Variational Bayesian Inference in Large Vector Autoregressions with Hierarchical Shrinkage

Deborah Gefang*
University of Leicester

Gary Koop†
University of Strathclyde

Aubrey Poon ‡
University of Strathclyde

Abstract

Many recent papers in macroeconomics have used large Vector Autoregressions (VARs) involving a hundred 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/.

*Email: dg171@leicester.ac.uk
†Email: gary.koop@strath.ac.uk
‡Email: aubrey.poon@strath.ac.uk

1 Introduction

Recent years have seen the emergence of a literature involving large Bayesian VARs. 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), Babura, Giannone and Lenza (2015), Koop and Korobilis (2016, 2018a), Jarocinski and Mackowiak (2017), Carriero, Clark and Marcellino (2016a,b,2018) and Chan (2018). The computational methods used in these papers fall into two general categories: i) those which use Markov Chain Monte Carlo (MCMC) methods and ii) those which avoid the use of MCMC methods by using natural conjugate priors (for which analytical results are available).[1] 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: MCMC methods are much slower than analytical ones. 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. 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.[2]

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 hierarchical priors which are increasingly used in the machine learning literature to ensure parsimony, 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, 2017). In this paper we show how an alternative approach, Variational Bayes (VB), can be

[1] There are also a few papers which use approximate methods, such as the discount factor methods of Koop and Korobilis (2013).

[2] In the US, the popular FRED data set, produced by the Federal Reserve Bank of St. Louis contains well over a hundred monthly variables and well over two hundred quarterly variables.

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. They can be used in Bayesian VAR forecasting exercises involving huge VARs.

In this paper, 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-prior. 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 highly 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 results that are virtually identical and are very close to one another with other priors. We 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 Blei, Kucukelbir and McAuliffe (2017) and Ormerod and Wand (2010). Here we explain the 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 θ. VB methods approximate this posterior with another simpler density $q(\theta)$ that is as close as possible to it in a Kullback-Leibler sense. Minimizing KL can be shown to be equivalent to maximizing the Evidence Lower Bound (ELBO):

$$ELBO = E(\log p(\theta, y)) - E(\log q(\theta)). \tag{1}$$

Thus, VB involves optimizing a function (the ELBO) which is typically much faster than doing MCMC.[3]

VB requires the choice of an approximating density, $q(\theta)$, and the derivation of the ELBO for that function. Computation is particularly easy if the former is taken from the so-called mean field variational family:

$$q(\theta) = \prod_{m=1}^{M} q_m(\theta_m), \tag{2}$$

where θ_m for $m = 1, .., M$ are the blocks of parameters which make up θ.

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

$$ELBO = E(\log p(y|\theta)) + \sum_{m=1}^{M} E(\log p(\theta_m)) - \sum_{m=1}^{M} E(\log q_m(\theta_m)). \tag{3}$$

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:

$$q_m(\theta_m) = exp[E(\log p(\theta_m|y, \theta_{-m}))], \tag{4}$$

where θ_{-m} denotes all parameters except for those in θ_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 (4) and, thus, in the ELBO in (3). The ELBO also involves the likelihood and prior and is easiest to evaluate if each approximating density has a known entropy. The expectations in (4) and (3) are calculated using optimization (not posterior simulation) in a similar fashion to the expectations maximization (EM) algorithm of Dempster, Laird and Rubin (1977).

[3]In some cases, the ELBO can be difficult to calculate. In such cases, a convergence criteria based on the parameters themselves is valid. That is, iterating until the estimated parameters of the approximating densities stop changing can be done. We have found this strategy to also work well with the models used in this paper.

3 Variational Bayes Methods for VARs Using Conventional Priors

There are many 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 is a popular Bayesian VAR software package that contains a range of popular priors including the Minnesota prior, the natural conjugate prior and the independent Normal-Wishart prior. We are not directly interested in any of these priors. 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. But VB methods for this prior have been developed in Hajargasht and Wozniak (2018). All of these are conventional, subjectively-elicited, non-hierarchical priors. In this paper, our interest lies in hierarchical priors which allow for automatic shrinkage in large VARs. All of these will be hierarchical extensions of a conventional prior. Hence, we begin with a conventional VAR prior in this section. For reasons outlined below, we do not work with the independent Normal-Wishart prior, but something closely related to it.

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

$$\mathbf{A}_0\mathbf{y}_t = \mathbf{b}_0 + \mathbf{B}_1\mathbf{y}_{t-1} + \ldots + \mathbf{B}_p\mathbf{y}_{t-p} + \epsilon_t, \epsilon_t \sim N(0, \Sigma), \tag{5}$$

for $t = 1, \ldots, T$ 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, $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ and $\mathbf{A}_0$ is an $n \times n$ lower triangular matrix with ones on the diagonal.

We can rewrite (5) as

$$\mathbf{y}_t = \mathbf{X}_t\beta + \mathbf{W}_t\mathbf{a} + \epsilon_t, \tag{6}$$

where $\mathbf{X}_t = \mathrm{I}_n \otimes [1, \mathbf{y}_{t-1}^{'}, \ldots, \mathbf{y}_{t-p}^{'}]$ is an $n \times K$ matrix, $\beta = vec([\mathbf{b}_0, \mathbf{B}_1, \ldots, \mathbf{B}_p]^{'})$ 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 y_t. Equation (6) can be written in terms of n

independent equations, with the i^{th} equation being:

$$y_{i,t} = \mathbf{z}_{i,t}\theta_i + \epsilon_{i,t}, \epsilon_{i,t} \sim N(0, \sigma_i^2). \tag{7}$$

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

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 (2016a), 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 (7) 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:[4]

$$\theta_i \sim N(0, \mathbf{V}_i), \tag{8}$$

$$\sigma_i^{-2} \sim G(\underline{\nu}, \underline{s}), \tag{9}$$

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 (4). For equation i, these are

[4]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.

$$q(\theta_i) \quad \sim \quad N(\bar{\theta}_i, \bar{\mathbf{V}}_i), \tag{10}$$

$$q(\sigma_i^{-2}) \sim G(\underline{\nu} + \frac{T}{2}, \bar{s}_i), \tag{11}$$

where

$$\bar{\mathbf{V}}_i = [(\frac{\underline{\nu} + \frac{T}{2}}{\bar{s}_i})\mathbf{Z}_i'\mathbf{Z}_i + \mathbf{V}_i^{-1})]^{-1}, \tag{12}$$

$$\bar{\theta}_i = (\frac{\underline{\nu} + \frac{T}{2}}{\bar{s}_i})\bar{\mathbf{V}}_i\mathbf{Z}_i'\mathbf{y}_i, \tag{13}$$

$$\bar{s}_i = \underline{s} + \frac{1}{2}\left\|\mathbf{y}_i - \mathbf{Z}_i\bar{\theta}_i\right\|^2 + \frac{1}{2}tr(\mathbf{Z}_i'\mathbf{Z}_i\bar{\mathbf{V}}_i). \tag{14}$$

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.[5] Beginning with an intialization 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 $ELBO_i$ is given in the Technical Appendix. This algorithm is done independently for each of the $i = 1, .., n$ 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

[5]Throughout this paper, we adopt a notational convention where upper bars denote quantities which are optimized in a VB algorithm.

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:

$$\mathbf{V}_i = \text{diag}(\tau_{i,1}, \ldots, \tau_{i,k_i}). \tag{15}$$

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

$$\tau_{i,j}^{-1} \sim G(\underline{a}_0, \underline{b}_0), \quad \text{for } j = 1, \ldots, k_i. \tag{16}$$

The VB approximating densities, $q(\theta_i)$ and $q(\sigma_i^2)$ are the same as (10) and (11) since their conditional posteriors (now additionally conditional on τ_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:

$$q(\tau_{i,j}^{-1}) \sim G(\underline{a}_0 + \frac{1}{2}, \frac{\bar{\theta}_{i,j}^2 + \bar{\mathbf{V}}_i^{jj}}{2} + \underline{b}_0), \tag{17}$$

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

$$\overline{\tau_{i,j}^{-1}} = \frac{\underline{a}_0 + \frac{1}{2}}{\frac{\bar{\theta}_{i,j}^2 + \bar{\mathbf{V}}_i^{jj}}{2} + \underline{b}_0}. \tag{18}$$

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

$$\tau_{i,j} \sim \tfrac{1}{\tau_{i,j}}, \quad \text{for } j = 1, \ldots \tau_{i,k_i}. \tag{19}$$

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 (15), but allows for a different treatment of the prior shrinkage parameters, τ_i. In particular, it assumes:

$$\tau_{i,j} \sim Exp(\tfrac{\lambda_{i,j}}{2}), \quad \text{for } j = 1, \ldots, k_i \tag{20}$$

with

$$\lambda_{i,j} \sim G(\underline{a}_0, \underline{b}_0). \tag{21}$$

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 τ_i and λ_i where $\lambda_i = (\lambda_{i,1}, \ldots, \lambda_{i,k_i})'$. These are

$$q(\tau_{i,j}^{-1}) \sim iG(\sqrt{\frac{\bar{\lambda}_{i,j}}{\bar{\theta}_{i,j}^2 + \bar{\mathbf{V}}_i^{jj}}}, \bar{\lambda}_{i,j}), \tag{22}$$

where iG denotes the inverse Gaussian distribution and

$$q(\lambda_{i,j}) \sim G(\underline{a}_0 + 1, 0.5\bar{\tau}_{i,j} + \underline{b}_0). \tag{23}$$

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

$$\overline{\tau_{i,j}^{-1}} = \sqrt{\frac{\bar{\lambda}_{i,j}}{\bar{\theta}_{i,j}^2 + \bar{\mathbf{V}}_i^{jj}}}, \tag{24}$$

$$\bar{\lambda}_{i,j} = \frac{\underline{a}_0 + 1}{0.5\bar{\tau}_{i,j} + \underline{b}_0}, \tag{25}$$

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

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 (7), (8) and (9), the horseshoe prior adds the assumptions that:

$$\mathbf{V}_i = \text{diag}(\lambda_{i,1}\tau_i, \ldots, \lambda_{i,k_i}\tau_i), \tag{26}$$

where the priors for the new parameters are

$$\lambda_{i,j}^{-1}|\nu_{i,j} \sim G(\frac{1}{2}, \frac{1}{\nu_{i,j}}), \tag{27}$$

$$\tau_i^{-1}|\xi_i \sim G(\frac{1}{2}, \frac{1}{\xi_i}), \tag{28}$$

$$\nu_{i,1}^{-1}, \ldots, \nu_{i,k_i}^{-1}, \xi_i^{-1} \sim G(\frac{1}{2}, 1) \tag{29}$$

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:

$$q(\lambda_{i,j}^{-1}) \sim G(1, \overline{\nu_{i,j}^{-1}} + \frac{\bar{\theta}_{i,j}^2 + \bar{\mathbf{V}}_i^{jj}}{2}\overline{\tau_i^{-1}}), \tag{30}$$

$$q(\tau_i^{-1}) \sim G(\frac{k_i+1}{2}, \overline{\xi_i^{-1}} + \frac{1}{2}\overline{\lambda_{i,j}^{-1}}\sum_{j=1}^{k_i}(\bar{\theta}_{i,j}^2 + \bar{\mathbf{V}}_i^{jj})), \tag{31}$$

$$q(\nu_{i,j}^{-1}) \sim G(1, 1 + \overline{\lambda_{i,j}^{-1}}) \tag{32}$$

and

$$q(\xi_i^{-1}) \sim G(1, 1 + \overline{\tau_i^{-1}}). \tag{33}$$

The terms which are updated in the VB iterations are (13), (14),

$$\overline{\lambda_{i,j}^{-1}} = \frac{1}{\overline{\nu_{i,j}^{-1}} + \overline{\tau_i^{-1}} \frac{\bar{\theta}_{i,j}^2 + \bar{\mathbf{V}}_i^{jj}}{2}}, \tag{34}$$

$$\overline{\tau_i^{-1}} = \frac{k_i + 1}{2\overline{\xi_i^{-1}} + [\overline{\lambda_{i,j}^{-1}} \sum_{j=1}^{k_i} (\bar{\theta}_{i,j}^2 + \bar{\mathbf{V}}_i^{jj})]}, \tag{35}$$

$$\overline{\nu_{i,j}^{-1}} = 1/(1 + \overline{\lambda_{i,j}^{-1}}), \tag{36}$$

$$\overline{\xi_i^{-1}} = 1/(1 + \overline{\tau_i^{-1}}). \tag{37}$$

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

4.4 SSVS

George, Sun and Ni (2008) develop MCMC methods for the VAR with the SSVS prior. The SSVS prior assumes that $\mathbf{V}_i = \text{diag}(v_{i,1}, \ldots, v_{i,k_i})$ and

$$v_{i,j} = \begin{cases} \underline{\kappa}_{i,j,0} & \text{if } \gamma_{i,j} = 0 \\ \underline{\kappa}_{i,j,1} & \text{if } \gamma_{i,j} = 1 \end{cases} \tag{38}$$

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^{th} coefficient in the i^{th} equation towards zero is used. The prior for $\gamma_i = (\gamma_{i,1}, \ldots, \gamma_{i,k_i})$ follows a Bernoulli distribution

$$P(\gamma_{i,j} = 1) = \underline{\pi}_{i,j}, \tag{39}$$

with

$$P(\gamma_{i,j} = 0) = 1 - \underline{\pi}_{i,j}. \tag{40}$$

The VB approximating densities for θ_i and σ_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 γ_i is

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

where

$$\bar{\pi}_{i,j} = \frac{\frac{1}{\underline{\kappa}_{i,j,1}} \exp(-\frac{\bar{\theta}^2_{i,j}+\mathbf{V}_i^{jj}}{2\underline{\kappa}^2_{i,j,1}})\underline{\pi}_{i,j}}{\frac{1}{\underline{\kappa}_{i,j,1}} \exp(-\frac{\bar{\theta}^2_{i,j}+\mathbf{V}_i^{jj}}{2\underline{\kappa}^2_{i,j,1}})\underline{\pi}_{i,j} + \frac{1}{\underline{\kappa}_{i,j,0}} \exp(-\frac{\bar{\theta}^2_{i,j}+\mathbf{V}_i^{jj}}{2\underline{\kappa}^2_{i,j,0}})(1-\underline{\pi}_{i,j})}. \tag{41}$$

Finally, we have

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

where

$$\bar{v}_{i,j} = \bar{\pi}_{i,j}\underline{\kappa}_{i,j,1} + (1-\bar{\pi}_{i,j,j})\underline{\kappa}_{i,j,0}. \tag{42}$$

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 an approximate 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

$$h_{i,t} = h_{i,t-1} + \zeta_{i,t}, \quad \zeta_{i,t} \sim N(0, \sigma^2_{h_i}), \tag{43}$$

$$\sigma^{-2}_{h_i} \sim G(\underline{a}_1, \underline{b}_1). \tag{44}$$

$$h_{i,0} \sim N(0, \underline{V}_{i,h}), \tag{45}$$

where the initial conditions $h_{i,0}$ are treated as parameters to be estimated. Chan and Eisenstat (2018)

and Chan and Hsiao (2014) provide the conditional posteriors for this model. We will not reproduce them here but note that these papers use MCMC methods involving Kim, Shephard and Chib (1998)'s auxiliary mixture sampler. The latter involves transforming the dependent variable in equation i as $y_{i,t}^* = \log(y_{i,t} - \mathbf{z}_{i,t}\theta_i)^2$ and approximating the error in the transformed equation by a mixture of seven normal distributions. This is a highly accurate approximation but the resulting MCMC algorithm is computationally slow. Accordingly, similar to Koop and Korobilis (2018b), we approximate this error by a single normal distribution. This allows for simple VB estimation of the volatilities involving only the normal distribution and, as demonstrated in Appendix B.2 of Koop and Korobilis (2018b), the approximation is a good one, except in the tails of the distribution. In this paper, instead of using the normal approximation of Koop and Korobilis (2018b), we use an alternative approximation taken from Chan and Eisenstat (2018) which we have found to work better and which should further improve the accuracy of the approximation. In macroeconomic applications such as ours, such an approximation should suffice and is definitely preferable to other simple estimators that have been used in this literature (e.g. the exponential weighted moving average). The VB estimator of stochastic volatility models of Tran, Nott and Kohn (2016) would be more accurate than our approach, but also much more computationally demanding and, thus, not feasible in our large VAR context.

The optimal VB approximating density for θ_i is given in (10) if we replace the homoskedastic error covariance matrix by the time-varying one.

The new approximating densities that arise when we add stochastic volatility are:

$$q(\boldsymbol{h}_i) \sim N(\bar{\mathbf{h}}_i, \bar{\mathbf{V}}_{h_i}), \tag{46}$$

$$q(h_{i,0}) \sim N(\bar{h}_{i,0}, \bar{V}_{h_{i,0}}), \tag{47}$$

and

$$q(\sigma_{h_i}^{-2}) \sim G(\underline{a}_1 + \frac{T}{2}, \bar{s}_{h_i}). \tag{48}$$

The arguments in these densities involve the errors $\hat{e}_i = (\mathbf{y}_i - \mathbf{z}_i\bar{\theta}_i)$ and use notation where

$$s_i = \frac{\bar{s}_{h_i}}{\underline{a}_1 + \frac{T}{2}} \tag{49}$$

and

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \ddots & 0 \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix}. \tag{50}$$

The VB algorithm, as it relates to stochastic volatility for the i^{th} equation then proceeds through the following steps:

1. Set $\bar{h}_{0,i} = \frac{(\underline{V}_{i,h}^{-1} + 1/s_i)^{-1}}{\frac{\bar{h}_{i,1}}{s_i}}$.
2. Input $\hat{e}_i$ and $\bar{h}_{0,i}$ into the M-Step of Algorithm 4 of Chan and Eisenstat (2018) which will produce updates for $\bar{\boldsymbol{h}}_i$ and $\bar{\mathbf{V}}_{h_i}$.
3. Using the updated $\bar{\boldsymbol{h}}_i$ and $\bar{\mathbf{V}}_{h_i}$, compute $\bar{s}_{h_i} = \underline{b}_1 + \frac{1}{2}[(\mathbf{H}\bar{\mathbf{h}}_i - \alpha)'(\mathbf{H}\bar{\mathbf{h}}_i - \alpha) + tr(\mathbf{H}\bar{\mathbf{V}}_{h_i}\mathbf{H}')]$ where $\alpha = (\bar{h}_{i,0}, 0, \ldots, 0)'$.
4. Compute the ELBO given in the Technical Appendix.
5. Repeat steps 1-4, until the change in the ELBO is very small.

A word is in order about our VB approximating density, $q(\boldsymbol{h}_i)$. The approximating density for all the model parameters falls in the mean field variational family, and thus our VB algorithm is a valid one. The fact that we are using a normal approximation for $q(\boldsymbol{h}_i)$ means, however, we are not working with the *optimal* VB density. Nevertheless, as we demonstrate below, the approximation is a good one. The M-step of Algorithm 4 of Chan and Eisenstat (2018) is a generic algorithm for obtaining the posterior mode of the volatilities in a stochastic volatility process (we use this posterior mode as $\bar{\mathbf{h}}_i$). It is an optimization algorithm. Details are provided in Appendix B.2 of Chan and Eisenstat (2018) but key points worth noting are that it involves two Hessian terms which are available in closed form and these terms can be used to obtain $\bar{\mathbf{V}}_{h_i}$. The optimization is done using the the Newton-Raphson method. In practice, we find that using this algorithm, which involves finding the mode of $q(\boldsymbol{h}_i)$, to work quickly and efficiently. In earlier versions of this paper, we used a different mean-based algorithm and occasionally found it to run into singularity problems, particularly in very high dimensional models. These problems do not occur with the present algorithm.

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. As we shall see in our forecasting exercise, Jeffreys' prior performs poorly. This inadequate shrinkage suggests that it is unsuitable for use in very large models. The horseshoe prior, too, involves no hyperparameters. We have found that it, too, forecasts poorly in very high dimensional models. We have found the reason for this to be that the prior for τ_i^{-1} given in (28) allocates too much prior probability to non-sparse regions of the parameter space. However, we have found that simply fixing τ_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

$$\bar{\tau}_i = \frac{1}{K+m}, \tag{51}$$

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 2018Q1 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 and medium data sets, 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 largely 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: 1) How much faster are VB methods than MCMC methods? and 2) How scaleable are VB methods? These are addressed in the following sub-section. The subsequent sub-section addresses the question: How accurate are VB methods? It does so by comparing VB (which is an approximate method) to MCMC (which, if a sufficient number of replications is taken, can be expected to be highly accurate) results. The final 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 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}.

Table 1: Computation time (in seconds)

	10 variables		**20 variables**		**100 variables**	
Homoskedastic						
Model	MCMC	VB	MCMC	VB	MCMC	VB
Normal Independent	19.6	0.3	53.8	0.3	1383.4	3.0
Horseshoe	49.0	1.0	103.6	1.7	1719.6	15.4
Lasso	260.0	0.8	455.2	1.3	2821.0	14.9
Adaptive Lasso	279.4	0.7	158.8	2.4	2783.8	4.0
t-priors	28.6	0.3	73.4	0.4	1510.6	3.0
SSVS	39.4	0.5	103.4	0.6	2343.0	2.0
Jeffreys'	39.8	0.5	113.0	0.9	1764.2	38.6
Heteroskedastic						
Normal Independent	131.4	4.6	315.0	15.2	3383.8	65.6
Horseshoe	159.6	13.6	346.6	19.4	3873.6	86.5
Lasso	358.0	7.1	682.6	23.2	4808.2	120.2
Adaptive Lasso	366.6	7.9	681.8	13.8	4316.8	89.4
t-priors	157.8	6.8	323.8	17.6	3470.2	85.6
SSVS	143.2	3.1	352.2	6.0	4360.8	48.4
Jeffreys'	145.0	6.7	333.4	12.2	3758.2	71.9

Note first that MCMC methods are very slow with the large data set. Even when using our equation-by-equation methods which, as noted by Carriero, Clark and Marcellino (2016), greatly speed up computation, 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.

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. The VAR without a hiearchical prior will tend to have faster computation since it has fewer parameters to estimate. But the addition of any of our hierchical priors does not lead to large increases in the computational

burden. Similarly, the inclusion of stochastic volatility will inevitably slow down computation. But this slowdown is relatively small for both VB and MCMC indicating our approximate method for including stochastic volatility is computationally fast.

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. Instead, for $n = 10$ we present Tables 2, 3 and 4 which summarize our findings. In particular, they produces summary statistics across all VAR coefficients or across all error covariance terms (i.e. the vector of parameters we call $\mathbf{a}$) of the absolute value of the difference between the MCMC and VB posterior means. For the homoskedastic version of each model we also present findings for the error variances. In the Empirical Appendix, figures are presented which plot VB and MCMC estimates (posterior means) for each individual VAR coefficient. It can be seen that they tend to be virtually the same.

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 0.02. 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.

The Empirical Appendix contains graphs which plot posterior means of impulse responses and (where relevant) stochastic volatilities for VB and MCMC approaches. These, too, in most cases are virtually identical and in the few cases where the two lines do not lie on top of one another tend to be quite similar. We highlight the fact that in the stochastic volatility figures, the MCMC estimates are based on a mixture of 7 Normal distributions whereas, as outlined in Section 4.5, the VB estimates use only one Normal distribution. Thus, our findings suggest that the approximation error inherent in the use of VB and that inherent in the use of one Normal distribution are both small.

Table 2: Absolute Value of Deviations between MCMC and VB - VAR coefficients

	Homoskedastic				Heteroskedastic			
Model	Median	10th Percentile	90th Percentile	Max	Median	10th Percentile	90th Percentile	Max
Normal Independent	0.00	0.00	0.00	0.01	0.01	0.00	0.03	0.23
Horseshoe	0.00	0.00	0.03	0.17	0.01	0.00	0.04	0.45
Lasso	0.01	0.00	0.03	0.22	0.01	0.00	0.03	0.14
Adaptive Lasso	0.00	0.00	0.01	0.06	0.01	0.00	0.03	0.34
t-priors	0.00	0.00	0.00	0.02	0.01	0.00	0.03	0.24
SSVS	0.00	0.00	0.04	0.70	0.01	0.00	0.05	0.68
Jeffreys'	0.00	0.00	0.13	0.72	0.04	0.00	0.15	1.08

Table 3: Absolute Value of Deviations between MCMC and VB - Covariance terms

	Homoskedastic				Heteroskedastic			
Model	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	0.13
Horseshoe	0.01	0.00	0.03	0.06	0.01	0.00	0.05	0.09
Lasso	0.01	0.00	0.04	0.08	0.02	0.00	0.04	0.09
Adaptive Lasso	0.00	0.00	0.01	0.01	0.01	0.00	0.04	0.12
t-priors	0.00	0.00	0.00	0.00	0.01	0.00	0.04	0.05
SSVS	0.00	0.00	0.04	0.10	0.01	0.00	0.05	0.12
Jeffreys'	0.02	0.00	0.14	0.20	0.03	0.01	0.18	0.32

Table 4: Absolute Value of Deviations between MCMC and VB - Error Variances

	Homoskedastic				Heteroskedastic			
Model	Median	10th Percentile	90th Percentile	Max	Median	10th Percentile	90th Percentile	Max
Normal Independent	0.00	0.00	0.00	0.00	-	-	-	-
Horseshoe	0.02	0.00	0.03	0.03	-	-	-	-
Lasso	0.00	0.00	0.00	0.00	-	-	-	-
Adaptive Lasso	0.00	0.00	0.01	0.01	-	-	-	-
t-priors	0.00	0.00	0.00	0.00	-	-	-	-
SSVS	0.00	0.00	0.00	0.02	-	-	-	-
Jeffreys'	0.06	0.01	0.14	0.16	-	-	-	-

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 this many variables, MCMC methods are not feasible and, hence, all results are based on VB methods. We use Mean Squared Forecast Errors (MSFEs) and log scores (i.e. averages of the log predictive densities) to evaluate forecast performance. We also present results for an AR(1) benchmark

and carry out the sign test of equal predictive accuracy of Diebold and Mariano (1995) against this benchmark. In the tables, ***, ** and * denote rejection of the null hypothesis of equal predictive accuracy of a model and the benchmark at the 1%, 5% and 10% level of significance, respectively.

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.

In general, we are finding VAR methods with hierarchical priors to forecast very well for GDP growth and the unemployment rate for both $h = 1$ and $h = 4$. Forecasts from the various hierarchical priors and VAR dimensions are almost always beating the AR(1) benchmark, usually by a large, statistically significant, amount. For inflation, which is often modelled using very parsimonious models, our results are less strong. However, if log-scores are used as the forecasting metric, then inflation forecasting results from the ten variable models with stochastic volatility are very good. This indicates the importance of the inclusion of stochastic volatility to get the correct modelling of the dispersion of the predictive density.

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 are inferior to the rest. These are Jeffreys' prior and the Lasso prior. For each of these there are a few cases where they forecast poorly, particularly in the $n = 100$ model. For instance, for Jeffreys' prior, forecasts of unemployment using the $n = 100$ are very poor. This prior clearly does not induce enough shrinkage in the large VAR. Note that, unlike the t-prior (of which it is a special case), Jeffreys' prior does not have a prior hyperparameter which can be used to provide an increasing amount of shrinkage as the VAR dimension increases. We have found incorporating such a property into a hierarchical shrinkage prior is important in obtaining good forecasts. The Lasso forecasts better than the Jeffreys' prior, but still forecasts poorly in some cases (e.g. GDP growth forecasts in the $n = 100$ case). 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 GDP forecasting, there is strong evidence that working with $n = 20$ leads to better

forecasts than working with $n = 10$ or an AR(1) model. However, 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 models with $n = 20$ are best. For inflation, with a single exception, smaller more parsimonious models are chosen. The single exception occurs for $h = 1$ where log-scores indicate that the 100 dimensional VAR with Adaptive Lasso prior is forecasting best.

For GDP growth and inflation we are finding strong evidence that stochastic volatility is present. That is, a comparison of models which are identical except that one contains stochastic volatility and the other does not reveals the former to have a better forecast metric.

Table 5: Forecasting Results for Real GDP

	Homoskedastic				Heteroskedastic			
Forecast Horizon	$h = 1$	$h = 4$	$h = 1$	$h = 4$	$h = 1$	$h = 4$	$h = 1$	$h = 4$
Model	MSFE		Log-scores		MSFE		Log-scores	
AR(1)	1.52	1.69	-1.40	-1.43	1.52	1.69	-1.40	-1.43
	100 variables model							
Normal Independent	1.00**	1.28***	-1.18***	-1.28	0.58***	0.90*******	-2.31	-1.67
Horseshoe	1.08**	1.41***	-1.21***	-1.32	0.72*******	**0.93*****	-1.01*******	**-1.17**
Lasso	1.59	1.62***	-1.42	-1.43	0.96*******	1.01*******	-1.14	-1.24
Adaptive Lasso	1.07**	1.39***	-1.22***	-1.31	0.70*******	1.78	-1.01*******	-1.22
t-prior	1.06**	1.46***	-1.21***	-1.33	0.70*******	1.19	-1.01*******	-1.24
SSVS	1.00**	1.29***	-1.18***	-1.28	0.57***	0.92***	-1.54*	-1.59
Jeffreys'	1.43**	1.58***	-1.37***	-1.40	0.83*******	1.00*******	-1.09	-1.22
	20 variables model							
Normal Independent	0.73**	1.35***	-1.02***	-1.25	**0.49*****	1.01***	**-0.83*****	-1.23
Horseshoe	**0.71****	1.30***	**-1.01*****	-1.24	0.50***	0.94***	**-0.83*****	-1.19
Lasso	0.74**	1.25***	-1.03***	**-1.23**	0.52***	**0.93*****	-0.85***	-1.18
Adaptive Lasso	0.72**	1.28***	-1.02***	-1.24	0.51***	0.94***	-0.85***	-1.19
t-prior	0.72**	1.29***	**-1.01*****	-1.24	0.50 ***	**0.93*****	-0.84***	-1.20
SSVS	0.72**	1.30***	**-1.01*****	-1.25	**0.49*****	1.06	-0.83***	-1.21
Jeffreys'	0.72**	1.32***	**-1.01*****	-1.24	**0.49*****	0.96***	**-0.83*****	-1.20
	10 variables model							
Normal Independent	1.22**	1.65	-1.28***	-1.38**	0.78***	1.07***	-1.06	-1.24
Horseshoe	1.20**	1.50*	-1.28***	-1.35***	0.77***	1.00***	-1.05**	-1.24
Lasso	1.21***	1.42***	-1.28***	-1.34	0.77***	0.96***	-1.05**	-1.22
Adaptive Lasso	1.19**	1.47**	-1.27***	-1.35*	0.76***	1.00***	-1.05**	-1.25
t-prior	1.19**	1.48***	-1.27***	-1.34***	0.77***	1.01***	-1.05	-1.26
SSVS	1.21**	1.57	-1.28***	-1.37***	0.77***	1.04***	-1.06	-1.27
Jeffreys'	1.19**	1.51*	-1.27***	-1.35***	0.76***	1.01***	-1.05***	-1.23

Table 6: Forecasting Results for Unemployment

	Homoskedastic				Heteroskedastic			
Forecast Horizon	$h=1$	$h=4$	$h=1$	$h=4$	$h=1$	$h=4$	$h=1$	$h=4$
Model	MSFE		Log-scores		MSFE		Log-scores	
AR(1)	1.01	1.73	-1.18	-1.39	1.01	1.73	-1.18	-1.39
	100 variables model							
Normal Independent	1.08	1.00***	-1.83	-1.41	1.39	1.01*	-3.60	**-1.33**
Horseshoe	1.21	1.02***	-1.75	-1.42	0.78	**0.89*** **	-1.59	-1.77
Lasso	0.89	1.04***	-1.35	-1.46	0.72	0.90***	-1.70	-1.84
Adaptive Lasso	0.89	0.93***	-1.82	-1.69	0.77	1.26	-1.36	-1.63
t-prior	0.88	0.97***	-1.85	-1.65	0.72	0.93***	-1.49	-1.77
SSVS	1.08	1.00***	-1.83	-1.43	1.38	1.06*	-3.79	-1.59
Jeffreys'	1.20	1.29***	-1.52	-1.56	1.29	1.43	-1.46	-1.65
	20 variables model							
Normal Independent	0.59**	1.52	-0.92**	-1.32	0.52**	1.09***	-0.98	-1.49
Horseshoe	0.54**	1.37*	-0.88**	-1.34	**0.45****	1.05**	**-0.85**	-1.58
Lasso	0.52***	1.11***	-0.86***	-1.32	0.48***	1.00***	-0.93	-1.73
Adaptive Lasso	**0.51****	1.24***	-0.85***	-1.36	0.48***	1.00**	-0.91	-1.60
t-prior	**0.51****	1.29**	**-0.84*****	-1.30	0.48**	1.00**	-0.92	-1.46
SSVS	0.58**	1.45	-0.91**	-1.29	0.50**	1.06**	-0.94	-1.45
Jeffreys'	0.57**	1.51	-0.90***	-1.28	0.48***	1.08*	-0.92	-1.52
	10 variables model							
Normal Independent	0.82***	1.36**	-1.08	-1.30	0.69***	1.15***	-1.03	-1.44
Horseshoe	0.74***	1.16***	-1.03	-1.33	0.65***	1.06***	-1.04	-1.54
Lasso	0.73***	1.12***	-1.02	-1.31	0.65***	1.03***	-1.07	-1.60
Adaptive Lasso	0.74***	1.18***	-1.03	-1.31	0.65***	1.07***	-1.04	-1.54
t-prior	0.75***	1.19***	-1.04	**-1.29**	0.65***	1.08***	-1.03	-1.48
SSVS	0.79***	1.28**	-1.06	-1.32	0.67***	1.12***	-1.03	-1.45
Jeffreys'	0.74***	1.15***	-1.03	**-1.29**	0.66***	1.08***	-1.07	-1.54

Table 7: Forecasting Results for PCE Inflation

	Homoskedastic				Heteroskedastic			
Forecast Horizon	$h=1$	$h=4$	$h=1$	$h=4$	$h=1$	$h=4$	$h=1$	$h=4$
Model	MSFE		Log-scores		MSFE		Log-scores	
AR(1)	**2.02**	**2.01**	-1.53	**-1.55**	**2.02**	**2.01**	-1.53	-1.55
	100 variables model							
Normal Independent	2.09	2.40	-1.56	-1.56	2.67	3.26	-1.73	-1.53
Horseshoe	2.20	3.23	-1.60	-1.66	3.30	2.41	-1.77	-1.52
Lasso	2.11	2.06	-1.56	-1.58	2.44	3.30	-1.64	-1.60
Adaptive Lasso	2.57	3.11	**-1.52**	-1.68	3.66	4.32	-1.72	-1.74
t-prior	3.74	5.64	-1.75	-1.89	3.14	3.93	-1.67	-1.70
SSVS	2.11	2.41	-1.56	-1.57	2.39	2.92	-1.86	-1.60
Jeffreys'	6.44	8.95	-2.08	-2.15	10.86	10.97	-2.11	-2.17
	20 variables model							
Normal Independent	2.58	2.70	-1.70	-1.62	2.60	2.64	-1.57	-1.55
Horseshoe	2.89	2.78	-1.75	-1.64	2.76	2.51	-1.59	-1.55
Lasso	2.66	2.35	-1.69	-1.59	2.63	2.44	-1.55	-1.51
Adaptive Lasso	2.76	2.62	-1.72	-1.62	2.63	2.49	-1.56	-1.55
t-prior	2.83	2.76	-1.73	-1.63	2.61	2.51	-1.55	-1.55
SSVS	2.57	2.67	-1.70	-1.62	2.61	2.61	-1.55	-1.53
Jeffreys'	2.87	3.06	-1.75	-1.66	2.68	2.55	-1.62	-1.57
	10 variables model							
Normal Independent	2.16	2.13	-1.58	-1.56	2.41	2.38	-1.44	**-1.49**
Horseshoe	2.14	2.07	-1.58	-1.57	2.33	2.31	-1.43	**-1.49**
Lasso	2.09	2.04	-1.54	-1.57	2.18	2.27	**-1.41**	-1.52
Adaptive Lasso	2.12	2.05	-1.56	-1.56	2.33	2.35	-1.42	-1.51
t-prior	2.12	2.06	-1.57	-1.56	2.33	2.32	-1.43	-1.50
SSVS	2.15	2.11	-1.56	-1.57	2.41	2.37	-1.44	-1.49
Jeffreys'	2.15	2.08	-1.58	-1.56	2.38	2.38	-1.44	-1.51

6 Conclusions

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 volatiility. In our empirical work, we have established that VB methods work well. That is, they are computationally efficient and scaleable. Estimation is very quick, even in VARs with hundreds of variables. Furthermore, they are accurate in the sense that they give very similar results to MCMC. Finally, our forecasting exercise suggests that hierarchical shrinkage and stochastic volatility, features that are rarely considered in the existing large VAR literature, are both useful additions to the

forecaster's toolbox.

References

[1] Blei, D., Kucukelbir, A. and J. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112, 859-877.

[2] 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.

[3] Banbura, M., Giannone, D. and Reichlin, L. (2010). Large Bayesian Vector Auto Regressions. *Journal of Applied Econometrics*, 25, 71-92.

[4] 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.

[5] Carriero, A., Clark, T. and Marcellino, M. (2016a). Common drifting volatility in large Bayesian VARs. *Journal of Business and Economic Statistics* 34, 375-390.

[6] Carriero, A., Clark, T. and Marcellino, M. (2016b). Large Vector Autoregressions with stochastic volatility and flexible priors. *Federal Reserve Bank of Cleveland Working Paper* no. 16-17.

[7] Carriero, A., Clark, T. and Marcellino, M. (2018). Measuring uncertainty and its impact on the economy. *Review of Economics and Statistics* forthcoming.

[8] Carriero, A., Kapetanios, G. and Marcellino, M. (2010). Forecasting exchange rates with a large Bayesian VAR. *International Journal of Forecasting* 25, 400-417.

[9] 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.

[10] Carvalho, C., Polson, N. and Scott, J. (2010). The horseshoe estimator for sparse signals. *Biometrika* 97 465-480.

[11] Chan, J. (2018). Large Bayesian VARs: A flexible Kronecker error covariance structure. *Journal of Business and Economic Statistics* forthcoming.

[12] Chan, J. and Eisenstat, E. (2018). Bayesian model comparison for time-varying parameter VARs with stochastic volatility. *Journal of Applied Econometrics* 33, 509-532.

[13] Chan, J. and Hsiao, C. (2014). Estimation of stochastic volatility models with heavy tails and serial dependence. In: I. Jeliazkov and X.-S. Yang (Eds.), *Bayesian Inference in the Social Sciences*, pages 159-180, John Wiley and Sons, Hoboken, New Jersey.

[14] Clark, T. (2011). Real-time density forecasts from BVARs with stochastic volatility. *Journal of Business and Economic Statistics* 29, 327-341.

[15] 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.

[16] Diebold, F. and Mariano, R. (1995). Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253-263.

[17] Dieppe, A., Legrand, R. and van Roye, B. (2016). The BEAR toolbox. *European Central Bank working paper* 1934.

[18] Gefang, D. (2014). Bayesian doubly adaptive elastic-net Lasso for VAR shrinkage. *International Journal of Forecasting* 30, 1-11.

[19] George, E., Sun, D. and Ni, S. (2008). Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics* 142, 553-580.

[20] 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.

[21] Hajargasht, G. and T. Wozniak (2018). Variational Bayes inference for large vector autoregressions. Manuscript.

[22] Jarocinski, M. and Mackowiak, B. (2017). Granger-causal-priority and choice of variables in vector autoregressions. *Review of Economics and Statistics* 99, 319-329.

[23] Kastner. G. and Huber, F. (2017) Sparse Bayesian Vector Autoregressions in huge dimensions. Manuscript available at https://arxiv.org/abs/1704.03239.

[24] 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.

[25] Koop, G. (2003). *Bayesian Econometrics*, John Wiley and Sons, Chichester.

[26] Koop, G. (2013). Forecasting with medium and large Bayesian VARs. *Journal of Applied Econometrics* 28, 177-203.

[27] Koop, G. and Korobilis, D. (2013). Large time-varying parameter VARs. *Journal of Econometrics* 177, 185-198.

[28] Koop, G. and Korobilis, D. (2016). Model uncertainty in panel vector autoregressive models. *European Economic Review* 81, 115-131.

[29] Koop, G. and Korobilis, D. (2018a). Forecasting with high dimensional panel VARs. *Oxford Bulletin of Economics and Statistics* forthcoming.

[30] Koop, G. and Korobilis, D. (2018b). Variational Bayes inference in high-dimensional time-varying parameter models. Manuscript available at https://sites.google.com/site/garykoop/research.

[31] Korobilis, D. (2013). VAR forecasting using Bayesian variable selection. *Journal of Applied Econometrics* 28, 204-230.

[32] Makalic, E., & Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. arXiv preprint arXiv:1508.03884.

[33] Ormerod, J. and M. Wand (2010). Explaining variational approximations. *American Statistician* 64, 140-153.

[34] Park, T. and G. Casella (2008). The Bayesian Lasso. *Journal of the American Statistical Association* 103, 681-686.

[35] Tran, M., Nott, D. and Kohn, R. (2016). Variational Bayes with intractable likelihood. Manuscript available at https://arxiv.org/abs/1503.08621v2.