Large Bayesian Vector Autoregressions

Joshua C. C. Chan

Abstract Bayesian vector autoregressions are widely used for macroeconomic forecasting and structural analysis. Until recently, however, most empirical work had considered only small systems with a few variables due to parameter proliferation concern and computational limitations. We first review a variety of shrinkage priors that are useful for tackling the parameter proliferation problem in large Bayesian VARs, followed by a detailed discussion of efficient sampling methods for overcoming the computational problem. We then give an overview of some recent models that incorporate various important model features into conventional large Bayesian VARs, including stochastic volatility, non-Gaussian and serially correlated errors. Efficient estimation methods for fitting these more flexible models are also discussed. These models and methods are illustrated using a forecasting exercise that involves a real-time macroeconomic dataset. The corresponding MATLAB code is also provided.[1]

1 Introduction

Vector autoregressions (VARs) are the workhorse models for empirical macroeconomics. They were introduced to economics by Sims (1980), and have since been widely adopted for macroeconomic forecasting and structural analysis. Despite their simple formulation, VARs tend to forecast well, and are used as the benchmark for comparing forecast performance of new models and methods. They are also used to better understand the impacts of structural shocks on key macroeconomic variables through the estimation of impulse response functions.

VARs tend to have a lot of parameters. Early work by Doan, Litterman, and Sims (1984) and Litterman (1986) on Bayesian methods that formally incorporate

Joshua C. C. Chan
Purdue University and UTS, e-mail: joshuacc.chan@gmail.com

[1] MATLAB code is available at `http://joshuachan.org/`

non-data information into informative priors are often found to greatly improve forecast performance. However, until recently, most empirical work had considered only small systems that rarely include more than a few dependent variables.

This has changed since the seminal work of Banbura, Giannone, and Reichlin (2010), who find that large Bayesian VARs with more than two dozens dependent variables forecast better than small VARs. This has generated a rapidly expanding literature on using large Bayesian VARs for forecasting and structural analysis; recent papers include Carriero, Kapetanios, and Marcellino (2009), Koop (2013) and Carriero, Clark, and Marcellino (2015a). Large Bayesian VARs thus provide an alternative to factor models that are traditionally used to handle large datasets (e.g., Stock and Watson, 2002; Forni, Hallin, Lippi, and Reichlin, 2003).

There are by now many extensions of small VARs that take into account salient features of macroeconomic data, the most important of which being time-varying volatility (Cogley and Sargent, 2005; Primiceri, 2005). How best to construct large VARs with time-varying volatility is an active research area, and has generated many new approaches, such as Koop and Korobilis (2013), Carriero, Clark, and Marcellino (2015b, 2016) and Chan (2018).

There are two key challenges in estimating large VARs. First, large VARs typically have far more parameters than observations. Without appropriate shrinkage or regularization, parameter uncertainty would make forecasts or any analysis unreliable. Second, estimation of large VARs involves manipulating large matrices and is typically computationally intensive. These two challenges are exacerbated when we extend large VARs to allow for more flexible error covariance structures, such as time-varying volatility.

In what follows, we first study methods to tackle these two challenges in the context of large homoscedastic VARs. We will then discuss a few recent models that incorporate stochastic volatility into large VARs and the associated estimation methods.

1.1 Vector Autoregressions

We first consider a standard homoscedastic VAR of order p. Let $\mathbf{y}_t = (y_{1t}, \ldots, y_{nt})'$ denote the $n \times 1$ vector of dependent variables at time t. Then, the basic VAR(p) is given by:

$$\mathbf{y}_t = \mathbf{b} + \mathbf{A}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{A}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\mathbf{b}$ is an $n \times 1$ vector of intercepts, $\mathbf{A}_1, \ldots, \mathbf{A}_p$ are $n \times n$ coefficient matrices and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \Sigma)$. In other words, the VAR(p) is simply a multiple-equation regression where the regressors are the lagged dependent variables. Specifically, there are n equations and each equation has $k = np + 1$ regressors—so there are a total of $nk = n^2 p + n$ VAR coefficients. With typical quarterly data, the number of VAR coefficients can be more than the number of observations when n is large.

The model in (1) runs from $t = 1$ to $t = T$, and it depends on the p initial conditions $\mathbf{y}_{-p+1}, \ldots, \mathbf{y}_0$. In principle these initial conditions can be modeled explicitly. Here all the analysis is done conditioned on these initial conditions. If the series is not too short, both approaches typically give similar results.

There are two common ways to stack the VAR(p) in (1) over $t = 1, \ldots, T$. In the first representation, we rewrite the VAR(p) as:

$$\mathbf{y}_t = \mathbf{X}_t \beta + \varepsilon_t,$$

where $\mathbf{X}_t = \mathbf{I}_n \otimes [1, \mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p}]$ with $\otimes$ denoting the Kronecker product and $\beta = \text{vec}([\mathbf{b}, \mathbf{A}_1, \ldots, \mathbf{A}_p]')$—i.e., the intercepts and VAR coefficient matrices are stacked by rows into a $nk \times 1$ vector. Furthermore, stacking $\mathbf{y} = (\mathbf{y}'_1, \ldots, \mathbf{y}'_T)'$, we obtain

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{2}$$

where $\mathbf{X} = (\mathbf{X}'_1, \ldots, \mathbf{X}'_T)'$ is a $Tn \times nk$ matrix of regressors and $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_T \otimes \Sigma)$.

In the second representation, we first stack the dependent variables into a $T \times n$ matrix $\mathbf{Y}$ so that its t-th row is $\mathbf{y}'_t$. Now, let $\mathbf{Z}$ be a $T \times k$ matrix of regressors, where the t-th row is $\mathbf{x}'_t = (1, \mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p})$. Next, let $\mathbf{A} = (\mathbf{b}, \mathbf{A}_1, \ldots, \mathbf{A}_p)'$ denote the $k \times n$ matrix of VAR coefficients. Then, we can write the VAR(p) as follows:

$$\mathbf{Y} = \mathbf{Z}\mathbf{A} + \mathbf{U}, \tag{3}$$

where $\mathbf{U}$ is a $T \times n$ matrix of innovations in which the t-th row is ε'_t. In terms of the first representation in (2), $\mathbf{y} = \text{vec}(\mathbf{Y}')$, $\beta = \text{vec}(\mathbf{A})$ and $\varepsilon = \text{vec}(\mathbf{U}')$. It follows that

$$\text{vec}(\mathbf{U}) \sim \mathcal{N}(\mathbf{0}, \Sigma \otimes \mathbf{I}_T). \tag{4}$$

1.2 Likelihood Functions

Next we derive the likelihood functions implied by the two equivalent representations of the VAR(p), namely (2) and (3).

Using the first representation of the VAR(p) in (2), we have

$$(\mathbf{y} \,|\, \beta, \Sigma) \sim \mathcal{N}(\mathbf{X}\beta, \mathbf{I}_T \otimes \Sigma).$$

Therefore, the likelihood function is given by:

$$\begin{aligned} p(\mathbf{y} \,|\, \beta, \Sigma) &= (2\pi)^{-\frac{Tn}{2}} |(\mathbf{I}_T \otimes \Sigma)|^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{I}_T \otimes \Sigma)^{-1}(\mathbf{y}-\mathbf{X}\beta)} \\ &= (2\pi)^{-\frac{Tn}{2}} |\Sigma|^{-\frac{T}{2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{I}_T \otimes \Sigma^{-1})(\mathbf{y}-\mathbf{X}\beta)}, \end{aligned} \tag{5}$$

where the second equality holds because $|\mathbf{I}_T \otimes \Sigma| = |\Sigma|^T$ and $(\mathbf{I}_T \otimes \Sigma)^{-1} = \mathbf{I}_T \otimes \Sigma^{-1}$.

Since the two representations of the VAR(p) are equivalent, the likelihood implied by (3) should be the same as in (5). In what follows we rewrite (5) in terms of $\mathbf{Y}$, $\mathbf{Z}$ and $\mathbf{A}$. To do that, we need the following results: for conformable matrices $\mathbf{B},\mathbf{C},\mathbf{D}$, we have

$$\text{vec}(\mathbf{BCD}) = (\mathbf{D}' \otimes \mathbf{B})\text{vec}(\mathbf{C}), \tag{6}$$

$$\text{tr}(\mathbf{B}'\mathbf{C}) = \text{vec}(\mathbf{B})'\text{vec}(\mathbf{C}), \tag{7}$$

$$\text{tr}(\mathbf{BCD}) = \text{tr}(\mathbf{CDB}) = \text{tr}(\mathbf{DBC}), \tag{8}$$

where $\text{tr}(\cdot)$ is the trace function.

Noting that $\mathbf{y} - \mathbf{X}\beta = \varepsilon = \text{vec}(\mathbf{U}')$, we now rewrite the quadratic form in (5) as

$$\begin{aligned}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{I}_T \otimes \Sigma^{-1})(\mathbf{y} - \mathbf{X}\beta) &= \text{vec}(\mathbf{U}')'(\mathbf{I}_T \otimes \Sigma^{-1})\text{vec}(\mathbf{U}') \\ &= \text{vec}(\mathbf{U}')'\text{vec}(\Sigma^{-1}\mathbf{U}') \\ &= \text{tr}(\mathbf{U}\Sigma^{-1}\mathbf{U}') \\ &= \text{tr}(\Sigma^{-1}\mathbf{U}'\mathbf{U}),\end{aligned}$$

where the second equality holds because of (6); the third equality holds because of (7); and the last equality holds because of (8). Using this representation of the quadratic form and $\mathbf{U} = \mathbf{Y} - \mathbf{ZA}$, the likelihood implied by the second representation in (3) is therefore given by

$$p(\mathbf{Y} \,|\, \mathbf{A}, \Sigma) = (2\pi)^{-\frac{Tn}{2}} |\Sigma|^{-\frac{T}{2}} \mathrm{e}^{-\frac{1}{2}\text{tr}\left(\Sigma^{-1}(\mathbf{Y}-\mathbf{ZA})'(\mathbf{Y}-\mathbf{ZA})\right)}. \tag{9}$$

2 Priors for Large Bayesian VARs

What makes Bayesian VARs Bayesian is the use of informative priors that incorporate non-data information. As mentioned in the introduction, VARs tend to have a lot of parameters, and large VARs exacerbate this problem. For example, a VAR(4) with $n = 20$ dependent variables has 1,620 VAR coefficients, which is much larger than the number of observations in typical quarterly datasets. Without informative priors or regularization, it is not even possible to estimate the VAR coefficients.

In this section we discuss a range of informative priors that are found useful in the context of large VARs. One common feature of these priors is that they aim to "shrink" an unrestricted VAR to one that is parsimonious and seemingly reasonable. These priors differ in how they achieve this goal, and whether they lead to analytical results or simpler Markov chain Monte Carlo (MCMC) algorithms for estimating the posterior distributions. In addition, they also differ in how easily they can be applied to more flexible VARs, such as VARs with stochastic volatility.

2.1 The Minnesota Prior

Shrinkage priors in the context of small VARs are first developed by Doan, Litterman, and Sims (1984) and Litterman (1986). Due to their affiliations with the University of Minnesota and the Federal Reserve Bank of Minneapolis at that time, this family of priors is commonly called Minnesota priors. It turns out that Minnesota priors can be directly applied to large VARs. This approach uses an approximation that leads to substantial simplifications in prior elicitation. Below we present a version discussed in Koop and Korobilis (2010).

To introduce the Minnesota priors, we use the first representation of the VAR with likelihood given in (5). Here the model parameters consist of two blocks: the VAR coefficients β and the error covariance matrix Σ. Instead of estimating Σ, the Minnesota prior replaces it with an estimate $\widehat{\Sigma}$ obtained as follows.

We first estimate each of the n equations of the VAR separately, ignoring the error covariances across equations. Let s_i^2 denote the standard OLS estimate of the error variance for the i-th equation. Then, we set $\widehat{\Sigma} = \text{diag}(s_1^2, \ldots, s_n^2)$. As we will see below, the main advantage of this approach is that it simplifies the computations—often MCMC is not needed for posterior analysis or forecasts. One main drawback, however, is that here we replace an unknown quantity Σ by a potentially crude estimate $\widehat{\Sigma}$. This approach therefore ignores parameter uncertainty—instead of tackling it by integrating out the unknown parameters with respect to the posterior distribution. As such, this approach often produces inferior density forecasts.

With Σ being replaced by an estimate, the only parameters are the VAR coefficients β. Now, consider the following normal prior for β:

$$\beta \sim \mathcal{N}(\beta_{\text{Minn}}, \mathbf{V}_{\text{Minn}}).$$

The Minnesota prior sets sensible values for β_{Minn} and $\mathbf{V}_{\text{Minn}}$ in a systematic manner. To explain the prior elicitation procedure, first note that β consists of three groups of parameters: intercepts, coefficients associated with a variable's own lags and coefficients associated with lags of other variables.

The prior mean β_{Minn} is typically set to zero for growth rates data, such as GDP growth rate or inflation rate. This prior mean provides shrinkage for VAR coefficients, and reflects the prior belief that growth rates data are typically not persistent. For levels data such as money supply or consumption level, β_{Minn} is set to be zero except the coefficients associated with the first own lag, which are set to be one. This prior incorporates the belief that levels data are highly persistent—particularly, it expresses the preference for a random walk specification. Other variants, such as specifying a highly persistent but stationary process, are also commonly used.

The Minnesota prior sets the prior covariance matrix $\mathbf{V}_{\text{Minn}}$ to be diagonal; the exact values of the diagonal elements in turn depend on three key hyperparameters, c_1, c_2 and c_3. Now consider the coefficients in the i-th equation. First, for a coefficient associated with the i-th variable's own lag l, $l = 1, \ldots, p$, its variance is set to be c_1/l^2. That is, the higher the lag length, the higher the degree of shrinkage (either to zero or to unity). Second, for a coefficient associated with the l-th lag of

variable $j, j \neq i$, its variance is set to be $c_2 s_i^2/(l^2 s_j^2)$. In other words, in additional to applying higher level of shrinkage to higher lag length, the prior variance also adjusts for the scales of the variables. Lastly, the variance of the intercept is set to be c_3. The Minnesota prior therefore turns a complicated prior elicitation task into setting only three hyperparameters. There are by now many different variants of the Minnesota prior; see, e.g., Kadiyala and Karlsson (1997) and Karlsson (2013) for additional discussion.

2.1.1 Estimation

Estimation under the Minnesota prior is straightforward; that is one of the main appeals of the Minnesota prior. Recall that Σ is replaced by an estimate $\widehat{\Sigma}$, and we only need to estimate β. Given the VAR representation in (2) and the normal prior $\beta \sim \mathcal{N}(\beta_{\text{Minn}}, \mathbf{V}_{\text{Minn}})$, standard linear regression results give

$$(\beta \,|\, \mathbf{y}) \sim \mathcal{N}(\widehat{\beta}, \mathbf{K}_\beta^{-1}),$$

where

$$\mathbf{K}_\beta = \mathbf{V}_{\text{Minn}}^{-1} + \mathbf{X}'(\mathbf{I}_T \otimes \widehat{\Sigma}^{-1})\mathbf{X}, \quad \widehat{\beta} = \mathbf{K}_\beta^{-1}\left(\mathbf{V}_{\text{Minn}}^{-1}\beta_{\text{Minn}} + \mathbf{X}'(\mathbf{I}_T \otimes \widehat{\Sigma}^{-1})\mathbf{y}\right),$$

and we have replaced Σ by the estimate $\widehat{\Sigma}$. In particular, the posterior mean of β is $\widehat{\beta}$, and we would only need to compute this once instead of tens of thousands of times within a Gibbs sampler.

When the number of variables n is large, however, computations might still be an issue because $\widehat{\beta}$ is of dimension $nk \times 1$ with $k = np + 1$. In those cases, inverting the $nk \times nk$ precision matrix $\mathbf{K}_\beta$ to obtain the covariance matrix $\mathbf{K}_\beta^{-1}$ is computational intensive. It turns out that to obtain $\widehat{\beta}$, one needs not compute the inverse $\mathbf{K}_\beta^{-1}$ explicitly. To that end, we introduce the following notations: given a non-singular square matrix $\mathbf{B}$ and a conformable vector $\mathbf{c}$, let $\mathbf{B}\backslash\mathbf{c}$ denote the unique solution to the linear system $\mathbf{B}\mathbf{z} = \mathbf{c}$, i.e., $\mathbf{B}\backslash\mathbf{c} = \mathbf{B}^{-1}\mathbf{c}$. When $\mathbf{B}$ is lower triangular, this linear system can be solved quickly by forward substitution. When $\mathbf{B}$ is upper triangular, it can be solved by backward substitution.[2]

Now, we first compute the Cholesky factor $\mathbf{C}_{\mathbf{K}_\beta}$ of $\mathbf{K}_\beta$ such that $\mathbf{K}_\beta = \mathbf{C}_{\mathbf{K}_\beta}\mathbf{C}'_{\mathbf{K}_\beta}$. Then, compute

$$\mathbf{C}'_{\mathbf{K}_\beta} \backslash \left(\mathbf{C}_{\mathbf{K}_\beta} \backslash (\mathbf{V}_{\text{Minn}}^{-1}\beta_{\text{Minn}} + \mathbf{X}'(\mathbf{I}_T \otimes \widehat{\Sigma}^{-1})\mathbf{y})\right)$$

by forward then backward substitution.[3] Then, by construction,

[2] Forward and backward substitutions are implemented in standard packages such as MATLAB, GAUSS and R. In MATLAB, for example, it is done by `mldivide(B,c)` or simply `B\c`.

[3] Since $\mathbf{V}_{\text{Minn}}$ is diagonal, its inverse is straightforward to compute.

$$
\begin{aligned}
&(\mathbf{C}'_{\mathbf{K}_\beta})^{-1}\mathbf{C}_{\mathbf{K}_\beta}^{-1}(\mathbf{V}_{\text{Minn}}^{-1}\beta_{\text{Minn}}+\mathbf{X}'(\mathbf{I}_T\otimes\widehat{\Sigma}^{-1})\mathbf{y})\\
&\qquad=(\mathbf{C}_{\mathbf{K}_\beta}\mathbf{C}'_{\mathbf{K}_\beta})^{-1}(\mathbf{V}_{\text{Minn}}^{-1}\beta_{\text{Minn}}+\mathbf{X}'(\mathbf{I}_T\otimes\widehat{\Sigma}^{-1})\mathbf{y})\\
&\qquad=\widehat{\beta}.
\end{aligned}
$$

This alternative way to obtain $\widehat{\beta}$ is substantially faster when n is large.

2.2 The Natural Conjugate Prior

The original Minnesota prior discussed in Section 2.1 replaces the error covariance matrix Σ with an estimate—and in doing so ignores parameter uncertainty associated with Σ. That approach substantially simplifies the computations at the expense of the quality of density forecasts. In this section we introduce the natural conjugate prior for the VAR coefficients and Σ. This prior retains much of the computational tractability of the Minnesota prior, but it explicitly treats Σ to be an unknown quantity to be estimated.

To introduce the natural conjugate prior, we use the second representation of the VAR with likelihood given in (9). Now the model parameters consist of two blocks: the error covariance matrix Σ as before and the VAR coefficients organized into the $k\times n$ matrix $\mathbf{A}$. The natural conjugate prior is a joint distribution for $(\text{vec}(\mathbf{A}),\Sigma)$. To describe its specific form, we first need to define the following distributions.

An $n\times n$ random matrix Ω is said to have an **inverse-Wishart distribution** with shape parameter $\nu>0$ and scale matrix $\mathbf{S}$ if its density function is given by

$$
f(\Omega;\nu,\mathbf{S})=\frac{|\mathbf{S}|^{\nu/2}}{2^{n\nu/2}\Gamma_n(\nu/2)}|\Omega|^{-\frac{\nu+n+1}{2}}\text{e}^{-\frac{1}{2}\text{tr}(\mathbf{S}\Omega^{-1})},
$$

where Γ_n is the multivariate gamma function. We write $\Omega\sim\mathscr{IW}(\nu,\mathbf{S})$. For $\nu>m+1$, $\mathbb{E}\Omega=\mathbf{S}/(\nu-m-1)$.

Next, an $m\times n$ random matrix $\mathbf{W}$ and an $n\times n$ random matrix Ω are said to have a **normal-inverse-Wishart distribution** with parameters $\mathbf{M},\mathbf{P},\mathbf{S}$ and ν if $(\text{vec}(\mathbf{W})\,|\,\Omega)\sim\mathscr{N}(\text{vec}(\mathbf{M}),\Omega\otimes\mathbf{P})$ and $\Omega\sim\mathscr{IW}(\nu,\mathbf{S})$. We write $(\mathbf{W},\Omega)\sim\mathscr{NIW}(\mathbf{M},\mathbf{P},\nu,\mathbf{S})$. The kernel of the normal-inverse-Wishart density function is given by

$$
f(\mathbf{W},\Omega;\mathbf{M},\mathbf{P},\nu,\mathbf{S})\propto|\Omega|^{-\frac{\nu+m+n+1}{2}}\text{e}^{-\frac{1}{2}\text{tr}(\Omega^{-1}(\mathbf{W}-\mathbf{M})'\mathbf{P}^{-1}(\mathbf{W}-\mathbf{M}))}\text{e}^{-\frac{1}{2}\text{tr}(\Omega^{-1}\mathbf{S})}. \quad (10)
$$

To derive this density function from the definition, first note that

$$\begin{aligned}[\text{vec}(\mathbf{W}-\mathbf{M})]'(\Omega\otimes\mathbf{P})^{-1}\text{vec}(\mathbf{W}-\mathbf{M}) &= [\text{vec}(\mathbf{W}-\mathbf{M})]'(\Omega^{-1}\otimes\mathbf{P}^{-1})\text{vec}(\mathbf{W}-\mathbf{M}) \\ &= [\text{vec}(\mathbf{W}-\mathbf{M})]'\text{vec}(\mathbf{P}^{-1}(\mathbf{W}-\mathbf{M})\Omega^{-1}) \\ &= \text{tr}((\mathbf{W}-\mathbf{M})'\mathbf{P}^{-1}(\mathbf{W}-\mathbf{M})\Omega^{-1}) \\ &= \text{tr}(\Omega^{-1}(\mathbf{W}-\mathbf{M})'\mathbf{P}^{-1}(\mathbf{W}-\mathbf{M})).\end{aligned}$$

In the above derivations, the second equality holds because of (6); the third equality holds because of (7); and the last equality holds because of (8).

Now, from the definition, the joint density function of $(\text{vec}(\mathbf{W}),\Omega)$ is given by

$$\begin{aligned}f(\mathbf{W},\Omega) &\propto |\Omega|^{-\frac{\nu+n+1}{2}}\mathrm{e}^{-\frac{1}{2}\text{tr}(\mathbf{S}\Omega^{-1})}\times|\Omega\otimes\mathbf{P}|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}[\text{vec}(\mathbf{W}-\mathbf{M})]'(\Omega\otimes\mathbf{P})^{-1}\text{vec}(\mathbf{W}-\mathbf{M})} \\ &= |\Omega|^{-\frac{\nu+m+n+1}{2}}\mathrm{e}^{-\frac{1}{2}\text{tr}(\Omega^{-1}\mathbf{S})}\mathrm{e}^{-\frac{1}{2}\text{tr}(\Omega^{-1}(\mathbf{W}-\mathbf{M})'\mathbf{P}^{-1}(\mathbf{W}-\mathbf{M}))},\end{aligned}$$

where we have used the fact that $|\Omega\otimes\mathbf{P}| = |\Omega|^m|\mathbf{P}|^n$. This proves that the joint density function of $(\text{vec}(\mathbf{W}),\Omega)$ has the form given in (10).

By construction, the marginal distribution of Ω is $\mathcal{IW}(\nu,\mathbf{S})$. It turns out that the marginal distribution of $\text{vec}(\mathbf{W})$ unconditional on Ω is a multivariate t distribution. For more details, see, e.g., Karlsson (2013).

Now we consider the following normal-inverse-Wishart prior on $(\mathbf{A},\Sigma)$:

$$\Sigma\sim\mathcal{IW}(\nu_0,\mathbf{S}_0),\quad (\text{vec}(\mathbf{A})\,|\,\Sigma)\sim\mathcal{N}(\text{vec}(\mathbf{A}_0),\Sigma\otimes\mathbf{V}_\mathbf{A}).$$

That is, $(\text{vec}(\mathbf{A}),\Sigma)\sim\mathcal{NIW}(\text{vec}(\mathbf{A}_0),\mathbf{V}_\mathbf{A},\nu_0,\mathbf{S}_0)$ with joint density function

$$p(\mathbf{A},\Sigma)\propto|\Sigma|^{-\frac{\nu_0+n+k+1}{2}}\mathrm{e}^{-\frac{1}{2}\text{tr}(\Sigma^{-1}\mathbf{S}_0)}\mathrm{e}^{-\frac{1}{2}\text{tr}\left(\Sigma^{-1}(\mathbf{A}-\mathbf{A}_0)'\mathbf{V}_\mathbf{A}^{-1}(\mathbf{A}-\mathbf{A}_0)\right)}. \tag{11}$$

It turns out that the joint posterior distribution of $(\mathbf{A},\Sigma)$ is also a normal-inverse-Wishart distribution, as shown in the next section. Hence, this prior is often called the natural conjugate prior. The hyperparameters of this prior are $\text{vec}(\mathbf{A}_0),\mathbf{V}_\mathbf{A},\nu_0$, and $\mathbf{S}_0$. Below we describe one way to elicit these hyperparameters.

One often sets a small value for ν_0 (say, $n+2$) so that the prior variance of Σ is large—i.e., the prior is relatively uninformative. Given ν_0, one then chooses a value for $\mathbf{S}_0$ to match the desired prior mean of Σ via the equality $\mathbb{E}\Sigma = \mathbf{S}_0/(\nu_0-n-1)$. As for $\text{vec}(\mathbf{A}_0)$ and $\mathbf{V}_\mathbf{A}$, their values are chosen to mimic the Minnesota prior. For example, $\text{vec}(\mathbf{A}_0)$ is typically set to zero for growth rates data. For levels data, $\text{vec}(\mathbf{A}_0)$ is set to be zero except the coefficients associated with the first own lag, which are set to be one.

Finally, to elicit $\mathbf{V}_\mathbf{A}$, first note that given Σ, the prior covariance matrix of $\text{vec}(\mathbf{A})$ is $\Sigma\otimes\mathbf{V}_\mathbf{A}$. This Kronecker structure implies cross-equation restrictions on the covariance matrix, which is more restrictive than the covariance matrix $\mathbf{V}_{\text{Minn}}$ under the Minnesota prior. However, the advantage of this Kronecker structure is that it can be exploited to speed up computations, which we will discuss in the next section.

Following the example of the Minnesota prior, we choose $\mathbf{V}_\mathbf{A}$ to induce shrinkage. Specifically, $\mathbf{V}_\mathbf{A}$ is assumed to be diagonal with diagonal elements $v_{\mathbf{A},ii} = c_1/(l^2 s_r^2)$ for a coefficient associated with the l-th lag of variable r and $v_{\mathbf{A},ii} = c_2$

for an intercept, where s_r^2 is the residual sample variance of an AR(p) model for the variable r. Similar to the Minnesota prior, we apply a higher degree of shrinkage for a coefficient associated with a higher lag length. But contrary to the Minnesota prior, here we cannot have different prior variances for a variable's own lag and the lag of a different variable due to the Kronecker structure.

2.2.1 Estimation

In this section we discuss the estimation of $\mathbf{A}$ and Σ under the natural conjugate prior. As mentioned earlier, the posterior distribution of $\mathbf{A}$ and Σ turns out to be the normal-inverse-Wishart distribution as well. To see this, we combine the likelihood given in (9) and the natural conjugate prior in (11) to get

$$\begin{aligned}
p(\mathbf{A},\Sigma\,|\,\mathbf{Y}) &\propto p(\mathbf{A},\Sigma)p(\mathbf{Y}\,|\,\mathbf{A},\Sigma) \\
&\propto |\Sigma|^{-\frac{\nu_0+n+k+1}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\mathbf{S}_0)} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}(\mathbf{A}-\mathbf{A}_0)'\mathbf{V}_\mathbf{A}^{-1}(\mathbf{A}-\mathbf{A}_0)\right)} \\
&\quad \times |\Sigma|^{-\frac{T}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}(\mathbf{Y}-\mathbf{Z}\mathbf{A})'(\mathbf{Y}-\mathbf{Z}\mathbf{A})\right)}. \\
&\propto |\Sigma|^{-\frac{\nu_0+n+k+T+1}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\mathbf{S}_0)} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}\left((\mathbf{A}-\mathbf{A}_0)'\mathbf{V}_\mathbf{A}^{-1}(\mathbf{A}-\mathbf{A}_0)+(\mathbf{Y}-\mathbf{Z}\mathbf{A})'(\mathbf{Y}-\mathbf{Z}\mathbf{A})\right)\right]}.
\end{aligned} \tag{12}$$

The last line looks almost like the kernel of the normal-inverse-Wishart density function in (10)—the only difference is that here we have two quadratic terms involving $\mathbf{A}$ instead of one. If we could somehow write the sum

$$(\mathbf{A}-\mathbf{A}_0)'\mathbf{V}_\mathbf{A}^{-1}(\mathbf{A}-\mathbf{A}_0)+(\mathbf{Y}-\mathbf{Z}\mathbf{A})'(\mathbf{Y}-\mathbf{Z}\mathbf{A})$$

as $(\mathbf{A}-\widehat{\mathbf{A}})'\mathbf{K}_\mathbf{A}(\mathbf{A}-\widehat{\mathbf{A}})$ for some $k\times n$ matrix $\widehat{\mathbf{A}}$ and $k\times k$ symmetric matrix $\mathbf{K}_\mathbf{A}$, then $p(\mathbf{A},\Sigma\,|\,\mathbf{Y})$ is a normal-inverse-Wishart density function.

To that end, below we do a matrix version of "completing the square":

$$\begin{aligned}
&(\mathbf{A}-\mathbf{A}_0)'\mathbf{V}_\mathbf{A}^{-1}(\mathbf{A}-\mathbf{A}_0)+(\mathbf{Y}-\mathbf{Z}\mathbf{A})'(\mathbf{Y}-\mathbf{Z}\mathbf{A}) \\
&\quad = (\mathbf{A}'\mathbf{V}_\mathbf{A}^{-1}\mathbf{A}-2\mathbf{A}'\mathbf{V}_\mathbf{A}^{-1}\mathbf{A}_0+\mathbf{A}_0'\mathbf{V}_\mathbf{A}^{-1}\mathbf{A}_0)+(\mathbf{A}'\mathbf{Z}'\mathbf{Z}\mathbf{A}-2\mathbf{A}'\mathbf{Z}'\mathbf{Y}+\mathbf{Y}'\mathbf{Y}) \\
&\quad = \mathbf{A}'(\mathbf{V}_\mathbf{A}^{-1}+\mathbf{Z}'\mathbf{Z})\mathbf{A}-2\mathbf{A}'(\mathbf{V}_\mathbf{A}^{-1}\mathbf{A}_0+\mathbf{Z}'\mathbf{Y})+\widehat{\mathbf{A}}'\mathbf{K}_\mathbf{A}\widehat{\mathbf{A}}-\widehat{\mathbf{A}}'\mathbf{K}_\mathbf{A}\widehat{\mathbf{A}}+\mathbf{A}_0'\mathbf{V}_\mathbf{A}^{-1}\mathbf{A}_0+\mathbf{Y}'\mathbf{Y} \\
&\quad = (\mathbf{A}-\widehat{\mathbf{A}})'\mathbf{K}_\mathbf{A}(\mathbf{A}-\widehat{\mathbf{A}})-\widehat{\mathbf{A}}'\mathbf{K}_\mathbf{A}\widehat{\mathbf{A}}+\mathbf{A}_0'\mathbf{V}_\mathbf{A}^{-1}\mathbf{A}_0+\mathbf{Y}'\mathbf{Y},
\end{aligned} \tag{13}$$

where

$$\mathbf{K}_\mathbf{A}=\mathbf{V}_\mathbf{A}^{-1}+\mathbf{Z}'\mathbf{Z},\quad \widehat{\mathbf{A}}=\mathbf{K}_\mathbf{A}^{-1}(\mathbf{V}_\mathbf{A}^{-1}\mathbf{A}_0+\mathbf{Z}'\mathbf{Y}).$$

Note that on the right-hand-side of the second equality, we judiciously add and subtract the term $\widehat{\mathbf{A}}'\mathbf{K}_\mathbf{A}\widehat{\mathbf{A}}$ so that we obtain one quadratic form in $\mathbf{A}$.

Now, substituting (13) into (12), we have

$$p(\mathbf{A},\Sigma\,|\,\mathbf{Y}) \propto |\Sigma|^{-\frac{\nu_0+n+k+T+1}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\mathbf{S}_0)} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}\left((\mathbf{A}-\widehat{\mathbf{A}})'\mathbf{K}_\mathbf{A}(\mathbf{A}-\widehat{\mathbf{A}})-\widehat{\mathbf{A}}'\mathbf{K}_\mathbf{A}\widehat{\mathbf{A}}+\mathbf{A}_0'\mathbf{V}_\mathbf{A}^{-1}\mathbf{A}_0+\mathbf{Y}'\mathbf{Y}\right)\right]}$$
$$= |\Sigma|^{-\frac{\nu_0+n+k+T+1}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\widehat{\mathbf{S}})} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}(\mathbf{A}-\widehat{\mathbf{A}})'\mathbf{K}_\mathbf{A}(\mathbf{A}-\widehat{\mathbf{A}})\right]},$$

where $\widehat{\mathbf{S}} = \mathbf{S}_0 + \mathbf{A}_0'\mathbf{V}_\mathbf{A}^{-1}\mathbf{A}_0 + \mathbf{Y}'\mathbf{Y} - \widehat{\mathbf{A}}'\mathbf{K}_\mathbf{A}\widehat{\mathbf{A}}$. Comparing this kernel with the normal-inverse-gamma density function in (10), we conclude that

$$(\mathbf{A},\Sigma\,|\,\mathbf{Y}) \sim \mathscr{N}\mathscr{I}\mathscr{W}(\widehat{\mathbf{A}},\mathbf{K}_\mathbf{A}^{-1},\nu_0+T,\widehat{\mathbf{S}}).$$

In particular, the posterior means of $\mathbf{A}$ and Σ are respectively $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{S}}/(\nu_0+T-1)$. Other posterior moments can often be found by using properties of the normal-inverse-Wishart distribution. When analytically results are not available, we can estimate the quantities of interest by generating draws from the posterior distribution $p(\mathbf{A},\Sigma\,|\,\mathbf{Y})$. Below we describe a computationally efficient way to obtain posterior draws.

Since $(\mathbf{A},\Sigma\,|\,\mathbf{Y}) \sim \mathscr{N}\mathscr{I}\mathscr{W}(\widehat{\mathbf{A}},\mathbf{K}_\mathbf{A}^{-1},\nu_0+T,\widehat{\mathbf{S}})$, we can sample $\mathbf{A}$ and Σ in two steps. First, we draw Σ marginally from $(\Sigma\,|\,\mathbf{Y}) \sim \mathscr{I}\mathscr{W}(\nu_0+T,\widehat{\mathbf{S}})$. Then, given the sampled Σ, we simulate from the conditional distribution

$$(\mathrm{vec}(\mathbf{A})\,|\,\mathbf{Y},\Sigma) \sim \mathscr{N}(\mathrm{vec}(\widehat{\mathbf{A}}),\Sigma\otimes\mathbf{K}_\mathbf{A}^{-1}).$$

Here note that the covariance matrix $\Sigma\otimes\mathbf{K}_\mathbf{A}^{-1}$ is of dimension $nk = n(np+1)$, and sampling from this normal distribution using conventional methods—e.g., computing the Cholesky factor of the covariance matrix $\Sigma\otimes\mathbf{K}_\mathbf{A}^{-1}$—would involve $\mathscr{O}(n^6)$ operations. This is especially computationally intensive when n is large. Here we consider an alternative method with complexity of the order $\mathscr{O}(n^3)$ only.

This more efficient approach exploits the Kronecker structure $\Sigma\otimes\mathbf{K}_\mathbf{A}^{-1}$ to speed up computation. In particular, it is based on an efficient sampling algorithm to draw from the matrix normal distribution.[4] We further improve upon this approach by avoiding the computation of the inverse of the $k\times k$ matrix $\mathbf{K}_\mathbf{A}$.

Recall that given a non-singular square matrix $\mathbf{B}$ and a conformable vector $\mathbf{c}$, we use the notation $\mathbf{B}\backslash\mathbf{c}$ to denote the unique solution to the linear system $\mathbf{B}\mathbf{z}=\mathbf{c}$, i.e., $\mathbf{B}\backslash\mathbf{c} = \mathbf{B}^{-1}\mathbf{c}$. Now, we first obtain the Cholesky decomposition $\mathbf{C}_{\mathbf{K}_\mathbf{A}}$ of $\mathbf{K}_\mathbf{A}$ such that $\mathbf{C}_{\mathbf{K}_\mathbf{A}}\mathbf{C}_{\mathbf{K}_\mathbf{A}}' = \mathbf{K}_\mathbf{A}$. Then compute

$$\mathbf{C}_{\mathbf{K}_\mathbf{A}}'\backslash(\mathbf{C}_{\mathbf{K}_\mathbf{A}}\backslash(\mathbf{V}_\mathbf{A}^{-1}\mathbf{A}_0+\mathbf{Z}'\mathbf{Y}))$$

by forward followed by backward substitution. By construction,

$$(\mathbf{C}_{\mathbf{K}_\mathbf{A}}')^{-1}(\mathbf{C}_{\mathbf{K}_\mathbf{A}}^{-1}(\mathbf{V}_\mathbf{A}^{-1}\mathbf{A}_0+\mathbf{Z}'\mathbf{Y})) = (\mathbf{C}_{\mathbf{K}_\mathbf{A}}'\mathbf{C}_{\mathbf{K}_\mathbf{A}})^{-1}(\mathbf{V}_\mathbf{A}^{-1}\mathbf{A}_0+\mathbf{Z}'\mathbf{Y}) = \widehat{\mathbf{A}}.$$

Next, let $\mathbf{C}_\Sigma$ be the Cholesky decomposition of Σ. Then, compute

[4] The algorithm of drawing from the matrix normal distribution is well-known, and is described in the textbook by Bauwens, Lubrano, and Richard (1999, p.320). This algorithm is adapted in Carriero, Clark, and Marcellino (2016) and Chan (2018) to estimate more flexible large Bayesian VARs.

$$\mathbf{W}_1 = \widehat{\mathbf{A}} + (\mathbf{C}'_{\mathbf{K}_\mathbf{A}} \backslash \mathbf{U})\mathbf{C}'_\Sigma,$$

where $\mathbf{U}$ is a $k \times n$ matrix of independent $\mathcal{N}(0,1)$ random variables. In the Appendix B we show that $\text{vec}(\mathbf{W}_1) \sim \mathcal{N}(\text{vec}(\widehat{\mathbf{A}}), \Sigma \otimes \mathbf{K}_\mathbf{A}^{-1})$ as desired.

Therefore, we have a computationally efficient way to sample from the posterior distribution $(\mathbf{A}, \Sigma \,|\, \mathbf{Y}) \sim \mathcal{NIW}(\widehat{\mathbf{A}}, \mathbf{K}_\mathbf{A}^{-1}, \nu_0 + T, \widehat{\mathbf{S}})$. Note also that the algorithm described above gives us an independent sample—unlike MCMC draws which are correlated by construction.

2.3 The Independent Normal and Inverse-Wishart Prior

The main advantage of the natural conjugate prior is that analytical results are available for posterior analysis and simulation is typically not needed. However, it comes at a cost of restricting the form of prior variances on the VAR coefficients. In this section we discuss an alternative joint prior for the VAR coefficients and covariance matrix that is more flexible.

To that end, we use the first representation of the VAR with likelihood given in (5). This joint prior on (β, Σ) is often called the independent normal and inverse-Wishart prior, because it assumes prior independence between β and Σ, i.e., $p(\beta, \Sigma) = p(\beta)p(\Sigma)$. More specifically, we consider the form

$$\beta \sim \mathcal{N}(\beta_0, \mathbf{V}_\beta), \quad \Sigma \sim \mathcal{IW}(\nu_0, \mathbf{S}_0)$$

with prior densities

$$p(\beta) = (2\pi)^{-\frac{nk}{2}} |\mathbf{V}_\beta|^{-\frac{1}{2}} \text{e}^{-\frac{1}{2}(\beta-\beta_0)'\mathbf{V}_\beta^{-1}(\beta-\beta_0)}, \tag{14}$$

$$p(\Sigma) = \frac{|\mathbf{S}_0|^{\nu_0/2}}{2^{n\nu_0/2}\Gamma_n(\nu_0/2)} |\Sigma|^{-\frac{\nu_0+n+1}{2}} \text{e}^{-\frac{1}{2}\text{tr}(\mathbf{S}_0\Sigma^{-1})}. \tag{15}$$

The hyperparameters of this prior are $\beta, \mathbf{V}_\beta, \nu_0$, and $\mathbf{S}_0$. The values for ν_0 and $\mathbf{S}_0$ can be chosen the same way as in the case of the natural conjugate prior. For β_0 and $\mathbf{V}_\beta$, we can set them to be the same as the Minnesota prior, i.e., $\beta_0 = \beta_{\text{Minn}}$ and $\mathbf{V}_\beta = \mathbf{V}_{\text{Minn}}$. Also note that in contrast to the natural conjugate prior, here $\mathbf{V}_\beta$, the prior covariance matrix of the VAR coefficients, is not required to have a Kronecker structure, and is therefore more flexible.

2.3.1 Estimation

As mentioned above, in contrast to the case of the natural conjugate prior, the posterior distribution under the independent normal and inverse-Wishart prior is non-standard, and posterior simulation is needed for estimation and forecasting. Below

we derive a Gibbs sampler to draw from the posterior distribution $p(\beta, \Sigma \,|\, \mathbf{y})$. To that end, we derive the two full conditional distributions $p(\beta \,|\, \mathbf{y}, \Sigma)$ and $p(\Sigma \,|\, \mathbf{y}, \beta)$.

Using the likelihood given in (5) and the prior on β given in (14), we note that standard linear regression results would apply. In fact, we have

$$(\beta \,|\, \mathbf{y}, \Sigma) \sim \mathcal{N}(\widehat{\beta}, \mathbf{K}_\beta^{-1}),$$

where

$$\mathbf{K}_\beta = \mathbf{V}_\beta^{-1} + \mathbf{X}'(\mathbf{I}_T \otimes \Sigma^{-1})\mathbf{X}, \quad \widehat{\beta} = \mathbf{K}_\beta^{-1}\left(\mathbf{V}_\beta^{-1}\beta_0 + \mathbf{X}'(\mathbf{I}_T \otimes \Sigma^{-1})\mathbf{y}\right).$$

The main difficulty of obtaining a draw from $\mathcal{N}(\widehat{\beta}, \mathbf{K}_\beta^{-1})$ using conventional methods is that computing the $n(np+1) \times n(np+1)$ inverse $\mathbf{K}_\beta^{-1}$ is very computationally intensive when n is large. But fortunately we can sample from $\mathcal{N}(\widehat{\beta}, \mathbf{K}_\beta^{-1})$ without computing $\mathbf{K}_\beta^{-1}$ explicitly. First, $\widehat{\beta}$ can be obtained by forward and backward substitution as before. Second, we can use an alternative algorithm to sample from $\mathcal{N}(\widehat{\beta}, \mathbf{K}_\beta^{-1})$ without computing $\mathbf{K}_\beta^{-1}$ explicitly.

Algorithm 1 (Sampling from the Normal Distribution Given the Precision Matrix).

To generate R independent draws from $\mathcal{N}(\mu, \mathbf{K}^{-1})$ of dimension m, carry out the following steps:

1. Compute the lower Cholesky factor $\mathbf{B}$ of $\mathbf{K}$ such that $\mathbf{K} = \mathbf{B}\mathbf{B}'$.
2. Generate $\mathbf{U} = (U_1, \ldots, U_m)'$ by drawing $U_1, \ldots, U_m \sim \mathcal{N}(0,1)$.
3. Return $\mathbf{W} = \mu + (\mathbf{B}')^{-1}\mathbf{U}$.
4. Repeat Steps 2 and 3 independently R times.

To check that $\mathbf{W} \sim \mathcal{N}(\mu, \mathbf{K}^{-1})$, we first note that $\mathbf{W}$ is an affine transformation of the normal random vector $\mathbf{U}$, so it has a normal distribution. It is easy to check that $\mathbb{E}\mathbf{W} = \mu$. The covariance matrix of $\mathbf{W}$ is

$$\mathrm{Cov}(\mathbf{W}) = (\mathbf{B}')^{-1}\mathrm{Cov}(\mathbf{U})((\mathbf{B}')^{-1})' = (\mathbf{B}')^{-1}(\mathbf{B})^{-1} = (\mathbf{B}\mathbf{B}')^{-1} = \mathbf{K}^{-1}.$$

Hence, $\mathbf{W} \sim \mathcal{N}(\mu, \mathbf{K}^{-1})$.

Using this algorithm to sample from $\mathcal{N}(\widehat{\beta}, \mathbf{K}_\beta^{-1})$ allows us to avoid the expensive computation of the inverse $\mathbf{K}_\beta^{-1}$. However, if $\mathbf{K}_\beta$ is a dense matrix, this algorithm still involves $\mathcal{O}(n^6)$ operations. Hence, it is expected to be much slower than simulations under the natural conjugate prior that involves only $\mathcal{O}(n^3)$ operations.

Next, we derive the conditional distribution $p(\Sigma \,|\, \mathbf{y}, \beta)$. First note that the likelihood in (5) can be equivalently written as

$$p(\mathbf{y} \,|\, \beta, \Sigma) = (2\pi)^{-\frac{Tn}{2}} |\Sigma|^{-\frac{T}{2}} \mathrm{e}^{-\frac{1}{2}\sum_{t=1}^{T}(\mathbf{y}_t - \mathbf{X}_t\beta)'\Sigma^{-1}(\mathbf{y}_t - \mathbf{X}_t\beta)}. \tag{16}$$

Now, combining (16) and the prior on Σ given in (15), we have

$$
\begin{aligned}
p(\Sigma \,|\, \mathbf{y}, \beta) &\propto p(\mathbf{y} \,|\, \beta, \Sigma) p(\Sigma) \\
&\propto |\Sigma|^{-\frac{T}{2}} \mathrm{e}^{-\frac{1}{2}\sum_{t=1}^{T}(\mathbf{y}_t - \mathbf{X}_t\beta)'\Sigma^{-1}(\mathbf{y}_t - \mathbf{X}_t\beta)} \times |\Sigma|^{-\frac{\nu_0+n+1}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}(\mathbf{S}_0\Sigma^{-1})} \\
&= |\Sigma|^{-\frac{\nu_0+n+T+1}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}(\mathbf{S}_0\Sigma^{-1})} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left[\sum_{t=1}^{T}(\mathbf{y}_t - \mathbf{X}_t\beta)(\mathbf{y}_t - \mathbf{X}_t\beta)'\Sigma^{-1}\right]} \\
&= |\Sigma|^{-\frac{\nu_0+n+T+1}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left[\left(\mathbf{S}_0 + \sum_{t=1}^{T}(\mathbf{y}_t - \mathbf{X}_t\beta)(\mathbf{y}_t - \mathbf{X}_t\beta)'\right)\Sigma^{-1}\right]},
\end{aligned}
$$

which is the kernel of an inverse-Wishart density function. In fact, we have

$$
(\Sigma \,|\, \mathbf{y}, \beta) \sim \mathscr{IW}\left(\nu_0 + T, \mathbf{S}_0 + \sum_{t=1}^{T}(\mathbf{y}_t - \mathbf{X}_t\beta)(\mathbf{y}_t - \mathbf{X}_t\beta)'\right). \tag{17}
$$

Hence, a Gibbs sampler can be constructed to simulate from the posterior distribution by repeatedly drawing from $p(\beta \,|\, \mathbf{y}, \Sigma)$ and $p(\Sigma \,|\, \mathbf{y}, \beta)$.

2.4 The Stochastic Search Variable Selection Prior

Another popular shrinkage prior for the VAR coefficients is the so-called stochastic search variable selection (SSVS) prior considered in George, Sun, and Ni (2008). It is based on the independent normal and inverse-Wishart prior, but it introduces a hierarchical structure for the normal prior on β. The main idea is to divide, in a data-based manner, the elements in β into two groups: in the first group the coefficients are shrunk strongly to zero, whereas they are not shrunk in the second group. In other words, the "variable selection" part is done by setting the coefficients in the first group to be close to zero, and only the variables in the second group are "selected". This partition is done stochastically in each iteration in the MCMC sampler, and hence "stochastic search".

Specifically, the elements of β are assumed to be independent, and each element β_j has a two-component mixture distribution with mixture weight $q_j \in (0,1)$:

$$
(\beta_j \,|\, q_j) \sim (1 - q_j)\phi(\beta_j; 0, \kappa_{0j}) + q_j\phi(\beta_j; 0, \kappa_{1j}),
$$

where $\phi(\cdot; \mu, \sigma^2)$ denotes the density function of the $\mathscr{N}(\mu, \sigma^2)$ distribution. The SSVS prior sets the first prior variance κ_{0j} to be "small"' and the second prior variance κ_{1j} to be large.

To see the partition more clearly, let us consider an equivalent latent variable representation by introducing the indicator $\gamma_j \in \{0,1\}$ with success probability q_j, i.e., $\mathbb{P}(\gamma_j = 1 \,|\, q_j) = q_j$. Then, we can rewrite the above prior as

$$
(\beta_j \,|\, \gamma_j) \sim (1 - \gamma_j)\mathscr{N}(0, \kappa_{0j}) + \gamma_j\mathscr{N}(0, \kappa_{1j}).
$$

Hence, when $\gamma_j = 0$, β_j is strongly shrunk to zero; when $\gamma_j = 1$, the prior on β_j is relatively non-informative.

Let $\gamma = (\gamma_1, \ldots, \gamma_{nk})'$. For later reference, we rewrite the joint prior $(\beta \,|\, \gamma)$ as:

$$(\beta \,|\, \gamma) \sim \mathcal{N}(\mathbf{0}, \Omega_\gamma),$$

where Ω_γ is diagonal with diagonal elements $(1-\gamma_j)\kappa_{0j} + \gamma_j \kappa_{1j}$, $j = 1, \ldots, nk$.

It remains to choose values for the prior variances κ_{0j} and κ_{1j}. There are various implementations, here we simply set $\kappa_{1j} = 10$ and κ_{0j} to be the j-th diagonal element of the Minnesota prior covariance matrix $\mathbf{V}_{\text{Minn}}$. As for Σ, we assume the inverse-Wishart prior:

$$\Sigma \sim \mathcal{IW}(\nu_0, \mathbf{S}_0).$$

It is possible to have a SSVS prior on Σ as well. See George, Sun, and Ni (2008) for further details. Finally, we set the mixture weight q_j to be 0.5, so that β_j has equal probabilities in each component. An alternative is to treat q_j as a model parameter to be estimated.

2.4.1 Estimation

Estimation involves only slight modifications of the 2-block Gibbs sampler under the independent normal and inverse-Wishart prior. In particular, here we construct a 3-block sampler to sequentially draw from $p(\Sigma \,|\, \mathbf{y}, \gamma, \beta)$, $p(\beta \,|\, \mathbf{y}, \gamma, \Sigma)$ and $p(\gamma | \mathbf{y}, \beta, \Sigma)$.

The full conditional distribution of Σ is inverse-Wishart, having the exact same form as given in (17). Next, the full conditional distribution of β is again normal:

$$(\beta \,|\, \mathbf{y}, \gamma, \Sigma) \sim \mathcal{N}(\widehat{\beta}, \mathbf{K}_\beta^{-1}),$$

where

$$\mathbf{K}_\beta = \Omega_\gamma^{-1} + \mathbf{X}'(\mathbf{I}_T \otimes \Sigma^{-1})\mathbf{X}, \quad \widehat{\beta} = \mathbf{K}_\beta^{-1}\mathbf{X}'(\mathbf{I}_T \otimes \Sigma^{-1})\mathbf{y}.$$

Sampling from this normal distribution can be done using Algorithm 1.

Finally, to draw from $p(\gamma | \mathbf{y}, \beta, \Sigma)$, note that $\gamma_1, \ldots, \gamma_{nk}$ are conditionally independent given the data and other parameters. In fact, we have $p(\gamma | \mathbf{y}, \beta, \Sigma) = \prod_{j=1}^{nk} p(\gamma_j \,|\, \beta_j)$. Moreover, each γ_j is a Bernoulli random variable and we only need need to compute its success probability. To that end, note that

$$\mathbb{P}(\gamma_j = 1 \,|\, \beta_j) \propto q_j \phi(\beta_j; 0, \kappa_{1j})$$

and

$$\mathbb{P}(\gamma_j = 0 \,|\, \beta_j) \propto (1-q_j)\phi(\beta_j; 0, \kappa_{0j}).$$

Hence, after normalization, we obtain

$$\mathbb{P}(\gamma_j = 1 \,|\, \beta_j) = \frac{q_j \phi(\beta_j; 0, \kappa_{1j})}{q_j \phi(\beta_j; 0, \kappa_{1j}) + (1-q_j)\phi(\beta_j; 0, \kappa_{0j})}.$$

3 Large Bayesian VARs with Time-Varying Volatility, Heavy Tails and Serial Dependent Errors

Despite the empirical success of large Bayesian VARs with standard error assumptions (e.g., homoscedastic, Gaussian and serially independent), there is a lot of recent work in developing flexible VARs with more general error distributions. These more flexible VARs are motivated by the empirical observations that features like time-varing volatility and non-Gaussian errors are useful for modeling a variety of macroeconomic time series.

In this section we study a few of these more flexible VARs, including VARs with heteroscedastic, non-Gaussian and serially correlated errors. To that end, we focus on the second representation of the VAR(p), which we reproduce below for convenience:

$$\mathbf{Y} = \mathbf{ZA} + \mathbf{U}, \qquad \text{vec}(\mathbf{U}) \sim \mathcal{N}(\mathbf{0}, \Sigma \otimes \mathbf{I}_T).$$

Note we can equivalently write the error specification as $\mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \Sigma)$, $t = 1, \ldots, T$. That is, the errors here are assumed to be independent, homoscedastic and Gaussian. Below we consider a variety of extensions of this basic VAR.

To motivate the framework, recall that the main difficulty in doing posterior simulation for large Bayesian VARs is the large number of VAR coefficients in $\mathbf{A}$. One key advantage of the natural conjugate prior on $(\mathbf{A}, \Sigma)$ is that the conditional distribution of $\mathbf{A}$ given Σ is Gaussian and its covariance matrix has a Kronecker product structure. This special feature can be exploited to dramatically speed up computation from $\mathcal{O}(n^6)$ to $\mathcal{O}(n^3)$, as described in Section 2.2.1.

It turns out that this Kronecker product structure in the conditional covariance matrix of $\mathbf{A}$ can be preserved for a wide class of flexible models. Specifically, Chan (2018) proposes the following VAR with a more general covariance structure:

$$\mathbf{Y} = \mathbf{ZA} + \mathbf{U}, \qquad \text{vec}(\mathbf{U}) \sim \mathcal{N}(\mathbf{0}, \Sigma \otimes \Omega), \tag{18}$$

where Ω is a $T \times T$ covariance matrix. Obviously, if $\Omega = \mathbf{I}_T$, (18) reduces to the standard VAR. Here the covariance matrix of $\text{vec}(\mathbf{U})$ is assumed to have the Kronecker product structure $\Sigma \otimes \Omega$. Intuitively, it separately models the cross-sectional and serial covariance structures of $\mathbf{U}$, which are governed by Σ and Ω respectively.

In the next few subsections, we first show that by choosing a suitable serial covariance structure Ω, the model in (18) includes a wide variety of flexible specifications. Section 3.4 then shows that the form of the error covariance matrix, namely $\Sigma \otimes \Omega$, leads to a Kronecker product structure in the conditional covariance matrix of $\mathbf{A}$. Again this special feature is used to dramatically speed up computation. The presentation below follows Chan (2018).

3.1 Common Stochastic Volatility

One of the most useful features for modeling macroeconomic time series is time-variance volatility. For example, the volatilities of a wide range of macroeconomic variables were substantially reduced at the start of the Great Moderation in the early 1980s. Models with homoscedastic errors would not be able to capture this feature of the data.

To allow for heteroscedastic errors, Carriero, Clark, and Marcellino (2016) introduce a large Bayesian VAR with a common stochastic volatility. In their setup, the error covariance matrix is scaled by a common, time-varying factor that can be interpreted as the overall macroeconomic volatility. More specifically, consider $\mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \mathrm{e}^{h_t}\Sigma)$ with the common volatility e^{h_t}. The log volatility h_t in turns follows a stationary AR(1) process:

$$h_t = \rho h_{t-1} + \varepsilon_t^h, \tag{19}$$

where $\varepsilon_t^h \sim \mathcal{N}(0, \sigma_h^2)$ and $|\rho| < 1$. Note that for identification purposes, this AR(1) process is assumed to have a zero unconditional mean.

One drawback of this setup is that the volatility specification is somewhat restrictive—all variances are scaled by a single factor and, consequently, they are always proportional to each other. On the other hand, there is empirical evidence, as shown in Carriero, Clark, and Marcellino (2016), that the volatilities of macroeconomic variables tend to move together. And specifying a common stochastic volatility is a parsimonious way to model that feature.

This common stochastic volatility model falls within the framework in (18) with $\Omega = \mathrm{diag}(\mathrm{e}^{h_1}, \ldots, \mathrm{e}^{h_T})$. Empirical applications that use this common stochastic volatility include Mumtaz (2016), Mumtaz and Theodoridis (2017) and Poon (2018).

3.2 Non-Gaussian Errors

Gaussian errors are often assumed for convenience rather than for deep theoretical reasons. In fact, some recent work has found that VARs with heavy-tailed error distributions, such as the t distribution, often forecast better than their counterparts with Gaussian errors (see, e.g., Cross and Poon, 2016; Chiu, Mumtaz, and Pinter, 2017).

Since many distributions can be written as a scale mixture of Gaussian distributions, the framework in (18) can accommodate various commonly-used non-Gaussian distributions. To see this, let $\Omega = \mathrm{diag}(\lambda_1, \ldots, \lambda_T)$. If each λ_t follows independently an inverse-gamma distribution $(\lambda_t \,|\, \nu) \sim \mathcal{IG}(\nu/2, \nu/2)$, then marginally $\mathbf{u}_t$ has a multivariate t distribution with mean vector $\mathbf{0}$, scale matrix Σ and degree of freedom parameter ν (see, e.g., Geweke, 1993).

If each λ_t has an independent exponential distribution with mean α, then marginally $\mathbf{u}_t$ has a multivariate Laplace distribution with mean vector $\mathbf{0}$ and covariance matrix $\alpha\Sigma$ (Eltoft, Kim, and Lee, 2006b). Other scale mixtures of Gaussian distributions can be defined similarly. For additional examples, see, e.g., Eltoft, Kim, and Lee (2006a).

3.3 Serially Dependent Errors

Instead of the conventional assumption of serially independent errors, the framework in (18) can also handle serially correlated errors, such as errors that follow an ARMA(p,q) process.

For a concrete example, suppose $\mathbf{u}_t$ follows the following MA(2) process:

$$\mathbf{u}_t = \varepsilon_t + \psi_1\varepsilon_{t-1} + \psi_2\varepsilon_{t-2},$$

where $\varepsilon_t \sim \mathcal{N}(\mathbf{0},\Sigma)$, ψ_1 and ψ_2 satisfy the invertibility conditions. This is nested within the general framework with

$$\Omega = \begin{pmatrix} \omega_0 & \omega_1 & \omega_2 & 0 & \cdots & 0 \\ \omega_1 & \omega_0 & \omega_1 & \ddots & \ddots & \vdots \\ \omega_2 & \omega_1 & \omega_0 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \omega_2 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \omega_1 \\ 0 & \cdots & 0 & \omega_2 & \omega_1 & \omega_0 \end{pmatrix},$$

where $\omega_0 = 1+\psi_1^2+\psi_2^2$, $\omega_1 = \psi_1(1+\psi_2)$ and $\omega_2 = \psi_2$.

One drawback of the above MA(2) specification is that each element of $\mathbf{u}_t$ must have the same MA coefficients (although their variances can be different). Put it differently, the framework in (18) cannot accommodate, for example, a general MA(2) process of the form

$$\mathbf{u}_t = \varepsilon_t + \Psi_1\varepsilon_{t-1} + \Psi_2\varepsilon_{t-2},$$

where Ψ_1 and Ψ_2 are $n\times n$ matrices of coefficients. This is because in this case the covariance matrix of $\mathrm{vec}(\mathbf{U})$ does not have a Kronecker structure—i.e., it cannot be written as $\Sigma\otimes\Omega$. Nevertheless, this restricted form of serial correlation might still be useful to capture persistence in the data.

Other more elaborate covariance structures can be constructed by combining different examples in previous sections. For example, suppose $\mathbf{u}_t$ follows an MA(1) stochastic volatility process of the form:

$$\mathbf{u}_t = \varepsilon_t + \psi_1\varepsilon_{t-1},$$

where $\varepsilon_t \sim \mathcal{N}(\mathbf{0}, \mathrm{e}^{h_t}\Sigma)$ and h_t has an AR(1) process as in (19). This is a multivariate generalization of the univariate moving average stochastic volatility models considered in Chan (2013). This model is a special case of the flexible Bayesian VAR in (18) with

$$\Omega = \begin{pmatrix} (1+\psi_1^2)\mathrm{e}^{h_1} & \psi_1\mathrm{e}^{h_1} & 0 & \cdots & 0 \\ \psi_1\mathrm{e}^{h_1} & \psi_1^2\mathrm{e}^{h_1}+\mathrm{e}^{h_2} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \psi_1^2\mathrm{e}^{h_{T-2}}+\mathrm{e}^{h_{T-1}} & \psi_1\mathrm{e}^{h_{T-1}} \\ 0 & \cdots & 0 & \psi_1\mathrm{e}^{h_{T-1}} & \psi_1^2\mathrm{e}^{h_{T-1}}+\mathrm{e}^{h_T} \end{pmatrix}.$$

3.4 Estimation

Next we discuss the estimation of the Bayesian VAR in (18) using MCMC methods. To keep the discussion general, we leave Ω unspecified and focus on the key step of jointly sampling both the VAR coefficients $\mathbf{A}$ and the cross-sectional covariance matrix Σ. Then, we take up various examples of Ω and provide estimation details for tackling each case.

Using a similar derivation as in Section 1.2, one can show that the likelihood of the VAR in (18) is given by

$$p(\mathbf{Y} \,|\, \mathbf{A}, \Sigma, \Omega) = (2\pi)^{-\frac{Tn}{2}} |\Sigma|^{-\frac{T}{2}} |\Omega|^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}(\mathbf{Y}-\mathbf{X}\mathbf{A})'\Omega^{-1}(\mathbf{Y}-\mathbf{X}\mathbf{A})\right)}. \tag{20}$$

Next, we assume a prior of the form $p(\mathbf{A}, \Sigma, \Omega) = p(\mathbf{A}, \Sigma)p(\Omega)$, i.e., the parameter blocks $(\mathbf{A}, \Sigma)$ and Ω are *a priori* independent. For $(\mathbf{A}, \Sigma)$, we adopt the natural conjugate prior:

$$\Sigma \sim \mathcal{IW}(\nu_0, \mathbf{S}_0), \quad (\mathrm{vec}(\mathbf{A}) \,|\, \Sigma) \sim \mathcal{N}(\mathrm{vec}(\mathbf{A}_0), \Sigma \otimes \mathbf{V}_\mathbf{A})$$

with joint density function given in (11).

Given the prior $p(\mathbf{A}, \Sigma, \Omega) = p(\mathbf{A}, \Sigma)p(\Omega)$, posterior draws can be obtained by sequentially sampling from: 1) $p(\mathbf{A}, \Sigma \,|\, \mathbf{Y}, \Omega)$; and 2) $p(\Omega \,|\, \mathbf{Y}, \mathbf{A}, \Sigma)$. Here we first describe how one can implement Step 1 efficiently. Depending on the covariance structure Ω, additional blocks might be needed to sample some extra hierarchical parameters. These steps are typically easy to implement as they amount to fitting a univariate time series model. We will discuss various examples below.

When $\Omega = \mathbf{I}_T$, the Bayesian VAR in (18) reduces to the conventional VAR with the natural conjugate prior. And in Section 2.2 we showed that $(\mathbf{A}, \Sigma \,|\, \mathbf{Y})$ has a normal-inverse-Wishart distribution. There we also discussed how we can draw from the normal-inverse-Wishart distribution efficiently. It turns out that similar derivations go through even with an arbitrary covariance matrix Ω. More specifically, it follows from (20) and (11) that

$$\begin{aligned} p(\mathbf{A},\Sigma\,|\,\mathbf{Y},\Omega) \propto & |\Sigma|^{-\frac{\nu_0+n+k+T}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\mathbf{S}_0)} \\ & \times \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}((\mathbf{A}-\mathbf{A}_0)'\mathbf{V}_\mathbf{A}^{-1}(\mathbf{A}-\mathbf{A}_0)+(\mathbf{Y}-\mathbf{Z}\mathbf{A})'\Omega^{-1}(\mathbf{Y}-\mathbf{Z}\mathbf{A}))\right)} \\ = & |\Sigma|^{-\frac{\nu_0+n+k+T}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\mathbf{S}_0)} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}(\mathbf{A}_0'\mathbf{V}_\mathbf{A}^{-1}\mathbf{A}_0+\mathbf{Y}'\Omega^{-1}\mathbf{Y}-\widehat{\mathbf{A}}'\mathbf{K}_\mathbf{A}\widehat{\mathbf{A}})\right)} \\ & \times \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}(\mathbf{A}-\widehat{\mathbf{A}})'\mathbf{K}_\mathbf{A}(\mathbf{A}-\widehat{\mathbf{A}})\right)}, \end{aligned}$$

where $\mathbf{K}_\mathbf{A} = \mathbf{V}_\mathbf{A}^{-1} + \mathbf{Z}'\Omega^{-1}\mathbf{Z}$ and $\widehat{\mathbf{A}} = \mathbf{K}_\mathbf{A}^{-1}(\mathbf{V}_\mathbf{A}^{-1}\mathbf{A}_0 + \mathbf{Z}'\Omega^{-1}\mathbf{Y})$. In the above derivations, we have "completed the square" and obtained:

$$\begin{aligned} & (\mathbf{A}-\mathbf{A}_0)'\mathbf{V}_\mathbf{A}^{-1}(\mathbf{A}-\mathbf{A}_0) + (\mathbf{Y}-\mathbf{Z}\mathbf{A})'\Omega^{-1}(\mathbf{Y}-\mathbf{Z}\mathbf{A}) \\ & \qquad = (\mathbf{A}-\widehat{\mathbf{A}})'\mathbf{K}_\mathbf{A}(\mathbf{A}-\widehat{\mathbf{A}}) + \mathbf{A}_0'\mathbf{V}_\mathbf{A}^{-1}\mathbf{A}_0 + \mathbf{Y}'\Omega^{-1}\mathbf{Y} - \widehat{\mathbf{A}}'\mathbf{K}_\mathbf{A}\widehat{\mathbf{A}}. \end{aligned}$$

If we let

$$\widehat{\mathbf{S}} = \mathbf{S}_0 + \mathbf{A}_0'\mathbf{V}_\mathbf{A}^{-1}\mathbf{A}_0 + \mathbf{Y}'\Omega^{-1}\mathbf{Y} - \widehat{\mathbf{A}}'\mathbf{K}_\mathbf{A}\widehat{\mathbf{A}},$$

then $(\mathbf{A},\Sigma\,|\,\mathbf{Y},\Omega)$ has a normal-inverse-Wishart distribution with parameters ν_0+T, $\widehat{\mathbf{S}}$, $\widehat{\mathbf{A}}$ and $\mathbf{K}_\mathbf{A}^{-1}$. We can then sample $(\mathbf{A},\Sigma\,|\,\mathbf{Y},\Omega)$ in two steps. First, sample Σ marginally from $(\Sigma\,|\,\mathbf{Y},\Omega) \sim \mathscr{IW}(\nu_0+T,\widehat{\mathbf{S}})$. Second, given the Σ sampled, simulate

$$(\mathrm{vec}(\mathbf{A})\,|\,\mathbf{Y},\Sigma,\Omega) \sim \mathscr{N}(\mathrm{vec}(\widehat{\mathbf{A}}),\Sigma\otimes\mathbf{K}_\mathbf{A}^{-1}).$$

As discussed in Section 2.2, we can sample from this normal distribution efficiently without explicitly computing the inverse $\mathbf{K}_\mathbf{A}^{-1}$.

Here we comment on a few computational details. Again, we need not compute the $T\times T$ inverse Ω^{-1} to obtain $\mathbf{K}_\mathbf{A}$, $\widehat{\mathbf{A}}$ or $\widehat{\mathbf{S}}$. As an example, consider computing the quadratic form $\mathbf{Z}'\Omega^{-1}\mathbf{Z}$. Let $\mathbf{C}_\Omega$ be the Cholesky factor of Ω such that $\mathbf{C}_\Omega\mathbf{C}_\Omega' = \Omega$. Then, $\mathbf{Z}'\Omega^{-1}\mathbf{Z}$ can be obtained via $\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}}$, where $\widetilde{\mathbf{Z}} = \mathbf{C}_\Omega\backslash\mathbf{Z}$.

This approach would work fine for an arbitrary Ω with dimension, say, less than 1000. For larger T, computing the Cholesky factor of Ω and performing the forward and backward substitution is likely to be time-consuming. Fortunately, for most models, Ω or Ω^{-1} are band matrices—i.e., sparse matrices whose nonzero elements are confined to a diagonal band. For example, Ω is diagonal—hence banded—for both the common stochastic volatility model and the t errors model. Moreover, Ω is banded for VARs with MA errors and Ω^{-1} is banded for AR errors.

This special structure of Ω or Ω^{-1} can be exploited to speed up computation. For instance, obtaining the Cholesky factor of a band $T\times T$ matrix with fixed bandwidth involves only $\mathscr{O}(T)$ operations (e.g., Golub and van Loan, 1983, p.156) as opposed to $\mathscr{O}(T^3)$ for a dense matrix of the same size. Similar computational savings can be obtained for operations such as multiplication, forward and backward substitution by using band matrix routines. We refer the readers to Chan (2013) for a more detailed discussion on computation involving band matrices.

Next, we take up various examples of Ω and provide the corresponding estimation details.

3.4.1 t Errors

As discussed in Section 3.2, a VAR with iid t errors falls within the framework in (18) with $\Omega = \text{diag}(\lambda_1, \dots, \lambda_T)$, where each λ_t follows an inverse-gamma distribution $(\lambda_t \,|\, \nu) \sim \mathcal{IG}(\nu/2, \nu/2)$. Unconditional on λ_t, $\mathbf{u}_t$ has a t distribution with degree of freedom parameter ν. Note that in this case Ω is diagonal and $\Omega^{-1} = \text{diag}(\lambda_1^{-1}, \dots, \lambda_T^{-1})$.

Let $p(\nu)$ denote the prior density function of ν. Then, posterior draws can be obtained by sequentially sampling from: 1) $p(\mathbf{A}, \Sigma \,|\, \mathbf{Y}, \Omega, \nu)$; 2) $p(\Omega \,|\, \mathbf{Y}, \mathbf{A}, \Sigma, \nu)$; and 3) $p(\nu \,|\, \mathbf{Y}, \mathbf{A}, \Sigma, \Omega)$. Step 1 can be implemented exactly as before. For Step 2, note that

$$p(\Omega \,|\, \mathbf{Y}, \mathbf{A}, \Sigma, \nu) = \prod_{t=1}^{T} p(\lambda_t \,|\, \mathbf{Y}, \mathbf{A}, \Sigma, \nu) \propto \prod_{t=1}^{T} \lambda_t^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2\lambda_t}\mathbf{u}_t'\Sigma^{-1}\mathbf{u}_t} \times \lambda_t^{-(\frac{\nu}{2}+1)} \mathrm{e}^{-\frac{\nu}{2\lambda_t}}$$

In other words, each λ_t is conditionally independent given other parameters and has an inverse-gamma distribution: $(\lambda_t \,|\, \mathbf{Y}, \mathbf{A}, \Sigma, \nu) \sim \mathcal{IG}((n+\nu)/2, (\mathbf{u}_t'\Sigma^{-1}\mathbf{u}_t + \nu)/2)$.

Lastly, ν can be sampled by an independence-chain Metropolis-Hastings step with the proposal distribution $\mathcal{N}(\widehat{\nu}, K_\nu^{-1})$, where $\widehat{\nu}$ is the mode of $\log p(\nu \,|\, \mathbf{Y}, \mathbf{A}, \Sigma, \Omega)$ and K_ν is the negative Hessian evaluated at the mode. For implementation details of this step, see Chan and Hsiao (2014).

3.4.2 Common Stochastic Volatility

Now, consider the common stochastic volatility model proposed in Carriero, Clark, and Marcellino (2016): $\mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \mathrm{e}^{h_t}\Sigma)$, where h_t follows an AR(1) process in (19). In this case $\Omega = \text{diag}(\mathrm{e}^{h_1}, \dots, \mathrm{e}^{h_T})$, which is also diagonal.

We assume independent truncated normal and inverse-gamma priors for ρ and σ_h^2: $\rho \sim \mathcal{N}(\rho_0, V_\rho)1(|\rho| < 1)$ and $\sigma_h^2 \sim \mathcal{IG}(\nu_{h0}, S_{h0})$. Then, posterior draws can be obtained by sampling from: 1) $p(\mathbf{A}, \Sigma \,|\, \mathbf{Y}, \Omega, \rho, \sigma_h^2)$; 2) $p(\Omega \,|\, \mathbf{Y}, \mathbf{A}, \Sigma, \rho, \sigma_h^2)$; 3) $p(\rho \,|\, \mathbf{Y}, \mathbf{A}, \Sigma, \Omega, \sigma_h^2)$; and 4) $p(\sigma_h^2 \,|\, \mathbf{Y}, \mathbf{A}, \Sigma, \Omega, \rho)$.

Step 1 again can be implemented exactly as before. For Step 2, note that

$$p(\Omega \,|\, \mathbf{Y}, \mathbf{A}, \Sigma, \rho, \sigma_h^2) = p(\mathbf{h} \,|\, \mathbf{Y}, \mathbf{A}, \Sigma, \rho, \sigma_h^2) \propto p(\mathbf{h} \,|\, \rho, \sigma_h^2) \prod_{t=1}^{T} p(\mathbf{y}_t \,|\, \mathbf{A}, \Sigma, h_t),$$

where $p(\mathbf{h} \,|\, \rho, \sigma_h^2)$ is a Gaussian density implied by the state equation,

$$\log p(\mathbf{y}_t \,|\, \mathbf{A}, \Sigma, h_t) = c_t - \frac{n}{2} h_t - \frac{1}{2} \mathrm{e}^{-h_t} \mathbf{u}_t'\Sigma^{-1}\mathbf{u}_t$$

and c_t is a constant not dependent on h_t. It is easy to check that

$$\frac{\partial}{\partial h_t}\log p(\mathbf{y}_t\,|\,\mathbf{A},\Sigma,h_t) = -\frac{n}{2}+\frac{1}{2}\mathrm{e}^{-h_t}\mathbf{u}_t'\Sigma^{-1}\mathbf{u}_t,$$
$$\frac{\partial^2}{\partial h_t^2}\log p(\mathbf{y}_t\,|\,\mathbf{A},\Sigma,h_t) = -\frac{1}{2}\mathrm{e}^{-h_t}\mathbf{u}_t'\Sigma^{-1}\mathbf{u}_t.$$

Then, one can implement a Newton-Raphson algorithm to obtain the mode of $\log p(\mathbf{h}\,|\,\mathbf{Y},\mathbf{A},\Sigma,\rho,\sigma_h^2)$ and compute the negative Hessian evaluated at the mode, which are denoted as $\widehat{\mathbf{h}}$ and $\mathbf{K}_{\mathbf{h}}$, respectively. Using $\mathscr{N}(\widehat{\mathbf{h}},\mathbf{K}_{\mathbf{h}}^{-1})$ as a proposal distribution, one can sample $\mathbf{h}$ directly using an acceptance-rejection Metropolis-Hastings step. We refer the readers to Chan (2017) and Chan and Jeliazkov (2009) for details. Finally, Steps 3 and 4 are standard and can be easily implemented (see., e.g., Chan and Hsiao, 2014).

3.4.3 MA(1) Errors

We now consider an example where Ω is not diagonal and we construct Ω using band matrices. More specifically, suppose each element of $\mathbf{u}_t$ follows the same MA(1) process:

$$u_{it} = \eta_{it}+\psi\eta_{i,t-1},$$

where $|\psi|<1$, $\eta_{it}\sim\mathscr{N}(0,1)$, and the process is initialized with $u_{i1}\sim\mathscr{N}(0,1+\psi^2)$. Stacking $\mathbf{u}_i=(u_{i1},\ldots,u_{iT})'$ and $\eta_i=(\eta_{i1},\ldots,\eta_{iT})'$, we can rewrite the MA(1) process as

$$\mathbf{u}_i = \mathbf{H}_\psi\eta_i,$$

where $\eta_i\sim\mathscr{N}(\mathbf{0},\mathbf{O}_\psi)$ with $\mathbf{O}_\psi=\mathrm{diag}(1+\psi^2,1,\ldots,1)$, and

$$\mathbf{H}_\psi = \begin{pmatrix} 1 & 0 & \cdots & 0\\ \psi & 1 & \cdots & 0\\ \vdots & \ddots & \ddots & \vdots\\ 0 & \cdots & \psi & 1\end{pmatrix}.$$

It follows that the covariance matrix of $\mathbf{u}_i$ is $\mathbf{H}_\psi\mathbf{O}_\psi\mathbf{H}_\psi'$. That is, $\Omega=\mathbf{H}_\psi\mathbf{O}_\psi\mathbf{H}_\psi'$ is a function of ψ only. Moreover, both $\mathbf{O}_\psi$ and $\mathbf{H}_\psi$ are band matrices. Notice also that for a general MA(q) process, one only needs to redefine $\mathbf{H}_\psi$ and $\mathbf{O}_\psi$ appropriately and the same procedure would apply.

Let $p(\psi)$ be the prior for ψ. Then, posterior draws can be obtained by sequentially sampling from: 1) $p(\mathbf{A},\Sigma\,|\,\mathbf{Y},\psi)$ and 2) $p(\psi\,|\,\mathbf{Y},\mathbf{A},\Sigma)$. Again, Step 1 can be carried out exactly the same as before. In implementing Step 1, we emphasize that products of the form $\mathbf{Z}'\Omega^{-1}\mathbf{Z}$ or $\mathbf{Z}'\Omega^{-1}\mathbf{Y}$ can be obtained without explicitly computing the inverse Ω^{-1}. Instead, since in this case Ω is a band matrix, its Cholesky factor $\mathbf{C}_\Omega$ can be obtained in $\mathscr{O}(T)$ operations. Then, to compute $\mathbf{Z}'\Omega^{-1}\mathbf{Z}$, one simply returns $\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}}$, where $\widetilde{\mathbf{Z}}=\mathbf{C}_\Omega\backslash\mathbf{Z}$.

For Step 2, $p(\psi \,|\, \mathbf{Y}, \mathbf{A}, \Sigma)$ is non-standard, but it can be evaluated quickly using the direct method in Chan (2013), which is more efficient than using the Kalman filter. Specifically, since the determinant $|\mathbf{H}_\psi| = 1$, it follows from (4) that the likelihood is given by

$$p(\mathbf{Y} \,|\, \mathbf{A}, \Sigma, \psi) = (2\pi)^{-\frac{Tn}{2}} |\Sigma|^{-\frac{T}{2}} (1+\psi^2)^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\widetilde{\mathbf{U}}'\mathbf{O}_\psi^{-1}\widetilde{\mathbf{U}}\right)},$$

where $\widetilde{\mathbf{U}} = \mathbf{H}_\psi^{-1}(\mathbf{Y} - \mathbf{Z}\mathbf{A})$, which can be obtained in $\mathscr{O}(T)$ operations since $\mathbf{H}_\psi$ is a band matrix. Therefore, $p(\psi \,|\, \mathbf{Y}, \mathbf{A}, \Sigma) \propto p(\mathbf{Y} \,|\, \mathbf{A}, \Sigma, \psi) p(\psi)$ can be evaluated quickly. Then, ψ is sampled using an independence-chain Metropolis-Hastings step as in Chan (2013).

3.4.4 AR(1) Errors

Here we consider an example where Ω is a full matrix, but Ω^{-1} is banded. Specifically, suppose each element of $\mathbf{u}_t$ follows the same AR(1) process:

$$u_{it} = \phi u_{i,t-1} + \eta_{it},$$

where $|\phi| < 1$, $\eta_{it} \sim \mathscr{N}(0,1)$, and the process is initialized with $u_{i1} \sim \mathscr{N}(0, 1/(1-\phi^2))$. Stacking $\mathbf{u}_i = (u_{i1}, \ldots, u_{iT})'$ and $\eta_i = (\eta_{i1}, \ldots, \eta_{iT})'$, we can rewrite the AR(1) process as

$$\mathbf{H}_\phi \mathbf{u}_i = \eta_i,$$

where $\eta_i \sim \mathscr{N}(\mathbf{0}, \mathbf{O}_\phi)$ with $\mathbf{O}_\phi = \mathrm{diag}(1/(1-\phi^2), 1, \ldots, 1)$, and

$$\mathbf{H}_\phi = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -\phi & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & -\phi & 1 \end{pmatrix}.$$

Since the determinant $|\mathbf{H}_\phi| = 1 \neq 0$, $\mathbf{H}_\phi$ is invertible. It follows that the covariance matrix of $\mathbf{u}_i$ is $\mathbf{H}_\phi^{-1}\mathbf{O}_\phi(\mathbf{H}_\phi')^{-1}$, or $\Omega^{-1} = \mathbf{H}_\phi'\mathbf{O}_\phi^{-1}\mathbf{H}_\phi$, where both $\mathbf{O}_\phi$ and $\mathbf{H}_\phi$ are band matrices.

Suppose we assume the truncated normal prior ϕ: $\phi \sim \mathscr{N}(\phi_0, V_\phi)1(|\phi| < 1)$. Then, posterior draws can be obtained by sampling from: 1) $p(\mathbf{A}, \Sigma \,|\, \mathbf{Y}, \phi)$; and 2) $p(\phi \,|\, \mathbf{Y}, \mathbf{A}, \Sigma)$. In implementing Step 1, products of the form $\mathbf{Z}'\Omega^{-1}\mathbf{Z}$ can be computed easily as the inverse Ω^{-1} is a band matrix.

For Step 2, $p(\phi \,|\, \mathbf{Y}, \mathbf{A}, \Sigma)$ is non-standard, but a good approximation can be obtained easily without numerical optimization. To that end, recall that

$$\mathbf{u}_t = \phi \mathbf{u}_{t-1} + \varepsilon_t,$$

where $\varepsilon_t \sim \mathscr{N}(\mathbf{0}, \Sigma)$, and the process is initialized by $\mathbf{u}_1 \sim \mathscr{N}(\mathbf{0}, \Sigma/(1-\phi^2))$. Then, consider the Gaussian proposal $\mathscr{N}(\widehat{\phi}, K_\phi^{-1})$, where $K_\phi = 1/V_\phi + \sum_{t=2}^T \mathbf{u}_{t-1}'\Sigma^{-1}\mathbf{u}_{t-1}$

and $\widehat{\phi} = K_{\phi}^{-1}(\phi_0/V_{\phi} + \sum_{t=2}^{T} \mathbf{u}_{t-1}' \Sigma^{-1} \mathbf{u}_t)$. With this proposal distribution, we can then implement an independence-chain Metropolis-Hastings step to sample ϕ.

4 Empirical Application: Forecasting with Large Bayesian VARs

In this section we consider a real-time macroeconomic forecasting exercise to illustrate the large Bayesian VARs and the associated estimation methods discussed in Section 3.

4.1 Data, Models and Priors

In our empirical application we use a real-time dataset considered in Chan (2018) that consists of 20 variables at quarterly frequency. The dataset includes a variety of standard macroeconomic and financial variables such as GDP, industrial production, inflation, interest rates and unemployment. The data are sourced from the Federal Reserve Bank of Philadelphia and the sample period is from 1964Q1 to 2015Q4. These variables are commonly used in applied work and are similar to the variables included in the large VARs in Banbura, Giannone, and Reichlin (2010) and Koop (2013). A detailed description of the variables and their transformations are provided in Appendix A.

We include a range of large Bayesian VARs combined with different prior specifications. For comparison, we also include a small Bayesian VAR using only four core variables: real GDP growth, industrial production, unemployment rate and PCE inflation. The full description other models are given in Table 1.

Table 1 A list of competing models.

Model	Description
BVAR-small	4-variable VAR with the Minnesota prior
BVAR-Minn	20-variable VAR with the Minnesota prior
BVAR-NCP	20-variable VAR with the natural conjugate prior
BVAR-IP	20-variable VAR with the independent prior
BVAR-SSVS	20-variable VAR with the SSVS prior
BVAR-CSV	20-variable VAR with a common stochastic volatility
BVAR-CSV-t	20-variable VAR with a common SV and t errors
BVAR-CSV-t-MA	20-variable VAR with a common SV and MA(1) t errors

Whenever possible, we choose the same priors for common parameters across models. For the Minnesota prior, we set $\beta_{\text{Minn}} = \mathbf{0}$ and the three hyperparameters for $\mathbf{V}_{\text{Minn}}$ are set to be $c_1 = 0.2^2, c_2 = 0.1^2$ and $c_3 = 10^2$. For the natural conjugate prior, we set $\mathbf{A}_0 = \mathbf{0}$ and the two hyperparameters for the covariance ma-

trix $\mathbf{V_A}$ are assumed to be $c_1 = 0.2^2$ and $c_2 = 10^2$. Moreover we set $\nu_0 = n+3$, $\mathbf{S}_0 = \text{diag}(s_1^2, \ldots, s_n^2)$, where s_i^2 denotes the standard OLS estimate of the error variance for the i-th equation.

For the common stochastic volatility model, we assume independent priors for σ_h^2 and ρ: $\sigma_h^2 \sim \mathcal{IG}(\nu_{h0}, S_{h0})$ and $\rho \sim \mathcal{N}(\rho_0, V_\rho)\mathbb{1}(|\rho| < 1)$, where we set $\nu_{h0} = 5$, $S_{h0} = 0.04$, $\rho_0 = 0.9$ and $V_\rho = 0.2^2$. These values imply that the prior mean of σ_h^2 is 0.1^2 and ρ is centered at 0.9. For the degree of freedom parameter ν under the t model, we consider a uniform prior on $(2, 50)$, i.e., $\nu \sim \mathcal{U}(2, 50)$. For the MA coefficient ψ under the MA model, we assume the truncated normal prior $\psi \sim \mathcal{N}(\psi_0, V_\psi)\mathbb{1}(|\psi| < 1)$ so that the MA process is invertible. We set $\psi_0 = 0$ and $V_\psi = 1$. The prior thus centers around 0 and has support within the interval $(-1, 1)$. Given the large prior variance, it is also relatively noninformative.

4.2 Forecast Evaluation Metrics

We perform a recursive out-of-sample forecasting exercise to evaluate the performance of the Bayesian VARs with different priors in terms of both point and density forecasts. We focus on four main variables: real GDP growth, industrial production, unemployment rate and PCE inflation.

We use each of the Bayesian VARs listed in Table 1 to produce both point and density m-step-ahead iterated forecasts with $m = 1$ and $m = 2$. Due to reporting lags, the real-time data vintage available at time t contains observations only up to quarter $t-1$. Hence, the forecasts are current quarter nowcasts and one-quarter-ahead forecasts. The evaluation period is from 1975Q1 to 2015Q4, and we use the 2017Q3 vintage to compute the actual outcomes.

Given the data up to time t, denoted as $\mathbf{Y}_{1:t}$, we obtain posterior draws given $\mathbf{Y}_{1:t}$. We then compute the predictive mean $\mathbb{E}(y_{i,t+m} \,|\, \mathbf{Y}_{1:t})$ as the point forecast for variable i, and the predictive density $p(y_{i,t+m} \,|\, \mathbf{Y}_{1:t})$ as the density forecast for the same variable. For many Bayesian VARs considered, neither the predictive mean nor the predictive density of $y_{i,t+m}$ can be computed analytically. If that is the case, we obtain them using predictive simulation. Next, we move one period forward and repeat the whole exercise with data $\mathbf{Y}_{1:t+1}$, and so on. These forecasts are then evaluated for $t = t_0, \ldots, T-m$.

For forecast evaluation metrics, let $y^{\text{o}}_{i,t+m}$ denote the actual value of the variable $y_{i,t+m}$. The metric used to evaluate the point forecasts is the root mean squared forecast error (RMSFE) defined as

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

To evaluate the density forecast $p(y_{i,t+m} \,|\, \mathbf{Y}_{1:t})$, one natural measure is the predictive likelihood $p(y_{i,t+m} = y^{\text{o}}_{i,t+m} \,|\, \mathbf{Y}_{1:t})$, i.e., the predictive density of $y_{i,t+m}$ evaluated at the actual value $y^{\text{o}}_{i,t+m}$. If the actual outcome $y^{\text{o}}_{i,t+m}$ is likely under the density

forecast, the value of the predictive likelihood will be large, and vise versa. See, e.g., Geweke and Amisano (2011) for a more detailed discussion of the predictive likelihood and its connection to the marginal likelihood. We evaluate the density forecasts using the average of log predictive likelihoods (ALPL):

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

For this metric, a larger value indicates better forecast performance.

4.3 Forecasting Results

For easy comparison, we report below the ratios of RMSFEs of a given model to those of the 4-variable Bayesian VAR using the core variables: real GDP growth, industrial production, unemployment rate and PCE inflation. Hence, values smaller than unity indicate better forecast performance than the benchmark. For the average of log predictive likelihoods, we report differences from that of the 4-variable Bayesian VAR. In this case, positive values indicate better forecast performance than the benchmark.

Tables 2–5 report the point and density forecast results for the four core variables. No single models or priors can outperform others for all variables in all horizons. However, there are a few consistent patterns in the forecasting results. First, consistent with the results in Banbura, Giannone, and Reichlin (2010) and Koop (2013), large VARs tend to forecast real variables better than the small VAR, whereas the small VAR does better than large models for PCE inflation in terms of point forecasts (see also Stock and Watson, 2007).

Table 2 Forecast performance relative to a 4-variable Bayesian VAR; GDP growth.

	relative RMSFE		relative ALPL	
	$m=1$	$m=2$	$m=1$	$m=2$
BVAR-Minn	0.95	0.98	−0.04	−0.10
BVAR-NCP	0.92	0.98	0.04	0.03
BVAR-IP	1.01	0.96	0.03	0.05
BVAR-SSVS	0.92	1.02	0.01	0.00
BVAR-CSV	0.95	0.94	0.13	0.09
BVAR-CSV-t	0.93	0.95	0.13	0.09
BVAR-CSV-t-MA	0.93	0.93	0.13	0.10

Table 3 Forecast performance relative to a 4-variable Bayesian VAR; industrial production.

	relative RMSFE		relative ALPL	
	$m = 1$	$m = 2$	$m = 1$	$m = 2$
BVAR-Minn	0.97	0.94	0.02	−0.02
BVAR-NCP	0.96	0.95	0.15	0.09
BVAR-IP	0.94	0.90	0.10	0.10
BVAR-SSVS	0.99	0.96	0.08	0.07
BVAR-CSV	0.89	0.90	0.27	0.17
BVAR-CSV-t	0.88	0.89	0.26	0.17
BVAR-CSV-t-MA	0.87	0.89	0.27	0.17

Table 4 Forecast performance relative to a 4-variable Bayesian VAR; unemployment rate.

	relative RMSFE		relative ALPL	
	$m = 1$	$m = 2$	$m = 1$	$m = 2$
BVAR-Minn	0.99	0.96	0.08	0.33
BVAR-NCP	0.99	0.96	0.11	0.30
BVAR-IP	1.02	0.99	−0.01	−0.03
BVAR-SSVS	1.02	1.01	−0.03	−0.07
BVAR-CSV	1.00	0.95	0.18	0.43
BVAR-CSV-t	0.99	0.95	0.16	0.40
BVAR-CSV-t-MA	0.98	0.96	0.16	0.37

Table 5 Forecast performance relative to a 4-variable Bayesian VAR; PCE inflation.

	relative RMSFE		relative ALPL	
	$m = 1$	$m = 2$	$m = 1$	$m = 2$
BVAR-Minn	1.04	1.06	−0.01	0.02
BVAR-NCP	1.04	1.06	−0.02	−0.01
BVAR-IP	1.02	1.00	−0.02	0.00
BVAR-SSVS	1.00	1.02	0.00	0.02
BVAR-CSV	1.04	1.04	0.09	0.11
BVAR-CSV-t	1.03	1.03	0.10	0.10
BVAR-CSV-t-MA	1.03	1.04	0.09	0.08

Second, among the four priors for large Bayesian VARs, the natural conjugate prior seems to perform well—even when it is not the best among the four, its performance is close to the best. See also a similar comparison in Koop (2013). Given that the natural conjugate prior can substantially speed up computations in posterior simulation, it might be justified to be used as the default in large systems.

Third, the results also show that large Bayesian VARs with more flexible error covariance structures tend to outperform the standard VARs. This is especially so for density forecasts. Our results are consistent with those in numerous studies, such as Clark (2011), D'Agostino, Gambetti, and Giannone (2013) and Clark and Ravazzolo

(2015), which find that small Bayesian VARs with stochastic volatility outperform their counterparts with only constant variance. Fourth, even though BVAR-CSV tends to forecast very well, in many instances its forecast performance can be further improved by using the t error distribution or adding an MA component.

Overall, these forecasting results show that large Bayesian VARs tend to forecast well relative to small systems. Moreover, their forecast performance can be further enhanced by allowing for stochastic volatility, heavy-tailed and serially correlated errors.

5 Further Reading

Koop and Korobilis (2010) and Karlsson (2013) are two excellent review papers that cover many of the topics discussed in Section 2. The presentation of the large Bayesian VARs with time-varying volatility, heavy-tailed distributions and serial dependent errors in Section 3 closely follows Chan (2018).

Developing large, flexible Bayesian VARs is an active research area and there are many different approaches. For instance, Koop and Korobilis (2013) consider an approximate method for forecasting using large time-varying parameter Bayesian VARs. Chan, Eisenstat, and Koop (2016) estimate a Bayesian VARMA containing 12 variables. Carriero, Clark, and Marcellino (2015b) propose an efficient method to estimate a 125-variable VAR with a standard stochastic volatility specification. Koop, Korobilis, and Pettenuzzo (2017) consider compressed VARs based on the random projection method. Ahelegbey, Billio, and Casarin (2016b,a) develop Bayesian graphical models for large VARs. Gefang, Koop, and Poon (2019) use variational approximation for estimating large Bayesian VARs with stochastic volatility.

Appendix A: Data

The real-time dataset for our forecasting application includes 13 macroeconomic variables that are frequently revised and 7 financial or survey variables that are not revised. The list of variables is given in Table 6. They are sourced from the Federal Reserve Bank of Philadelphia and cover the quarters from 1964Q1 to 2015Q4. All monthly variables are converted to quarterly frequency by averaging the three monthly values within the quarter.

Table 6 Description of variables used in the recursive forecasting exercise.

Variable	Transformation
Real GNP/GDP	$400\Delta \log$
Real Personal Consumption Expenditures: Total	$400\Delta \log$
Real Gross Private Domestic Investment: Nonresidential	$400\Delta \log$
Real Gross Private Domestic Investment: Residential	$400\Delta \log$
Real Net Exports of Goods and Services	no transformation
Nominal Personal Income	$400\Delta \log$
Industrial Production Index: Total	$400\Delta \log$
Unemployment Rate	no transformation
Nonfarm Payroll Employment	$400\Delta \log$
Indexes of Aggregate Weekly Hours: Total	$400\Delta \log$
Housing Starts	$400\Delta \log$
Price Index for Personal Consumption Expenditures, Constructed	$400\Delta \log$
Price Index for Imports of Goods and Services	$400\Delta \log$
Effective Federal Funds Rate	no transformation
1-Year Treasury Constant Maturity Rate	no transformation
10-Year Treasury Constant Maturity Rate	no transformation
Moody's Seasoned Baa Corporate Bond Minus Federal Funds Rate	no transformation
ISM Manufacturing: PMI Composite Index	no transformation
ISM Manufacturing: New Orders Index	no transformation
S&P 500	$400\Delta \log$

Appendix B: Sampling from the Matrix Normal Distribution

Suppose we wish to sample from $\mathcal{N}(\text{vec}(\widehat{\mathbf{A}}), \Sigma \otimes \mathbf{K}_{\mathbf{A}}^{-1})$. Let $\mathbf{C}_{\mathbf{K}_{\mathbf{A}}}$ and $\mathbf{C}_{\Sigma}$ be the Cholesky decompositions of $\mathbf{K}_{\mathbf{A}}$ and Σ respectively. We wish to show that if we construct

$$\mathbf{W}_1 = \widehat{\mathbf{A}} + (\mathbf{C}'_{\mathbf{K}_{\mathbf{A}}} \backslash \mathbf{U})\mathbf{C}'_{\Sigma},$$

where $\mathbf{U}$ is a $k \times n$ matrix of independent $\mathcal{N}(0,1)$ random variables, then $\text{vec}(\mathbf{W}_1)$ has the desired distribution. To that end, we make use of some standard results on the matrix normal distribution (see, e.g., Bauwens, Lubrano, and Richard, 1999, pp. 301-302).

A $p \times q$ random matrix $\mathbf{W}$ is said to have a **matrix normal distribution** $\mathcal{MN}(\mathbf{M}, \mathbf{Q} \otimes \mathbf{P})$ for covariance matrices $\mathbf{P}$ and $\mathbf{Q}$ of dimensions $p \times p$ and $q \times q$, respectively, if $\text{vec}(\mathbf{W}) \sim \mathcal{N}(\text{vec}(\mathbf{M}), \mathbf{Q} \otimes \mathbf{P})$. Now suppose $\mathbf{W} \sim \mathcal{MN}(\mathbf{M}, \mathbf{Q} \otimes \mathbf{P})$ and define $\mathbf{V} = \mathbf{CWD} + \mathbf{E}$. Then, $\mathbf{V} \sim \mathcal{MN}(\mathbf{CMD} + \mathbf{E}, (\mathbf{D}'\mathbf{QD}) \otimes (\mathbf{CPC}'))$.

Recall that $\mathbf{U}$ is a $k \times n$ matrix of independent $\mathcal{N}(0,1)$ random variables. Hence, $\mathbf{U} \sim \mathcal{MN}(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{I}_k)$. Using the previous result with $\mathbf{C} = (\mathbf{C}'_{\mathbf{K}_{\mathbf{A}}})^{-1}$, $\mathbf{D} = \mathbf{C}'_{\Sigma}$ and $\mathbf{E} = \widehat{\mathbf{A}}$, it is easy to see that $\mathbf{W}_1 \sim \mathcal{MN}(\widehat{\mathbf{A}}, \Sigma \otimes \mathbf{K}_{\mathbf{A}}^{-1})$. Finally, by definition we have $\text{vec}(\mathbf{W}_1) \sim \mathcal{N}(\text{vec}(\widehat{\mathbf{A}}), \Sigma \otimes \mathbf{K}_{\mathbf{A}}^{-1})$.

References

AHELEGBEY, D. F., M. BILLIO, AND R. CASARIN (2016a): "Bayesian Graphical Models for Structural Vector Autoregressive Processes," *Journal of Applied Econometrics*, 31(2), 357–386.

——— (2016b): "Sparse Graphical Multivariate Autoregression: A Bayesian approach," *Annals of Economics and Statistics*, 123/124, 1–30.

BANBURA, M., D. GIANNONE, AND L. REICHLIN (2010): "Large Bayesian vector auto regressions," *Journal of Applied Econometrics*, 25(1), 71–92.

BAUWENS, L., M. LUBRANO, AND J. RICHARD (1999): *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press, New York.

CARRIERO, A., T. E. CLARK, AND M. MARCELLINO (2015a): "Bayesian VARs: Specification Choices and Forecast Accuracy," *Journal of Applied Econometrics*, 30(1), 46–73.

——— (2015b): "Large Vector Autoregressions with asymmetric priors and time varying volatilities," *Working Paper, School of Economics and Finance, Queen Mary University of London*.

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

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.

——— (2017): "The Stochastic Volatility in Mean Model with Time-Varying Parameters: An Application to Inflation Modeling," *Journal of Business and Economic Statistics*, 35(1), 17–28.

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

CHAN, J. C. C., E. EISENSTAT, AND G. KOOP (2016): "Large Bayesian VARMAs," *Journal of Econometrics*, 192(2), 374–390.

CHAN, J. C. C., AND C. Y. L. HSIAO (2014): "Estimation of Stochastic Volatility Models with Heavy Tails and Serial Dependence," in *Bayesian Inference in the Social Sciences*, ed. by I. Jeliazkov, and X.-S. Yang. John Wiley & Sons, Hoboken.

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, 101–120.

CHIU, C. J., H. MUMTAZ, AND G. PINTER (2017): "Forecasting with VAR models: Fat tails and stochastic volatility," *International Journal of Forecasting*, 33(4), 1124–1143.

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

CLARK, T. E., AND F. RAVAZZOLO (2015): "Macroeconomic Forecasting Performance under alternative specifications of time-varying volatility," *Journal of Applied Econometrics*, 30(4), 551–575.

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., 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.

ELTOFT, T., T. KIM, AND T. LEE (2006a): "Multivariate Scale Mixture of Gaussians Modeling," in *Independent Component Analysis and Blind Signal Separation*, ed. by J. Rosca, D. Erdogmus, J. Principe, and S. Haykin, vol. 3889 of *Lecture Notes in Computer Science*, pp. 799–806. Springer Berlin Heidelberg.

——— (2006b): "On the multivariate Laplace distribution," *Signal Processing Letters, IEEE*, 13(5), 300–303.

FORNI, M., M. HALLIN, M. LIPPI, AND L. REICHLIN (2003): "Do financial variables help forecasting inflation and real activity in the euro area?," *Journal of Monetary Economics*, 50(6), 1243 – 1255.

GEFANG, D., G. KOOP, AND A. POON (2019): "Variational Bayesian inference in large Vector Autoregressions with hierarchical shrinkage," *CAMA Working Paper*.

GEORGE, E. I., D. SUN, AND S. NI (2008): "Bayesian stochastic search for VAR model restrictions," *Journal of Econometrics*, 142(1), 553–580.

GEWEKE, J. (1993): "Bayesian Treatment of the Independent Student-*t* Linear Model," *Journal of Applied Econometrics*, 8(S1), S19–S40.

GEWEKE, J., AND G. AMISANO (2011): "Hierarchical Markov Normal Mixture Models with Applications to Financial Asset Returns," *Journal of Applied Econometrics*, 26, 1–29.

GOLUB, G. H., AND C. F. VAN LOAN (1983): *Matrix computations*. Johns Hopkins University Press, Baltimore.

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.

KOOP, G., D. KOROBILIS, AND D. PETTENUZZO (2017): "Bayesian compressed vector autoregression," *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.

MUMTAZ, H. (2016): "The Evolving Transmission of Uncertainty Shocks in the United Kingdom," *Econometrics*, 4(1), 16.

MUMTAZ, H., AND K. THEODORIDIS (2017): "The Changing Transmission of Uncertainty Shocks in the U.S.," *Journal of Business and Economic Statistics*.

POON, A. (2018): "Assessing the synchronicity and nature of Australian state business cycles," *Economic Record*, 94(307), 372–390.

PRIMICERI, G. E. (2005): "Time Varying Structural Vector Autoregressions and Monetary Policy," *Review of Economic Studies*, 72(3), 821–852.

SIMS, C. A. (1980): "Macroeconomics and reality," *Econometrica*, 48, 1–48.

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

——— (2007): "Why has U.S. inflation become harder to forecast?," *Journal of Money Credit and Banking*, 39, 3–33.