Large Bayesian VARs: A Flexible Kronecker Error Covariance Structure

Joshua C.C. Chan*
Research School of Economics,
Australian National University

September 2015

Abstract

We introduce a class of large Bayesian vector autoregressions (BVARs) that allows for non-Gaussian, heteroscedastic and serially dependent innovations. To make estimation computationally tractable, we exploit a certain Kronecker structure of the likelihood implied by this class of models. We propose a unified approach for estimating these models using Markov chain Monte Carlo (MCMC) methods. In an application that involves 20 macroeconomic variables, we find that these BVARs with more flexible covariance structures outperform the standard variant with independent, homoscedastic Gaussian innovations in both in-sample model-fit and out-of-sample forecast performance.

Keywords: stochastic volatility, non-Gaussian, ARMA, forecasting

JEL classification codes: C11, C51, C53

*Financial support by the Australian Research Council via a Discovery Early Career Researcher Award (DE150100795) is gratefully acknowledged. We would also like to thank Gary Koop and Todd Clark for their insightful comments.

1 Introduction

Vector autoregressions (VARs) are widely used for macroeconomic forecasting and structural analysis. VARs tend to have a lot of parameters, and Bayesian methods that formally incorporate prior information to provide shrinkage are often found to greatly improve forecast performance (e.g., Doan, Litterman, and Sims, 1984; Litterman, 1986). Until recently, most empirical work had considered only small systems that rarely include more than ten dependent variables. This has changed since the seminal work of Banbura, Giannone, and Reichlin (2010), who find that large Bayesian VARs (BVARs) with more than 20 dependent variables forecast better than small VARs. This has generated a rapidly expanding literature on using large BVARs for forecasting; recent papers include Carriero, Kapetanios, and Marcellino (2009), Koop (2013) and Carriero, Clark, and Marcellino (2015b). Large BVARs 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 is a wide variety of extensions of small VARs that take into account important features of macroeconomic data, such as time-varying volatility (Cogley and Sargent, 2005; Primiceri, 2005). Some recent papers have considered similar extensions for large BVARs. For example, Koop and Korobilis (2013) propose an approximate method for forecasting using large time-varying parameter BVARs. Chan, Eisenstat, and Koop (2015) estimate a Bayesian VARMA containing 12 variables. Carriero, Clark, and Marcellino (2015a) propose a fast algorithm to estimate a large BVAR with a common stochastic volatility. These extensions are all found to outperform BVARs with homoscedastic and independent innovations.

We contribute to this emerging literature on flexible large BVARs by proposing a framework that allows for non-Gaussian, heteroscedastic and serially dependent innovations. More specifically, we build upon the computational approach in Carriero et al. (2015a) that exploits a certain Kronecker structure of the likelihood implied by their model to speed up the sampling of the large dimensional VAR coefficients. As they emphasize in their paper, without this Kronecker structure it is computationally intractable to estimate BVARs above a handful of variables. It turns out that the same computational shortcut can be applied to a much wider class of models. We further improve upon their algorithm so that it can be applied to much larger systems.

We first explicitly state the required Kronecker structure of the likelihood for the proposed approach to be applicable. We then give a few examples that can be used to model important features of macroeconomic data—these include models with heavy-tailed innovations, autoregressive moving average innovations, and the common stochastic volatility model of Carriero et al. (2015a). The proposed framework also extends the univariate moving average stochastic volatility models of Chan (2013) to a multivariate setting. All these diverse models can be estimated using a unified approach.

In our application we consider a dataset of 20 quarterly variables—it includes a variety

of standard macroeconomic variables such as GDP, inflation, interest rates, and money supply. To illustrate the usefulness of the proposed framework, we fit this dataset using BVARs with various flexible error covariance structures. In particular, we consider three types of innovations: t innovations, innovations with a common stochastic volatility and moving average innovations. The full sample estimation results show that all these features are useful in improving the performance of the standard BVAR. The most important gain is to allow for a common stochastic volatility, followed by using t innovations and adding a moving average component. In addition, one can further improve the performance of the common stochastic volatility model by allowing for t or MA innovations. In a recursive out-of-sample forecasting exercise, we show that these more flexible BVARs also provide better point and density forecasts.

The rest of this paper is organized as follows. Section 2 first introduces a general framework for modeling the error covariance structure that lends itself to fast computation. We then discuss in Section 3 a unified approach to estimate these more flexible BVARs using Markov chain Monte Carlo (MCMC) methods. Section 4 considers an application that involves 20 macroeconomic variables. We first present full sample estimation results and results from a model comparison exercise using the marginal likelihood. It is followed by a recursive out-of-sample forecasting exercise that assesses the performance of the proposed BVARs. Lastly, Section 5 concludes and briefly discusses some future research directions.

2 Covariance with a General Kronecker Structure

In this section we introduce a general framework for modeling the covariance structure of a large BVAR. Specifically, we consider a class of covariance matrices with a certain Kronecker structure that would allow us to model non-Gaussian, heteroscedastic and serially dependent innovations.

To set the stage, let $\mathbf{y}_t$ be an $n \times 1$ vector of variables that is observed over the periods $t = 1, \ldots, T$. Consider the following generic VAR(p) model:

$$\mathbf{y}_t = \mathbf{a}_0 + \mathbf{A}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{A}_p \mathbf{y}_{t-p} + \mathbf{u}_t,$$

where $\mathbf{a}_0$ is an $n \times 1$ vector of intercepts and $\mathbf{A}_1, \ldots, \mathbf{A}_p$ are all $n \times n$ coefficient matrices. Let $\mathbf{x}_t' = (1, \mathbf{y}_{t-1}', \ldots, \mathbf{y}_{t-p}')$ be a $1 \times k$ vector of an intercept and lags with $k = 1 + np$. Then, stacking the observations over $t = 1, \ldots, T$, we have

$$\mathbf{Y} = \mathbf{X}\mathbf{A} + \mathbf{U}, \tag{1}$$

where $\mathbf{A} = (\mathbf{a}_0, \mathbf{A}_1, \ldots, \mathbf{A}_p)'$ is $k \times n$, and the matrices $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{U}$ are respectively of dimensions $T \times n$, $T \times k$ and $T \times n$. In a standard VAR the innovations $\mathbf{u}_1, \ldots, \mathbf{u}_T$ are assumed to be independent and identically distributed (iid) as $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. More succinctly, we write $\text{vec}(\mathbf{U}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{I}_T)$, where $\mathbf{\Sigma}$ is an $n \times n$ covariance matrix, $\mathbf{I}_T$ is the identity

matrix of dimension T, $\otimes$ is the Kronecker product and the vec$(\cdot)$ operator converts the matrix into a column vector by stacking the columns.

Here we consider the following more general covariance structure:

$$\text{vec}(\mathbf{U}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{\Omega}), \tag{2}$$

where $\mathbf{\Omega}$ is a $T \times T$ covariance matrix. Intuitively, we separately model the cross-sectional and serial covariance structures of $\mathbf{Y}$, which are governed by $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ respectively. By choosing a suitable serial covariance structure $\mathbf{\Omega}$, the model in (1)–(2) includes a wide variety of flexible specifications. Below we list a few examples.

1. **Non-Gaussian innovations**. Since many distributions can be written as a scale mixture of Gaussian distributions, the proposed framework accommodates various commonly-used non-Gaussian distributions.

   To see this, let $\mathbf{\Omega} = \text{diag}(\lambda_1, \ldots, \lambda_T)$. 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 $\mathbf{\Sigma}$ 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 $\alpha\mathbf{\Sigma}$ (Eltoft, Kim, and Lee, 2006a). Other scale mixtures of Gaussian distributions can be defined similarly. For additional examples, see, e.g., Eltoft, Kim, and Lee (2006b).

2. **Heteroscedastic innovations**. Time-varying volatility can be modeled by specifying a suitable diagonal $\mathbf{\Omega}$. For example, the BVAR with a common drifting volatility considered in Carriero et al. (2015a) is nested within the proposed framework. Specifically, they consider $\mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \mathrm{e}^{h_t}\mathbf{\Sigma})$, where the log volatility follows an AR(1) process:

   $$h_t = \rho h_{t-1} + \varepsilon_t^h \tag{3}$$

   with $\varepsilon_t^h \sim \mathcal{N}(0, \sigma_h^2)$ and $|\rho| < 1$. This model falls within the proposed framework with $\mathbf{\Omega} = \text{diag}(\mathrm{e}^{h_1}, \ldots, \mathrm{e}^{h_T})$.

3. **Serially dependent innovations**. Innovations with ARMA(p, q) structure can be easily handled. For example, suppose $\mathbf{u}_t$ follows the following MA(2) process:

   $$\mathbf{u}_t = \boldsymbol{\varepsilon}_t + \psi_1 \boldsymbol{\varepsilon}_{t-1} + \psi_2 \boldsymbol{\varepsilon}_{t-2},$$

   where $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, ψ_1 and ψ_2 satisfy the invertibility conditions. This is a special

case of the proposed framework with

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

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

In the next section we will discuss how one can construct $\boldsymbol{\Omega}$ efficiently from elementary matrices.

Of course, these are just a few simple examples. More elaborate covariance structures can be constructed by combining different elements discussed above. For example, suppose $\mathbf{u}_t$ follows an MA(1) stochastic volatility process of the form:

$$\mathbf{u}_t = \boldsymbol{\varepsilon}_t + \psi_1 \boldsymbol{\varepsilon}_{t-1},$$

where $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathrm{e}^{h_t}\boldsymbol{\Sigma})$ and h_t has an AR(1) process as in (3). This is a multivariate extension of the univariate moving average stochastic volatility models proposed in Chan (2013). This model is a special case of the general framework with

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

While the modeling framework in (1)–(2) is flexible and includes various recently proposed models as special cases, it is important to understand its limitations. One crucial assumption embedded in (2) is that each element of the innovation $\mathbf{u}_t$ must have the same univariate time series model (though their variances can be different). That means our framework cannot accommodate, for example, a general MA(1) process of the form

$$\mathbf{u}_t = \boldsymbol{\varepsilon}_t + \boldsymbol{\Psi}_1 \boldsymbol{\varepsilon}_{t-1},$$

where $\boldsymbol{\Psi}_1$ is an $n \times n$ matrix of coefficients. This is because in this case the covariance matrix of $\mathrm{vec}(\mathbf{U})$ does not have a Kronecker structure. Consequently, some of the analytical results we derive next would not hold.

3 Bayesian Estimation

In this section we discuss the estimation of the proposed BVARs with the covariance structure in (2) using MCMC methods. We first introduce a fast and simple way to jointly sample both the VAR coefficients $\mathbf{A}$ and the cross-sectional covariance matrix $\mathbf{\Sigma}$ for an arbitrary serial covariance matrix $\mathbf{\Omega}$. Then, we take up various examples of $\mathbf{\Omega}$ and provide estimation details for tackling each case.

3.1 Posterior Analysis

We first provide a general framework to estimate the BVAR in (1)–(2). Here we leave the covariance structure $\mathbf{\Omega}$ unspecified and discuss how one can efficiently sample $\mathbf{A}$ and $\mathbf{\Sigma}$ in one block. In the next subsection we will consider various examples of $\mathbf{\Omega}$ and how this general framework can be modified to handle those cases.

First, note that the likelihood implied by the model in (1)–(2) is given by

$$p(\mathbf{Y} \mid \mathbf{A}, \mathbf{\Sigma}, \mathbf{\Omega}) = (2\pi)^{-\frac{Tn}{2}} |\mathbf{\Sigma}|^{-\frac{T}{2}} |\mathbf{\Omega}|^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}(\mathbf{Y}-\mathbf{XA})'\mathbf{\Omega}^{-1}(\mathbf{Y}-\mathbf{XA})\right)}, \tag{4}$$

where $\mathrm{tr}(\cdot)$ is the trace operator. Consider a prior of the form $p(\mathbf{A}, \mathbf{\Sigma}, \mathbf{\Omega}) = p(\mathbf{A}, \mathbf{\Sigma})p(\mathbf{\Omega})$, i.e., the parameter blocks $(\mathbf{A}, \mathbf{\Sigma})$ and $\mathbf{\Omega}$ are *a priori* independent. For $(\mathbf{A}, \mathbf{\Sigma})$, we adopt a standard normal-inverse-Wishart prior (see, e.g., Kadiyala and Karlsson, 1997):

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

with joint density function

$$p(\mathbf{A}, \mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-\frac{\nu_0+n+k}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{S}_0)} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}(\mathbf{A}-\mathbf{A}_0)'\mathbf{V}_{\mathbf{A}}^{-1}(\mathbf{A}-\mathbf{A}_0)\right)}. \tag{5}$$

The prior covariance matrix $\mathbf{V}_{\mathbf{A}}$ is chosen to induce shrinkage. The exact form is given in Section 4. Here we leave it unspecified.

Given the natural conjugate prior for $(\mathbf{A}, \mathbf{\Sigma})$, posterior draws can be obtained by sequentially sampling from: 1) $p(\mathbf{A}, \mathbf{\Sigma} \mid \mathbf{Y}, \mathbf{\Omega})$; and 2) $p(\mathbf{\Omega} \mid \mathbf{Y}, \mathbf{A}, \mathbf{\Sigma})$. Depending on the covariance structure $\mathbf{\Omega}$, 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. Various examples are given in the next subsection.

Here we describe how one can implement Step 1 of sampling from the high-dimensional density $p(\mathbf{A}, \mathbf{\Sigma} \mid \mathbf{Y}, \mathbf{\Omega})$ efficiently. It is well known that when $\mathbf{\Omega} = \mathbf{I}_T$, $(\mathbf{A}, \mathbf{\Sigma} \mid \mathbf{Y})$ has a normal-inverse-Wishart distribution—in this case analytical results are available and no simulation is necessary. It turns out that the same derivations go through even with an

arbitrary covariance matrix $\mathbf{\Omega}$. To see this, it follows from (4) and (5) that

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

where

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

and we have used the fact that

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

In other words, $(\mathbf{A}, \mathbf{\Sigma} \,|\, \mathbf{Y}, \mathbf{\Omega})$ has a normal-inverse-Wishart distribution with parameters $\nu_0 + T$, $\widehat{\mathbf{S}}$, $\widehat{\mathbf{A}}$ and $\mathbf{K}_\mathbf{A}^{-1}$, where

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

Hence, we can sample $(\mathbf{A}, \mathbf{\Sigma} \,|\, \mathbf{Y}, \mathbf{\Omega})$ in two steps. First, we sample $\mathbf{\Sigma}$ marginally from $(\mathbf{\Sigma} \,|\, \mathbf{Y}, \mathbf{\Omega}) \sim \mathcal{IW}(\widehat{\mathbf{S}}, \nu_0 + T)$. Then, given the $\mathbf{\Sigma}$ drawn we sample

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

Since the covariance matrix $\mathbf{\Sigma} \otimes \mathbf{K}_\mathbf{A}^{-1}$ is of dimension $nk = n(np+1)$, sampling from this high-dimensional density using conventional methods would involve $\mathcal{O}(n^6p^3)$ operations. This can be very time consuming when n is large. We adopt a computational shortcut considered in Carriero et al. (2015a) to our setting. More specifically, we exploit the Kronecker structure $\mathbf{\Sigma} \otimes \mathbf{K}_\mathbf{A}^{-1}$ to speed up computation. Consequently, we can drastically reduce the complexity of the problem to $\mathcal{O}(n^3p^3)$ operations. We further improve upon this approach by avoiding the computation of the inverse of the $k \times k$ matrix $\mathbf{K}_\mathbf{A}$.

To that end, we introduce the following notations: given a lower (upper) triangular non-singular matrix $\mathbf{B}$ and a conformable vector $\mathbf{c}$, let $\mathbf{B}\backslash\mathbf{c}$ denote the unique solution to the triangular system $\mathbf{B}\mathbf{z} = \mathbf{c}$ obtained by forward (backward) substitution, i.e., $\mathbf{B}\backslash\mathbf{c} = \mathbf{B}^{-1}\mathbf{c}$.[1] The number of operations needed is of the same order as computing the multiplication $\mathbf{B}\mathbf{c}$. 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

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

by forward followed by backward substitution. By construction,

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

[1] 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`.

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

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

where $\mathbf{Z}$ is a $k \times n$ matrix of independent $\mathcal{N}(0,1)$ random variables. In the Appendix we show that

$$\text{vec}(\mathbf{W}_1) \sim \mathcal{N}(\text{vec}(\widehat{\mathbf{A}}), \mathbf{\Sigma} \otimes \mathbf{K}_{\mathbf{A}}^{-1})$$

as desired.

In what follows, we comment on a few computational details. First, one need not compute the $T \times T$ inverse $\mathbf{\Omega}^{-1}$ to obtain $\mathbf{K_A}$, $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{S}}$. For example, one can calculate $\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}$ by $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$, where $\widetilde{\mathbf{X}} = \mathbf{C}_{\mathbf{\Omega}} \backslash \mathbf{X}$ and $\mathbf{C}_{\mathbf{\Omega}}$ is the Cholesky factor of $\mathbf{\Omega}$ such that $\mathbf{C}_{\mathbf{\Omega}}\mathbf{C}'_{\mathbf{\Omega}} = \mathbf{\Omega}$. This approach would work fine for an arbitrary $\mathbf{\Omega}$ with dimension, say, less than 1000—this case includes most quarterly and monthly macroeconomic datasets. For a larger T, computing the Cholesky factor of $\mathbf{\Omega}$ and performing the forward and backward substitution is likely to be time-consuming.

Fortunately, for many interesting cases, $\mathbf{\Omega}$ (or $\mathbf{\Omega}^{-1}$) are band matrices—i.e., sparse matrices whose nonzero elements are confined to a diagonal band. For example, $\mathbf{\Omega}$ is diagonal for both t innovations and the case of a common stochastic volatility. Moreover, $\mathbf{\Omega}$ is banded for MA innovations and $\mathbf{\Omega}^{-1}$ is banded for AR innovations. We will discuss these examples in more detail in Section 3.2. This special structure of $\mathbf{\Omega}$ (or $\mathbf{\Omega}^{-1}$) 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 $\mathcal{O}(T)$ operations (e.g., Golub and van Loan, 1983, p.156) as opposed to $\mathcal{O}(T^3)$ for a full matrix of the same size.[2] We refer the readers to Chan (2013) for a more detailed discussion on computation involving band matrices.

3.2 Examples for $\mathbf{\Omega}$

In this section we consider various specific examples of $\mathbf{\Omega}$ and discuss how one can augment the general sampling scheme above to handle each case.

Example 1. Independent t innovations

As discussed in Section 3.1, the case of iid t distributed innovations falls within the proposed framework. Specifically, if we assume $\mathbf{\Omega} = \text{diag}(\lambda_1, \ldots, \lambda_T)$ and each λ_t follows an inverse-gamma distribution $(\lambda_t \,|\, \nu) \sim \mathcal{IG}(\nu/2, \nu/2)$, then marginally $\mathbf{u}_t$ has a t distribution with degree of freedom parameter ν. Note that in this case $\mathbf{\Omega}$ is diagonal and $\mathbf{\Omega}^{-1} = \text{diag}(\lambda_1^{-1}, \ldots, \lambda_T^{-1})$.

[2]Similar computational savings can be generated for operations such as multiplication, forward and backward substitution by using band matrix routines, which are implemented in standard packages such as MATLAB, GAUSS and R.

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

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

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

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 $\log p(\nu \,|\, \mathbf{Y}, \mathbf{A}, \mathbf{\Sigma}, \mathbf{\Omega})$ and K_ν is the negative Hessian evaluated at the mode. For implementation details of this step, see Chan and Hsiao (2014).

Example 2. Independent innovations with a common stochastic volatility

Next, consider the common drifting volatility model proposed in Carriero et al. (2015a): $\mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \mathrm{e}^{h_t}\mathbf{\Sigma})$, where h_t follows an AR(1) process in (3), which is reproduced here for convenience: $h_t = \rho h_{t-1} + \varepsilon_t^h$, where $\varepsilon_t^h \sim \mathcal{N}(0, \sigma_h^2)$. This model falls within the proposed framework with $\mathbf{\Omega} = \mathrm{diag}(\mathrm{e}^{h_1}, \ldots, \mathrm{e}^{h_T})$, 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 $\sigma_h^2 \sim \mathcal{IG}(\nu_h, S_h)$. Then, posterior draws can be obtained by sampling from: 1) $p(\mathbf{A}, \mathbf{\Sigma} \,|\, \mathbf{Y}, \mathbf{\Omega}, \rho, \sigma_h^2)$; 2) $p(\mathbf{\Omega} \,|\, \mathbf{Y}, \mathbf{A}, \mathbf{\Sigma}, \rho, \sigma_h^2)$; 3) $p(\rho \,|\, \mathbf{Y}, \mathbf{A}, \mathbf{\Sigma}, \mathbf{\Omega}, \sigma_h^2)$; and 4) $p(\sigma_h^2 \,|\, \mathbf{Y}, \mathbf{A}, \mathbf{\Sigma}, \mathbf{\Omega}, \rho)$.

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

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

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

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

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

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

Then, one can implement a Newton-Raphson algorithm to obtain the mode of the log density $\log p(\mathbf{h} \,|\, \mathbf{Y}, \mathbf{A}, \mathbf{\Sigma}, \rho, \sigma_h^2)$ and the negative Hessian evaluated at the mode, which are denoted as $\widehat{\mathbf{h}}$ and $\mathbf{K}_\mathbf{h}$, respectively. Using $\mathcal{N}(\widehat{\mathbf{h}}, \mathbf{K}_\mathbf{h}^{-1})$ as a proposal distribution, one can sample $\mathbf{h}$ directly using an acceptance-rejection Metropolis-Hastings step. We refer the readers to Chan (2015) for details. Finally, Steps 3 and 4 are standard and can be easily implemented (see., e.g., Chan and Hsiao, 2014).

Example 3. MA(1) innovations

We now consider an example where $\mathbf{\Omega}$ is not diagonal and we construct $\mathbf{\Omega}$ 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 $|\psi| < 1$, $\eta_{it} \sim \mathcal{N}(0,1)$, and the process is initialized with $u_{i1} \sim \mathcal{N}(0, 1+\psi^2)$. Stacking $\mathbf{u}_i = (u_{i1}, \ldots, u_{iT})'$ and $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iT})'$, we can rewrite the MA(1) process as

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

where $\boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{O}_\psi)$ with $\mathbf{O}_\psi = \text{diag}(1+\psi^2, 1, \ldots, 1)$, 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, $\mathbf{\Omega} = \mathbf{H}_\psi \mathbf{O}_\psi \mathbf{H}_\psi'$ 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) $p(\mathbf{A}, \mathbf{\Sigma} \,|\, \mathbf{Y}, \psi)$ and 2) $p(\psi \,|\, \mathbf{Y}, \mathbf{A}, \mathbf{\Sigma})$. Again, Step 1 can be carried out exactly the same as before. In implementing Step 1, we emphasize that when one computes products of the form $\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}$ or $\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{Y}$, one needs not obtain the inverse $\mathbf{\Omega}^{-1}$, which is a time-consuming step. Instead, since in this case $\mathbf{\Omega}$ is a band matrix, its Cholesky factor $\mathbf{C}_{\mathbf{\Omega}}$ can be obtained in $\mathcal{O}(T)$ operations. Then, to compute $\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}$, one simply return $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$, where $\widetilde{\mathbf{X}} = \mathbf{C}_{\mathbf{\Omega}} \backslash \mathbf{X}$.

For Step 2, $p(\psi \,|\, \mathbf{Y}, \mathbf{A}, \mathbf{\Sigma})$ 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

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

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

Example 4. AR(1) Innovations

Here we consider an example where $\mathbf{\Omega}$ is a full matrix, but $\mathbf{\Omega}^{-1}$ 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 $|\phi| < 1$, $\eta_{it} \sim \mathcal{N}(0,1)$, and the process is initialized with $u_{i1} \sim \mathcal{N}(0, 1/(1-\phi^2))$. Stacking $\mathbf{u}_i = (u_{i1}, \ldots, u_{iT})'$ and $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iT})'$, we can rewrite the AR(1) process as

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

where $\boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{O}_\phi)$ with $\mathbf{O}_\phi = \text{diag}(1/(1-\phi^2), 1, \ldots, 1)$, 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 $\boldsymbol{\Omega}^{-1} = \mathbf{H}_\phi'\mathbf{O}_\phi^{-1}\mathbf{H}_\phi$, where both $\mathbf{O}_\phi$ and $\mathbf{H}_\phi$ are band matrices.

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

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

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

where $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, and the process is initialized by $\mathbf{u}_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}/(1-\phi^2))$. Then, consider the Gaussian proposal $\mathcal{N}(\widehat{\phi}, K_\phi^{-1})$, where $K_\phi = 1/V_\phi + \sum_{t=2}^T \mathbf{u}_{t-1}'\boldsymbol{\Sigma}^{-1}\mathbf{u}_{t-1}$ and $\widehat{\phi} = K_\phi^{-1}(\phi_0/V_\phi + \sum_{t=2}^T \mathbf{u}_{t-1}'\boldsymbol{\Sigma}^{-1}\mathbf{u}_t)$. With this proposal distribution, we can then implement an independence-chain Metropolis-Hastings step to sample ϕ.

The above are only a few simple examples of BVARs which fall within the proposed framework. More elaborate models can be estimated by combining different examples given above. This is demonstrated in the next section.

4 Application

To illustrate the usefulness of the proposed BVARs with a variety of flexible covariance structures, we consider an application using a dataset of 20 macroeconomic variables at quarterly frequency. We first present full sample estimation results and show that the proposed extensions of standard BVARs fit the data substantially better (even after taking into account the added model complexity). Then, in a recursive out-of-sample forecasting exercise we compare the forecast performance of the BVARs at various forecast horizons.

4.1 Competing Models

We consider a variety of BVARs that fall within the proposed framework. All the models have the same conditional mean as specified in (1)—they only differ in the distributional assumptions on the innovations $\mathbf{U}$. Following standard practice we fix the lag length to $p = 4$. The primary goal of this exercise is not to find the best model *per se*. Rather, our objective is to investigate if the more flexible covariance structures improve model-fit and forecast performance. To that end, we consider three types of innovations: non-Gaussian, heteroscedastic and serially dependent innovations.

For non-Gaussian innovations, we choose the multivariate t distribution as it is a popular specification that can better handle outliers than the Gaussian distribution. Specifically, the innovations $\mathbf{u}_t$ are assumed to be independent $\mathcal{N}(\mathbf{0}, \lambda_t \boldsymbol{\Sigma})$ distributed with $\lambda_t \sim \mathcal{IG}(\nu/2, \nu/2)$. For heteroscedastic innovations, we consider the common stochastic volatility model in Carriero et al. (2015a): $\mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \mathrm{e}^{h_t}\boldsymbol{\Sigma})$, where the log volatility follows an AR(1) process: $h_t = \rho h_{t-1} + \varepsilon_t^h$. For serially dependent innovations, we consider a simple MA(1) structure: $\mathbf{u}_t = \boldsymbol{\varepsilon}_t + \psi \boldsymbol{\varepsilon}_{t-1}$, where $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

We also include different combinations of the three error covariance structures. More specifically, we denote a standard BVAR with iid Gaussian innovations as simply BVAR. BVARs with t innovations have a suffix -t. We use suffixes -CSV and -MA to denote models with a common stochastic volatility and MA(1) innovations, respectively. For example, BVAR-t-CSV represents a BVAR that has a common stochastic volatility and t innovations: $\mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \lambda_t \mathrm{e}^{h_t}\boldsymbol{\Sigma})$, where $\lambda_t \sim \mathcal{IG}(\nu/2, \nu/2)$ and h_t follows an AR(1) process.

We summarize all the specifications in Table 1.

Table 1: A list of competing models.

Model	Description
BVAR	standard BVAR with iid Gaussian innovations
BVAR-t	BVAR with t innovations
BVAR-CSV	BVAR with a common stochastic volatility
BVAR-MA	BVAR with MA(1) Gaussian innovations
BVAR-t-CSV	BVAR with a common stochastic volatility and t innovations
BVAR-t-MA	BVAR with MA(1) t innovations
BVAR-CSV-MA	BVAR with a common stochastic volatility and MA(1) innovations

4.2 Data and Priors

In our application we use a dataset of 20 variables at quarterly frequency sourced from the Federal Reserve Bank of St. Louis. It covers the quarters 1959Q1 to 2013Q4 and includes a variety of standard macroeconomic variables such as GDP, inflation, interest

rates, unemployment and money supply. A detailed description of the variables and their transformations are provided in Appendix B.

For easy comparison, we choose exactly the same priors for the common parameters across models. In particular, for all models we adopt the normal-inverse-Wishart prior for $(\mathbf{A}, \mathbf{\Sigma})$: $\mathbf{\Sigma} \sim \mathcal{IW}(\mathbf{S}_0, \nu_0)$, and $(\text{vec}(\mathbf{A}) \,|\, \mathbf{\Sigma}) \sim \mathcal{N}(\text{vec}(\mathbf{A}_0), \mathbf{\Sigma} \otimes \mathbf{V}_\mathbf{A})$. We set $\mathbf{A}_0 = \mathbf{0}$, and the covariance matrix $\mathbf{V}_\mathbf{A}$ is assumed to be diagonal with diagonal elements $v_{\mathbf{A},ii} = \kappa_1/(l^2 \widehat{s}_r)$ for a coefficient associated to lag l of variable r and $v_{\mathbf{A},ii} = \kappa_2$ for an intercept, where $\widehat{s}_r$ is the sample variance of an AR(4) model for the variable r. Further we set $\nu_0 = n + 3$, $\mathbf{S}_0 = \mathbf{I}_n$, $\kappa_1 = 0.2^2$ and $\kappa_2 = 10^2$. Intuitively, the coefficient associated to a lag l variable is shrunk more heavily to zero as the lag length increases, but intercepts are not shrunk to zero. For a more detailed discussion of this natural conjugate prior, see, e.g., Koop and Korobilis (2010) or Karlsson (2013).

For the MA coefficient ψ, 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 distribution thus centers around 0 and has support within the interval $(-1, 1)$. Given the large prior variance, it is also relatively noninformative. 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. Finally, we consider a uniform prior on $(2, 100)$ for the degree of freedom parameter ν, i.e., $\nu \sim \mathcal{U}(2, 100)$.

4.3 Computation Efficiency

In this section we briefly discuss some computation issues and the scalability of the proposed algorithms. In the context of large BVARs, the most time-consuming step is the joint sampling of $\mathbf{A}$ and $\mathbf{\Sigma}$. Using the Kronecker structure $\mathbf{\Sigma} \otimes \mathbf{\Omega}$—even with arbitrary covariance matrices $\mathbf{\Sigma}$ and $\mathbf{\Omega}$—one can substantially speed up the computation. In particular, instead of manipulating covariance matrices of dimension $nk \times nk$ with $k = np + 1$, the dimension of the problem is reduced to $k \times k$. For $n = 100$ and $p = 4$, the latter involves computing Cholesky factors of 401×401 matrices—which can be done fairly quickly—instead of 40100×40100 matrices. In other words, the proposed algorithms can handle over $n = 100$ variables.

In the time dimension, if we leave $\mathbf{\Omega}$ to be an arbitrary $T \times T$ covariance matrix, the proposed algorithms can easily handle $T < 1000$ observations, which includes most quarterly and monthly datasets. For higher-frequency data with $T > 1000$, manipulating arbitrary $T \times T$ covariance matrix can be time-consuming. In those cases one would need to impose more structure on $\mathbf{\Omega}$, such as an ARMA model for the time series structure. Then, $\mathbf{\Omega}$ or its inverse can be decomposed into band matrices. Since manipulating band matrices involves only $\mathcal{O}(T)$ operations, in this case there is essentially no limitation on the dimension T.

In what follows we document the computation times to fit the various models listed in

Table 1. As a benchmark for comparison, we note that Carriero et al. (2015a) estimate a 14-variable BVAR with 4 lags and a common stochastic volatility. They report that their algorithm takes about 40 minutes to obtain 10000 posterior draws.[3] For all our models we have 20 variables and 4 lags. All the algorithms are implemented using MATLAB on a desktop with an Intel Core i7-870 @2.93 GHz processor. The results are reported in Table 2.

Table 2: Time taken to obtain 10000 posterior draws for various BVARs with 20 variables and 4 lags.

Model	Time (minutes)
BVAR	0.21
BVAR-t	0.46
BVAR-CSV	0.46
BVAR-MA	0.83
BVAR-t-CSV	0.68
BVAR-t-MA	1.17
BVAR-CSV-MA	1.16

As is evident from the results, the proposed approach is very fast. For example, obtaining 10000 posterior draws for the common stochastic volatility model takes only half a minute. More complex models take only slightly over a minute. Hence, the proposed algorithms are applicable to very high-dimensional BVARs.

Our proposed sampler is different from that in Carriero et al. (2015a) in three main ways. First, we sample $\mathbf{A}$ and $\mathbf{\Sigma}$ jointly, whereas they sample $\mathbf{A}$ given $\mathbf{\Sigma}$ followed by drawing $\mathbf{\Sigma}$ given $\mathbf{A}$. Second, they sample the log volatilities using the auxiliary mixture sampler of Kim, Shepherd, and Chib (1998). In the case of a common stochastic volatility, there are n measurement equations with Tn additional auxiliary variables. In contrast, we sample the log volatilities directly using an independence-chain Metropolis-Hastings step as in Chan (2015). Lastly, we vectorize all the operations, as this approach is much faster in programming environments optimized for matrix operations, such as MATLAB, GAUSS and R.

To further assess the efficiency of the proposed samplers, we compute the inefficiency factors of the posterior draws, defined as

$$1 + 2\sum_{l=1}^{L} r_l,$$

where r_l is the sample autocorrelation at lag length l and L is chosen to be large enough so that the autocorrelation tapers off. In the ideal case where the posterior draws are

[3]The authors have indicated that if some supplementary calculations—such as storage and computing quantities for the marginal likelihood—are removed, the computation time can be substantially reduced.

independent, the corresponding inefficiency factor is 1. In general, the inefficiency factor measures how many extra draws are needed to give results equivalent to this ideal case. For example, an inefficiency factor of 50 indicates that roughly 5000 posterior draws are required to give the same information as 100 independent draws.

As an example, we report in Figure 1 the inefficiency factors corresponding to the posterior draws of $\mathbf{A}$, $\mathbf{\Sigma}$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_T)'$ and $\mathbf{h} = (h_1, \ldots, h_T)'$ under the BVAR-t-CSV model. Note that there are nk inefficiency factors associated with $\mathbf{A}$, $n(n+1)/2$ values for $\mathbf{\Sigma}$, and T values for each $\boldsymbol{\lambda}$ and $\mathbf{h}$. To present the information visually, boxplots are reported, where the middle line of the box denotes the median, while the lower and upper lines represent respectively the 25- and the 75-percentiles. The whiskers extend to the maximum and minimum.

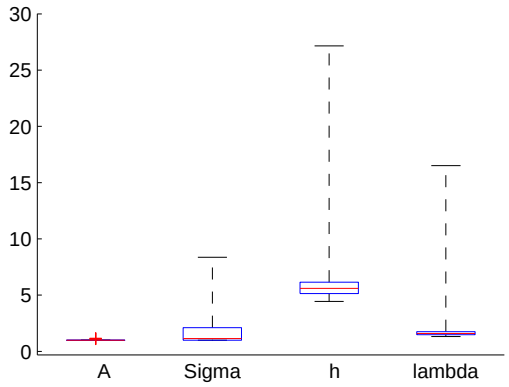

Figure 1: Boxplots of the inefficiency factors corresponding to the posterior draws of $\mathbf{A}$, $\mathbf{\Sigma}$ (Sigma), $\mathbf{h}$ and $\boldsymbol{\lambda}$ (lambda) under the BVAR-t-CSV model.

For example, the boxplot associated with $\mathbf{h}$ indicates that more than 75% of the log volatilities have inefficiency factors less than 10, and the maximum is less than 30. These values are comparable to those obtained from a standard univariate stochastic volatility model estimated by the auxiliary mixture sampler of Kim, Shepherd, and Chib (1998). The inefficiency factors of the model parameters ν, ρ and σ_h^2 are 34, 11 and 46, respectively. All in all, these results suggest that the proposed sampler is quite efficient in terms of producing posterior draws that are not highly autocorrelated.

4.4 Full Sample Estimation Results

In this section we first present the empirical results for the BVARs listed in Table 1, obtained using the full sample from 1959Q1 to 2013Q4. It is then followed by a formal model comparison exercise using the marginal likelihood. In the next section we compare the models in a recursive out-of-sample forecasting exercise.

All the estimates below are based on 30000 posterior draws after a burn-in period of 5000. A key parameter of interest is the MA coefficient ψ—if the posterior density of ψ is concentrated around zero, it would indicate that the MA component might not be necessary. Figure 2 reports the marginal posterior densities of ψ, i.e., $p(\psi \mid \mathbf{Y})$, for the three models that have the MA component. These densities are estimated using the Monte Carlo methods described in Chan (2013).

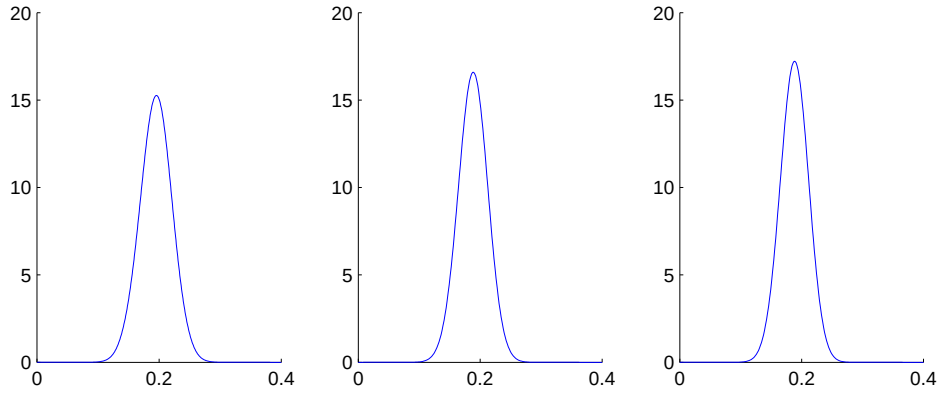

Figure 2: Posterior densities of the MA(1) coefficient ψ under the BVAR-MA (left panel), BVAR-t-MA (middle panel), and BVAR-CSV-MA (right panel) models.

The results are very similar across the three models despite differences in the covariance structure. In particular, all the posterior modes are around 0.2, and there is essentially no mass around zero. This indicates that even after allowing for 4 lags in the dependent variables, there is still considerable positive autocorrelation in the innovations, highlighting the empirical relevance of adding an MA component.

Next, Figure 3 presents the posterior means of the common stochastic volatility—in standard deviations $\exp(h_t/2)$—for the three models: BVAR-CSV, BVAR-CSV-MA and BVAR-t-CSV. The estimates obtained under the BVAR-CSV and BVAR-CSV-MA models are almost identical, and they are similar to those obtained in Carriero et al. (2015a). In particular, we also find substantial time-variation in the common stochastic volatility—volatility is relatively high in the 1970s and early 1980s, and it drops gradually throughout the 1980s and 1990s, until it picks up again during the Global Financial Crisis.

Interestingly, the estimates under the BVAR-t-CSV model are similar to those of the previous two models only in low-volatility periods. During high-volatility periods, the estimates for BVAR-t-CSV are substantially smaller. For instance, during the Global Financial Crisis, the volatility estimates for BVAR-t-CSV are less than 2, compared to about 3.5 for BVAR-CSV. These contrasting results highlight how different models accommodate large shocks. More specifically, both BVAR-CSV and BVAR-CSV-MA have only one main channel to handle large shocks—by increasing the volatility. In contrast, BVAR-t-CSV has the extra channel of making the tails of the distribution heavier—by allowing 'outliers' to appear more often. Our results are broadly similar to the findings

in Cúrdia, Del Negro, and Greenwald (2014)—under their dynamic stochastic general equilibrium (DSGE) model they also find that heavy tails reduce the amount of variation in the stochastic volatility estimates.

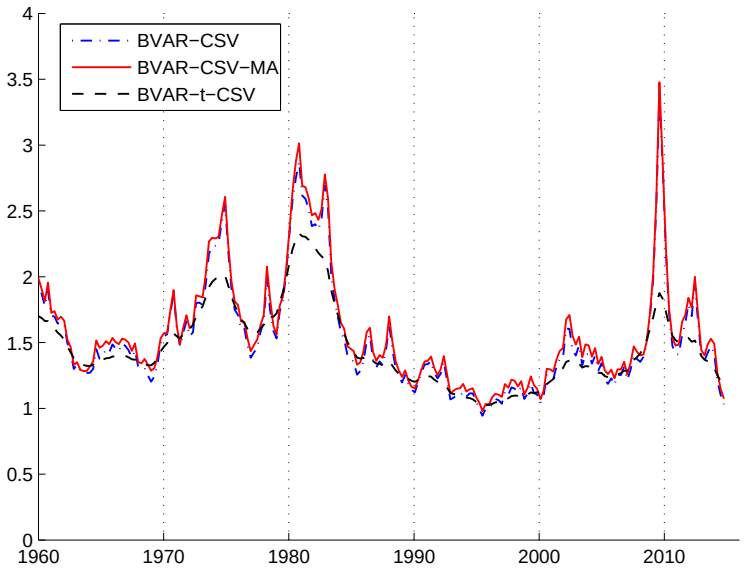

Figure 3: Posterior means of the common stochastic volatility (in standard deviations $\exp(h_t/2)$) under the BVAR-CSV, BVAR-CSV-MA and BVAR-t-CSV models.

These two channels have different implications on the occurrence of large shocks. For instance, if the volatility in the current period is high, large shocks are more likely to occur in the next period as next period's volatility also tends to be high (due to the AR(1) log volatility process). On the other hand, the probability of the occurrence of 'outliers' is time-invariant. The results in Figure 2 suggest that BVAR-t-CSV uses both of these channels to accommodate large shocks—it increases the volatility moderately and makes the tails heavier than those of the Gaussian (see also below for the results on the degree of freedom parameter ν).

In Figure 4 we plot the marginal posterior densities of the degree of freedom parameter ν, i.e., $p(\nu \mid \mathbf{Y})$, for the three models with t innovations. The results for BVAR-t and BVAR-t-MA are similar. Specifically, most of the mass for both densities is concentrated between 5 and 10, indicating that the occurrence of outliers is relatively frequent. This is not surprising in light of the above discussion on the common stochastic volatility estimates—given that the volatility has changed over time, both BVAR-t and BVAR-t-MA that assume constant variance accommodates large shocks in high-volatility periods by making the tails of the distribution heavier.

In contrast, for BVAR-t-CSV that allows for a common stochastic volatility, the posterior density of ν is concentrated between 10 and 30. By allowing for time-varying volatility,

the occurrence of outliers becomes less frequent—even though the tails are still heavier than those of the Gaussian. This mirrors the results reported in Figure 3—models with Gaussian innovations drastically increase the volatility during periods of large shocks, whereas the volatility estimates for BVAR-t-CSV are substantially smaller for the same periods. We present further evidence below that both features—time-varying volatility and heavy tails—are favored by the data.

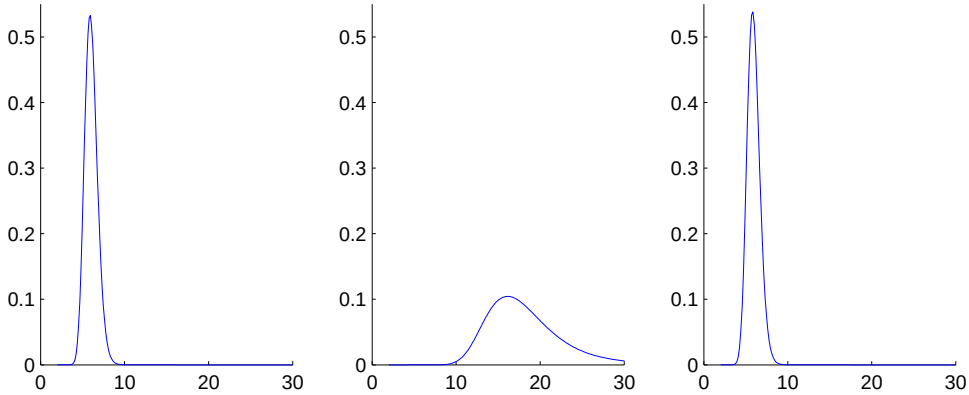

Figure 4: Posterior densities of the degree of freedom parameter ν under the BVAR-t (left panel), BVAR-t-CSV (middle panel), and BVAR-t-MA (right panel) models.

The above estimation results may be viewed as suggestive evidence showing the empirical relevance of the more general covariance structures. In what follows, we consider a formal Bayesian model comparison exercise to compare various BVARs using the marginal likelihood (see, e.g., Koop, 2003). For the standard BVAR with the natural conjugate prior, the marginal likelihood is available analytically (see, e.g., Karlsson, 2013). For other BVARs, the marginal likelihoods are estimated using the Chib's method (Chib, 1995).[4]

For latent variable models, such as stochastic volatility models, the Chib's method can be implemented in two ways: one involves the *conditional* likelihood—the conditional density of the data given the latent variables; the other uses the *integrated* likelihood—the marginal density of the data obtained by integrating out the latent variables. In many applications the first approach is adopted as the conditional likelihood is often easy to evaluate, whereas the integrated likelihood is not. However, some recent papers, such as Frühwirth-Schnatter and Wagner (2008) and Chan and Grant (2015), have pointed out that marginal likelihood estimates based on the conditional likelihood can be extremely inaccurate, whereas those based on the integrated likelihood seem to perform well.[5] As

[4]Even though some of the full conditional densities are non-standard, they can still be quickly evaluated. For instance, the full conditional density of the MA coefficient ψ can be evaluated using the Monte Carlo methods described in Chan (2013).

[5]In the related context of computing the deviance information criterion, Li, Zeng, and Yu (2012) give theoretical arguments why the conditional likelihood should not be used.

such, our marginal likelihood estimates are based on the integrated likelihood. Details of integrated likelihood evaluation for various BVARs are given in Appendix A.

The marginal likelihood estimates reported in Table 3 are based on 10 parallel chains each of which is of length 10000. A few broad conclusions can be drawn. First, compared to a standard BVAR with independent, homoscedastic Gaussian innovations, the three extensions BVAR-t, BVAR-CSV and BVAR-MA all fit the data substantially better. The most important improvement is to allow for a common stochastic volatility, followed by using t innovations and adding an MA component. For example, the log Bayes factor in favor of BVAR-CSV against the standard BVAR is about 215. In other words, if two models are equally probable *a priori*, the posterior probability of BVAR-CSV is 2.4×10^{93} times larger than that of the standard BVAR, indicating overwhelming evidence in favor of allowing for time-varying volatility.

Table 3: Estimated log marginal likelihoods and the associated numerical standard errors for various BVARs.

Model	Log marginal likelihood	NSE
BVAR	-8727	–
BVAR-t	-8545	1.12
BVAR-CSV	-8513	0.50
BVAR-MA	-8704	0.04
BVAR-t-CSV	-8500	0.51
BVAR-t-MA	-8515	0.73
BVAR-CSV-MA	-8484	0.27

Second, one can further improve the model performance of BVAR-CSV by allowing for heavier tails via t innovations. In particular, the log Bayes factor in favor of BVAR-t-CSV against BVAR-CSV is about 13. The same conclusion holds when one compares BVAR-t-MA to BVAR-MA. Hence, there is clear evidence that the data prefer the variants with t innovations. This is consistent with the results in Cúrdia et al. (2014), who find that in the context of DSGE models, allowing for t innovations further improves model-fit even in the presence of stochastic volatility.

Third, consistent with the estimates of ψ reported in Figure 2, there is strong support for the MA component. For instance, the log Bayes factors comparing the two pairs—BVAR-CSV-MA against BVAR-CSV and BVAR-t-MA against BVAR-MA—are both about 30, indicating overwhelming evidence in favor of allowing for the MA dynamics.

Overall, the three additional features—heavy-tailedness, common stochastic volatility and serial dependence—are all empirically important. Models with these features substantially fit the data better. In the next section, we show that these features also help improve both point and density forecasts.

4.5 Forecasting Results

In this section we perform a recursive out-of-sample forecasting exercise to evaluate the performance of the proposed BVARs in terms of both point and density forecasts. We use all the 20 variables to forecast four target variables, namely, real GDP growth, unemployment, Fed funds rate and CPI inflation. The evaluation period is from 1985Q1 to 2013Q4.

We use each of BVARs listed in Table 1 to produce both point and density m-step-ahead iterated forecasts with $m = 1, 2, 4$ and 8. Specifically, given the data up to time t, denoted as $\mathbf{Y}_{1:t}$, we implement the MCMC samplers in Section 3 to obtain posterior draws given $\mathbf{Y}_{1:t}$. Then, we 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. Next, we move one period forwad and repeat the whole exercise with data $\mathbf{Y}_{1:t+1}$, and so forth. These forecasts are then evaluated for $t = t_0, \ldots, T - m$.

For general BVARs, neither the predictive mean nor the predictive density of $y_{i,t+m}$ can be computed analytically. Instead, they are obtained using predictive simulation; see, e.g., Chan (2013) for more details. As for forecast evaluation metrics, let $y^{\text{o}}_{i,t+m}$ denote the observed value of the variable $y_{i,t+m}$ that is known at time $t+m$. 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 observed value $y^{\text{o}}_{i,t+m}$. Clearly, 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:

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

For this metric, a larger value indicates better forecast performance. For easy comparison, we report below the ratios of RMSFEs of a given model to those of the standard BVAR. 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 standard BVAR. In this case, positive values indicate better forecast performance than the benchmark.

Tables 4 and 5 report the point and density forecast results for the two nominal variables: Fed funds rate and CPI inflation. It is evident that BVARs with more flexible covariance

structures tend to outperform the standard BVAR at various forecast horizons. For example, the BVAR-CSV model of Carriero et al. (2015a) with a common stochastic volatility forecasts substantially better than the benchmark for both GDP growth and inflation.

Table 4: Forecast performance relative to the standard BVAR; Fed funds rate.

	relative RMSFE			
	$m = 1$	$m = 2$	$m = 4$	$m = 8$
BVAR-t	0.93	0.88	0.86	0.93
BVAR-CSV	0.99	0.88	0.83	0.89
BVAR-MA	0.95	0.96	0.97	0.98
BVAR-t-CSV	0.98	0.87	0.83	0.90
BVAR-t-MA	0.86	0.85	0.84	0.92
BVAR-CSV-MA	0.90	0.85	0.81	0.88
	average log predictive likelihoods			
	$m = 1$	$m = 2$	$m = 4$	$m = 8$
BVAR-t	0.12	0.12	0.08	0.00
BVAR-CSV	0.11	0.20	0.17	0.03
BVAR-MA	0.04	0.02	0.02	0.01
BVAR-t-CSV	0.10	0.18	0.15	0.02
BVAR-t-MA	0.18	0.13	0.09	0.00
BVAR-CSV-MA	0.15	0.19	0.17	0.04

Table 5: Forecast performance for relative to the standard BVAR; CPI inflation.

	relative RMSFE			
	$m = 1$	$m = 2$	$m = 4$	$m = 8$
BVAR-t	0.95	0.90	0.88	0.87
BVAR-CSV	0.94	0.89	0.87	0.88
BVAR-MA	1.00	0.99	0.99	0.99
BVAR-t-CSV	0.94	0.89	0.87	0.87
BVAR-t-MA	0.95	0.90	0.87	0.87
BVAR-CSV-MA	0.94	0.88	0.86	0.87
	average log predictive likelihoods			
	$m = 1$	$m = 2$	$m = 4$	$m = 8$
BVAR-t	0.28	0.24	0.22	0.12
BVAR-CSV	0.27	0.25	0.17	0.13
BVAR-MA	0.00	0.01	0.05	0.00
BVAR-t-CSV	0.28	0.25	0.19	0.13
BVAR-t-MA	0.28	0.26	0.23	0.11
BVAR-CSV-MA	0.26	0.25	0.18	0.12

This is consistent with the results in numerous studies, such as Clark (2011), D'Agostino,

Gambetti, and Giannone (2013) and Clark and Ravazzolo (2014), which find that small BVARs with stochastic volatility forecast substantially better than their counterparts with only constant variance. Interestingly, our results also show that BVAR-t with t innovations provides similar forecast performance compared to BVAR-CSV. Moreover, allowing for either t innovations or stochastic volatility seems to be sufficient, as having both components does not seem to further improve forecast performance. These results are broadly consistent with the findings in Clark and Ravazzolo (2014) and Chiu et al. (2015).

In contrast, forecast performance of either BVAR-t or BVAR-CSV can be further improved by adding an MA component. For instance, the BVAR-t-MA model tends to forecast Fed funds rate best for short horizons, whereas BVAR-CSV-MA tends to perform better for longer horizons.

Next, we present in Tables 6 and 7 the point and density forecast results for the two real variables: real GDP growth and unemployment rate. Compared to previous results, improvements over the standard BVAR are harder to find. For example, for forecasting the unemployment rate, no models can consistently outperform the benchmark in terms of point and density forecasts. However, for GDP growth, most models perform as well as the benchmark for short horizons, and they tend to be better at longer horizons. For instance, at $m = 8$ horizon, BVAR-CSV reduces the RMSFE of BVAR by about 5%. In addition, the RMSFEs of both BVAR-t-MA and BVAR-CSV-MA can be further reduced to about 93% of that of BVAR.

Overall, these results show that the more flexible BVARs can deliver substantial improvements in both point and density forecasts for nominal variables, and to a less extent for real variables.

Table 6: Forecast performance relative to the standard BVAR; real GDP growth.

relative RMSFE				
	$m = 1$	$m = 2$	$m = 4$	$m = 8$
BVAR-t	0.98	1.02	1.00	0.95
BVAR-CSV	0.99	1.01	1.00	0.95
BVAR-MA	1.00	1.00	0.99	0.98
BVAR-t-CSV	0.98	1.01	1.00	0.94
BVAR-t-MA	0.98	1.01	0.99	0.93
BVAR-CSV-MA	0.98	1.01	0.99	0.93
average log predictive likelihoods				
	$m = 1$	$m = 2$	$m = 4$	$m = 8$
BVAR-t	0.05	0.05	0.07	0.10
BVAR-CSV	0.04	0.04	0.04	0.09
BVAR-MA	-0.01	0.00	0.00	0.01
BVAR-t-CSV	0.05	0.05	0.06	0.10
BVAR-t-MA	0.05	0.05	0.08	0.12
BVAR-CSV-MA	0.03	0.04	0.04	0.10

Table 7: Forecast performance relative to the standard BVAR; unemployment rate.

relative RMSFE				
	$m = 1$	$m = 2$	$m = 4$	$m = 8$
BVAR-t	1.03	1.04	1.07	1.06
BVAR-CSV	1.02	1.02	1.05	1.03
BVAR-MA	0.99	0.99	1.00	1.00
BVAR-t-CSV	1.03	1.03	1.05	1.03
BVAR-t-MA	1.01	1.04	1.08	1.05
BVAR-CSV-MA	1.01	1.04	1.08	1.04
average log predictive likelihoods				
	$m = 1$	$m = 2$	$m = 4$	$m = 8$
BVAR-t	0.01	0.03	0.02	-0.05
BVAR-CSV	-0.07	0.03	0.04	0.00
BVAR-MA	0.01	-0.01	-0.01	-0.06
BVAR-t-CSV	-0.07	0.02	0.04	-0.08
BVAR-t-MA	0.02	0.02	0.04	-0.04
BVAR-CSV-MA	-0.08	-0.02	0.02	-0.02

5 Concluding Remarks and Future Research

With the aim of expanding the toolkit for empirical macroeconomists, we have introduced a new class of large BVARs that allows for non-Gaussian, heteroscedastic and serially dependent innovations. In general it is challenging to estimate large BVARs due to the high-dimensional VAR coefficients. The main advantage of the proposed framework is that the implied likelihood has a certain Kronecker structure that can be exploited to vastly speed up computation.

We have demonstrated the empirical relevance of this new class of models with an application using 20 macroeconomic variables. The model comparison results show that the data overwhelmingly favor these more flexible BVARs over the standard variant with independent, homoscedastic Gaussian innovations. In a recursive forecasting exercise, we find that the new models also deliver improvements in both point and density forecasts.

For future research, it would be worthwhile to further extend these models to allow for richer covariance structures. More specifically, in our framework all the variables need to have the same time series model to preserve the Kronecker structure in the likelihood. One possible way forward is to incorporate a factor model to induce a richer serial dependence structure, while maintaining the Kronecker structure so that the same type of computation shortcut can be applied.

Appendix A: Estimation Details

A1: Sampling $(\mathbf{A}, \boldsymbol{\Sigma})$

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

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

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

A2: Integrated Likelihood Evaluation

In this section we discuss how one can evaluate the integrated likelihoods of the proposed BVARs. To that end, we first introduce some generic notations. Let $\boldsymbol{\theta}$ denote the model parameters and let $\mathbf{z}$ be the latent variables or states.[6] Then, the *integrated* likelihood or *observed-data* likelihood is defined as

$$p(\mathbf{Y} \,|\, \boldsymbol{\theta}) = \int p(\mathbf{Y} \,|\, \boldsymbol{\theta}, \mathbf{z}) p(\mathbf{z} \,|\, \boldsymbol{\theta}) \mathrm{d}\mathbf{z},$$

where $f(\mathbf{Y} \,|\, \boldsymbol{\theta}, \mathbf{z})$ is the *conditional* likelihood. In most latent variable models, the conditional likelihood is easy to evaluate by construction, whereas the evaluation of the integrated likelihood is typically difficult due to the high-dimensional integration.

First, the integrated likelihood of BVAR-MA is available analytically as there are no latent variables. In this case, $\boldsymbol{\theta}$ consists of $\mathbf{A}$, $\boldsymbol{\Sigma}$ and ψ. It follows from (4) that the likelihood is given by

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

[6] One crucial difference between model parameters and latent variables is that the number of parameters does not increase with T, whereas the number of latent variables does.

where $\widetilde{\mathbf{U}} = \mathbf{H}_\psi^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{A})$, $\mathbf{O}_\psi = \text{diag}(1+\psi^2, 1, \ldots, 1)$, and

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

Next, we derive the integrated likelihood of BVAR-t-MA. For this model $\boldsymbol{\Omega} = \mathbf{H}_\psi \mathbf{O}_{\boldsymbol{\lambda},\psi} \mathbf{H}_\psi'$, where $\mathbf{O}_{\boldsymbol{\lambda},\psi} = \text{diag}((1+\psi^2)\lambda_1, \lambda_2, \ldots, \lambda_T)$. Using our generic notations, in this case $\boldsymbol{\theta}$ consists of $\mathbf{A}$, $\boldsymbol{\Sigma}$ and ν, while the latent variables are $\mathbf{z} = \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_T)'$. It follows from (4) that the conditional likelihood is given by

$$p(\mathbf{Y} \,|\, \boldsymbol{\lambda}, \mathbf{A}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{Tn}{2}} |\boldsymbol{\Sigma}|^{-\frac{T}{2}} (1+\psi^2)^{-\frac{n}{2}} \prod_{t=1}^{T} \left(\lambda_t^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2\lambda_t} \widetilde{\mathbf{u}}_t' \boldsymbol{\Sigma}^{-1} \widetilde{\mathbf{u}}_t} \right),$$

where $\widetilde{\mathbf{u}}_t'$ is the t-th row of $\widetilde{\mathbf{U}}$ for $t = 2, \ldots, T$, and $\widetilde{\mathbf{u}}_1'$ is the first row of $\widetilde{\mathbf{U}}$ divided by $\sqrt{1+\psi^2}$. Given the prior $(\lambda_t \,|\, \nu) \sim \mathcal{IG}(\nu/2, \nu/2)$, the integrated likelihood can be computed as follows:

$$\begin{aligned} p(\mathbf{Y} \,|\, \mathbf{A}, \boldsymbol{\Sigma}, \nu) &= \int p(\mathbf{Y} \,|\, \boldsymbol{\lambda}, \mathbf{A}, \boldsymbol{\Sigma}, \nu) p(\boldsymbol{\lambda} \,|\, \nu) \mathrm{d}\boldsymbol{\lambda} \\ &= (2\pi)^{-\frac{Tn}{2}} |\boldsymbol{\Sigma}|^{-\frac{T}{2}} (1+\psi^2)^{-\frac{n}{2}} \frac{\left(\frac{\nu}{2}\right)^{\frac{T\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)^T} \int \prod_{t=1}^{T} \left(\lambda_t^{-(\frac{n+\nu}{2}+1)} \mathrm{e}^{-\frac{1}{2\lambda_t}(\widetilde{\mathbf{u}}_t' \boldsymbol{\Sigma}^{-1} \widetilde{\mathbf{u}}_t + \nu)} \right) \mathrm{d}\boldsymbol{\lambda} \\ &= (\nu\pi)^{-\frac{Tn}{2}} |\boldsymbol{\Sigma}|^{-\frac{T}{2}} (1+\psi^2)^{-\frac{n}{2}} \left(\frac{\Gamma\left(\frac{n+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \right)^T \prod_{t=1}^{T} \left(1 + \frac{1}{\nu} \widetilde{\mathbf{u}}_t' \boldsymbol{\Sigma}^{-1} \widetilde{\mathbf{u}}_t \right)^{-\frac{n+\nu}{2}}, \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function, and we have used the fact that

$$\int \lambda_t^{-(\frac{n+\nu}{2}+1)} \mathrm{e}^{-\frac{1}{2\lambda_t}(\widetilde{\mathbf{u}}_t' \boldsymbol{\Sigma}^{-1} \widetilde{\mathbf{u}}_t + \nu)} \mathrm{d}\lambda_t = \Gamma\left(\frac{n+\nu}{2}\right) \left(\frac{\nu + \widetilde{\mathbf{u}}_t' \boldsymbol{\Sigma}^{-1} \widetilde{\mathbf{u}}_t}{2} \right)^{-\frac{n+\nu}{2}}.$$

Note that the integrated likelihood of BVAR-t can be obtained simply by setting $\psi = 0$.

Now, we move on to the integrated likelihood of BVAR-CSV. Using our generic notations, in this model $\boldsymbol{\theta}$ consists of $\mathbf{A}$, $\boldsymbol{\Sigma}$, ρ and σ_h^2, while the latent variables are $\mathbf{z} = \mathbf{h} = (h_1, \ldots, h_T)'$. That is, to obtain the integrated likelihood we need to integrate out $\mathbf{h}$. Unfortunately, this cannot be done analytically. Instead, we estimate the integrated likelihood using the importance sampling approach in Chan and Grant (2014). To do so, we need three ingredients: the prior $p(\mathbf{h} \,|\, \rho, \sigma_h^2)$ implied by the state equation (3), the conditional likelihood $p(\mathbf{Y} \,|\, \mathbf{A}, \boldsymbol{\Sigma}, \mathbf{h})$ and a good importance sampling density.

First, $p(\mathbf{h} \,|\, \rho, \sigma_h^2)$ is Gaussian, and an explicit expression can be found in, e.g., Chan and Grant (2014). Next, it follows from (4) that the conditional likelihood is given by

$$p(\mathbf{Y} \,|\, \mathbf{A}, \boldsymbol{\Sigma}, \mathbf{h}) = (2\pi)^{-\frac{Tn}{2}} |\boldsymbol{\Sigma}|^{-\frac{T}{2}} \mathrm{e}^{-\frac{n}{2}\sum_{t=1}^{T} h_t} \mathrm{e}^{-\frac{1}{2}\mathrm{e}^{-h_t} \mathbf{u}_t' \boldsymbol{\Sigma}^{-1} \mathbf{u}_t}.$$

For the choice of the importance sampling density, note that the ideal zero-variance importance sampling density for estimating the integrated likelihood is the conditional density $p(\mathbf{h} \,|\, \mathbf{Y}, \mathbf{A}, \mathbf{\Sigma}, \rho, \sigma_h^2)$. However, this density cannot be used directly as the normalizing constant is unknown. To proceed, we approximate this conditional density using a Gaussian density. Let $\widehat{\mathbf{h}}$ and $\mathbf{K}_{\mathbf{h}}$ denote respectively the mode and the negative Hessian evaluated at the mode of $\log p(\mathbf{h} \,|\, \mathbf{Y}, \mathbf{A}, \mathbf{\Sigma}, \rho, \sigma_h^2)$. Then, we use the $\mathcal{N}(\widehat{\mathbf{h}}, \mathbf{K}_{\mathbf{h}}^{-1})$ density as the importance sampling density. The parameters $\widehat{\mathbf{h}}$ and $\mathbf{K}_{\mathbf{h}}$ can be obtained as discussed in Section 3.2. With all these ingredients, we can then use the importance sampling approach in Chan and Grant (2014) to evaluate the integrated likelihood of BVAR-CSV.

Next, to evaluate the integrated likelihood of BVAR-CSV-MA, only small modifications are needed. In this case, the prior $p(\mathbf{h} \,|\, \rho, \sigma_h^2)$ is exactly the same as before. Now, $\mathbf{\Omega} = \mathbf{H}_\psi \mathbf{O}_{\mathbf{h},\psi} \mathbf{H}_\psi'$, where $\mathbf{O}_{\mathbf{h},\psi} = \text{diag}((1+\psi^2)\text{e}^{h_1}, \text{e}^{h_2}, \dots, \text{e}^{h_T})$. Then, the conditional likelihood is given by

$$p(\mathbf{Y} \,|\, \mathbf{A}, \mathbf{\Sigma}, \mathbf{h}) = (2\pi)^{-\frac{Tn}{2}} |\mathbf{\Sigma}|^{-\frac{T}{2}} (1+\psi^2)^{-\frac{n}{2}} \text{e}^{-\frac{n}{2}\sum_{t=1}^T h_t} \text{e}^{-\frac{1}{2}\text{e}^{-h_t} \widetilde{\mathbf{u}}_t' \mathbf{\Sigma}^{-1} \widetilde{\mathbf{u}}_t},$$

where $\widetilde{\mathbf{u}}_t$ is defined as above. Moreover, a Gaussian approximation of the full conditional density $p(\mathbf{h} \,|\, \mathbf{Y}, \mathbf{A}, \mathbf{\Sigma}, \psi, \rho, \sigma_h^2)$ can be obtained similarly, which is then used as the importance sampling density.

Finally, we consider the integrated likelihood evaluation for BVAR-t-CSV. In this case, we need to integrated out both $\boldsymbol{\lambda}$ and $\mathbf{h}$. It turns out we can integrate out $\boldsymbol{\lambda}$ analytically, and $\mathbf{h}$ is then integrated out by importance sampling as discussed before. First, the prior $p(\mathbf{h} \,|\, \rho, \sigma_h^2)$ is exactly the same as before. Using a similar derivation as in the BVAR-t-MA case, the partial conditional likelihood (marginal of $\boldsymbol{\lambda}$) is given by

$$p(\mathbf{Y} \,|\, \mathbf{h}, \mathbf{A}, \mathbf{\Sigma}, \nu) = (\nu\pi)^{-\frac{Tn}{2}} |\mathbf{\Sigma}|^{-\frac{T}{2}} \text{e}^{-\frac{n}{2}\sum_{t=1}^T h_t} \left(\frac{\Gamma\left(\frac{n+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \right)^T \prod_{t=1}^T \left(1 + \frac{1}{\nu} \text{e}^{-h_t} \mathbf{u}_t' \mathbf{\Sigma}^{-1} \mathbf{u}_t \right)^{-\frac{n+\nu}{2}}.$$

The last ingredient we need is an importance sampling density. In this case, the ideal zero-variance importance sampling density is the conditional density $p(\mathbf{h} \,|\, \mathbf{Y}, \mathbf{A}, \mathbf{\Sigma}, \nu, \rho, \sigma_h^2)$ marginal of $\boldsymbol{\lambda}$. As before, we approximate this with a Gaussian density with mean vector $\widehat{\mathbf{h}}$ and precision matrix $\mathbf{K}_{\mathbf{h}}$, where $\widehat{\mathbf{h}}$ and $\mathbf{K}_{\mathbf{h}}$ are respectively the mode and the negative Hessian evaluated at the mode of $\log p(\mathbf{h} \,|\, \mathbf{Y}, \mathbf{A}, \mathbf{\Sigma}, \nu, \rho, \sigma_h^2)$.

Appendix B: Data

All the variables are sourced from the Federal Reserve Bank of St. Louis and cover the quarters 1959Q1 to 2013Q4. Table 8 lists the 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 8: Description of variables used in the BVARs.

Variable	Transformation
Real gross domestic product	$400\Delta \log$
Consumer price index	$400\Delta \log$
Effective Federal funds rate	no transformation
M2 money stock	$400\Delta \log$
Personal income	$400\Delta \log$
Real personal consumption expenditure	$400\Delta \log$
Industrial production index	$400\Delta \log$
Civilian unemployment rate	no transformation
Housing starts	log
Producer price index	$400\Delta \log$
Personal consumption expenditures: chain-type price index	$400\Delta \log$
Average hourly earnings: manufacturing	$400\Delta \log$
M1 money stock	$400\Delta \log$
10-Year Treasury constant maturity rate	no transformation
Real gross private domestic investment	$400\Delta \log$
All employees: total nonfarm	$400\Delta \log$
ISM manufacturing: PMI composite index	no transformation
ISM manufacturing: new orders index	no transformation
Business sector: real output per hour of all Persons	$400\Delta \log$
Real stock prices (S& P 500 index divided by PCE index)	$100\Delta \log$

References

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

L. Bauwens, M. Lubrano, and J. Richard. *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press, New York, 1999.

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

A. Carriero, T. E. Clark, and M. Marcellino. Common drifting volatility in large Bayesian VARs. *Journal of Business and Economic Statistics*, 2015a. Forthcoming.

A. Carriero, T. E. Clark, and M. Marcellino. Bayesian VARs: Specification choices and forecast accuracy. *Journal of Applied Econometrics*, 30(1):46–73, 2015b.

J. C. C. Chan. Moving average stochastic volatility models with application to inflation forecast. *Journal of Econometrics*, 176(2):162–172, 2013.

J. C. C. Chan. The stochastic volatility in mean model with time-varying parameters: An application to inflation modeling. *Journal of Business and Economic Statistics*, 2015. Forthcoming.

J. C. C. Chan and A. L. Grant. Issues in comparing stochastic volatility models using the deviance information criterion. *CAMA Working Paper*, 2014.

J. C. C. Chan and A. L. Grant. Pitfalls of estimating the marginal likelihood using the modified harmonic mean. *Economics Letters*, 131:29–33, 2015.

J. C. C. Chan and C. Y. L. Hsiao. Estimation of stochastic volatility models with heavy tails and serial dependence. In I. Jeliazkov and X.-S. Yang, editors, *Bayesian Inference in the Social Sciences*. John Wiley & Sons, Hoboken, 2014.

J. C. C. Chan, E. Eisenstat, and G. Koop. Large Bayesian VARMAs. Working Paper, 2015.

S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90:1313–1321, 1995.

C. J. Chiu, H. Mumtaz, and G. Pinter. Forecasting with VAR models: Fat tails and stochastic volatility. Technical report, CReMFi, School of Economics and Finance, QMUL, 2015.

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

T. E. Clark and F. Ravazzolo. Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *Journal of Applied Econometrics*, 2014. Forthcoming.

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

V. Cúrdia, M. Del Negro, and D. L. Greenwald. Rare shocks, great recessions. *Journal of Applied Econometrics*, 29(7):1031–1052, 2014.

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

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

T. Eltoft, T. Kim, and T. Lee. On the multivariate Laplace distribution. *Signal Processing Letters, IEEE*, 13(5):300–303, 2006a.

T. Eltoft, T. Kim, and T. Lee. Multivariate scale mixture of Gaussians modeling. In J. Rosca, D. Erdogmus, J. Principe, and S. Haykin, editors, *Independent Component Analysis and Blind Signal Separation*, volume 3889 of *Lecture Notes in Computer Science*, pages 799–806. Springer Berlin Heidelberg, 2006b.

M. Forni, M. Hallin, M. Lippi, and L. Reichlin. Do financial variables help forecasting inflation and real activity in the euro area? *Journal of Monetary Economics*, 50(6): 1243 – 1255, 2003.

S. Frühwirth-Schnatter and H. Wagner. Marginal likelihoods for non-Gaussian models using auxiliary mixture sampling. *Computational Statistics and Data Analysis*, 52(10): 4608–4624, 2008.

J. Geweke. Bayesian treatment of the independent Student-t linear model. *Journal of Applied Econometrics*, 8(S1):S19–S40, 1993.

J. Geweke and G. Amisano. Hierarchical Markov normal mixture models with applications to financial asset returns. *Journal of Applied Econometrics*, 26:1–29, 2011.

G. H. Golub and C. F. van Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, 1983.

K. Kadiyala and S. Karlsson. Numerical methods for estimation and inference in Bayesian VAR-models. *Journal of Applied Econometrics*, 12(2):99–132, 1997.

S. Karlsson. Forecasting with Bayesian vector autoregressions. In G. Elliott and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 2 of *Handbook of Economic Forecasting*, pages 791–897. Elsevier, 2013.

S. Kim, N. Shepherd, and S. Chib. Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies*, 65(3):361–393, 1998.

G. Koop. *Bayesian Econometrics*. Wiley & Sons, New York, 2003.

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

G. Koop and D. Korobilis. Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends in Econometrics*, 3(4):267–358, 2010.

G. Koop and D. Korobilis. Large time-varying parameter VARs. *Journal of Econometrics*, 177(2):185–198, 2013.

Y. Li, T. Zeng, and J. Yu. Robust deviance information criterion for latent variable models. *SMU Economics and Statistics Working Paper Series*, 2012.

R. Litterman. Forecasting with Bayesian vector autoregressions—five years of experience. *Journal of Business and Economic Statistics*, 4:25–38, 1986.

G. E. Primiceri. Time varying structural vector autoregressions and monetary policy. *Review of Economic Studies*, 72(3):821–852, 2005.

J. H. Stock and M. W. Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics*, 20:147–162, 2002.