indicates the median value. In line with the theoretical optimality results proved elsewhere in the literature (e.g. van der Pas et al. (2016)), we find that the GL priors dominate the Minnesota prior when recovering the true parameters from the sparse DGP. This is true for both the lagged and contemporaneous coefficients. In line with our conjecture however, we find that this result does not extend to the case of dense environments. Instead, in that case the Minnesota prior provides competitive results. While this analysis by no means provides a formal proof of our claim, it is at least suggestive of the notion that the Minnesota prior is able to provide competitive forecasts compared to the GL priors because commonly used macroeconomic data sets are not sparse, but instead dense. Since such a process requires the development of sophisticated statistical tools, we leave this open as an important avenue of future research.

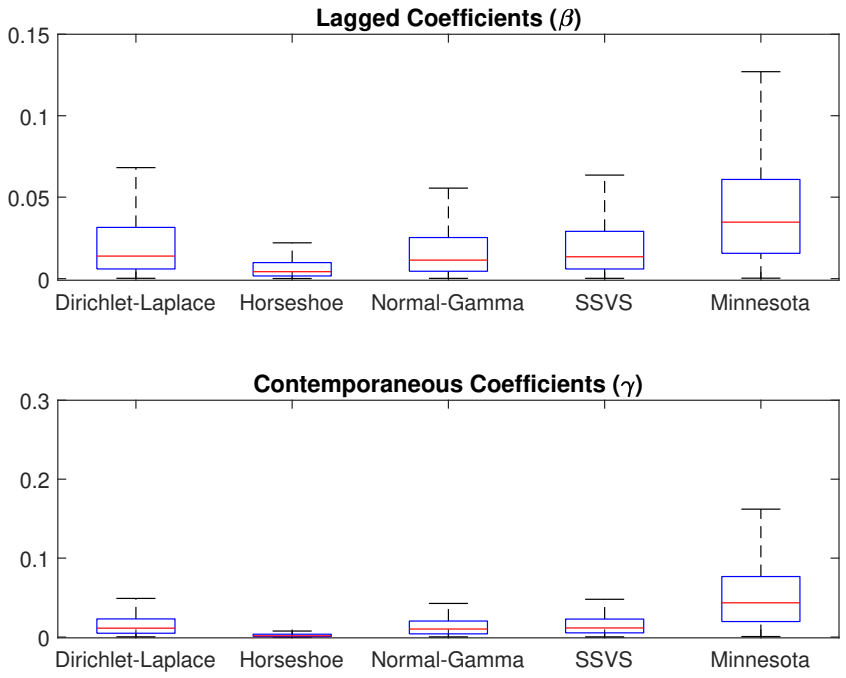

Figure 1: Box plots of the Mean Absolute Deviation (MAD) statistics over for the BVAR coefficients from a 20 variable homoscedastic BVAR over 10 sparse samples.

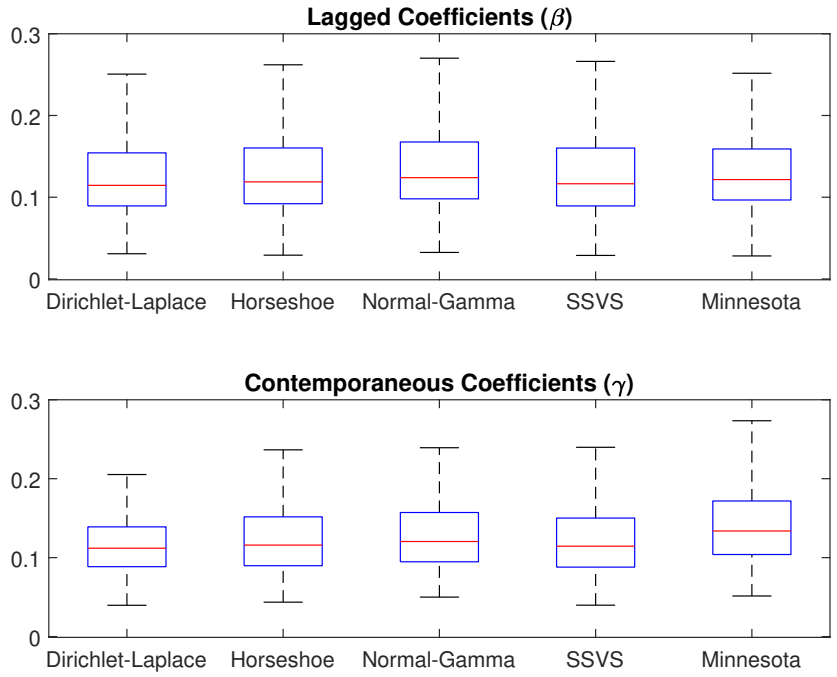

Figure 2: Box plots of the Mean Absolute Deviation (MAD) statistics over for the contemporaneous coefficients from a 20 variable homoscedastic BVAR over 10 dense samples.

6 Additional Analysis

In this section, we present various robustness checks and some additional analysis of the posterior draws and computation times for estimating the BVAR-SV model with each prior. To conserve space, we present the tables containing results from the robustness exercises in the online appendix.

6.1 Robustness

6.1.1 Sub-Sample Analysis

The four-decade-long forecast evaluation period used in the main analysis contains periods which have been categorized by different degrees of volatility and structural change. For instance, during the mid-1980s to early 2000s, many key US macroeconomic variables, including real GDP exhibited relatively low volatility, a period that Stock and Watson (2002b) termed the Great Moderation.

While the presence of stochastic volatility in the BVAR controls for volatility changes, the existence of structural breaks may violate the constant parameter assumption, and therefore distort the relative forecast results. As a robustness check, we repeat both the in-sample and out-of-sample forecasting exercise across decades: 1979-1989, 1990-1999, 2000-2010, and 2010-2018.

The results in Section 1 of the online appendix show that the GL priors often outperform the Minnesota prior in the 2000s; however, the improvements are not statistically significant based on the Diebold-Mariano test. Moreover, the Minnesota prior performs well in the other three decades. Nonetheless, since the 2000s contained the Great Recession, a period which was notoriously unpredictable with conventional models, it raises the idea of utilizing GL priors to estimate more sophisticated BVARs with time-varying coefficients. Since this is a non-trivial extension, we leave this as an important avenue for future research.[6]

6.1.2 Order Effects

In the forecasting exercise, the three key macroeconomic variables of interest were ordered first in the BVAR-SV. Since $\mathbf{A}_0$ in (2.1) is lower triangular, the variable order in the BVAR-SV model will influence the posterior distribution and may consequently affect the relative forecasting performance of the models. To test this hypothesis, we re-estimate the models with these variables ordered last, rather than first. For completeness, in addition to the full sample robustness check, we also present results for each of the sub-samples considered in the previous robustness check.

The results in Section 2 of the online appendix show the relative forecast performance of the GL priors to the Minnesota prior is robust over the full sample. In particular, no one prior regularly outperforms the Minnesota prior. Thus, the main conclusion of the paper holds. Interestingly, however, we again find that the GL priors often beat the Minnesota prior in the 2000s; however, this improvement is seldom significant according to the Diebold-Mariano test. Moreover, we also find that the Minnesota prior performs well in the other three decades.

[6]This topic has been recently undertaken in a working paper by Huber et al. (2019). Using a similar data set to the one employed in this paper, they find that a time-varying coefficient BVAR-SV model with GL priors can provide more accurate forecasts relative to the time-invariant model. Interestingly, however, they do not consider the hierarchical Minnesota prior used in this paper.

6.1.3 Non-Hierarchical Minnesota Prior

As a final robustness check, it is interesting also to compare the relative performance of a simpler non-hierarchical Minnesota prior to the hierarchical Minnesota prior used in the main paper. To that end, we specify conventional user choice of the two hyperparameters, i.e., $\pi_1 = 0.2$ and $\pi_2 = 0.9$, and re-estimate the models across the full sample and each of the sub-samples mentioned above.

In line with Giannone et al. (2015), the results in Section 3 of the online appendix show that the hierarchical structure improves forecast performance. This result holds for both the full sample and each of the sub-samples. The main conclusion of the paper is consequently unchanged if we also consider this simpler model.

6.2 Posterior Distributions

A key lesson from our results is that it is important to make use of hierarchical priors that avoid the well-known pitfalls of user-specified hyperparameters. It is, therefore, interesting to examine the posterior distributions of the key hyperparameters and compare them to conventional user choices. To this end, we present histograms of the posterior distribution of each hyperparameter in both large and small models. In the exercise, we use a loop of 20,000 draws and data over the full sample period. Since the Horseshoe and SSVS priors do not have hyperparameters we do not provide any plots for those priors.

The posterior draws for the two hyperparameters in the Minnesota prior are provided in Figure 4. User choices for these parameters vary in the literature but are typically close to values of $\pi_1 = 0.2$ and $\pi_2 = 0.9$. The posterior mode for π_1 is around 0.11 in the large model and 0.3 in the small one. The relative size of these numbers reflects the idea that a greater degree of regularization is preferred as the model size increases. Interestingly, the value of π_2 is very close to zero in both models (0.015 in the large and less than 0.001 in the small). This suggests that the data is in favor of much more regularization of the cross-lags compared to the conventional choice of $\pi_2 = 0.9$.

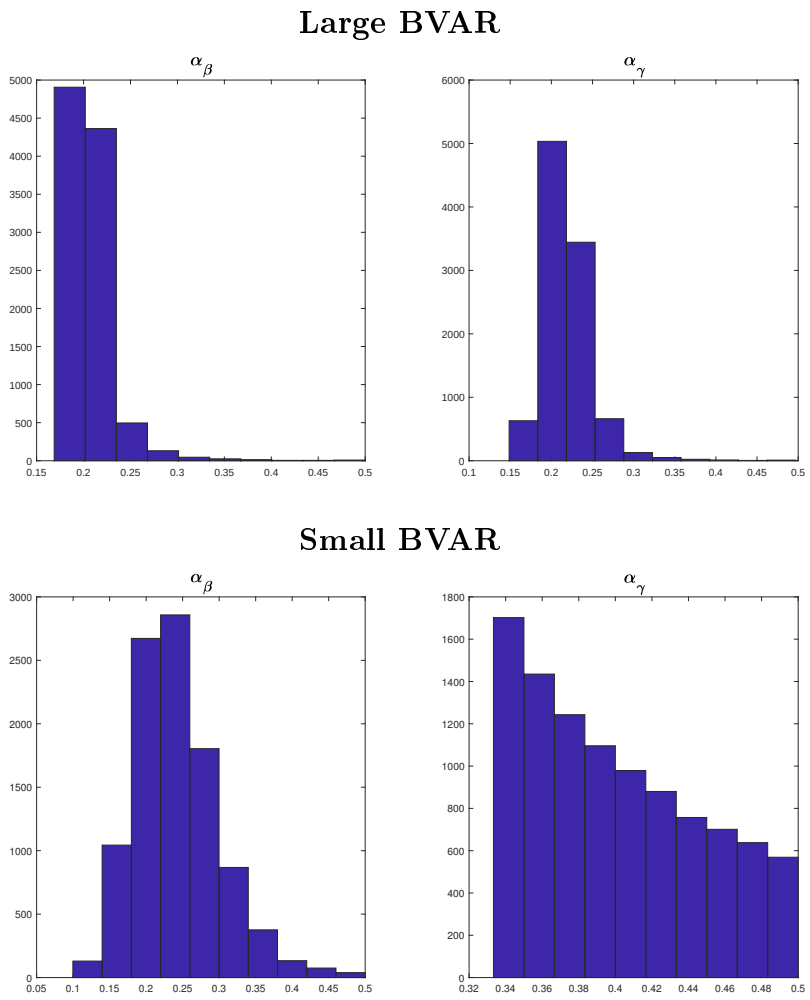

Figure 3: Posterior distributions of the hyperparameters: α_β (left column) and α_γ (right column), in the Dirichlet-Laplace prior for the large (top row) and small (bottom row) BVARs.

Next, the posterior draws of the sole hyperparameter in the Dirichlet-Laplace prior are shown in Figure 3. The posterior mode tends to be very different from conventional user choices. Recall from Section 3.3.1 that the two common choices of the sole hyperparameter are $\alpha_\beta = k_\beta^{-1}$ and 0.5. The results from Figure 3 show that the preferred value, according to the data, is somewhere in between these two values. For instance, in the case of α_β, the posterior mode is around 0.2 for the large model and 0.25 in the small one. This suggests that setting $\alpha_\beta = k_\beta^{-1}$ would result in over-shrinking the BVAR coefficients towards zero, while setting $\alpha_\beta = 0.5$ would result in too little regularization. A similar case can be made for α_γ[7]. We also note that in either case, the posterior

[7]Note that we impose two independent DL priors for the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. We use α_γ to denote the hyperparameter of the DL prior for the contemporaneous coefficients $\boldsymbol{\gamma}$.

mode of the key tuning parameter gets closer to zero as the size of the BVAR-SVs gets larger. This supports the idea that a greater degree of regularization is preferred as the model size increases.

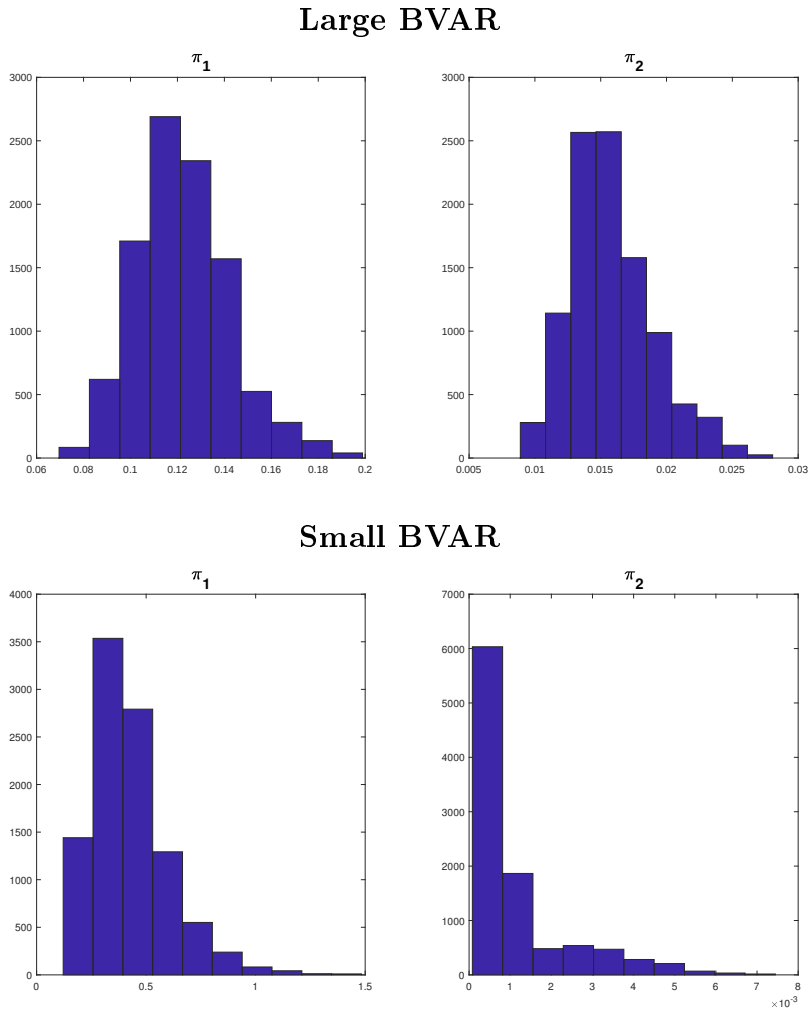

Figure 4: Posterior distributions of the hyperparameters: π_1 (left column) and π_2 (right column), in the Minnesota prior for the large (top row) and small (bottom row) BVARs.

Finally, the posterior draws of the key hyperparameters of the Normal-Gamma prior (ϑ_ψ^β and ϑ_ψ^γ)[8] are provided in Figure 5. While there are no conventional values of this parameter (previous studies have always used a hierarchical structure), it is still interesting to determine whether the posterior draws are close to one, as such a result would suggest that the Normal-Gamma prior reduces to the double exponential prior (see the discussion in Section 3.3.3). It turns out that the values are

[8]As we assume an independent NG prior on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, we use ϑ_ψ^γ to denote the hyperparameter of the NG prior for the contemporaneous coefficients $\boldsymbol{\gamma}$.

much smaller than unity. For instance, the posterior mode for $\vartheta_{\psi}^{\gamma}$ is around 0.16 in the large model and 0.3 in the small one. This suggests that the Normal-Gamma prior is inducing substantially more regularization than a simpler double exponential prior in both small and large models.

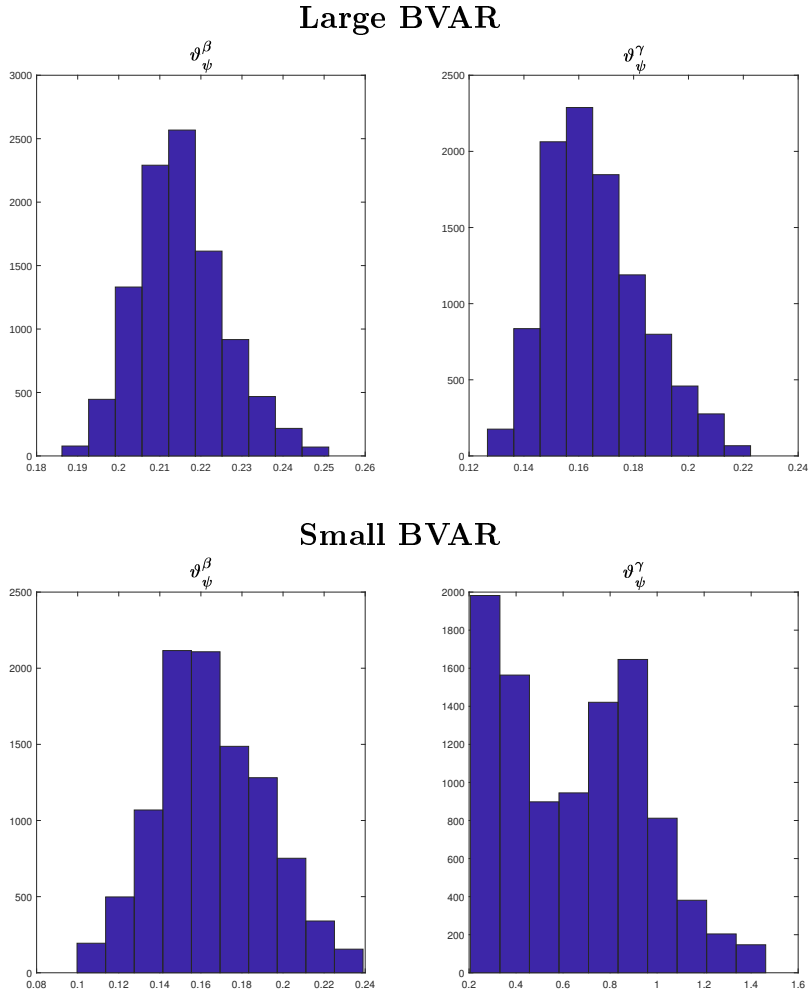

Figure 5: Posterior distributions of the hyperparameters: ϑ_{ψ}^{β} (left column) and $\vartheta_{\psi}^{\gamma}$ (right column), in the Normal-Gamma prior for the large (top row) and small (bottom row) BVARs.

6.3 Computation Times

To assess the additional computational cost of estimating the hyperparameters in the hierarchical structure, Table 7 reports the computation times (in seconds) to obtain 20,000 posterior draws from a BVAR-SV model with four lags under the NG, DL and Minnesota priors[9]. The estimation

[9]Note that the Horseshoe and SSVS have no estimated hyperparameters and consequently no non-hierarchical structure.

period is the full sample, and the models are estimated using Matlab on a desktop with an Intel Core i7-7700 @3.60 GHz processor and 16GB memory. Our results show that the additional computational cost for estimating the hyperparameters is not overly demanding. For instance, the greatest increase in computational time when estimating the hierarchical model compared to its non-hierarchical counterpart is around one minute, i.e., 290.8 seconds to 349.3 seconds.

Table 7: The computation times (in seconds) to obtain 20,000 posterior draws under each prior on the full sample. All BVARs have p = 4 lags.

	$n = 3$	$n = 10$	$n = 21$
Benchmark Models			
BVARSV-NG	24.1	91.7	295.5
BVARSV-DL	25.8	107.7	349.3
BVARSV-M	23.0	73.6	210.5
Non-hierarchical structure			
BVARSV-NG	23.1	86.2	279.6
BVARSV-DL	24.0	90.2	290.8
BVARSV-M	20.4	66.6	200.4

7 Conclusion

We have questioned whether a class of recently proposed global-local hierarchical priors can improve upon conventional benchmarks when forecasting macroeconomic variables. Using US data with a focus on real GDP growth, inflation, and the Federal Funds Rate over the period 1960Q1 to 2018Q2, the results suggested that the answer is no. Despite its simplicity, a hierarchical Minnesota prior provided competitive in- and out-of-sample fit of all variables. This finding indicates that the Minnesota prior remains a solid practical choice when forecasting macroeconomic variables.

In light of existing optimality results in sparse environments, a possible explanation for this result is that macroeconomic data is not sparse, but instead dense. Indeed, the results from a simple simulation study, revealed while global-local priors are superior in recovering the true model parameters in a sparse environment, the Minnesota prior is competitive in dense environments. Of course, such an exercise is not a formal proof that commonly used macroeconomic data sets are dense;

however, it is suggestive of such a phenomenon. Since such a process requires the development of sophisticated statistical tools, we leave this as an important avenue of future research.

References

Bańbura, Marta, Domenico Giannone, and Lucrezia Reichlin, "Large Bayesian vector auto regressions," *Journal of Applied Econometrics*, 2010, *25* (1), 71–92.

Bhattacharya, Anirban, Debdeep Pati, Natesh S Pillai, and David B Dunson, "Dirichlet–Laplace priors for optimal shrinkage," *Journal of the American Statistical Association*, 2015, *110* (512), 1479–1490.

Carriero, Andrea, Todd E Clark, and Massimiliano Marcellino, "Common drifting volatility in large Bayesian VARs," *Journal of Business & Economic Statistics*, 2016, *34* (3), 375–390.

_ , _ , and _ , "Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors," *Journal of Econometrics*, 2019.

Carvalho, Carlos M, Nicholas G Polson, and James G Scott, "Handling sparsity via the horseshoe," in "Artificial Intelligence and Statistics" 2009, pp. 73–80.

_ , _ , and _ , "The horseshoe estimator for sparse signals," *Biometrika*, 2010, *97* (2), 465–480.

Castillo, Ismaël, Johannes Schmidt-Hieber, Aad Van der Vaart et al., "Bayesian linear regression with sparse priors," *The Annals of Statistics*, 2015, *43* (5), 1986–2018.

Chan, Joshua CC, "Large Bayesian VARs: A flexible Kronecker error covariance structure," *Journal of Business & Economic Statistics*, 2018, pp. 1–12.

Chan, Joshua C.C. and Cody Y.L. Hsiao, *Estimation of Stochastic Volatility Models with Heavy Tails and Serial Dependence*, John Wiley and Sons Inc.,

Chan, Joshua CC and Eric Eisenstat, "Bayesian model comparison for time-varying parameter VARs with stochastic volatility," *Journal of Applied Econometrics*, 2018.

_ **and Ivan Jeliazkov**, "Efficient simulation and integrated likelihood estimation in state space models," *International Journal of Mathematical Modelling and Numerical Optimisation*, 2009, *1* (1-2), 101–120.

Chan, Joshua, Liana Jacobi, and Dan Zhu, "How sensitive are VAR forecasts to prior hyperparameters? An automated sensitivity analysis," *To be published in Advances in Econometrics*, 2018.

Clark, Todd E and Francesco Ravazzolo, "Macroeconomic forecasting performance under alternative specifications of time-varying volatility," *Journal of Applied Econometrics*, 2015, *30* (4), 551–575.

Cogley, Timothy and Thomas J Sargent, "Drifts and volatilities: monetary policies and outcomes in the post WWII US," *Review of Economic dynamics*, 2005, *8* (2), 262–302.

Devroye, Luc, "Random variate generation for the generalized inverse Gaussian distribution," *Statistics and Computing*, 2014, *24* (2), 239–246.

Follett, Lendie and Cindy Yu, "Achieving Parsimony in Bayesian VARs with the Horseshoe Prior," *arXiv preprint arXiv:1709.07524*, 2017.

George, Edward I, Dongchu Sun, and Shawn Ni, "Bayesian stochastic search for VAR model restrictions," *Journal of Econometrics*, 2008, *142* (1), 553–580.

Geweke, John, "Bayesian econometrics and forecasting," *Journal of Econometrics*, 2001, *100* (1), 11–15.

Giannone, Domenico, Michele Lenza, and Giorgio E Primiceri, "Prior selection for vector autoregressions," *Review of Economics and Statistics*, 2015, *97* (2), 436–451.

_ **,** _ **,** _ **, and** _ , "Economic predictions with big data: the illusion of sparsity," Technical Report, Federal Reserve Bank of New York 2018.

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

Griffin, Jim E, Philip J Brown et al., "Inference with normal-gamma prior distributions in regression problems," *Bayesian Analysis*, 2010, *5* (1), 171–188.

Huber, Florian and Martin Feldkircher, "Adaptive shrinkage in Bayesian vector autoregressive models," *Journal of Business & Economic Statistics*, 2017, pp. 1–13.

Huber, Florian, Gary Koop, and Luca Onorante, "Inducing Sparsity and Shrinkage in Time-Varying Parameter Models," *arXiv e-prints*, May 2019, p. arXiv:1905.10787.

Karlsson, Sune, *Handbook of Economic Forecasting*, Vol. 2, Elsevier,

Kastner, Gregor and Florian Huber, "Sparse Bayesian vector autoregressions in huge dimensions," *arXiv preprint arXiv:1704.03239*, 2017.

Kim, Sangjoon, Neil Shephard, and Siddhartha Chib, "Stochastic volatility: likelihood inference and comparison with ARCH models," *The Review of Economic Studies*, 1998, *65* (3), 361–393.

Koop, Gary and Dimitris Korobilis, "Large time-varying parameter VARs," *Journal of Econometrics*, 2013, *177* (2), 185–198.

_ , _ et al., "Bayesian multivariate time series methods for empirical macroeconomics," *Foundations and Trends® in Econometrics*, 2010, *3* (4), 267–358.

Koop, Gary M, "Forecasting with medium and large Bayesian VARs," *Journal of Applied Econometrics*, 2013, *28* (2), 177–203.

Litterman, Robert B, "Techniques of forecasting using vector autoregressions," Working paper 115, Federal Reserve Bank of Minneapolis 1979.

_ , "A Bayesian procedure for forecasting with vector autoregressions," Mimeo, Massachusetts Institute of Technology 1980.

Makalic, Enes and Daniel F Schmidt, "A simple sampler for the horseshoe estimator," *IEEE Signal Processing Letters*, 2016, *23* (1), 179–182.

McCracken, Michael W and Serena Ng, "FRED-MD: A monthly database for macroeconomic research," *Journal of Business & Economic Statistics*, 2016, *34* (4), 574–589.

Mol, Christine De, Domenico Giannone, and Lucrezia Reichlin, "Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?," *Journal of Econometrics*, 2008, *146* (2), 318–328.

Park, Trevor and George Casella, "The bayesian lasso," *Journal of the American Statistical Association*, 2008, *103* (482), 681–686.

Polson, Nicholas G and James G Scott, "Shrink globally, act locally: Sparse Bayesian regularization and prediction," *Bayesian statistics*, 2010, *9*, 501–538.

Ritter, Christian and Martin A Tanner, "Facilitating the Gibbs sampler: the Gibbs stopper and the griddy-Gibbs sampler," *Journal of the American Statistical Association*, 1992, *87* (419), 861–868.

Stock, James H and Mark W Watson, "Forecasting using principal components from a large number of predictors," *Journal of the American statistical association*, 2002, *97* (460), 1167–1179.

_ **and** _ , "Has the business cycle changed and why?," *NBER macroeconomics annual*, 2002, *17*, 159–218.

_ **and** _ , "Macroeconomic forecasting using diffusion indexes," *Journal of Business & Economic Statistics*, 2002, *20* (2), 147–162.

_ **and Mark Watson**, "Dynamic factor models," *Oxford handbook on economic forecasting*, 2011.

van der Pas, SL, J-B Salomond, Johannes Schmidt-Hieber et al., "Conditions for posterior contraction in the sparse normal means problem," *Electronic journal of statistics*, 2016, *10* (1), 976–1000.

Wu, Jing Cynthia and Fan Dora Xia, "Measuring the macroeconomic impact of monetary policy at the zero lower bound," *Journal of Money, Credit and Banking*, 2016, *48* (2-3), 253–291.