Asymmetric Conjugate Priors for Large Bayesian VARs

Joshua C.C. Chan*
Purdue University

November 2019

Abstract

Large Bayesian VARs are now widely used in empirical macroeconomics. One popular shrinkage prior in this setting is the natural conjugate prior as it facilitates posterior simulation and leads to a range of useful analytical results. This is, however, at the expense of modeling flexibility, as it rules out cross-variable shrinkage—i.e., shrinking coefficients on lags of other variables more aggressively than those on own lags. We develop a prior that has the best of both worlds: it can accommodate cross-variable shrinkage, while maintaining many useful analytical results, such as a closed-form expression of the marginal likelihood. This new prior also leads to fast posterior simulation—for a BVAR with 100 variables and 4 lags, obtaining 10,000 posterior draws takes less than half a minute on a standard desktop. In a forecasting exercise, we show that a data-driven asymmetric prior outperforms two useful benchmarks: a data-driven symmetric prior and a subjective asymmetric prior.

Keywords: shrinkage prior, forecasting, marginal likelihood, optimal hyperparameters, structural VAR

JEL classifications: C11, C52, C55, E37, E47

*I thank Todd Clark and Gary Koop for many constructive comments and useful suggestions that have improved a previous version of the paper.

1 Introduction

Large Bayesian vector autoregressions (BVARs) have become increasingly popular in empirical macroeconomics for forecasting and structural analysis since the influential work by Banbura, Giannone, and Reichlin (2010). Prominent examples include Carriero, Kapetanios, and Marcellino (2009), Koop (2013), Koop and Korobilis (2013) and Korobilis and Pettenuzzo (2019). VARs tend to have a lot of parameters, and the key that makes these highly parameterized VARs useful is the introduction of shrinkage priors. For large BVARs, one commonly adopted prior is the natural conjugate prior, which has a few advantages over alternatives. First, this prior is conjugate, and consequently it gives rise to a range of useful analytical results, including a closed-form expression of the marginal likelihood.[1] Second, the posterior covariance matrix of the VAR coefficients under this prior has a Kronecker product structure, which can be used to speed up computations.

On the other hand, a key limitation of the natural conjugate prior is that the prior covariance matrix of the VAR coefficients needs to have a Kronecker product structure, which implies cross-equation restrictions that might not be reasonable. In particular, this Kronecker structure requires symmetric treatment of own lags and lags of other variables. In many applications one might wish to shrink the coefficients on other variables' lags more strongly to zero than those of own lags. This cross-variable shrinkage, however, cannot be implemented using the natural conjugate prior due to this Kronecker structure. Carriero, Clark, and Marcellino (2015) summarize this dilemma between computational convenience and prior flexibility as: "While the pioneering work of Litterman (1986) suggested it was useful to have cross-variable shrinkage, it has become more common to estimate larger models without cross-variable shrinkage, in order to have a Kronecker structure that speeds up computations and facilitates simulation."

We develop a prior that solves this dilemma—this new prior allows asymmetric treatment between own lags and lags of other variables, while it maintains many useful analytical results, such as a closed-form expression of the marginal likelihood. In addition, we

[1] An analytical expression for the marginal likelihood is valuable for many purposes. First, it is useful for model selection—e.g., choosing the lag length in BVARs. Second, it can be used to select prior hyperparameters that control the degree of shrinkage. Examples include Del Negro and Schorfheide (2004), Schorfheide and Song (2015) and Carriero, Clark, and Marcellino (2015). This approach of selecting hyperparameters is incorporated in the BEAR MATLAB toolbox developed by the European Central Bank (Dieppe, Legrand, and Van Roye, 2016).

exploit these analytical results to develop an efficient method to simulate directly from the posterior distribution—we obtain *independent* posterior draws and avoid Markov chain Monte Carlo (MCMC) methods altogether. For a BVAR with 100 variables and 4 lags, simulating 10,000 posterior draws under this new asymmetric conjugate prior takes less than 30 seconds.

To develop this asymmetric conjugate prior, we first write the BVAR in the structural form, under which the error covariance matrix is diagonal. We then adopt an equation-by-equation estimation approach similar to that in Carriero, Clark, and Marcellino (2019). In particular, we assume that the parameters are *a priori* independent across equations—i.e., the joint prior density is a product of densities, each for the set of parameters in each equation. Under this setup, we show that if the VAR coefficients and the error variance in each equation follows a normal-inverse-gamma prior, the posterior distribution has the same form—i.e., it is a product of normal-inverse-gamma densities.

To help elicit the hyperparameters in this asymmetric conjugate prior, we prove that if we assume a standard inverse-Wishart prior on the reduced-form error covariance matrix, the implied prior on the structural-form impact matrix and error variances is a product of normal-inverse-gamma densities. Hence, using this proposition, we can first elicit the hyperparameters in the reduced-form prior, which is often more natural, and then obtain the implied hyperparameters in the structural-form prior. In addition, this proposition implies that the proposed prior—with carefully chosen hyperparameters—is invariant to reordering of the dependent variables.

We illustrate the empirical relevance of the proposed asymmetric conjugate prior with a forecasting exercise that involves 21 US quarterly macroeconomic and financial variables. More specifically, we use the analytical expression of the marginal likelihood under the asymmetric conjugate prior to obtain the optimal hyperparameters. We show that this data-based asymmetric prior outperforms two important benchmarks: 1) a data-based symmetric prior that rules out cross-variable shrinkage; and 2) an asymmetric prior in which the hyperparameters are chosen subjectively as in Carriero, Clark, and Marcellino (2015).

The rest of the paper is organized as follows. We first introduce in Section 2 a reparameterization of the reduced-form BVAR and the new asymmetric conjugate prior. We then derive the associated posterior distribution and the marginal likelihood. Section 3

discusses a few extensions of the standard BVAR, and outlines the corresponding sampling schemes. It is followed by a macroeconomic forecasting exercise to illustrate the usefulness of the proposed prior in Section 4. Lastly, Section 5 concludes and briefly discusses some future research directions.

2 Bayesian VARs and Conjugate Priors

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

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

where $\widetilde{\mathbf{b}}$ is an $n \times 1$ vector of intercepts, $\widetilde{\mathbf{B}}_1, \ldots, \widetilde{\mathbf{B}}_p$ are $n \times n$ VAR coefficient matrices and $\widetilde{\boldsymbol{\Sigma}}$ is a full covariance matrix.

The parameters in this model can be naturally divided into two blocks: the error covariance matrix $\widetilde{\boldsymbol{\Sigma}}$ and the matrix of intercepts and VAR coefficients, i.e., $\widetilde{\mathbf{B}} = (\widetilde{\mathbf{b}}, \widetilde{\mathbf{B}}_1, \cdots, \widetilde{\mathbf{B}}_p)'$. Under this parameterization, there is a conjugate prior on $(\widetilde{\mathbf{B}}, \widetilde{\boldsymbol{\Sigma}})$, namely, the normal-inverse-Wishart distribution:

$$\widetilde{\boldsymbol{\Sigma}} \sim \mathcal{IW}(\widetilde{\nu}_0, \widetilde{\mathbf{S}}_0), \quad (\mathrm{vec}(\widetilde{\mathbf{B}}) \,|\, \widetilde{\boldsymbol{\Sigma}}) \sim \mathcal{N}(\mathrm{vec}(\widetilde{\mathbf{B}}_0), \widetilde{\boldsymbol{\Sigma}} \otimes \widetilde{\mathbf{V}}),$$

where $\otimes$ denotes the Kronecker product, $\mathrm{vec}(\cdot)$ vectorizes a matrix by stacking the columns from left to right and $\mathcal{IW}$ denotes the inverse-Wishart distribution. This prior is commonly called the natural conjugate prior and can be traced back to Zellner (1971). For textbook treatment of this prior and the associated posterior distribution, see, e.g., Koop and Korobilis (2010), Karlsson (2013) or Chan (2019).

The main advantage of the natural conjugate prior is that it gives rise to a range of analytical results. For example, the associated posterior and one-step-ahead predictive distributions are both known; the marginal likelihood is also available in closed-form. These analytical results are useful for a variety of purposes. For instance, the closed-form expression of the marginal likelihood under the natural conjugate prior can be used to calculate optimal hyperparameters, as is done in Del Negro and Schorfheide (2004),

Schorfheide and Song (2015) and Carriero, Clark, and Marcellino (2015). The analytical expression of the posterior distribution of $(\widetilde{\mathbf{B}}, \widetilde{\mathbf{\Sigma}})$ can be used to develop efficient sampling algorithms to estimate more flexible Bayesian VARs. Examples include Carriero, Clark, and Marcellino (2016) and Chan (2018).

On the other hand, one key drawback of the natural conjugate prior is that the prior covariance matrix of $\text{vec}(\widetilde{\mathbf{B}})$ is restrictive—to be conjugate it needs to have the Kronecker product structure $\widetilde{\mathbf{\Sigma}} \otimes \widetilde{\mathbf{V}}$, which implies cross-equation restrictions on the covariance matrix. In particular, this structure requires symmetric treatment of own lags and lags of other variables. In many situations one might want to shrink the coefficients on lags of other variables more strongly to zero than those of own lags. This prior belief, however, cannot be implemented using the natural conjugate prior due to the Kronecker structure.

Here we develop a prior that solves this dilemma: this new prior allows asymmetric treatment between own lags and lags of other variables, while it maintains many useful analytical results. In what follows, we first consider a reparameterization of the reduced-form VAR in (1). We introduce in Section 2.2 the new asymmetric conjugate prior and discuss its properties. We then derive the associated posterior distribution and discuss an efficient sampling scheme in Section 2.3. Finally, we give an analytical expression of the marginal likelihood in Section 2.4.

2.1 The Bayesian VAR in Structural Form

In this section we introduce a reparameterization of the reduced-form VAR in (1) and derive the associated likelihood function. To that end, we first write the VAR in the following structural form:

$$\mathbf{A}\mathbf{y}_t = \mathbf{b} + \mathbf{B}_1\mathbf{y}_{t-1} + \cdots + \mathbf{B}_p\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t^y, \quad \boldsymbol{\varepsilon}_t^y \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \tag{2}$$

where $\mathbf{b}$ is an $n \times 1$ vector of intercepts, $\mathbf{B}_1, \ldots, \mathbf{B}_p$ are $n \times n$ VAR coefficient matrices, $\mathbf{A}$ is an $n \times n$ lower triangular matrix with ones on the diagonal and $\mathbf{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ is diagonal. Since the covariance matrix $\mathbf{\Sigma}$ is diagonal, we can estimate this recursive system equation by equation without loss of efficiency.[2] It is easy to see that we can

[2]Carriero, Clark, and Marcellino (2019) pioneer a similar equation-by-equation estimation approach to estimate a large VAR with a standard stochastic volatility specification. However, they use the reduced-

recover the reduced-form parameters by setting $\widetilde{\mathbf{b}} = \mathbf{A}^{-1}\mathbf{b}$, $\widetilde{\mathbf{B}}_j = \mathbf{A}^{-1}\mathbf{B}_j, j = 1, \ldots, p$ and $\widetilde{\mathbf{\Sigma}} = \mathbf{A}^{-1}\mathbf{\Sigma}(\mathbf{A}^{-1})'$.

For later reference, we 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 $\mathbf{A}$, i.e., $\boldsymbol{\alpha}_i = (A_{i,1}, \ldots, A_{i,i-1})'$. We then follow Chan and Eisenstat (2018) to rewrite the i-th equation of the system in (2) as:

$$y_{i,t} = \widetilde{\mathbf{w}}_{i,t}\boldsymbol{\alpha}_i + \widetilde{\mathbf{x}}_t\boldsymbol{\beta}_i + \varepsilon^y_{i,t}, \quad \varepsilon^y_{i,t} \sim \mathcal{N}(0, \sigma_i^2),$$

where $\widetilde{\mathbf{w}}_{i,t} = (-y_{1,t}, \ldots, -y_{i-1,t})$ and $\widetilde{\mathbf{x}}_t = (1, \mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p})$. Note that $y_{i,t}$ depends on the contemporaneous variables $y_{1,t}, \ldots, y_{i-1,t}$. But since the system is triangular, when we perform the change of variables from $\boldsymbol{\varepsilon}^y_t$ to $\mathbf{y}_t$ to obtain the likelihood function, the corresponding Jacobian has unit determinant and the likelihood function has the usual Gaussian form.

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

$$y_{i,t} = \mathbf{x}_{i,t}\boldsymbol{\theta}_i + \varepsilon^y_{i,t}, \quad \varepsilon^y_{i,t} \sim \mathcal{N}(0, \sigma_i^2),$$

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 (2) as a system of n independent regressions. Moreover, 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.[3]

To derive the likelihood function, we further stack $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,T})'$ and define $\mathbf{X}_i$ and $\boldsymbol{\varepsilon}^y_i$ similarly. Hence, we can rewrite the above equation as follows:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\theta}_i + \boldsymbol{\varepsilon}^y_i, \quad \boldsymbol{\varepsilon}^y_i \sim \mathcal{N}(\mathbf{0}, \sigma_i^2\mathbf{I}_T).$$

form parameterization in (1), whereas here we use the structural form in (2). As we will see below, the latter parameterization has the advantage of having a convenient representation as n independent regressions and it consequently leads to a more efficient sampling scheme. Ando and Zellner (2010) also consider a similar reparameterization of the reduced-form VAR that allows equation-by-equation estimation. But in their implementation they need to switch between two parameterizations, which makes estimation more cumbersome.

[3]This more efficient blocking scheme has been used previously in the literature. For example, Eisenstat, Chan, and Strachan (2016) use it to speed up computations in the context of time-varying parameter VARs with stochastic volatility.

Finally, let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_n')'$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_n^2)'$. Then, the likelihood function of the VAR in (2) is given by

$$p(\mathbf{y} \,|\, \boldsymbol{\theta}, \boldsymbol{\sigma}^2) = \prod_{i=1}^{n} p(\mathbf{y}_i \,|\, \boldsymbol{\theta}_i, \sigma_i^2) = \prod_{i=1}^{n} (2\pi\sigma_i^2)^{-\frac{T}{2}} \mathrm{e}^{-\frac{1}{2\sigma_i^2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta}_i)'(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta}_i)}. \tag{3}$$

In other words, the likelihood function is the product of n Gaussian densities.

2.2 Asymmetric Conjugate Priors

Next we introduce a conjugate prior on $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ that allows differential treatment between prior variances on own lags versus others. We assume that the parameters are *a priori* independent across equations, i.e., $p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2) = \prod_{i=1}^{n} p(\boldsymbol{\theta}_i, \sigma_i^2)$. Furthermore, we consider a normal-inverse-gamma prior for each pair $(\boldsymbol{\theta}_i, \sigma_i^2), i = 1, \ldots, n$:

$$(\boldsymbol{\theta}_i \,|\, \sigma_i^2) \sim \mathcal{N}(\mathbf{m}_i, \sigma_i^2 \mathbf{V}_i), \qquad \sigma_i^2 \sim \mathcal{IG}(\nu_i, S_i), \tag{4}$$

and we write $(\boldsymbol{\theta}_i, \sigma_i^2) \sim \mathcal{NIG}(\mathbf{m}_i, \mathbf{V}_i, \nu_i, S_i)$. In other words, the prior density of $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ is given by

$$p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2) = \prod_{i=1}^{n} c_i (\sigma_i^2)^{-\left(\nu_i + 1 + \frac{k_i}{2}\right)} \mathrm{e}^{-\frac{1}{\sigma_i^2}\left(S_i + \frac{1}{2}(\boldsymbol{\theta}_i - \mathbf{m}_i)'\mathbf{V}_i^{-1}(\boldsymbol{\theta}_i - \mathbf{m}_i)\right)}, \tag{5}$$

where $c_i = (2\pi)^{-\frac{k_i}{2}} |\mathbf{V}_i|^{-\frac{1}{2}} S_i^{\nu_i} / \Gamma(\nu_i)$.

Since the prior variance of each element of $\boldsymbol{\theta}_i$ is controlled by the corresponding diagonal element of $\mathbf{V}_i$, it is obvious that this prior can accommodate different prior variances between own lags versus others. As we will show in the next section, this prior is also conjugate. To distinguish this from the natural conjugate prior, we call the prior in (4) the asymmetric conjugate prior. The hyperparameters of the asymmetric conjugate prior are $\mathbf{m}_i, \mathbf{V}_i, \nu_i$ and S_i, $i = 1, \ldots, n$. Below we describe one way to elicit these hyperparameters.

Recall that for each equation i there are three types of parameters: $\boldsymbol{\alpha}_i$, the free parameters in the impact matrix $\mathbf{A}$; $\boldsymbol{\beta}_i$, the structural-form VAR coefficients; and σ_i^2, the structural-form error variance. For the hyperparameters associated with the conditional prior of $\boldsymbol{\beta}_i$, we follow Sims and Zha (1998), who consider Minnesota-type shrinkage priors for VAR

coefficients in the structural form.[4] For the hyperparameters associated with $\boldsymbol{\alpha}_i$ and σ_i^2, a good way to elicit them is less obvious. One concern is that an arbitrary choice of the hyperparameters for $\boldsymbol{\alpha}_i$ and σ_i^2 would induce some unreasonable prior on the reduced-form error covariance matrix $\widetilde{\boldsymbol{\Sigma}} = \mathbf{A}^{-1}\boldsymbol{\Sigma}(\mathbf{A}^{-1})'$. In particular, the induced prior on $\widetilde{\boldsymbol{\Sigma}}$ might not be invariant to reordering—e.g., the prior variance of the i-th reduced-form error depends on its position in the n-tuple. This problem is especially acute for large systems due to the fact that $\mathbf{A}^{-1}$ is lower triangular.

To avoid this potential non-invariance problem, we instead specify a prior on the reduced-form error covariance matrix $\widetilde{\boldsymbol{\Sigma}}$. And given this prior on $\widetilde{\boldsymbol{\Sigma}}$, we then derive the implied prior on $\boldsymbol{\alpha}_i$ and $\sigma_i^2, i = 1, \ldots, n$. To that end, we consider a standard inverse-Wishart prior on $\widetilde{\boldsymbol{\Sigma}}$ centered around $\mathbf{S} = \text{diag}(s_1^2, \ldots, s_n^2)$, where s_i^2 denotes the sample variance of the residuals from an AR(4) model for the variable $i, i = 1, \ldots, n$. More precisely, $\widetilde{\boldsymbol{\Sigma}} \sim \mathcal{IW}(\nu_0, \mathbf{S})$ with $\nu_0 = n + 2$. This prior on $\widetilde{\boldsymbol{\Sigma}}$ is commonly used in the literature (e.g., in Kadiyala and Karlsson, 1997; Carriero, Clark, and Marcellino, 2015). It turns out that, quite remarkably, the implied prior on $\boldsymbol{\alpha}_i$ and σ_i^2 is normal-inverse-gamma. The following proposition and corollary summarize this result.

Proposition 1. *Consider the following normal-inverse-gamma priors on the diagonal elements of* $\boldsymbol{\Sigma}$ *and the lower triangular elements of* $\mathbf{A}$*:*

$$\sigma_i^2 \sim \mathcal{IG}\left(\frac{\nu_0 + i - n}{2}, \frac{s_i^2}{2}\right), \quad i = 1, \ldots, n, \tag{6}$$

$$(A_{i,j} \,|\, \sigma_i^2) \sim \mathcal{N}\left(0, \frac{\sigma_i^2}{s_j^2}\right), \quad 1 \leqslant j < i \leqslant n,\ i = 2, \ldots, n. \tag{7}$$

Then, $\widetilde{\boldsymbol{\Sigma}}^{-1} = \mathbf{A}'\boldsymbol{\Sigma}^{-1}\mathbf{A}$ *has the Wishart distribution* $\widetilde{\boldsymbol{\Sigma}}^{-1} \sim \mathcal{W}(\nu_0, \mathbf{S}^{-1})$*, where* $\mathbf{S} = \text{diag}(s_1^2, \ldots, s_n^2)$*. It follows that* $\widetilde{\boldsymbol{\Sigma}} \sim \mathcal{IW}(\nu_0, \mathbf{S})$*.*

The proof is given in Appendix C. Since the mapping $\widetilde{\boldsymbol{\Sigma}}^{-1} = \mathbf{A}'\boldsymbol{\Sigma}^{-1}\mathbf{A}$ is one-to-one, the converse of Proposition 1 is also true.

Corollary 1. *Using the same notations as in Proposition 1, if* $\widetilde{\boldsymbol{\Sigma}} \sim \mathcal{IW}(\nu_0, \mathbf{S})$*, then the implied priors on* $A_{i,j}$ *and* σ_i^2*,* $i = 1, \ldots, n, j = 1, \ldots, i-1$*, are the normal-inverse-gamma distributions given in* (6) *and* (7)*.*

[4]The original Minnesota priors on the reduced-form VAR coefficients were first developed by Doan, Litterman, and Sims (1984) and Litterman (1986).

The proof is given in Appendix C. In addition, Proposition 1 holds for the more general case where $\mathbf{S}$ is any symmetric positive definite matrix. That is, the induced priors on the structural-form variances are independent gamma distributions and the conditional priors of the free elements of $\mathbf{A}$ are normal distributions. But unlike the previous case with diagonal $\mathbf{S}$, here the free elements in the same row of $\mathbf{A}$ are correlated. We summarize the results in the following corollary.[5] Its proof is given in Appendix C.

Corollary 2. *Suppose $\widetilde{\mathbf{\Sigma}} \sim \mathcal{IW}(\nu_0, \mathbf{R})$, where $\mathbf{R}$ is a symmetric positive definite matrix. Factor $\widetilde{\mathbf{\Sigma}}^{-1} = \mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C}$ and $\mathbf{R}^{-1} = \mathbf{L}'\mathbf{S}^{-1}\mathbf{L}$, where $\mathbf{C}$ and $\mathbf{L}$ are lower triangular matrices with ones on the main diagonal, $\mathbf{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_n^2)$ and $\mathbf{S} = diag(s_1^2, \ldots, s_n^2)$ are diagonal matrices. Let $\mathbf{c}_i$ denote the free elements of the i-th row of $\mathbf{C}$, i.e., $\mathbf{c}_i = (C_{i,1}, \ldots, C_{i,i-1})'$, and let $\mathbf{L}_{1:i-1}$ denote the $(i-1)\times(i-1)$ matrix that consists of the first $(i-1)$ rows and columns of $\mathbf{L}$. Similarly define $\mathbf{l}_i$ and $\mathbf{S}_{1:i-1}$. Then, the implied priors on $\mathbf{c}_i$ and σ_i^2 are*

$$\sigma_i^2 \sim \mathcal{IG}\left(\frac{\nu_0 + i - n}{2}, \frac{s_i^2}{2}\right), \quad i = 1, \ldots, n, \tag{8}$$

$$(\mathbf{c}_i \,|\, \sigma_i^2) \sim \mathcal{N}\left(\mathbf{l}_i, \sigma_i^2 \mathbf{L}_{1:i-1}' \mathbf{S}_{1:i-1}^{-1} \mathbf{L}_{1:i-1}\right), \quad i = 2, \ldots, n. \tag{9}$$

Since the mapping $\widetilde{\mathbf{\Sigma}}^{-1} = \mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C}$ is one-to-one, the converse is also true. That is, if we assume that $(\mathbf{c}_i, \sigma_i^2)$ follows the normal-inverse-gamma distributions given in (8) *and* (9)*, then the implied prior on $\widetilde{\mathbf{\Sigma}}$ is inverse-Wishart: $\widetilde{\mathbf{\Sigma}} \sim \mathcal{IW}(\nu_0, \mathbf{R})$.*

Hence, the above proposition and corollaries give us a guide to elicit the hyperparameters associated with $\boldsymbol{\alpha}_i$ and σ_i^2. In particular, we set $\nu_i = 1 + i/2$ and $S_i = s_i^2/2$. For the elements of $\mathbf{m}_i$ and $\mathbf{V}_i$ associated with $\boldsymbol{\alpha}_i$, they are discussed below.

Recall that the elements of $\mathbf{m}_i$ and $\mathbf{V}_i$ associated with $\boldsymbol{\beta}_i$ are elicited along the lines of Sims and Zha (1998). Together with Proposition 1, one can set $\mathbf{m}_i = \mathbf{0}$ to shrink the VAR coefficients to zero for growth rates data; for level data, $\mathbf{m}_i$ is set to be zero as well except the coefficient associated with the first own lag, which is set to be one.

To elicit $\mathbf{V}_i$, recall that $\mathbf{V}_i$ is the ratio of the prior covariance matrix of $\boldsymbol{\theta}_i$ relative to the error variance σ_i^2. Similar to the Minnesota prior, here we assume $\mathbf{V}_i$ to be diagonal with

[5] Chan and Jeliazkov (2009) have shown a similar result. However, their proof is not sufficiently constructive and they did not give an explicit mapping between the inverse-Wishart parameters and the normal-inverse-gamma parameters.

the k-th diagonal element $V_{i,kk}$ set to be:

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

The hyperparameter κ_1 controls the overall shrinkage strength for coefficients on own lags, whereas κ_2 controls those on lags of other variables. These two hyperparameters will play a key role in the empirical analysis, and we will select them optimally by maximizing the associated marginal likelihood. We set $\kappa_3 = 1$ as per Proposition 1. Lastly, we fix $\kappa_4 = 100$, which implies essentially no shrinkage for the intercepts.[6]

2.3 Posterior Distribution and Efficient Sampling

In this section we first derive the posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ under the asymmetric conjugate prior and show that it has indeed the same form as the prior. Then, we describe an efficient method for posterior simulation.

Since both the likelihood in (3) and the prior in (5) have the product form, we can estimate each pair $(\boldsymbol{\theta}_i, \sigma_i^2)$ separately. More specifically, the posterior distribution of

[6] In principle one can select κ_3 and κ_4 optimally as well, but the corresponding optimization is more costly to solve. More generally, high-dimensional numerical optimization using derivative-free methods is time consuming. One feasible alternative is to use Automatic Differentiation to obtain the relevant partial derivatives, which are then fed to numerical optimization routines that use these partial derivatives to more efficiently find the maximizer. See Chan, Jacobi, and Zhu (2019) for an example.

$(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ is given by:

$$
\begin{aligned}
p(\boldsymbol{\theta},\boldsymbol{\sigma}^2 \,|\, \mathbf{y}) &\propto p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2) p(\mathbf{y} \,|\, \boldsymbol{\theta}, \boldsymbol{\sigma}^2) = \prod_{i=1}^{n} p(\boldsymbol{\theta}_i, \sigma_i^2) p(\mathbf{y}_i \,|\, \boldsymbol{\theta}_i, \sigma_i^2) \\
&= \prod_{i=1}^{n} c_i (\sigma_i^2)^{-\left(\nu_i + 1 + \frac{k_i}{2}\right)} \mathrm{e}^{-\frac{1}{\sigma_i^2}\left(S_i + \frac{1}{2}(\boldsymbol{\theta}_i - \mathbf{m}_i)' \mathbf{V}_i^{-1} (\boldsymbol{\theta}_i - \mathbf{m}_i)\right)} \times (2\pi\sigma_i^2)^{-\frac{T}{2}} \mathrm{e}^{-\frac{1}{2\sigma_i^2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta}_i)'(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta}_i)} \\
&= \prod_{i=1}^{n} c_i (2\pi)^{-\frac{T}{2}} (\sigma_i^2)^{-\left(\nu_i + \frac{T+k_i}{2} + 1\right)} \mathrm{e}^{-\frac{1}{\sigma_i^2}\left(S_i + \frac{1}{2}\left(\boldsymbol{\theta}_i'(\mathbf{V}_i^{-1} + \mathbf{X}_i'\mathbf{X}_i)\boldsymbol{\theta}_i - 2\boldsymbol{\theta}_i'(\mathbf{V}_i^{-1}\mathbf{m}_i + \mathbf{X}_i'\mathbf{y}_i) + \mathbf{m}_i'\mathbf{V}_i^{-1}\mathbf{m}_i + \mathbf{y}_i'\mathbf{y}_i\right)\right)} \\
&= \prod_{i=1}^{n} c_i (2\pi)^{-\frac{T}{2}} (\sigma_i^2)^{-\left(\nu_i + \frac{T+k_i}{2} + 1\right)} \mathrm{e}^{-\frac{1}{\sigma_i^2}\left(\widehat{S}_i + \frac{1}{2}(\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_i)' \mathbf{K}_{\boldsymbol{\theta}_i} (\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_i)\right)},
\end{aligned}
$$

where $\mathbf{K}_{\boldsymbol{\theta}_i} = \mathbf{V}_i^{-1} + \mathbf{X}_i'\mathbf{X}_i$, $\widehat{\boldsymbol{\theta}}_i = \mathbf{K}_{\boldsymbol{\theta}_i}^{-1}(\mathbf{V}_i^{-1}\mathbf{m}_i + \mathbf{X}_i'\mathbf{y}_i)$ and $\widehat{S}_i = S_i + (\mathbf{y}_i'\mathbf{y}_i + \mathbf{m}_i'\mathbf{V}_i^{-1}\mathbf{m}_i - \widehat{\boldsymbol{\theta}}_i'\mathbf{K}_{\boldsymbol{\theta}_i}\widehat{\boldsymbol{\theta}}_i)/2$. Hence, the posterior distribution is a product of n normal-inverse-gamma distributions and we have:

$$
(\boldsymbol{\theta}_i, \sigma_i^2 \,|\, \mathbf{y}) \sim \mathcal{NIG}\left(\widehat{\boldsymbol{\theta}}_i, \mathbf{K}_{\boldsymbol{\theta}_i}^{-1}, \nu_i + \frac{T}{2}, \widehat{S}_i\right), \quad i = 1, \ldots, n. \tag{10}
$$

Using properties of the normal-inverse-gamma distribution, it is easy to see that the posterior means of $\boldsymbol{\theta}_i$ and σ_i^2 are respectively $\widehat{\boldsymbol{\theta}}_i$ and $\widehat{S}_i/(\nu_i + T/2 - 1)$. Other posterior moments can also be obtained by using similar properties of the normal-inverse-gamma distribution. For other quantities of interest where analytical results are not available, we can estimate them by posterior simulation. For example, the h-step-ahead predictive distribution of $\mathbf{y}_{T+h}$ is non-standard. But we can obtain posterior draws from $p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2 \,|\, \mathbf{y})$ to construct the h-step-ahead predictive distribution.

In what follows, we outline an efficient method to simulate a sample of size R from the posterior distribution. Here we can directly generate *independent* draws from the posterior distribution as opposed to MCMC draws that are correlated by construction. First, note that $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2 \,|\, \mathbf{y})$ is a product of n normal-inverse-gamma distributions as given in (10). Thus we can sample each pair $(\boldsymbol{\theta}_i, \sigma_i^2 \,|\, \mathbf{y})$ individually. Next, we can sample $(\boldsymbol{\theta}_i, \sigma_i^2 \,|\, \mathbf{y})$ in two steps. First, we draw σ_i^2 marginally from $(\sigma_i^2 \,|\, \mathbf{y}) \sim \mathcal{IG}(\nu_i + T/2, \widehat{S}_i)$. Then, given the σ_i^2 sampled, we obtain $\boldsymbol{\theta}_i$ from the conditional distribution

$$
(\boldsymbol{\theta}_i \,|\, \mathbf{y}, \sigma_i^2) \sim \mathcal{N}(\widehat{\boldsymbol{\theta}}_i, \sigma_i^2 \mathbf{K}_{\boldsymbol{\theta}_i}^{-1}).
$$

Here the covariance matrix $\sigma_i^2 \mathbf{K}_{\boldsymbol{\theta}_i}^{-1}$ is of dimension $k_i = np + i$. When n is large, sampling from this normal distribution using conventional methods—based on the Cholesky factor of $\sigma_i^2 \mathbf{K}_{\boldsymbol{\theta}_i}^{-1}$—is computationally intensive for two reasons. First, inverting the $k_i \times k_i$ matrix $\mathbf{K}_{\boldsymbol{\theta}_i}$ to obtain the covariance matrix $\sigma_i^2 \mathbf{K}_{\boldsymbol{\theta}_i}^{-1}$ is computationally costly. Second, the Cholesky factor of the covariance matrix needs to be computed R times—once for each draw of σ_i^2 from the marginal distribution. It turns out that both of these computationally intensive steps can be avoided.

To that end, 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.[7] Now, compute the Cholesky factor $\mathbf{C}_{\mathbf{K}_{\boldsymbol{\theta}_i}}$ of $\mathbf{K}_{\boldsymbol{\theta}_i}$ such that $\mathbf{K}_{\boldsymbol{\theta}_i} = \mathbf{C}_{\mathbf{K}_{\boldsymbol{\theta}_i}} \mathbf{C}'_{\mathbf{K}_{\boldsymbol{\theta}_i}}$. Note that this needs to be done only once. Let $\mathbf{u}$ be a $k_i \times 1$ vector of independent sample from $\mathcal{N}(0, \sigma_i^2)$. Then, return

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

which has the $\mathcal{N}(\widehat{\boldsymbol{\theta}}_i, \sigma_i^2 \mathbf{K}_{\boldsymbol{\theta}_i}^{-1})$ distribution.[8] Finally, we can further speed up the computations by vectorizing all operations to obtain R posterior draws instead of using for-loops.

This sampling scheme is more efficient than the method in Carriero, Clark, and Marcellino (2019), who propose estimating the reduced-form parameters equation-by-equation. The main reason is that their method requires computing the Cholesky factor of every sampled reduced-form covariance matrix (e.g., a total of R times for R draws), whereas we only need to do it once. This difference becomes more important when n becomes larger, as number of operations for computing the Cholesky factor of an $n \times n$ matrix is $\mathcal{O}(n^3)$.

To get a sense of how long it takes to obtain posterior draws using the proposed algorithm, we fit Bayesian VARs of different sizes, each with $p = 4$ lags. The algorithm is implemented using MATLAB on a desktop with an Intel Core i7-7700 @3.60 GHz processor and 64GB memory. The computation times (in seconds) to obtain 10,000 posterior draws of $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ are reported in Table 1. As it is evident from the table, the proposed

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

[8] Note that $\widehat{\boldsymbol{\theta}}_i$ can be obtained similarly without explicitly computing the inverse of $\mathbf{K}_{\boldsymbol{\theta}_i}$. Specifically, it is easy to see that $\widehat{\boldsymbol{\theta}}_i$ can be calculated as: $\mathbf{C}'_{\mathbf{K}_{\boldsymbol{\theta}_i}} \backslash (\mathbf{C}_{\mathbf{K}_{\boldsymbol{\theta}_i}} \backslash (\mathbf{V}_i^{-1}\mathbf{m}_i + \mathbf{X}_i'\mathbf{y}_i))$ by forward then backward substitution. Also note that since $\mathbf{V}_i^{-1}$ is diagonal, its inverse is straightforward to compute.

method is fast and scales well. It also compares favorably to the algorithm in Carriero, Clark, and Marcellino (2019), especially when n is large. For example, for a large BVAR with $n = 100$ variables, the proposed method takes about half a minute to obtain 10,000 posterior draws. In comparison, using the algorithm in Carriero, Clark, and Marcellino (2019) takes about 43 minutes.

Table 1: The computation times (in seconds) to obtain 10,000 posterior draws of $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ 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	1.3	6.8	28
CCM	58	238	2,574

2.4 The Marginal Likelihood

In this section we provide an analytical expression of the marginal likelihood. This closed-form expression is useful for a range of purposes, such as obtaining optimal hyperparamaters or designing efficient estimation algorithms for more flexible Bayesian VARs.

To prevent arithmetic underflow and overflow, we evaluate the marginal likelihood in log scale. Given the likelihood function in (3) and the asymmetric conjugate prior in (5), the associated log marginal likelihood of the VAR has the following analytical expression:

$$\begin{aligned}\log p(\mathbf{y}) = &-\frac{Tn}{2}\log(2\pi) + \sum_{i=1}^{n}\Bigg[-\frac{1}{2}\left(\log|\mathbf{V}_i| + \log|\mathbf{K}_{\boldsymbol{\theta}_i}|\right) \\ &+ \log\Gamma\left(\nu_i + \frac{T}{2}\right) + \nu_i \log S_i - \log\Gamma(\nu_i) - \left(\nu_i + \frac{T}{2}\right)\log\widehat{S}_i\Bigg].\end{aligned} \tag{11}$$

The details of the derivation are given in Appendix B. The above expression is straightforward to evaluate. We only note that to compute the log determinant $\log|\mathbf{K}_{\boldsymbol{\theta}_i}|$, it is numerically more stable to first compute its Cholesky factor $\mathbf{C}_{\mathbf{K}_{\boldsymbol{\theta}_i}}$ and return $2\sum \log c_{ii}$, where c_{ii} is the i-th diagonal element of the $\mathbf{C}_{\mathbf{K}_{\boldsymbol{\theta}_i}}$.

3 Extensions

In this section we briefly discuss how we can use the above analytical results and the efficient sampling scheme in more general settings. Suppose we augment our BVAR in (3) to the model $p(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma})$ with the additional parameter vector $\boldsymbol{\gamma}$. Further, consider the prior $p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}) = p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2 \mid \boldsymbol{\gamma})p(\boldsymbol{\gamma})$, where $p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2 \mid \boldsymbol{\gamma})$ is the asymmetric conjugate prior that could potentially depend on $\boldsymbol{\gamma}$ and the marginal prior $p(\boldsymbol{\gamma})$ is left unspecified for now. Before we discuss some efficient posterior samplers, we first give two examples that fit into this framework.

In our first example, we augment the BVAR by treating the hyperparameters κ_1 and κ_2 as parameters to be estimated. That is, $\boldsymbol{\gamma} = (\kappa_1, \kappa_2)'$. This extension is useful as it takes into account the parameter uncertainty of κ_1 and κ_2 (see also Giannone, Lenza, and Primiceri, 2015). This extension is considered in the empirical application. In our second example, we extend the BVAR by adding an MA(1) component to each equation:

$$
\begin{aligned}
y_{i,t} &= \mathbf{x}_{i,t}\boldsymbol{\theta}_i + \varepsilon^y_{i,t}, \\
\varepsilon^y_{i,t} &= u_{i,t} + \psi_i u_{i,t-1},
\end{aligned}
$$

where $u_{i,t} \sim \mathcal{N}(0, \sigma_i^2), t = 1, \ldots, T, i = 1, \ldots, n$. In this case, $\boldsymbol{\gamma} = (\psi_1, \ldots, \psi_n)'$. This extension is motivated by the empirical finding that allowing for moving average errors often improves forecast performance (see, e.g., Chan, 2013, 2018).

Both examples fit into the framework with likelihood $p(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma})$ and prior $p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma})$. One natural posterior sampler is to construct a Markov chain by sequentially sampling from $p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2 \mid \mathbf{y}, \boldsymbol{\gamma})$ and $p(\boldsymbol{\gamma} \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\sigma}^2)$. The first density is a product of normal-inverse-gamma densities, and we can efficiently obtain a draw from it as described before. The second density depends on the model, but it is often easy to sample from.[9]

Alternatively, a more efficient approach is the collapsed sampler that samples from $p(\boldsymbol{\gamma} \mid \mathbf{y})$. This sampling scheme is typically more efficient as it integrates out the high-dimensional parameters $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ analytically. The density $p(\boldsymbol{\gamma} \mid \mathbf{y})$ can be evaluated quickly

[9] For our first example with a low-dimensional $\boldsymbol{\gamma} = (\kappa_1, \kappa_2)'$, an independent-chain Metropolis-Hastings algorithm can be easily constructed. For our second example with $\boldsymbol{\gamma} = (\psi_1, \ldots, \psi_n)'$, it turns out that we can factor $p(\boldsymbol{\gamma} \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\sigma}^2) = p(\boldsymbol{\psi} \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\sigma}^2) = \prod_{i=1}^n p(\psi_i \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\sigma}^2)$. Then, each ψ_i can be simulated using the method in Chan (2013).

since

$$p(\boldsymbol{\gamma} \,|\, \mathbf{y}) \propto p(\mathbf{y} \,|\, \boldsymbol{\gamma}) p(\boldsymbol{\gamma}),$$

where $p(\mathbf{y} \,|\, \boldsymbol{\gamma})$ is the 'marginal likelihood' of the standard BVAR. For instance, for the first example with $\boldsymbol{\gamma} = (\kappa_1, \kappa_2)'$, the quantity $p(\mathbf{y} \,|\, \boldsymbol{\gamma})$ is exactly as the analytical expression given in (11). Finally, given the posterior draws of $\boldsymbol{\gamma}$, we can obtain the posterior draws of $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ from $p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2 \,|\, \mathbf{y}, \boldsymbol{\gamma})$.

4 Application: Forecasting with Large BVARs

We consider a forecasting exercise using large Bayesian VARs to illustrate the usefulness of the proposed asymmetric conjugate prior. After describing the macroeconomic dataset in Section 4.1, we first present the full sample results in Section 4.2, highlighting the empirical relevance of allowing for different levels of shrinkage on own lags and other lags. We then show in Section 4.3 that the proposed asymmetric conjugate prior leads to substantial gains in forecast performance over the symmetric prior that rules out cross-variable shrinkage.

4.1 Data

The dataset for our forecasting exercise consists of 21 US quarterly variables and the sample period is from 1959Q1 to 2018Q4. The dataset is constructed from the FRED-QD database at the Federal Reserve Bank of St. Louis as described in McCracken and Ng (2016). We use 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.

4.2 Full Sample Results

In this section we use the full sample to obtain the optimal hyperparameters κ_1 and κ_2 that maximize the log marginal likelihood given in (11). The optimal hyperparameters

and the associated log marginal likelihood are reported in Table 2. For comparison, we also consider two useful benchmarks. In the first case we set $\kappa_1 = \kappa_2 = \kappa$ and maximize the log marginal likelihood with respect to κ only. This benchmark mimics the standard practice of using the natural conjugate prior that does not distinguish between own lags and lags of other variables. We refer to this version as the symmetric prior. The second benchmark is a set of subjectively chosen values that apply cross-variable shrinkage. In particular, we follow Carriero, Clark, and Marcellino (2015) and consider $\kappa_1 = 0.04$ and $\kappa_2 = 0.0016$. This second benchmark is referred to as the subjective prior.

Table 2: Optimal values of the hyperparameters κ_1 and κ_2 under the symmetric prior ($\kappa_1 = \kappa_2$), the subjective prior (Carriero, Clark, and Marcellino, 2015) and the proposed asymmetric prior.

	Symmetric prior	Subjective prior	Asymmetric prior
κ_1	0.039	0.04	0.406
κ_2	0.039	0.0016	0.009
log-ML	−9,436	−9,372	−9,201

Under the symmetric prior with the restriction that $\kappa_1 = \kappa_2$, the optimal hyperparameter value is 0.039, which is very close a widely used value in the literature of 0.04 (e.g., Sims and Zha, 1998; Carriero, Clark, and Marcellino, 2015; Chan, 2018). However, if we allow κ_1 and κ_2 to be different, we obtain very different results: the optimal value for κ_1 increases more than ten-folds to 0.406, whereas the optimal value of κ_2 reduces to 0.009. These results suggest that the data prefers shrinking the coefficients on lags of other variables much more aggressively to zero than those on own lags. This makes intuitive sense as one would expect, on average, a variable's own lags would contain more information about its future evolution than lags of other variables. By relaxing the restriction that $\kappa_1 = \kappa_2$, the log marginal likelihood increases by 235. If we were to test the hypothesis that $\kappa_1 = \kappa_2$, this large difference in log marginal likelihood values would have decidedly reject it.

In addition, the optimal values of κ_1 and κ_2 under the asymmetric prior are also very different from those of the subjective prior. In particular, the values of κ_1 and κ_2 under the asymmetric prior are, respectively, 10 and 5.6 times larger than the latter. By selecting the values of κ_1 and κ_2 optimally, one can increase the log marginal likelihood value by 171. These results suggest that the subjective prior might have shrunk the coefficients

too aggressively.

We have so far taken the empirical Bayes approach of choosing hyperparameter values by maximizing the log marginal likelihood. A fully Bayesian approach would specify proper priors on κ_1 and κ_2 and obtain the corresponding posterior distribution. The latter approach has the additional advantage of being able to quantity parameter uncertainty of κ_1 and κ_2. In view of this, we take a fully Bayesian approach and treat κ_1 and κ_2 as parameters to be estimated. Specifically, we assume a uniform prior on the unit square $(0,1) \times (0,1)$ for κ_1 and κ_2, and compute the marginal posterior distribution of κ_1 and κ_2. The contour plot of the joint posterior density is reported in Figure 1.

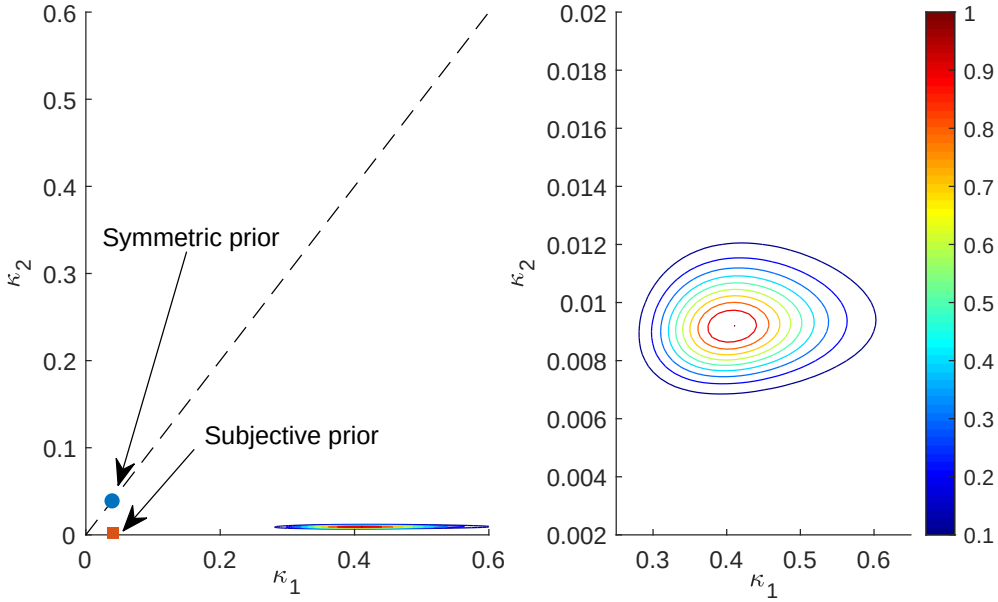

Figure 1: Contour plot of the joint posterior density of κ_1 and κ_2. The mode of the density is normalized to one for easy comparison.

As the contour plot in the right panel shows, most of the mass of κ_1 lies between 0.3 and 0.6, whereas the mass of κ_2 is mostly between 0.007 and 0.012. These results confirm the conclusion that one should shrink the coefficients on other lags much more aggressively to zero than those on own lags. Moreover, since there is virtually no mass along the diagonal line $\kappa_1 = \kappa_2$, requiring them to be the same as in the natural conjugate prior appears to be too restrictive. As a comparison, we also plot the values of κ_1 and κ_2 under the symmetric and subjective priors in the left panel. As is evident from the figure, the

values under both priors are far from the high-density region of the posterior distribution.

Overall, the full sample results indicate that the optimal hyperparameter values could be very different from some subjectively chosen values commonly used in empirical work. In addition, these results also highlight the importance of allowing for different levels of shrinkage on own versus other lags—and therefore the empirical relevance of the proposed asymmetric conjugate prior.[10]

4.3 Forecasting Results

In this section we evaluate the forecast performance of BVARs with the proposed asymmetric conjugate prior relative to two alternative priors: the symmetric prior that restricts $\kappa_1 = \kappa_2$ and the subjective prior that fixes $\kappa_1 = 0.04$ and $\kappa_2 = 0.0016$. Our sample period is from 1959Q1 to 2018Q4 and the evaluation period starts at 1985Q1 and runs 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 obtain the optimal hyperparameters in the symmetric and asymmetric priors by maximizing the marginal likelihood as given in (11). We evaluate both point and density forecasts. We use the conditional expectation $\mathbb{E}(y_{i,t+h} \,|\, \mathbf{y}_{1:t})$ as the h-step-ahead point forecast for variable i and the predictive density $p(y_{i,t+h} \,|\, \mathbf{y}_{1:t})$ as the corresponding density forecast.

The metric used to evaluate the point forecasts from model M is the root mean squared forecast error (RMSFE) defined as

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

where $y_{i,t+h}^{o}$ is the actual observed value of $y_{i,t+h}$. For RMSFE, a smaller value indicates better forecast performance. To evaluate the density forecasts, the metric we use is the

[10] We also compare the posterior estimates under this asymmetric conjugate prior on the structural-form parameters with those under the more standard independent normal and inverse-Wishart priors on the reduced-form parameters. The results are reported in Appendix D. The reduced-form estimates under both priors are mostly similar.

average of log predictive likelihoods (ALPL):

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

where $p(y_{i,t+h} = y_{i,t+h}^{\text{o}} \,|\, \mathbf{y}_{1:t})$ 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

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

and the percentage gains in terms of ALPL:

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

Figure 2 reports the forecasting results from the BVARs with the asymmetric conjugate prior relative to the benchmark symmetric prior. The top panel shows the percentage gains in RMSFE for all 21 variables, whereas the bottom panel presents the corresponding results in ALPL. For 1-step-ahead point forecasts, the asymmetric prior outperforms the benchmark for all variables except one. For a few variables, such as capacity utilization and 3-month treasury bill rate, the former outperforms the benchmark by more than 10%.

For 4-step-ahead point forecasts, the asymmetric 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 3.41% and 0.72%, respectively. Results for density forecasts are similar, though for two variables—a credit spread variable and PPI—the 4-step-ahead density forecasts under the asymmetric prior are noticeably worse than the benchmark. Overall, the asymmetric prior performs well: the median percentage gains in ALPL for 1- and 4-step-ahead forecasts are, respectively, 2.5% and 0.47%.

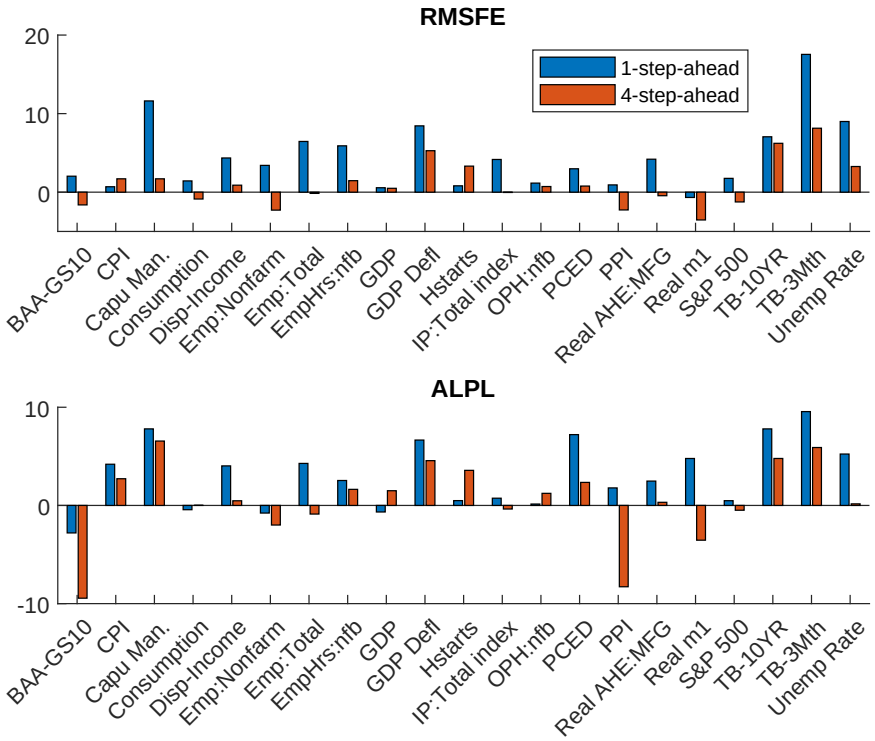

Figure 2: Forecasting results from BVAR with the asymmetric conjugate prior versus BVAR with the symmetric prior that assumes $\kappa_1 = \kappa_2$. The top panel shows the percentage gains in root mean squared forecast error of the asymmetric conjugate prior. The bottom panel presents the percentage gains in the average of log predictive likelihoods.

Next, we compare the forecast performance of the asymmetric prior with that of the subjective prior, and the results are reported in Figure 3. For the 1-step-ahead forecast horizon, the asymmetric prior substantially outperforms the subjective prior for a majority of variables for both point and density forecasts. The median percentage gains in RMSFE and ALPL are 2% and 3.6%, respectively.

For 4-step-ahead forecasts, the results are mixed. The asymmetric prior performs better than the benchmark for density forecast, but for point forecasts it is a bit worse. The median percentage gains in ALPL and RMSFE are 1.8% and -0.23%, respectively. This could reflect the fact that the hyperparameters under the asymmetric prior are chosen by maximizing the marginal likelihood—which can be interpreted as a one-step-ahead density forecast metric. Hence, the hyperparameters under the asymmetric prior are optimized for one-step-ahead forecast performance, and they do not necessary do as well

for longer forecast horizons compared to some subjectively chosen hyperparameter values.

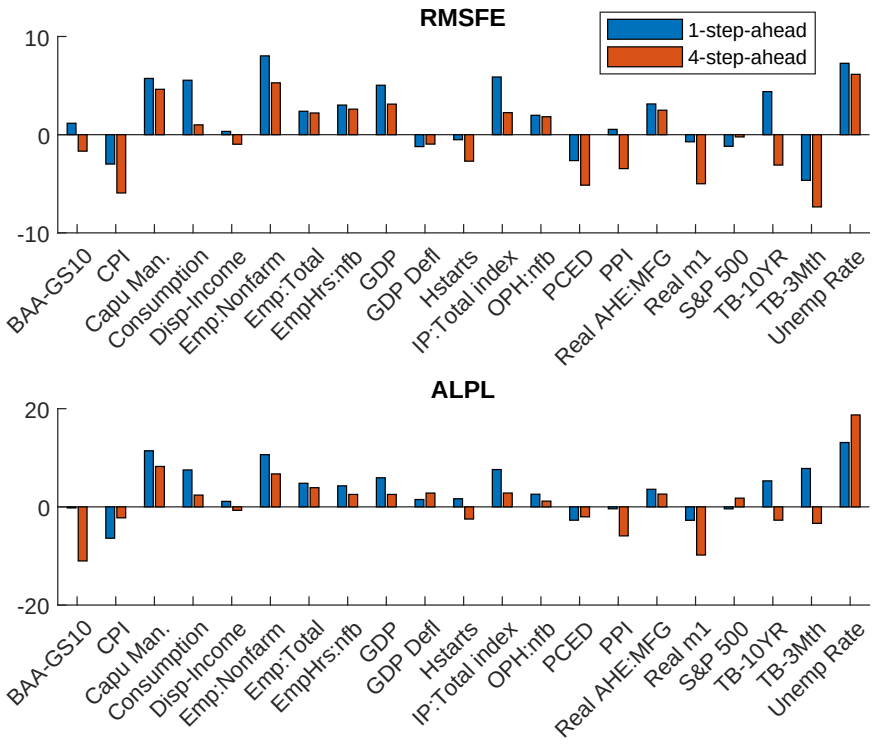

Figure 3: Forecasting results from BVAR with the asymmetric conjugate prior versus BVAR with the subjective prior that fixes $\kappa_1 = 0.04$ and $\kappa_2 = 0.0016$. The top panel shows the percentage gains in root mean squared forecast error of the asymmetric conjugate prior. The bottom panel presents the percentage gains in the average of log predictive likelihoods.

Lastly, we investigate if there are forecast performance gains by estimating the hyperparameters κ_1 and κ_2, as opposed to choosing their values by maximizing the marginal likelihood. In theory, the former approach is more desirable as it takes into account of parameter uncertainty. To that end, we compare the forecast performance of the two versions of the asymmetric conjugate prior. In the first version, we select the hyperparameters κ_1 and κ_2 by maximizing the marginal likelihood as before. In the second version, we estimate them using a uniform prior over the square $(0, 1) \times (0, 1)$.

The results are reported in Figure 4. For point forecasts, both versions have essentially the same performance. For density forecasts, they perform very similarly as well, though for a few variables the difference in ALPL can be as large as 4%. Hence, for computational

reason, the asymmetric prior in which the hyperparameters are fixed at optimal values might be a good default.

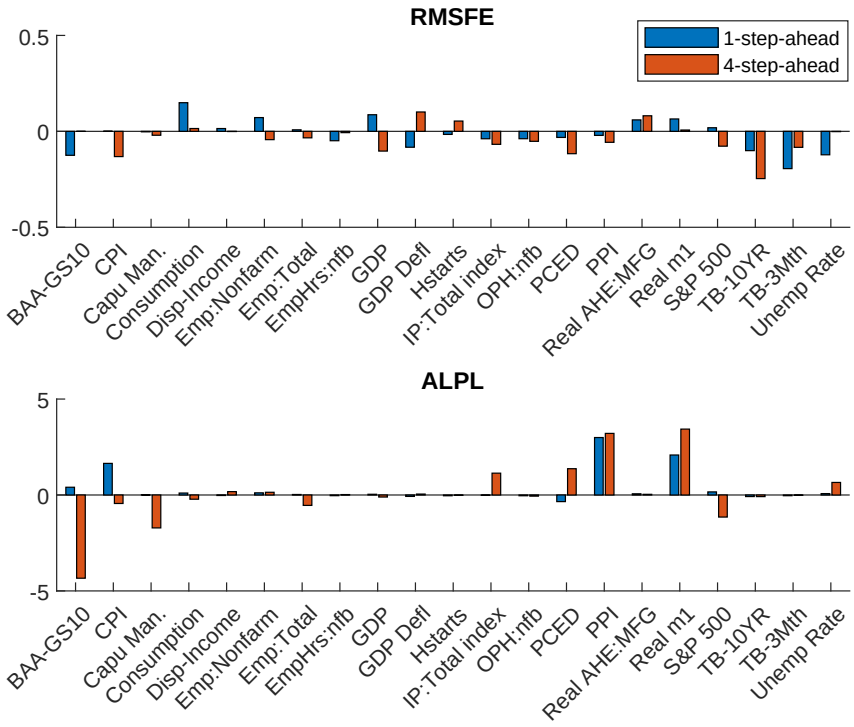

Figure 4: Forecasting results from the asymmetric conjugate prior where the hyperparameters are chosen by maximizing the marginal likelihood versus the version where the hyperparameters are estimated. The top panel shows the percentage gains in root mean squared forecast error of the asymmetric conjugate prior. The bottom panel presents the percentage gains in the average of log predictive likelihoods.

5 Concluding Remarks and Future Research

We have developed a new asymmetric conjugate prior for large BVARs that can accommodate cross-variable shrinkage, while maintaining many useful analytically results as the natural conjugate prior. Using a large US dataset, we demonstrated that the gains in forecast accuracy of this new prior can be substantial compared to convectional benchmarks. In particular, we showed that allowing for cross-variable shrinkage can lead to notable improvement in forecast performance. Our findings therefore highlight the

empirical relevance of this new asymmetric prior.

There is now a large empirical literature that shows that models with stochastic volatility tend to forecast substantially better (Clark, 2011; D'Agostino, Gambetti, and Giannone, 2013; Cross and Poon, 2016). In future work, it would be useful to develop similar efficient posterior samplers for large BVARs with stochastic volatility, such as the models in Eisenstat, Chan, and Strachan (2018) and Carriero, Clark, and Marcellino (2019). In addition, it would be interesting to use Proposition 1 to work out a way to construct a multivariate stochastic volatility model that is invariant to reordering of the variables.

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 21 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 the forecasting exercise.

Variable	Transformation
Real Gross Domestic Product	$400\Delta \log$
Personal Consumption Expenditures	$400\Delta \log$
Real Disposable Personal Income	$400\Delta \log$
Industrial Production Index	$400\Delta \log$
Capacity Utilization: Manufacturing (SIC)	no transformation
All Employees: Total nonfarm	$400\Delta \log$
Civilian Employment	$400\Delta \log$
Civilian Unemployment Rate	no transformation
Nonfarm Business Section: Hours of All Persons	$400\Delta \log$
Housing Starts: Total	$400\Delta \log$
Personal Consumption Expenditures: Chain-type Price index	$400\Delta \log$
Gross Domestic Product: Chain-type Price index	$400\Delta \log$
Consumer Price Index for All Urban Consumers: All Items	$400\Delta \log$
Producer Price Index for All commodities	$400\Delta \log$
Real Average Hourly Earnings of Production and Nonsupervisory Employees: Manufacturing, deflated by Core PCE	$400\Delta \log$
Nonfarm Business Section: Real Output Per Hour of All Persons	$400\Delta \log$
3-Month Treasury Bill: Secondary Market Rate	no transformation
10-Year Treasury Constant Maturity Rate	no transformation
Moody's Seasoned Baa Corporate Bond Yield Relative to Yield on 10-Year Treasury Constant Maturity	no transformation
Real M1 Money Stock	$400\Delta \log$
S&P's Common Stock Price Index : Composite	$400\Delta \log$

Appendix B: Derivation of the Marginal Likelihood

In this appendix we prove that the marginal likelihood of the VAR(p) under the asymmetric conjugate prior in (4) has the following expression:

$$p(\mathbf{y}) = \prod_{i=1}^{n} (2\pi)^{-\frac{T}{2}} |\mathbf{V}_i|^{-\frac{1}{2}} |\mathbf{K}_{\boldsymbol{\theta}_i}|^{-\frac{1}{2}} \frac{\Gamma(\nu_i + T/2) S_i^{\nu_i}}{\Gamma(\nu_i) \widehat{S}_i^{\nu_i + \frac{T}{2}}}.$$

This result follows from direct computation:

$$\begin{aligned}
p(\mathbf{y}) = \prod_{i=1}^{n} p(\mathbf{y}_i) &= \prod_{i=1}^{n} \int p(\boldsymbol{\theta}_i, \sigma_i^2) p(\mathbf{y}_i \,|\, \boldsymbol{\theta}_i, \sigma^2) \mathrm{d}(\boldsymbol{\theta}_i, \sigma_i^2) \\
&= \prod_{i=1}^{n} c_i (2\pi)^{-\frac{T}{2}} \int (\sigma_i^2)^{-\left(\nu_i + \frac{T+k_i}{2} + 1\right)} \mathrm{e}^{-\frac{1}{\sigma_i^2}\left(\widehat{S}_i + \frac{1}{2}(\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_i)' \mathbf{K}_{\boldsymbol{\theta}_i} (\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_i)\right)} \mathrm{d}(\boldsymbol{\theta}_i, \sigma_i^2) \\
&= \prod_{i=1}^{n} c_i (2\pi)^{-\frac{T}{2}} (2\pi)^{\frac{k_i}{2}} |\mathbf{K}_{\boldsymbol{\theta}_i}^{-1}|^{\frac{1}{2}} \frac{\Gamma(\nu_i + T/2)}{\widehat{S}_i^{\nu_i + \frac{T}{2}}} \\
&= \prod_{i=1}^{n} (2\pi)^{-\frac{T}{2}} |\mathbf{V}_i|^{-\frac{1}{2}} |\mathbf{K}_{\boldsymbol{\theta}_i}|^{-\frac{1}{2}} \frac{\Gamma(\nu_i + T/2) S_i^{\nu_i}}{\Gamma(\nu_i) \widehat{S}_i^{\nu_i + \frac{T}{2}}}.
\end{aligned}$$

where $c_i = (2\pi)^{-\frac{k_i}{2}} |\mathbf{V}_i|^{-\frac{1}{2}} S_i^{\nu_i} / \Gamma(\nu_i)$, $\mathbf{K}_{\boldsymbol{\theta}_i} = \mathbf{V}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i$, $\widehat{\boldsymbol{\theta}}_i = \mathbf{K}_{\boldsymbol{\theta}_i}^{-1} (\mathbf{V}_i^{-1} \mathbf{m}_i + \mathbf{X}_i' \mathbf{y}_i)$ and $\widehat{S}_i = S_i + (\mathbf{y}_i' \mathbf{y}_i + \mathbf{m}_i' \mathbf{V}_i^{-1} \mathbf{m}_i - \widehat{\boldsymbol{\theta}}_i' \mathbf{K}_{\boldsymbol{\theta}_i} \widehat{\boldsymbol{\theta}}_i)/2$. In the above derivation we have used the fact that

$$\int (\sigma_i^2)^{-\left(\nu_i + \frac{T+k_i}{2} + 1\right)} \mathrm{e}^{-\frac{1}{\sigma_i^2}\left(\widehat{S}_i + \frac{1}{2}(\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_i)' \mathbf{K}_{\boldsymbol{\theta}_i} (\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_i)\right)} = (2\pi)^{\frac{k_i}{2}} |\mathbf{K}_{\boldsymbol{\theta}_i}^{-1}|^{\frac{1}{2}} \frac{\Gamma(\nu_i + T/2)}{\widehat{S}_i^{\nu_i + \frac{T}{2}}}.$$

The above equality holds because the quantity on the right-hand side is the normalizing constant of the $(\boldsymbol{\theta}_i, \sigma_i^2) \sim \mathcal{NIG}(\mathbf{m}_i, \mathbf{V}_i, \nu_i, S_i)$ distribution.

Appendix C: Proofs of Proposition and Corollaries

This appendix provides proofs of Proposition 1 and two corollaries in the main text. We first record in the following lemma the determinant of the Jacobian of transformation from the structural-form parameterization to the reduced-form parameterization.[11] This lemma was proved in Chan and Jeliazkov (2009), and we include it here for convenience. The proof uses the differential forms approach that is equivalent to calculating the Jacobian (see, e.g., Theorem 2.1.1 in Muirhead, 1982).

Lemma 1. Suppose $\mathbf{W}$ is a $n \times n$ positive definite matrix and $\mathbf{W} = \mathbf{T}'\widetilde{\mathbf{T}}\mathbf{T}$, where $\mathbf{T}$ is a lower triangular matrix with ones on the main diagonal and $\widetilde{\mathbf{T}}$ is a diagonal matrix with positive diagonal elements. Denote the lower diagonal elements of $\mathbf{T}$ by $t_{ij},\ 1 \leqslant j < i \leqslant n$, and the diagonal elements of $\widetilde{\mathbf{T}}$ by $t_{ii},\ i = 1, \ldots, n,$. Let $(\mathrm{d}\mathbf{W})$ denote the differential form $(\mathrm{d}\mathbf{W}) \equiv \bigwedge_{i \geqslant j} \mathrm{d}w_{ij}$ and similarly define $(\mathrm{d}\mathbf{T}) \equiv \bigwedge_{i \geqslant j} \mathrm{d}t_{ij}$. Then we have

$$(\mathrm{d}\mathbf{W}) = \prod_{i=1}^{n} t_{ii}^{i-1} (\mathrm{d}\mathbf{T}).$$

In other words, the determinant of the Jacobian of the transformation from $\mathbf{T}'\widetilde{\mathbf{T}}\mathbf{T}$ to $\mathbf{W}$ is $\prod_{i=1}^{n} t_{ii}^{-i+1}$.

Proof of the lemma: By the definition $\mathbf{W} = \mathbf{T}'\widetilde{\mathbf{T}}\mathbf{T}$, we have

$$\begin{pmatrix} w_{11} & w_{21} & \ldots & w_{n1} \\ w_{21} & w_{22} & \ldots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \ldots & w_{nn} \end{pmatrix} = \begin{pmatrix} 1 & t_{21} & \ldots & t_{n1} \\ 0 & 1 & \ldots & t_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} \begin{pmatrix} t_{11} & 0 & \ldots & 0 \\ 0 & t_{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & t_{nn} \end{pmatrix} \begin{pmatrix} 1 & 0 & \ldots & 0 \\ t_{21} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \ldots & 1 \end{pmatrix}.$$

[11] For a different structural-form parameterization where the error covariance matrix is the identity matrix and the impact matrix is triangular with free diagonal elements, Zha (1999) derives the determinant of the Jacobian of transformation from the reduced-form parameterization.

Hence, we can express each w_{ij} in terms of $\{t_{ij}\}$:

$$w_{ii} = t_{ii} + \sum_{j=i+1}^{n} t_{ji}^2 t_{jj}, \quad i = 1, \ldots, n, \tag{12}$$

$$w_{ij} = t_{ij} t_{ii} + \sum_{k=i+1}^{n} t_{ki} t_{kj} t_{kk}, \quad 1 \leqslant j < i \leqslant n. \tag{13}$$

Next, we take differentials of these two equations so that we can write the differential form ($\mathrm{d}\mathbf{W}$) in terms of ($\mathrm{d}\mathbf{T}$). Since we are going to take the exterior product of these differentials and the exterior products of repeated differentials are zero, there is no need to keep track of differentials in t_{ij} that have previously occurred. Therefore, we take differentials of (12) and (13) and ignore those that have previously occurred:

$$\begin{aligned}
\mathrm{d}w_{nn} &= \mathrm{d}t_{nn} \\
\mathrm{d}w_{n,n-1} &= \mathrm{d}t_{nn}\mathrm{d}t_{n,n-1} + \ldots \\
&\vdots \\
\mathrm{d}w_{n1} &= t_{nn}\mathrm{d}t_{n1} + \ldots \\
\mathrm{d}w_{n-1,n-1} &= \mathrm{d}t_{n-1,n-1} + \ldots \\
&\vdots \\
\mathrm{d}w_{11} &= \mathrm{d}t_{11} + \ldots
\end{aligned}$$

Finally, taking exterior products gives

$$\bigwedge_{i \geqslant j} \mathrm{d}w_{ij} = t_{nn}^{n-1} t_{n-1,n-1}^{n-2} \cdots t_{22} \bigwedge_{i \geqslant j} \mathrm{d}t_{ij}$$

as claimed. □

Proof of Proposition 1: Assume the same notation as in Lemma 1. To prove Proposition 1, we consider the case where

$$t_{ii} \sim \mathcal{G}\left(\frac{\nu_0 + i - n}{2}, \frac{s_i^2}{2}\right), \quad i = 1, \ldots, n, \tag{14}$$

$$(t_{ij} \,|\, t_{ii}) \sim \mathcal{N}\left(0, \frac{t_{ii}^{-1}}{s_j^2}\right), \quad 1 \leqslant j < i \leqslant n,\ i = 2, \ldots, n. \tag{15}$$

More specifically, we will show that the density of $\mathbf{W} = \mathbf{T}'\widetilde{\mathbf{T}}\mathbf{T}$ is the same as that of the Wishart distribution $\mathcal{W}(\nu, \mathbf{S}^{-1})$, where $\mathbf{S} = \text{diag}(s_1^2, \ldots, s_n^2)$. Then, if we let $t_{ii} = 1/\sigma_i^2$ and $A_{i,j} = t_{ij}$, we have $\widetilde{\mathbf{\Sigma}}^{-1} = \mathbf{A}'\mathbf{\Sigma}^{-1}\mathbf{A} \sim \mathcal{W}(\nu_0, \mathbf{S}^{-1})$.

To prove the proposition, we first compute the determinant of $\mathbf{W}$ and the trace $\text{tr}(\mathbf{SW})$. Since the determinant of $\mathbf{T}$ is 1, we have

$$|\mathbf{W}| = |\widetilde{\mathbf{T}}| = \prod_{i=1}^{n} t_{ii}.$$

Next, using (12), we have

$$\begin{aligned}
\text{tr}(\mathbf{SW}) &= \sum_{i=1}^{n} w_{ii} s_i^2 \\
&= \sum_{i=1}^{n} t_{ii} s_i^2 + \sum_{i=1}^{n} \sum_{j=i+1}^{n} t_{ji}^2 t_{jj} s_i^2 \\
&= \sum_{i=1}^{n} t_{ii} s_i^2 + \sum_{j=2}^{n} \sum_{i=1}^{j-1} t_{ji}^2 t_{jj} s_i^2 \\
&= \sum_{i=1}^{n} t_{ii} s_i^2 + \sum_{i=2}^{n} \sum_{j=1}^{i-1} t_{ij}^2 t_{ii} s_j^2,
\end{aligned}$$

where we change the order of the double summations in the third equality and interchange the dummy indices i and j in the last equality.

Now, it follows from the distributional assumptions in (14) and (15) that the kernel of the joint density of $\mathbf{T}$ and $\widetilde{\mathbf{T}}$ is

$$\begin{aligned}
&\prod_{i=1}^{n} t_{ii}^{\frac{\nu_0+i-n}{2}-1} \text{e}^{-\frac{s_i^2}{2} t_{ii}} \times \prod_{i=2}^{n} t_{ii}^{\frac{i-1}{2}} \text{e}^{-\frac{1}{2}\sum_{j=1}^{i-1} t_{ij}^2 t_{ii} s_j^2} \\
&= \left(\prod_{i=1}^{n} t_{ii}^{\frac{\nu_0-n-1}{2}+(i-1)} \right) \text{e}^{-\frac{1}{2}\left(\sum_{i=1}^{n} t_{ii} s_i^2 + \sum_{i=2}^{n} \sum_{j=1}^{i-1} t_{ij}^2 t_{ii} s_j^2\right)}.
\end{aligned}$$

Next, we derive the kernel of the density of $\mathbf{W}$. By the lemma, the determinant of the Jacobian is $\prod_{i=1}^{n} t_{ii}^{-i+1}$. Substituting $\text{tr}(\mathbf{W})$ and $|\mathbf{W}|$ into the above expression and

multiplying the determinant of the Jacobian, we obtain the kernel of the density of $\mathbf{W}$:

$$|\mathbf{W}|^{\frac{\nu_0-n-1}{2}}\mathrm{e}^{-\frac{1}{2}\mathrm{tr}(\mathbf{SW})},$$

which is the kernel of the Wishart density $\mathcal{W}(\nu_0, \mathbf{S}^{-1})$. □

Proof of Corollary 1: Here we use the same notation as in Proposition 1. Assume $\mathbf{W} \sim \mathcal{W}(\nu, \mathbf{S}^{-1})$, and let $\mathbf{W} = \mathbf{T}'\widetilde{\mathbf{T}}\mathbf{T}$, where $\mathbf{T}$ and $\widetilde{\mathbf{T}}$ are given in Lemma 1. If we can show that t_{ii} and $(t_{ij} \,|\, t_{ii})$ follow the same normal-gamma distributions given in (14) and (15), respectively, then we are done. Since the transformation between $\mathbf{W}$ and $\mathbf{T}'\widetilde{\mathbf{T}}\mathbf{T}$ is one-to-one, the proof essentially just "reverses" the equalities given in Proposition 1. More specifically, in the proof of Proposition 1 we showed that $|\mathbf{W}| = \prod_{i=1}^{n} t_{ii}$ and

$$\mathrm{tr}(\mathbf{SW}) = \sum_{i=1}^{n} t_{ii}s_i^2 + \sum_{i=2}^{n}\sum_{j=1}^{i-1} t_{ij}^2 t_{ii}s_j^2.$$

Also, by Lemma 1, the determinant of the Jacobian of transformation is $\prod_{i=1}^{n} t_{ii}^{i-1}$. Hence, the kernel of the joint distribution of $t_{ii}, i = 1, \ldots, n$, and $t_{ij}, 1 \leqslant j < i \leqslant n,\ i = 2, \ldots, n$ is given by:

$$\begin{aligned}
&|\mathbf{W}|^{\frac{\nu_0-n-1}{2}}\mathrm{e}^{-\frac{1}{2}\mathrm{tr}(\mathbf{SW})} \times \prod_{i=1}^{n} t_{ii}^{i-1} \\
&\quad = \left(\prod_{i=1}^{n} t_{ii}^{\frac{\nu_0-n-1}{2}+(i-1)}\right)\mathrm{e}^{-\frac{1}{2}\left(\sum_{i=1}^{n} t_{ii}s_i^2 + \sum_{i=2}^{n}\sum_{j=1}^{i-1} t_{ij}^2 t_{ii}s_j^2\right)} \\
&\quad = \prod_{i=1}^{n} t_{ii}^{\frac{\nu_0+i-n}{2}-1}\mathrm{e}^{-\frac{s_i^2}{2}t_{ii}} \times \prod_{i=2}^{n} t_{ii}^{\frac{i-1}{2}}\mathrm{e}^{-\frac{1}{2}\sum_{j=1}^{i-1} t_{ij}^2 t_{ii}s_j^2}.
\end{aligned}$$

It follows that $t_{ii}, i = 1, \ldots, n$, are independent gamma random variables given in (14). Moreover, conditional on t_{ii}, $t_{ij}, 1 \leqslant j < i$, are independent normal variables given in (15).

Proof of Corollary 2: Suppose $\widetilde{\mathbf{\Sigma}} \sim \mathcal{IW}(\nu_0, \mathbf{R})$, where $\mathbf{R}$ is a symmetric positive definite matrix. Factor $\mathbf{R}^{-1} = \mathbf{L}'\mathbf{S}^{-1}\mathbf{L}$, where $\mathbf{L}$ is lower triangular with ones on the main diagonal and $\mathbf{S}$ is diagonal. Since $\widetilde{\mathbf{\Sigma}}^{-1} \sim \mathcal{W}(\nu_0, \mathbf{R}^{-1})$, by the properties of the Wishart distribution, we have $(\mathbf{L}')^{-1}\widetilde{\mathbf{\Sigma}}^{-1}\mathbf{L}^{-1} \sim \mathcal{W}(\nu_0, \mathbf{S}^{-1})$. Now, applying Corollary 1, we obtain $(\mathbf{L}')^{-1}\widetilde{\mathbf{\Sigma}}^{-1}\mathbf{L}^{-1} = \mathbf{A}'\mathbf{\Sigma}^{-1}\mathbf{A}$, where $\mathbf{A}$ is lower triangular with ones on the main diagonal and $\mathbf{\Sigma}$ is diagonal. The diagonal elements of $\mathbf{\Sigma}$ and the lower triangular elements

of $\mathbf{A}$ follow the normal-inverse-gamma distributions:

$$\sigma_i^2 \sim \mathcal{IG}\left(\frac{\nu_0 + i - n}{2}, \frac{s_i^2}{2}\right), \quad i = 1, \ldots, n,$$
$$(A_{i,j} \,|\, \sigma_i^2) \sim \mathcal{N}\left(0, \frac{\sigma_i^2}{s_j^2}\right), \quad 1 \leqslant j < i \leqslant n,\ i = 2, \ldots, n.$$

Letting $\mathbf{C} = \mathbf{AL}$, we can write $\widetilde{\mathbf{\Sigma}}^{-1} = \mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C}$. Since both $\mathbf{A}$ and $\mathbf{L}$ are lower triangular with ones on the main diagonal, so is $\mathbf{C}$. It remains to show that $\mathbf{c}_i$, the free elements of the i-th row of $\mathbf{C}$, follows the normal distribution in (9). Since $\mathbf{C}' = \mathbf{L}'\mathbf{A}'$, we can write $\mathbf{c}_i$ in terms of $\mathbf{A}$ and $\mathbf{L}$ as:

$$\mathbf{c}_i = \mathbf{l}_i + \mathbf{L}'_{1:i-1}\mathbf{a}_i,$$

where $\mathbf{l}_i$ and $\mathbf{a}_i$ are respectively the free elements of the i-th row of $\mathbf{L}$ and $\mathbf{A}$, and $\mathbf{L}_{1:i-1}$ is the $(i-1) \times (i-1)$ matrix that consists of the first $(i-1)$ rows and columns of $\mathbf{L}$. Since $\mathbf{c}_i$ is an affine transformation of the normal vector $\mathbf{a}_i$, conditional on σ_i^2, $\mathbf{c}_i$ is normally distributed with mean vector $\mathbf{l}_i$ and covariance matrix $\sigma_i^2\mathbf{L}'_{1:i-1}\mathbf{S}^{-1}_{1:i-1}\mathbf{L}_{1:i-1}$, where $\mathbf{S}_{1:i-1}$ is the submatrix consisting of the first $(i-1)$ rows and columns of $\mathbf{S} = \operatorname{diag}(s_1^2, \ldots, s_n^2)$.

Appendix D: Comparison with Independent Normal and Inverse-Wishart Prior

This appendix compares the posterior estimates under the asymmetric conjugate prior on the structural-form parameters with those under the independent normal and inverse-Wishart priors on the reduced-form parameters. More specifically, we first obtain posterior draws of the structural-form parameters under the asymmetric conjugate prior, and transform them into the reduced-form parameters $\widetilde{\boldsymbol{\Sigma}}$ and $\widetilde{\boldsymbol{\beta}} = \text{vec}([\widetilde{\mathbf{b}}, \widetilde{\mathbf{B}}_1, \ldots, \widetilde{\mathbf{B}}_p]')$. Then, we obtain posterior draws of $\widetilde{\boldsymbol{\Sigma}}$ and $\widetilde{\boldsymbol{\beta}}$ under the independent priors $\widetilde{\boldsymbol{\Sigma}} \sim \mathcal{IW}(\nu_0, \mathbf{S})$ and $\widetilde{\boldsymbol{\beta}} \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$, where $\mathbf{V} = \text{diag}(\mathbf{V}_1, \ldots, \mathbf{V}_n)$. The results are reported in Figure 5. The two priors give almost identical estimates of $\widetilde{\boldsymbol{\Sigma}}$. Moreover, the estimates of $\widetilde{\boldsymbol{\beta}}$ are also very similar.

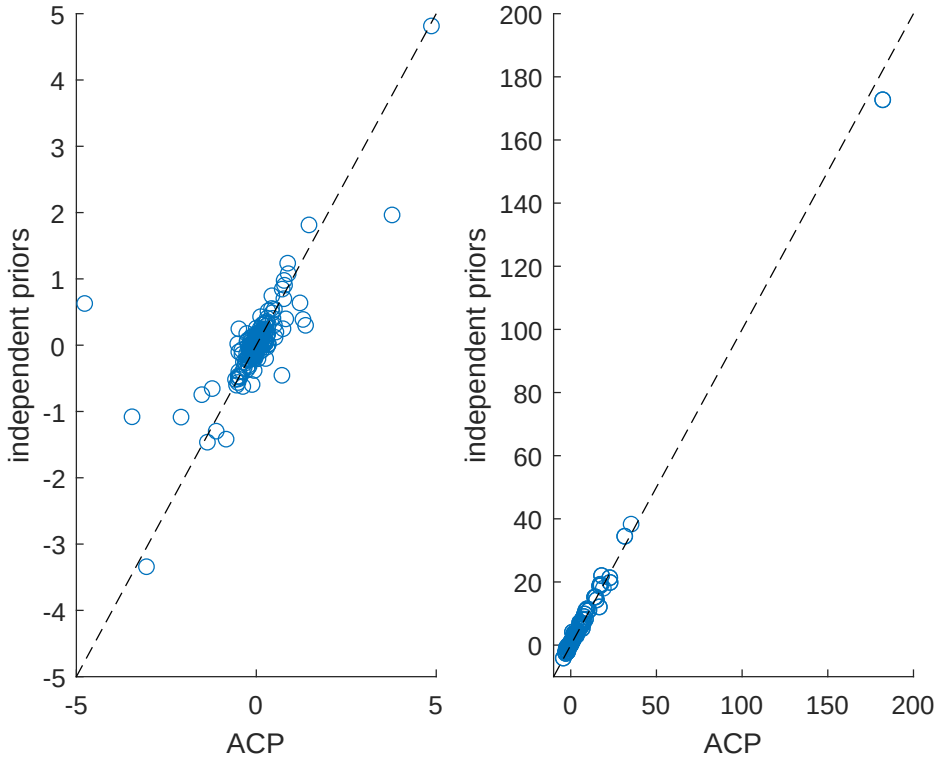

Figure 5: Scatter plots of the VAR coefficients (left panel) and the free elements of the covariance matrix (right panel).

References

Ando, T., and A. Zellner (2010): "Hierarchical Bayesian analysis of the seemingly unrelated regression and simultaneous equations models using a combination of direct Monte Carlo and importance sampling techniques," *Bayesian Analysis*, 5(1), 65–95.

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

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.

Chan, J. C. C. (2013): "Moving Average Stochastic Volatility Models with Application to Inflation Forecast," *Journal of Econometrics*, 176(2), 162–172.

——— (2018): "Large Bayesian VARs: A Flexible Kronecker Error Covariance Structure," *Journal of Business and Economic Statistics*, forthcoming.

——— (2019): "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): "MCMC Estimation of Restricted Covariance Matrix," *Journal of Computational and Graphical Statistics*, 18, 457–480.

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.

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.

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

Dieppe, A., R. Legrand, and B. Van Roye (2016): "The BEAR toolbox," *ECB Working Paper 1934*.

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

Eisenstat, E., J. C. C. Chan, and R. W. Strachan (2016): "Stochastic Model Specification Search for Time-Varying Parameter VARs," *Econometric Reviews*, 35(8-10), 1638–1665.

——— (2018): "Reducing Dimensions in a Large TVP-VAR," *Working Paper series 18-37, Rimini Centre for Economic Analysis*.

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

Kadiyala, K., and S. Karlsson (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.

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.

Muirhead, R. J. (1982): *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Inc.

Schorfheide, F., and D. Song (2015): "Real-Time Forecasting With a Mixed-Frequency VAR," *Journal of Business and Economic Statistics*, 33(3), 366–380.

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

Zellner, A. (1971): *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York.

Zha, T. (1999): "Block recursion and structural vector autoregressions," *Journal of Econometrics*, 90(2), 291–316.