Received: 20 October 2016 | Revised: 7 November 2017

DOI: 10.1002/jae.2617

RESEARCH ARTICLE

WILEY Journal of APPLIED ECONOMETRICS

Bayesian model comparison for time-varying parameter VARs with stochastic volatility

Joshua C. C. Chan[1] | Eric Eisenstat[2]

[1]Economics Discipline Group, University of Technology Sydney, NSW Broadway, Australia

[2]School of Economics, University of Queensland, Brisbane, Queensland, Australia

Correspondence
Joshua C. C. Chan, Economics Discipline Group, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia.
Email: joshuacc.chan@gmail.com

Funding information
Australian Research Council via a Discovery Early Career Researcher Award, Grant/Award Number: DE150100795

Summary

We develop importance sampling methods for computing two popular Bayesian model comparison criteria, namely, the marginal likelihood and the deviance information criterion (DIC) for time-varying parameter vector autoregressions (TVP-VARs), where both the regression coefficients and volatilities are drifting over time. The proposed estimators are based on the integrated likelihood, which are substantially more reliable than alternatives. Using US data, we find overwhelming support for the TVP-VAR with stochastic volatility compared to a conventional constant coefficients VAR with homoskedastic 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 outperforms the more general model with time-varying parameters.

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 interrelationships 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 (2015). Despite the empirical success of these flexible time-varying models, an emerging literature has expressed concerns about their potential overparametrization.[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 the 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.

[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, for example, Chan, Koop, Leon-Gonzalez, and Strachan (2012), Nakajima and West (2013), and Belmonte, Koop, and Korobilis (2014).

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 is more empirically relevant, one direct approach is to perform a model comparison exercise—for example, 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.

Our contributions are twofold. On the methodological side, we develop importance sampling methods for computing two popular Bayesian model comparison criteria, namely, the marginal likelihood and the deviance information criterion (DIC) for TVP-VARs with stochastic volatility. 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 marginal likelihood estimates computed using the modified harmonic mean (Gelfand & 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—that is, 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.[4]

On the empirical side, we illustrate the proposed methodology by a model comparison exercise using a standard set of macroeconomic variables for the USA. 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. 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. 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.

In this paper we have focused on Bayesian model comparison, but the integrated likelihood estimators can be used in other settings, such as in developing more efficient Markov chain Monte Carlo (MCMC) samplers or designing reversible jump MCMC algorithms to explore models of different dimensions.[5] 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. We then conduct a small Monte Carlo experiment in Section 5 to assess how the proposed algorithms perform in selecting the correct models. Section 6 evaluates the evidence in support of the TVP-VARs in explaining the US data. Lastly, Section 8 concludes and briefly discusses some future research directions.

[3]For instance, in one example that involves US consumer price index 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 million, the estimate is −502.70—the finite sample bias is still substantial.

[4]The associated MATLAB code is available at http://joshuachan.org/code.html.

[5]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—for example, 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.

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 μ_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 $\boldsymbol{\Sigma}_t = \text{diag}(\exp(h_{1t}), \ldots, \exp(h_{nt}))$.[6] The log volatilities $\mathbf{h}_t = (h_{1t}, \ldots, h_{nt})'$ evolve according to the following random walk:

$$\mathbf{h}_t = \mathbf{h}_{t-1} + \boldsymbol{\zeta}_t, \quad \boldsymbol{\zeta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_h), \tag{2}$$

where the initial conditions $\mathbf{h}_0$ are treated as parameters to be estimated.

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: $\boldsymbol{\beta}_t = \text{vec}((\mu_t, \mathbf{B}_{1t}, \ldots \mathbf{B}_{pt})')$. 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 Equation 1 as follows:

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

where $\tilde{\mathbf{X}}_t = \mathbf{I}_n \otimes (1, \mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p})$ 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 ith 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:

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\theta}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t), \tag{3}$$

where $\mathbf{X}_t = (\tilde{\mathbf{X}}_t, \mathbf{W}_t)$ and $\boldsymbol{\theta}_t = (\boldsymbol{\beta}'_t, \boldsymbol{\gamma}'_t)'$ is of dimension $k_\theta = k_\beta + k_\gamma$. This representation is used in Eisenstat, Chan, and Strachan (2016) 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 vector of time-varying parameters $\boldsymbol{\theta}_t$ in turn follows the following random walk process:

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

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

To complete the model specification, below we give the details of the priors on the model parameters. 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—that is, $\boldsymbol{\Sigma}_\theta = \text{diag}(\sigma^2_{\theta 1}, \ldots, \sigma^2_{\theta k_\theta})$ and $\boldsymbol{\Sigma}_h = \text{diag}(\sigma^2_{h1}, \ldots, \sigma^2_{hn})$.[7] The diagonal elements of $\boldsymbol{\Sigma}_\theta$ and $\boldsymbol{\Sigma}_h$ are independently distributed as

[6] 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.

[7] 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 $\boldsymbol{\Sigma}_h$.

TABLE 1 List of competing models

TVP-SV	Time-varying parameter VAR with SV in Equations 2–(4)
TVP	Same as TVP-SV but $\mathbf{h}_t = \mathbf{h}_0$ and $\boldsymbol{\gamma}_t = \boldsymbol{\gamma}_0$
TVP-R1-SV	Same as TVP-SV but $\boldsymbol{\beta}_t = \boldsymbol{\beta}_0$
TVP-R2-SV	Same as TVP-SV but $\boldsymbol{\gamma}_t = \boldsymbol{\gamma}_0$
TVP-R3-SV	Same as TVP-SV but only the intercepts are time varying
CVAR-SV	Same as TVP-SV but $\boldsymbol{\beta}_t = \boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_t = \boldsymbol{\gamma}_0$
CVAR	Constant coefficients VAR with $\boldsymbol{\theta}_t = \boldsymbol{\theta}_0$ and $\mathbf{h}_t = \mathbf{h}_0$
RS-VAR	Regime-switching VAR in Equation 5
RS-VAR-R1	Same as RS-VAR but $(\mathbf{B}_{0j}, \mathbf{B}_{1j}, \dots, \mathbf{B}_{pj})$ are the same across regimes
RS-VAR-R2	Same as RS-VAR but $\boldsymbol{\Sigma}_j$ are the same across regimes

$$\sigma_{\theta i}^2 \sim \mathcal{IG}(\nu_{\theta i}, S_{\theta i}), \quad \sigma_{hj}^2 \sim \mathcal{IG}(\nu_{hj}, S_{hj}), \quad i = 1, \dots, k_\theta, \; j = 1, \dots, k_h.$$

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_{hj}^2 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 $\sigma_{\theta i}^2$ is 0.01^2 if it is associated with a VAR coefficient and 0.1^2 for an intercept.

The model as specified in Equations 2–4 can be fitted using MCMC 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), which 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 Equations 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 (referred to as TVP), as well as a version that has stochastic volatility but with constant coefficients (CVAR-SV).

Next, to investigate the individual contributions of the two groups of time-varying coefficients, we consider 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 by TVP-R1-SV and the latter by TVP-R2-SV. Note that TVP-R2-SV is the model proposed in Cogley and Sargent (2005). In addition, we also consider a variant in which only the intercepts are time varying—hence this is a restricted version of TVP-R1-SV. This version is denoted by TVP-R3-SV.

For comparison, we also consider the class of regime-switching VARs similar to those in Sims and Zha (2006). More specifically, let $S_t \in \{1, \dots, r\}$ denote the regime indicator at time t, where r is the number of regimes. Then, a regime-switching VAR is given by

$$\mathbf{B}_{0S_t}\mathbf{y}_t = \boldsymbol{\mu}_{S_t} + \mathbf{B}_{1S_t}\mathbf{y}_{t-1} + \cdots + \mathbf{B}_{pS_t}\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{S_t}), \tag{5}$$

where $(\mathbf{B}_{0j}, \mathbf{B}_{1j}, \dots, \mathbf{B}_{pj})$ and $\boldsymbol{\Sigma}_j$ for $j = 1, \dots, r$ are regime-specific parameters. The regime indicator S_t is assumed to follow a Markov process with transition probability $\mathbb{P}(S_t = j \,|\, S_{t-1} = i) = p_{ij}$. This regime-switching VAR is denoted by RS-VAR.

In addition, we consider two restricted versions of RS-VAR in which either the VAR coefficients $(\mathbf{B}_{0j}, \mathbf{B}_{1j}, \dots, \mathbf{B}_{pj})$ or the covariance matrices $\boldsymbol{\Sigma}_j$ are the same across regimes. The former is denoted by RS-VAR-R1 and the latter by RS-VAR-R2. Lastly, we also include a VAR with both constant coefficients and covariance matrix, which we simply refer to as CVAR.

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 $p(\mathbf{y}|\boldsymbol{\psi}_k, M_k)$ and a prior on the model-specific parameter vector $\boldsymbol{\psi}_k$ denoted by $p(\boldsymbol{\psi}_k|M_k)$.

A natural Bayesian model comparison criterion is the *Bayes factor* in favor of M_i against M_j, defined as

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

where

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

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:

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

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, that is, $p(M_i) = p(M_j)$, 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 $p(\mathbf{y}|\boldsymbol{\psi})$.

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

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

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

$$p_D = \overline{D(\boldsymbol{\psi})} - D(\tilde{\boldsymbol{\psi}}), \tag{6}$$

where

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

is the posterior mean deviance and $\tilde{\boldsymbol{\psi}}$ is an estimate of $\boldsymbol{\psi}$, which is typically taken as the posterior mean or mode. 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:

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

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, & Titterington, 2006). In particular, suppose we augment the model $p(\mathbf{y}|\boldsymbol{\psi})$ with a vector of latent variables $\mathbf{z}$ with density $p(\mathbf{z}|\boldsymbol{\psi})$ such that

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

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

An alternative DIC can then be defined in terms of the conditional likelihood, which has been used in numerous applications (e.g., Abanto-Valle, Bandyopadhyay, Lachos, & Enriquez, 2010; Brooks & Prokopczuk, 2013; Mumtaz & Surico, 2012; Yu & Meyer, 2006). 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 & 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.[8] 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, for example, 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 employ 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, for example, 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 (Chan & Eisenstat, 2015; Frühwirth-Schnatter, 1995) or auxiliary mixture sampling (Frühwirth-Schnatter & Wagner, 2008). We build on this line of research by extending importance sampling methods to the more complex setting of TVP-VARs with stochastic volatility.

For observed-data DIC estimation in the context of nonlinear state-space models, the literature is relatively scarce. The main difficulty is the evaluation of the integrated likelihood—the marginal density of the data unconditional on the latent states. Earlier work, such as Durbin and Koopman (1997), McCausland (2012), and Chan and Grant (2016b), considers integrated likelihood estimation only for univariate stochastic volatility models. Here we develop algorithms suitable for high-dimensional stochastic volatility models.

Section 4.1 first discusses the estimation of the observed-data DIC. The main step is to compute the average of the integrated likelihoods over the posterior draws. There we present a fast routine to evaluate the integrated likelihood—marginal of the time-varying coefficients and log volatilities. It involves an importance sampling algorithm that first integrates out the time-varying coefficients analytically, followed by integrating out the log volatilities using Monte Carlo.

Next, marginal likelihood computation is discussed in Section 4.2. In addition to integrating out the latent states, marginal likelihood computation requires 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.

4.1 | DIC estimation

We now discuss the estimation of the observed-data DIC. Let $\boldsymbol{\psi} = (\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$ denote the model parameters in the TVP-SV model and let $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_T')'$ denote the data. Recall that the DIC is defined as

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

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

where $\overline{D(\boldsymbol{\psi})}$ is the posterior mean of the deviance and $p_D = \overline{D(\boldsymbol{\psi})} - D(\tilde{\boldsymbol{\psi}})$ is the effective number of parameters, with $\tilde{\boldsymbol{\psi}}$ being the posterior mean.

The main challenge in computing the DIC is the calculation of the first term:

$$\overline{D(\boldsymbol{\psi})} = -2\mathbb{E}_{\boldsymbol{\psi}}[\log p(\mathbf{y}\,|\,\boldsymbol{\psi})\,|\,\mathbf{y}] = -2\mathbb{E}_{\boldsymbol{\psi}}[\log p(\mathbf{y}\,|\,\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)\,|\,\mathbf{y}].$$

To evaluate the posterior expectation, one could obtain the average of the log integrated likelihood $\log p(\mathbf{y}\,|\,\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$ over the posterior draws. However, this is computationally challenging as the evaluation of the integrated likelihood is non-trivial—it involves a very high-dimensional integration:

$$p(\mathbf{y}\,|\,\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) = \int p(\mathbf{y}\,|\,\boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) p(\boldsymbol{\theta}, \mathbf{h}\,|\,\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) \mathrm{d}(\boldsymbol{\theta}, \mathbf{h}),$$

where $\mathbf{h} = (\mathbf{h}_1', \ldots, \mathbf{h}_T')'$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_T')'$.

Below we develop an importance sampling algorithm for estimating the integrated likelihood in two steps. In the first step, we integrate out the time-varying coefficients $\boldsymbol{\theta}$ analytically. In the second step, we use importance sampling to integrate out the log volatilities $\mathbf{h}$. 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 (2016), and Chan and Grant (2016b).

To implement the first step, recall that $\boldsymbol{\theta}$ and $\mathbf{h}$ evolve according to independent random walks given in Equations 4 and 2, respectively. Therefore, they are conditionally independent:

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

Now, we write the integrated likelihood as

$$\begin{aligned} p(\mathbf{y}\,|\,\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) &= \int p(\mathbf{y}\,|\,\boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) p(\boldsymbol{\theta}\,|\,\boldsymbol{\Sigma}_\theta, \boldsymbol{\theta}_0) p(\mathbf{h}\,|\,\boldsymbol{\Sigma}_h, \mathbf{h}_0) \mathrm{d}(\boldsymbol{\theta}, \mathbf{h}) \\ &= \int p(\mathbf{y}\,|\,\mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) p(\mathbf{h}\,|\,\boldsymbol{\Sigma}_h, \mathbf{h}_0) \mathrm{d}\mathbf{h}. \end{aligned} \tag{7}$$

Both terms in the integrand in Equation 7 have an analytical expression. The first term is the density of the data marginal of $\boldsymbol{\theta}$; its closed-form expression is given in Appendix B. The second term is the prior density of $\mathbf{h}$ implied by Equation 2—it is in fact Gaussian, and its closed-formed expression is given in Appendix A.

In the second step, we further integrate out $\mathbf{h}$ using importance sampling. Specifically, since both terms in the integrand in Equation 7 can be evaluated quickly, we can then estimate the integrated likelihood using importance sampling:

$$\hat{p}(\mathbf{y}\,|\,\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) = \frac{1}{M} \sum_{i=1}^{M} \frac{p(\mathbf{y}\,|\,\mathbf{h}^i, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) p(\mathbf{h}^i\,|\,\boldsymbol{\Sigma}_h, \mathbf{h}_0)}{g(\mathbf{h}^i; \mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)}, \tag{8}$$

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

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 Equation 7. The ideal zero-variance importance sampling density in this case is the marginal density of $\mathbf{h}$ unconditional on $\boldsymbol{\theta}$:

$$p(\mathbf{h}\,|\,\mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) = \frac{p(\mathbf{h}, \boldsymbol{\theta}\,|\,\mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)}{p(\boldsymbol{\theta}\,|\,\mathbf{y}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)}. \tag{9}$$

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 as follows. We first develop an expectation maximization (EM) algorithm (for a textbook treatment see, e.g., Kroese, Taimre, & Botev, 2011) to locate the mode of $p(\mathbf{h}|\mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$, say, $\hat{\mathbf{h}}$. Then, we obtain the *negative* Hessian of this density evaluated at the mode, say, $\mathbf{K}_h$. The mode and negative Hessian are then used, respectively, as the mean and precision matrix of the Gaussian approximation. That is, the importance sampling density is $\mathcal{N}(\hat{\mathbf{h}}, \mathbf{K}_h^{-1})$. We leave the technical details to Appendix A. We summarize the estimation of the integrated likelihood in Algorithm 1.

Algorithm 1. (Integrated likelihood estimation)
Given the parameters $\boldsymbol{\Sigma}_\theta$, $\boldsymbol{\Sigma}_h$, $\boldsymbol{\theta}_0$ and $\boldsymbol{h}_0$, complete the following two steps:

1. *Obtain the mean $\hat{\mathbf{h}}$ and precision matrix $\boldsymbol{K}_h$ of the Gaussian importance sampling density.*
2. *For $i = 1, \dots, M$, simulate $\mathbf{h}^i \sim \mathcal{N}(\hat{\mathbf{h}}, \mathbf{K}_h^{-1})$ using the method in Chan and Jeliazkov (2009) and compute the average:*

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

To ensure that an importance sampling estimator works well, a requirement is that the variance of the importance sampling weights should be finite. However, checking this requirement analytically in high-dimensional settings such as ours is difficult. One way to ensure that this finite-variance condition holds is to modify the importance sampling estimator $g(\mathbf{h}) = g(\mathbf{h}; \mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$ to include an additional mixture component as proposed by Hesterberg (1995). More specifically, for $\delta \in (0, 1)$, consider the mixture density

$$g_\delta(\mathbf{h}) = \delta p(\mathbf{h} \,|\, \boldsymbol{\Sigma}_h, \mathbf{h}_0) + (1 - \delta) g(\mathbf{h}).$$

That is, samples are taken from the prior density $p(\mathbf{h}|\boldsymbol{\Sigma}_h, \mathbf{h}_0)$ with probability δ; otherwise, we draw from the original importance sampling density $g(\mathbf{h})$.

If we assume that for fixed data $\mathbf{y}$ and parameter values $(\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$, the conditional likelihood is bounded in $\mathbf{h}$; that is, there exists a constant C such that $p(\mathbf{y}|\mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) \leq C$ for all $\mathbf{h}$ (this condition holds for the stochastic volatility models we consider), then the importance sampling weight is bounded by

$$\begin{aligned} \frac{p(\mathbf{y} \,|\, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) p(\mathbf{h} \,|\, \boldsymbol{\Sigma}_h, \mathbf{h}_0)}{g_\delta(\mathbf{h}; \mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)} &\leq \frac{p(\mathbf{y} \,|\, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) p(\mathbf{h} \,|\, \boldsymbol{\Sigma}_h, \mathbf{h}_0)}{\delta p(\mathbf{h} \,|\, \boldsymbol{\Sigma}_h, \mathbf{h}_0)} \\ &\leq \frac{C}{\delta}. \end{aligned}$$

Hence the variance of the importance sampling weights corresponding to $g_\delta(\mathbf{h})$ is finite. In our applications we experiment with both $g(\mathbf{h})$ and $g_\delta(\mathbf{h})$, and they give very similar results.[9]

Finally, given Algorithm 1, we can estimate the DIC using the following Algorithm 2.

Algorithm 2. (DIC estimation)
The DIC can be estimated by the following steps:

1. *Obtain N sets of posterior draws $\boldsymbol{\psi}^i = (\boldsymbol{\Sigma}_\theta^i, \boldsymbol{\Sigma}_h^i, \boldsymbol{\theta}_0^i, \mathbf{h}_0^i)$ for $i = 1, \dots, N$.*
2. *For $i = 1, \dots, N$, compute the integrated likelihood $\hat{p}(\mathbf{y} \,|\, \boldsymbol{\psi}^i) = \hat{p}(\mathbf{y} \,|\, \boldsymbol{\Sigma}_\theta^i, \boldsymbol{\Sigma}_h^i, \boldsymbol{\theta}_0^i, \mathbf{h}_0^i)$ using Algorithm 1. Then, average the log integrated likelihoods to obtain $\overline{D(\boldsymbol{\psi})}$.*
3. *Given $\overline{D(\boldsymbol{\psi})}$ and the posterior mean $\bar{\boldsymbol{\psi}}$ obtained from the posterior draws, compute p_D.*
4. *Finally, return DIC $= \overline{D(\boldsymbol{\psi})} + p_D$.*

4.2 | Marginal likelihood estimation

In this section we discuss the marginal likelihood estimation of the TVP-SV model. We use the importance sampling approach to evaluate the integrated likelihood discussed in the previous section in conjunction with an improved version

[9]For example, using the US data in Section 6 we evaluate the log integrated likelihood of the TVP-SV model at the posterior means of the parameters. For the original estimator, we obtain an estimate of −1074.3; for the modified estimator with $\delta = 0.05$, we have −1074.5.

of the classic cross-entropy method. More specifically, the cross-entropy method is originally developed for rare-event simulation by Rubinstein (1997, 1999) using a multilevel procedure to construct the optimal importance sampling density (see also Rubinstein & 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 $p(\boldsymbol{\psi}|\mathbf{y}) = p(\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0|\mathbf{y})$. 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 key idea is to locate a density that is "close" to this ideal importance sampling density, which we denote by $f^* = f^*(\boldsymbol{\psi}) = p(\boldsymbol{\psi}|\mathbf{y})$. To that end, consider a parametric family $\mathcal{F} = \{f(\boldsymbol{\psi}; \mathbf{v})\}$ indexed by the parameter vector $\mathbf{v}$. We then find the density $f(\boldsymbol{\psi}; \mathbf{v}^*) \in \mathcal{F}$ such that it is the "closest" to f^*.

One convenient measure of closeness between densities is the *Kullback–Leibler divergence* or the *cross-entropy distance*. Specifically, let f_1 and f_2 be two probability density functions. Then, the cross-entropy distance from f_1 to f_2 is defined as follows:

$$\mathcal{D}(f_1, f_2) = \int f_1(\mathbf{x}) \log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \mathrm{d}\mathbf{x}.$$

Given this measure, we locate the density $f(\cdot; \mathbf{v}) \in \mathcal{F}$ such that $\mathcal{D}(f^*, f(\cdot; \mathbf{v}))$ is minimized:

$$\begin{aligned} \mathbf{v}_{\text{ce}}^* &= \arg\min_{\mathbf{v}} \mathcal{D}(f^*, f(\cdot; \mathbf{v})) \\ &= \arg\min_{\mathbf{v}} \left(\int f^*(\boldsymbol{\psi}) \log f^*(\boldsymbol{\psi}) \mathrm{d}\boldsymbol{\psi} - p(\mathbf{y})^{-1} \int p(\mathbf{y}\,|\,\boldsymbol{\psi}) p(\boldsymbol{\psi}) \log f(\boldsymbol{\psi}; \mathbf{v}) \mathrm{d}\boldsymbol{\psi} \right), \end{aligned}$$

where we used the fact that $f^*(\boldsymbol{\psi}) = p(\mathbf{y}|\boldsymbol{\psi})p(\boldsymbol{\psi})/p(\mathbf{y})$. Since the first term does not depend on $\mathbf{v}$, solving the CE minimization problem is equivalent to finding

$$\mathbf{v}_{\text{ce}}^* = \arg\max_{\mathbf{v}} \int p(\mathbf{y}\,|\,\boldsymbol{\psi}) p(\boldsymbol{\psi}) \log f(\boldsymbol{\psi}; \mathbf{v}) \mathrm{d}\boldsymbol{\psi}.$$

In practice, this optimization problem is often difficult to solve analytically. Instead, we consider its stochastic counterpart:

$$\hat{\mathbf{v}}_{\text{ce}}^* = \arg\max_{\mathbf{v}} \frac{1}{R} \sum_{r=1}^{R} \log f(\boldsymbol{\psi}^r; \mathbf{v}), \tag{10}$$

where $\boldsymbol{\psi}^1, \ldots, \boldsymbol{\psi}^R$ are posterior draws. In other words, $\hat{\mathbf{v}}_{\text{ce}}^*$ is exactly the maximum likelihood estimate for $\mathbf{v}$ if we treat $f(\boldsymbol{\psi}; \mathbf{v})$ as the likelihood function with parameter vector $\mathbf{v}$ and $\boldsymbol{\psi}^1, \ldots, \boldsymbol{\psi}^R$ as an observed sample. Since finding the maximum likelihood estimate is a standard problem, solving Equation 10 is typically easy. In particular, analytical solutions to Equation 10 can be found explicitly for the exponential family (e.g., Rubinstein & Kroese, 2004, p. 70).

As for the choice of parametric family $\mathcal{F}$, it is often chosen so that each member $f(\boldsymbol{\psi}; \mathbf{v})$ is a product of densities, for example, $f(\boldsymbol{\psi}; \mathbf{v}) = f(\boldsymbol{\psi}_1; \mathbf{v}_1) \times \cdots \times f(\boldsymbol{\psi}_B; \mathbf{v}_B)$, where $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_B)$ and $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_B)$. In that case, one can reduce the possibly high-dimensional maximization problem (Equation 10) into B low-dimensional problems, which can then be readily solved. For example, for the TVP-SV model, we divide $\boldsymbol{\psi} = (\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$ into four natural blocks and consider the parametric family

$$\mathcal{F} = \left\{ \prod_{i=1}^{k_\theta} f_{\mathcal{IG}}(\sigma_{\theta i}^2; \nu_{1,\theta i}, \nu_{2,\theta i}) \prod_{j=1}^{k_h} f_{\mathcal{IG}}(\sigma_{hj}^2; \nu_{1,hj}, \nu_{2,hj}) f_{\mathcal{N}}(\boldsymbol{\theta}_0; \mathbf{v}_{1,\boldsymbol{\theta}_0}, \mathbf{v}_{2,\boldsymbol{\theta}_0}) f_{\mathcal{N}}(\mathbf{h}_0; \mathbf{v}_{1,\mathbf{h}_0}, \mathbf{v}_{2,\mathbf{h}_0}) \right\},$$

where $f_{\mathcal{IG}}$ and $f_{\mathcal{N}}$ are the inverse-gamma and Gaussian densities, respectively. Given this choice of parametric family, the CE minimization problem in Equation 10 can be readily solved.[10]

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

Once the optimal density $f(\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0; \mathbf{v}^*)$ is obtained, it is used to construct the importance sampling estimator:

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

where $(\boldsymbol{\Sigma}_\theta^1, \boldsymbol{\Sigma}_h^1, \boldsymbol{\theta}_0^1, \mathbf{h}_0^1), \ldots, (\boldsymbol{\Sigma}_\theta^N, \boldsymbol{\Sigma}_h^N, \boldsymbol{\theta}_0^N, \mathbf{h}_0^N)$ are independent draws from the optimal importance sampling density $f(\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0; \mathbf{v}^*)$ and $p(\mathbf{y}|\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$ is the integrated likelihood, which can be estimated using the estimator in Equation 8.

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 reader to Chan and Eisenstat (2015) for a more thorough discussion. 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 R posterior draws and use them to solve the CE minimization problem in Equation 10 to obtain the optimal importance sampling density $f(\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0; \mathbf{v}^*)$.*
2. *For $j = 1, \ldots, N$, simulate $(\boldsymbol{\Sigma}_\theta^j, \boldsymbol{\Sigma}_h^j, \boldsymbol{\theta}_0^j, \mathbf{h}_0^j) \sim f(\boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0; \mathbf{v}^*)$ and compute the average:*

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

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 1.

Since Algorithm 3 has two nested importance sampling steps, it falls within the importance sampling squared (IS^2) framework in Tran, Scharth, Pitt, and Kohn (2014). Following their recommendation, the simulation size of the inner importance sampling loop—i.e., the importance sampling step for estimating the integrated likelihood—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).

5 | A MONTE CARLO STUDY

In this section we conduct a small Monte Carlo experiment to assess how the proposed algorithms perform in selecting the correct models. Specifically, from each of the three models—CVAR, CVAR-SV, and TVP-SV—we generate one dataset of $n = 3$ variables and $T = 300$ observations. For each dataset, we estimate the log marginal likelihoods and DICs for the same three models.

For CVAR, we set the diagonal elements of the error covariance matrix to be 0.5. The intercepts are generated uniformly from the set $\{-10, -9, \ldots, 0, \ldots, 9, 10\}$. The diagonal elements of the ith VAR coefficient matrix are independent $\mathcal{U}(-0.3/i, 0.3/i)$ and the off-diagonal elements are $\mathcal{U}(-0.1/i, 0.1/i)$ with $i = 1, \ldots, p$. The free elements of the impact matrix are generated independently from the standard normal distribution.

For CVAR-SV, the VAR coefficients are generated as in the CVAR case. We set the diagonal elements of $\boldsymbol{\Sigma}_h$ to be 0.01 and generate the log volatilities according to the state equation (Equation 2). Lastly, for TVP-SV, we set the diagonal elements of $\boldsymbol{\Sigma}_\theta$ to be 0.001 and generate the time-varying coefficients according to the state equation (Equation 4). The log volatilities are generated as in the CVAR-SV case.

Each log marginal likelihood estimate is based on 10,000 evaluations of the integrated likelihood, where the importance sampling density is constructed using 20,000 posterior draws after a burn-in period of 5,000. Each DIC estimate (and the corresponding numerical standard error) is computed using 10 parallel chains; each consists of 20,000 posterior draws after a burn-in period of 5,000. The integrated likelihood is evaluated every 20th post burn-in draw—a total of 10,000 evaluations. To calculate the plug-in estimate $D(\tilde{\boldsymbol{\psi}})$ in Equation 6, where $\tilde{\boldsymbol{\psi}}$ is the vector of posterior means, 500 draws are used for the integrated likelihood evaluation. The results are reported in Table 2.

In each case, both the log marginal likelihood and DIC are able to select the true data-generating process. In addition, both criteria penalize model complexity when it is not needed. For example, when the data are generated from the CVAR (second column of the table), both CVAR-SV and TVP-SV perform worse than the simpler homoskedastic CVAR.

TABLE 2 Log marginal likelihood and DIC estimates for three Monte Carlo experiments

	Log marginal likelihood		
	DGP1: CVAR	**DGP2: CVAR-SV**	**DGP3: TVP-SV**
CVAR	**−1042.8**	−1180.2	−1810.8
	(0.003)	(0.004)	(0.003)
CVAR-SV	−1059.6	**−1116.2**	−1729.4
	(0.004)	(0.004)	(0.01)
TVP-SV	−1130.4	−1201.3	**−1514.2**
	(0.08)	(0.04)	(0.18)
		DIC	
CVAR	**1913.4**	2186.0	3436.3
	(0.21)	(0.14)	(0.10)
CVAR-SV	1945.8	**2057.6**	3287.0
	(0.25)	(0.27)	(0.28)
TVP-SV	2099.3	2239.7	**2857.5**
	(0.62)	(0.33)	(0.51)

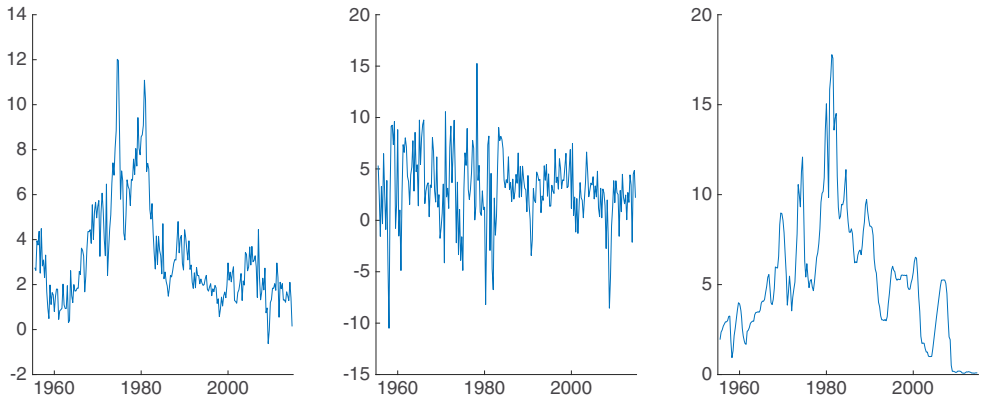

FIGURE 1 Plots of the GDP deflator inflation (left), real GDP growth (middle), and interest rate equation (right) [Colour figure can be viewed at wileyonlinelibrary.com]

6 | DATA AND EMPIRICAL RESULTS

In this section we compare a number of VARs that involve quarterly data on the gross domestic product (GDP) deflator, real GDP, and short-term interest rate for the USA. These three variables are commonly used in forecasting (e.g., Banbura, Giannone, & Reichlin, 2010; Koop, 2013) and small dynamic stochastic general equilibrium models (e.g., An & Schorfheide, 2007). The data on real GDP and the GDP deflator are sourced from the Federal Reserve Bank of St. Louis economic database. They are then transformed to annualized growth rates. The short-term interest rate is the effective Federal Funds rate, which is also obtained from the Federal Reserve Bank of St. Louis economic database. The sample period covers the quarters 1954:Q3 to 2014:Q4. The data are plotted in Figure 1.

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 "nonsystematic policy actions" that capture both policy mistakes and interest rate movements that are responses to variables other than inflation and GDP growth. In our baseline results we set the lag length to be $p = 2$.[11]

We compute the log marginal likelihoods and DICs for the competing models listed in Table 1. Each log marginal likelihood estimate is based on 10,000 evaluations of the integrated likelihood, where the importance sampling density is constructed using 20,000 posterior draws after a burn-in period of 5,000. Each DIC estimate (and the corresponding numerical standard error) is computed using 10 parallel chains; each consists of 20,000 posterior draws after a burn-in

[11] In Appendix C we report results with $p = 3$ lags. For all models, the additional lag makes both model selection criteria worse, suggesting that two lags seem to be sufficient.

TABLE 3 Log marginal likelihood and DIC estimates for various time-varying VARs (numerical standard errors in parentheses)

	TVP-SV	TVP	TVP-R1-SV	TVP-R2-SV	TVP-R3-SV	CVAR-SV	CVAR
log-ML	−1180.2	−1303.7	**−1171.0**	−1177.5	−1172.1	−1171.7	−1337.7
	(0.12)	(0.12)	(0.04)	(0.26)	(0.07)	(0.05)	(0.003)
DIC	2215.5	2400.4	2154.9	2202.8	2166.5	**2148.9**	2503.1
	(1.01)	(1.79)	(0.29)	(0.50)	(0.50)	(0.36)	(0.17)
p_D	29.9	29.2	31.9	28.7	30.0	31.9	26.8
	(0.26)	(0.86)	(0.15)	(0.21)	(0.50)	(0.25)	(0.08)

TABLE 4 Log marginal likelihood for various regime-switching VARs (numerical standard errors in parentheses)

	RS-VAR ($r = 2$)	RS-VAR-R1 ($r = 2$)	RS-VAR-R2 ($r = 2$)	RS-VAR ($r = 3$)	RS-VAR-R1 ($r = 3$)	RS-VAR-R2 ($r = 3$)
log-ML	−1277.5	**−1226.3**	−1293.5	−1289.4	−1231.1	−1324.6
	(0.06)	(0.01)	(0.03)	(0.04)	(0.14)	(1.73)

period of 5,000. The integrated likelihood is evaluated every 20th post burn-in draw—a total of 10,000 evaluations. To calculate the plug-in estimate $D(\tilde{\boldsymbol{\psi}})$ in Equation 6, where $\tilde{\boldsymbol{\psi}}$ is the vector of posterior means, 500 draws are used for the integrated likelihood evaluation. To give an indication of computation time, we implement the algorithms using MATLAB on a standard desktop with an Intel Core i5-4590S @ 3.0 GHz processor and 8 GB of RAM. It takes 2 minutes to compute the marginal likelihood for the CVAR-SV and 205 minutes for the TVP-SV.

The model comparison results are reported in Tables 3 and 4. 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:

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

where $p(\mathbf{y}_{t+1}|\mathbf{y}_{1:t}, M_k)$ is the *predictive likelihood* under model M_k. Each predictive likelihood $p(\mathbf{y}_{t+1}|\mathbf{y}_{1:t}, M_k)$ 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.[12] Using five independent runs, we obtain an estimate of −1171.6 with a numerical standard error of 0.71. This is essentially identical to our estimate of −1171.7 from the proposed importance sampling approach.

Table 3 reports the model comparison results for the time-varying VARs as well as the standard CVAR with constant coefficients and homoskedastic innovations. A few broad conclusions can be drawn from these results. Firstly, compared to the standard CVAR, the TVP-SV with both time-varying parameters and stochastic volatility is overwhelmingly favored by the data—for example, the Bayes factor in favor of the latter model is 2.5×10^{68}. 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.

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 4,900, indicating overwhelming support for the former model; the difference in DICs is 66.6 in favor of the former. In contrast to the findings in Koop et al. (2009), our results suggest that when stochastic volatility is allowed time variation in the VAR coefficients is not important in explaining the data.[13] 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.

[12]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.

[13]To check the robustness of this conclusion, in Appendix C we provide additional results using a more diffuse prior on $\boldsymbol{\Sigma}_\theta$, the error covariance matrix of the VAR coefficients. For all models with time-varying parameters, this more diffuse prior makes both model selection criteria worse, suggesting that the data do not favor substantial time variation in the VAR coefficients.

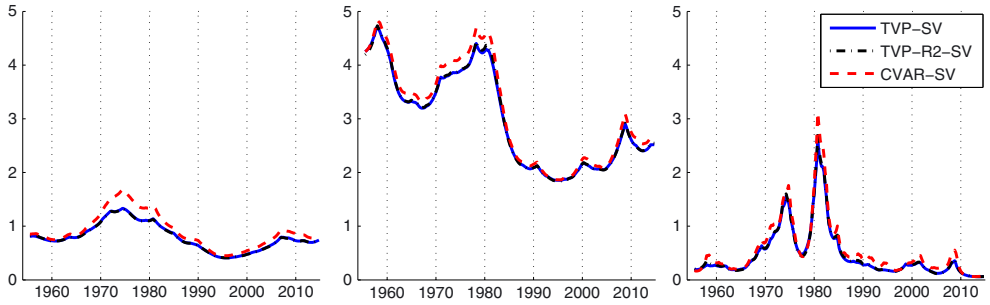

FIGURE 2 Estimated standard deviation of the innovation in the inflation equation (left), GDP growth equation (middle), and interest rate equation (right) [Colour figure can be viewed at wileyonlinelibrary.com]

The three restricted versions of TVP-SV, namely, TVP-R1-SV, TVP-R2-SV and TVP-R3-SV, all compare more favorably to the more general TVP-SV, but they receive similar support to CVAR-SV. Perhaps it is surprising that allowing for only time-varying intercepts, as in TVP-R3-SV, does not substantially improve model fit—even though one might expect structural breaks in mean for variables like inflation or interest rate. Indeed, there seems to be some time variation in the estimated intercepts (see Figure C1 in Appendix C). However, the associated credible intervals are relatively wide, so that one cannot draw a definitive conclusion.

Our results thus support the so-called "Sims critique" (see Sims, 2001) that the earlier finding of time variation in VAR coefficients was due to the failure to account for heteroskedasticity in a TVP model with a constant covariance matrix. 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 the first and the second models—the marginal likelihood slightly prefers TVP-R1-SV, whereas the DIC favors CVAR-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 et al. (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 (2017) and Chan and Eisenstat (2015), who find that a time-varying parameter VAR with constant variance compares favorably with a constant coefficients VAR.

In addition, Table 4 reports the marginal likelihood estimates for various regime-switching VARs.[14] Recall that RS-VAR denotes the model where both VAR coefficients and variances can differ across regimes, whereas RS-VAR-R1 and RS-VAR-R2 allow only variances or VAR coefficients to be different, respectively. Our results broadly support the conclusions in Sims and Zha (2006). In particular, according to the marginal likelihood, the top models are RS-VAR-R1 (with two and three regimes), followed by the more general RS-VAR. In other words, the VAR variances seem to be different across regimes but not the VAR coefficients.

Finally, comparing the two main classes of models—VARs with stochastic volatility and regime-switching VARs—we find strong evidence in favor of the former. In particular, the Bayes factor in favor of CVAR-SV against the best regime-switching VAR, RS-VAR-R1 with two regimes, is 5.2×10^{23}, indicating overwhelming support for the former model. Hence the data support the conclusion that the volatility process is better modeled as a random walk with a gradual drift instead of a process with discrete breaks.

We plot the posterior means of the standard deviations of the innovations for selected models in Figure 2. 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 homoskedastic 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

[14]We do not report DIC estimates for the regime-switching models because the posterior distributions under these models typically have multiple modes. As such, the choice of plug-in estimate in computing the DIC can be problematic.

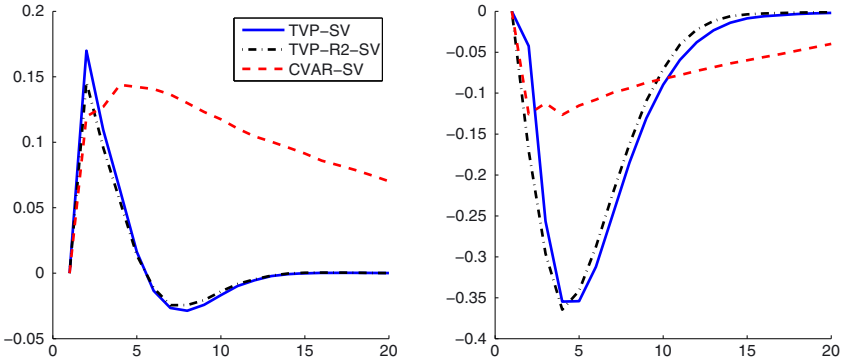

FIGURE 3 Impulse responses of inflation (left) and GDP growth (right) to monetary shock at 2014:Q4 [Colour figure can be viewed at wileyonlinelibrary.com]

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 3 we plot the impulse responses of inflation and GDP growth to a 1% monetary shock. In particular, a "positive" shock here refers to an increase in the policy rate. For the TVP models, the VAR coefficients used to compute the impulse responses are fixed at the 2014:Q4 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.

7 | EXTENSIONS

In this paper we have focused on TVP-VARs, but the proposed algorithms can be readily applied to other similar settings. Below we discuss a general framework in which the proposed algorithms are applicable. We then outline how the algorithms can be modified to fit that framework.

Let $\mathbf{y}_t$ denote a vector of observations at time t and let $\boldsymbol{\psi}$ represent a vector of time-invariant parameters. Consider a state-space model with two types of states: $\boldsymbol{\alpha}_t$ and λ_t. Specifically, suppose the measurement equation is characterized by the Gaussian density $p(\mathbf{y}|\boldsymbol{\alpha}_t, \lambda_t, \boldsymbol{\psi})$, where the equation is linear in $\boldsymbol{\alpha}_t$ but not in λ_t. The latent states in turn proceed from the following VAR(1) processes:

$$\boldsymbol{\alpha}_t = \Phi_\alpha \boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t^\alpha, \quad \boldsymbol{\eta}_t^\alpha \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\alpha), \tag{12}$$

$$\lambda_t = \Phi_\lambda \lambda_{t-1} + \eta_t^\lambda, \quad \eta_t^\lambda \sim \mathcal{N}(0, \Sigma_\lambda), \tag{13}$$

where the initial conditions $\boldsymbol{\alpha}_0$ and λ_0 are treated as unknown parameters and are included in $\boldsymbol{\psi}$.

This general framework naturally includes the TVP-VAR with VAR(1) state equations as a special case, in which $\boldsymbol{\alpha}_t = \boldsymbol{\theta}_t$ and $\lambda_t = \mathbf{h}_t$. Another example is a dynamic factor model with stochastic volatility. More specifically, consider

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{A}\mathbf{f}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t),$$

where μ is a vector of intercepts, $\boldsymbol{\Sigma}_t = \text{diag}(\exp(h_{1t}), \ldots, \exp(h_{nt}))$ is the time-varying covariance matrix, and $\mathbf{f}_t$ is a vector of latent factors. Further, the latent factors are assumed to proceed from the following VAR(1) process:

$$\mathbf{f}_t = \Phi_f \mathbf{f}_{t-1} + \boldsymbol{\eta}_t^f, \quad \boldsymbol{\eta}_t^f \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_f),$$

and the log volatilities $\mathbf{h} = (\mathbf{h}_1', \ldots, \mathbf{h}_T')'$ follow a random walk process as before. Then, this dynamic factor model can be written as a special case of the general framework with $\boldsymbol{\alpha}_t = \mathbf{f}_t$ and $\lambda_t = \mathbf{h}_t$.

Next, we outline how one can estimate the integrated likelihood implied by this more general framework. Given the VAR state equations 12 and 13, the prior densities of $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_T')'$ and $\lambda = (\lambda_1', \ldots, \lambda_T')'$, denoted by $p(\boldsymbol{\alpha}|\boldsymbol{\psi})$ and $p(\lambda|\boldsymbol{\psi})$, are both Gaussian. In addition, due to the VAR(1) structure, the precision matrices of the Gaussian densities are also banded.

Since $p(\mathbf{y}_t|\boldsymbol{\alpha}_t, \boldsymbol{\lambda}_t, \boldsymbol{\psi})$ is Gaussian and linear in $\boldsymbol{\alpha}_t$, one can compute analytically the (partial) integrated likelihood:

$$p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\psi}) = \int p(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\psi}) p(\boldsymbol{\alpha} \mid \boldsymbol{\psi}) \mathrm{d}\boldsymbol{\alpha}.$$

To integrate out $\boldsymbol{\lambda}$ by importance sampling, as before we can approximate the ideal zero-variance importance sampling density, in this case $p(\boldsymbol{\lambda}|\mathbf{y}, \boldsymbol{\psi})$, by a Gaussian approximation. That is, we use the EM algorithm to locate the mode of $\log p(\boldsymbol{\lambda} \mid \mathbf{y}, \boldsymbol{\psi})$, which is used as the mean vector of the Gaussian approximation. We then compute the associated negative Hessian evaluated at the mode, and use it as the precision matrix. Then Algorithm 1 can be implemented as before.

8 | 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 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 the DIC, a constant coefficients VAR with stochastic volatility receives similar support compared to the more general model of Primiceri.

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—for example, 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 et al. (2014) and Eisenstat, Chan, and Strachan (2016).

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, & 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 (2016).

ACKNOWLEDGMENTS

Joshua Chan would like to acknowledge financial support from 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.

OPEN RESEARCH BADGES

This article has earned an Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at [http://qed.econ.queensu.ca/jae/2018-v33.4/chan-eisenstat/].

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Clark, T. E., & Ravazzolo, F. (2015). Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *Journal of Applied Econometrics*, *30*(4), 551–575.

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

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

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

Del Negro, M., & Primiceri, G. E. (2015). Time-varying structural vector autoregressions and monetary policy: A corrigendum. *Review of Economic Studies*, *82*(4), 1342–1345.

Djegnéné, B., & McCausland, W. J. (2016). The HESSIAN method for models with leverage-like effects. *Journal of Financial Econometrics*, *13*(3), 722–755.

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

Eisenstat, E., Chan, J. C. C., & Strachan, R. W. (2016). Stochastic model specification search for time-varying parameter VARs. *Econometric Reviews*, *35*(8–10), 1638–1665.

Eisenstat, E., & Strachan, R. W. (2016). Modelling inflation volatility. *Journal of Applied Econometrics*, *31*(5), 805–820.

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

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

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

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

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

Grant, A. L. (2017). The early millennium slowdown: Replicating the Peersman (2005) results. *Journal of Applied Econometrics*, *32*(1), 224–232.

Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions.*Technometrics*, *37*(2), 185–194.

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

Koop, G. (2003). *Bayesian econometrics*. New York, NY: Wiley.

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

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

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

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

Kroese, D. P., & Chan, J. C. C. (2014). *Statistical modeling and computation*. New York, NY: Springer.

Kroese, D. P., Taimre, T., & Botev, Z. I. (2011). *Handbook of Monte Carlo methods*. New York, NY: Wiley.

Li, Y., Zeng, T., & Yu, J. (2012). Robust deviance information criterion for latent variable models. (SMU Economics and Statistics Working Paper Series), Singapore Mangement University, Singapore.

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

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

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

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

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

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

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

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

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

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

Rubinstein, R. Y., & Kroese, D. P. (2004). *The cross-entropy method: A unified approach to combinatorial optimization Monte-Carlo simulation, and machine learning*. New York, NY: Springer.

Sims, C. A. (2001). Comment on Sargent and Cogley's "evolving US postwar inflation dynamics". *NBER Macroeconomics Annual*, *16*, 373–379.

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

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

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

Uhlig, H. (1997). Bayesian vector autoregressions with stochastic volatility. *Econometrica*, *65*(1), 59–73.

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

How to cite this article: Chan JCC, Eisenstat E. Bayesian model comparison for time-varying parameter VARs with stochastic volatility. *J Appl Econ*. 2018;33:509–532. https://doi.org/10.1002/jae.2617

APPENDIX A: ESTIMATION DETAILS

In this Appendix we outline the estimation details for fitting the model in Equations (2)–(4) and other restricted models.

A.1 | Estimation of TVP-SV

For notational convenience, let $\mathbf{y} = (\mathbf{y}_1', \dots, \mathbf{y}_T')'$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \dots, \boldsymbol{\theta}_T')'$. 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 Equation 3 as a seemingly unrelated regression:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{A1}$$

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

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

Then, we can rewrite Equation 4 as

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

where $\tilde{\boldsymbol{\alpha}}_\theta = (\boldsymbol{\theta}_0', \mathbf{0}, \ldots, \mathbf{0})'$ 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 $\boldsymbol{\alpha}_\theta = \mathbf{H}_\theta^{-1} \tilde{\boldsymbol{\alpha}}_\theta$. Using standard linear regression results, one can show that (see e.g., Kroese & Chan, 2014), Corollary 8.1):

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

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

$$\mathbf{K}_\theta = \mathbf{H}_\theta' \mathbf{S}_\theta^{-1} \mathbf{H}_\theta + \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X}, \quad \mathbf{d}_\theta = \mathbf{H}_\theta' \mathbf{S}_\theta^{-1} \mathbf{H}_\theta \boldsymbol{\alpha}_\theta + \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{y}. \tag{A2}$$

Note that the precision matrix $\mathbf{K}_\theta$ is a band matrix—that is, 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 $\mathcal{N}(\hat{\boldsymbol{\theta}}, \mathbf{K}_\theta^{-1})$ efficiently.

To implement step 2, we can apply the auxiliary mixture sampler of Kim et al. (1998) in conjunction with 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:

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

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

A.2 | Estimation of other restricted models

Here we outline the estimation of various restricted versions of TVP-SV. To start, consider TVP-SV-R1 with the restrictions $\boldsymbol{\beta}_t = \boldsymbol{\beta}_0$ for $t = 1, \ldots, T$. Then, the model can be written as

$$\mathbf{y}_t = \tilde{\mathbf{X}}_t \boldsymbol{\beta}_0 + \mathbf{W}_t \boldsymbol{\gamma}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t),$$

where $\tilde{\mathbf{X}}_t = \mathbf{I}_n \otimes (1, \mathbf{y}_{t-1}', \ldots, \mathbf{y}_{t-p}')$ and $\mathbf{W}_t$ is an $n \times k_\gamma$ matrix that contains appropriate elements of $-\mathbf{y}_t$ as defined in the

main text. Recall that the prior on $\theta_0 = (\beta_0', \gamma_0')'$ is $\theta_0 \sim \mathcal{N}(\mathbf{a}_\theta, \mathbf{V}_\theta)$. Let $\mathbf{a}_\beta$ and $\mathbf{a}_\gamma$ denote respectively the prior means of β_0 and γ_0. Similarly, define $\mathbf{V}_\beta$ and $\mathbf{V}_\gamma$.

Then, the posterior sampler for TVP-SV can be modified to fit this restricted version. More specifically, we obtain posterior draws by sequentially drawing from

1. $p(\gamma|\mathbf{y}, \mathbf{h}, \Sigma_\gamma, \Sigma_h, \beta_0, \gamma_0, \mathbf{h}_0)$;
2. $p(\mathbf{h}|\mathbf{y}, \gamma, \Sigma_\gamma, \Sigma_h, \beta_0, \gamma_0, \mathbf{h}_0)$;
3. $p(\Sigma_\gamma, \Sigma_h|\mathbf{y}, \gamma, \mathbf{h}, \beta_0, \gamma_0, \mathbf{h}_0)$;
4. $p(\beta_0, \gamma_0, \mathbf{h}_0|\mathbf{y}, \gamma, \mathbf{h}, \Sigma_\gamma, \Sigma_h)$.

To implement step 1, define $\mathbf{W} = \text{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_T)$ and $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1', \ldots, \tilde{\mathbf{X}}_T')'$, $\tilde{\alpha}_\gamma = (\gamma_0', \mathbf{0}, \ldots, \mathbf{0})'$ and $\mathbf{S}_\gamma = \mathbf{I}_T \otimes \Sigma_\gamma$. Further, let $\mathbf{H}_\gamma$ denote the first difference matrix of appropriate dimension. Then, using a similar derivation as in the previous section, one can show that

$$(\gamma\,|\,\mathbf{y}, \mathbf{h}, \Sigma_\gamma, \Sigma_h, \beta_0, \gamma_0, \mathbf{h}_0) \sim \mathcal{N}(\hat{\gamma}, \mathbf{K}_\gamma^{-1}),$$

where

$$\mathbf{K}_\gamma = \mathbf{H}_\gamma' \mathbf{S}_\gamma^{-1} \mathbf{H}_\gamma + \mathbf{W}' \Sigma^{-1} \mathbf{W}, \quad \hat{\gamma} = \mathbf{K}_\gamma^{-1} \left(\mathbf{H}_\gamma' \mathbf{S}_\gamma^{-1} \mathbf{H}_\gamma \alpha_\gamma + \mathbf{W}' \Sigma^{-1} (\mathbf{y} - \tilde{\mathbf{X}} \beta_0) \right).$$

We then use the precision sampler of Chan and Jeliazkov (2009) to sample from this Gaussian distribution.

Steps 2 and 3 can be carried out similarly as before. For step 4, first note that β_0, γ_0, and $\mathbf{h}_0$ are conditionally independent given the data and other parameters.

In fact, they are conditionally independent Gaussian random vectors:

$$(\beta_0\,|\,\mathbf{y}, \gamma, \mathbf{h}, \Sigma_\gamma, \Sigma_h) \sim \mathcal{N}(\hat{\beta}_0, \mathbf{K}_{\beta_0}^{-1}),$$
$$(\gamma_0,\,|\,\mathbf{y}, \gamma, \mathbf{h}, \Sigma_\gamma, \Sigma_h) \sim \mathcal{N}(\hat{\gamma}_0, \mathbf{K}_{\gamma_0}^{-1}),$$
$$(\mathbf{h}_0\,|\,\mathbf{y}, \gamma, \mathbf{h}, \Sigma_\gamma, \Sigma_h) \sim \mathcal{N}(\hat{\mathbf{h}}_0, \mathbf{K}_{\mathbf{h}_0}^{-1}),$$

where $\mathbf{K}_{\beta_0} = \mathbf{V}_\beta^{-1} + \tilde{\mathbf{X}}' \Sigma^{-1} \tilde{\mathbf{X}}$, $\hat{\beta}_0 = \mathbf{K}_{\beta_0}^{-1} (\mathbf{V}_\beta^{-1} \mathbf{a}_\beta + \tilde{\mathbf{X}}' \Sigma^{-1} (\mathbf{y} - \mathbf{W}\gamma))$, $\mathbf{K}_{\gamma_0} = \mathbf{V}_\gamma^{-1} + \Sigma_\gamma^{-1}$, $\hat{\gamma}_0 = \mathbf{K}_{\gamma_0}^{-1} (\mathbf{V}_\gamma^{-1} \mathbf{a}_\gamma + \Sigma_\gamma^{-1} \gamma_1)$, $\mathbf{K}_{\mathbf{h}_0} = \mathbf{V}_h^{-1} + \Sigma_h^{-1}$ and $\hat{\mathbf{h}}_0 = \mathbf{K}_{\mathbf{h}_0}^{-1} (\mathbf{V}_h^{-1} \mathbf{a}_h + \Sigma_h^{-1} \mathbf{h}_1)$.

For TVP-SV-R2 with the restrictions $\gamma_t = \gamma_0$ for $t = 1, \ldots, T$, we can write the model as

$$\mathbf{y}_t = \tilde{\mathbf{X}}_t \beta_t + \mathbf{W}_t \gamma_0 + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(\mathbf{0}, \Sigma_t).$$

Hence, compared to TVP-SV-R1, the roles of β_t and γ_t are now swapped, and we can use the same Gibbs sampler to estimate TVP-SV-R2.

For CVAR-SV, we impose the restrictions $\beta_t = \beta_0$ and $\gamma_t = \gamma_0$ for $t = 1, \ldots, T$. The model is therefore

$$\mathbf{y}_t = \tilde{\mathbf{X}}_t \beta_0 + \mathbf{W}_t \gamma_0 + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(\mathbf{0}, \Sigma_t). \tag{A3}$$

We can then obtain posterior draws by sequentially drawing from

1. $p(\mathbf{h}|\mathbf{y}, \gamma, \Sigma_\gamma, \Sigma_h, \beta_0, \gamma_0, \mathbf{h}_0)$;
2. $p(\Sigma_h|\mathbf{y}, \mathbf{h}, \beta_0, \gamma_0, \mathbf{h}_0)$;
3. $p(\beta_0, \gamma_0, \mathbf{h}_0|\mathbf{y}, \mathbf{h}, \Sigma_h)$.

Steps 1 and 2 can be carried similarly as before. For step 3, first note that $\theta_0 = (\beta_0', \gamma_0')'$ and $\mathbf{h}_0$ are conditionally independent given the data and other parameters. We can sample $\mathbf{h}_0$ from its conditional distribution as before. For sampling θ_0, we first stack A3 over $t = 1, \ldots, T$:

$$\mathbf{y} = \mathbf{Z}\theta_0 + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(\mathbf{0}, \Sigma_t),$$

where

$$\mathbf{Z} = \begin{pmatrix} \tilde{\mathbf{X}}_1 & \mathbf{W}_1 \\ \vdots & \vdots \\ \tilde{\mathbf{X}}_T & \mathbf{W}_T \end{pmatrix}.$$

With the prior $\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{a}_\theta, \mathbf{V}_\theta)$, the conditional distribution of $\boldsymbol{\theta}_0$ is therefore

$$(\boldsymbol{\theta}_0 \,|\, \mathbf{y}, \mathbf{h}, \boldsymbol{\Sigma}_h) \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_0, \mathbf{K}_{\theta_0}^{-1}),$$

where $\mathbf{K}_{\theta_0} = \mathbf{V}_\theta^{-1} + \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}$ and $\hat{\boldsymbol{\theta}}_0 = \mathbf{K}_{\theta_0}(\mathbf{V}_\theta^{-1}\mathbf{a}_\theta + \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{y})$. For other simpler restricted models, we only need to modify the above algorithms to fit them.

APPENDIX B: DETAILS ON INTEGRATED LIKELIHOOD ESTIMATION

In this Appendix we provide the technical details for estimating the integrated likelihood outlined in Section 4.1. Recall that the integrated likelihood estimation consists of two steps. In the first step we integrate out $\boldsymbol{\theta}$ analytically:

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

In the second step, we integrate out $\mathbf{h}$ using importance sampling.

B.1 | Analytical integration with respect to $\boldsymbol{\theta}$

To implement the first step, 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 $\boldsymbol{\theta}$:

$$p(\mathbf{y} \,|\, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) = \int p(\mathbf{y} \,|\, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) p(\boldsymbol{\theta} \,|\, \boldsymbol{\Sigma}_\theta, \boldsymbol{\theta}_0) \mathrm{d}\boldsymbol{\theta}.$$

We showed in Appendix A that $(\mathbf{y} \,|\, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) = (\mathbf{y} \,|\, \mathbf{h}, \boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta}, \boldsymbol{\Sigma})$ 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 to that in Chan and Grant (2016a), one can obtain the log density as follows:

$$\begin{aligned} \log p(\mathbf{y} \,|\, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0) = &-\frac{Tn}{2}\log(2\pi) - \frac{1}{2}\mathbf{1}_{nT}'\mathbf{h} - \frac{T}{2}\log|\boldsymbol{\Sigma}_\theta| - \frac{1}{2}\log|\mathbf{K}_\theta| \\ &- \frac{1}{2}\left(\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y} + \boldsymbol{\alpha}_\theta'\mathbf{H}_\theta'\mathbf{S}_\theta^{-1}\mathbf{H}_\theta\boldsymbol{\alpha}_\theta - \mathbf{d}_\theta'\mathbf{K}_\theta^{-1}\mathbf{d}_\theta\right), \end{aligned} \tag{B1}$$

where $\mathbf{1}_{nT}$ is a $Tn \times 1$ column of ones, $\mathbf{K}_\theta$ and $\mathbf{d}_\theta$ are given in Equation A2 in Appendix A. This expression can also be derived using the identity

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

and setting $\boldsymbol{\theta} = \mathbf{0}$.

Since $\mathbf{K}_\theta$, $\boldsymbol{\Sigma}$, $\mathbf{H}_\theta$ and $\mathbf{S}_\theta$ are all band matrices, the expression in Equation B1 can be evaluated quickly; see Chan and Grant (2016a) for computational details.

B.2 | Importance sampling for integrating out h

Next, we discuss the second step that integrates out $\mathbf{h}$ using importance sampling. The ideal zero-variance importance sampling density in this case is the marginal density of$\mathbf{h}$ unconditional on $\boldsymbol{\theta}$—that is, $p(\mathbf{h}|\mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$. But this density cannot be used as an importance sampling density because its normalization constant is unknown. We therefore approximate it using a Gaussian density, which is then used as the importance sampling density.

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

To that end, we first use the EM algorithm to find the maximum of the log marginal density $\log p(\mathbf{h} \,|\, \mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \boldsymbol{\theta}_0, \mathbf{h}_0)$.

To implement the E-step, we compute the following conditional expectation:

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

where the expectation is taken with respect to $p(\boldsymbol{\theta}\,|\,\mathbf{y}, \tilde{\mathbf{h}}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \theta_0, \mathbf{h}_0)$ for an arbitrary vector $\tilde{\mathbf{h}}$. As shown in Appendix A,

$$(\boldsymbol{\theta}\,|\,\mathbf{y}, \tilde{\mathbf{h}}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \theta_0, \mathbf{h}_0) \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, \mathbf{K}_\theta^{-1}),$$

where $\hat{\boldsymbol{\theta}} = \mathbf{K}_\theta^{-1}\mathbf{d}_\theta$, and $\mathbf{d}_\theta$ and $\mathbf{K}_\theta$ are given in Equation A2. Note that both the mean vector $\hat{\boldsymbol{\theta}}$ and precision matrix $\mathbf{K}_\theta$ are functions of $\tilde{\mathbf{h}}$—that is, $\hat{\boldsymbol{\theta}}$ and $\mathbf{K}_\theta$ are computed using $\tilde{\mathbf{h}}$.

Then, an explicit expression of $\mathcal{Q}(\mathbf{h}\,|\,\tilde{\mathbf{h}})$ can be derived as follows:

$$\begin{aligned}\mathcal{Q}(\mathbf{h}\,|\,\tilde{\mathbf{h}}) &= -\frac{1}{2}(\mathbf{h}-\boldsymbol{\alpha}_h)'\mathbf{H}_h'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}_h^{-1})\mathbf{H}_h(\mathbf{h}-\boldsymbol{\alpha}_h) - \frac{1}{2}\mathbf{1}_{nT}'\mathbf{h} \\ &\quad - \frac{1}{2}\mathrm{tr}\left(\mathrm{diag}(\mathrm{e}^{-\mathbf{h}})\mathbb{E}_{\theta|\tilde{\mathbf{h}}}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})'\right]\right) + c_1 \\ &= -\frac{1}{2}(\mathbf{h}-\boldsymbol{\alpha}_h)'\mathbf{H}_h'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}_h^{-1})\mathbf{H}_h(\mathbf{h}-\boldsymbol{\alpha}_h) - \frac{1}{2}\mathbf{1}_{nT}'\mathbf{h} \\ &\quad - \frac{1}{2}\mathrm{tr}\left(\mathrm{diag}(\mathrm{e}^{-\mathbf{h}})\left(\mathbf{X}\mathbf{K}_\theta^{-1}\mathbf{X}' + (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\theta}})(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\theta}})'\right)\right) + c_1,\end{aligned} \tag{B2}$$

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

In the M-step, we maximize the function $\mathcal{Q}(\mathbf{h}\,|\,\tilde{\mathbf{h}})$ 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

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

and the Hessian is

$$\mathbf{H}_\mathcal{Q} = -\mathbf{H}_h'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}_h^{-1})\mathbf{H}_h - \frac{1}{2}\mathrm{diag}\left(\mathrm{e}^{-\mathbf{h}} \odot \hat{\mathbf{z}}\right), \tag{B3}$$

where $\odot$ denotes the entry-wise product, $\hat{\mathbf{z}} = (s_1^2 + \hat{\varepsilon}_1^2, \ldots, s_{nT}^2 + \hat{\varepsilon}_{nT}^2)'$, s_i^2 is the ith diagonal element of $\mathbf{X}\mathbf{K}_\theta^{-1}\mathbf{X}'$ and $\hat{\varepsilon}_i$ is the ith element of $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}$.

We emphasize that both $\mathbf{g}_\mathcal{Q}$ and $\mathbf{H}_\mathcal{Q}$ can be computed efficiently using sparse and band matrix algorithms.[15] 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.

Given the E- and M-steps above, 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 jth iteration, we obtain $\mathbf{g}_\mathcal{Q}$ and $\mathbf{H}_\mathcal{Q}$, where both $\hat{\boldsymbol{\theta}}$ and $\mathbf{K}_\theta$ are evaluated using $\mathbf{h}^{(j-1)}$. Then, we compute

$$\mathbf{h}^{(j)} = \arg\max_{\mathbf{h}} \mathcal{Q}(\mathbf{h}\,|\,\mathbf{h}^{(j-1)}),$$

using the Newton–Raphson method. We repeat the E- and M-steps until some convergence criterion is met—for example, 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 density $p(\mathbf{h}|\mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \theta_0, \mathbf{h}_0)$, which is denoted by $\hat{\mathbf{h}}$. We summarize the EM algorithm in Algorithm 4.

Algorithm 4. (EM algorithm to obtain the mode of $p(\mathbf{h}\,|\,\mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \theta_0, \mathbf{h}_0)$)
Suppose we have an initial guess $\boldsymbol{h}^{(0)}$ 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 $j = 1, 2, \ldots$:

[15] 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 $\mathbf{U}$.

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

Computing the Hessian of $\log p(\mathbf{h}\,|\,\mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \theta_0, \mathbf{h}_0)$

After obtaining the mode $\hat{\mathbf{h}}$ of the log marginal density $\log p(\mathbf{h}\,|\,\mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \theta_0, \mathbf{h}_0)$, next we compute the Hessian evaluated at $\hat{\mathbf{h}}$. If we take the log of both sides of Equation 9 and then take the expectation with respect to $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \theta_0, \mathbf{h}_0)$, we obtain the identity

$$\log p(\mathbf{h}\,|\,\mathbf{y}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \theta_0, \mathbf{h}_0) = \mathcal{Q}(\mathbf{h}\,|\,\mathbf{h}) + \mathcal{H}(\mathbf{h}\,|\,\mathbf{h}), \tag{B4}$$

where $\mathcal{H}(\mathbf{h}\,|\,\mathbf{h}) = -\mathbb{E}_{\theta|\mathbf{h}}\left[\log p(\boldsymbol{\theta}\,|\,\mathbf{y}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \theta_0, \mathbf{h}_0)\right]$. It follows that the Hessian of the log marginal density evaluated at$\hat{\mathbf{h}}$ is simply the sum of the Hessians of $\mathcal{Q}$ and $\mathcal{H}$ with $\mathbf{h} = \hat{\mathbf{h}}$. The former comes out as a by-product of the EM algorithm; an analytical expression is given in Equation (B3). The latter is computed below.

To that end, we first derive an explicit expression of $\mathcal{H}(\mathbf{h}\,|\,\mathbf{h})$:

$$\begin{aligned}
\mathcal{H}(\mathbf{h}\,|\,\mathbf{h}) &= -\mathbb{E}_{\theta|\mathbf{h}}\left[\log p(\boldsymbol{\theta}\,|\,\mathbf{y}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \theta_0, \mathbf{h}_0)\right] \\
&= \frac{kT}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_\theta| + \frac{1}{2}\mathbb{E}_{\theta|\mathbf{h}}\left[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\mathbf{K}_\theta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right] \\
&= -\frac{1}{2}\log|\mathbf{X}'\mathrm{diag}(\mathrm{e}^{-\mathbf{h}})\mathbf{X} + \mathbf{H}_\theta'\mathbf{S}_\theta^{-1}\mathbf{H}_\theta| + c_2,
\end{aligned}$$

where c_2 is a constant not dependent on $\mathbf{h}$. Note that under $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{h}, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_h, \theta_0, \mathbf{h}_0)$, the quadratic form $(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\mathbf{K}_\theta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$ is a chi-squared random variable and its expectation does not depend on $\mathbf{h}$.

To compute the Hessian of $\mathcal{H}$, we first introduce some notations. Let $\mathbf{x}_i$ be a $kT \times 1$ vector that consists of the elements in the ith row of $\mathbf{X}$. With a slight abuse of notations, we let h_i denote the ith 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 $\mathbf{Z} = \mathrm{diag}(\mathrm{e}^{-\mathbf{h}})\mathbf{X}\mathbf{K}_\theta^{-1}\mathbf{X}'$.

Finally, let $\mathbf{H}_\mathcal{Q}$ denote the Hessian of $\mathcal{Q}(\mathbf{h}\,|\,\mathbf{h})$ evaluated at $\mathbf{h} = \hat{\mathbf{h}}$. Then, the negative Hessian of the log marginal density of $\mathbf{h}$ evaluated at $\mathbf{h} = \hat{\mathbf{h}}$ is simply $\mathbf{K}_h = -(\mathbf{H}_\mathcal{Q} + \mathbf{H}_\mathcal{H})$, which is used as the precision matrix of the Gaussian approximation.

APPENDIX C: ADDITIONAL RESULTS

C.1 | Prior sensitivity analysis

In this section we provide additional results using a different set of priors to assess our main conclusion that when stochastic volatility is allowed time variation in the VAR coefficients is not important in explaining the data.

In particular, we consider larger prior means for the error variance of VAR coefficients $\mathbf{\Sigma}_\theta$. Recall that in the baseline results we set the hyperparameters so that the prior mean of $\sigma_{\theta i}^2$ is 0.01^2 if it is associated with a VAR coefficient and 0.1^2

TABLE C1 Log marginal likelihood and DIC estimates for various time-varying VARs (numerical standard errors in parentheses) under an alternative prior on the error variance of VAR coefficients

	TVP-SV	TVP	TVP-R1-SV	TVP-R2-SV	TVP-R3-SV	CVAR-SV	CVAR
log-ML	−1385.5	−1531.8	−1216.6	−1352.8	−1224.4	−1171.7	−1337.7
	(0.22)	(0.03)	(0.11)	(0.29)	(0.25)	(0.05)	(0.003)
DIC	2600.4	2819.6	2249.3	2537.4	2262.7	2148.9	2503.1
	(0.52)	(0.43)	(0.82)	(0.73)	(0.99)	(0.36)	(0.17)
p_D	26.7	28.6	33.7	27.4	35.2	31.9	26.8
	(0.47)	(0.15)	(0.75)	(0.45)	(0.90)	(0.25)	(0.08)

TABLE C2 Log marginal likelihood and DIC estimates for various time-varying VARs (numerical standard errors in parentheses) with $p = 3$

	TVP-SV	TVP	TVP-R1-SV	TVP-R2-SV	TVP-R3-SV	CVAR-SV	CVAR
log-ML	−1213.3	−1322.6	−1193.7	−1210.4	−1197.4	−1193.8	−1349.9
	(0.20)	(0.18)	(0.07)	(0.16)	(0.09)	(0.04)	(0.004)
DIC	2235.1	2390.9	2144.1	2223.3	2163.5	2136.1	2474.8
	(0.59)	(2.90)	(0.36)	(0.70)	(1.05)	(0.35)	(0.04)
p_D	39.7	36.0	41.6	39.3	42.3	40.7	35.7
	(0.36)	(0.86)	(0.26)	(0.55)	(0.95)	(0.22)	(0.02)

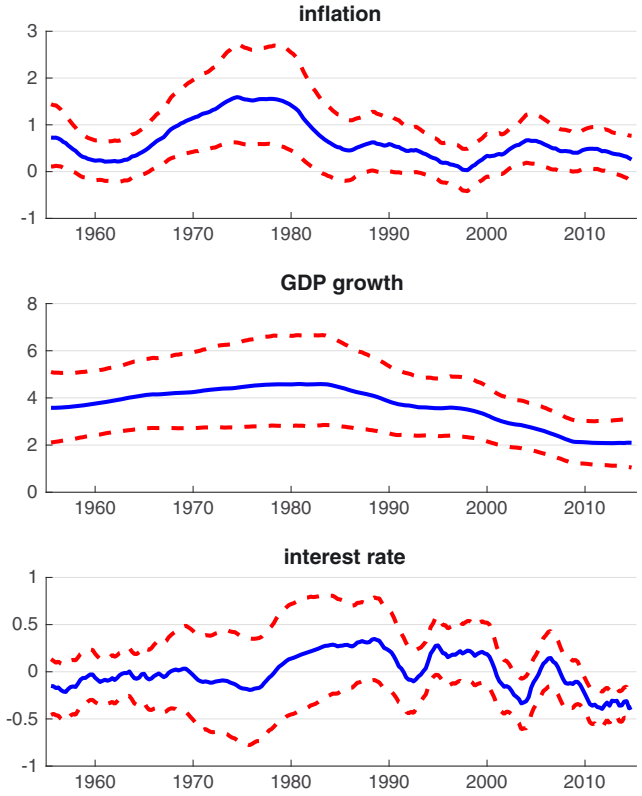

FIGURE C1 Estimated time-varying intercepts under TVP-R3-SV for the inflation equation (top), the GDP growth equation (middle), and the interest rate equation (bottom). The dotted lines denote 90% credible intervals [Colour figure can be viewed at wileyonlinelibrary.com]

for an intercept. Here the hyperparameters are chosen so that the prior means are 0.1^2 and 1^2, respectively. All other priors are the same as in Section 2.

The results are reported in Table C1 (the values for CVAR-SV and CVAR are the same as before). For all models with time-varying parameters, this alternative prior makes both model selection criteria worse, suggesting that the data do not favor substantial time-variation in the VAR coefficients.

C.2 | Results from VARs with $p = 3$ lags

In the baseline results reported in the main text, we set the number of lags to be two. Here we present the log marginal likelihood and DIC estimates for VARs with three lags in Table C2. For all models, the additional lag makes both model selection criteria worse, suggesting that two lags seem to be sufficient.

C.3 | Results from TVP-R3-SV

Figure C1 plots the posterior means of the time-varying intercepts under the TVP-R3-SV model as well as the corresponding 90% credible intervals.