

# Forecasting Using Variational Bayesian Inference in Large Vector Autoregressions with Hierarchical Shrinkage

## Abstract

Many recent papers in macroeconomics have used large vector autoregressions (VARs) involving 100 or more dependent variables. With so many parameters to estimate, Bayesian prior shrinkage is vital in achieving reasonable results. Computational concerns currently limit the range of priors used and render difficult the addition of empirically important features such as stochastic volatility to the large VAR. In this paper, we develop variational Bayes methods for large VARs which overcome the computational hurdle and allow for Bayesian inference in large VARs with a range of hierarchical shrinkage priors and with time-varying volatilities. We demonstrate the computational feasibility and good forecast performance of our methods in an empirical application involving a large quarterly US macroeconomic data set.

*Keywords:* Variational Inference, Vector Autoregression, Stochastic Volatility, Hierarchical Prior, Forecasting

*JEL Classifications:* C11, C32, C53

The Technical, Empirical and Data Appendices referenced in this paper are available at <https://sites.google.com/site/garykoop/>.

# 1 Introduction

This paper develops variational Bayesian (VB) methods for large Bayesian VARs with hierarchical shrinkage priors and multivariate stochastic volatility and shows them to have a much lower computational burden than the Markov chain Monte Carlo (MCMC) methods which are currently predominant in the Bayesian VAR literature. We demonstrate that VB methods are accurate and are scaleable. They can be used in practice even in very large VARs.

To explain why VB methods can be a useful tool for the researcher working with Bayesian VARs, note that in recent years we have seen the emergence of a literature which uses VARs with a large number of dependent variables. The seminal paper was Banbura, Giannone and Reichlin (2010). Subsequently large VARs have been used in many empirical applications in macroeconomics and finance; see, among many others, Bloor and Matheson (2010), Carriero, Kapetanios and Marcellino (2010), Carriero, Kapetanios and Marcellino (2012), Koop (2013), Gefang (2014), Giannone, Lenza, Momferatou and Onorante (2014), Banbura, Giannone and Lenza (2015), Koop and Korobilis (2016, 2019), Jarocinski and Mackowiak (2017), Carriero, Clark and Marcellino (2016, 2018, 2019) and Chan (2018). The computational methods used in these papers fall into two general categories: i) those which use MCMC methods and ii) those which avoid the use of MCMC methods by using natural conjugate priors (for which analytical results are available). It is noteworthy that papers in category i) tend to use VARs which are much smaller than papers in category ii). For instance, Banbura, Giannone and Reichlin (2010) use a natural conjugate prior and work with 131 variables whereas Chan (2018) uses MCMC methods and works with 20 variables. The reason for this is largely computational: the time taken to carry out Bayesian inference or prediction in models which require the use of MCMC methods is much greater than that taken when using models for which analytical results are available.

In the large VAR literature there is a growing realization that it is computationally difficult (if not impossible) to use MCMC methods with 100 or more variables, especially in the context of a recursive forecasting exercise where MCMC methods are used repeatedly on an expanding or rolling window of data. However, macroeconomic researchers currently wish to work with over 100 variables and it is easy to imagine that, in the near future, they will want to work with many more.<sup>1</sup>

If MCMC methods cannot be used with large VARs, then there is a risk that the large Bayesian VAR literature will not be able to expand to the increasingly large datasets that economists wish to work with. This is because the natural conjugate approaches which provide analytical results have their limitations. In particular, empirically-necessary extensions of the VAR such as adding stochastic volatility are not possible with the natural conjugate prior. Nor is it possible, using the natural conjugate prior, to accommodate the

---

<sup>1</sup>In the US, the popular FRED-MD and FRED-QD data sets, produced by the Federal Reserve Bank of St. Louis, contain well over 100 monthly variables and well over 200 quarterly variables, respectively.

hierarchical priors which are increasingly used in the machine learning literature to ensure shrinkage and sparsity. The VAR literature has typically used MCMC methods to handle such extensions (see, e.g., George, Sun and Ni, 2008, Koop, 2013, Korobilis, 2013 and Kastner and Huber, 2021).

In this paper we show how an alternative approach, VB, can be used for Bayesian inference in cases where MCMC methods are computationally-infeasible. VB methods will be discussed in the next section, but their key properties are that they provide an approximation to the Bayesian posterior and predictive distributions in the VAR and are computationally much faster than MCMC methods, which makes VB a useful substitute for MCMC in Bayesian VAR forecasting exercises involving huge VARs.

We develop VB methods for a range of hierarchical shrinkage priors that are popular in the machine learning literature and have been used in regression or with small or medium sized VARs. These include the horseshoe, priors which fall in the least absolute shrinkage and selection operator (LASSO) class, the stochastic search variable selection (SSVS) prior and adaptive shrinkage Jeffreys' and t-priors. Our methods allow for automatic shrinkage on the VAR error covariances as well as the VAR coefficients themselves. We also develop VB methods which can be used to add stochastic volatility to any of the VARs with hierarchical shrinkage.

In an empirical exercise involving a large data set of quarterly US macroeconomic variables we show that VB methods are accurate and forecast well. In particular, we demonstrate the accuracy of VB methods using a data set of 10 variables. We show that, for some of the shrinkage priors, MCMC methods and VB methods produce mean squared forecast error (MSFE) results that are virtually identical. For the remainder of the priors, results are very close to one another. We do, however, find VB tends to underestimate predictive variances slightly. We also demonstrate the forecasting performance of VB methods using a large data set of 100 variables. In this dimension, MCMC methods are not feasible, but we show good forecasting performance can be obtained using VB methods.

## 2 Variational Bayesian Inference

VB methods have been growing in popularity as a practical way of doing Bayesian inference in models for which MCMC would be too computationally demanding. The basic theory justifying VB is provided in many papers including Ormerod and Wand (2010), You, Ormerod and Muller (2014) and Blei, Kucukelbir and McAuliffe (2017). Here we explain the basic theory and necessary ingredients to use VB methods in practice in a general context where  $p(\theta|y)$  is the posterior of interest involving data  $y$  and parameters  $\theta$ . VB methods approximate this posterior with another simpler density  $q(\theta)$  that is as close as possible to it in a Kullback-Leibler (KL) sense. It is important to be clear from the outset that VB methods are approximate in the sense that  $q(\theta)$  is

not the same as  $p(\theta|y)$ . This is the type of approximation which we refer to in the following material. MCMC methods are only exact if an infinite number of draws are taken and, thus, in practice are also approximate.

VB requires the choice of a class of approximating densities,  $q(\theta)$  (e.g. the normal density is a popular choice). The optimal VB density is the one within this class that has parameters chosen so as to make the VB density as close as possible to the posterior. For instance, the mean and variance-covariance matrix of the normal are estimated so as to minimize the KL distance between the normal approximation and the posterior. Minimizing KL can be shown to be equivalent to maximizing the evidence lower bound (ELBO):

$$ELBO = E(\log p(\theta, y)) - E(\log q(\theta)),$$

where the expectations are taken with respect to  $q(\theta)$ . Thus, VB involves optimizing a function (the ELBO), which is typically much faster than doing MCMC.

Computation is particularly easy if the class of approximating densities is taken from the so-called mean field variational family:

$$q(\theta) = \prod_{m=1}^M q_m(\theta_m),$$

where  $\theta_m$  for  $m = 1, \dots, M$  are the blocks of parameters which make up  $\theta$ .

If we assume the prior for the blocks of parameters are independent, then the ELBO can be written as:

$$ELBO = E(\log p(y|\theta)) + \sum_{m=1}^M E(\log p(\theta_m)) - \sum_{m=1}^M E(\log q_m(\theta_m)). \quad (1)$$

Note that the  $-E(\log q_m(\theta_m))$  terms are the entropies of each approximating density and, thus, working with densities with known entropies is convenient when doing VB inference.

Within this family, it can be proved (see, for instance, Section 2.2 of Ormerod and Wand, 2010) that the optimal choice for  $q_m(\theta_m)$  involves the full conditional posterior densities used in a Gibbs sampler and is given by:

$$q_m(\theta_m) = \exp[E(\log p(\theta_m|y, \theta_{-m}))], \quad (2)$$

where  $\theta_{-m}$  denotes all parameters except for those in  $\theta_m$  and the expectation is taken over  $q(\theta_{-m})$ .

Thus, VB is particularly easy for any model which admits Gibbs sampling. The full conditional posteriors used in Gibbs sampling appear in (2) and, thus, in the ELBO in (1). The ELBO also involves the likelihood and prior and is easy to evaluate if each approximating density has a known entropy. The expectations in (2) and (1) are calculated using optimization (not posterior simulation) in a similar fashion to the expectations

maximization (EM) algorithm of Dempster, Laird and Rubin (1977).

The theoretical properties of VB methods depend to some extent on the specific class of models under consideration. However, You, Omerod and Muller (2014) derive the theoretical properties of VB for linear regression models with independent normal inverse Gamma priors.<sup>2</sup> They prove that VB point estimates are statistically valid in a frequentist sense. That is, they are consistent and provide asymptotically valid standard errors. All the models in this paper use hierarchical normal priors for regression coefficients which are independent of the error variances. The prior hierarchies depend on model-specific prior hyperparameters. Thus, conditional on the prior hyperparameters, the theory of You, Omerod and Muller (2014) applies immediately to the VB methods used in this paper. And, since the influence of the hierarchical prior will vanish asymptotically, their theoretical results will also apply unconditionally. Thus, we have a strong frequentist justification for our Bayesian methods.

Note that the MCMC type of approximation error is under the control of the researcher and can be made arbitrarily small by taking a sufficient number of draws; and MCMC methods are sometimes referred to as ‘exact’ as a consequence. That ‘exactness’, however, comes at a cost in terms of computational speed, which becomes infeasible when the number of draws required to efficiently explore a very large parameter space entails a prohibitive computational cost. By contrast, VB methods are typically much faster. This speed advantage, coupling with VB’s nice statistical properties, makes VB an effective alternative approach to estimating models with too many parameters that are infeasible to estimate using MCMC.

### 3 Variational Bayes Methods for VARs Using Conventional Priors

In this section, we describe VB methods for a conventional, non-hierarchical, prior. We do this not because it is important in and of itself, but because this is a key building block for our VB algorithms using hierarchical shrinkage priors. That is, there are many non-hierarchical priors which have been used with VARs in the past which involve subjectively-elicited prior hyperparameters. See, for instance, Dieppe, Legrand and van Roye (2016) which discusses a range of popular priors including the Minnesota prior, the natural conjugate prior and the independent normal-Wishart prior. For the Minnesota and natural conjugate priors, analytical posterior and predictive results are available. Hence, MCMC methods are not required and they can be used with large VARs. However, these priors have restrictive properties and cannot easily be extended (e.g. to allow for stochastic volatility) without resort to MCMC methods. With the independent normal-Wishart prior MCMC methods are required.<sup>3</sup> All of these are conventional, subjectively-elicited, non-hierarchical priors.

---

<sup>2</sup>The fact that the prior exhibits independence between regression coefficients and the error variance is important. Theoretical properties of VB for natural conjugate priors which do not exhibit such independence are different.

<sup>3</sup>VB methods for this prior have been developed in Hajargasht and Wozniak (2018).

In this paper, our interest lies in developing fast VB approaches to estimating large VARs using hierarchical priors which allow for automatic shrinkage. We stress that, due to the huge computational cost, it is not feasible to use MCMC to estimate such models. But all our priors will be hierarchical extensions of a conventional prior. Hence, we begin with a conventional VAR prior in this section.

Throughout this paper, we work with the following VAR (or extensions of it):

$$\mathbf{A}_0 \mathbf{y}_t = \mathbf{b}_0 + \mathbf{B}_1 \mathbf{y}_{t-1} + \dots + \mathbf{B}_p \mathbf{y}_{t-p} + \epsilon_t, \epsilon_t \sim N(0, \Sigma), \quad (3)$$

for  $t = 1, \dots, T$  where  $\mathbf{y}_t$  is an  $n \times 1$  vector of endogenous variables,  $\mathbf{b}_0$  is a  $n \times 1$  vector of intercept terms,  $\mathbf{B}_i$  is the  $n \times n$  matrix of lag  $i$  VAR coefficients,  $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$  and  $\mathbf{A}_0$  is an  $n \times n$  lower triangular matrix with ones on the diagonal.

We can rewrite (3) as

$$\mathbf{y}_t = \mathbf{X}_t \beta + \mathbf{W}_t \mathbf{a} + \epsilon_t, \quad (4)$$

where  $\mathbf{X}_t = \mathbf{I}_n \otimes [1, \mathbf{y}'_{t-1}, \dots, \mathbf{y}'_{t-p}]$  is an  $n \times K$  matrix,  $\beta = \text{vec}([\mathbf{b}_0, \mathbf{B}_1, \dots, \mathbf{B}_p]')$  is  $K \times 1$  vector of coefficients,  $\mathbf{a}$  consists of the free elements of  $\mathbf{A}_0$  stacked by rows with  $\mathbf{W}_t$  being the  $n \times m$  matrix containing the appropriate contemporaneous elements of  $\mathbf{y}_t$ . Equation (4) can be written in terms of  $n$  independent equations, with the  $i^{\text{th}}$  equation being:

$$y_{i,t} = \mathbf{z}_{i,t} \theta_i + \epsilon_{i,t}, \epsilon_{i,t} \sim N(0, \sigma_i^2). \quad (5)$$

where  $\mathbf{z}_{i,t}$  is a row vector with  $k_i$  elements and  $\theta_i$  is a vector containing the elements of  $\beta$  and  $\mathbf{a}$  pertaining to the  $i^{\text{th}}$  equation. Below we also use notation where  $\mathbf{Z}_i = (\mathbf{z}_{i,1}, \dots, \mathbf{z}_{i,T})'$ ,  $\mathbf{y}_i = (y_{i,1}, \dots, y_{i,T})'$  and  $\epsilon_i = (\epsilon_{i,1}, \dots, \epsilon_{i,T})'$ .

There are two advantages of writing the VAR in this form. The first advantage is computational. This specification allows for equation-by-equation estimation of the VAR. This breaks the task of working with the huge  $K$  dimensional vector of VAR coefficients into that of working with  $n$  smaller  $k_i$  dimensional sets of regression coefficients. As documented in, e.g., Carriero, Clark and Marcellino (2016), working directly with the posterior covariance matrix for all the VAR coefficients jointly involves  $O(n^6)$  manipulations whereas equation-by-equation estimation reduces this to  $O(n^4)$ . For large values of  $n$  the computational benefits of this are huge. Secondly, the elements of  $\mathbf{A}_0$  relate to the error covariances of the reduced form VAR (i.e. the latter covariance can be written as  $(\mathbf{A}_0^{-1})\Sigma(\mathbf{A}_0^{-1})'$ ). When  $n$  is large, the number of error covariances can be large and it can be desirable to shrink many of them to zero. Using the specification in (5) means that this shrinkage can easily be done using the same prior as is used on the VAR coefficients.

The prior we use for the parameters in the  $i^{th}$  equation is:<sup>4</sup>

$$\theta_i \sim N(0, \mathbf{V}_i), \tag{6}$$

$$\sigma_i^{-2} \sim G(\underline{\nu}, \underline{s}), \tag{7}$$

where  $G$  denotes the Gamma distribution. We call this the normal independent prior.

Textbook derivations for the normal linear regression model with independent normal-Gamma prior (e.g. chapter 3 of Koop, 2003) can be used to derive the full conditional posteriors. You, Ormerod and Muller (2014) derive the VB approximating densities using these full conditional posteriors and (2). For equation  $i$ , these are

$$q(\theta_i) \sim N(\bar{\theta}_i, \bar{\mathbf{V}}_i), \tag{8}$$

$$q(\sigma_i^{-2}) \sim G(\underline{\nu} + \frac{T}{2}, \bar{s}_i), \tag{9}$$

where

$$\bar{\mathbf{V}}_i = [(\frac{\underline{\nu} + \frac{T}{2}}{\bar{s}_i}) \mathbf{Z}'_i \mathbf{Z}_i + \mathbf{V}_i^{-1}]^{-1},$$

$$\bar{\theta}_i = (\frac{\underline{\nu} + \frac{T}{2}}{\bar{s}_i}) \bar{\mathbf{V}}_i \mathbf{Z}'_i \mathbf{y}_i, \tag{10}$$

$$\bar{s}_i = \underline{s} + \frac{1}{2} \|\mathbf{y}_i - \mathbf{Z}_i \bar{\theta}_i\|^2 + \frac{1}{2} tr(\mathbf{Z}'_i \mathbf{Z}_i \bar{\mathbf{V}}_i). \tag{11}$$

Note that the VB approximating densities depend on three arguments:  $\bar{\theta}_i$ ,  $\bar{\mathbf{V}}_i$  and  $\bar{s}_i$ . These are optimized in an iterative process.<sup>5</sup> Beginning with an initialization of any two of these, the algorithm iterates using the preceding formulae. After each iteration,  $ELBO_i$  is calculated. Iteration continues until the increase in  $ELBO_i$  between the  $j^{th}$  and  $(j - 1)^{th}$  iteration is less than some convergence criterion. The formula for

---

<sup>4</sup>We adopt a notational convention where prior hyperparameters selected by the researcher are denoted using lower bars. We do not adopt this convention for  $\mathbf{V}_i$  since, in the next section, we will use a hierarchical structure which means it will depend on other parameters. Our notation also assumes that most prior hyperparameters are chosen to be the same in every equation. This can be trivially relaxed by adding  $i$  subscripts to the prior hyperparameters.

<sup>5</sup>Throughout this paper, we adopt a notational convention where upper bars denote quantities which are optimized in a VB algorithm.

$ELBO_i$  is given in the Technical Appendix. This algorithm is done independently for each of the  $i = 1, \dots, n$  equations, which means it can be parallelized to increase computational efficiency.

## 4 Variational Bayes Methods for the VAR with Hierarchical Shrinkage Priors

We have emphasized the fact that, with large VARs, over-parameterization concerns can be serious and, thus, Bayesian prior shrinkage is desirable. In this section, we develop VB methods for a range of priors which do this shrinkage in an automatic fashion. These priors are all hierarchical and have been used in the machine learning literature. These all are hierarchical extensions of the VAR and prior of the preceding section. That is, whereas the prior of the preceding section depended on hyperparameters chosen by the researcher, in this section we will work with priors that involve a hierarchical structure and require less input from the researcher. But, conditional on a particular hierarchy, all the theoretical results derived above still hold and we will draw upon them in this section.

### 4.1 Adaptive shrinkage t-prior

The adaptive shrinkage t-prior, as used in, e.g., Korobilis (2013), adopts the same prior at the first level of the hierarchy as the conventional prior of Section 3. However, the prior covariance matrix for the coefficients in equation  $i$  becomes:

$$\mathbf{V}_i = \text{diag}(\tau_{i,1}, \dots, \tau_{i,k_i}). \quad (12)$$

The degree of shrinkage is controlled by  $\tau_i = (\tau_{i,1}, \dots, \tau_{i,k_i})'$  which are treated as unknown parameters. The prior for each of these is

$$\tau_{i,j}^{-1} \sim G(\underline{a}_0, \underline{b}_0), \quad \text{for } j = 1, \dots, k_i.$$

The VB approximating densities,  $q(\theta_i)$  and  $q(\sigma_i^2)$  are the same as (8) and (9) since their conditional posteriors (now additionally conditional on  $\tau_i$ ) are the same as in the preceding section. Hence, we only need to derive  $q(\tau_i^{-1})$ . Given the form of the conditional posterior for  $\tau_{i,j}$  given in Korobilis (2013), we can derive:

$$q(\tau_{i,j}^{-1}) \sim G(\underline{a}_0 + \frac{1}{2}, \frac{\bar{\theta}_{i,j}^2 + \bar{\mathbf{V}}_i^{jj}}{2} + \underline{b}_0),$$



where  $\bar{\mathbf{V}}_i^{jj}$  is the  $(j, j)^{th}$  element of  $\bar{\mathbf{V}}_i$ , Thus, the new term that VB updates is

$$\overline{\tau_{i,j}^{-1}} = \frac{\underline{a}_0 + \frac{1}{2}}{\frac{\bar{\theta}_{i,j}^2 + \bar{\mathbf{V}}_i^{jj}}{2} + \underline{b}_0}.$$

As before, VB iterates over  $\bar{\theta}_i$ ,  $\bar{\mathbf{V}}_i$  and  $\bar{s}_i$ , but now we additionally have to iterate over  $\overline{\tau_{i,j}^{-1}}$ . The ELBO used to assess convergence is given in the Technical Appendix.

We also use the adaptive shrinkage Jeffreys' prior (see Korobilis, 2013) which takes the form

$$\tau_{i,j} \sim \frac{1}{\tau_{i,j}}, \quad \text{for } j = 1, \dots, \tau_{i,k_i}.$$

This can be viewed as a special case of the adaptive shrinkage t-prior with  $\underline{a}_0 = \underline{b}_0 = 0$ .

## 4.2 The Adaptive LASSO

The adaptive LASSO maintains the prior covariance matrix given in (12), but allows for a different treatment of the prior shrinkage parameters,  $\tau_i$ . In particular, it assumes:

$$\tau_{i,j} \sim \text{Exp}\left(\frac{\lambda_{i,j}}{2}\right), \quad \text{for } j = 1, \dots, k_i$$

with

$$\lambda_{i,j} \sim G(\underline{a}_0, \underline{b}_0).$$

With this hierarchical shrinkage prior, the optimal VB approximating densities for  $q(\theta_i)$  and  $q(\sigma_i^{-2})$  are the same as in Section 3, but we now add approximating densities for  $\tau_i$  and  $\lambda_i$  where  $\lambda_i = (\lambda_{i,1}, \dots, \lambda_{i,k_i})'$ . These are

$$q(\tau_{i,j}^{-1}) \sim iG\left(\sqrt{\frac{\bar{\lambda}_{i,j}}{\bar{\theta}_{i,j}^2 + \bar{\mathbf{V}}_i^{jj}}}, \bar{\lambda}_{i,j}\right),$$

where  $iG$  denotes the inverse Gaussian distribution and

$$q(\lambda_{i,j}) \sim G(\underline{a}_0 + 1, 0.5\bar{\tau}_{i,j} + \underline{b}_0).$$

These involve the following terms to be iterated in the VB algorithm:

$$\overline{\tau_{i,j}^{-1}} = \sqrt{\frac{\bar{\lambda}_{i,j}}{\bar{\theta}_{i,j}^2 + \bar{\mathbf{V}}_i^{jj}}},$$

$$\bar{\lambda}_{i,j} = \frac{a_0 + 1}{0.5\bar{\tau}_{i,j} + b_0},$$

and  $\bar{\tau}_{i,j}$  denotes  $\frac{1}{\tau_{i,j}^{-1}}$ .

The evidence lower bound is given in the Technical Appendix. In our empirical section, we also use the Bayesian LASSO of Park and Casella (2008). This is the same as the adaptive LASSO but sets  $\lambda_{i,j} = \lambda_i$ , so that we now have a global shrinkage parameter which is the same for all coefficients in equation  $i$ .

### 4.3 Horseshoe prior

Another popular hierarchical shrinkage prior is the horseshoe prior of Carvalho, Polson and Scott (2010). It has attractive theoretical properties, including an ability to adapt to different patterns of sparsity and has been found to be quite robust.

To the equation-by-equation VAR set up involving (5), (6) and (7), the horseshoe prior adds the assumptions that:

$$\mathbf{V}_i = \text{diag}(\lambda_{i,1}\tau_i, \dots, \lambda_{i,k_i}\tau_i),$$

where the priors for the new parameters are

$$\lambda_{i,j}^{-1}|\nu_{i,j} \sim G\left(\frac{1}{2}, \frac{1}{\nu_{i,j}}\right),$$

$$\tau_i^{-1}|\xi_i \sim G\left(\frac{1}{2}, \frac{1}{\xi_i}\right),$$

$$\nu_{i,1}^{-1}, \dots, \nu_{i,k_i}^{-1}, \xi_i^{-1} \sim G\left(\frac{1}{2}, 1\right)$$

and  $i$  indexes equations and  $j$  indexes coefficients.

The optimal  $q(\theta_i)$  and  $q(\sigma_i^{-2})$  are the same as in preceding sub-sections. The conditional posteriors for the remaining parameters using the horseshoe prior can be found in Makalic and Schmidt (2015). These can be used to derive:

$$q(\lambda_{i,j}^{-1}) \sim G\left(1, \overline{\nu_{i,j}^{-1}} + \frac{\bar{\theta}_{i,j}^2 + \bar{\mathbf{V}}_i^{jj}}{2} \overline{\tau_i^{-1}}\right),$$

$$q(\tau_i^{-1}) \sim G\left(\frac{k_i + 1}{2}, \overline{\xi_i^{-1}} + \frac{1}{2} \overline{\lambda_{i,j}^{-1}} \sum_{j=1}^{k_i} (\bar{\theta}_{i,j}^2 + \bar{\mathbf{V}}_i^{jj})\right),$$

$$q(\nu_{i,j}^{-1}) \sim G\left(1, 1 + \overline{\lambda_{i,j}^{-1}}\right)$$

and

$$q(\xi_i^{-1}) \sim G(1, 1 + \overline{\tau_i^{-1}}).$$

The terms which are updated in the VB iterations are (10), (11),

$$\overline{\lambda_{i,j}^{-1}} = \frac{1}{\nu_{i,j}^{-1} + \tau_i^{-1} \frac{\overline{\theta_{i,j}^2} + \overline{\mathbf{V}_i^{jj}}}{2}},$$

$$\overline{\tau_i^{-1}} = \frac{k_i + 1}{2\xi_i^{-1} + [\overline{\lambda_{i,j}^{-1}} \sum_{j=1}^{k_i} (\overline{\theta_{i,j}^2} + \overline{\mathbf{V}_i^{jj}})]},$$

$$\overline{\nu_{i,j}^{-1}} = 1/(1 + \overline{\lambda_{i,j}^{-1}}),$$

$$\overline{\xi_i^{-1}} = 1/(1 + \overline{\tau_i^{-1}}).$$

These values can be plugged into the formula for  $\mathbf{V}_i$  and used to update  $\overline{\mathbf{V}_i}$ . The formula for the evidence lower bound used to assess convergence is given in the Technical Appendix.

#### 4.4 SSVS

One of the widely used hierarchical shrinkage priors is the SSVS prior which assumes that  $\mathbf{V}_i = \text{diag}(v_{i,1}, \dots, v_{i,k_i})$  and

$$v_{i,j} = \begin{cases} \underline{\kappa}_{i,j,0} & \text{if } \gamma_{i,j} = 0 \\ \underline{\kappa}_{i,j,1} & \text{if } \gamma_{i,j} = 1 \end{cases}$$

where  $\underline{\kappa}_{i,j,0}$  is chosen to be large and  $\underline{\kappa}_{i,j,1}$  to be small. In words, if  $\gamma_{i,j} = 1$  then a prior which strongly shrinks the  $j^{\text{th}}$  coefficient in the  $i^{\text{th}}$  equation towards zero is used. The prior for  $\gamma_i = (\gamma_{i,1}, \dots, \gamma_{i,k_i})$  follows a Bernoulli distribution

$$P(\gamma_{i,j} = 1) = \underline{\pi}_{i,j},$$

with

$$P(\gamma_{i,j} = 0) = 1 - \underline{\pi}_{i,j}.$$

The VB approximating densities for  $\theta_i$  and  $\sigma_i^2$  are the same as in the preceding section. The remaining approximating densities can be derived based on posterior conditionals given in George, Sun and Ni (2008).

The approximating density for  $\gamma_i$  is

$$q(\gamma_i) \propto \text{Bernoulli}(\bar{\pi}_{i,j})$$

where

$$\bar{\pi}_{i,j} = \frac{\frac{1}{\kappa_{i,j,1}} \exp\left(-\frac{\bar{\theta}_{i,j}^2 + \mathbf{V}_i^{jj}}{2\kappa_{i,j,1}^2}\right) \pi_{i,j}}{\frac{1}{\kappa_{i,j,1}} \exp\left(-\frac{\bar{\theta}_{i,j}^2 + \mathbf{V}_i^{jj}}{2\kappa_{i,j,1}^2}\right) \pi_{i,j} + \frac{1}{\kappa_{i,j,0}} \exp\left(-\frac{\bar{\theta}_{i,j}^2 + \mathbf{V}_i^{jj}}{2\kappa_{i,j,0}^2}\right) (1 - \pi_{i,j})}.$$

Finally, we have

$$\mathbf{V}_i = \text{diag}(\bar{v}_{i,1}, \dots, \bar{v}_{i,k_i}).$$

where

$$\bar{v}_{i,j} = \bar{\pi}_{i,j} \kappa_{i,j,1} + (1 - \bar{\pi}_{i,j}) \kappa_{i,j,0}.$$

The evidence lower bound for the VAR with SSVS prior is given in the Technical Appendix.

## 4.5 Adding Stochastic Volatility to the VAR

Many papers, using many different macroeconomic data sets, have found stochastic volatility to be an important feature and that failing to take it into account can lead to poor forecasting performance (see, e.g., Clark, 2011). Thus, it is important to develop methods for adding stochastic volatility to the VAR using any of the priors in the preceding sub-sections. In this sub-section, we do so with a VB method.

We assume the model is the same as in any of the preceding sub-sections, except that the error variance in equation  $i$  is now  $\exp(h_{i,t})$  where

$$h_{i,t} = h_{i,t-1} + \zeta_{i,t}, \quad \zeta_{i,t} \sim N(0, \sigma_{h_i}^2),$$

$$\sigma_{h_i}^{-2} \sim G(\underline{a}_1, \underline{b}_1).$$

$$h_{i,0} \sim N(0, \underline{V}_{i,h}),$$

where the initial conditions  $h_{i,0}$  are treated as parameters to be estimated.

The auxiliary mixture sampler of Kim, Shephard and Chib (1998) is a popular way of doing MCMC with stochastic volatility models. We use this sampler for our MCMC results, but it is very slow precluding its use in large models. Instead we use VB assuming the VB density of the log-volatilities is normal. There are several ways of obtaining the mean and variance of this single normal distribution using VB methods. In a recent

paper, Chan and Yu (2020) compare different VB methods involving normal approximations for stochastic volatility models and propose a new one. They demonstrate, both theoretically and in practice, that their new algorithm is more accurate than previous algorithms. Accordingly, this is the VB algorithm we use in this paper.

The Technical Appendix provides complete details of this algorithm. Here we outline the justification for and steps involved in it. Let  $\mathbf{h}_i = (h_{i,1}, \dots, h_{i,T})'$ . From (2) the optimal VB density for  $\mathbf{h}_i$  is

$$q_{h_i}(\mathbf{h}_i) \propto \exp \left[ E \left( \log p(\mathbf{h}_i | y, \theta_i, h_{i,0}, \sigma_{h_i}^2) \right) \right].$$

Chan and Yu (2020) derive this unrestricted optimal VB density and note that it is not normal. They use a normal optimal VB density which is as close as possible to this unrestricted optimal VB density in a KL sense. The minimization of the KL distance is done using the Newton-Raphson method. This minimization problem can be done quickly since, as shown by Chan and Yu (2020), the Hessian of the objective function is banded and fast band matrix routines can be exploited. Note that this is a global approximation to the joint distribution of the entire vector of log-volatilities in equation  $i$  and Chan and Yu (2020) show, in a Monte Carlo study, the approximation to be very accurate.

It is worth noting that other VB approximations for stochastic volatility models have been proposed. The theoretical properties of some of these are discussed in Frazier, Loaiza-Maya and Martin (2021). We highlight this paper since it compares the method of Chan and Yu (2020) to one based on the methods of Loaiza-Maya, Smith, Nott and Danaher (2021). It discusses some possible theoretical weaknesses of the Chan and Yu (2020) method (i.e. a lack of Bayesian consistency), but also concludes that in terms of prediction (particularly at shorter horizons) it performs almost as well as the more accurate method of Loaiza-Maya, Smith, Nott and Danaher (2021). However, the latter method includes an MCMC step for drawing the volatilities thus adding substantially to the computational burden. Given the need for fast computation in our high dimensional models, in this paper we use the methods of Chan and Yu (2020). However, it would be simple to modify our methods by replacing the blocks of the algorithm which use the methods of Chan and Yu (2020) with blocks based on Loaiza-Maya, Smith, Nott and Danaher (2021).

## 4.6 Choice of Prior Hyperparameters

With the exception of the horseshoe prior and Jeffreys' prior, our priors involve hyperparameters which must be selected. The Technical Appendix provides the values we use for these. Here we describe the general issues which infuse our choices. In extensive experimentation, we have found it is not acceptable to simply use the same choices for all VAR dimensions. This is unsurprising. Each equation in the VAR has  $np + 1$  right-hand

side variables most of which are probably unimportant. As VAR dimension increases the number of right-hand side increases and the need for a prior which induces sparsity increases. Our prior hyperparameter choices reflect this. We have found that working with relatively non-informative priors is fine if  $n = 10$  or even 20, but not with  $n = 100$ . Accordingly, for the t-prior, both variants of the LASSO and the SSVS prior, our prior hyperparameters depend on  $n$  and  $p$  and induce a higher degree of shrinkage in larger models.

For the non-informative Jeffreys' prior, adding increasing shrinkage as  $n$  increases is not possible. In an earlier version of this paper we found large VAR models adopting Jeffreys' prior to forecast very poorly. This inadequate shrinkage suggests that it is unsuitable for use in VAR of very large dimensions. The horseshoe prior, too, involves no hyperparameters. There is evidence that high-dimensional VAR using horseshoe prior, too, forecasts poorly. We have found the reason for this to be that the prior for  $\tau_i^{-1}$  given in (4.3) allocates too much prior probability to non-sparse regions of the parameter space. However, we have found that simply fixing  $\tau_i$  to a value which implies tighter shrinkage as VAR dimension increases work much better. The results in the empirical section of this paper reflect such an approach and, for the large VAR with  $n = 100$ , set

$$\bar{\tau}_i = \frac{1}{K + m},$$

where  $K$  is the number of VAR coefficients and  $m$  is the number of elements in  $\mathbf{a}$ . Adopting the same strategy for Jeffreys' prior also improves forecast performance and, thus, in our empirical section we do so.

## 5 Empirical Work

In this section, we present evidence on the performance of VB methods with various hierarchical priors using quarterly US data from 1959Q4 through 2019Q4 taken from the Federal Reserve Bank of St. Louis' FRED-QD data set. All variables are transformed to stationarity following recommendations in the FRED-QD data base. All variables are standardized to have mean zero and standard deviation one.

We present results from VARs of various dimensions. Our small/medium/large data sets contains  $n = 10/20/100$  variables. The list of variables in each data set is given in the Data Appendix. The main justification for use of VB methods is that their computational burden is potentially much less than MCMC methods. This motivates our choice of VAR dimensions. With our small data set, MCMC computation is not that onerous and we can do extensive comparisons between VB and MCMC methods. With the large data set, a huge computation burden results when using MCMC methods and, hence, we focus on VB methods with this data set.

In this empirical exercise we aim to answer several questions. Two of these relate to computation time: How

much faster are VB methods than MCMC methods? and How scaleable are VB methods? These are addressed in the following sub-section. The subsequent sub-section addresses the question: How accurate are estimates produced by VB methods? It does so by comparing posteriors produced by VB (which is an approximate method) to those produced by MCMC (which, if a sufficient number of replications is taken, can be expected to be highly accurate). The third sub-section is a recursive forecasting exercise which addresses the question: How good are VB methods combined with hierarchical priors at macroeconomic forecasting? This sub-section offers a detailed comparison of forecast performance for the various priors and VAR dimensions. The final sub-section addresses the question: How accurate are the predictive densities produced by VB? It does so by comparing predictive densities produced by VB and MCMC using the small data set with a particular focus on the tails of the predictive densities.

The results in this paper use  $p = 1$  lag with results for longer lag lengths being put in the Empirical Appendix. In the forecasting exercise, different lag length choices tend to lead to very similar forecast performance but  $p = 1$  tends to forecast slightly better than longer lag lengths.

## 5.1 Computation Time

Table 1 presents computation time in seconds to estimate a model using a standard desktop with an Intel Core i7-7700 @ 3.6GHz processor and 16 GB of RAM. For MCMC methods we take 22,000 draws and discard an initial 2,000 burn-in draws. These values lead to convergence as assessed by standard MCMC diagnostics. For VB methods, we judge convergence to have occurred when the change in the ELBO is less than  $10^{-4}$ . Computation time is likely to depend mostly on VAR dimension and, hence, we present results for the VARs of different dimensions. Remember that estimation proceeds one equation at a time and that the number of VAR coefficients in the  $i^{th}$  equation is  $k_i = n * p + i$ . But the number of prior hyperparameters to be estimated in the prior covariance matrix  $\mathbf{V}_i$  can also have an impact on computation time. Accordingly, Table 1 lists the number of these to be estimated for the  $i^{th}$  equation for each prior.

Table 1: Computation times (in seconds)

Models	No. params. in prior cov.	10 variables		20 variables		100 variables	
		MCMC	VB	MCMC	VB	MCMC	VB
Homoskedastic							
normal-independent	0	30.4	0.2	69.5	0.3	1541.9	1.3
Horseshoe	$k_i + 1$	41.9	0.6	91.8	1.1	1756.2	10.3
LASSO	$k_i$	123.8	0.6	242.0	1.0	2649.2	11.8
Adaptive LASSO	$2k_i$	136.4	0.7	244.7	0.7	2671.6	3.3
t-prior	$k_i$	34.6	0.2	73.1	0.3	1603.2	1.3
SSVS	$2k_i$	33.2	0.2	71.0	0.3	1559.1	1.7
Jeffreys'	$k_i$	65.0	0.3	136.8	0.4	1976.7	17.3
Heteroskedastic							
normal-independent	0	182.7	5.1	399.4	8.0	3100.7	110.1
Horseshoe	$k_i + 1$	186.5	6.7	416.3	10.6	3327.2	89.4
LASSO	$k_i$	253.1	6.5	533.7	12.0	3609.8	145.5
Adaptive LASSO	$2k_i$	255.9	6.0	543.3	9.3	3575.5	125.5
t-prior	$k_i$	181.6	5.0	420.7	7.6	3406.6	110.5
SSVS	$2k_i$	174.9	4.6	421.0	7.3	3244.9	64.4
Jeffreys'	$k_i$	175.5	3.3	402.9	6.1	3481.0	181.1

The general picture here is that VB is much faster than MCMC. MCMC methods are very slow with the large data set. To estimate a single model takes roughly an hour for any of the priors considered in this paper when stochastic volatility is added. Clearly, running an extensive recursive forecasting exercise by repeatedly re-running the MCMC algorithm on an expanding window of data would lead to a huge computational burden. In contrast VB methods are much faster with the time for estimating a single model being a minute or two for the various priors with the large data set. The reason for this is largely due to the fact that the optimization problems within VB converge very quickly, usually within a few iterations. In contrast, MCMC involves taking a large number (here 22000) draws. The computation time for each iteration is roughly comparable to the computation time for each draw. The speed improvements of VB are simply due to its having fewer iterations than MCMC has draws.

VB methods are also found to be scaleable in the sense that the computational time is increasing roughly at a linear rate with  $n$  (e.g. computation times for the 100 variable models are roughly 10 times as big as those for 10 variable models). In contrast, MCMC methods are less scaleable with the computational burden increasing at a greater than linear rate. In this paper, we are working with a maximum of  $n = 100$ . With this value, MCMC methods are just feasible. But for larger values of  $n$  (e.g.  $n = 200$  or more) that researchers are



interested in working with, our results suggest VB methods are practical whereas MCMC methods are not.

The computation times for both VB and MCMC are similar across hierarchical priors indicating that the number of parameters in the prior covariance to be estimated has only a small impact on computation time. The VAR without a hierarchical prior will tend to have faster computation since it has fewer parameters to estimate.<sup>6</sup> But the addition of any of our hierarchical priors does not lead to large increases in the computational burden. Interestingly, the relatively parameter-rich SSVS prior is leading to relatively fast computation. This is because the variable inclusion indicators have Bernoulli distributions which can be handled very quickly.

Table 1 also shows that the inclusion of stochastic volatility will inevitably slow down computation. This slowdown is, proportionally, larger for VB than for MCMC. Consider, for instance, the 100 variables models. If we compare the heteroskedastic to homoskedastic versions of each model it can be seen that VB computation times are up to 100 times larger for the former than the latter. However, for MCMC the computation times approximately double. Hence, the computational benefits of using VB lie mostly in its faster estimation of huge numbers of VAR coefficients with lesser benefits arising from the VB treatment of stochastic volatility. Nonetheless, the most important revelation of our exercises is that even with SV incorporated, VB is still much faster than MCMC.

## 5.2 The Accuracy of VB

The accuracy of VB estimation can be investigated by comparing VB results to MCMC results. We have done extensive comparisons and found VB to be highly accurate. For the sake of brevity we do not report a full set of results for our many different priors, VAR dimensions and parameters. We illustrate our findings in Figures 1, 2, 3 and 4. These figures are for the small VAR-SV with adaptive LASSO prior. Figures 1, 2 and 3 produce parameter estimates for three key variables. They plot VB point estimates<sup>7</sup> and MCMC posterior means for each individual VAR coefficient, covariance term (i.e. the vector of parameters we call  $\mathbf{a}$ ) and the volatilities. Figure 4 compares impulse responses to a monetary policy shock, where in the VB's case, the impulse responses are derived using the point estimates of the parameters, while in the case of MCMC, the plotted impulse responses are the posterior means of the impulse responses computed in each iterations. The Empirical Appendix contains comparable figures for the other priors.

All of these figures show that VB and MCMC posterior means are virtually identical and in the few cases they are not they are at least very similar. Given that the stochastic volatility model is not a normal linear

<sup>6</sup>The exception to this arises in the heteroskedastic model and is due to slower convergence of the Newton-Raphson optimization that is required to find the global normal approximation in the Chan and Yu (2020) algorithm. This optimization is required for both VB and MCMC.

<sup>7</sup>VB point estimates are the means of the VB approximating densities.

state space model, it is possible that using a normal approximation to the posterior of the volatilities will be poor. But Figure 3 indicates that this is not the case.

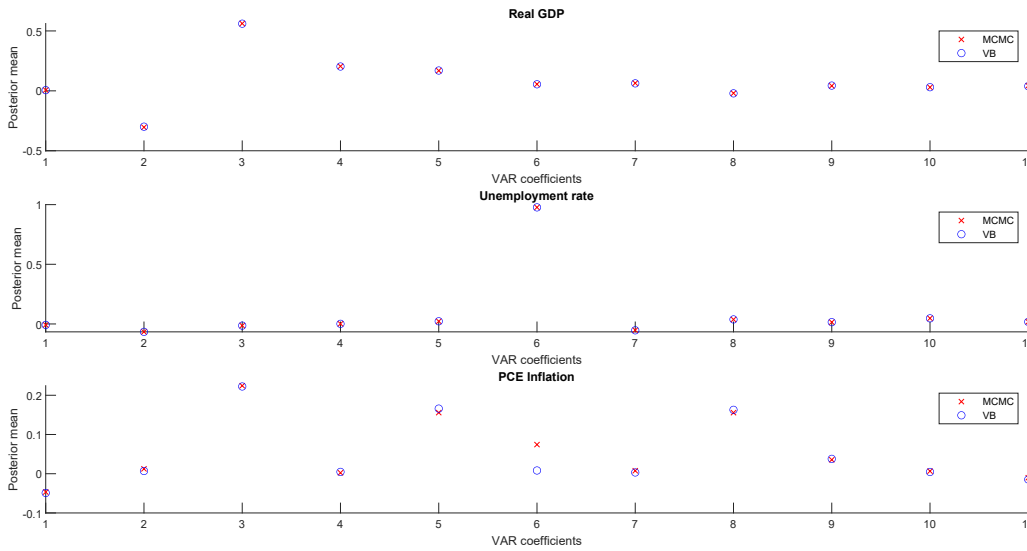


Figure 1: VAR Coefficients: Comparison of MCMC posterior means and VB point estimates

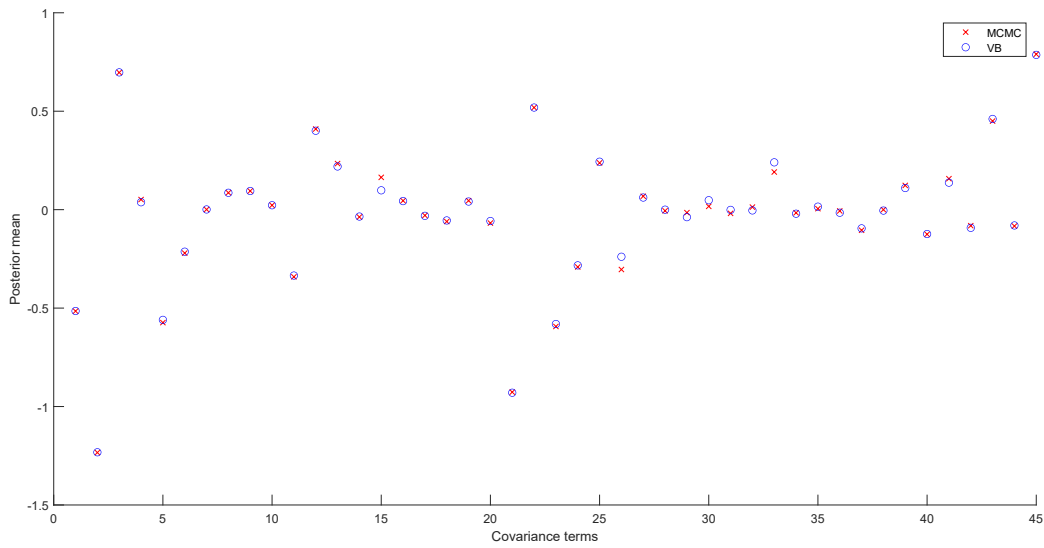


Figure 2: Covariances: Comparison of MCMC posterior means and VB point estimates

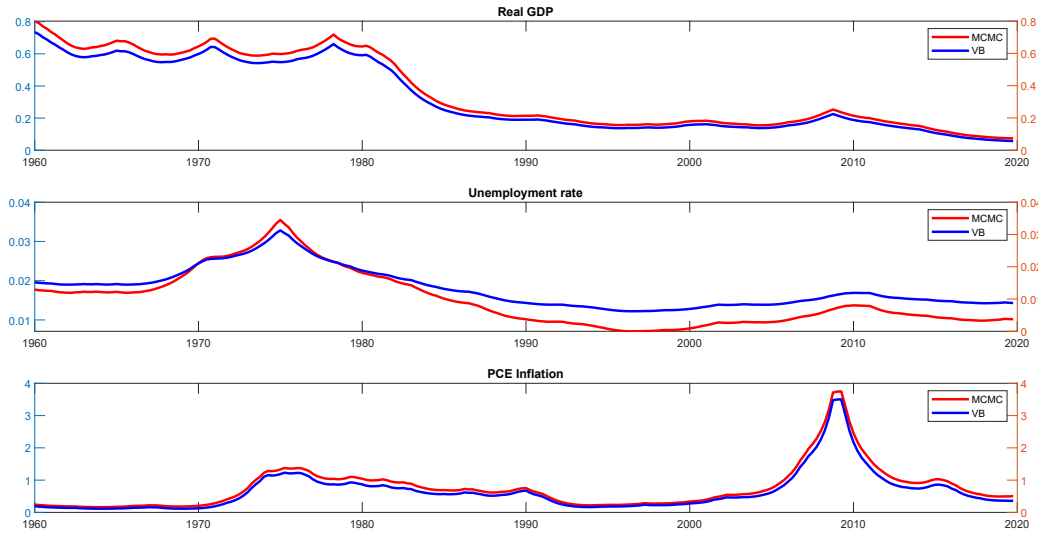


Figure 3: Volatilities: Comparison of MCMC posterior means and VB point estimates

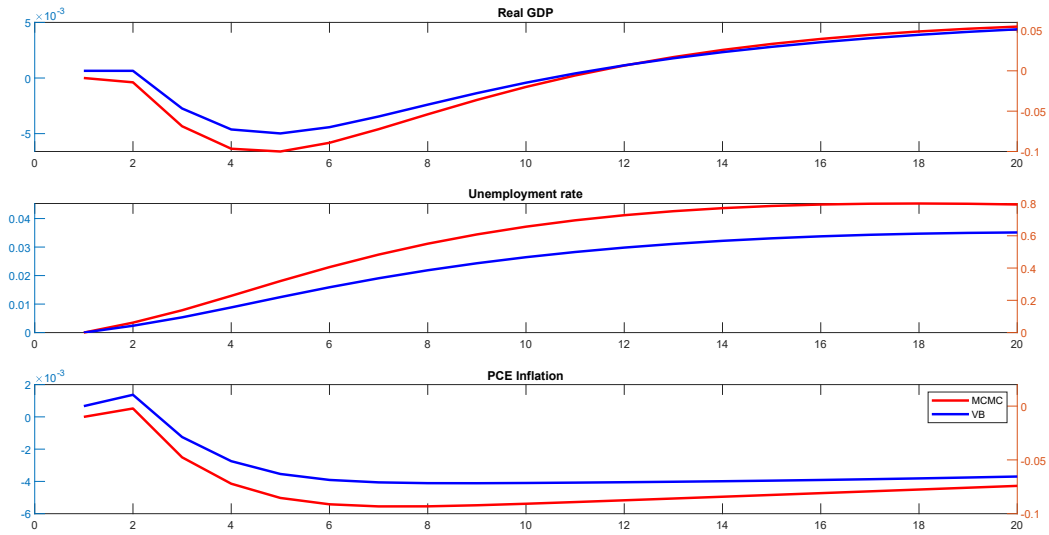


Figure 4: Impulse Responses: Comparison of results derived via MCMC and VB

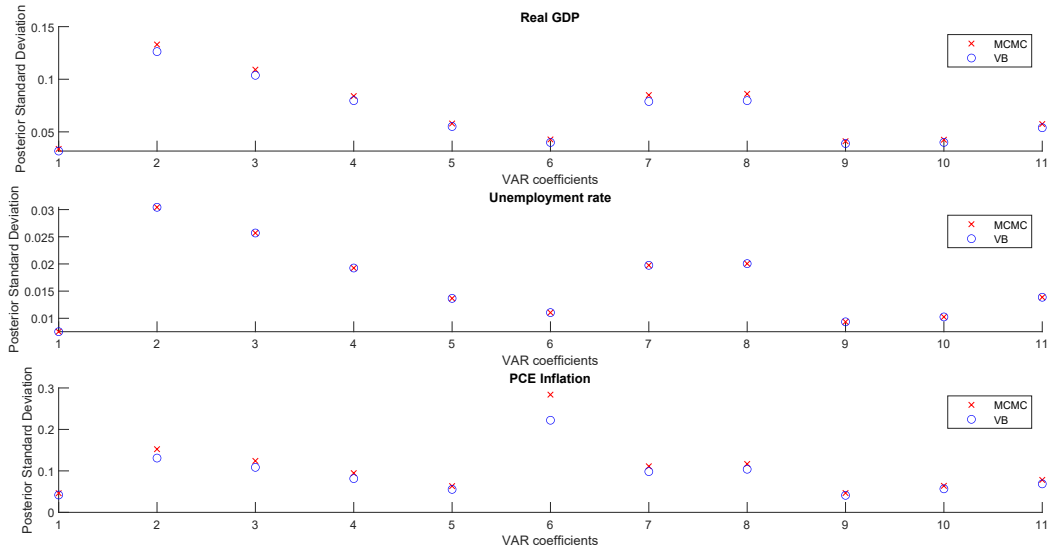


Figure 5: VAR Coefficients: Comparison of MCMC posterior standard deviations and VB standard deviations

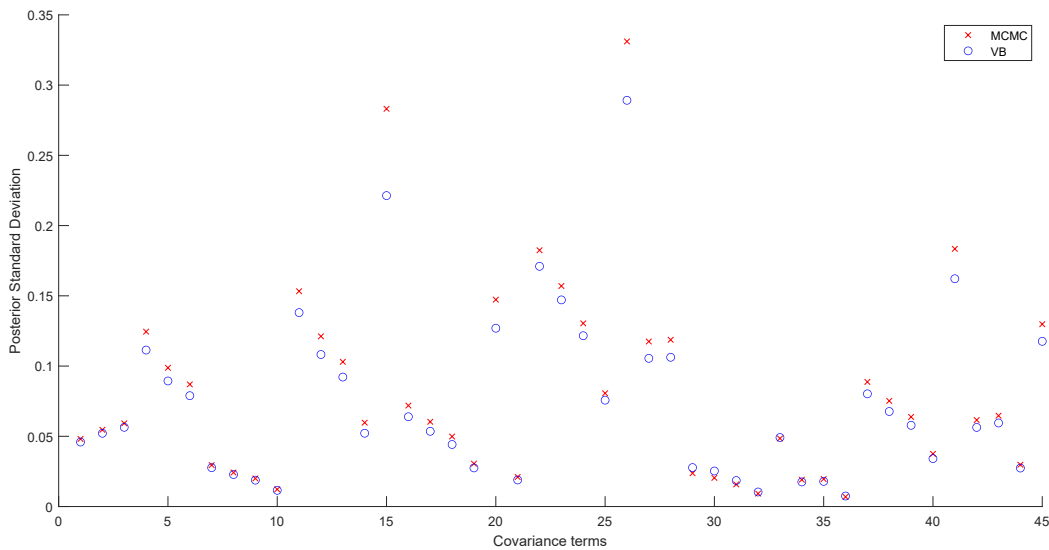


Figure 6: Covariances: Comparison of MCMC posterior standard deviations and VB standard deviations

As a summary of the accuracy of VB for all the priors, we present Tables 2, 3 and 4. These tables, which are for models with  $n = 10$ , contain summary statistics across all VAR coefficients or across all error covariance terms of the absolute value of the difference between the MCMC and VB estimates. For the homoskedastic version of each model we also present findings for the error variances.

Tables 2, 3 and 4 show VB results to be very accurate. The median of our divergence measure is very small

for every prior. With one exception, it is never greater than 0.01. The one exception is for Jeffreys' prior, but even here the median absolute divergence between VB and MCMC is small. For most of the priors, the maximum divergence is also very small. Again, the main exception is Jeffreys' prior which has a small number of coefficients where the divergence is larger. Overall, we are finding VB to be highly accurate.

Table 2: Absolute Value of Deviations between MCMC posterior means and VB point estimates- VAR coefficients

Models	Homoskedastic				Heteroskedastic			
	Median	10th percentile	90th percentile	Max	Median	10th percentile	90th percentile	Max
normal-independent	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06
Horseshoe	0.01	0.00	0.02	0.06	0.01	0.00	0.03	0.11
LASSO	0.00	0.00	0.02	0.16	0.00	0.00	0.02	0.10
Adaptive LASSO	0.00	0.00	0.01	0.07	0.00	0.00	0.01	0.07
t-prior	0.00	0.00	0.00	0.02	0.00	0.00	0.01	0.04
SSVS	0.00	0.00	0.03	0.69	0.00	0.00	0.02	0.26
Jeffreys'	0.00	0.00	0.10	0.26	0.02	0.00	0.10	0.57

Table 3: Absolute Value of Deviations between MCMC posterior means and VB point estimates - Covariance terms

Models	Homoskedastic				Heteroskedastic			
	Median	10th percentile	90th percentile	Max	Median	10th percentile	90th percentile	Max
normal-independent	0.00	0.00	0.00	0.00	0.01	0.00	0.03	0.06
Horseshoe	0.01	0.00	0.04	0.10	0.01	0.00	0.06	0.12
LASSO	0.00	0.00	0.08	0.16	0.01	0.00	0.04	0.09
Adaptive LASSO	0.00	0.00	0.02	0.07	0.01	0.00	0.02	0.07
t-prior	0.00	0.00	0.00	0.02	0.00	0.00	0.03	0.07
SSVS	0.01	0.00	0.17	0.69	0.01	0.00	0.10	0.26
Jeffreys'	0.01	0.00	0.14	0.24	0.03	0.00	0.24	0.58

Table 4: Absolute Value of Deviations between MCMC posterior means and VB point estimates - Error Variances

Models	Homoskedastic				Heteroskedastic			
	Median	10th percentile	90th percentile	Max	Median	10th percentile	90th percentile	Max
normal-independent	0.00	0.00	0.00	0.00	-	-	-	-
Horseshoe	0.00	0.00	0.01	0.02	-	-	-	-
LASSO	0.00	0.00	0.00	0.00	-	-	-	-
Adaptive LASSO	0.00	0.00	0.00	0.00	-	-	-	-
t-prior	0.00	0.00	0.00	0.00	-	-	-	-
SSVS	0.00	0.00	0.01	0.01	-	-	-	-
Jeffreys'	0.02	0.01	0.05	0.06	-	-	-	-

This sub-section discusses the accuracy of VB by comparing VB point estimates to MCMC posterior means. It is well known that VB methods have a tendency to underestimate posterior variances, see for example Giordano, Broderick and Jordan (2018). This is shown in Figures 5 and 6, which present the posterior standard deviations that are associated with the posterior means plotted in Figures 1 and 2. It can be seen that the VB standard deviations are consistently somewhat smaller than the MCMC ones. We would expect underestimated posterior variances to lead to under-estimated predictive variances which could have an impact on forecast performance. This is an issue we will investigate in the following sub-section

### 5.3 Forecasting Comparison

In this sub-section, we carry out a forecasting exercise using our small, medium and large data sets. We forecast three variables: GDP growth, inflation (based on the PCE price index) and the unemployment rate for forecast horizons  $h = 1$  and 4. The forecast evaluation period begins in 1990Q1. We remind the reader that, with our larger data sets, MCMC methods are not feasible and, hence, all results involving the larger data sets are based on VB methods only. We use MSFEs and average log scores (i.e. averages of the log predictive distributions)<sup>8</sup> to evaluate forecast performance. To benchmark our results, we use individual AR(1)-GARCH(1,1) models for the three variables being forecast.<sup>9</sup> Both MSFEs and average log scores are particular sample realizations. To examine whether a model can forecast better than the benchmark in population, we carry out the one-sided sign test of equal predictive accuracy defined in sub-section 1.2.1 of Diebold and Mariano (1995). In the tables, \*\*\*, \*\* and \* denote rejection of the null hypothesis of equal predictive accuracy of a model and the AR(1)-GARCH(1,1) benchmark at the 1%, 5% and 10% level of significance, respectively. Rejection of the

<sup>8</sup>In this paper the average log score is used as a positively-oriented score, i.e., a model with larger average log score is deemed to have better probabilistic forecast than another model with smaller average log score.

<sup>9</sup>We use non-informative prior Bayesian methods to estimate and forecast with the AR(1)-GARCH(1,1) models. Computation is done using the MCMC algorithm of Chan and Grant (2016).

null hypothesis means that the VAR is forecasting better than the benchmark.

Tables 5, 6 and 7 present results for GDP growth, inflation and the unemployment rate, respectively. The most important point about these tables is that we were able to produce them. That is, the use of VB methods means that it is computationally feasible to carry out a large VAR forecasting exercise using models with hierarchical shrinkage priors and stochastic volatility.

Table 5: Forecasting results for Real GDP

Forecast Horizon	Homoskedastic				Heteroskedastic			
	$h = 1$	$h = 4$	$h = 1$	$h = 4$	$h = 1$	$h = 4$	$h = 1$	$h = 4$
Models	MSFE		Ave. Log Scores		MSFE		Ave. Log Scores	
AR-GARCH(1,1)	0.26	0.31	-1.23	-1.46	0.26	0.31	-1.23	-1.46
<b>100 Variables Models</b>								
normal-independent	0.23	0.28***	-1.31	-1.36	0.24	0.28***	-1.20***	-1.51
Horseshoe	0.24	0.32	-1.21	-1.63	0.27	0.44	-2.43	-4.86
LASSO	0.26	0.28**	-1.35	-1.37	0.24	0.28**	-16.35	-14.01
Adaptive LASSO	0.19*	0.27**	-1.17	-1.41	0.18	0.28*	-1.15***	-1.90
t-prior	0.20*	0.27**	-1.18	-1.44	0.19	0.28	-1.25	-2.02
SSVS	0.19	0.30	-1.16	-1.38	0.19	0.29	-1.23***	-1.97
Jeffreys'	0.20	0.28**	-1.28	-1.39	0.21**	0.28**	-1.20***	-1.62
<b>20 Variables Models</b>								
normal-independent	0.20*	0.30	-1.22	-1.38	0.19**	0.30	-3.51	-2.99
Horseshoe	0.19**	0.35	-1.15	-1.60	0.19**	0.31	-1.16***	-2.05
LASSO	0.20***	0.30	-1.22	-1.41	0.19	0.29	-1.11***	-1.66
Adaptive LASSO	0.18	0.33	-1.15	-1.55	0.18*	0.28	-1.12***	-1.89
t-prior	0.19*	0.37	-1.16	-1.65	0.19*	0.32	-1.18***	-2.24
SSVS	0.20	0.38	-1.18	-1.73	0.20**	0.33	-1.27	-2.39
Jeffreys'	0.19*	0.35	-1.15	-1.63	0.19***	0.30***	-1.14***	-1.92
<b>10 Variables Models</b>								
normal-independent	0.24	0.32	-1.27	-1.45	0.22	0.32	-1.19***	-1.76
Horseshoe	0.23	0.36	-1.25	-1.56	0.21	0.32	-1.17***	-1.81
LASSO	0.22	0.31	-1.25	-1.45	0.20	0.30**	-1.15***	-1.60
Adaptive LASSO	0.22	0.33	-1.24	-1.52	0.20	0.30	-1.14***	-1.66
t-prior	0.23	0.37	-1.26	-1.57	0.21	0.33	-1.18***	-1.86
SSVS	0.24	0.39	-1.27	-1.62	0.21	0.34	-1.21***	-2.01
Jeffreys'	0.22	0.33	-1.24	-1.53	0.20	0.30	-1.17***	-1.72

Table 6: Forecasting results for the Unemployment rate

Forecast Horizon	Homoskedastic				Heteroskedastic			
	$h = 1$	$h = 4$	$h = 1$	$h = 4$	$h = 1$	$h = 4$	$h = 1$	$h = 4$
Models	MSFE		Ave. Log Scores		MSFE		Ave. Log Scores	
AR-GARCH(1,1)	0.07	0.27	-0.34	-5.80	0.07	0.27	-0.34	-5.80
<b>100 Variables Models</b>								
normal-independent	0.55	0.66	-1.65	-1.73	0.54	0.64	-3.68	-20.71
Horseshoe	0.02***	0.14***	0.12***	-7.57	0.02***	0.24**	-1.23	-30.11
LASSO	0.02***	0.16***	0.11***	-6.69	0.02***	0.17***	-7.97	-26.97
Adaptive LASSO	0.02***	0.15***	0.13***	-7.37	0.02***	0.15***	-0.21***	-14.26
t-prior	0.02***	0.15***	0.12***	-7.43	0.02***	0.15**	-0.32***	-15.72
SSVS	0.05	0.25	-0.57	-6.95	0.08	0.35	-1.67	-13.30
Jeffreys'	0.01***	0.14***	0.14***	-5.48	0.02***	0.15***	0.04***	-10.36
<b>20 Variables Models</b>								
normal-independent	0.04	0.25	-0.35	-5.07	0.05	0.28	-3.13	-13.16
Horseshoe	0.01***	0.14***	0.18***	-5.43	0.01***	0.14***	0.16***	-12.36
LASSO	0.01***	0.15***	0.19***	-5.37	0.01***	0.14***	0.16***	-12.86
Adaptive LASSO	0.01***	0.14***	0.18***	-5.49	0.01***	0.14***	0.16***	-12.14
t-prior	0.01***	0.14***	0.17***	-5.45	0.01***	0.15***	0.15***	-12.97
SSVS	0.01***	0.14***	0.15**	-5.27	0.01***	0.14***	0.14***	-12.80
Jeffreys'	0.01***	0.14***	0.16***	-5.27	0.01***	0.14***	0.18***	-11.64
<b>10 Variables Models</b>								
normal-independent	0.02***	0.16***	0.07***	-4.94	0.02***	0.16***	0.03***	-10.62
Horseshoe	0.02***	0.16***	0.05**	-4.98	0.02***	0.15***	0.06***	-10.08
LASSO	0.02***	0.16***	0.05***	-4.84	0.02***	0.15***	0.08***	-9.64
Adaptive LASSO	0.02***	0.16***	0.05***	-4.97	0.02***	0.15***	0.06***	-9.87
t-prior	0.02***	0.17***	0.04**	-5.00	0.02***	0.16***	0.07***	-10.02
SSVS	0.02***	0.17***	0.04**	-5.10	0.02***	0.16***	0.06***	-10.60
Jeffreys'	0.02***	0.16***	0.04***	-5.05	0.02***	0.15***	0.05***	-9.69



Table 7: Forecasting results for PCE Inflation

Forecast Horizon	Homoskedastic				Heteroskedastic			
	$h = 1$	$h = 4$	$h = 1$	$h = 4$	$h = 1$	$h = 4$	$h = 1$	$h = 4$
Models	MSFE		Ave. Log Scores		MSFE		Ave. Log Scores	
AR-GARCH(1,1)	1.36	1.17	-1.97	-2.11	1.36	1.17	-1.97	-2.11
<b>100 Variables Models</b>								
normal-independent	1.17***	1.17	-2.07	-2.13	1.15***	1.18	-1.94***	-2.19
Horseshoe	1.14	1.16	-2.35	-2.77	1.13	1.38	-3.30	-7.51
LASSO	1.16***	1.17	-2.05	-2.11*	1.15***	1.19	-21.37	-25.33
Adaptive LASSO	1.25***	1.15*	-2.29	-2.36	1.19**	1.16***	-3.00	-5.55
t-prior	1.26**	1.17**	-2.24	-2.35	1.20**	1.17**	-2.92	-5.33
SSVS	1.16**	1.17	-2.22	-2.34	1.12***	1.16	-1.99	-2.24
Jeffreys'	1.23	1.19	-2.13	-2.21	1.22	1.19	-3.95	-3.05
<b>20 Variables Models</b>								
normal-independent	1.19**	1.17	-2.10	-2.14	1.16***	1.17	-2.12	-3.86
Horseshoe	1.21	1.17	-2.13	-2.20	1.20**	1.18	-1.97***	-2.20
LASSO	1.16***	1.17	-2.05	-2.11***	1.17***	1.17	-1.93***	-2.15
Adaptive LASSO	1.20	1.17	-2.12	-2.20	1.22*	1.17	-1.98	-2.22
t-prior	1.22	1.20	-2.13	-2.23	1.19	1.20	-2.01	-2.26
SSVS	1.28	1.20	-2.16	-2.25	1.28	1.19	-2.06	-2.28
Jeffreys'	1.21	1.18	-2.13	-2.22	1.20	1.18	-1.99	-2.21
<b>10 Variables Models</b>								
normal-independent	1.23	1.16	-2.12	-2.15	1.19***	1.16	-1.95***	-2.17
Horseshoe	1.24**	1.17	-2.13	-2.17	1.19***	1.15*	-1.96***	-2.19
LASSO	1.18***	1.17	-2.06	-2.11	1.18***	1.17	-1.96***	-2.16
Adaptive LASSO	1.23**	1.16	-2.13	-2.16	1.18***	1.16*	-1.96***	-2.18
t-prior	1.25**	1.17	-2.13	-2.18	1.19***	1.16*	-1.96***	-2.18
SSVS	1.27	1.17	-2.14	-2.18	1.20**	1.16**	-1.96***	-2.15
Jeffreys'	1.24	1.16	-2.13	-2.17	1.19***	1.15	-1.97***	-2.16

In general, we are finding VARs with hierarchical priors to forecast very well for the unemployment rate and, to a lesser extent and with some exceptions, for GDP growth and inflation relative to the benchmark AR(1)-GARCH(1,1) model.

For GDP, VARs and VAR-SVs with various hierarchical priors almost all produce smaller MSFEs than the benchmark for both  $h = 1$  and  $h = 4$ . In many cases the forecast improvements are statistically significant. There is evidence that models of  $n = 20$  forecast better than either larger ( $n = 100$ ) or smaller ( $n = 10$ ) models when using MSFEs. Results from average log scores are mixed. VAR-SVs of all dimensions are

typically producing better log-scores for  $h = 1$  but for  $h = 4$  the forecasts tend to be slightly worse than the benchmark. The homoskedastic VARs are forecasting slightly better than the benchmark for  $n = 20$  and  $n = 100$  but are roughly the same for the small VAR. The LASSO is forecasting very poorly for the large VAR-SV (but not for the homoskedastic VARs).

For unemployment, in most cases VAR and VAR-SV models yield smaller MSFEs than that of the benchmark model at the 1% significance level. Forecasts for  $h = 1$  measured by average log scores show a similar pattern, apart from in the cases of VAR-SVs with  $n = 100$ . The good forecasting results produced by the various shrinkage priors for VARs of various dimensions, however, are elusive when we examine the average log scores for the  $h = 4$  forecasts. In particular, for the  $h = 4$  forecasts, VAR-SV models are substantially worse than the benchmark despite the fact that the relevant MSFEs tend to be substantially better than the benchmark. This pattern (which is repeated to a lesser extent with some other priors with other variables) is due to the fact that our iterative forecasts involve simulating volatility processes  $h = 4$  periods out of sample. Occasionally this produces unduly large volatilities, especially for VAR-SVs with  $n = 100$ . It is interesting that this occurs for the more persistent variables (unemployment and inflation) rather than for the less persistent variable (GDP growth).

For inflation, the general pattern is that models with hierarchical shrinkage priors produce good point forecasts, but density forecasts for larger models are often beaten by the benchmark. The contrast between the results of MSFE and average log scores is particularly strong notable for the model with  $n = 100$ . Measured by MSFE, VAR-SVs with  $n = 100$  give better  $h = 1$  forecasts than their homoskedastic counterparts in all cases. In addition, MSFE values show that all the VAR and VAR-SV models outperform the benchmark model for  $h = 1$  forecast. MSFE results associated with the  $h = 4$  forecast horizon, however, show that benchmark model is only outperformed in a few cases. By contrast, values of average log scores show that VAR and VAR-SV models do not outperform the benchmark except for two cases: when LASSO and normal-independent priors are used to estimate VAR-SVs with  $n = 20$  and  $n = 100$ , respectively, for  $h = 1$  forecast. Comparing average log scores also indicate that VAR-SVs with  $n = 100$  tend to be the worst performing models with the LASSO forecasting particularly poorly.

A comparison of results across the different hierarchical priors indicates that most of the different approaches are leading to quite similar forecast performance. So we cannot provide a recommendation of one prior which is particularly well suited for working with large VARs. However, there are two priors which in a few cases are inferior to the rest, especially when measured by the average log score. These are the LASSO prior and, to a lesser extent, the horseshoe prior. The former produces poor forecasts of GDP when used with the VAR-SV with  $n = 100$ . The latter forecasts the unemployment rate poorly, also for the  $n = 100$  case. These are two of the simplest priors, lacking the more sophisticated prior hierarchies of the other priors which may be useful in

the high-dimensional models where VB methods would be required. Remember that the LASSO is a special case of the adaptive LASSO and involves a single global shrinkage parameter common to all coefficients in each equation. Clearly there are cases where this is too restrictive and the more flexible adaptive LASSO is to be preferred.

With regards to VAR dimension, we are finding some evidence of the benefits of working with larger VARs. For unemployment and GDP growth forecasting, there is evidence that working with  $n = 20$  leads to better forecasts than working with  $n = 10$ , in the form of smaller MSFEs, higher average log scores and even significant Diebold-Mariano test results. However, for GDP, moving to  $n = 100$  leads to a slight deterioration in forecast performance relative to  $n = 20$ . For unemployment, in some cases we are finding  $n = 100$  to forecast best and in the remainder of cases models with  $n = 20$  are best. For inflation, with the exception of the  $h = 4$  average log scores noted above, the different VAR dimensions are leading to similar forecasts.

In the previous discussion of results, we noted some cases where the results of MSFEs and average log scores were not consistent with one another. Of course, the former relate to point forecasts and the latter to density forecasts so it is always possible that they may tell different stories. VB methods are known to under-estimate posterior variances and so it is worth considering what effect this is having on our density forecasts. The following sub-section offers a more detailed investigation of VB forecast performance in the tails of distributions. Here we briefly note that we have investigated predictive variances and their impact on average log scores in our small data set using the VAR-SV and the horseshoe prior. A comparison of VB and MCMC estimates of predictive variances as well as average log scores is available in the Empirical Appendix. What we find is that, for much of the time, VB is underestimating posterior variances and, as expected, this feeds into an underestimation of predictive variances. However, the magnitude of this under-estimate is typically not large.

## 5.4 Comparing VB to MCMC Forecasts

In this sub-section, we provide a comparison of VB to MCMC forecasts using the small data set for which MCMC computation is practical. In addition to presenting MCMC-based MSFEs and average log scores, we provide evidence on the forecast performance of VB in the tails of the distribution. It is in relation to the tails and higher moments of the distribution where the normal approximation we use with our VB methods could lead to inaccuracies. It is well-known that global-local shrinkage priors can lead to posteriors which depart substantially from normality. For instance, Betancourt and Girolami (2015) document funnel shaped posteriors when working with the horseshoe prior. Of course, the normal approximation relates to the posterior, not the predictive. Particularly when stochastic volatility is added, it is possible to get very non-normal fat tailed

predictive densities even if the VAR coefficients are modelled using a normal likelihood and a normal prior (see, e.g., Carriero, Clark and Marcellino 2020). Nevertheless it is worthwhile to see how well our VB methods do in modelling the tails of predictive densities. We compare tail forecasts using the quantile score. This is the standard method of evaluating tail forecast performance. Following Gneiting and Ranjan (2011), we define the quantile score for quantile  $\tau$  as

$$QS_{\tau i,t} = (y_{it} - \mathbb{Q}_{\tau i,t}) (\tau - \mathbb{I}\{y_{it} \leq \mathbb{Q}_{\tau i,t}\}),$$

where  $\mathbb{Q}_{\tau i,t}$  is the predictive quantile of the  $i^{\text{th}}$  variable.  $\mathbb{I}\{y_{it} \leq \mathbb{Q}_{\tau i,t}\}$  has a value of 1 if the realized value is at or below the predictive quantile and 0 otherwise. We evaluate the QS in the upper and lower tails by setting  $\tau = 0.9$  and  $\tau = 0.1$ , respectively.

Tables 8, 9 and 10 contain our results for the comparison of VB and MCMC. The comparison of quantile scores can be done directly off these tables. The comparison of MSFEs and average log scores can be done by comparing the bottom panels of these tables to the VB results for the 10 variables model in Tables 5, 6 and 7, respectively.

In terms of the MSFEs and quantile scores, we are finding that MCMC and VB are producing results which, with a few exceptions, are very similar. This result holds for both tails of the predictive density. However, when looking at average log scores, there are a few cases where more substantive differences occur between MCMC and VB. Consider, for instance,  $h = 1$  forecasts of real GDP growth. When using VB we found these to be consistently better than the benchmark for all the priors. However, when using MCMC the comparable numbers in Table 5 show a forecast performance which is roughly the same and sometimes slightly worse than the benchmark.

The differences between our MSFE and average log score findings in the MCMC versus VB comparison can partly be explained by VB's tendency to under-estimate posterior variances. As noted above and documented in the Empirical Appendix for the  $n = 10$  case, this is leading VB to produce predictive variances which are slightly smaller than those produced by MCMC. In relation to models with stochastic volatility, it is worth stressing that the MCMC results use the auxiliary mixture sampler of Kim, Shephard and Chib (1998) whereas the VB results use the algorithm of Chan and Yu (2020). This is an additional reason for VB and MCMC results to be different and it is notable that the difference between VB and MCMC results is less for the homoskedastic models than for models with stochastic volatility.

The nature of the forecast metrics used also can account for some of the reason why VB and MCMC are producing similar results for quantile scores and MSFEs, but somewhat less similar results when using average log scores. The latter involve evaluating the predictive density at a realized value of a variable whereas the

former do not. Evaluating a predictive density can be very sensitive to approximation error. For example, a slight change to a point forecast will not change an MSFE much, but a slight change in an estimated predictive density can have a more substantial impact on an average log score.

The preceding discussion suggests that VB works better as a method for producing point forecasts as opposed to density forecasts. But it is also worth noting that, in several cases, VB is producing better average log scores than MCMC. For instance, VB's slight underestimation of predictive variances is actually benefitting density forecast performance for the unemployment rate for  $h = 1$ . This indicates that, at least for certain variables, that some of our shrinkage priors may not be doing enough to overcome the over-parameterization problems of large VARs. In these cases, VB's tendency to underestimate posterior variances is acting like an additional form of shrinkage.

Table 8: Comparison of VB and MCMC Forecasting Results Using the Small Data Set: Results for Real GDP Measured by Quantile Score, MSFE and Average Log Score

Forecast Horizon	Homoskedastic				Heteroskedastic			
	$h = 1$	$h = 4$	$h = 1$	$h = 4$	$h = 1$	$h = 4$	$h = 1$	$h = 4$
Models	Quantile Scores - 10%		Quantile Scores - 90%		Quantile Scores - 10%		Quantile Scores - 90%	
<b>VB</b>								
normal-independent	0.19	0.21	0.20	0.22	0.18	0.23	0.16	0.19
Horseshoe	0.18	0.22	0.20	0.24	0.17	0.22	0.15	0.19
LASSO	0.19	0.21	0.19	0.22	0.18	0.22	0.15	0.19
Adaptive LASSO	0.18	0.21	0.19	0.23	0.17	0.22	0.15	0.19
t-prior	0.18	0.22	0.20	0.24	0.17	0.22	0.15	0.19
SSVS	0.18	0.22	0.20	0.25	0.18	0.23	0.15	0.19
Jeffreys'	0.18	0.21	0.20	0.23	0.17	0.22	0.15	0.19
<b>MCMC</b>								
normal-independent	0.19	0.21	0.20	0.22	0.18	0.22	0.17	0.20
Horseshoe	0.18	0.21	0.20	0.24	0.17	0.21	0.16	0.20
LASSO	0.18	0.22	0.20	0.24	0.17	0.22	0.16	0.20
Adaptive LASSO	0.18	0.22	0.20	0.25	0.17	0.21	0.16	0.21
t-prior	0.18	0.22	0.20	0.25	0.17	0.21	0.16	0.21
SSVS	0.18	0.22	0.20	0.24	0.17	0.21	0.16	0.20
Jeffreys'	0.18	0.22	0.21	0.26	0.17	0.22	0.16	0.21
<b>MCMC - MSFE and Ave. Log Scores</b>								
	MSFE		Ave. Log Scores		MSFE		Ave. Log Scores	
normal-independent	0.24	0.32	-1.28	-1.47	0.22	0.32	-1.25	-1.80
Horseshoe	0.22	0.33	-1.26	-1.55	0.20	0.30	-1.21	-1.82
LASSO	0.23	0.35	-1.28	-1.58	0.21	0.32	-1.24	-1.89
Adaptive LASSO	0.23	0.36	-1.28	-1.61	0.21	0.33	-1.24	-1.95
t-prior	0.23	0.36	-1.28	-1.62	0.21	0.33	-1.25	-1.95
SSVS	0.22	0.35	-1.27	-1.59	0.20	0.32	-1.23	-1.92
Jeffreys'	0.25	0.40	-1.31	-1.71	0.22	0.36	-1.29	-2.14

Table 9: Comparison of VB and MCMC Forecasting Results Using the Small Data Set: Results for Unemployment Rate Measured by Quantile Score, MSFE and Average Log Score

Forecast Horizon	Homoskedastic				Heteroskedastic			
	$h = 1$	$h = 4$	$h = 1$	$h = 4$	$h = 1$	$h = 4$	$h = 1$	$h = 4$
Models	Quantile Scores - 10%		Quantile Scores - 90%		Quantile Scores - 10%		Quantile Scores - 90%	
<b>VB</b>								
normal-independent	0.04	0.09	0.06	0.19	0.05	0.10	0.05	0.19
Horseshoe	0.05	0.11	0.06	0.19	0.05	0.10	0.05	0.19
LASSO	0.05	0.10	0.06	0.19	0.05	0.10	0.05	0.19
Adaptive LASSO	0.05	0.11	0.06	0.19	0.05	0.10	0.05	0.19
t-prior	0.05	0.11	0.06	0.19	0.05	0.10	0.05	0.19
SSVS	0.05	0.11	0.06	0.19	0.05	0.11	0.05	0.19
Jeffreys'	0.05	0.11	0.06	0.19	0.05	0.10	0.05	0.19
<b>MCMC</b>								
normal-independent	0.04	0.10	0.06	0.19	0.04	0.09	0.05	0.19
Horseshoe	0.05	0.11	0.06	0.19	0.04	0.10	0.05	0.19
LASSO	0.05	0.11	0.06	0.19	0.04	0.10	0.05	0.19
Adaptive LASSO	0.05	0.12	0.06	0.19	0.05	0.10	0.05	0.19
t-prior	0.05	0.11	0.06	0.19	0.05	0.10	0.05	0.19
SSVS	0.05	0.11	0.06	0.19	0.04	0.10	0.05	0.19
Jeffreys'	0.05	0.12	0.06	0.19	0.04	0.11	0.05	0.19
<b>MCMC - MSFE and Ave. Log Scores</b>								
	MSFE		Ave. Log Scores		MSFE		Ave. Log Scores	
normal-independent	0.02	0.16	0.04	-5.11	0.02	0.16	0.03	-9.29
Horseshoe	0.02	0.16	0.02	-5.20	0.02	0.15	0.04	-9.09
LASSO	0.02	0.16	0.02	-5.24	0.02	0.15	0.05	-9.20
Adaptive LASSO	0.02	0.16	0.02	-5.30	0.02	0.15	0.03	-9.41
t-prior	0.02	0.16	0.02	-5.28	0.02	0.15	0.04	-9.35
SSVS	0.02	0.16	0.02	-5.26	0.02	0.15	0.04	-9.32
Jeffreys'	0.02	0.18	0.01	-5.57	0.02	0.17	0.02	-9.89

Table 10: Comparison of VB and MCMC Forecasting Results Using the Small Data Set: Results for PCE Inflation Measured by Quantile Score, MSFE and Average Log Score

Forecast Horizon	Homoskedastic				Heteroskedastic			
	$h = 1$	$h = 4$	$h = 1$	$h = 4$	$h = 1$	$h = 4$	$h = 1$	$h = 4$
Models	Quantile Scores - 10%		Quantile Scores - 90%		Quantile Scores - 10%		Quantile Scores - 90%	
<b>VB</b>								
normal-independent	0.37	0.40	0.38	0.36	0.38	0.42	0.34	0.37
Horseshoe	0.37	0.40	0.40	0.36	0.37	0.41	0.34	0.37
LASSO	0.39	0.40	0.37	0.36	0.40	0.42	0.32	0.38
Adaptive LASSO	0.38	0.40	0.38	0.36	0.38	0.41	0.34	0.38
t-prior	0.37	0.40	0.40	0.36	0.37	0.42	0.35	0.38
SSVS	0.37	0.40	0.40	0.36	0.38	0.41	0.35	0.37
Jeffreys'	0.37	0.40	0.39	0.36	0.38	0.41	0.35	0.38
<b>MCMC</b>								
normal-independent	0.37	0.40	0.39	0.36	0.37	0.42	0.34	0.39
Horseshoe	0.37	0.40	0.38	0.36	0.38	0.42	0.34	0.39
LASSO	0.37	0.40	0.39	0.36	0.37	0.42	0.34	0.39
Adaptive LASSO	0.36	0.40	0.39	0.37	0.38	0.42	0.35	0.39
t-prior	0.36	0.40	0.39	0.37	0.38	0.42	0.35	0.39
SSVS	0.36	0.40	0.39	0.37	0.38	0.41	0.35	0.39
Jeffreys'	0.37	0.40	0.40	0.37	0.39	0.41	0.36	0.39
<b>MCMC - MSFE and Ave. Log Scores</b>								
	MSFE		Ave. Log Scores		MSFE		Ave. Log Scores	
normal-independent	1.22	1.17	-2.13	-2.17	1.17	1.16	-1.99	-2.21
Horseshoe	1.23	1.16	-2.13	-2.17	1.17	1.16	-2.00	-2.23
LASSO	1.23	1.17	-2.14	-2.18	1.18	1.16	-2.00	-2.22
Adaptive LASSO	1.25	1.17	-2.15	-2.20	1.18	1.16	-2.01	-2.22
t-prior	1.25	1.17	-2.15	-2.20	1.18	1.16	-2.01	-2.22
SSVS	1.24	1.17	-2.14	-2.19	1.18	1.16	-2.01	-2.22
Jeffreys'	1.30	1.17	-2.18	-2.22	1.22	1.16	-2.01	-2.21

## 6 Conclusions and Further Discussion

The computational demands of a Bayesian analysis using large VARs can be very large, or even prohibitive, when MCMC methods are used; and empirically interesting versions of large VARs involving hierarchical shrinkage priors have, in the past, required use of MCMC methods. In response to this situation, we have developed VB methods for VARs with a range of hierarchical shrinkage priors with stochastic volatility.



The two important issues that require investigation when using VB methods are computational efficiency and accuracy. In our empirical work, we have established that VB methods are very computationally efficient and scalable. Estimation is very quick, even in VARs with hundreds of variables.

Our findings in terms of accuracy are more nuanced. In terms of point estimates and point forecasts, we have established that VB methods are very accurate. In terms of higher moments, it is well-known that VB methods tend to under-estimate posterior variances. In our empirical work, we have investigated the impact this has on predictive variances and log-scores and find it to be small but non-negligible. Overall, we establish that it is possible to successfully forecast in large VARs using VB methods in a manner that is impossible using MCMC.

This paper is directed at the reader interested in forecasting with large VARs with global-local shrinkage priors, possibly with stochastic volatility. Of course, with large data sets involving hundreds of variables there are other models that have been used in the past. The main competitors to large VARs are factor models (i.e. the dynamic factor model or the factor-augmented VAR). With factor models, there is less need for prior shrinkage and computationally efficient methods such as VB since the parameter vector is low-dimensional and conventional MCMC methods can be used. The question as to whether a factor model or a large VAR is to be preferred is an application-specific one. For some data sets the large VAR might be preferred and for others a factor model might be preferred. However, for large US macroeconomic data sets, papers such as Banbura et al. (2010) have found the large VAR to forecast better than the factor model. This suggests that large VARs should remain a popular tool in the macroeconomic forecaster's toolbox. This paper has established that VB is a fast and effective way of producing forecasts using them.

## References

- [1] Banbura, M., Giannone, D. and Reichlin, L. (2010). Large Bayesian Vector Auto Regressions. *Journal of Applied Econometrics*, 25, 71-92.
- [2] Banbura, M., Giannone, D. and Lenza, M. (2015). Conditional forecasts and scenario analysis with vector autoregressions for large cross-sections. *International Journal of Forecasting*, 31, 3, 739-756.
- [3] Betancourt, M. and Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79, 2-4
- [4] Blei, D., Kucukelbir, A. and J. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112, 859-877.
- [5] Bloor, C. and Matheson, T. (2010). Analysing shock transmission in a data-rich environment: a large BVAR for New Zealand. *Empirical Economics* 39, 537-558.
- [6] Carriero, A., Clark, T. and Marcellino, M. (2016). Common drifting volatility in large Bayesian VARs. *Journal of Business and Economic Statistics* 34, 375-390.
- [7] Carriero, A., Clark, T. and Marcellino, M. (2018). Measuring uncertainty and its impact on the economy. *Review of Economics and Statistics* 100, 799-815.
- [8] Carriero, A., Clark, T. and Marcellino, M. (2019). Large Bayesian Vector Autoregressions with stochastic volatility and non-conjugate priors. *Journal of Econometrics* 212, 137-154.
- [9] Carriero, A., Clark, T. and Marcellino, M. (2020). Capturing macroeconomic tail risks with Bayesian Vector Autoregressions. *Federal Reserve Bank of Cleveland Working Paper* 20-02.
- [10] Carriero, A., Kapetanios, G. and Marcellino, M. (2010). Forecasting exchange rates with a large Bayesian VAR. *International Journal of Forecasting* 25, 400-417.
- [11] Carriero, A., Kapetanios, G. and Marcellino, M. (2012). Forecasting government bond yields with large Bayesian vector autoregressions. *Journal of Banking & Finance* 36,7, 2026-2047.
- [12] Carvalho, C., Polson, N. and Scott, J. (2010). The horseshoe estimator for sparse signals. *Biometrika* 97 465-480.
- [13] Chan, J. C., and Grant, A. L. (2016). Modeling energy price dynamics: GARCH versus stochastic volatility. *Energy Economics*, 54, 182-189.

- [14] Chan, J. (2018). Large Bayesian VARs: A flexible Kronecker error covariance structure. *Journal of Business and Economic Statistics* forthcoming.
- [15] Chan, J. and Yu, X. (2020). Fact and accurate variational inference for large Bayesian VARs with stochastic volatility. Manuscript available at <https://joshuachan.org/papers/VB-SV.pdf>.
- [16] Clark, T. (2011). Real-time density forecasts from BVARs with stochastic volatility. *Journal of Business and Economic Statistics* 29, 327-341.
- [17] Dempster, A. P., Laird, N. M. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39, 1-38.
- [18] Diebold, F. and Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13, 253-263.
- [19] Dieppe, A., Legrand, R. and van Roye, B. (2016). The BEAR toolbox. *European Central Bank working paper* 1934.
- [20] Frazier, D. T., Loaiza-Maya R., and Martin, G. M. (2021). A note on the accuracy of Variational Bayes in state space models: Inference and prediction. <https://arxiv.org/abs/2106.12262>
- [21] Gefang, D. (2014). Bayesian doubly adaptive elastic-net LASSO for VAR shrinkage. *International Journal of Forecasting* 30, 1-11.
- [22] George, E., Sun, D. and Ni, S. (2008). Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics* 142, 553-580.
- [23] Ghosh, S., Yao, J. and Doshi-Velez, F. (2019). Model selection in Bayesian neural networks via Horseshoe priors. *Journal of Machine Learning Research* 20, 1-46.
- [24] Giannone, D., Lenza, M., Momferatou, D. and Onorante, L. (2014). Short-term inflation projections: a Bayesian vector autoregressive approach. *International Journal of Forecasting* 30, 635-644.
- [25] Giordano, R., Broderick, T. and Jordan, M. (2018). Covariances, robustness and Variational Bayes. *Journal of Machine Learning Research* 19, 1-49.
- [26] Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold and quantile weighted scoring rules. *Journal of Business and Economic Statistics* 29 (3), 411-422.
- [27] Hajargasht, G. and Wozniak, T. (2018). Variational Bayes inference for large vector autoregressions. Manuscript.

- [28] Jarocinski, M. and Mackowiak, B. (2017). Granger-causal-priority and choice of variables in vector autoregressions. *Review of Economics and Statistics* 99, 319-329.
- [29] Kastner, G. and Huber, F. (2021) Sparse Bayesian Vector Autoregressions in huge dimensions. *Journal of Forecasting* 39, 1142-1165.
- [30] Kim, S., Shephard, N. and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *The Review of Economic Studies* 65, 361-393.
- [31] Koop, G. (2003). *Bayesian Econometrics*, John Wiley and Sons, Chichester.
- [32] Koop, G. (2013). Forecasting with medium and large Bayesian VARs. *Journal of Applied Econometrics* 28, 177-203.
- [33] Koop, G. and Korobilis, D. (2016). Model uncertainty in panel vector autoregressive models. *European Economic Review* 81, 115-131.
- [34] Koop, G. and Korobilis, D. (2018). Variational Bayes inference in high-dimensional time-varying parameter models. Manuscript available at <https://sites.google.com/site/garykoop/research>.
- [35] Koop, G. and Korobilis, D. (2019). Forecasting with high dimensional panel VARs. *Oxford Bulletin of Economics and Statistics* 81, 937-959.
- [36] Korobilis, D. (2013). VAR forecasting using Bayesian variable selection. *Journal of Applied Econometrics* 28, 204-230.
- [37] Loaiza-Maya, R., Smith, M., Nott, D. and Danaher, P. (2021). Fast and accurate variational inference for models with many latent variables. *Journal of Econometrics*, forthcoming.
- [38] Makalic, E., & Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. arXiv preprint arXiv:1508.03884.
- [39] Ormerod, J. and M. Wand (2010). Explaining variational approximations. *American Statistician* 64, 140-153.
- [40] Park, T. and G. Casella (2008). The Bayesian LASSO. *Journal of the American Statistical Association* 103, 681-686.
- [41] You, C., Ormerod, J. and Muller, S. (2014). On variational Bayes estimation and variational information criteria for linear regression models. *Australian and New Zealand Journal of Statistics* 56, 73-87.