Bayesian Model Comparison for Time-Varying Parameter VARs with Stochastic Volatility

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

Eric Eisenstat
School of Economics,
University of Queensland

October 2016

Abstract

We develop importance sampling methods for computing two popular Bayesian model comparison criteria, namely, the marginal likelihood and deviance information criterion (DIC) for TVP-VARs with stochastic volatility. The proposed estimators are based on the integrated likelihood, which are substantially more reliable than alternatives. Using US and Australian data, we find overwhelming support for the TVP-VAR with stochastic volatility compared to a conventional constant coefficients VAR with homoscedastic innovations. Most of the gains, however, appear to have come from allowing for stochastic volatility rather than time variation in the VAR coefficients or contemporaneous relationships. Indeed, according to both criteria, a constant coefficients VAR with stochastic volatility receives similar support as the more general model with time-varying parameters.

Keywords: state space, marginal likelihood, deviance information criterion, Great Moderation

JEL classifications: C11, C52, E32, E52

*Joshua Chan would like to acknowledge financial support by the Australian Research Council via a Discovery Early Career Researcher Award (DE150100795). We would also like to thank Angelia Grant for her excellent research assistance.

1 Introduction

Since the seminal work of Cogley and Sargent (2001, 2005) and Primiceri (2005), the time-varying parameter vector autoregression (TVP-VAR) with stochastic volatility has become a benchmark for analyzing the evolving inter-relationships between multiple macroeconomic variables.[1] In addition, models with time-varying parameters and stochastic volatility are often found to forecast better than their constant-coefficient counterparts, as demonstrated in papers such as Clark (2011), D'Agostino, Gambetti, and Giannone (2013) and Clark and Ravazzolo (2014). Despite the empirical success of these flexible time-varying models, an emerging literature has expressed concerns about their potential over-parameterization.[2] This new development highlights the need for model comparison techniques. For instance, one might wish to compare a general TVP-VAR with stochastic volatility to various restricted models to see if all forms of time variation are required.

Model comparison techniques for these TVP-VARs are also needed when one wishes to test competing hypotheses. For example, there is an ongoing debate about the causes of the Great Moderation—the widespread, historically unprecedented stability in most developed economies between early 1980s and mid 2000s. A number of authors, including Cogley and Sargent (2001) and Boivin and Giannoni (2006), have argued that the monetary policy regime is an important factor in explaining the Great Moderation. Under this explanation, one would expect that the monetary policy transmission mechanism would be markedly different during the Great Moderation compared to earlier decades. This in turn would manifest itself in changes in the reduced-form VAR coefficients.

On the other hand, other researchers such as Sims and Zha (2006) and Benati (2008) have emphasized that the volatility of exogenous shocks has changed over time, and this alone may be sufficient to explain the Great Moderation. To assess which of these two explanations are more empirically relevant, one direct approach is to perform a model comparison exercise—e.g., comparing a TVP-VAR with constant variance against a constant coefficients VAR with stochastic volatility—to see which model is more favored by the data.

Given these considerations, we develop importance sampling methods for computing two popular Bayesian model comparison criteria, namely, the marginal likelihood and deviance information criterion (DIC). The former evaluates how likely it is for the observed data to have occurred given the model, whereas the latter trades off between model fit and model complexity. There are earlier attempts to formally compare these TVP-VARs. For instance, Koop et al. (2009) compute the marginal likelihood using the harmonic mean of a conditional likelihood—the conditional density of the data given the log-volatilities but marginal of the time-varying parameters. However, recent work has shown that this approach can be extremely inaccurate. For example, Chan and Grant (2015) find that the

[1] For example, recent papers include Benati (2008), Koop, Leon-Gonzalez, and Strachan (2009),Koop and Korobilis (2013) and Liu and Morley (2014).

[2] See, e.g., Chan, Koop, Leon-Gonzalez, and Strachan (2012), Nakajima and West (2013) and Belmonte, Koop, and Korobilis (2014).

marginal likelihood estimates computed using the modified harmonic mean (Gelfand and Dey, 1994) of the conditional likelihood can have a substantial bias and tend to select the wrong model.[3] Frühwirth-Schnatter and Wagner (2008) conclude the same when Chib's method (Chib, 1995) is used. In a related context, Millar (2009) and Chan and Grant (2016b) provide Monte Carlo evidence that the DIC based on the conditional likelihood almost always favors the most complex models.

In contrast, our proposed estimators are based on the integrated likelihood—i.e., the conditional density of the data marginal of all the latent states. As such, the proposed estimators have good theoretical properties and are substantially more stable in practice. Specifically, integrated likelihood evaluation is achieved by integrating out the time-varying parameters analytically, while the log-volatilities are integrated out numerically via importance sampling. A key novel feature of our approach is that it is based on band and sparse matrix algorithms instead of the conventional Kalman filter, which markedly reduces the computational costs. Our approach builds upon earlier work on DIC and marginal likelihood estimation for TVP-VARs (but without stochastic volatility) developed in Chan and Grant (2016a) and Chan and Eisenstat (2015). The extension to multivariate stochastic volatility models is nontrivial as it involves high-dimensional Monte Carlo integration.

We illustrate the proposed methodology by a model comparison exercise using a standard set of macroeconomic variables for the US and Australia. Specifically, we evaluate the support for various TVP-VARs with or without stochastic volatility, with the aim of contributing to the "good luck" versus "good policy" debate. The main results can be summarized as follows. For US data, the model of Primiceri (2005)—with both time-varying parameters and stochastic volatility—is overwhelmingly favored by the data compared to a conventional VAR according to both criteria. However, most of the gains appear to have come from allowing for stochastic volatility rather than time variation in the VAR coefficients or contemporaneous relationships. In fact, both criteria prefer a constant coefficients VAR with stochastic volatility against the more general model of Primiceri (2005).

This suggests that the time variation in the variance of exogenous shocks is empirically more important than changes in the monetary policy regime, lending support for the good luck hypothesis of the Great Moderation. These results also provide empirical support for the modeling approach of Carriero, Clark, and Marcellino (2016), who construct large constant coefficients VARs with a variety of stochastic volatility specifications. Results for Australia are similar: the data overwhelmingly favor the model of Primiceri (2005) against a conventional VAR with homoscedastic innovations. In addition, a constant coefficients VAR with stochastic volatility receives similar support as the model of Primiceri (2005).

In this paper we have focused on Bayesian model comparison, but the integrated likelihood

[3]For instance, in one example that involves US CPI inflation, they show that the log marginal likelihood of an unobserved components model should be -591.94, but the modified harmonic mean estimate is -494.62 (the associated numerical standard error is 1.32). Even when the number of draws is increased to ten millions, the estimate is -502.70—the finite sample bias is still substantial.

estimators can be used in other settings, such as in developing more efficient MCMC samplers or designing reversible jump MCMC algorithms to explore models of different dimensions.[4] The rest of this paper is organized as follows. In Section 2 we introduce the class of TVP-VARs we wish to compare. We give an overview of the two Bayesian model comparison criteria—the marginal likelihood and DIC in Section 3. Section 4 discusses the estimation of the two criteria for the competing models. To that end, we first introduce a fast routine to evaluate the integrated likelihood, which is followed by a discussion of an adaptive importance sampling approach known as the improved cross-entropy method for marginal likelihood computation. Section 5 evaluates the evidence in support of the TVP-VARs in explaining the US and Australian data. Lastly, Section 6 concludes and briefly discusses some future research directions.

2 TVP-VARs with Stochastic Volatility

In this section we outline the class of models we wish to compare. We first discuss the most general model; other models are then specified as restricted versions of this general model. To that end, let $\mathbf{y}_t$ be an $n \times 1$ vector of observations. Consider the following TVP-VAR with stochastic volatility:

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

where $\boldsymbol{\mu}_t$ is an $n \times 1$ vector of time-varying intercepts, $\mathbf{B}_{1t}, \ldots, \mathbf{B}_{pt}$ are $n \times n$ VAR coefficient matrices, $\mathbf{B}_{0t}$ is an $n \times n$ lower triangular matrix with ones on the diagonal and $\boldsymbol{\Sigma}_t = \text{diag}(\exp(h_{1t}), \ldots, \exp(h_{nt}))$.[5]

For the purpose of model comparison, we separate the time-varying parameters into two groups. The first group consists of the $k_\beta \times 1$ vector of time-varying intercepts and coefficients associated with the lagged observations: $\boldsymbol{\beta} = \text{vec}((\boldsymbol{\mu}_t, \mathbf{B}_{1t}, \ldots \mathbf{B}_{pt})')$. The second group is the $k_\gamma \times 1$ vector of time-varying coefficients that characterize the contemporaneous relationships among the variables, which we denote as $\boldsymbol{\gamma}_t$— it consists of the free elements of $\mathbf{B}_{0t}$ stacked by rows. Note that $k_\beta = n(np+1)$ and $k_\gamma = n(n-1)/2$. With these two groups of parameters defined, we can rewrite (1) as:

$$\mathbf{y}_t = \widetilde{\mathbf{X}}_t\boldsymbol{\beta}_t + \mathbf{W}_t\boldsymbol{\gamma}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t),$$

[4] In addition to the two Bayesian model selection criteria considered in this paper, another possibility is to construct a reversible jump MCMC algorithm to compute the posterior model probabilities. Primiceri (2005) uses this strategy to compare various choices of hyperparameter values. Note that in his setting, the dimensions of all the models considered are the same—the models only differ in their hyperparameters. For our problem, the dimensions of the models can be very different. As such, to compute the transition probability, e.g., from a constant coefficients VAR to one with stochastic volatility, one would need to evaluate the integrated likelihood of the latter model. Hence, our proposed method would also be useful if such an approach is desired. We leave this possibility to future research.

[5] Note that this TVP-VAR is written in the structural form and is therefore different from the reduced-form formulation in Primiceri (2005). However, reduced-form coefficients can be easily recovered from the structural-form coefficients.

where $\widetilde{\mathbf{X}}_t = \mathbf{I}_n \otimes (1, \mathbf{y}'_{t-1}, \cdots, \mathbf{y}'_{t-p})$ and $\mathbf{W}_t$ is an $n \times k_\gamma$ matrix that contains appropriate elements of $-\mathbf{y}_t$. For example, when $n = 3$, $\mathbf{W}_t$ has the form

$$\mathbf{W}_t = \begin{pmatrix} 0 & 0 & 0 \\ -y_{1t} & 0 & 0 \\ 0 & -y_{1t} & -y_{2t} \end{pmatrix},$$

where y_{it} is the i-th element of $\mathbf{y}_t$ for $i = 1, 2$. In the application we will investigate the empirical relevance of allowing time variation in each group of parameters.

Finally, the above model can be further written as a generic state space model:

$$\mathbf{y}_t = \mathbf{X}_t \boldsymbol{\theta}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t), \tag{2}$$

where $\mathbf{X}_t = (\widetilde{\mathbf{X}}_t, \mathbf{W}_t)$ and $\boldsymbol{\theta}_t = (\boldsymbol{\beta}'_t, \boldsymbol{\gamma}'_t)'$ is of dimension $k_\theta = k_\beta + k_\gamma$. This representation is first used in Eisenstat, Chan, and Strachan (2015) to improve the efficiency of the sampler by drawing $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$ jointly—instead of the conventional approach in Primiceri (2005) that samples $\boldsymbol{\beta}_t$ given $\boldsymbol{\gamma}_t$ followed by sampling $\boldsymbol{\gamma}_t$ given $\boldsymbol{\beta}_t$. Moreover, it also allows us to integrate out both $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$ analytically, which is important for the method of integrated likelihood evaluation described later.

The time-varying parameters $\boldsymbol{\theta}_t$ and log-volatilities $\mathbf{h}_t = (h_{1t}, \ldots, h_{nt})'$ in turn follow the following random walk processes:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\theta), \tag{3}$$

$$\mathbf{h}_t = \mathbf{h}_{t-1} + \boldsymbol{\zeta}_t, \quad \boldsymbol{\zeta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_h). \tag{4}$$

We treat the initial conditions $\boldsymbol{\theta}_0$ and $\mathbf{h}_0$ as parameters to be estimated.

The model (2)–(4) can be fitted using Markov chain Monte Carlo methods. In particular, it is conventionally estimated using Kalman filter-based algorithms in conjunction with the auxiliary mixture sampler of Kim, Shepherd, and Chib (1998). In contrast, here we adopt the precision sampler of Chan and Jeliazkov (2009) that is based on fast band and sparse matrix routines and is more efficient than Kalman filter-based algorithms. We modify the algorithm of Primiceri (2005) as discussed in Del Negro and Primiceri (2015). Estimation details are given in Appendix A.

We denote the general model in (2)–(4) as TVP-SV. To investigate what features are the most important in explaining the observed data, we consider a variety of restricted versions of this general model in the model comparison exercise. The competing models are listed in Table 1. More specifically, to examine the role of time-varying volatility, we consider a model with only drifting coefficients but no stochastic volatility, as well as a version that has stochastic volatility but with constant coefficients. The former is referred to as TVP and the latter as CVAR-SV.

Next, to investigate the individual contributions of the two groups of time-varying coefficients, we have two variants of the general model in which either $\boldsymbol{\beta}_t$ or $\boldsymbol{\gamma}_t$ is restricted to be time-invariant—the former is denoted as TVP-R1-SV and the latter as TVP-R2-SV.

Note that TVP-R2-SV is the model proposed in Cogley and Sargent (2005). Lastly, we also include a VAR with both constant coefficients and variance, which we simply refer to as CVAR.

Table 1: List of competing models.

TVP-SV	the time-varying parameter VAR with SV in (2)–(4)
TVP	same as TVP-SV but without SV
TVP-R1-SV	same as TVP-SV but $\boldsymbol{\beta}_t$ is restricted to be time-invariant
TVP-R2-SV	same as TVP-SV but $\boldsymbol{\gamma}_t$ is restricted to be time-invariant
CVAR-SV	the constant coefficients VAR with SV
CVAR	the constant coefficients VAR without SV

3 Bayesian Model Comparison Criteria

In this section we give an overview of the two Bayesian model comparison criteria—the marginal likelihood and the deviance information criterion—which we will use to compare the models outlined in Section 2.

To set the stage, suppose we wish to compare a collection of models $\{M_1, \dots, M_K\}$, where each model M_k is formally defined by a likelihood function $p(\mathbf{y} \mid \boldsymbol{\psi}_k, M_k)$ and a prior on the model-specific parameter vector $\boldsymbol{\psi}_k$ denoted by $p(\boldsymbol{\psi}_k \mid M_k)$. A natural Bayesian model comparison criterion is the *Bayes factor* in favor of M_i against M_j, defined as

$$\mathrm{BF}_{ij} = \frac{p(\mathbf{y} \mid M_i)}{p(\mathbf{y} \mid M_j)},$$

where

$$p(\mathbf{y} \mid M_k) = \int p(\mathbf{y} \mid \boldsymbol{\psi}_k, M_k) p(\boldsymbol{\psi}_k \mid M_k) \mathrm{d}\boldsymbol{\psi}_k$$

is the *marginal likelihood* under model M_k, $k = i, j$. The marginal likelihood can be interpreted as a density forecast from the model evaluated at the observed data $\mathbf{y}$—hence, if the observed data are likely under the model, the corresponding marginal likelihood would be "large" and vice versa. Therefore, if BF_{ij} is larger than 1, observed data are more likely under model M_i than model M_j, which is viewed as evidence in favor of M_i.

Furthermore, the Bayes factor is related to the *posterior odds ratio* between the two models as follows:

$$\frac{\mathbb{P}(M_i \mid \mathbf{y})}{\mathbb{P}(M_j \mid \mathbf{y})} = \frac{\mathbb{P}(M_i)}{\mathbb{P}(M_j)} \times \mathrm{BF}_{ij},$$

where $\mathbb{P}(M_i)/\mathbb{P}(M_j)$ is the prior odds ratio. It follows that if both models are equally probable *a priori*, i.e., $p(M_i) = p(M_j)$, the posterior odds ratio between the two models is then equal to the Bayes factor. In that case, if, for example, $\mathrm{BF}_{ij} = 10$, then model M_i is 10 times more likely than model M_j given the data. For a more detailed discussion of

the Bayes factor and its role in Bayesian model comparison, see Koop (2003) or Kroese and Chan (2014). From here onwards we suppress the model indicator; for example we denote the likelihood by $p(\mathbf{y} \mid \boldsymbol{\psi})$.

The Bayes factor is conceptually simple and has a natural interpretation. However, one drawback is that it is relatively sensitive to the prior distributions. An alternative Bayesian model selection criterion that is relatively insensitive to the priors is the deviance information criterion (DIC) introduced in the seminal paper by Spiegelhalter, Best, Carlin, and van der Linde (2002). This criterion can be viewed as a tradeoff between model fit and model complexity. It is based on the *deviance*, which is defined as

$$D(\boldsymbol{\psi}) = -2 \log p(\mathbf{y} \mid \boldsymbol{\psi}) + 2 \log h(\mathbf{y}),$$

where $h(\mathbf{y})$ is some fully specified standardizing term that is a function of the data alone.

Model complexity is measured by the *effective number of parameters* p_D of the model, which is defined to be

$$p_D = \overline{D(\boldsymbol{\psi})} - D(\widetilde{\boldsymbol{\psi}}), \tag{5}$$

where

$$\overline{D(\boldsymbol{\psi})} = -2\mathbb{E}_{\boldsymbol{\psi}}[\log p(\mathbf{y} \mid \boldsymbol{\psi}) \mid \mathbf{y}] + 2 \log h(\mathbf{y})$$

is the posterior mean deviance and $\widetilde{\boldsymbol{\psi}}$ is an estimate of $\boldsymbol{\psi}$, which is typically taken as the posterior mean or mode.[6] The difference between the number of parameters (i.e., cardinality of $\boldsymbol{\psi}$) and p_D may be viewed as a measure of shrinkage of the posterior estimates towards the prior means; see Spiegelhalter et al. (2002) for a more detailed discussion.

Then, the DIC is defined as the sum of the posterior mean deviance, which can be used as a Bayesian measure of model fit or adequacy, and the effective number of parameters:

$$\text{DIC} = \overline{D(\boldsymbol{\psi})} + p_D.$$

For model comparison, the function $h(\mathbf{y})$ is often set to be unity for all models. Given a set of competing models for the data, the preferred model is the one with the minimum DIC value.

We note that there are alternative definitions of the DIC depending on different concepts of the likelihood (Celeux, Forbes, Robert, and Titterington, 2006). In particular, suppose we augment the model $p(\mathbf{y} \mid \boldsymbol{\psi})$ with a vector of latent variables $\mathbf{z}$ with density $p(\mathbf{z} \mid \boldsymbol{\psi})$ such that

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

where $p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{z})$ is the *conditional likelihood* and $p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$ is the *complete-data likelihood*. To avoid ambiguity, we refer to the likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$ as the *observed-data likelihood* or the *integrated likelihood*.

[6] Following the recommendation of Spiegelhalter et al. (2002), in the empirical application we use the posterior means based on parameterizations obeying approximate likelihood normality. In particular, we take log of all variance parameters so that they take values on the real line.

An alternative DIC can then be defined in terms of the conditional likelihood, which has been used in numerous applications (e.g., Yu and Meyer, 2006; Abanto-Valle, Bandyopadhyay, Lachos, and Enriquez, 2010; Mumtaz and Surico, 2012; Brooks and Prokopczuk, 2013). However, this variant has recently been criticized on both theoretical and practical grounds. Li, Zeng, and Yu (2012) argue that the conditional DIC should not be used as a model selection criterion, as the conditional likelihood of the augmented data is nonregular and hence invalidates the standard asymptotic arguments that are needed to justify the DIC. On practical grounds, Millar (2009) and Chan and Grant (2016b) provide Monte Carlo evidence that the conditional DIC almost always favors overfitted models, whereas the original version based on the integrated likelihood works well.

Relatedly, one could in principle compute the marginal likelihood using the conditional likelihood instead of the integrated likelihood. For instance, one could estimate the marginal likelihood using the modified harmonic mean (Gelfand and Dey, 1994) of the conditional likelihood. However, Chan and Grant (2015) find that this approach does not work well in practice, as the resulting estimates have substantial bias and tend to select the wrong model. Frühwirth-Schnatter and Wagner (2008) reach the same conclusion when Chib's method is used in conjunction with the conditional likelihood. Given these findings, the calculation of both the marginal likelihood and DIC in this paper are based on the integrated likelihood.

One main difficulty of the proposed approach is that the integrated likelihood for models with stochastic volatility typically does not have a closed-form expression.[7] In fact, its evaluation is nontrivial as it requires integrating out the high-dimensional time-varying coefficients and log-volatilities. In principle one can use, e.g., the auxiliary particle filter of Pitt and Shephard (1999) to evaluate the integrated likelihood for general nonlinear state space models. In practice, however, the auxiliary particle filter is computationally intensive and it is infeasible to be employed in our settings with a large number of latent states. To overcome this problem, we develop an efficient importance sampling estimator for evaluating the integrated likelihood in the next section.

4 Marginal Likelihood and DIC Estimation

In this section we discuss the estimation of the marginal likelihood and DIC for TVP-VARs. Marginal likelihood estimation has generated a large literature; see, e.g., Friel and Wyse (2012) and Ardia, Baştürk, Hoogerheide, and van Dijk (2012) for a recent review. There are several papers dealing specifically with marginal likelihood estimation for Gaussian and non-Gaussian state space models using importance sampling (Frühwirth-Schnatter, 1995; Chan and Eisenstat, 2015) or auxiliary mixture sampling (Frühwirth-Schnatter and Wagner, 2008). We build on this line of research by extending the importance sampling methods to the more complex setting of TVP-VARs with stochastic volatility.

[7] One notable exception is the stochastic volatility of Uhlig (1997).

As mentioned in the previous section, the key ingredient in computing both the marginal likelihood and DIC is a fast routine to evaluate the integrated likelihood—the marginal density of the data unconditional on the time-varying coefficients and log-volatilities. In Section 4.1 we first propose an importance sampling algorithm for estimating the integrated likelihood. We show that one can integrate out the time-varying coefficients analytically; the log-volatilities can then be integrated out by Monte Carlo. Our approach extends earlier work on integrated likelihood evaluation for various univariate stochastic volatility models, including Durbin and Koopman (1997), Koopman and Hol Uspensky (2002), Frühwirth-Schnatter and Wagner (2008), McCausland (2012), Djegnéné and McCausland (2014) and Chan and Grant (2016b).

Once we can quickly evaluate the integrated likelihood, the DIC can then be obtained by simply averaging the integrated likelihood over the posterior draws. For marginal likelihood computation, we need an extra importance sampling step to integrate out the parameters. We adopt an adaptive importance sampling approach known as the improved cross-entropy method for this purpose, which is discussed in Section 4.2. Since we have an outer and an inner importance sampling steps

4.1 Integrated Likelihood Estimation

An intermediate goal is to develop an importance sampling estimator for the integrated likelihood:

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) &= \int p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) p(\boldsymbol{\theta}, \mathbf{h} \mid \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) \mathrm{d}(\boldsymbol{\theta}, \mathbf{h}) \\ &= \int p(\mathbf{y} \mid \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) p(\mathbf{h} \mid \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) \mathrm{d}\mathbf{h}, \end{aligned} \tag{6}$$

where $\mathbf{y} = (\mathbf{y}_1', \dots, \mathbf{y}_T')'$, $\mathbf{h} = (\mathbf{h}_1', \dots, \mathbf{h}_T')'$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \dots, \boldsymbol{\theta}_T')'$. The second term in the integrand in (6) is the prior density of $\mathbf{h}$ implied by (4), which is Gaussian (see Appendix A). The first term is the density of the data marginal of $\boldsymbol{\theta}$, which has an analytical expression and can be evaluated quickly using band matrix algorithms (see Appendix B for details).

Since both terms can be evaluated quickly, we can then estimate the integrated likelihood using importance sampling:

$$\widehat{p(\mathbf{y} \mid \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)} = \frac{1}{M} \sum_{i=1}^{M} \frac{p(\mathbf{y} \mid \mathbf{h}^i, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) p(\mathbf{h}^i \mid \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)}{g(\mathbf{h}^i; \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)}, \tag{7}$$

where $\mathbf{h}^1, \dots, \mathbf{h}^M$ are draws from the importance sampling density g that might depend on the parameters.

The choice of the importance sampling density g is of vital importance as it determines the variance of the estimator. In general, we wish to find g so that it well approximates

the integrand in (6). The ideal zero-variance importance sampling density in this case is the marginal density of $\mathbf{h}$ unconditional on $\boldsymbol{\theta}$:

$$p(\mathbf{h} \,|\, \mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) = \frac{p(\mathbf{h}, \boldsymbol{\theta} \,|\, \mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)}{p(\boldsymbol{\theta} \,|\, \mathbf{y}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)}. \tag{8}$$

However, this density cannot be used as an importance sampling density because its normalization constant is unknown. To proceed, we approximate $p(\mathbf{h} \,|\, \mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$ using a Gaussian density, which is then used as the importance sampling density. This Gaussian approximation is obtained in two steps. First, we develop an expectation-maximization (EM) algorithm (for a textbook treatment see, e.g., Kroese, Taimre, and Botev, 2011) to locate the mode of $p(\mathbf{h} \,|\, \mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$. Second, we obtain the Hessian of this density evaluated at the mode. The mode and negative Hessian are then used, respectively, as the mean and precision matrix of the Gaussian approximation.

Now, we describe the EM algorithm in more detail. To implement the E-step, we obtain the following conditional expectation (see Appendix B for an explicit expression):

$$\mathcal{Q}(\mathbf{h} \,|\, \widetilde{\mathbf{h}}) = \mathbb{E}_{\boldsymbol{\theta}|\widetilde{\mathbf{h}}} \left[\log p(\mathbf{h}, \boldsymbol{\theta} \,|\, \mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) \right],$$

where the expectation is taken with respect to $p(\boldsymbol{\theta} \,|\, \mathbf{y}, \widetilde{\mathbf{h}}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$ for an arbitrary vector $\widetilde{\mathbf{h}}$. It can be shown that this conditional density is Gaussian:

$$(\boldsymbol{\theta} \,|\, \mathbf{y}, \widetilde{\mathbf{h}}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) \sim \mathcal{N}(\widehat{\boldsymbol{\theta}}, \mathbf{K}_\theta^{-1}),$$

where $\widehat{\boldsymbol{\theta}} = \mathbf{K}_\theta^{-1} \mathbf{d}_\theta$ with

$$\mathbf{K}_\theta = \mathbf{H}_\theta' \mathbf{S}_\theta^{-1} \mathbf{H}_\theta + \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X}, \qquad \mathbf{d}_\theta = \mathbf{H}_\theta' \mathbf{S}_\theta^{-1} \mathbf{H}_\theta \boldsymbol{\alpha}_\theta + \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{y}. \tag{9}$$

Note that both the mean vector $\widehat{\boldsymbol{\theta}}$ and precision matrix $\mathbf{K}_\theta$ are functions of $\widetilde{\mathbf{h}}$—that is, $\widehat{\boldsymbol{\theta}}$ and $\mathbf{K}_\theta$ are computed using $\widetilde{\mathbf{h}}$.

In the M-step, we maximize the function $\mathcal{Q}(\mathbf{h} \,|\, \widetilde{\mathbf{h}})$ with respect to $\mathbf{h}$. This can be done using the Newton-Raphson method (see, e.g., Kroese et al., 2011). The gradient is given by

$$\mathbf{g}_\mathcal{Q} = -\mathbf{H}_h' (\mathbf{I}_T \otimes \boldsymbol{\Sigma}_h^{-1}) \mathbf{H}_h (\mathbf{h} - \boldsymbol{\alpha}_h) - \frac{1}{2} \left(\mathbf{1}_{nT} - \mathrm{e}^{-\mathbf{h}} \odot \widehat{\mathbf{z}} \right),$$

and the Hessian is

$$\mathbf{H}_\mathcal{Q} = -\mathbf{H}_h' (\mathbf{I}_T \otimes \boldsymbol{\Sigma}_h^{-1}) \mathbf{H}_h - \frac{1}{2} \operatorname{diag} \left(\mathrm{e}^{-\mathbf{h}} \odot \widehat{\mathbf{z}} \right),$$

where $\odot$ denotes the entry-wise product, $\widehat{\mathbf{z}} = (s_1^2 + \widehat{\varepsilon}_1^2, \ldots, s_{nT}^2 + \widehat{\varepsilon}_{nT}^2)'$, s_i^2 is the i-th diagonal element of $\mathbf{X} \mathbf{K}_\theta^{-1} \mathbf{X}'$ and $\widehat{\varepsilon}_i$ is the i-th element of $\mathbf{y} - \mathbf{X} \widehat{\boldsymbol{\theta}}$ with $\boldsymbol{\alpha}_h = \mathbf{H}_h^{-1} \widetilde{\boldsymbol{\alpha}}_h$, $\widetilde{\boldsymbol{\alpha}}_h = (\mathbf{h}_0', \mathbf{0}, \ldots, \mathbf{0})'$ and

$$\mathbf{H}_h = \begin{pmatrix} \mathbf{I}_n & \mathbf{0} & \cdots & \mathbf{0} \\ -\mathbf{I}_n & \mathbf{I}_n & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \cdots & -\mathbf{I}_n & \mathbf{I}_n \end{pmatrix}.$$

We emphasize that both $\mathbf{g}_{\mathcal{Q}}$ and $\mathbf{H}_{\mathcal{Q}}$ can be computed efficiently using sparse and band matrix algorithms.[8] Also note that the Hessian $\mathbf{H}_{\mathcal{Q}}$ is negative definite for all $\mathbf{h}$. This guarantees fast convergence of the Newton-Raphson method.

Finally, the EM algorithm can be implemented as follows. We initialize the algorithm with $\mathbf{h} = \mathbf{h}^{(0)}$ for some constant vector $\mathbf{h}^{(0)}$. At the j-th iteration, we obtain $\mathbf{g}_{\mathcal{Q}}$ and $\mathbf{H}_{\mathcal{Q}}$, where both $\widehat{\boldsymbol{\theta}}$ and $\mathbf{K}_\theta$ are evaluated using $\mathbf{h}^{(j-1)}$. Then, we compute

$$\mathbf{h}^{(j)} = \underset{\mathbf{h}}{\operatorname{argmax}}\, \mathcal{Q}(\mathbf{h} \,|\, \mathbf{h}^{(j-1)}),$$

using the Newton-Raphson method. We repeat the E- and M-steps until some convergence criterion is met, e.g., the norm between consecutive $\mathbf{h}^{(j)}$ is less than a predetermined tolerance value. At the end of the EM algorithm, we obtain the mode of the marginal density of $\mathbf{h}$ in (8), which is denoted by $\widehat{\mathbf{h}}$. We summarize the EM algorithm in Algorithm 1.

Algorithm 1 EM algorithm to obtain the mode of $p(\mathbf{h} \,|\, \mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$.

Suppose we have an initial guess $\mathbf{h}^{(0)}$ and error tolerance levels ε_1 and ε_2, say, $\varepsilon_1 = \varepsilon_2 = 10^{-4}$. The EM algorithm consists of iterating the following steps for $j = 1, 2, \ldots$.

1. E-Step: Given the current value $\mathbf{h}^{(j-1)}$, compute $\mathbf{K}_\theta$, $\mathbf{d}_\theta$ and $\widehat{\mathbf{z}}$
2. M-Step: Maximize $\mathcal{Q}(\mathbf{h} \,|\, \mathbf{h}^{(j-1)})$ with respect to $\mathbf{h}$ by the Newton-Raphson method. Set $\mathbf{h}^{(0,j-1)} = \mathbf{h}^{(j-1)}$ and iterate the following steps for $k = 1, 2, \ldots$.
   (a) Compute $\mathbf{g}_{\mathcal{Q}}$ and $\mathbf{H}_{\mathcal{Q}}$ using $\mathbf{K}_\theta$, $\mathbf{d}_\theta$ and $\widehat{\mathbf{z}}$ obtained in the E-step, and set $\mathbf{h} = \mathbf{h}^{(k-1,j-1)}$
   (b) Update $\mathbf{h}^{(k,j-1)} = \mathbf{h}^{(k-1,j-1)} - \mathbf{H}_{\mathcal{Q}}^{-1}\mathbf{g}_{\mathcal{Q}}$
   (c) If, for example, $\|\mathbf{h}^{(k,j-1)} - \mathbf{h}^{(k-1,j-1)}\| < \varepsilon_1$, terminate the iteration and set $\mathbf{h}^{(j)} = \mathbf{h}^{(k,j-1)}$.
3. Stopping condition: If, for example, $\|\mathbf{h}^{(j)} - \mathbf{h}^{(j-1)}\| < \varepsilon_2$, terminate the algorithm.

Next, we describe how one can compute the Hessian of the marginal density of $\mathbf{h}$ evaluated at the mode $\widehat{\mathbf{h}}$. If we take the log of both sides of (8) and then take the expectation with respect to $p(\boldsymbol{\theta} \,|\, \mathbf{y}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$, we obtain the identity

$$\log p(\mathbf{h} \,|\, \mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) = \mathcal{Q}(\mathbf{h} \,|\, \mathbf{h}) + \mathcal{H}(\mathbf{h} \,|\, \mathbf{h}), \tag{10}$$

where $\mathcal{H}(\mathbf{h} \,|\, \mathbf{h}) = -\mathbb{E}_{\boldsymbol{\theta}|\mathbf{h}}\left[\log p(\boldsymbol{\theta} \,|\, \mathbf{y}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)\right]$. An explicit expression of $\mathcal{H}(\mathbf{h} \,|\, \mathbf{h})$ is given in Appendix B.

[8] In particular, note that we only need the diagonal elements of $\mathbf{X}\mathbf{K}_\theta^{-1}\mathbf{X}'$. Since $\mathbf{K}_\theta$ is a band matrix, its Cholesky factor $\mathbf{L}_\theta$ such that $\mathbf{L}_\theta\mathbf{L}_\theta' = \mathbf{K}_\theta$ can be obtained quickly. Then, $\mathbf{U} = \mathbf{L}_\theta^{-1}\mathbf{X}$ can be computed by solving the linear system $\mathbf{L}_\theta\mathbf{U} = \mathbf{X}$ for $\mathbf{U}$. Finally, the diagonal elements of $\mathbf{X}\mathbf{K}_\theta^{-1}\mathbf{X}'$ are the row sums of squares of $\mathbf{U}$.

It follows from (10) that the Hessian of the log marginal density of $\mathbf{h}$ evaluated at $\widehat{\mathbf{h}}$ is simply the sum of the Hessians of $\mathcal{Q}$ and $\mathcal{H}$ with $\mathbf{h} = \widehat{\mathbf{h}}$. The former comes out as a by-product of the EM algorithm. As for the latter term, we show in Appendix B that the Hessian of $\mathcal{H}(\mathbf{h} \,|\, \mathbf{h})$ is given by

$$\mathbf{H}_{\mathcal{H}} = -\frac{1}{2}\mathbf{Z}' \odot (\mathbf{I}_{nT} - \mathbf{Z}),$$

where $\mathbf{Z} = \text{diag}(\text{e}^{-\mathbf{h}})\mathbf{X}\mathbf{K}_\theta^{-1}\mathbf{X}'$ and $\mathbf{K}_\theta$ is given in (9).

Finally, the Gaussian approximation is centered around the mode $\widehat{\mathbf{h}}$ with precision matrix $\mathbf{K}_h = -(\mathbf{H}_{\mathcal{Q}} + \mathbf{H}_{\mathcal{H}})$, i.e., $\mathcal{N}(\widehat{\mathbf{h}}, \mathbf{K}_h^{-1})$. Note that in general $\mathbf{K}_h$ is a full matrix as $\mathbf{H}_{\mathcal{H}}$ is full.

Algorithm 2 Integrated likelihood estimation.

Given the parameters $\boldsymbol{\Sigma}_\theta$, $\boldsymbol{\Sigma}_h$, $\boldsymbol{\theta}_0$ and $\mathbf{h}_0$, the integrated likelihood can be estimated by the following steps.

1. Obtain $\widehat{\mathbf{h}}$ and $\mathbf{H}_{\mathcal{Q}}$ using Algorithm 1.
2. Compute $\mathbf{H}_{\mathcal{H}}$ using $\mathbf{h} = \widehat{\mathbf{h}}$ and set $\mathbf{K}_h = -(\mathbf{H}_{\mathcal{Q}} + \mathbf{H}_{\mathcal{H}})$.
3. For $i = 1, \ldots, M$, simulate $\mathbf{h}^i \sim \mathcal{N}(\widehat{\mathbf{h}}, \mathbf{K}_h^{-1})$ using the method in Chan and Jeliazkov (2009) and compute the average

$$\widehat{p(\mathbf{y} \,|\, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)} = \frac{1}{M}\sum_{i=1}^{M} \frac{p(\mathbf{y} \,|\, \mathbf{h}^i, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)p(\mathbf{h}^i \,|\, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)}{g(\mathbf{h}^i; \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)}.$$

4.2 Improved Cross-Entropy Method

In the previous section we discuss an importance sampling approach to evaluate the integrated likelihood. Here it is used in conjunction with an improved version of the classic cross-entropy method to estimate the marginal likelihood. More specifically, the cross-entropy method is originally developed for rare-event simulation by Rubinstein (1997, 1999) using a multi-level procedure to construct the optimal importance sampling density (see also Rubinstein and Kroese, 2004, for a book-length treatment). Chan and Kroese (2012) later show that the optimal importance sampling density can be obtained more accurately in one step using MCMC. This new variant is applied in Chan and Eisenstat (2015) for marginal likelihood estimation. In what follows, we outline the main ideas.

First, to estimate the marginal likelihood, the ideal zero-variance importance sampling density is the posterior density $p(\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0 \,|\, \mathbf{y})$. Unfortunately, this density is only known up to a constant and therefore cannot be used directly in practice. Nevertheless, it provides a good benchmark to obtain a suitable importance sampling density.

The idea now is to locate a density that is "close" to the ideal importance sampling density. Operationally, we find the density within a convenient family of distributions such that its Kullback-Leibler divergence—or the cross-entropy distance—to the ideal density is minimized.[9] Once the optimal density is obtained, it is used to construct the importance sampling estimator:

$$\widehat{p(\mathbf{y})} = \frac{1}{N}\sum_{j=1}^{N} \frac{p(\mathbf{y} \mid \boldsymbol{\Sigma}_\theta^j, \boldsymbol{\Sigma}_h^j, \boldsymbol{\theta}_0^j, \mathbf{h}_0^j) p(\boldsymbol{\Sigma}_\theta^j, \boldsymbol{\Sigma}_h^j, \boldsymbol{\theta}_0^j, \mathbf{h}_0^j)}{g(\boldsymbol{\Sigma}_\theta^j, \boldsymbol{\Sigma}_h^j, \boldsymbol{\theta}_0^j, \mathbf{h}_0^j)}, \tag{11}$$

where $p(\mathbf{y} \mid \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$ is the integrated likelihood that can be estimated using the estimator in (7) and $(\boldsymbol{\Sigma}_\theta^1, \boldsymbol{\Sigma}_h^1, \boldsymbol{\theta}_0^1, \mathbf{h}_0^1), \ldots, (\boldsymbol{\Sigma}_\theta^N, \boldsymbol{\Sigma}_h^N, \boldsymbol{\theta}_0^N, \mathbf{h}_0^N)$ are draws from $g(\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$.

The main advantage of this importance sampling approach is that it is easy to implement and the numerical standard error of the estimator is readily available. We refer the readers to Chan and Eisenstat (2015) for technical details. We summarize the algorithm in Algorithm 3.

Algorithm 3 Marginal likelihood estimation via the improved cross-entropy method.

The marginal likelihood $p(\mathbf{y})$ can be estimated by the following steps.

1. Obtain the optimal importance sampling density $g(\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$ using the improved cross-entropy method described in Chan and Eisenstat (2015).

2. For $j = 1, \ldots, N$, simulate $(\boldsymbol{\Sigma}_\theta^j, \boldsymbol{\Sigma}_h^j, \boldsymbol{\theta}_0^j, \mathbf{h}_0^j) \sim g(\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$ and compute the average

$$\widehat{p(\mathbf{y})} = \frac{1}{N}\sum_{j=1}^{N} \frac{p(\mathbf{y} \mid \widehat{\boldsymbol{\Sigma}_\theta^j, \boldsymbol{\Sigma}_h^j}, \boldsymbol{\theta}_0^j, \mathbf{h}_0^j) p(\boldsymbol{\Sigma}_\theta^j, \boldsymbol{\Sigma}_h^j, \boldsymbol{\theta}_0^j, \mathbf{h}_0^j)}{g(\boldsymbol{\Sigma}_\theta^j, \boldsymbol{\Sigma}_h^j, \boldsymbol{\theta}_0^j, \mathbf{h}_0^j)},$$

   where the integrated likelihood estimate $\widehat{p(\mathbf{y} \mid \boldsymbol{\Sigma}_\theta^j, \boldsymbol{\Sigma}_h^j}, \boldsymbol{\theta}_0^j, \mathbf{h}_0^j)$ is computed using Algorithm 2.

Since Algorithm 3 has two nested importance steps, it falls within the importance sampling squared (IS^2) framework in Tran, Scharth, Pitt, and Kohn (2014). Following their recommendation, the simulation size M of the inner importance sampling is chosen adaptively so that the variance of the log integrated likelihood is around 1. See also the discussion in Pitt, dos Santos Silva, Giordani, and Kohn (2012).

[9] See also Frühwirth-Schnatter (1995), which constructs a different importance sampling density by using a mixture of full conditional distributions given the latent states.

5 Data and Empirical Results

In this section we compare a number of VARs that involve quarterly data on the GDP deflator, real GDP, and short-term interest rate for the US and Australia. These three variables are commonly used in forecasting (e.g., Banbura, Giannone, and Reichlin, 2010; Koop, 2013) and small DSGE models (e.g., An and Schorfheide, 2007). The data on real GDP and the GDP deflator are sourced from the Federal Reserve Bank of St. Louis economic database and the Australian Bureau of Statistics. They are then transformed to annualized growth rates.

The short-term interest rate is the effective Federal Funds rate for the US and the rate on 3-month bank accepted bills/negotiable certificates of deposit for Australia. These series are sourced from the Federal Reserve Bank of St. Louis and the Reserve Bank of Australia respectively. The sample period covers the quarters 1954Q3 to 2014Q4 for the US and 1969Q3 to 2014Q4 for Australia. Following Primiceri (2005), we order the interest rate last and treat it as the monetary policy instrument. The identified monetary policy shocks are interpreted as "non-systematic policy actions" that capture both policy mistakes and interest rate movements that are responses to variables other than inflation and GDP growth.

For each dataset, we compute the log marginal likelihoods and DICs for the competing models listed in Table 1. Each log marginal likelihood estimate is based on 10000 evaluations of the integrated likelihood, where the importance sampling density is constructed using 20000 posterior draws after a burn-in period of 5000.

Each DIC estimate (and the corresponding numerical standard error) is computed using 10 parallel chains, each consists of 20000 posterior draws after a burn-in period of 5000. The integrated likelihood is evaluated every 20-th post burn-in draw—a total of 10000 evaluations. To calculate the plug-in estimate $D(\widetilde{\boldsymbol{\psi}})$ in (5), where $\widetilde{\boldsymbol{\psi}}$ is the vector of posterior means, 500 draws are used for the integrated likelihood evaluation.

The model comparison results for US data are reported in Table 2. For comparison, we also compute the marginal likelihood of the CVAR-SV model using a brute-force approach. Specifically, let $\mathbf{y}_{1:t} = (\mathbf{y}_1', \ldots, \mathbf{y}_t')'$ denote all the data up to time t. Then, we can factor the marginal likelihood of model M_k as follows:

$$p(\mathbf{y} \mid M_k) = p(\mathbf{y}_1 \mid M_k) \prod_{t=1}^{T-1} p(\mathbf{y}_{t+1} \mid \mathbf{y}_{1:t}, M_k),$$

where $p(\mathbf{y}_{t+1} \mid \mathbf{y}_{1:t}, M_k)$ is the *predictive likelihood* under model M_k. Each predictive likelihood $p(\mathbf{y}_{t+1} \mid \mathbf{y}_{1:t}, M_k)$ is not available analytically, but it can be estimated with an MCMC run using data $\mathbf{y}_1, \ldots, \mathbf{y}_t$. Hence, to estimate the marginal likelihood this way would require a total of $T-1$ separate MCMC runs, which is generally very time-consuming.[10] Using five independent runs, we obtain an estimate of -1171.6 with a

[10] Computing the marginal likelihood of the CVAR-SV model using this approach is feasible, but it is too time-consuming for other more complex stochastic volatility models.

numerical standard error of 0.71. This is essentially identical to our estimate of −1171.7 from the proposed importance sampling approach.

Next, we return to the model comparison results in Table 2. A few broad conclusions can be drawn from this exercise. Firstly, compared to the standard CVAR, the TVP-SV with both time-varying parameters and stochastic volatility is overwhelmingly favored by the data—e.g., the Bayes factor in favor of the latter model is 3×10^{66}. However, most of the gains in model fit appear to have come from allowing for stochastic volatility rather than time variation in the VAR coefficients or contemporaneous relationships.

Table 2: Log marginal likelihood and DIC estimates for the competing VARs (numerical standard errors in parentheses); US data.

	TVP-SV	TVP	TVP-R1-SV	TVP-R2-SV	CVAR-SV	CVAR
log-ML	−1184.6	−1303.7	−1172.8	−1171.9	**−1171.7**	−1337.7
	(0.09)	(0.12)	(0.04)	(0.26)	(0.05)	(0.003)
DIC	2220.1	2400.4	2158.0	2205.9	**2148.9**	2503.1
	(0.51)	(1.79)	(0.22)	(0.35)	(0.36)	(0.17)
p_D	29.9	29.2	31.9	28.7	31.9	26.8
	(0.26)	(0.86)	(0.15)	(0.21)	(0.25)	(0.08)

In fact, the most general TVP-SV is not the best model according to both criteria. For instance, the Bayes factor in favor of CVAR-SV against TVP-SV is about 4×10^5, indicating overwhelming support for the former model; the difference in DICs is 71.2 in favor of the former. In contrast to the findings in Koop, Leon-Gonzalez, and Strachan (2009), our results suggest that when stochastic volatility is allowed, time variation in the VAR coefficients is unimportant in explaining the data. This conclusion is in line with Primiceri (2005), who computes posterior model probabilities of different hyperparameter values. His selected model is the one that implies the smallest prior variances in the state equation for the time-varying parameters. Our findings also complement the results in Sims and Zha (2006), who consider various regime-switching models and find the best model to be the one that allows time variation in disturbance variances only.

Secondly, the two model comparison criteria mostly agree in the ranking of the models. The only disagreement is in the second and third models—the marginal likelihood slightly prefers TVP-R2-SV, whereas the DIC favors TVP-R1-SV at the margin—and the order for the remaining models is exactly the same for both criteria. Given these results, one may feel comfortable using CVAR-SV as the default model. This also provides a feasible route to construct flexible high-dimensional VARs. In particular, one can consider a constant coefficients VAR with constant impact matrix and shrinkage priors as in Banbura et al. (2010) and Koop (2013), but extend the diagonal covariance matrix to allow for stochastic volatility; see Carriero, Clark, and Marcellino (2016) for such a modeling approach.

Thirdly, when the covariance matrix is restricted to be constant (comparing CVAR and

TVP), allowing for time variation in the parameters improves model fit. This finding supports the conclusion in Cogley and Sargent (2001), who find substantial time variation in the VAR coefficients in a model with constant variance. In addition, our finding is also in line with the model comparison results in Grant (2015) and Chan and Eisenstat (2015), who find that a TVP-VAR with constant variance compares favorably to a constant coefficients VAR.

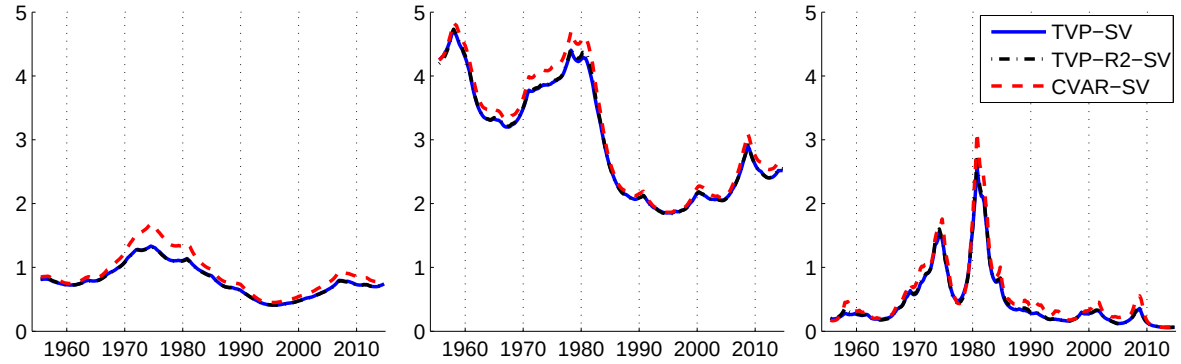

Figure 1: Estimated standard deviation of the innovation in the inflation equation (left), GDP growth equation (middle) and interest rate equation (right); US data.

We plot the posterior means of the standard deviations of the innovations for selected models in Figure 1. The volatilities of the innovations are typically quite high in the 1970s, followed by a marked decline during the Great Moderation, until they increase again following the aftermath of the Great Recession. Given these drastic changes in volatilities, it is no surprise that models that assume homoscedastic innovations cannot fit the data well. In addition, it is interesting to note that the volatility estimates are remarkably similar under the three models—although those of the CVAR-SV are slightly larger in the 1970s. This may reflect that some parameter instability in the VAR coefficients is treated as an increase in variance under CVAR-SV.

In Figure 2 we plot the impulse responses of inflation and GDP growth to a one percent monetary shock. For the TVP models, the VAR coefficients used to compute the impulse responses are fixed at the 2014Q4 estimates. The two TVP models give very similar impulse response functions, whereas those from the constant coefficients model are quite different. For example, the impulse response of inflation under the constant coefficients model is much more persistent than those of the two TVP models, highlighting the importance of performing model selection or model averaging.

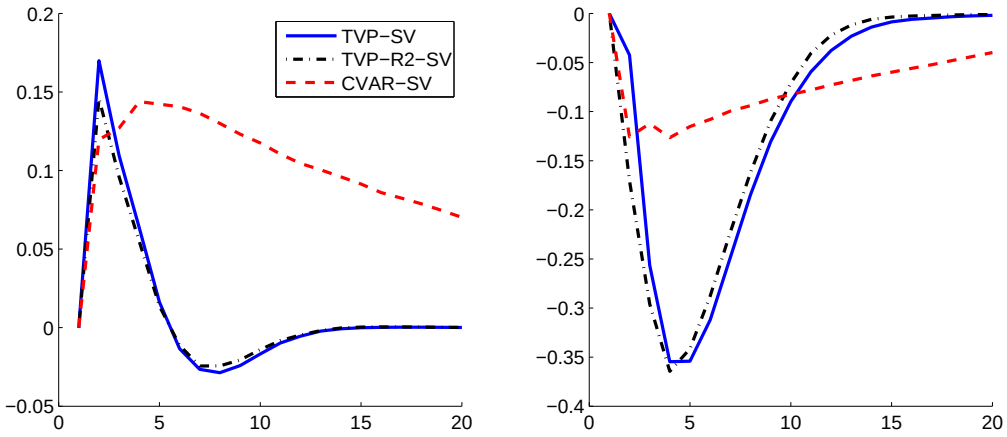

Figure 2: Impulse responses of inflation (left) and GDP growth (right) to monetary shock at 2014Q4; US data.

Next, we report the model comparison results for Australian data in Table 3. The conclusions are broadly similar to those drawn from US data, but there are a few interesting differences. Specifically, the data again overwhelmingly favor TVP-SV with both time-varying parameters and stochastic volatility against the standard CVAR according to both criteria. Furthermore, both criteria indicate that allowing for stochastic volatility alone markedly improves model fit.

In contrast to US results, however, here the two criteria disagree in the ranking of the top four models. The marginal likelihood ranks TVP-R2-SV first followed by CVAR-SV and TVP-SV in second and third. It appears that allowing for time-varying parameters further improves model fit even in the presence of stochastic volatility for Australian data.

Table 3: Log marginal likelihood and DIC estimates for the competing VARs (numerical standard errors in parentheses); Australian data.

	TVP-SV	TVP	TVP-R1-SV	TVP-R2-SV	CVAR-SV	CVAR
log-ML	−1228.9	−1341.6	−1233.4	**−1219.0**	−1228.0	−1359.2
	(0.08)	(0.04)	(0.03)	(0.06)	(0.02)	(0.004)
DIC	2326.9	2525.7	2298.0	2308.7	**2278.7**	2535.0
	(0.42)	(0.46)	(0.29)	(0.59)	(0.15)	(0.14)
p_D	29.2	26.4	30.9	28.6	31.6	26.7
	(0.20)	(0.26)	(0.10)	(0.28)	(0.07)	(0.07)

The DIC tells a slightly different story. The top three models according to the DIC are CVAR-SV, TVP-R1-SV and TVP-R2-SV. Here the most parsimonious model does the best, indicating that time variation in parameters is not needed. This difference in ranking may reflect a different penalty for model complexity—the DIC seems to penalize

model complexity more heavily than the marginal likelihood. All in all, we conclude that for Australian data, allowing for stochastic volatility is of first order importance. There is mixed evidence for time-varying parameters, but both TVP-SV and CVAR-SV have comparable support.

We plot the posterior means of the standard deviations of the innovations for selected models in Figure 3. Similar to the US results, the volatilities of the innovations are relatively large in the 1970s, but they gradually decline throughout the 1980s, 1990s and early 2000s. In contrast to US results, however, the Great Recession seemingly only affects the volatility of inflation; its effects on the volatility of GDP and interest rate are barely noticeable.

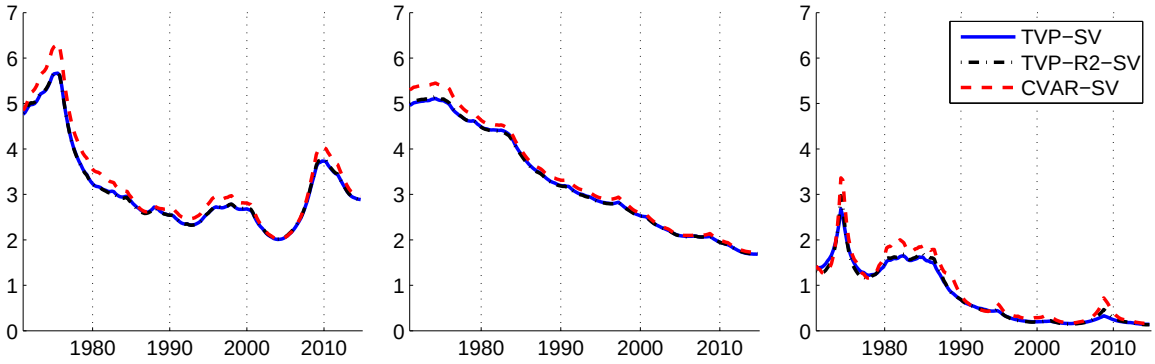

Figure 3: Estimated standard deviation of the innovation in the inflation equation (left), GDP growth equation (middle) and interest rate equation (right); Australian data.

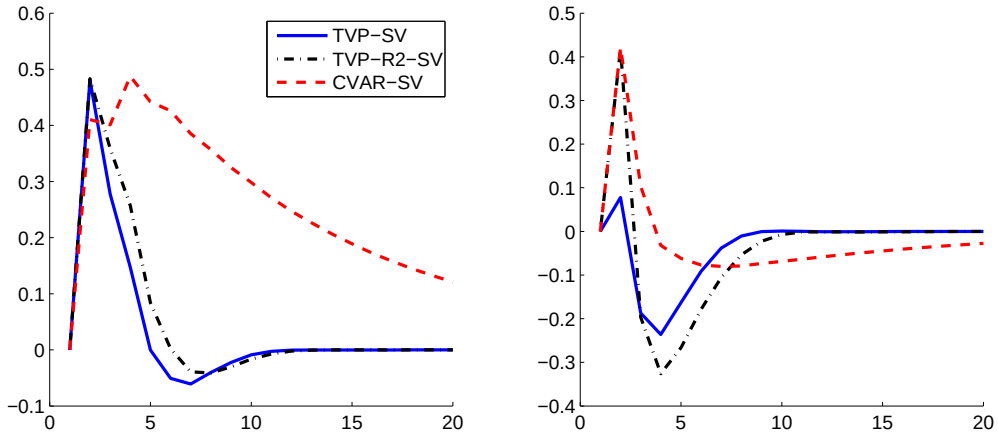

Figure 4: Impulse responses of inflation (left) and GDP growth (right) to monetary shock at 2014Q4; Australian data.

Next, we report the impulse responses of inflation and GDP growth to a one percent monetary shock in Figure 4. The impulse response of inflation is qualitatively similar

to that of the US—the two TVP models give similar responses, whereas the response of inflation under CVAR-SV is more persistent. However, the responses of GDP growth are quite different among the three models. Interestingly, all three models show an initial positive GDP growth after a positive monetary shock—in contrast to US results where the response of GDP growth is always negative—suggesting Australia faces different types of monetary shocks. Since the impulse responses are quite different under the models, this again highlights the importance of model selection and model averaging.

6 Concluding Remarks and Future Research

We have developed importance sampling estimators for evaluating the integrated likelihoods of TVP-VARs with stochastic volatility. The proposed methods are then used to compute the marginal likelihood and DIC in a model comparison exercise. Using US and Australian data, we find overwhelming support for the model of Primiceri (2005) against a conventional VAR. Nevertheless, most of the gains appear to have come from allowing for stochastic volatility rather than time variation in the VAR coefficients or contemporaneous relationships. Indeed, according to both the marginal likelihood and DIC, a constant coefficients VAR with stochastic volatility receives similar support compared to the more general model of Primiceri (2005).

However, our results do not rule out the possibility that a model in which some of the VAR coefficients are constant while others are time-varying might perform even better. To investigate this possibility, one could build upon the proposed methods of integrated likelihood evaluation to construct a reversible jump MCMC to explore the vast model space of hybrid models—e.g., we can have a model in which only one equation has time-varying coefficients or only the nominal variables have stochastic volatility. This provides an alternative to the stochastic model specification search approach of Frühwirth-Schnatter and Wagner (2010), which has been extended to TVP-VARs in Belmonte, Koop, and Korobilis (2014) and Eisenstat, Chan, and Strachan (2015).

In addition, it would also be interesting to compare large TVP-VARs. Since the number of model choices vastly increases in large systems, such a model comparison exercise would provide useful guidelines for practitioners. In particular, it would be useful to understand the effects of various shrinkage priors recently proposed in the literature. One line of investigation would be to compute the effective number of parameters and DICs for models with these shrinkage priors to see which one receives more support from the data.

Furthermore, the proposed importance sampling estimators for integrated likelihoods can be used in other settings, such as in developing more efficient MCMC samplers (e.g., as an input for particle MCMC methods; see Andrieu, Doucet, and Holenstein 2010) or designing reversible jump MCMC algorithms to explore models of different dimensions. We leave these possibilities for future research. Moreover, we have only considered TVP-VARs with simple stochastic volatility processes. It would be useful to develop similar

importance sampling methods for other richer stochastic volatility models, such as those in Eisenstat and Strachan (2015).

Appendix A: Priors and Estimation Details

In this appendix we outline the priors and provide the estimation details for fitting the model in (2)–(4).

The priors of the initial conditions $\boldsymbol{\theta}_0$ and $\mathbf{h}_0$ are both Gaussian: $\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{a}_\theta, \mathbf{V}_\theta)$ and $\mathbf{h}_0 \sim \mathcal{N}(\mathbf{a}_h, \mathbf{V}_h)$. Moreover, we assume that the error covariance matrices for the state equations are diagonal, i.e., $\boldsymbol{\Sigma}_\theta = \text{diag}(\sigma^2_{\theta 1}, \ldots, \sigma^2_{\theta k_\theta})$ and $\boldsymbol{\Sigma}_h = \text{diag}(\sigma^2_{h1}, \ldots, \sigma^2_{hn})$.[11] The elements of $\boldsymbol{\Sigma}_\theta$ and $\boldsymbol{\Sigma}_h$ are independently distributed as

$$\sigma^2_{\theta i} \sim \mathcal{IG}(\nu_{\theta i}, S_{\theta i}), \qquad \sigma^2_{hj} \sim \mathcal{IG}(\nu_{hj}, S_{hj}), \qquad i = 1, \ldots, k_\theta,\ j = 1, \ldots, k_h.$$

In particular, we set the hyperparameters to be $\mathbf{a}_\theta = \mathbf{0}$, $\mathbf{V}_\theta = 10 \times \mathbf{I}_{k_\theta}$, $\mathbf{a}_h = \mathbf{0}$ and $\mathbf{V}_h = 10 \times \mathbf{I}_n$. For the degree of freedom parameters, they are assumed to be small: $\nu_{\theta i} = \nu_{hj} = 5$. The scale parameters are set so that the prior mean of σ^2_{hj} is 0.1^2. In other words, the difference between consecutive log-volatilities is within 0.2 with probability of about 0.95. Similarly, the implied prior mean of $\sigma^2_{\theta i}$ is 0.01^2 if it is associated with a VAR coefficient and 0.1^2 for an intercept.

For notational convenience, let $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_T')'$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_T')'$. Then, posterior draws can be obtained by sequentially sampling from the following full conditional distributions:

1. $p(\boldsymbol{\theta} \,|\, \mathbf{y}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$;
2. $p(\mathbf{h} \,|\, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$;
3. $p(\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h \,|\, \mathbf{y}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\theta}_0, \mathbf{h}_0)$;
4. $p(\boldsymbol{\theta}_0, \mathbf{h}_0 \,|\, \mathbf{y}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h)$.

To implement Step 1, we first show that the conditional distribution of $\boldsymbol{\theta}$ is Gaussian. To that end, rewrite (2) as a seemingly unrelated regression:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{12}$$

where $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1', \ldots, \boldsymbol{\varepsilon}_T')'$, $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_T)$ and $\mathbf{X} = \text{diag}(\mathbf{X}_1, \ldots, \mathbf{X}_T)$. Next, let $\mathbf{H}_\theta$ denote the first difference matrix, i.e.,

$$\mathbf{H}_\theta = \begin{pmatrix} \mathbf{I}_{k_\theta} & \mathbf{0} & \cdots & \mathbf{0} \\ -\mathbf{I}_{k_\theta} & \mathbf{I}_{k_\theta} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \cdots & -\mathbf{I}_{k_\theta} & \mathbf{I}_{k_\theta} \end{pmatrix}.$$

[11] This diagonal assumption is made for simplicity and all the proposed algorithms apply to the case with general covariance matrices. In fact, the algorithms for integrated likelihood evaluation in Appendix B are presented for general covariance matrices $\boldsymbol{\Sigma}_\theta$ and $\boldsymbol{\Sigma}_h$.

Then, we can rewrite (3) as

$$\mathbf{H}_\theta \boldsymbol{\theta} = \widetilde{\boldsymbol{\alpha}}_\theta + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{S}_\theta),$$

where $\widetilde{\boldsymbol{\alpha}}_\theta = (\boldsymbol{\theta}_0', \mathbf{0}, \ldots, \mathbf{0})'$ and $\mathbf{S}_\theta = \mathbf{I}_T \otimes \boldsymbol{\Sigma}_\theta$. Or equivalently

$$(\boldsymbol{\theta} \,|\, \boldsymbol{\Sigma}_\theta, \boldsymbol{\theta}_0) \sim \mathcal{N}(\boldsymbol{\alpha}_\theta, (\mathbf{H}_\theta' \mathbf{S}_\theta^{-1} \mathbf{H}_\theta)^{-1}),$$

where $\boldsymbol{\alpha}_\theta = \mathbf{H}_\theta^{-1} \widetilde{\boldsymbol{\alpha}}_\theta$. Using standard linear regression results, one can show that (see, e.g., Kroese and Chan, 2014, Corollary 8.1):

$$(\boldsymbol{\theta} \,|\, \mathbf{y}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) \sim \mathcal{N}(\widehat{\boldsymbol{\theta}}, \mathbf{K}_\theta^{-1}),$$

where $\widehat{\boldsymbol{\theta}} = \mathbf{K}_\theta^{-1} \mathbf{d}_\theta$, and $\mathbf{K}_\theta$ and $\mathbf{d}_\theta$ are given in (9). Note that the precision matrix $\mathbf{K}_\theta$ is a band matrix—i.e., the nonzero elements are all confined within a narrow band along the main diagonal. As such, the precision sampler of Chan and Jeliazkov (2009) can be used to sample from $\mathcal{N}(\widehat{\boldsymbol{\theta}}, \mathbf{K}_\theta^{-1})$ efficiently.

To implement Step 2, we can apply the auxiliary mixture sampler of Kim et al. (1998) in conjunction of the precision sampler to sequentially draw each slice of $\mathbf{h}_{i\bullet} = (h_{i1}, \ldots, h_{iT})'$, $i = 1, \ldots, n$. Next, the elements of $\boldsymbol{\Sigma}_\theta$ and $\boldsymbol{\Sigma}_h$ are conditionally independent and follow inverse-gamma distributions:

$$(\sigma_{\theta i}^2 \,|\, \mathbf{y}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\theta}_0, \mathbf{h}_0) \sim \mathcal{IG}\left(\nu_{\theta i} + \frac{T}{2}, S_{\theta i} + \frac{1}{2}\sum_{t=1}^{T} (\theta_{it} - \theta_{i,t-1})^2\right), \qquad i = 1, \ldots, k_\theta,$$

$$(\sigma_{hj}^2 \,|\, \mathbf{y}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\theta}_0, \mathbf{h}_0) \sim \mathcal{IG}\left(\nu_{hj} + \frac{T}{2}, S_{hj} + \frac{1}{2}\sum_{t=1}^{T} (h_{jt} - h_{j,t-1})^2\right), \qquad j = 1, \ldots, k_h.$$

Lastly, $\boldsymbol{\theta}_0$ and $\mathbf{h}_0$ are conditionally independent and follow Gaussian distributions:

$$(\boldsymbol{\theta}_0 \,|\, \mathbf{y}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h) \sim \mathcal{N}(\widehat{\boldsymbol{\theta}}_0, \mathbf{K}_{\boldsymbol{\theta}_0}^{-1}), \qquad (\mathbf{h}_0 \,|\, \mathbf{y}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h) \sim \mathcal{N}(\widehat{\mathbf{h}}_0, \mathbf{K}_{\mathbf{h}_0}^{-1}),$$

where $\mathbf{K}_{\boldsymbol{\theta}_0} = \mathbf{V}_\theta^{-1} + \boldsymbol{\Sigma}_\theta^{-1}$, $\widehat{\boldsymbol{\theta}}_0 = \mathbf{K}_{\boldsymbol{\theta}_0}^{-1}(\mathbf{V}_\theta^{-1}\mathbf{a}_\theta + \boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{\theta}_1)$, $\mathbf{K}_{\mathbf{h}_0} = \mathbf{V}_h^{-1} + \boldsymbol{\Sigma}_h^{-1}$ and $\widehat{\mathbf{h}}_0 = \mathbf{K}_{\mathbf{h}_0}^{-1}(\mathbf{V}_h^{-1}\mathbf{a}_h + \boldsymbol{\Sigma}_h^{-1}\mathbf{h}_1)$.

Appendix B: Technical Details for Integrated Likelihood Estimation

In this appendix we provide the technical details for estimating the integrated likelihood outlined in Section 4.1.

We first give an analytical expression of the marginal density $p(\mathbf{y} \,|\, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$ unconditional of $\boldsymbol{\theta}$. We have showed in Appendix A that $(\mathbf{y} \,|\, \mathbf{h}, \boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and $(\boldsymbol{\theta} \,|\, \boldsymbol{\Sigma}_\theta, \boldsymbol{\theta}_0) \sim \mathcal{N}(\boldsymbol{\alpha}_\theta, (\mathbf{H}_\theta' \mathbf{S}_\theta^{-1} \mathbf{H}_\theta)^{-1})$. Then, using a similar derivation as in Chan and Grant (2016a), one can obtain the log-density as follows:

$$\begin{aligned} \log p(\mathbf{y} \,|\, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) = &-\frac{Tn}{2}\log(2\pi) - \frac{T}{2}\log|\boldsymbol{\Sigma}_\theta| - \frac{1}{2}\mathbf{1}_{nT}'\mathbf{h} - \frac{1}{2}\log|\mathbf{K}_\theta| \\ &- \frac{1}{2}\left(\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y} + \boldsymbol{\alpha}_\theta'\mathbf{H}_\theta'\mathbf{S}_\theta^{-1}\mathbf{H}_\theta\boldsymbol{\alpha}_\theta - \mathbf{d}_\theta'\mathbf{K}_\theta^{-1}\mathbf{d}_\theta\right), \end{aligned} \tag{13}$$

where $\mathbf{1}_{nT}$ is a $Tn \times 1$ column of ones, $\mathbf{K}_\theta$ and $\mathbf{d}_\theta$ are given in (9). Since $\mathbf{K}_\theta$, $\boldsymbol{\Sigma}$, $\mathbf{H}_\theta$ and $\mathbf{S}_\theta$ are all band matrices, the expression in (13) can be evaluated quickly; see Chan and Grant (2016a) for computational details.

Then, an explicit expression of $\mathcal{Q}(\mathbf{h} \,|\, \widetilde{\mathbf{h}})$ is given as follows:

$$\begin{aligned} \mathcal{Q}(\mathbf{h} \,|\, \widetilde{\mathbf{h}}) = &-\frac{1}{2}(\mathbf{h}-\boldsymbol{\alpha}_h)'\mathbf{H}_h'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}_h^{-1})\mathbf{H}_h(\mathbf{h}-\boldsymbol{\alpha}_h) - \frac{1}{2}\mathbf{1}_{nT}'\mathbf{h} \\ &- \frac{1}{2}\mathrm{tr}\left(\mathrm{diag}(\mathrm{e}^{-\mathbf{h}})\mathbb{E}_{\boldsymbol{\theta}|\widetilde{\mathbf{h}}}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})'\right]\right) + c_1 \\ = &-\frac{1}{2}(\mathbf{h}-\boldsymbol{\alpha}_h)'\mathbf{H}_h'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}_h^{-1})\mathbf{H}_h(\mathbf{h}-\boldsymbol{\alpha}_h) - \frac{1}{2}\mathbf{1}_{nT}'\mathbf{h} \\ &- \frac{1}{2}\mathrm{tr}\left(\mathrm{diag}(\mathrm{e}^{-\mathbf{h}})\left(\mathbf{X}\mathbf{K}_\theta^{-1}\mathbf{X}' + (\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\theta}})(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\theta}})'\right)\right) + c_1, \end{aligned} \tag{14}$$

where $\mathrm{tr}(\cdot)$ is the trace operator, c_1 is a constant not dependent on $\mathbf{h}$,

$$\mathbf{H}_h = \begin{pmatrix} \mathbf{I}_n & \mathbf{0} & \cdots & \mathbf{0} \\ -\mathbf{I}_n & \mathbf{I}_n & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \cdots & -\mathbf{I}_n & \mathbf{I}_n \end{pmatrix}.$$

and $\boldsymbol{\alpha}_h = \mathbf{H}_h^{-1}\widetilde{\boldsymbol{\alpha}}_h$ with $\widetilde{\boldsymbol{\alpha}}_h = (\mathbf{h}_0', \mathbf{0}, \ldots, \mathbf{0})'$.

Next, we derive the Hessian of $\mathcal{H}(\mathbf{h} \,|\, \mathbf{h})$. First, $\mathcal{H}(\mathbf{h} \,|\, \mathbf{h})$ has the follow explicit expression:

$$\begin{aligned} \mathcal{H}(\mathbf{h} \,|\, \mathbf{h}) &= -\mathbb{E}_{\boldsymbol{\theta}|\mathbf{h}}\left[\log p(\boldsymbol{\theta} \,|\, \mathbf{y}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)\right] \\ &= \frac{kT}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_\theta| + \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}|\mathbf{h}}\left[(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})'\mathbf{K}_\theta(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\right] \\ &= -\frac{1}{2}\log|\mathbf{X}'\mathrm{diag}(\mathrm{e}^{-\mathbf{h}})\mathbf{X} + \mathbf{H}_\theta'\mathbf{S}_\theta^{-1}\mathbf{H}_\theta| + c_2, \end{aligned}$$

where c_2 is a constant not dependent on $\mathbf{h}$. Note that under $p(\boldsymbol{\theta} \,|\, \mathbf{y}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$, the quadratic form $(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})' \mathbf{K}_\theta (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})$ is a chi-squared random variable and its expectation does not depend on $\mathbf{h}$.

To compute the Hessian of $\mathcal{H}(\mathbf{h} \,|\, \mathbf{h})$, we first introduce some notations. Let $\mathbf{x}_i$ be a $kT \times 1$ vector that consists of the elements in the i-th row of $\mathbf{X}$. With a slight abuse of notations, we let h_i denote the i-th element of $\mathbf{h}$. Then, it is easy to check that

$$\frac{\partial}{\partial h_i} \mathbf{K}_\theta = \frac{\partial}{\partial h_i} \mathbf{X}' \mathrm{diag}(\mathrm{e}^{-\mathbf{h}}) \mathbf{X} = \frac{\partial}{\partial h_i} \sum_{j=1}^{nT} \mathrm{e}^{-h_j} \mathbf{x}_j \mathbf{x}_j' = -\mathrm{e}^{-h_i} \mathbf{x}_i \mathbf{x}_i'.$$

Next, using standard results of matrix differentiation, we obtain

$$\begin{aligned}
\frac{\partial}{\partial h_i} \mathcal{H}(\mathbf{h} \,|\, \mathbf{h}) &= -\frac{1}{2} \mathrm{tr} \left(\mathbf{K}_\theta^{-1} \frac{\partial \mathbf{K}_\theta}{\partial h_i} \right) = \frac{1}{2} \mathrm{e}^{-h_i} \mathbf{x}_i' \mathbf{K}_\theta^{-1} \mathbf{x}_i, \\
\frac{\partial^2}{\partial h_i^2} \mathcal{H}(\mathbf{h} \,|\, \mathbf{h}) &= -\frac{1}{2} \left(\mathrm{e}^{-h_i} \mathbf{x}_i' \mathbf{K}_\theta^{-1} \mathbf{x}_i + \mathrm{e}^{-h_i} \mathbf{x}_i' \mathbf{K}_\theta^{-1} \frac{\partial \mathbf{K}_\theta}{\partial h_i} \mathbf{K}_\theta^{-1} \mathbf{x}_i \right) \\
&= -\frac{1}{2} \mathrm{e}^{-h_i} \mathbf{x}_i' \mathbf{K}_\theta^{-1} \mathbf{x}_i (1 - \mathrm{e}^{-h_i} \mathbf{x}_i' \mathbf{K}_\theta^{-1} \mathbf{x}_i), \\
\frac{\partial^2}{\partial h_i \partial h_j} \mathcal{H}(\mathbf{h} \,|\, \mathbf{h}) &= \frac{1}{2} \mathrm{e}^{-(h_i + h_j)} \mathbf{x}_i' \mathbf{K}_\theta^{-1} \mathbf{x}_j \mathbf{x}_j' \mathbf{K}_\theta^{-1} \mathbf{x}_i.
\end{aligned}$$

In matrix form, the Hessian of $\mathcal{H}(\mathbf{h} \,|\, \mathbf{h})$ is therefore

$$\mathbf{H}_{\mathcal{H}} = -\frac{1}{2} \mathbf{Z}' \odot (\mathbf{I}_{nT} - \mathbf{Z}),$$

where $\mathbf{Z} = \mathrm{diag}(\mathrm{e}^{-\mathbf{h}}) \mathbf{X} \mathbf{K}_\theta^{-1} \mathbf{X}'$.

References

C. A. Abanto-Valle, D. Bandyopadhyay, V. H. Lachos, and I. Enriquez. Robust Bayesian analysis of heavy-tailed stochastic volatility models using scale mixtures of normal distributions. *Computational Statistics and Data Analysis*, 54(12):2883–2898, 2010.

S. An and F. Schorfheide. Bayesian analysis of DSGE models. *Econometric Reviews*, 26 (2-4):113–172, 2007.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B*, 72(3):269–342, 2010.

D. Ardia, N. Baştürk, L. Hoogerheide, and H. K. van Dijk. A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood. *Computational Statistics and Data Analysis*, 56(11):3398–3414, 2012.

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

M. Belmonte, G. Koop, and D. Korobilis. Hierarchical shrinkage in time-varying coefficients models. *Journal of Forecasting*, 33(1):80–94, 2014.

L. Benati. The "great moderation" in the United Kingdom. *Journal of Money, Credit and Banking*, 40(1):121–147, 2008.

J. Boivin and M. P. Giannoni. Has monetary policy become more effective? *The Review of Economics and Statistics*, 88(3):445–462, 2006.

C. Brooks and M. Prokopczuk. The dynamics of commodity prices. *Quantitative Finance*, 13(4):527–542, 2013.

A. Carriero, T. E. Clark, and M. G. Marcellino. Common drifting volatility in large Bayesian VARs. *Journal of Business and Economic Statistics*, 34(3):375–390, 2016.

G. Celeux, F. Forbes, C. P. Robert, and D. M. Titterington. Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–674, 2006.

J. C. C. Chan and E. Eisenstat. Marginal likelihood estimation with the cross-entropy method. *Econometric Reviews*, 34(3):256–285, 2015.

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

J. C. C. Chan and A. L. Grant. Fast computation of the deviance information criterion for latent variable models. *Computational Statistics and Data Analysis*, 100:847–859, 2016a.

J. C. C. Chan and A. L. Grant. On the observed-data deviance information criterion for volatility modeling. *Journal of Financial Econometrics*, 14(4):772–802, 2016b.

J. C. C. Chan and I. Jeliazkov. Efficient simulation and integrated likelihood estimation in state space models. *International Journal of Mathematical Modelling and Numerical Optimisation*, 1(1):101–120, 2009.

J. C. C. Chan and D. P. Kroese. Improved cross-entropy method for estimation. *Statistics and Computing*, 22(5):1031–1040, 2012.

J. C. C. Chan, G. Koop, R. Leon-Gonzalez, and R. Strachan. Time varying dimension models. *Journal of Business and Economic Statistics*, 30(3):358–367, 2012.

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

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

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

T. Cogley and T. J. Sargent. Evolving post-world war II US inflation dynamics. *NBER Macroeconomics Annual*, 16:331–388, 2001.

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

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

M. Del Negro and G. E. Primiceri. Time-varying structural vector autoregressions and monetary policy: a corrigendum. *Review of Economic Studies*, 2015. Forthcoming.

B. Djegnéné and W. J. McCausland. The HESSIAN method for models with leverage-like effects. *Journal of Financial Econometrics*, 2014. Forthcoming.

J. Durbin and S. J. Koopman. Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, 84:669–684, 1997.

E. Eisenstat and R. W. Strachan. Modelling inflation volatility. *Journal of Applied Econometrics*, 2015. Forthcoming.

E. Eisenstat, J. C. C. Chan, and R. W. Strachan. Stochastic model specification search for time-varying parameter VARs. *Econometric Reviews*, 2015. Forthcoming.

N. Friel and J. Wyse. Estimating the evidence—a review. *Statistica Neerlandica*, 66(3): 288–308, 2012.

S. Frühwirth-Schnatter. Bayesian model discrimination and Bayes factors for linear Gaussian state space models. *Journal of the Royal Statistical Society Series B*, 57(1):237–246, 1995.

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

S. Frühwirth-Schnatter and H. Wagner. Stochastic model specification search for Gaussian and partial non-Gaussian state space models. *Journal of Econometrics*, 154:85–100, 2010.

A. E. Gelfand and D. K. Dey. Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society Series B*, 56(3):501–514, 1994.

A. L. Grant. The early millennium slowdown: Replicating the Peersman (2005) results. *Journal of Applied Econometrics*, 2015. Forthcoming.

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

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

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

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

G. Koop, R. Leon-Gonzalez, and R. W. Strachan. On the evolution of the monetary policy transmission mechanism. *Journal of Economic Dynamics and Control*, 33(4): 997–1017, 2009.

S. J. Koopman and E. Hol Uspensky. The stochastic volatility in mean model: Empirical evidence from international stock markets. *Journal of Applied Econometrics*, 17(6): 667–689, 2002.

D. P. Kroese and J. C. C. Chan. *Statistical Modeling and Computation*. Springer, New York, 2014.

D. P. Kroese, T. Taimre, and Z. I. Botev. *Handbook of Monte Carlo Methods*. John Wiley and Sons, New York, 2011.

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

Y. Liu and J. Morley. Structural evolution of the U.S. economy. *Journal of Economic Dynamics and Control*, 42:50–68, 2014.

W. J. McCausland. The HESSIAN method: Highly efficient simulation smoothing, in a nutshell. *Journal of Econometrics*, 168(2):189–206, 2012.

R. B. Millar. Comparison of hierarchical Bayesian models for overdispersed count data using DIC and Bayes factors. *Biometrics*, 65(3):962–969, 2009.

H. Mumtaz and P. Surico. Evolving international inflation dynamics: World and country-specific factors. *Journal of the European Economic Association*, 10(4):716–734, 2012.

J. Nakajima and M. West. Bayesian analysis of latent threshold dynamic models. *Journal of Business and Economic Statistics*, 31(2):151–164, 2013.

M. K. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599, 1999.

M. K. Pitt, R. dos Santos Silva, P. Giordani, and R. Kohn. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151, 2012.

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

R .Y. Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99:89–112, 1997.

R. Y. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, 2:127–190, 1999.

R. Y. Rubinstein and D. P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization Monte-Carlo Simulation, and Machine Learning*. Springer-Verlag, New York, 2004.

C. A. Sims and T. Zha. Were there regime switches in U.S. monetary policy? *American Economic Review*, 96(1):54–81, 2006.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, 64(4): 583–639, 2002.

M.-N. Tran, M. Scharth, M. K. Pitt, and R. Kohn. Importance sampling squared for Bayesian inference in latent variable models. *Available at SSRN 2386371*, 2014.

Harald Uhlig. Bayesian vector autoregressions with stochastic volatility. *Econometrica*, pages 59–73, 1997.

J. Yu and R. Meyer. Multivariate stochastic volatility models: Bayesian estimation and model comparison. *Econometric Reviews*, 25(2-3):361–384, 2006.