Minnesota-Type Adaptive Hierarchical Priors for Large Bayesian VARs

Joshua C.C. Chan[*]
Purdue University

August 2019

Abstract

Large Bayesian VARs with stochastic volatility are increasingly used in empirical macroeconomics. The key to make these highly parameterized VARs useful is the use of shrinkage priors. We develop a family of priors that captures the best features of two prominent classes of shrinkage priors: adaptive hierarchical priors and Minnesota priors. Like the adaptive hierarchical priors, these new priors ensure that only 'small' coefficients are strongly shrunk to zero, while 'large' coefficients remain intact. At the same time, these new priors can also incorporate many useful features of the Minnesota priors, such as cross-variable shrinkage and shrinking coefficients on higher lags more aggressively. We introduce a fast posterior sampler to estimate BVARs with this family of priors—for a BVAR with 25 variables and 4 lags, obtaining 10,000 posterior draws takes about 3 minutes on a standard desktop. In a forecasting exercise, we show that these new priors outperform both adaptive hierarchical priors and Minnesota priors.

Keywords: shrinkage prior, forecasting, stochastic volatility, structural VAR

JEL classifications: C11, C52, C55, E37

[*] I thank Gary Koop and Aubrey Poon for many constructive comments and useful suggestions that have improved a previous version of the paper.

1 Introduction

Vector autoregressions (VARs) are the main workhorse in empirical macroeconomics, and increasingly large Bayesian VARs are used after the influential work by Banbura, Giannone, and Reichlin (2010).[1] VARs tend to be highly parameterized, and the key to make these VARs useful is the introduction of shrinkage priors. The most prominent of these are the Minnesota prior (Doan, Litterman, and Sims, 1984; Litterman, 1986) and its modern variants (see, e.g., Kadiyala and Karlsson, 1993, 1997; Giannone, Lenza, and Primiceri, 2015). More recently, adaptive hierarchical shrinkage priors with good theoretical properties have been introduced to the large VAR settings. Examples include the normal-gamma prior in Huber and Feldkircher (2019) and the horseshoe prior in Follett and Yu (2019).[2] While the Minnesota prior has the undesirable property of shrinking all VAR coefficients, these adaptive hierarchical priors tend to leave 'large' coefficients intact and only shrink 'small' coefficients strongly to zero.

Despite this good theoretical property, empirically these adaptive hierarchical priors do not seem to forecast better than a variant of the Minnesota prior where the hyperparameters are selected based on the data. This is, for example, demonstrated in a recent forecasting exercise by Cross, Hou, and Poon (2019). One reason for this surprising result could be because under these adaptive hierarchical priors, all VAR coefficients are treated identically—e.g., a coefficient on the first lag has the same prior distribution as that of the fourth lag. In contrast, the Minnesota prior incorporates many plausible prior beliefs, such as cross-variable shrinkage—i.e., coefficients on lags of other variables are shrunk more aggressively than those of own lags—and the prior belief that variables of higher lags are less important.[3]

We introduce a class of priors that captures the best features of both the adaptive hierarchical priors and the Minnesota prior. Like the adaptive hierarchical priors, these new

[1] Important examples include Carriero, Kapetanios, and Marcellino (2009), Koop (2013), Koop and Korobilis (2013), Banbura, Giannone, Modugno, and Reichlin (2013), Carriero, Clark, and Marcellino (2015) and Carriero, Clark, and Marcellino (2016).

[2] For more examples of these adaptive hierarchical priors in the context of large Bayesian VARs, see Kastner and Huber (2018), Korobilis and Pettenuzzo (2019) and Gefang, Koop, and Poon (2019).

[3] There is empirical support for these prior beliefs. For example, Chan (2019a) finds cross-variable shrinkage to be critical in improving forecasting performance. Chan, Jacobi, and Zhu (2019) find evidence that shrinking coefficients associated with higher lags more strongly increases the value of the marginal likelihood.

priors have both heavy tails and substantial mass around zero. These features ensure that only 'small' coefficients are strongly shrunk to zero, while 'large' coefficients remain intact. At the same time, these new priors can also incorporate many useful features of the Minnesota prior, such as cross-variable shrinkage and shrinking coefficients on higher lags more aggressively.

To estimate large BVARs with this new family of priors, we introduce a reparameterization of the standard reduced-form BVAR with stochastic volatility. Specifically, we rewrite the BVAR in the structural form, where the time-varying error covariance matrices are diagonal. Hence, we can treat the structural BVAR as a system of n independent regressions, which substantially speeds up computations. This approach is similar to the equation-by-equation estimation approach in Carriero, Clark, and Marcellino (2019), which is designed for the reduced-form parameterization. Since under our parameterization there is no need to obtain the 'orthogonalized' shocks at each iteration as in Carriero, Clark, and Marcellino (2019), our approach is substantially faster. For example, for a BVAR with 25 variables and 4 lags, simulating 10,000 posterior draws using the proposed posterior sampler takes about 3 minutes on a standard desktop.

We illustrate the empirical relevance of the proposed Minnesota-type adaptive hierarchical priors with a forecasting exercise that involves 23 US quarterly macroeconomic and financial variables. More specifically, we consider a Minnesota-type normal-gamma prior, and show that this new prior outperforms two important benchmarks: 1) a standard normal-gamma prior that treats all coefficients identically; and 2) a Minnesota prior where the hyperparameters are estimated from the data. These results suggest that both the Minnesota prior and the normal-gamma prior have useful features, and combining them gives us the best of both worlds.

The rest of the paper is organized as follows. We first introduce in Section 2 a reparameterization of the reduced-form BVAR with stochastic volatility. We then outline various shrinkage priors for large BVARs, including the Minnesota prior and some recently introduced adaptive hierarchical priors. Then, Section 3 develops the new class of Minnesota-type adaptive hierarchical priors that combines the best features of popular priors. Section 4 describes an efficient posterior simulator to estimate the BVAR with the proposed Minnesota-type adaptive hierarchical priors. It is followed by a macroeconomic forecasting exercise to illustrate the usefulness of the proposed priors in Section 5. Lastly, Section 6 concludes and briefly discusses some future research directions.

2 Bayesian VARs and Shrinkage Priors

In this section we first provide some background on Bayesian VARs with stochastic volatility. We then outline various shrinkage priors for large BVARs, including the Minnesota prior and some recently introduced adaptive hierarchical priors.

Let $\mathbf{y}_t = (y_{1,t}, \dots, y_{n,t})'$ be an $n \times 1$ vector of endogenous variables at time t. A standard reduced-form VAR can be written as:

$$\mathbf{y}_t = \widetilde{\mathbf{b}} + \widetilde{\mathbf{B}}_1 \mathbf{y}_{t-1} + \cdots + \widetilde{\mathbf{B}}_p \mathbf{y}_{t-p} + \widetilde{\boldsymbol{\varepsilon}}_t^y, \quad \widetilde{\boldsymbol{\varepsilon}}_t^y \sim \mathcal{N}(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}_t), \tag{1}$$

where $\widetilde{\mathbf{b}}$ is an $n \times 1$ vector of intercepts and $\widetilde{\mathbf{B}}_1, \dots, \widetilde{\mathbf{B}}_p$ are $n \times n$ VAR coefficient matrices. Here we allow the covariance matrix $\widetilde{\boldsymbol{\Sigma}}_t$ to be time-varying, as a large empirical literature has demonstrated that this is an important feature for improving forecasting performance (Clark, 2011; D'Agostino, Gambetti, and Giannone, 2013; Cross and Poon, 2016; Chan, 2018).

More specifically, we follow Cogley and Sargent (2005) and Carriero, Clark, and Marcellino (2019) to decompose the inverse covariance matrix, or the precision matrix, as $\widetilde{\boldsymbol{\Sigma}}_t^{-1} = \mathbf{B}_0' \boldsymbol{\Sigma}_t^{-1} \mathbf{B}_0$, where $\boldsymbol{\Sigma}_t = \text{diag}(\mathrm{e}^{h_{1,t}}, \dots, \mathrm{e}^{h_{n,t}})$ is a diagonal matrix and $\mathbf{B}_0$ is a lower triangular matrix with ones on the main diagonal. Each log-volatility $h_{i,t}$ for $i = 1, \dots, n$ in turn follows an independent random walk process:

$$h_{i,t} = h_{i,t-1} + \varepsilon_{i,t}^h, \quad \varepsilon_{i,t}^h \sim \mathcal{N}(0, \sigma_{h,i}^2) \tag{2}$$

for $t = 1, \dots, T$, where the initial condition $h_{i,0}$ is treated as an unknown parameter to be estimated.

2.1 The Bayesian VAR in Structural Form

Next, we introduce a reparameterization of the reduced-form VAR in (1) that facilitates posterior simulation. In a nutshell, this reparameterization allows us to rewrite the VAR as n independent regressions, and it leads to a more efficient sampling scheme. Consequently, we are able to improve upon the pioneering equation-by-equation estimation approach proposed in Carriero, Clark, and Marcellino (2019). The relative gains in posterior simulation are illustrated in Section 4.

Now, left multiply the reduced-form VAR in (1) by $\mathbf{B}_0$ to obtain the following structural form:

$$\mathbf{B}_0\mathbf{y}_t = \mathbf{b} + \mathbf{B}_1\mathbf{y}_{t-1} + \cdots + \mathbf{B}_p\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}^y_t, \quad \boldsymbol{\varepsilon}^y_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t). \tag{3}$$

It is easy to see that we can recover the reduced-form parameters by setting $\widetilde{\mathbf{b}} = \mathbf{B}_0^{-1}\mathbf{b}$ and $\widetilde{\mathbf{B}}_j = \mathbf{B}_0^{-1}\mathbf{B}_j, j = 1, \ldots, p$. Since the covariance matrix $\boldsymbol{\Sigma}_t$ in this structural form is diagonal, we can estimate this recursive system equation by equation without loss of efficiency.

To write the structural VAR in (3) as n independent regressions, we first introduce some notations. Let b_i denote the i-th element of $\mathbf{b}$ and let $\mathbf{b}_{j,i}$ represent the i-th row of $\mathbf{B}_j$. Then, $\boldsymbol{\beta}_i = (b_i, \mathbf{b}_{1,i}, \ldots, \mathbf{b}_{p,i})'$ is the intercept and VAR coefficients for the i-th equation. Furthermore, let $\boldsymbol{\alpha}_i$ denote the free elements in the i-th row of the impact matrix $\mathbf{B}_0$. We then follow Chan and Eisenstat (2018) to write the i-th equation of the system in (3) as:

$$y_{i,t} = \widetilde{\mathbf{w}}_{i,t}\boldsymbol{\alpha}_i + \widetilde{\mathbf{x}}_t\boldsymbol{\beta}_i + \varepsilon^y_{i,t}, \quad \varepsilon^y_{i,t} \sim \mathcal{N}(0, \mathrm{e}^{h_{i,t}}),$$

where $\widetilde{\mathbf{w}}_{i,t} = (-y_{1,t}, \ldots, -y_{i-1,t})$ and $\widetilde{\mathbf{x}}_t = (1, \mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p})$. Here we have a recursive system in which $y_{i,t}$ depends on the contemporaneous variables $y_{1,t}, \ldots, y_{i-1,t}$. But since the system is recursive, the Jacobian of the change of variables from $\boldsymbol{\varepsilon}^y_t$ to $\mathbf{y}_t$ has unit determinant, and therefore the likelihood function has the usual Gaussian form.

Finally, let $\mathbf{x}_{i,t} = (\widetilde{\mathbf{w}}_{i,t}, \widetilde{\mathbf{x}}_t)$. We can further simplify the i-th equation as:

$$y_{i,t} = \mathbf{x}_{i,t}\boldsymbol{\theta}_i + \varepsilon^y_{i,t}, \quad \varepsilon^y_{i,t} \sim \mathcal{N}(0, \mathrm{e}^{h_{i,t}}), \tag{4}$$

where $\boldsymbol{\theta}_i = (\boldsymbol{\alpha}'_i, \boldsymbol{\beta}'_i)'$ is of dimension $k_i = np + i$. Hence, we have rewritten the structural VAR in (3) as a system of n independent regressions. This representation facilitates equation-by-equation estimation, as we will discuss in detail in Section 4. In addition, by stacking the elements of the impact matrix $\boldsymbol{\alpha}_i$ and the VAR coefficients $\boldsymbol{\beta}_i$, we can sample them together to improve efficiency.

2.2 The Minnesota Prior

In this section we outline a data-based Minnesota prior on the structural VAR in (4). We will then use this version of the Minnesota prior to construct the proposed adaptive hierarchical priors in Section 3. For a general discussion of the Minnesota prior, see, e.g., Koop and Korobilis (2010), Karlsson (2013) or Chan (2019b).

Early works on shrinkage priors for small and medium VARs were developed by Doan, Litterman, and Sims (1984) and Litterman (1986). This family of priors, and many variants developed later, have come to be collectively called the Minnesota priors. In the original version, the prior is placed on the reduced-form VAR coefficients. Sims and Zha (1998) later formulated a version for structural VARs, which we will follow here. More specifically, we assume that the VAR coefficients $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_n')'$ are *a priori* independent across equations, and each $\boldsymbol{\theta}_i$, for $i = 1, \ldots, n$, has a normal prior:

$$\boldsymbol{\theta}_i \sim \mathcal{N}(\mathbf{m}_i, \mathbf{V}_i). \tag{5}$$

For growth rates data, we set $\mathbf{m}_i = \mathbf{0}$ to shrink the VAR coefficients to zero. For level data, $\mathbf{m}_i$ is set to be zero as well except for the coefficient associated with the first own lag, which is set to be one. Next, for $\mathbf{V}_i$, we assume it to be diagonal with the k-th diagonal element $V_{i,kk}$ set to be:

$$V_{i,kk} = \begin{cases} \frac{\kappa_1}{l^2}, & \text{for the coefficient on the } l\text{-th lag of variable } i, \\ \frac{\kappa_2 s_i^2}{l^2 s_j^2}, & \text{for the coefficient on the } l\text{-th lag of variable } j, j \neq i, \\ \frac{\kappa_3 s_i^2}{s_j^2}, & \text{for the } j\text{-th element of } \boldsymbol{\alpha}_i, \\ \kappa_4 s_i^2, & \text{for the intercept,} \end{cases}$$

where s_r^2 denotes the sample variance of the residuals from an AR(4) model for the variable $r, r = 1, \ldots, n$.

The prior covariance matrix $\mathbf{V}_i$ depends on four hyperparameters, namely, $\kappa_1, \ldots, \kappa_4$, that control the degree of shrinkage for different types of coefficients. For simplicity, we set $\kappa_3 = 1$ and $\kappa_4 = 100$. These values imply moderate shrinkage for the coefficients on the contemporaneous variables (the same magnitude as the residual variance) and essentially no shrinkage for the intercepts.

The hyperparameter κ_1 controls the overall shrinkage strength for coefficients on own lags, whereas κ_2 controls those on lags of other variables. We treat them as unknown parameters to be estimated. This is motivated by a few recent papers, such as Carriero, Clark, and Marcellino (2015) and Giannone, Lenza, and Primiceri (2015), which show that by selecting hyperparameters that control the overall shrinkage strength in a data-based fashion, one can substantially improve forecast performance. Also note that here we allow κ_1 and κ_2 to be different, as one might expect that coefficients on lags of other variables would be on average smaller than those on own lags. In fact, Chan (2019a) finds empirical evidence in support of this so-called cross-variable shrinkage. Finally, in both cases coefficients on higher lags are shrunk more strongly to zero, at a rate of $1/l^2$ for $l = 1, \ldots, p$.

2.3 Global-Local Adaptive Shrinkage Priors

Despite its empirical success, the Minnesota prior has been recently criticized as not being sufficiently adaptive. Ideally, a shrinkage prior should shrink only 'small' coefficients to zero, while leaving 'large' coefficients intact. But the Minnesota prior—being a normal prior with very thin tails—substantially shrinks both types of coefficients. For example, Griffin and Brown (2010) show that under the normal prior, the posterior mean of a regression coefficient does not converge to the least squares estimate when the latter approaches infinite. In contrast, if the tails of the prior distribution are heavier than those of the normal, the posterior mean of a regression coefficient converges to the least squares estimate as the latter approaches infinite.

In view of this problem of the normal prior, Polson and Scott (2010) consider a class of scale mixtures of normals priors called the global-local adaptive shrinkage priors:

$$\begin{aligned}(\theta_{i,j} \,|\, \tau, \psi_{i,j}) &\sim \mathcal{N}(0, \tau\psi_{i,j}),\\ \psi_{i,j} &\sim F_\psi(\psi_{i,j}),\\ \tau &\sim F_\tau(\tau),\end{aligned}$$

for $i = 1, \ldots, n, j = 1, \ldots, k_i$. Each $\psi_{i,j}$ is a local variance component associated with the coefficient $\theta_{i,j}$, whereas τ is a global variance component that is common to all coefficients. By using different mixing distribution $F_\psi(\cdot)$, this framework includes a wide variety of distributions that have heavier tails than those of the normal. Prominent examples include the t prior (Geweke, 1993), the normal-gamma prior (Griffin and Brown, 2010) and the horseshoe prior (Carvalho, Polson, and Scott, 2010).

While these global-local priors have the desirable property of shrinking only ‘small’ coefficients strongly to zero, one key drawback is that they treat all coefficients identically. In the context of large BVARs, one might wish to apply cross-variable shrinkage, or to shrink coefficients on higher lags more aggressively to zero than those of the first lag. However, these prior beliefs cannot be implemented within the current framework.

3 Minnesota-Type Adaptive Hierarchical Priors

In this section we introduce a new class of adaptive hierarchical priors that combines the best features of the Minnesota prior and the global-local priors. Similar to the global-local priors, these new priors are scale mixtures of normals with heavy tails, which ensures good theoretical properties. At the same time, they can incorporate many useful features of the Minnesota prior, such as cross-variable shrinkage and shrinking coefficients on higher lags more aggressively.

To formulate this new family of priors, let $C_{i,j}$ be positive constants. Then, consider the following prior on $\theta_{i,j}$:

$$(\theta_{i,j} \,|\, \kappa_1, \kappa_2, \psi_{i,j}) \sim \mathcal{N}(m_{i,j}, \kappa_{i,j}\psi_{i,j}C_{i,j}), \tag{6}$$

where $\psi_{i,j} \sim F_\psi(\psi_{i,j})$ for some suitably chosen distribution $F_\psi(\cdot)$ as in the global-local priors, $\kappa_{i,j} = \kappa_1$ for coefficients on own lags, $\kappa_{i,j} = \kappa_2$ for coefficients on other lags, and $\kappa_{i,j} = 1$ otherwise.

The setup in (6) nests both the Minnesota prior specified in Section 2.2 and the global-local prior discussed in Section 2.3. To verify the former claim, for each $i = 1, \ldots, n$

define the constants $C_{i,j}$ associated with the VAR coefficients $\theta_{i,j}, j = 1, \ldots, k_i$ as follows:

$$C_{i,j} = \begin{cases} \frac{1}{l^2}, & \text{for the coefficient on the } l\text{-th lag of variable } i, \\ \frac{s_i^2}{l^2 s_j^2}, & \text{for the coefficient on the } l\text{-th lag of variable } j, j \neq i, \\ \frac{\kappa_3 s_i^2}{s_j^2}, & \text{for the } j\text{-th element of } \boldsymbol{\alpha}_i, \\ \kappa_4 s_i^2, & \text{for the intercept.} \end{cases} \tag{7}$$

Then, it is easy to see that if all the local variances are degenerated at 1, i.e., $\psi_{i,j} \equiv 1$, this new family of priors reduces to the Minnesota prior. Hence, (6) can be viewed as a generalization of the Minnesota prior by introducing a local variance component that allows the marginal prior distribution of $\theta_{i,j}$ to have heavier tails than those of the normal.

If instead we set $C_{i,j} \equiv 1$ and $\kappa_1 \equiv \kappa_2 \equiv \tau$, then this family of priors becomes the standard global-local priors. Hence, (6) can also be interpreted as a generalization of the global-local priors that allows for non-identical distributions. By introducing the constants $C_{i,j}$, we can incorporate richer prior beliefs on the VAR coefficients, such as those useful features of the Minnesota prior.

In what follows, we use the constants $C_{i,j}$ defined in (7) as the benchmark. In addition, we set all the $\psi_{i,j}$'s associated with the intercepts and $\boldsymbol{\alpha}_i$ to be one, although it is straightforward to treat them as random variables.

4 Bayesian Estimation

We now describe an efficient posterior simulator to estimate the structural VAR in (3) with the proposed Minnesota-type adaptive hierarchical prior given in (6). For concreteness, we focus on the Minnesota-type normal-gamma prior, as it seems to perform best in the context of forecasting using large BVARs. We emphasize that our framework can handle all global-local priors, including the horseshoe prior and the Dirichlet-Laplace prior (Bhattacharya, Pati, Pillai, and Dunson, 2015). Estimating BVARs with other global-local priors requires only minor modifications of the proposed sampler.

More specifically, we adapt the parameterization of the normal-gamma prior in Huber and Feldkircher (2019), and consider the following Minnesota-type normal-gamma prior:

$$(\theta_{i,j} \,|\, \kappa_1, \kappa_2, \psi_{i,j}) \sim \mathcal{N}(m_{i,j}, 2\kappa_{i,j}\psi_{i,j}C_{i,j}), \tag{8}$$

$$\psi_{i,j} \sim \mathcal{G}(\nu_\psi, \nu_\psi/2), \tag{9}$$

where $\mathcal{G}(a, b)$ denotes the gamma distribution with mean a/b, $C_{i,j}$ is given in (7), $\kappa_{i,j} = \kappa_1$ for coefficients on own lags, $\kappa_{i,j} = \kappa_2$ for coefficients on other lags, and $\kappa_{i,j} = 1$ otherwise. This parameterization of the normal-gamma prior includes the Bayesian Lasso (Park and Casella, 2008) as a special case with $\nu_\psi = 1$.

To complete the model specification, we assume gamma priors for the hyperparameters κ_1, κ_2 and ν_ψ: $\kappa_j \sim \mathcal{G}(c_{1,j}, c_{2,j}), j = 1, 2$ and $\nu_\psi \sim \mathcal{G}(d_1, d_2)$. We set $c_{1,1} = c_{1,2} = 1$, $c_{2,1} = 1/0.04$ and $c_{2,2} = 1/0.04^2$. These values imply that the prior means of κ_1 and κ_2 are 0.04 and 0.04^2 respectively, which are the fixed values used in Carriero, Clark, and Marcellino (2015) for κ_1 and κ_2. We set $d_1 = d_2 = 1$, implying prior mean of 1 for ν_ψ. Finally, for the parameters in the state equation of $h_{i,j}$, we assume $h_{i,0} \sim \mathcal{N}(a_{h,i}, V_{h,i})$ and $\sigma^2_{h,i} \sim \mathcal{IG}(\nu_{h,i}, S_{h,i}), i = 1, \ldots, n$.

Next, we derive the (conditional) likelihood function of the structural VAR given the latent variables. To that end, we stack $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,T})'$ and $\mathbf{h}_i = (h_{i,1}, \ldots, h_{i,T})'$ over $t = 1, \ldots, T$, and define $\mathbf{X}_i$ and $\boldsymbol{\varepsilon}^y_i$ similarly. Then, we rewrite (4) in matrix form:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\theta}_i + \boldsymbol{\varepsilon}^y_i, \quad \boldsymbol{\varepsilon}^y_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_{\mathbf{h}_i}),$$

where $\boldsymbol{\Omega}_{\mathbf{h}_i} = \mathrm{diag}(\mathrm{e}^{h_{i,1}}, \ldots, \mathrm{e}^{h_{i,T}})$. Finally, let $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_n)'$ and $\mathbf{h} = (\mathbf{h}'_1, \ldots, \mathbf{h}'_T)'$. Then, the likelihood function of the VAR in (3) is given by

$$p(\mathbf{y} \,|\, \boldsymbol{\theta}, \mathbf{h}) = \prod_{i=1}^{n} p(\mathbf{y}_i \,|\, \boldsymbol{\theta}_i, \mathbf{h}_i) = \prod_{i=1}^{n} (2\pi)^{-\frac{T}{2}} \mathrm{e}^{-\frac{1}{2}\mathbf{1}'_T\mathbf{h}_i - \frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta}_i)'\boldsymbol{\Omega}^{-1}_{\mathbf{h}_i}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta}_i)}, \tag{10}$$

where $\mathbf{1}_T$ is a $T \times 1$ column of ones.

4.1 Posterior Simulator

To introduce the posterior sampler, we first define a few terms. Let $\boldsymbol{\psi}_i$ denote the free elements of $\psi_{i,j}, j = 1, \ldots, k_i$ and stack $\boldsymbol{\psi} = (\boldsymbol{\psi}_1', \ldots, \boldsymbol{\psi}_n')'$. Further, let $\boldsymbol{\Sigma}_h = (\sigma_{h,1}^2, \ldots, \sigma_{h,n}^2)'$, $\mathbf{h}_0 = (h_{1,0}, \ldots, h_{n,0})'$ and $\boldsymbol{\kappa} = (\kappa_1, \kappa_2)'$. Given the above priors and the likelihood in (10), we can simulate from the joint posterior distribution using the following posterior sampler that sequentially samples from:

1. $p(\boldsymbol{\theta}_i \,|\, \mathbf{y}, \mathbf{h}, \boldsymbol{\psi}, \boldsymbol{\kappa}, \nu_\psi, \mathbf{h}_0, \boldsymbol{\Sigma}_h)$, $i = 1, \ldots, n$;

2. $p(\boldsymbol{\psi}_i \,|\, \mathbf{y}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\kappa}, \nu_\psi, \mathbf{h}_0, \boldsymbol{\Sigma}_h)$, $i = 1, \ldots, n$;

3. $p(\boldsymbol{\kappa} \,|\, \mathbf{y}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\psi}, \nu_\psi, \mathbf{h}_0, \boldsymbol{\Sigma}_h)$;

4. $p(\nu_\psi \,|\, \mathbf{y}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\psi}, \boldsymbol{\kappa}, \mathbf{h}_0, \boldsymbol{\Sigma}_h)$;

5. $p(\mathbf{h}_i \,|\, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\kappa}, \nu_\psi, \mathbf{h}_0, \boldsymbol{\Sigma}_h)$, $i = 1, \ldots, n$;

6. $p(\mathbf{h}_0 \,|\, \mathbf{y}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\psi}, \boldsymbol{\kappa}, \nu_\psi, \boldsymbol{\Sigma}_h)$;

7. $p(\boldsymbol{\Sigma}_h \,|\, \mathbf{y}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\psi}, \boldsymbol{\kappa}, \nu_\psi, \mathbf{h}_0)$.

Steps 4-7 are standard and we leave the details to Appendix B. Here we focus on the first three steps.

<u>**Step 1**</u>. Note that the likelihood function in (10) can be written as a product of n Gaussian densities, each depends only on $(\boldsymbol{\theta}_i, \mathbf{h}_i)$. And since the priors on $\boldsymbol{\theta}_i$ are independent across equations, we can sample $\boldsymbol{\theta}_i$ equation by equation without loss of efficiency. To that end, let $\mathbf{m}_i = (m_{i,1}, \ldots, m_{i,k_i})'$. Then, we can rewrite the conditional prior of $\boldsymbol{\theta}_i$ in (8) as:

$$(\boldsymbol{\theta}_i \,|\, \kappa_1, \kappa_2, \boldsymbol{\psi}_i) \sim \mathcal{N}(\mathbf{m}_i, \mathbf{V}_i),$$

where $\mathbf{V}_i = \mathrm{diag}(2\kappa_{i,1}\psi_{i,1}C_{i,1}, \ldots, 2\kappa_{i,k_i}\psi_{i,k_i}C_{i,k_i})$. Combining the above prior and the likelihood in (10), we have

$$(\boldsymbol{\theta}_i \,|\, \mathbf{y}, \mathbf{h}, \boldsymbol{\psi}, \boldsymbol{\kappa}, \nu_\psi, \mathbf{h}_0, \boldsymbol{\Sigma}_h) \sim \mathcal{N}(\widehat{\boldsymbol{\theta}}_i, \mathbf{K}_{\boldsymbol{\theta}_i}^{-1}),$$

where

$$\mathbf{K}_{\boldsymbol{\theta}_i} = \mathbf{V}_i^{-1} + \mathbf{X}_i'\boldsymbol{\Omega}_{\mathbf{h}_i}^{-1}\mathbf{X}_i, \quad \widehat{\boldsymbol{\theta}}_i = \mathbf{K}_{\boldsymbol{\theta}_i}^{-1}(\mathbf{V}_i^{-1}\mathbf{m}_i + \mathbf{X}_i'\boldsymbol{\Omega}_{\mathbf{h}_i}^{-1}\mathbf{y}_i).$$

Here the covariance matrix $\mathbf{K}_{\boldsymbol{\theta}_i}^{-1}$ is of dimension $k_i = np + i$. Conventional methods to sample from the normal distribution require the Cholesky factor of $\mathbf{K}_{\boldsymbol{\theta}_i}^{-1}$. When n is large, computing the covariance matrix $\mathbf{K}_{\boldsymbol{\theta}_i}^{-1}$ explicitly by inverting $\mathbf{K}_{\boldsymbol{\theta}_i}$ is computationally intensive. It turns out we do not need an explicit expression of the inverse $\mathbf{K}_{\boldsymbol{\theta}_i}^{-1}$.

To explain the method, we introduce the following notations: given a non-singular square matrix $\mathbf{F}$ and a conformable vector $\mathbf{d}$, let $\mathbf{F}\backslash\mathbf{d}$ denote the unique solution to the linear system $\mathbf{F}\mathbf{z} = \mathbf{d}$, i.e., $\mathbf{F}\backslash\mathbf{d} = \mathbf{F}^{-1}\mathbf{d}$. When $\mathbf{F}$ is lower triangular, this linear system can be solved quickly by forward substitution; when $\mathbf{F}$ is upper triangular, it can be solved by backward substitution.[4] Now, compute the Cholesky factor $\mathbf{C}_{\mathbf{K}_{\boldsymbol{\theta}_i}}$ of $\mathbf{K}_{\boldsymbol{\theta}_i}$ such that $\mathbf{K}_{\boldsymbol{\theta}_i} = \mathbf{C}_{\mathbf{K}_{\boldsymbol{\theta}_i}}\mathbf{C}'_{\mathbf{K}_{\boldsymbol{\theta}_i}}$. It is easy to verify that $\widehat{\boldsymbol{\theta}}_i$ can be calculated as: $\mathbf{C}'_{\mathbf{K}_{\boldsymbol{\theta}_i}}\backslash(\mathbf{C}_{\mathbf{K}_{\boldsymbol{\theta}_i}}\backslash(\mathbf{V}_i^{-1}\mathbf{m}_i + \mathbf{X}_i'\boldsymbol{\Omega}_{\mathbf{h}_i}^{-1}\mathbf{y}_i))$ by forward then backward substitution. Next, let $\mathbf{u}$ be a $k_i \times 1$ vector of independent $\mathcal{N}(0,1)$ random variables. Then, return

$$\widehat{\boldsymbol{\theta}}_i + \mathbf{C}'_{\mathbf{K}_{\boldsymbol{\theta}_i}}\backslash\mathbf{u},$$

which has the $\mathcal{N}(\widehat{\boldsymbol{\theta}}_i, \mathbf{K}_{\boldsymbol{\theta}_i}^{-1})$ distribution.

<u>**Step 2**</u>. First, note that the free elements of $\boldsymbol{\psi}_i$ are conditionally independent and we can sample them one by one without loss of efficiency. Next, combining (8) and (9), we obtain

$$\begin{aligned} p(\psi_{i,j}\,|\,\mathbf{y},\boldsymbol{\theta},\mathbf{h},\boldsymbol{\kappa},\nu_\psi,\mathbf{h}_0,\boldsymbol{\Sigma}_h) &\propto \psi_{i,j}^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{4\kappa_{i,j}C_{i,j}\psi_{i,j}}(\theta_{i,j}-m_{i,j})^2} \times \psi_{i,j}^{\nu_\psi-1}\mathrm{e}^{-\frac{\nu_\psi}{2}\psi_{i,j}} \\ &= \psi_{i,j}^{\nu_\psi-\frac{1}{2}-1}\mathrm{e}^{-\frac{1}{2}\left(\nu_\psi\psi_{i,j}+\psi_{i,j}^{-1}\frac{(\theta_{i,j}-m_{i,j})^2}{2\kappa_{i,j}C_{i,j}}\right)}, \end{aligned}$$

which is the kernel of a generalized inverse Gaussian distribution. More precisely, we have

$$(\psi_{i,j}\,|\,\mathbf{y},\boldsymbol{\theta},\mathbf{h},\boldsymbol{\kappa},\nu_\psi,\mathbf{h}_0,\boldsymbol{\Sigma}_h) \sim \mathcal{GIG}\left(\nu_\psi - \frac{1}{2}, \nu_\psi, \frac{(\theta_{i,j}-m_{i,j})^2}{2\kappa_{i,j}C_{i,j}}\right).$$

<u>**Step 3**</u>. Note that κ_1 and κ_2 only appear in their priors $\kappa_j \sim \mathcal{G}(c_{1,j}, c_{2,j}), j = 1,2$, and in (8) (recall $\kappa_{i,j} = \kappa_1$ for coefficients on own lags and $\kappa_{i,j} = \kappa_2$ for coefficients on other lags). To sample κ_1 and κ_2, first define the index set S_{κ_1} that collects all the indexes (i,j) such that $\theta_{i,j}$ is a coefficient associated with an own lag. That is,

[4] Forward and backward substitutions are implemented in standard packages such as MATLAB, GAUSS and R. For instance, in MATLAB it is done by `mldivide(F, d)` or simply $\mathbf{F}\backslash\mathbf{d}$.

$S_{\kappa_1} = \{(i,j) : \theta_{i,j}$ is a coefficient associated with an own lag$\}$. Similarly, define S_{κ_2} as the set that collects all the indexes (i,j) such that $\theta_{i,j}$ is a coefficient associated with a lag of other variables. It is easy to check that the numbers of elements in S_{κ_1} and S_{κ_2} are respectively np and $(n-1)np$. Then, we have

$$\begin{aligned} p(\kappa_1 \,|\, \mathbf{y}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\psi}, \nu_\psi, \mathbf{h}_0, \boldsymbol{\Sigma}_h) &\propto \prod_{(i,j)\in S_{\kappa_1}} \kappa_1^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{4\kappa_1 C_{i,j}\psi_{i,j}}(\theta_{i,j}-m_{i,j})^2} \times \kappa_1^{c_{1,1}-1} \mathrm{e}^{-\kappa_1 c_{2,1}} \\ &= \kappa_1^{c_{1,1}-\frac{np}{2}-1} \mathrm{e}^{-\frac{1}{2}\left(2c_{2,1}\kappa_1 + \kappa_1^{-1}\sum_{(i,j)\in S_{\kappa_1}} \frac{(\theta_{i,j}-m_{i,j})^2}{2\psi_{i,j}C_{i,j}}\right)}, \end{aligned}$$

which is the kernel of the $\mathcal{GIG}\left(c_{1,1} - \frac{np}{2}, 2c_{2,1}, \sum_{(i,j)\in S_{\kappa_1}} \frac{(\theta_{i,j}-m_{i,j})^2}{2\psi_{i,j}C_{i,j}}\right)$ distribution. Similarly, we have

$$(\kappa_2 \,|\, \mathbf{y}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\psi}, \nu_\psi, \mathbf{h}_0, \boldsymbol{\Sigma}_h) \sim \mathcal{GIG}\left(c_{1,2} - \frac{(n-1)np}{2}, 2c_{2,2}, \sum_{(i,j)\in S_{\kappa_2}} \frac{(\theta_{i,j}-m_{i,j})^2}{2\psi_{i,j}C_{i,j}}\right).$$

The implementation details of the remaining steps are given in Appendix B.

4.2 A Numerical Comparison

In this section we report the estimation times for BVARs of different sizes under the Minnesota-type normal-gamma prior using the proposed sampler. For comparison, we also report the estimation times of the same models using the algorithm in Carriero, Clark, and Marcellino (2019). To have a fair comparison, when we implement their algorithm, we also avoid any explicit computation of inverse precision matrices whenever possible. Moreover, to simulate the log-volatilities, we use the more efficient precision sampler in Chan and Jeliazkov (2009) instead of Kalman filter-based methods.

The main difference between the two approaches is that Carriero, Clark, and Marcellino (2019) is designed for the reduced-form parameterization in (1), whereas the proposed sampler is for the structural-form parameterization in (3). A key advantage of the latter parameterization is that the VAR can be readily written as n separate regressions, and there is no need to obtain the 'orthogonalized' shocks at each MCMC iteration, as is required in the algorithm in Carriero, Clark, and Marcellino (2019). Hence, one would expect the proposed sampler would run faster. In addition, the algorithm in Carriero, Clark, and Marcellino (2019) requires an extra block to sample the free elements of the impact matrix $\mathbf{B}_0$, whereas the proposed sampler simulates them jointly with the VAR coefficients. Hence, the proposed sampler is expected to induce less autocorrelation in the Markov chain (at the expense of slower computation time).

Table 1 reports the computation times (in minutes) to obtain 10,000 posterior draws. All the BVARs have $p = 4$ lags, and the algorithms are implemented using Matlab on a desktop with an Intel Core i7-7700 @3.60 GHz processor and 64GB memory. As it is evident from the table, the proposed method is fast and scales well. For example, for a BVAR with $n = 25$ variables, simulating 10,000 posterior draws using the proposed posterior sampler takes about 3 minutes; when $n = 100$, it takes about 72 minutes. The proposed algorithm also compares favorably to the algorithm in Carriero, Clark, and Marcellino (2019), with a speed-up between 25% to about 3 times.

Table 1: The computation times (in minutes) to obtain 10,000 posterior draws under the Minnesota-type normal-gamma prior using the proposed method compared to the method in Carriero, Clark, and Marcellino (2019). All BVARs have $p = 4$ lags.

	$n = 25$	$n = 50$	$n = 100$
proposed method	2.9	10.2	71.8
CCM	8.2	20.4	90.0

5 Application: Forecasting with Large BVARs

We consider a forecasting exercise using large BVARs to illustrate the usefulness of the proposed Minnesota-type adaptive hierarchical priors. We first describe the macroeconomic dataset in Section 5.1, which is followed by some full sample results in Section 5.2. We then compare in Section 5.3 the forecast performance of the proposed Minnesota-type normal-gamma prior with two important benchmarks: the normal-gamma prior and the data-based Minnesota prior.

5.1 Data

We use a dataset that consists of 23 US quarterly variables with a sample period from 1959Q1 to 2018Q4. It is constructed from the FRED-QD database at the Federal Reserve Bank of St. Louis as described in McCracken and Ng (2016). The dataset contains a range of standard macroeconomic and financial variables, such as Real GDP, industrial production, inflation rates, labor market variables and interest rates. They are transformed to stationarity, typically to growth rates. The complete list of variables and how they are transformed is given in Appendix A.

5.2 Full Sample Results

Next, we present some full sample results that highlight the differences of the proposed Minnesota-type normal-gamma prior compared to the normal-gamma prior and the data-based Minnesota prior. We first report in Table 2 the posterior estimates of κ_1 and κ_2 under the three priors.

Table 2: Posterior means and standard deviations (in parenthesis) of κ_1 and κ_2 under the normal-gamma prior ($\kappa_1 = \kappa_2$), the Minnesota prior and the proposed Minnesota-type normal-gamma prior.

	normal-gamma	Minnesota	Minnesota-type normal-gamma
κ_1	0.0007	0.093	0.041
	(0.0001)	(0.0152)	(0.0171)
κ_2	0.0007	0.0028	0.0006
	(0.0001)	(0.0003)	(0.0001)
ν_ψ	0.13	–	0.15
	(0.004)	–	(0.012)

Recall that the normal-gamma does not distinguish own lags versus other lags and it restricts $\kappa_1 = \kappa_2$. Under the normal-gamma prior, the posterior mean is 0.0007, implying aggressive global shrinkage. This is a general feature of the family of global-local priors—strong global shrinkage handles the noise, while the local variance component detects the signals (see Polson and Scott, 2010, for more discussion). However, if we allow κ_1 and κ_2 to be different as in the proposed Minnesota-type normal-gamma prior, we obtain very different results: the posterior mean of κ_1 increases about 58 times to 0.041, whereas the posterior mean of κ_2 reduces to 0.0006. These estimates suggest that the data prefers shrinking the coefficients on lags of other variables much more strongly to zero than those on own lags. This is consistent with the prior belief that, on average, a variable's own lags contain more information about its future evolution than lags of other variables. These results highlight the empirical relevance of allowing for cross-variable shrinkage.

In addition, the posterior means of κ_1 and κ_2 under the Minnesota prior are both substantially larger than those of the Minnesota-type normal-gamma prior. Specifically, the posterior means of κ_1 and κ_2 under the Minnesota prior are, respectively, 2.3 and 4.7 times larger than the latter. This makes intuitive sense as the only difference between the two priors is the addition of the local variance component in the latter. By allowing for a local variance component to handle 'large' coefficients, the global component can shrink all coefficients more aggressively. Hence, the estimates of κ_1 and κ_2 under the Minnesota-type normal-gamma prior are both smaller.

We also report in Table 2 the posterior estimates of ν_ψ for the two BVARs with the normal-gamma prior. Recall that when $\nu_\psi = 1$, the normal-gamma prior reduces to the standard Bayesian Lasso prior. In both cases the estimates are far from unity with small posterior standard deviations. These estimates indicate that the Bayesian Lasso prior might be too restrictive.

Overall, the full sample results suggest that the features of both the Minnesota prior and the normal-gamma prior are empirically useful. In particular, the results highlight the importance of the addition of a local variance component, as well as allowing for different levels of shrinkage on own versus other lags. Hence, these results show the empirical relevance of the proposed Minnesota-type normal-gamma prior.

5.3 Forecasting Results

Next, we evaluate the forecast performance of BVARs with the proposed Minnesota-type normal-gamma prior relative to two alternative priors: the normal-gamma and the Minnesota prior. The sample period is from 1959Q1 to 2018Q4, and the forecast performance of the models is evaluated from 1985Q1 till the end of the sample. In each recursive forecasting iteration, we use only the data up to time t, denoted as $\mathbf{y}_{1:t}$, to estimate the models. We evaluate both point and density forecasts, and we use the conditional expectation $\mathbb{E}(y_{i,t+m} \,|\, \mathbf{y}_{1:t})$ as the m-step-ahead point forecast for variable i and the predictive density $p(y_{i,t+m} \,|\, \mathbf{y}_{1:t})$ as the corresponding density forecast.

We use the root mean squared forecast error (RMSFE) to evaluate the point forecasts from model M, which is defined as:

$$\text{RMSFE}^{M}_{i,m} = \sqrt{\frac{\sum_{t=t_0}^{T-m}(y^{\text{o}}_{i,t+m} - \mathbb{E}(y_{i,t+m} \,|\, \mathbf{y}_{1:t}))^2}{T-m-t_0+1}},$$

where $y^{\text{o}}_{i,t+m}$ is the observed value of $y_{i,t+h}$. To evaluate the density forecasts, the metric we use is the average of log predictive likelihoods (ALPL):

$$\text{ALPL}^{M}_{i,m} = \frac{1}{T-m-t_0+1}\sum_{t=t_0}^{T-m}\log p(y_{i,t+m} = y^{\text{o}}_{i,t+m} \,|\, \mathbf{y}_{1:t}),$$

where $p(y_{i,t+m} = y^{\text{o}}_{i,t+m} \,|\, \mathbf{y}_{1:t})$ is the predictive likelihood. For this metric, a larger value indicates better forecast performance.

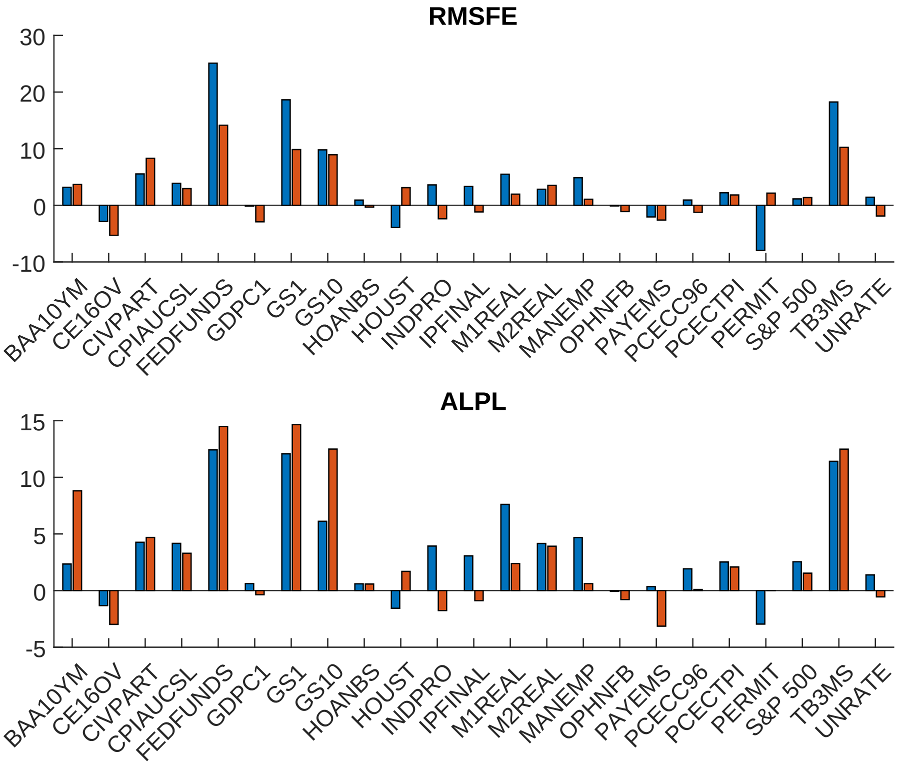
To compare the forecast performance of model M against the benchmark B, we follow Carriero, Clark, and Marcellino (2015) to report the percentage gains in terms of RMSFE, defined as

$$100 \times (1 - \text{RMSFE}^{M}_{i,m}/\text{RMSFE}^{B}_{i,m}),$$

and the percentage gains in terms of ALPL:

$$100 \times (\text{ALPL}^{M}_{i,m} - \text{ALPL}^{B}_{i,m}).$$

Figure 1 reports the forecasting results from the BVARs with the proposed Minnesota-type normal-gamma prior relative to the benchmark normal-gamma prior. The top panel shows the percentage gains in RMSFE for all 23 variables, whereas the bottom panel presents the corresponding results in ALPL. For 1-step-ahead point forecasts, the Minnesota-type normal-gamma prior outperforms the benchmark for most of the variables. For a few variables, such as real output per hour for nonfarm business section, Fed funds rate and 3-month treasury bill rate, the former outperforms the benchmark between 18%-25%. For 4-step-ahead point forecasts, the Minnesota-type normal-gamma prior similarly outperforms the benchmark, though the gains are more modest. The median percentage gains in RMSFE for 1- and 4-step-ahead forecasts are 2.8% and 1.8%, respectively, whereas the mean percentage gains are 4.1% and 2.4%. Results for density forecasts are similar: the median percentage gains in ALPL for 1- and 4-step-ahead forecasts are, respectively, 2.5% and 1.5%, while the the mean percentage gains are 3.5% and 3.2%. These results demonstrate that by incorporating richer prior beliefs such as cross-variable shrinkage, one can substantially improve the forecast performance of the normal-gamma prior.


RMSFE

ALPL

Figure 1: Forecasting results from BVAR with the Minnesota-type normal-gamma prior versus BVAR with the normal-gamma. The top panel shows the percentage gains in root mean squared forecast error of the Minnesota-type normal-gamma prior. The bottom panel presents the percentage gains in the average of log predictive likelihoods.

Next, we compare the forecast performance of the Minnesota-type normal-gamma prior with that of the Minnesota prior, and the results are reported in Figure 2. For the 1-step-ahead point forecasts, the Minnesota-type normal-gamma prior outperforms the Minnesota prior for a majority of variables. The median and mean percentage gains in

RMSFE are 0.5% and 0.9%, respectively. For 4-step-ahead point forecasts, the results are more mixed. The median percentage gains in RMSFE is 0.5%, whereas the mean percentage gains is -0.26%. However, the Minnesota-type normal-gamma prior performs better than the benchmark for density forecasts for both 1- and 4-step-ahead forecast horizons. For example, The mean percentage gains in ALPL are 1.6% and 1.7%, respectively. Overall, these results show that one can improve the forecast performance of the Minnesota prior by the addition of a local variance component as in the normal-gamma prior.

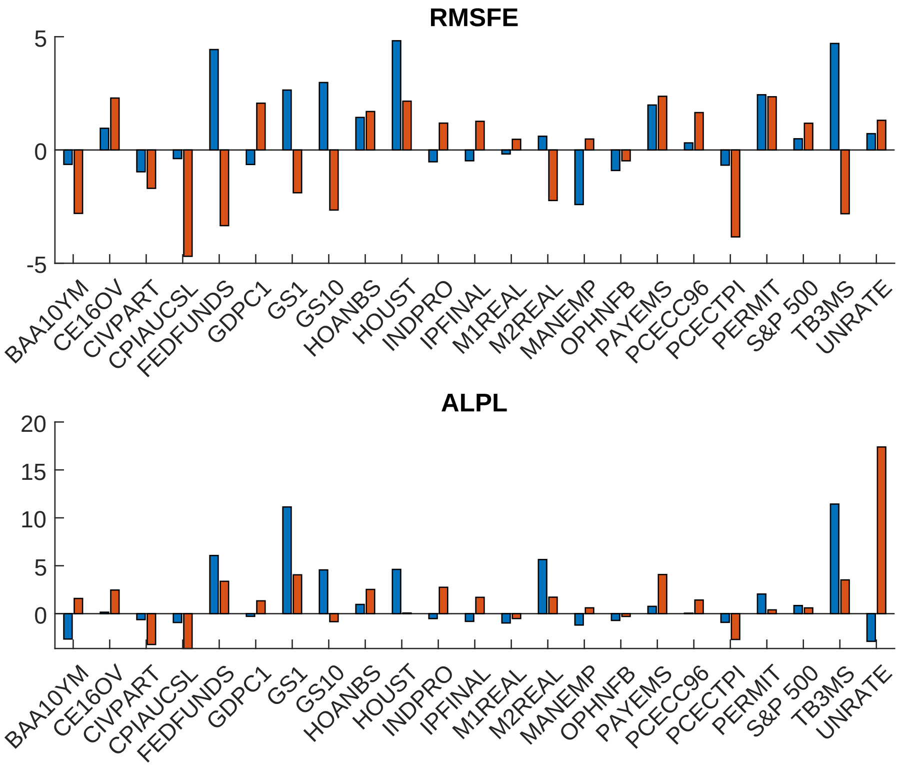


Figure 2: Forecasting results from BVAR with the Minnesota-type normal-gamma prior versus BVAR with the Minnesota prior. The top panel shows the percentage gains in root mean squared forecast error of the Minnesota-type normal-gamma prior. The bottom panel presents the percentage gains in the average of log predictive likelihoods.

6 Concluding Remarks and Future Research

We have developed a new family of shrinkage priors that combines the useful features of both the Minnesota prior and the global-local priors. Using a large US dataset, we demonstrated that the gains in forecast accuracy of these new priors can be substantial compared to the Minnesota prior and the global-local priors. In particular, our results highlighted the importance of allowing for cross-variable shrinkage, as well as the addition of a local variance component.

The family of Minnesota priors has stood the test of time, and continues to be an important benchmark. In future work, it would be useful to develop more flexible Minnesota-type priors, where key hyperparameters are selected by the data. For example, coefficients on lags under a standard Minnesota prior are shrunk at a rate of $1/l^2$, where l is the lag length. It would be interesting to consider other types of shrinkage.

Appendix A: Data

The dataset is sourced from the FRED-QD database at the Federal Reserve Bank of St. Louis (McCracken and Ng, 2016). It covers the quarters from 1959Q1 to 2018Q4. Table 3 lists the 23 quarterly variables and describes how they are transformed. For example, $\Delta \log$ is used to denote the first difference in the logs, i.e., $\Delta \log x = \log x_t - \log x_{t-1}$.

Table 3: Description of variables used in empirical application.

Variable	Mnemonic	Transformation
Real Gross Domestic Product	GDPC1	$400\Delta \log$
Personal Consumption Expenditures	PCECC96	$400\Delta \log$
Industrial Production Index	INDPRO	$400\Delta \log$
Industrial Production: Final Products	IPFINAL	$400\Delta \log$
All Employees: Total nonfarm	PAYEMS	$400\Delta \log$
All Employees: Manufacturing	MANEMP	$400\Delta \log$
Civilian Employment	CE16OV	$400\Delta \log$
Civilian Labor Force Participation Rate	CIVPART	no transformation
Civilian Unemployment Rate	UNRATE	no transformation
Nonfarm Business Section: Hours of All Persons	HOANBS	$400\Delta \log$
Housing Starts: Total	HOUST	$400\Delta \log$
New Private Housing Units Authorized by Building Permits	PERMIT	$400\Delta \log$
Personal Consumption Expenditures: Chain-type Price index	PCECTPI	$400\Delta \log$
Consumer Price Index for All Urban Consumers: All Items	CPIAUCSL	$400\Delta \log$
Nonfarm Business Section: Real Output Per Hour of All Persons	OPHNFB	$400\Delta \log$
Effective Federal Funds Rate	FEDFUNDS	no transformation
3-Month Treasury Bill: Secondary Market Rate	TB3MS	no transformation
1-Year Treasury Constant Maturity Rate	GS1	no transformation
10-Year Treasury Constant Maturity Rate	GS10	no transformation
Moody's Seasoned Baa Corporate Bond Yield Relative to Yield on 10-Year Treasury Constant Maturity	BAA10YM	no transformation
Real M1 Money Stock	M1REAL	$400\Delta \log$
Real M2 Money Stock	M2REAL	$400\Delta \log$
S&P's Common Stock Price Index : Composite	S&P 500	$400\Delta \log$

Appendix B: Estimation Details

In this appendix we provide estimation details of the structural VAR in (3) with the proposed Minnesota-type adaptive hierarchical prior given in (6). In the main text we discuss Steps 1-3. Here we describe the details of the remaining steps.

Step 4. To implement Step 4, we first derive the log conditional density of ν_ψ: $\log p(\nu_\psi \,|\, \mathbf{y}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\psi}, \boldsymbol{\kappa}, \mathbf{h}_0, \boldsymbol{\Sigma}_h) = p(\nu_\psi \,|\, \boldsymbol{\psi})$. It follows from (9) and the prior $\nu_\psi \sim \mathcal{G}(d_1, d_2)$ that we have:

$$\begin{aligned}\log p(\nu_\psi \,|\, \boldsymbol{\psi}) = n^2 p(\nu_\psi \log(\nu_\psi/2) - \log\Gamma(\nu_\psi)) + (\nu_\psi - 1)\sum \log \psi_{i,j} \\ - \frac{\nu_\psi}{2}\sum \psi_{i,j} - (d_1 - 1)\log \nu_\psi - d_2\nu_\psi + c_1,\end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function and c_1 is a normalization constant. It is easy to check that the first and second derivatives of this log-density with respect to ν_ψ are given by

$$\begin{aligned}\frac{\mathrm{d}\log p(\nu_\psi \,|\, \boldsymbol{\psi})}{\mathrm{d}\nu_\psi} &= n^2 p(\log(\nu_\psi/2) + 1 - \Psi(\nu_\psi)) + \sum \log \psi_{i,j} \\ &\quad - \frac{1}{2}\sum \psi_{i,j} - (d_1 - 1)\nu_\psi^{-1} - d_2, \\ \frac{\mathrm{d}^2 \log p(\nu_\psi \,|\, \boldsymbol{\psi})}{\mathrm{d}\nu_\psi^2} &= n^2 p(\nu_\psi^{-1} + (d_1 - 1)\nu_\psi^{-2},\end{aligned}$$

where $\Psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log\Gamma(x)$ and $\Psi'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\Psi(x)$ are respectively the digamma and trigamma functions.

Since the first and second derivatives can be evaluated quickly, we can maximize $\log p(\nu_\psi \,|\, \boldsymbol{\psi})$ using Newton-Raphson method and obtain the mode and the negative Hessian evaluated at the mode, denoted as $\widehat{\nu}_\psi$ and K_{ν_ψ}, respectively. Then, we implement an independence-chain Metropolis-Hastings step with proposal distribution $\mathcal{N}(\widehat{\nu}_\psi, K_{\nu_\psi}^{-1})$.

Step 5. Step 5 is straightforward to implement as (4) is already in the form of a univariate regression. Specifically, we can directly apply the auxiliary mixture sampler in Kim, Shephard, and Chib (1998) in conjunction with the precision sampler of Chan and Jeliazkov (2009) to simulate $(\mathbf{h}_i \,|\, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\kappa}, \nu_\psi, \mathbf{h}_0, \boldsymbol{\Sigma}_h)$ for $i = 1, \ldots, n$.

Step 6. Step 6 is also straightforward, as the full conditional distribution of $\mathbf{h}_0$ is Gaussian:

$$(\mathbf{h}_0 \,|\, \mathbf{y}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\psi}, \boldsymbol{\kappa}, \nu_\psi, \boldsymbol{\Sigma}_h) \sim \mathcal{N}(\widehat{\mathbf{h}}_0, \mathbf{K}_{\mathbf{h}_0}^{-1}),$$

where $\mathbf{K}_{\mathbf{h}_0} = \mathbf{V}_h^{-1} + \boldsymbol{\Sigma}_h^{-1}$ and $\widehat{\mathbf{h}}_0 = \mathbf{K}_{\mathbf{h}_0}^{-1}(\mathbf{V}_h^{-1}\mathbf{a}_h + \boldsymbol{\Sigma}_h^{-1}\mathbf{h}_1)$ with $\mathbf{h}_1 = (h_{1,1}, \ldots, h_{n,1})'$.

Step 7. Lastly, the elements of $\boldsymbol{\Sigma}_h$ are conditionally independent and follow inverse-gamma distributions:

$$(\sigma^2_{h,i} \,|\, \mathbf{y}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\psi}, \boldsymbol{\kappa}, \nu_\psi, \mathbf{h}_0) \sim \mathcal{IG}\left(\nu_{h,i} + \frac{T}{2}, S_{h,i} + \frac{1}{2}\sum_{t=1}^{T}(h_{i,t} - h_{i,t-1})^2\right)$$

for $i = 1, \ldots, n$.

References

Banbura, M., D. Giannone, M. Modugno, and L. Reichlin (2013): “Now-casting and the real-time data flow,” in *Handbook of Economic Forecasting*, vol. 2, pp. 195–237. Elsevier.

Banbura, M., D. Giannone, and L. Reichlin (2010): “Large Bayesian vector auto regressions,” *Journal of Applied Econometrics*, 25(1), 71–92.

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

Carriero, A., T. E. Clark, and M. G. Marcellino (2015): “Bayesian VARs: Specification Choices and Forecast Accuracy,” *Journal of Applied Econometrics*, 30(1), 46–73.

——— (2016): “Common drifting volatility in large Bayesian VARs,” *Journal of Business and Economic Statistics*, 34(3), 375–390.

——— (2019): “Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors,” *Journal of Econometrics*, Forthcoming.

Carriero, A., G. Kapetanios, and M. Marcellino (2009): “Forecasting exchange rates with a large Bayesian VAR,” *International Journal of Forecasting*, 25(2), 400–417.

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

Chan, J. C. C. (2018): “Large Bayesian VARs: A Flexible Kronecker Error Covariance Structure,” *Journal of Business and Economic Statistics*, Forthcoming.

——— (2019a): “Asymmetric Conjugate Priors for Large Bayesian VARs,” *CAMA Working Paper 51/2019.*

——— (2019b): “Large Bayesian Vector Autoregressions,” *CAMA Working Paper 19/2019.*

Chan, J. C. C., and E. Eisenstat (2018): “Comparing Hybrid Time-Varying Parameter VARs,” *Economics Letters*, 171, 1–5.

Chan, J. C. C., L. Jacobi, and D. Zhu (2019): “Efficient Selection of Hyperparameters in Large Bayesian VARs Using Automatic Differentiation,” *CAMA Working Paper 46/2019.*

Chan, J. C. C., and I. Jeliazkov (2009): "Efficient Simulation and Integrated Likelihood Estimation in State Space Models," *International Journal of Mathematical Modelling and Numerical Optimisation*, 1(1), 101–120.

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

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

Cross, J., C. Hou, and A. Poon (2019): "Macroeconomic forecasting with large Bayesian VARs: Global-local priors and the illusion of sparsity," *Working Paper*.

Cross, J., and A. Poon (2016): "Forecasting structural change and fat-tailed events in Australian macroeconomic variables," *Economic Modelling*, 58, 34–51.

D'Agostino, A., L. Gambetti, and D. Giannone (2013): "Macroeconomic forecasting and structural change," *Journal of Applied Econometrics*, 28, 82–101.

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

Follett, L., and C. Yu (2019): "Achieving Parsimony in Bayesian VARs with the Horseshoe Prior," *Econometrics and Statistics*, 11, 130–144.

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

Geweke, J. (1993): "Bayesian Treatment of the Independent Student-t Linear Model," *Journal of Applied Econometrics*, 8, S19–S40.

Giannone, D., M. Lenza, and G. E. Primiceri (2015): "Prior selection for vector autoregressions," *Review of Economics and Statistics*, 97(2), 436–451.

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

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

Kadiyala, R. K., and S. Karlsson (1993): "Forecasting with generalized Bayesian vector auto regressions," *Journal of Forecasting*, 12(3-4), 365–378.

——— (1997): "Numerical Methods for Estimation and inference in Bayesian VAR-models," *Journal of Applied Econometrics*, 12(2), 99–132.

KARLSSON, S. (2013): "Forecasting with Bayesian vector autoregressions," in *Handbook of Economic Forecasting*, ed. by G. Elliott, and A. Timmermann, vol. 2 of *Handbook of Economic Forecasting*, pp. 791–897. Elsevier.

KASTNER, G., AND F. HUBER (2018): "Sparse Bayesian vector autoregressions in huge dimensions," *arXiv preprint arXiv:1704.03239*.

KIM, S., N. SHEPHARD, AND S. CHIB (1998): "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models," *Review of Economic Studies*, 65(3), 361–393.

KOOP, G. (2013): "Forecasting with medium and large Bayesian VARs," *Journal of Applied Econometrics*, 28(2), 177–203.

KOOP, G., AND D. KOROBILIS (2010): "Bayesian Multivariate Time Series Methods for Empirical Macroeconomics," *Foundations and Trends in Econometrics*, 3(4), 267–358.

——— (2013): "Large time-varying parameter VARs," *Journal of Econometrics*, 177(2), 185–198.

KOROBILIS, D., AND D. PETTENUZZO (2019): "Adaptive hierarchical priors for high-dimensional vector autoregressions," *Journal of Econometrics*, forthcoming.

LITTERMAN, R. (1986): "Forecasting With Bayesian Vector Autoregressions — Five Years of Experience," *Journal of Business and Economic Statistics*, 4, 25–38.

MCCRACKEN, M. W., AND S. NG (2016): "FRED-MD: A monthly database for macroeconomic research," *Journal of Business and Economic Statistics*, 34(4), 574–589.

PARK, T., AND G. CASELLA (2008): "The Bayesian Lasso," *Journal of the American Statistical Association*, 103(482), 681–686.

POLSON, N. G., AND J. G. SCOTT (2010): "Shrink globally, act locally: Sparse Bayesian regularization and prediction," *Bayesian statistics*, 9, 501–538.

SIMS, C. A., AND T. ZHA (1998): "Bayesian methods for dynamic multivariate models," *International Economic Review*, 39(4), 949–968.