On the Observed-Data Deviance Information Criterion for Volatility Modeling*

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

Angelia L. Grant
Research School of Economics
Australian National University

February 2016

Abstract

We propose importance sampling algorithms based on fast band matrix routines for estimating the observed-data likelihoods for a variety of stochastic volatility models. This is motivated by the problem of computing the deviance information criterion (DIC)—a popular Bayesian model comparison criterion that comes in a few variants. While the DIC based on the conditional likelihood—obtained by conditioning on the latent variables—is widely used for comparing stochastic volatility models, recent studies have argued against its use on both theoretical and practical grounds. Indeed, we show via a Monte Carlo study that the conditional DIC tends to favor overfitted models, whereas the DIC based on the observed-data likelihood—calculated using the proposed importance sampling algorithms—seems to perform well. We demonstrate the methodology with an application involving daily returns on the Standard & Poors (S&P) 500 index.

Keywords: Bayesian model comparison, nonlinear state space, DIC, jumps, moving average, leverage, heavy tails, S&P 500

JEL classification: C11, C15, C52, C58

*A previous version of this paper was circulated under the title "Issues in Comparing Stochastic Volatility Models Using the Deviance Information Criterion". We thank the Editor Federico Bandi for his insightful comments on repositioning the paper, and an Associate Editor and two anonymous referees for their many detailed suggestions that have greatly improved the paper. We also thank seminar participants at the Melbourne Bayesian Econometrics Workshop for helpful comments and suggestions. All remaining errors are, of course, our own. Financial support from the Australian Research Council via a Discovery Early Career Researcher Award (DE150100795) is gratefully acknowledged. Email addresses: joshua.chan@anu.edu.au and angelia.grant@anu.edu.au.

1 Introduction

Stochastic volatility models are widely used for modeling financial time series, and have more recently become important in macroeconometric modeling following the seminal work of Cogley and Sargent (2005) and Primiceri (2005). As a result, there is now a large and growing family of flexible stochastic volatility models.[1] Given the wide range of model candidates, it has become increasingly important to be able to discriminate between competing models for a given application.

One popular metric for Bayesian model comparison is the deviance information criterion (DIC) proposed by Spiegelhalter, Best, Carlin, and van der Linde (2002). For latent variable models, Celeux, Forbes, Robert, and Titterington (2006) point out that there are numerous alternative definitions of the DIC depending on different concepts of the likelihood. In particular, the DIC based on the conditional likelihood—obtained by conditioning on the latent variables—has been widely used for comparing stochastic volatility models due to its easy computation and its implementation in standard statistical packages, including WinBUGS.[2] In contrast, the DIC based on the observed-data likelihood—obtained by integrating out the latent variables—is rarely used as the observed-data likelihoods for stochastic volatility models are generally difficult to evaluate.

We propose importance sampling algorithms—based on fast band matrix routines—for evaluating the observed-data likelihoods under a variety of stochastic volatility models, with the aim of obtaining observed-data DICs for these models. This is motivated by recent studies that argue against the use of the conditional DIC on both theoretical and practical grounds. Li, Zeng, and Yu (2012) argue that 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) provides a Monte Carlo study using Poisson models in which the conditional DIC almost always favors an overfitted model. Using examples that involve macroeconomic and financial data, Chan and Grant (2014) show that the numerical standard errors of the conditional DICs are typically too large to be useful for comparing models.

A key feature of our approach is that it draws on recent advances in band matrix algorithms rather than using the conventional Kalman filter. This approach builds upon earlier work on Markov chain Monte Carlo (MCMC) algorithms for linear Gaussian state space models (Rue, 2001; Chan and Jeliazkov, 2009; McCausland, Miller, and Pelletier, 2011) and various nonlinear state space models (McCausland, 2012; Chan, Koop, and Potter, 2013; Djegnéné and McCausland, 2014). Instead of posterior simulation, we construct efficient importance sampling estimators for the observed-data likelihood. In this

[1] See, e.g., Chib, Nardari, and Shephard (2002), Koopman and Hol Uspensky (2002), Jensen and Maheu (2010), Nakajima and Omori (2012), Chan (2013), Mumtaz and Zanetti (2013), Eisenstat and Strachan (2014) and Carriero, Clark, and Marcellino (2015), to name but a few examples.

[2] This version of the DIC has been used to compare a wide variety of stochastic volatility models in empirical applications; recent studies include Berg, Meyer, and Yu (2004), Yu and Meyer (2006), Abanto-Valle, Bandyopadhyay, Lachos, and Enriquez (2010), Vo (2011), Mumtaz and Surico (2012), Tsiotas (2012), Brooks and Prokopczuk (2013) and Wang, Choy, and Chan (2013).

paper we focus on stochastic volatility models, but the proposed approach is applicable more broadly to general nonlinear state space models. Using these importance sampling estimators, we show that the observed-data DIC can be accurately estimated. This therefore extends our earlier work (Chan and Grant, 2014) on evaluating the observed-data DIC for linear Gaussian state space models to nonlinear settings.

We show in a Monte Carlo study that the conditional DIC tends to prefer overfitted stochastic volatility models. This is an important finding given that the conditional DIC is widely used in empirical applications. In contrast, the observed-data DIC based on the proposed importance sampling estimators seems to be able to select the correct model. This result is not surprising as standard asymptotic arguments for justifying the DIC apply to the observed-data DIC.

In the empirical application that involves daily returns on the S&P 500, we find that according to the observed-data DIC, the leverage effect, t innovations, volatility feedback and moving average components all seem to be useful additions to the standard stochastic volatility model. The same conclusion holds if the marginal likelihood is used as the model selection criterion.

The rest of this paper is organized as follows. Section 2 first discusses the marginal likelihood and then outlines two definitions of the DIC. In Section 3 we discuss various stochastic volatility models that are widely used in the literature and their estimation. In Section 4 we propose importance sampling algorithms for estimating the observed-data likelihoods for stochastic volatility models. The proposed methods are demonstrated via a Monte Carlo study in Section 5. Moreover, the behavior of the conditional and observed-data DICs are examined. Section 6 illustrates the methodology with an application involving daily returns on the S&P 500. Further applications and directions for future research are discussed in Sections 7 and 8.

2 Bayesian Model Comparison Criteria

In this section we give an overview of two popular Bayesian model comparison criteria—the marginal likelihood and the deviance information criterion. To set the stage, suppose we wish to compare a collection of models $\{M_1, \ldots, M_K\}$, where each model M_k is formally defined by a likelihood function $p(\mathbf{y} \mid \boldsymbol{\theta}_k, M_k)$ and a prior on the model-specific parameter vector $\boldsymbol{\theta}_k$ denoted by $p(\boldsymbol{\theta}_k \mid M_k)$. A natural Bayesian model comparison criterion is the *marginal likelihood*, defined as:

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

The marginal likelihood can be interpreted as a joint 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. To see this,

arrange the data as $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_T')'$ and 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 as follows:

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

where $p(\mathbf{y}_{t+1} \mid \mathbf{y}_{1:t}, M_k)$ is the *predictive likelihood*, which is basically a one-step-ahead density forecast for $\mathbf{y}_{t+1}$.

The marginal likelihood is conceptually simple and has a natural interpretation. However, one drawback is that it is relatively sensitive to the prior distribution. This can be seen from the factorization in (1). For example, the predictive likelihood $p(\mathbf{y}_1 \mid M_k)$ depends entirely on the prior distribution and not on the data. More generally, the component $p(\mathbf{y}_{t+1} \mid \mathbf{y}_{1:t}, M_k)$ is likely to be heavily influenced by the prior distribution when t is small. In what follows we will discuss an alternative Bayesian model selection criterion that is relatively insensitive to the priors. For notational convenience, from here onwards we suppress the model indicator; for example we denote the likelihood by $p(\mathbf{y} \mid \boldsymbol{\theta})$.

The seminal paper by Spiegelhalter et al. (2002) introduces and develops the concept of deviance information criterion (DIC) for model comparison. This criterion is based on the *deviance*, which is defined as

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

where $p(\mathbf{y} \mid \boldsymbol{\theta})$ is the likelihood function and $h(\mathbf{y})$ is some fully specified standardizing term that is a function of the data alone. The *effective number of parameters* p_D of the parametric model is defined to be

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\widetilde{\boldsymbol{\theta}}),$$

where

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

is the posterior mean deviance and $\widetilde{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$, which is typically taken as the posterior mean or mode. Then, the *deviance information criterion* 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 that measures model complexity:

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

Hence, the DIC may be viewed as a trade-off between model adequacy and complexity. For model comparison, the function $h(\mathbf{y})$ is often set to be unity for all models. Therefore, the DIC becomes

$$\text{DIC} = -4\mathbb{E}_{\boldsymbol{\theta}}[\log p(\mathbf{y} \mid \boldsymbol{\theta}) \mid \mathbf{y}] + 2 \log p(\mathbf{y} \mid \widetilde{\boldsymbol{\theta}}).$$

Given a set of competing models for the data, the preferred model is the one with the minimum DIC value.

For latent variable models, such as stochastic volatility models, Celeux et al. (2006) point out that there are numerous alternative definitions of the DIC depending on different concepts of the likelihood. In particular, suppose we augment the model $p(\mathbf{y} \mid \boldsymbol{\theta})$ with a vector of latent variables $\mathbf{z}$ with density $p(\mathbf{z} \mid \boldsymbol{\theta})$ such that

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

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

Naturally, one can define the DIC using the observed-data likelihood and we call this the observed-data DIC:

$$\mathrm{DIC}_{\mathrm{obs}} = -4\mathbb{E}_{\boldsymbol{\theta}}[\log p(\mathbf{y} \mid \boldsymbol{\theta}) \mid \mathbf{y}] + 2 \log p(\mathbf{y} \mid \widehat{\boldsymbol{\theta}}), \tag{2}$$

where the estimate $\widetilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is set as the posterior mode $\widehat{\boldsymbol{\theta}}$. The term $\mathbb{E}_{\boldsymbol{\theta}}[\log p(\mathbf{y} \mid \boldsymbol{\theta}) \mid \mathbf{y}]$ can be estimated by averaging the log-observed-data likelihoods $\log p(\mathbf{y} \mid \boldsymbol{\theta})$ over the posterior draws of $\boldsymbol{\theta}$. In addition, the posterior mode $\widehat{\boldsymbol{\theta}}$ is often approximated by the draw that has the highest value of $p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$ among the posterior draws, where $p(\boldsymbol{\theta})$ is the prior density. It is clear from (2) that the observed-data DIC depends on the prior only via its effect on the posterior distribution. In situations where the likelihood information dominates, one would expect that the observed-data DIC is insensitive to different prior distributions.

One main difficulty in computing $\mathrm{DIC}_{\mathrm{obs}}$ is that the observed-data likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$ is typically time-consuming to evaluate for a wide variety of latent variable models (although important exceptions exist, see, e.g., Chan and Grant, 2014). Since the latent variable structure is usually chosen so that the conditional likelihood $p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{z})$ is available in closed-form, one can alternatively define the DIC using the conditional likelihood and we refer to this version as the conditional DIC:

$$\mathrm{DIC}_{\mathrm{con}} = -4\mathbb{E}_{\boldsymbol{\theta},\mathbf{z}}[\log p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{Z}) \mid \mathbf{y}] + 2 \log p(\mathbf{y} \mid \widehat{\mathbf{z}}, \widehat{\boldsymbol{\theta}}), \tag{3}$$

where $(\widehat{\mathbf{z}}, \widehat{\boldsymbol{\theta}})$ is the joint maximum a posteriori (MAP) estimate of the pair $(\mathbf{z}, \boldsymbol{\theta})$ given the data $\mathbf{y}$.[3] As before, the expectation $\mathbb{E}_{\boldsymbol{\theta},\mathbf{z}}[\log p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{Z}) \mid \mathbf{y}]$ can be estimated by averaging the log-conditional likelihoods $\log p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{z})$ over the posterior draws of the pair $(\mathbf{z}, \boldsymbol{\theta})$. Moreover, the joint MAP estimate can be approximated by the best pair among the posterior draws, i.e., the pair that has the highest value of $p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$.

[3] Celeux et al. (2006) list eight versions of the DIC depending on different concepts of the likelihood and the estimates of $\boldsymbol{\theta}$. For example, one can define DICs using the posterior mean instead of the posterior mode used in this paper. The observed-data DIC in (2) and the conditional DIC in (3) correspond to DIC_2 and DIC_7 in Celeux et al. (2006), respectively. Celeux et al. (2006) study the behavior of the various DICs in the context of two classes of latent variable models: random effect models and mixture models. For both classes of models, the observed-data likelihood can be computed analytically. But this is not the case for stochastic volatility models, which require importance sampling.

Following the influential paper by Berg, Meyer, and Yu (2004), the conditional DIC is widely used for comparing stochastic volatility models, whereas the observed-data DIC is not computed in practice due to the difficulty in evaluating the observed-data likelihood. However, despite its popularity, in Section 5 we show via a Monte Carlo study that the conditional DIC tends to favor overfitted models. In contrast, the observed-data DIC seems to perform well and is better able to choose the correct model. It is also worthwhile to note that there are various new Bayesian approaches for model comparison and hypothesis testing, such as those developed in Li and Yu (2012), Li et al. (2012) and Li, Zeng, and Yu (2014).

3 Stochastic Volatility Models

In this section, we first discuss various stochastic volatility models that are widely used in the literature for modeling financial and macroeconomic time series. Then we outline some efficient algorithms for fitting these models that build on fast band matrix routines.

3.1 The Models

We consider seven different stochastic volatility models. The first model is the standard stochastic volatility model, which we denote as SV:

$$y_t = \mu + \varepsilon_t^y, \qquad \varepsilon_t^y \sim \mathcal{N}(0, \mathrm{e}^{h_t}), \tag{4}$$

$$h_t = \mu_h + \phi_h(h_{t-1} - \mu_h) + \varepsilon_t^h, \qquad \varepsilon_t^h \sim \mathcal{N}(0, \omega_h^2). \tag{5}$$

The log-volatility h_t follows a stationary AR(1) process with $|\phi_h| < 1$ and is initialized with $h_1 \sim \mathcal{N}(\mu_h, \omega_h^2/(1-\phi_h^2))$.

Under the second model, which we refer to as SV2, the observation equation is the same as in (4), but instead of the log-volatility h_t following an AR(1) process as in (5), it follows a stationary AR(2) process:

$$h_t = \mu_h + \phi_h(h_{t-1} - \mu_h) + \rho_h(h_{t-2} - \mu_h) + \varepsilon_t^h, \qquad \varepsilon_t^h \sim \mathcal{N}(0, \omega_h^2), \tag{6}$$

where we assume the roots of the characteristic polynomial associated with (ϕ_h, ρ_h) lie outside the unit circle. Further, the process is initialized by

$$h_1, h_2 \sim \mathcal{N}\left(\mu_h, \frac{(1-\rho_h)\omega_h^2}{(1+\rho_h)((1-\rho_h)^2 - \phi_h^2)}\right).$$

The third model allows for the possibility of infrequent "jumps" in the data series, which may be important for high frequency financial data. Under the stochastic volatility model with jumps (SVJ), the observation equation becomes:

$$y_t = \mu + k_t q_t + \varepsilon_t^y, \qquad \varepsilon_t^y \sim \mathcal{N}(0, \mathrm{e}^{h_t}), \tag{7}$$

where q_t is a Bernoulli jump random variable with success probability $\mathbb{P}(q_t = 1) = \kappa$ and the jump size k_t is modeled as $\log(1 + k_t) \sim \mathcal{N}(-0.5\delta^2, \delta^2)$ so that its expectation is zero. The log-volatility h_t follows the same AR(1) process as in (5).

Another variant is the stochastic volatility in mean (SVM) model of Koopman and Hol Uspensky (2002), which is often used to study volatility feedback. Specifically, under the SVM model, the stochastic volatility enters the observation equation as a covariate:

$$y_t = \mu + \alpha \mathrm{e}^{h_t} + \varepsilon_t^y, \qquad \varepsilon_t^y \sim \mathcal{N}(0, \mathrm{e}^{h_t}). \tag{8}$$

As before, the log-volatility follows the same AR(1) process as in (5).

The next model considered is a version of the stochastic volatility models with moving average innovations in Chan (2013). Specifically, consider the following first-order moving average model with stochastic volatility:

$$y_t = \mu + \varepsilon_t^y, \tag{9}$$
$$\varepsilon_t^y = u_t + \psi u_{t-1}, \qquad u_t \sim \mathcal{N}(0, \mathrm{e}^{h_t}), \tag{10}$$

where we assume that $u_0 = 0$ and the invertibility condition is satisfied, i.e., $|\psi| < 1$. Again the log-volatility h_t is assumed to follow the AR(1) process as in (5). This stochastic volatility model is referred to as SVMA.

The sixth model is the stochastic volatility model with leverage (see, e.g., Yu, 2005; Omori, Chib, Shephard, and Nakajima, 2007):

$$y_t = \mu + \varepsilon_t^y, \tag{11}$$
$$h_{h+1} = \mu_h + \phi_h(h_t - \mu_h) + \varepsilon_t^h, \tag{12}$$

where the innovations ε_t^y and ε_t^h jointly follow a bivariate normal distribution:

$$\begin{pmatrix} \varepsilon_t^y \\ \varepsilon_t^h \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathrm{e}^{h_t} & \rho \mathrm{e}^{\frac{1}{2}h_t}\omega_h \\ \rho \mathrm{e}^{\frac{1}{2}h_t}\omega_h & \omega_h^2 \end{pmatrix}\right).$$

By allowing for a nonzero correlation between the innovations, this model can accommodate the often-observed negative correlation between the return at time t and the realized volatility at time $t + 1$ for financial data. This stochastic volatility model is referred to as SVL.

Lastly, we consider the stochastic volatility model with t innovations in the observation equation, which is denoted as SVt. Since the t distribution can be written as a scale mixture of Gaussian distributions (see, e.g., Geweke, 1993), the SVt model has the following latent variable representation:

$$y_t = \mu + \varepsilon_t^y, \quad \varepsilon_t^y \sim \mathcal{N}(0, \mathrm{e}^{h_t}\lambda_t), \tag{13}$$

where the latent variables $\lambda_1, \ldots, \lambda_T$ are iid $\mathcal{IG}(\nu/2, \nu/2)$ distributed and $\mathcal{IG}(\cdot, \cdot)$ denotes the inverse-gamma distribution. The log-volatility h_t is again assumed to follow the AR(1)

process as in (5). Since the t distribution has heavier tails than the Gaussian, the SVt model allows for a more frequent occurrence of outliers compared to the standard SV model. We summarize the seven stochastic volatility models in Table 1.

We now discuss the set of priors considered under each of the models. For the standard SV, we assume the following independent priors for μ, μ_h, ϕ_h and ω_h^2:

$$\begin{aligned} \mu &\sim \mathcal{N}(\mu_0, V_\mu), & \mu_h &\sim \mathcal{N}(\mu_{h0}, V_{\mu_h}), \\ \phi_h &\sim \mathcal{N}(\phi_{h0}, V_{\phi_h})\mathbb{1}(|\phi_h| < 1), & \omega_h^2 &\sim \mathcal{IG}(\nu_h, S_h). \end{aligned} \tag{14}$$

Note that we impose the stationarity condition $|\phi_h| < 1$ through the prior on ϕ_h. For the SV2, we use the same priors for μ, μ_h and ω_h^2 as in (14), but replace the prior for ϕ_h with a prior for $\boldsymbol{\theta}_h = (\phi_h, \rho_h)'$: $\boldsymbol{\theta}_h \sim \mathcal{N}(\boldsymbol{\theta}_{h0}, \mathbf{V}_{\boldsymbol{\theta}_h})\mathbb{1}(\boldsymbol{\theta}_h \in \mathbf{A})$, where $\mathbf{A} \subset \mathbb{R}^2$ is the set where the roots of the characteristic polynomial defined by $\boldsymbol{\theta}_h$ lie outside the unit circle.

Table 1: List of stochastic volatility models.

Model	Description
SV	standard stochastic volatility model where h_t follows a stationary AR(1)
SV2	same as SV but h_t follows a stationary AR(2)
SVJ	same as SV but the observation equation contains a "jump" component
SVM	same as SV but h_t enters the observation equation as a covariate
SVMA	same as SV but the observation innovation follows an MA(1)
SVL	same as SV but the observation and transition innovations are correlated
SVt	same as SV but the observation innovations are t distributed

For each of the remaining models, the priors for μ, μ_h, ϕ_h and ω_h^2 are exactly the same as in (14). In addition, under the SVJ, the jump intensity κ is assumed to have a beta distribution and the jump variance δ follows a log-normal distribution: $\kappa \sim \mathcal{B}(k_a, k_b)$ and $\log \delta \sim \mathcal{N}(\delta_0, V_\delta)$. For the SVM, the coefficient of the volatility is assumed to have a normal distribution: $\alpha \sim \mathcal{N}(\alpha_0, V_\alpha)$. Next, both the MA(1) coefficient in the SVMA and the correlation coefficient in the SVL have normal distributions truncated within the unit interval: $\psi \sim \mathcal{N}(\psi_0, V_\psi)\mathbb{1}(|\psi| < 1)$ and $\rho \sim \mathcal{N}(\rho_0, V_\rho)\mathbb{1}(|\rho| < 1)$. Lastly, the prior for the degree of freedom parameter ν in the SVt is assumed to be uniform on $(2, 100)$: $\nu \sim \mathcal{U}(2, 100)$. We assume $\nu > 2$ to ensure that the first two moments of the t distribution exist.

3.2 Bayesian Estimation

In this section, we discuss a general approach for fitting all the stochastic volatility models in Section 3.1. The main difficulty in the estimation is the step where one simulates from the joint distribution of $\mathbf{h} = (h_1, \ldots, h_T)'$ conditional on the data and other model parameters, as the observation equation is nonlinear in $\mathbf{h}$. A key feature of our approach

is that it builds upon fast band and sparse matrix algorithms rather than using the conventional Kalman filter. Recent papers using the former approach include Rue (2001) for linear Gaussian Markov random fields; Chan and Jeliazkov (2009) and McCausland et al. (2011) for linear Gaussian state space models; Rue, Martino, and Chopin (2009) for nonlinear Markov random fields; and McCausland (2012), Djegnéné and McCausland (2014) and Chan (2015) for nonlinear state space models.

More specifically, our approach exploits the special structure of the problem, namely, that the Hessian of the log-conditional density of $\mathbf{h}$ is a band matrix—i.e., it contains only a few nonzero elements along a narrow diagonal band. This feature is important in developing efficient sampling algorithms. In addition, the same approach can be used for obtaining efficient importance sampling estimators as discussed in Section 4. For concreteness, we focus on the estimation of the standard stochastic volatility model in (4)–(5), with modifications of the main algorithm for fitting the other models discussed in Appendix A. Let $\mathbf{y} = (y_1, \dots, y_T)'$. Then posterior draws can be obtained by sequentially sampling from:

1. $p(\mathbf{h} \,|\, \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2)$;
2. $p(\mu \,|\, \mathbf{y}, \mathbf{h}, \mu_h, \phi_h, \omega_h^2) = p(\mu \,|\, \mathbf{y}, \mathbf{h})$;
3. $p(\mu_h \,|\, \mathbf{y}, \mu, \mathbf{h}, \phi_h, \omega_h^2) = p(\mu_h \,|\, \mathbf{h}, \phi_h, \omega_h^2)$;
4. $p(\omega_h^2 \,|\, \mathbf{y}, \mu, \mathbf{h}, \mu_h, \phi_h) = p(\omega_h^2 \,|\, \mathbf{h}, \mu_h, \phi_h)$;
5. $p(\phi_h \,|\, \mathbf{y}, \mu, \mathbf{h}, \mu_h, \omega_h^2) = p(\phi_h \,|\, \mathbf{h}, \mu_h, \omega_h^2)$.

In Step 1, the joint conditional density $p(\mathbf{h} \,|\, \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2)$ is high-dimensional and non-standard. For the standard stochastic volatility model, this step can be accomplished using the auxiliary mixture sampler of Kim, Shepherd, and Chib (1998). However, this approach is model specific and cannot be easily generalized to estimate other stochastic volatility models such as the SVM. As a result, we discuss a direct method to simulate from this density using the acceptance-rejection Metropolis-Hastings algorithm (see, e.g., Tierney, 1994). More specifically, we note that the Hessian of $\log p(\mathbf{h} \,|\, \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2)$ is a band matrix. Consequently, using fast band matrix routines we can quickly obtain a Gaussian approximation as a proposal density. Furthermore, by construction, the precision matrix—i.e., the inverse of the covariance matrix—of the Gaussian proposal density is also a band matrix. As such, candidate draws can be obtained quickly via the precision sampler in Chan and Jeliazkov (2009) instead of Kalman filter-based algorithms. The computation details are given in Appendix A.

Steps 2, 3 and 4 can be easily completed, as all the conditional distributions are standard. In particular, it is easy to check that

$$(\mu \,|\, \mathbf{y}, \mathbf{h}) \sim \mathcal{N}(\widehat{\mu}, D_\mu), \quad (\mu_h \,|\, \mathbf{h}, \phi_h, \omega_h^2) \sim \mathcal{N}(\widehat{\mu}_h, D_{\mu_h}), \quad (\omega_h^2 \,|\, \mathbf{h}, \mu_h, \phi_h) \sim \mathcal{IG}(\nu_h + T/2, \widetilde{S}_h),$$

where $\widetilde{S}_h = S_h + ((1-\phi_h^2)(h_1-\mu_h)^2 + \sum_{t=2}^{T}(h_t - \mu_h - \phi_h(h_{t-1}-\mu_h))^2)/2$,

$$D_\mu^{-1} = V_\mu^{-1} + \sum_{t=1}^{T} \mathrm{e}^{-h_t}, \quad \widehat{\mu} = D_\mu(V_\mu^{-1}\mu_0 + \sum_{t=1}^{T} \mathrm{e}^{-h_t} y_t),$$
$$D_{\mu_h}^{-1} = V_{\mu_h}^{-1} + \mathbf{X}_{\mu_h}' \boldsymbol{\Sigma}_{\mathbf{h}}^{-1} \mathbf{X}_{\mu_h}, \quad \widehat{\mu}_h = D_{\mu_h}(V_{\mu_h}^{-1}\mu_{h0} + \mathbf{X}_{\mu_h}' \boldsymbol{\Sigma}_{\mathbf{h}}^{-1} \mathbf{z}_{\mu_h}),$$

with $\mathbf{X}_{\mu_h} = (1, 1-\phi_h, \ldots, 1-\phi_h)'$, $\mathbf{z}_{\mu_h} = (h_1, h_2 - \phi_h h_1, \ldots, h_T - \phi_h h_{T-1})'$ and $\boldsymbol{\Sigma}_{\mathbf{h}} = \mathrm{diag}(\omega_h^2/(1-\phi_h^2), \omega_h^2, \ldots, \omega_h^2)$.

Lastly, one can sample from $p(\phi_h \,|\, \mathbf{h}, \mu_h, \omega_h^2)$ using an independence-chain Metropolis-Hastings step with proposal $\mathcal{N}(\widehat{\phi}_h, D_{\phi_h})\mathbb{1}(|\phi_h| < 1)$, where $D_{\phi_h}^{-1} = V_{\phi_h}^{-1} + \mathbf{X}_{\phi_h}'\mathbf{X}_{\phi_h}/\omega_h^2$ and $\widehat{\phi}_h = D_{\phi_h}(V_{\phi_h}^{-1}\phi_{h0} + \mathbf{X}_{\phi_h}'\mathbf{z}_{\phi_h}/\omega_h^2)$, with $\mathbf{X}_{\phi_h} = (h_1 - \mu_h, \ldots, h_{T-1} - \mu_h)'$ and $\mathbf{z}_{\phi_h} = (h_2 - \mu_h, \ldots, h_T - \mu_h)'$.

4 Importance Sampling for the Observed-Data Likelihoods

The popularity of the conditional DIC for comparing stochastic volatility models is partly due to its straightforward computation and its implementation in standard software such as WinBUGS. On the other hand, computing the observed-data DIC is less straightforward. In a recent paper, Chan and Grant (2014) derive analytical expressions for the observed-data likelihoods for a variety of linear latent variable models. However, for the stochastic volatility models discussed in Section 3, the observed-data likelihoods are not available in closed-form. One option, at least in principle, is the auxiliary particle filter proposed in Pitt and Shephard (1999), which can be used to evaluate the observed-data likelihood for general nonlinear state space models. In practice, however, the auxiliary particle filter is computationally intensive and it might not be feasible to be employed in our setting as the observed-data likelihood needs to be evaluated tens of thousands of times. To overcome this difficulty, in this section we consider fast algorithms for estimating the observed-data likelihoods for stochastic volatility models using importance sampling (see, e.g., Kroese, Taimre, and Botev, 2011, Chapter 9.7).

Recall that the observed-data or integrated likelihood is given by

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

where $p(\mathbf{y} \,|\, \boldsymbol{\theta}, \mathbf{z})$ is the conditional likelihood and $p(\mathbf{z} \,|\, \boldsymbol{\theta})$ is the prior density of the latent variables $\mathbf{z}$. Let $g(\mathbf{z})$ be a density that dominates $p(\mathbf{y} \,|\, \boldsymbol{\theta}, \mathbf{z})p(\mathbf{z} \,|\, \boldsymbol{\theta})$, i.e., $g(\mathbf{z}) = 0$ implies $p(\mathbf{y} \,|\, \boldsymbol{\theta}, \mathbf{z})p(\mathbf{z} \,|\, \boldsymbol{\theta}) = 0$. Then, the observed-data likelihood can be rewritten as

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

Hence, if $\mathbf{Z}_1, \ldots, \mathbf{Z}_R$ are independent samples from the *importance density* $g(\mathbf{z})$, then

$$\widehat{p(\mathbf{y} \mid \boldsymbol{\theta})} = \frac{1}{R} \sum_{i=1}^{R} \frac{p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{Z}_i) p(\mathbf{Z}_i \mid \boldsymbol{\theta})}{g(\mathbf{Z}_i)} \tag{15}$$

is an unbiased, simulation-consistent estimator of the observed-data likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$. Since the samples are independent, a numerical standard error of this importance sampling estimator, $\widehat{s}/\sqrt{R}$, can be easily computed, where $\widehat{s}$ is the sample standard deviation of the importance sampling weights $p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{Z}_i) p(\mathbf{Z}_i \mid \boldsymbol{\theta}) / g(\mathbf{Z}_i), i = 1, \ldots, R$. In addition, it is often more convenient to work in the logarithmic scale. Therefore it might also be of interest to obtain a numerical standard error of $\log \widehat{p(\mathbf{y} \mid \boldsymbol{\theta})}$. This can be done using either the delta method or the batch means method (see, e.g., Kroese et al., 2011).

The quality of the importance sampling estimator in (15) depends critically on the choice of the importance density $g(\mathbf{z})$. It can be shown that the conditional density of the latent variables $p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \propto p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{z}) p(\mathbf{z} \mid \boldsymbol{\theta})$ gives rise to a zero-variance importance sampling estimator (see, e.g., Kroese et al., 2011, Chapter 9.7.1). An obvious difficulty is that the evaluation of the optimal importance density $p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$ is not possible for stochastic volatility models as the normalization constant is not known. However, it provides guidance for choosing a good importance density. In particular, we would like to choose $g(\mathbf{z})$ to be "close" to the optimal importance density $p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$. In what follows, we focus on the standard stochastic volatility model, with the importance densities for the other stochastic volatility models discussed in Appendix B.

For the standard stochastic volatility model in (4)–(5), the latent variables are the log-volatilities $\mathbf{h}$. Therefore, we wish to approximate the conditional density

$$p(\mathbf{h} \mid \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2) \propto p(\mathbf{y} \mid \mu, \mathbf{h}) p(\mathbf{h} \mid \mu_h, \phi_h, \omega_h^2)$$

to obtain a good importance density $g(\mathbf{h})$ for the estimator in (15). In fact, we have already discussed such an approximation when we outlined the estimation of the stochastic volatility model in Section 3.2. Specifically, we considered (for details see Appendix A) the Gaussian approximation with mean vector $\widehat{\mathbf{h}}$ and precision matrix $\mathbf{K}_{\mathbf{h}}$, where $\widehat{\mathbf{h}}$ is the mode of $p(\mathbf{h} \mid \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2)$ and $\mathbf{K}_{\mathbf{h}}$ is the negative Hessian evaluated at the mode. Note that this approximating Gaussian density is the same as the one proposed in Durbin and Koopman (1997), although we obtain the approximation via band matrix routines instead of the Kalman filter.

In addition, note that $\mathbf{K}_{\mathbf{h}}$ is a band matrix. As such, draws from $\mathcal{N}(\widehat{\mathbf{h}}, \mathbf{K}_{\mathbf{h}}^{-1})$ can be efficiently obtained using the precision sampler in Chan and Jeliazkov (2009), where the computation cost of obtaining an additional draw is only $\mathcal{O}(T)$. This is a crucial feature as multiple draws from the high-dimensional importance density are required to construct the estimator in (15). In addition, this importance density can be quickly evaluated at any point as its precision matrix $\mathbf{K}_{\mathbf{h}}$ is a band matrix. Choices of importance densities for the other stochastic volatility models are discussed in Appendix B.

For the importance sampling estimators to work well, a requirement is that the variance of the importance sampling weights should be finite. While this requirement may be

checked analytically in simple problems, checking it in high-dimensional settings such as ours is difficult. One strategy to ensure this finite-variance condition holds is to modify the importance sampling estimator $g(\mathbf{z})$ to include an additional mixture component as proposed by Hesterberg (1995). More specifically, for $\gamma \in (0, 1)$, consider the mixture density

$$g_\gamma(\mathbf{z}) = \gamma p(\mathbf{z} \mid \boldsymbol{\theta}) + (1 - \gamma) g(\mathbf{z}),$$

i.e., with probability γ, samples are taken from the prior density $p(\mathbf{z} \mid \boldsymbol{\theta})$; otherwise, we draw from the original importance sampling density $g(\mathbf{z})$. If we assume that for fixed $\mathbf{y}$ and $\boldsymbol{\theta}$, the conditional likelihood is bounded in $\mathbf{z}$, i.e., there exists a constant $N_{\mathbf{y},\boldsymbol{\theta}}$ such that $p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{z}) \leqslant N_{\mathbf{y},\boldsymbol{\theta}}$ for all $\mathbf{z}$ (this condition holds for the stochastic volatility models we consider), then the importance sampling weight is bounded by:

$$\frac{p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{z}) p(\mathbf{z} \mid \boldsymbol{\theta})}{g_\gamma(\mathbf{z})} \leqslant \frac{p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{z}) p(\mathbf{z} \mid \boldsymbol{\theta})}{\gamma p(\mathbf{z} \mid \boldsymbol{\theta})} \leqslant \frac{N_{\mathbf{y},\boldsymbol{\theta}}}{\gamma}.$$

Hence, the variance of the importance sampling weights corresponding to $g_\gamma(\mathbf{z})$ is finite. In our applications we experiment with both $g(\mathbf{z})$ and $g_\gamma(\mathbf{z})$, and they give very similar results.[4]

5 A Monte Carlo Study

In this section, we first examine the behavior of the conditional and observed-data DICs via a simulation study. Then, we investigate the sensitivity of the marginal likelihood and observed-data DIC to different prior distributions.

To assess whether the conditional and observed-data DICs are able to pick the correct model from which the data are generated, we simulate data from three models: a constant variance model where observations are drawn independently from $\mathcal{N}(0, \sigma^2)$, the SV model and the SVJ model. Three hundred datasets each comprised of $T = 1000$ observations are produced from each of these three models. For each dataset, we estimate the three models by constructing Markov chains of length 20000 after a burn-in period of 1000.

To compute the observed-data likelihoods for the two stochastic volatility models, we sample $R = 50$ draws from the importance density at every iteration of the MCMC run. In choosing the sample size R, there is the obvious trade-off between faster computation time and more accurate estimates. We have experimented with different sample sizes and $R = 50$ seems to be enough (e.g., the numerical standard error of the observed-data likelihood estimate is less than 0.5 when the estimate is about -1000).

The parameter values are chosen to be comparable to those estimated from financial daily returns data (measured in decimals). They are also similar to those used in other

[4]For example, using the S&P 500 data (for details see Section 6) we evaluate the observed-data likelihood of the standard stochastic volatility model at the posterior means of the parameters. We obtain 2914.8 and 2914.7 (in log) using the original estimator and the modified estimator with $\gamma = 0.05$, respectively.

simulation studies, such as those in Chib et al. (2002) and Berg et al. (2004). In particular, we set $\mu = 0$ for all models. Parameters for the log-volatility transition are set to be $\mu_h = -10$, $\phi_h = 0.97$ and $\omega_h^2 = 0.2^2$ for both the SV and SVJ models. Moreover, parameters for the jump component are selected to be $\kappa = 0.03$ and $\delta = 0.03$. Finally, σ^2 is set so that it is comparable to the variance in the stochastic volatility models: $\sigma^2 = \mathrm{e}^{\mu_h} = \mathrm{e}^{-10}$.

The priors discussed in Section 3.1 are considered. We choose the same hyperparameters for parameters that are common across models. Moreover, the hyperparameters are selected so that the implied prior means are similar to the estimates from typical financial daily returns data. In particular, we have $\mu_0 = 0$, $\mu_{h0} = -10$, $V_\mu = V_{\mu_h} = 10$, $\phi_{h0} = 0.97$, $V_{\phi_h} = 0.1^2$, $\nu_h = 5$, $S_h = 0.16$, $k_a = 2$, $k_b = 100$, $\delta_0 = -3.07$ and $V_\delta = 0.149$. These values imply $\mathbb{E}\mu = 0$, $\mathbb{E}\mu_h = -10$, $\mathbb{E}\phi_h = 0.908$, $\mathbb{E}\omega_h^2 = 0.2^2$, $\mathbb{E}\kappa = 0.0196$ and $\mathbb{E}\delta = 0.05$.

In the first experiment, 300 datasets are generated from the constant variance (Const-Var) model. Given each dataset both the conditional and observed-data DICs of the two stochastic volatility models are computed. They are then compared to the (observed-data) DIC of the Const-Var model. Specifically, we subtract the latter DIC from the DICs of both the SV and SVJ models, and the results are reported in Figure 1.

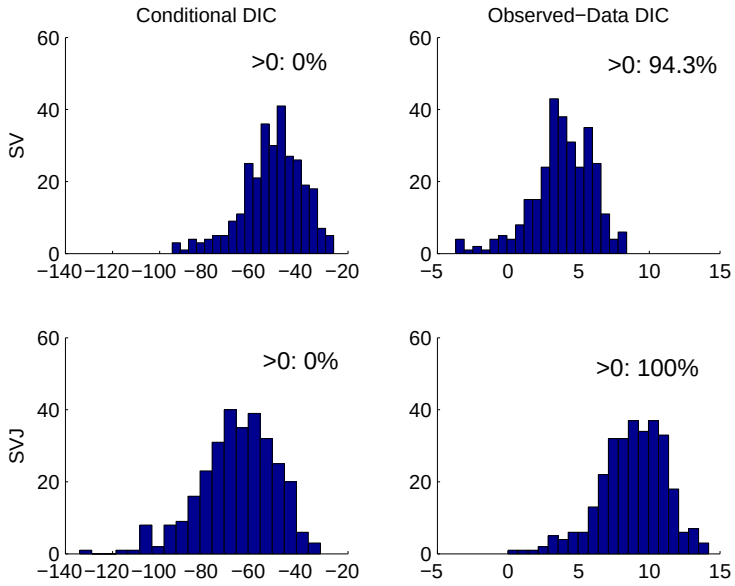

Figure 1: DICs of a given model minus that of the correct model (the Const-Var model). A positive value indicates that the correct model is favored.

Recall that a model is preferred if it has a smaller DIC value. Hence, according to the conditional DIC both the SV and SVJ models are favored relative to the correct model for all the generated datasets. In contrast, for the majority of datasets (94.3% and 100% for the SV and SVJ models, respectively), the observed-data DIC favors the correct model.

It is worth noting that among the two stochastic volatility models, the conditional DIC tends to prefer the more complex SVJ model.

In the second experiment, datasets are generated from the SV model, which includes the Const-Var model as a special case and is also nested within the SVJ model. The DICs relative to the SV model are reported in Figure 2. Both the conditional and observed-data DICs favor the SV model relative to the Const-Var model. However, the conditional DIC favors the overfitted SVJ model: for 100% of the datasets the SVJ model is preferred relative to the correct model. In contrast, the observed-data DIC favors the correct model for 98.3% of the datasets. It is also interesting to note that the differences in observed-data DICs between the SV and the SVJ models are small compared to the differences between the SV and Const-Var models, reflecting a small penalty for overfit relative to "underfit"—a model's inability to fit the data well.

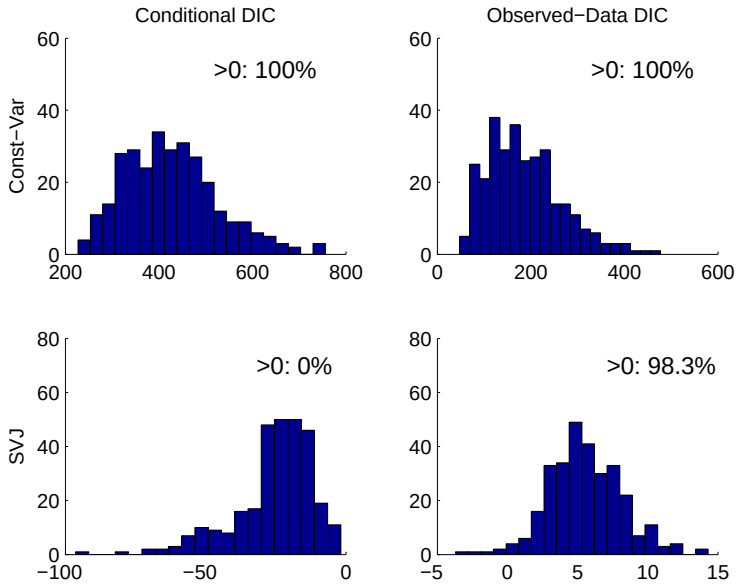

Figure 2: DICs of a given model minus that of the correct model (the SV model). A positive value indicates that the correct model is favored.

In the last simulation experiment, we generate datasets from the SVJ model, which includes both the Const-Var and SV models as special cases. As before, we report the DICs relative to the correct model, and the results are presented in Figure 3. In this example, both the conditional and observed-data DICs tend to pick the correct, more general SVJ model. In particular, comparing between SV and SVJ, the conditional DIC prefers the correct model for 97.7% of the datasets while the figure for the observed-data DIC is 99.7% of the datasets.

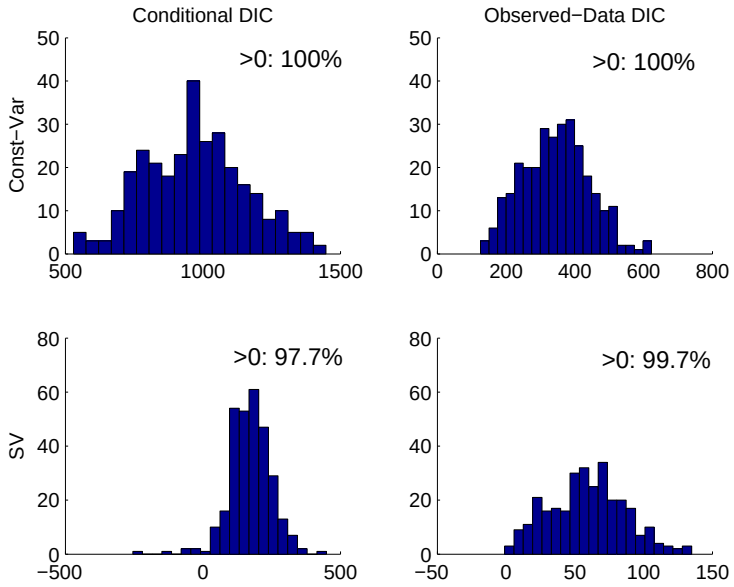

Figure 3: DICs of a given model minus that of the correct model (the SVJ model). A positive value indicates that the correct model is favored.

Overall, this Monte Carlo study provides evidence that the conditional DIC tends to pick overfitted models whereas the observed-data DIC seems to perform well.

Next, we report the effective number of parameters p_D computed using the conditional likelihood and the observed-data likelihood; we call the former conditional p_D and the latter observed-data p_D. As discussed in Section 2, the DIC may be viewed as a trade-off between model adequacy and complexity, where model complexity is measured by p_D. When prior information is dominated by the likelihood, one can show that (see, e.g., Li et al., 2012) $p_D = p + o(1)$, where p is the number of parameters. In other words, when likelihood information dominates, one expects that p_D is close to p, and the difference reflects the amount of prior information.

Using the 300 datasets from the first experiment (i.e., data generated from the constant variance model), we compute the conditional and observed-data p_D for the SV and SVJ models. The results are reported in Figure 4. Recall that the SV model has 4 parameters and the SVJ model has 6. The observed-data p_D for both models are quite close to the actual number of parameters, confirming the theoretical results. However, the conditional p_D for both models are all negative, indicating a negative penalty for model complexity, which is more difficult to justify.

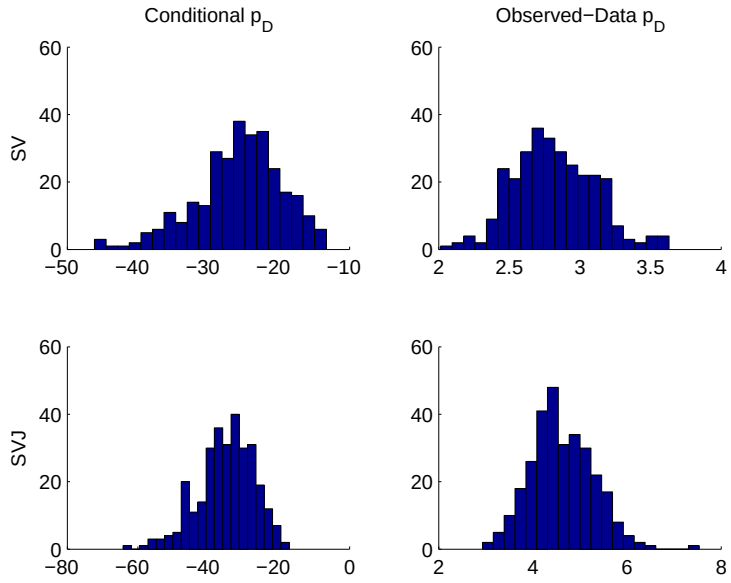

Figure 4: The effective number of parameters p_D computed using the conditional and observed-data likelihoods.

Lastly, we investigate the sensitivity of both the marginal likelihood and observed-data DIC to different prior distributions. As discussed in Section 2, one drawback of the marginal likelihood is that it is relatively sensitive to different prior assumptions, whereas the observed-data DIC is not. To that end, we generate a dataset of $T = 1000$ observations from the constant variance (Const-Var) model using the parameters described in the first set of experiments above. We then estimate the marginal likelihoods and the observed-data DICs for two models: Const-Var and SV. We assume exactly the same priors as before, except that we vary the degree of freedom hyperparameter ν_h for the parameter ω_h^2. The results are reported in Table 2.

Table 2: Observed-data DICs and log marginal likelihoods for the simulated dataset generated from the Const-Var model.

	Const-Var	SV				
		$\nu_h = 0.1$	$\nu_h = 1$	$\nu_h = 2$	$\nu_h = 5$	$\nu_h = 10$
Observed-data DIC	-7257.5	-7251.4	-7251.3	-7252.2	-7253.4	-7255.3
Log marginal likelihood	3618.7	3611.4	3614.3	3617.7	3628.6	3648.8

By varying ν_h from 0.1 to 10, the observed-data DIC changes from -7251.4 to -7255.3, a difference of 3.9.[5] In contrast, the difference in the log marginal likelihoods is 37.4. In

[5]Recall that the observed-data DIC is defined in terms of the log likelihood. Hence, this difference is in log scale.

addition, when one assumes $\nu_h = 0.1$, the marginal likelihood favors the correct Const-Var model. However, when $\nu_h \geqslant 5$, it overwhelmingly prefers the SV model. On the other hand, the observed-data DIC picks the correct model for the range of hyperparamters considered.

6 An Empirical Application

In this section we illustrate the methodology for estimating the observed-data likelihood with an application that involves the daily returns (in decimals) on the S&P 500 index. The sample period is January 2007 to December 2012, with a total of $T = 1509$ observations. The time series plot of the data is presented in Figure 5. We estimate the stochastic volatility models listed in Table 1 using the S&P 500 data, and we assess which model fits the data best while taking model complexity into account.

In addition, we consider two versions of the SVt model, which we label as SVt-1 and SVt-2. In the first version, the conditional likelihood is implied by the latent variable representation given in (13)—i.e., the latent variables are $h_1, \ldots, h_T$ and $\lambda_1, \ldots, \lambda_T$. In the second version, we directly assume that the innovation ε_t^y follows a t distribution, and hence the latent variables are $h_1, \ldots, h_T$ only. We show below that the different choices of latent variables lead to very different conditional DIC values.

We use the priors given in Section 3.1 and set the same hyperparameters for parameters that are common across models. For the SV and SVJ models, the same hyperparameters as in the Monte Carlo study in Section 5 are used. For the remaining models, we choose $\rho_{h0} = 0$, $V_{\rho_h} = 1$, $\psi_0 = 0$, $V_\psi = 1$, $\alpha_0 = 0$, $V_\alpha = 100^2$, $\rho_0 = 0$ and $V_\rho = 1$.

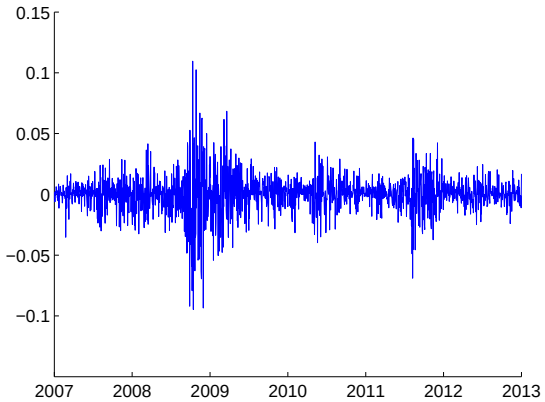

Figure 5: Daily returns on S&P 500 (in decimals) from January 2007 to December 2012.

For each model, we run 10 parallel chains, each of which is of length 10000 after a burn-in period of 1000, with a total of 100000 posterior draws. For each chain, we compute

the corresponding DIC estimate and report the grand mean of these 10 DIC estimates. The associated numerical standard error is obtained by dividing the sample standard deviation by $\sqrt{10}$. To compute the observed-data likelihood, we obtain $R = 50$ draws from the importance density at each MCMC iteration.

The estimated DICs and their numerical standard errors are reported in Table 3. For comparison, we also provide marginal likelihood estimates for each model. They are obtained using the adaptive importance sampling method called the improved cross-entropy method considered in Chan and Eisenstat (2015). This approach requires evaluation of the observed-data likelihood, which is done using our proposed methodology with again $R = 50$ draws from the importance sampling density. For this criterion a higher value indicates a more preferred model.

A few broad conclusions may be drawn from this model comparison exercise. Firstly, while the observed-data and conditional DICs agree on the ranking of the top two models, their rankings of the rest of the models differ substantially. For example, the SVMA model is ranked third by the observed-data DIC, whereas the conditional DIC ranks it as the worst—instead it prefers the SVJ model, which is ranked as the second to last by the observed-data DIC. Hence, erroneous conclusions might be drawn if the conditional DIC favors overfitted models, as suggested by the Monte Carlo study in Section 5. In fact, the ranking by the marginal likelihood more generally supports that of the observed-data DIC rather than the conditional DIC.

Secondly, the conditional DICs of the two versions of the SVt model are very different. For the SVt-1—where the latent variables are $h_1, \ldots, h_T$ and $\lambda_1, \ldots, \lambda_T$—the estimated conditional DIC is -9495.7. However, the conditional DIC estimate is -9291.8 for the SVt-2, under which the latent variables are $h_1, \ldots, h_T$. Accordingly, the ranking changes from the second to the second to last. This gives a particularly stark example of how the conditional DIC depends critically on how the latent variables are defined.

Thirdly, all three criteria rank the SVL model as the best model, followed by the SVt model. These results are in line with the general finding that leverage effects and heavy-tailed distributions are important for modeling equity returns.

Fourthly, according to the observed-data DIC and marginal likelihood, both the volatility feedback (SVM) and moving average (SVMA) components seem to be useful additions to the basic SV model. In contrast, the jump component and the AR(2) transition for the log-volatility are not as important in modeling the returns on the S&P 500. It is interesting to note that even though both the t distribution and the jump component aim to allow for a more frequent occurrence of "outliers" than the Gaussian distribution, the data prefer the former but not the latter. The key difference between the two is that the jump component is essentially a mixture of a Gaussian distribution and a discrete distribution, whereas the t distribution is a (continuous) scale mixture of Gaussian distributions. The latter turns out to fit the distribution of outliers better.

Table 3: Estimated DICs and log marginal likelihoods (numerical standard errors in parentheses).

	Observed-data DIC	Rank	Conditional DIC	Rank	Log marginal likelihood	Rank
SV	−9080.8 (0.56)	5	−9305.0 (6.18)	6	4532.9 (0.02)	6
SV2	−9057.5 (1.20)	7	−9315.4 (5.55)	5	4531.9 (0.06)	7
SVJ	−9079.5 (1.04)	6	−9342.2 (26.6)	3	4533.0 (0.03)	5
SVM	−9085.7 (0.28)	4	−9316.1 (5.23)	4	4534.4 (0.01)	3
SVMA	−9087.8 (0.51)	3	−9296.7 (5.06)	8	4533.4 (0.03)	4
SVL	−9145.2 (0.49)	1	−9776.9 (76.1)	1	4560.5 (0.02)	1
SVt-1	−9097.2 (0.61)	2	−9495.7 (7.72)	2	4537.6 (0.01)	2
SVt-2	−9097.2 (0.61)	2	−9291.8 (3.97)	7	4537.6 (0.01)	2

Lastly, the numerical standard errors of the conditional DICs are typically quite large, even after averaging 100000 posterior draws. This highlights the need to report numerical standard errors of the conditional DICs, which is often not done in empirical research. On the other hand, the observed-data DICs are much more accurately estimated.

Next, we report in Table 4 the effective number of parameters for each of the stochastic volatility models—computed using both the observed-data and conditional likelihoods. As before, we call the former version observed-data p_D and the latter conditional p_D. As discussed in the Monte Carlo study in Section 5, when prior information is dominated by the likelihood, one expects p_D to be close to the actual number of parameters. The observed-data p_D estimates for all models are positive, whereas many conditional p_D estimates are negative, indicating a negative penalty for model complexity. The latter counter-intuitive result casts doubt on the suitability of using the conditional p_D as a measure of model complexity. Lastly, it is interesting to note that the observed-data p_D indicates that the SVt model is the least complex model, whereas the SV, SVM, SVMA and SVL all have similar model complexity. One interpretation is that by allowing for t innovations in the observation equation, the prior under SVt has a stronger influence and therefore makes the model less complex.

Table 4: Estimated effective numbers of parameters computed using the observed-data and conditional likelihoods (numerical standard errors in parentheses).

	# of parameters	Observed-data p_D	Conditional p_D
SV	4	10.4	−22.7
		(0.45)	(3.89)
SV2	5	21.1	−11.3
		(0.53)	(6.83)
SVJ	6	15.4	9.7
		(1.11)	(18.4)
SVM	5	11.0	−18.9
		(0.36)	(4.79)
SVMA	5	10.3	−27.2
		(0.51)	(4.40)
SVL	5	10.1	953.1
		(0.42)	(56.6)
SVt-1	5	7.4	−101.0
		(0.36)	(6.85)
SVt-2	5	7.4	−30.7
		(0.36)	(3.94)

Now, we report the posterior means and standard deviations of the parameters in Table 5. The parameters governing the transition of the log-volatility have similar estimates across models. In particular, all show high persistence with the posterior mean of ϕ_h estimated to be between 0.95 to 0.987. In addition, an AR(1) transition seems to be sufficient given that the posterior mean of the AR(2) coefficient ρ_h is very small (0.022), which also supports the ranking of the observed-data DIC—it ranks the SV2 model below the SV model (the conditional DIC ranks the SV2 model higher, but the numerical standard errors are too large to be conclusive).

Interestingly, the posterior estimates of κ, α and ψ all seem to support the ranking of the observed-data DIC (but not that of the conditional DIC). For example, recall that when $\psi = 0$, the SVMA model reduces to the SV model. Since the observed-data DIC favors the SVMA model relative to the SV model, one would expect that the posterior distribution of ψ has little mass around zero. In fact, the 95% credible interval of ψ is estimated to be $(-0.126, -0.020)$, which excludes 0. Similarly, when $\alpha = 0$, the SVM model reduces to the SV model. The 95% credible interval of α is estimated to be $(-9.411, -1.158)$, which is consistent with the ranking of the observed-data DIC that favors the SVM model over the SV model. However, the observed-data DIC does not seem to be able to discriminate between the SV and SVJ models, which is reflected in the small posterior mean of κ relative to its posterior standard deviation.

Table 5: Parameter posterior means and standard deviations (in parentheses).

	SV	SV2	SVJ	SVM	SVMA	SVL	SVt
μ	0.0008	0.0009	0.0008	0.0013	0.0008	0.0005	0.0009
	(0.0002)	(0.0002)	(0.0002)	(0.0003)	(0.0002)	(0.0002)	(0.0002)
μ_h	−9.109	−9.161	−9.168	−8.832	−9.113	−9.234	−9.324
	(0.431)	(0.567)	(0.477)	(0.967)	(0.438)	(0.261)	(0.476)
ϕ_h	0.985	0.950	0.986	0.984	0.985	0.976	0.987
	(0.006)	(0.091)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
ω_h^2	0.039	0.059	0.037	0.040	0.038	0.052	0.036
	(0.008)	(0.015)	(0.008)	(0.008)	(0.008)	(0.010)	(0.008)
ρ_h	–	0.022	–	–	–	–	–
	–	(0.096)	–	–	–	–	–
κ	–	–	0.017	–	–	–	–
	–	–	(0.015)	–	–	–	–
δ	–	–	0.026	–	–	–	–
	–	–	(0.010)	–	–	–	–
α	–	–	–	−5.224	–	–	–
	–	–	–	(2.10)	–	–	–
ψ	–	–	–	–	−0.073	–	–
	–	–	–	–	(0.027)	–	–
ρ	–	–	–	–	–	−0.742	–
	–	–	–	–	–	(0.058)	–
ν	–	–	–	–	–	–	11.83
	–	–	–	–	–	–	(5.87)

7 Further Applications

In their seminal paper, Durbin and Koopman (1997) show how importance sampling estimators can be constructed to evaluate the observed-data likelihood of non-Gaussian state space models. Their algorithms use the Kalman filter to compute the approximating density and to obtain importance sampling draws. In contrast, our approach is based on fast band matrix routines, which require far less computations and avoid the forward-filtering-backward-smoothing loops.

We have focused on univariate stochastic volatility models in this paper, but the proposed approach is applicable to more general state space models. In particular, Durbin and Koopman (1997) consider models where the observations come from an exponential family distribution (e.g., a Poisson count model) and where the observation equation is linear but the observation innovations are non-Gaussian (e.g., an unobserved components model with t innovations). Our approach can be applied to those settings.

In addition, we discuss below in more detail a class of nonlinear latent factor models to

which our approach can be applied. This is motivated by a large and growing literature on forecasting with many predictors. One popular approach to extract useful information from large datasets is to use factor models (e.g., Stock and Watson, 2002; Forni, Hallin, Lippi, and Reichlin, 2003). In typical applications, a small number of factors can account for much of the variation in the economic and financial aggregates. Consequently, a simple dynamic factor model often provides better forecasts than competing methods. So far the literature has focused on linear factor models; nonlinear forecasting with many predictors remains mostly unexplored (Stock and Watson, 2006). Hence, it would be interesting to investigate if allowing for nonlinearities would improve forecasting performance, especially during volatile periods.

Consider a nonlinear latent factor model of the form:

$$\mathbf{y}_t = \mathbf{\Lambda}\mathbf{g}(\mathbf{b}_t) + \boldsymbol{\varepsilon}_t^y, \qquad \boldsymbol{\varepsilon}_t^y \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_y),$$

where $\mathbf{y}_t$ is an $n \times 1$ vector of observations, $\mathbf{\Lambda}$ is a matrix of factor loadings, $\mathbf{b}_t$ is an $m \times 1$ vector of latent factors, $\mathbf{g}$ is a vector-valued function and $\mathbf{\Sigma}_y$ is diagonal. To complete the model specification, assume that $\mathbf{b}_t$ follows a stationary autoregressive process:

$$\mathbf{b}_t = \mathbf{\Phi}_b \mathbf{b}_{t-1} + \boldsymbol{\varepsilon}_t^b, \qquad \boldsymbol{\varepsilon}_t^b \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_b),$$

where both $\mathbf{\Phi}_b$ and $\mathbf{\Sigma}_b$ are diagonal. This setup therefore generalizes the usual dynamic factor model to allow for nonlinear terms in $\mathbf{b}_t$.

Below we provide some computational details for sampling the factors $\mathbf{b} = (\mathbf{b}_1', \ldots, \mathbf{b}_T')'$. This step is difficult as $\mathbf{b}$ is high-dimensional and its full conditional distribution is non-standard. The goal is to approximate the full conditional density of $\mathbf{b}$ using a Gaussian density. This provides a multivariate generalization of the methods discussed earlier in this paper. In principle this step can be implemented using forward-filtering-backward-smoothing methods based on Durbin and Koopman (1997). But as discussed above, this approach is expected to be slower than the proposed method based on fast band matrix routines.

To approximate $p(\mathbf{b} \,|\, \mathbf{y}, \mathbf{\Lambda}, \mathbf{\Phi}_b, \mathbf{\Sigma}_y, \mathbf{\Sigma}_b, \mathbf{b}_0)$ using a Gaussian density, note that

$$p(\mathbf{b} \,|\, \mathbf{y}, \mathbf{\Lambda}, \mathbf{\Phi}_b, \mathbf{\Sigma}_y, \mathbf{\Sigma}_b, \mathbf{b}_0) \propto p(\mathbf{y} \,|\, \mathbf{\Lambda}, \mathbf{b}, \mathbf{\Sigma}_y) p(\mathbf{b} \,|\, \mathbf{\Phi}_b, \mathbf{\Sigma}_b, \mathbf{b}_0).$$

We first show that $p(\mathbf{b} \,|\, \mathbf{\Phi}_b, \mathbf{\Sigma}_b, \mathbf{b}_0)$ is a Gaussian density. To that end, let

$$\mathbf{H}_{\mathbf{\Phi}_b} = \begin{pmatrix} \mathbf{I}_m & \mathbf{0} & \cdots & \mathbf{0} \\ -\mathbf{\Phi}_b & \mathbf{I}_m & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \cdots & -\mathbf{\Phi}_b & \mathbf{I}_m \end{pmatrix}.$$

Then, we have

$$(\mathbf{b} \,|\, \mathbf{\Phi}_b, \mathbf{\Sigma}_b, \mathbf{b}_0) \sim \mathcal{N}(\boldsymbol{\delta}_b, (\mathbf{H}_{\mathbf{\Phi}_b}' \mathbf{S}_b^{-1} \mathbf{H}_{\mathbf{\Phi}_b})^{-1}),$$

where $\mathbf{S}_b = \mathbf{I}_T \otimes \boldsymbol{\Sigma}_b$ and $\boldsymbol{\delta}_b = \mathbf{H}_{\boldsymbol{\Phi}_b}^{-1}\widetilde{\boldsymbol{\delta}}_b$ with $\widetilde{\boldsymbol{\delta}}_b = (\mathbf{b}_0', \mathbf{0}, \ldots, \mathbf{0})'$. Hence, its log-density is given by

$$\log p(\mathbf{b} \mid \boldsymbol{\Phi}_b, \boldsymbol{\Sigma}_b, \mathbf{b}_0) = -\frac{1}{2}(\mathbf{b}'\mathbf{H}_{\boldsymbol{\Phi}_b}'\mathbf{S}_b^{-1}\mathbf{H}_{\boldsymbol{\Phi}_b}\mathbf{b} - 2\mathbf{b}'\mathbf{H}_{\boldsymbol{\Phi}_b}'\mathbf{S}_b^{-1}\mathbf{H}_{\boldsymbol{\Phi}_b}\boldsymbol{\delta}_b) + c_1, \tag{16}$$

where c_1 is a constant independent of $\mathbf{b}$.

Next, we approximate $\log p(\mathbf{y} \mid \boldsymbol{\Lambda}, \mathbf{b}, \boldsymbol{\Sigma}_y)$ by a second-order Taylor expansion in $\mathbf{b}$. To that end, let b_{jt} denote the j-th element of $\mathbf{b}_t$. Now, expand $\log p(\mathbf{y} \mid \boldsymbol{\Lambda}, \mathbf{b}, \boldsymbol{\Sigma}_y) = \sum_{t=1}^{T} \log p(\mathbf{y}_t \mid \boldsymbol{\Lambda}, \mathbf{b}_t, \boldsymbol{\Sigma}_y)$ around a given point $\widetilde{\mathbf{b}} \in \mathbb{R}^{Tm}$ (e.g., the posterior mode):

$$\begin{aligned}\log p(\mathbf{y} \mid \boldsymbol{\Lambda}, \mathbf{b}, \boldsymbol{\Sigma}_y) \approx & \log(\mathbf{y} \mid \boldsymbol{\Lambda}, \widetilde{\mathbf{b}}, \boldsymbol{\Sigma}_y) + (\mathbf{b} - \widetilde{\mathbf{b}})'\mathbf{f} - \frac{1}{2}(\mathbf{b} - \widetilde{\mathbf{b}})'\mathbf{G}(\mathbf{b} - \widetilde{\mathbf{b}}) \\ = & -\frac{1}{2}(\mathbf{b}'\mathbf{G}\mathbf{b} - 2\mathbf{b}'(\mathbf{f} + \mathbf{G}\widetilde{\mathbf{b}})) + c_2,\end{aligned} \tag{17}$$

where c_2 is a constant independent of $\mathbf{b}$, $\mathbf{f} = (\mathbf{f}_1', \ldots, \mathbf{f}_T')'$ and $\mathbf{G} = \operatorname{diag}(\mathbf{G}_1, \ldots, \mathbf{G}_T)$ with

$$f_{jt} \equiv \frac{\partial}{\partial b_{jt}} \log p(\mathbf{y}_t \mid \boldsymbol{\Lambda}, \mathbf{b}_t, \boldsymbol{\Sigma}_y)|_{\mathbf{b}=\widetilde{\mathbf{b}}}, \qquad G_{jk,t} \equiv \frac{\partial^2}{\partial b_{jt}\partial b_{kt}} \log p(\mathbf{y}_t \mid \boldsymbol{\Lambda}, \mathbf{b}_t, \boldsymbol{\Sigma}_y)|_{\mathbf{b}=\widetilde{\mathbf{b}}}.$$

That is, $\mathbf{G}$ is block-diagonal (hence a band matrix) where the (j,k)-th element of the t-th block is $G_{jk,t}$. Note also that $\mathbf{G}$ is the negative Hessian of the log-density evaluated at $\widetilde{\mathbf{b}}$. It follows that by combining (16) and (17), we have

$$\begin{aligned}\log p(\mathbf{b} \mid \mathbf{y}, \boldsymbol{\Lambda}, \boldsymbol{\Phi}_b, \boldsymbol{\Sigma}_y, \boldsymbol{\Sigma}_b, \mathbf{b}_0) &= \log p(\mathbf{y} \mid \boldsymbol{\Lambda}, \mathbf{b}, \boldsymbol{\Sigma}_y) + \log p(\mathbf{b} \mid \boldsymbol{\Phi}_b, \boldsymbol{\Sigma}_b, \mathbf{b}_0) + c_3, \\ &\approx -\frac{1}{2}(\mathbf{b}'\mathbf{K}_\mathbf{b}\mathbf{b} - 2\mathbf{b}'\mathbf{k}_\mathbf{b}) + c_4,\end{aligned} \tag{18}$$

where c_3 and c_4 are constants independent of $\mathbf{b}$, $\mathbf{K}_\mathbf{b} = \mathbf{H}_{\boldsymbol{\Phi}_b}'\mathbf{S}_b^{-1}\mathbf{H}_{\boldsymbol{\Phi}_b} + \mathbf{G}$ and $\mathbf{k}_\mathbf{b} = \mathbf{f} + \mathbf{G}\widetilde{\mathbf{b}} + \mathbf{H}_{\boldsymbol{\Phi}_b}'\mathbf{S}_b^{-1}\mathbf{H}_{\boldsymbol{\Phi}_b}\boldsymbol{\delta}_b$. The expression in (18) is the log-kernel of the $\mathcal{N}(\widehat{\mathbf{b}}, \mathbf{K}_\mathbf{b}^{-1})$ density, where $\widehat{\mathbf{b}} = \mathbf{K}_\mathbf{b}^{-1}\mathbf{k}_\mathbf{b}$. Since $\mathbf{H}_{\boldsymbol{\Phi}_b}$, $\mathbf{S}_b$ and $\mathbf{G}$ are all band matrices, so is $\mathbf{K}_\mathbf{b}$. This Gaussian approximation can then used as the proposal density in the acceptance-rejection Metropolis-Hastings algorithm. For implementation details, see Appendix A.

8 Concluding Remarks and Future Research

We have proposed novel importance sampling algorithms for estimating the observed-data likelihoods under a variety of stochastic volatility models, with the goal of computing the observed-data DICs. It is illustrated via a Monte Carlo study that the observed-data DICs based on the proposed importance sampling estimators are able to select the correct model, whereas the conditional DICs tend to favor overfitted models. In the empirical application involving daily returns on the S&P 500, we find that according to the observed-data DIC, the leverage effect, t innovations, volatility feedback and moving

average components all seem to be useful additions to the standard SV model. Moreover, the marginal likelihood and the estimation results support the model ranking of the observed-data DIC but not that of the conditional DIC.

The proposed importance sampling estimators for observed-data likelihoods can be used in other settings, such as for developing more efficient MCMC algorithms (e.g., as an input for particle MCMC methods; see Andrieu, Doucet, and Holenstein 2010). We leave these possibilities for future research. In addition, we have only considered a few popular univariate stochastic volatility models. It would be useful to develop similar importance sampling algorithms for more complex multivariate stochastic volatility models, such as the time-varying parameter vector autoregression of Primiceri (2005). More broadly, our proposed approach can be applied to general nonlinear state space models, such as those discussed in Section 7. We also leave these extensions for future research.

Appendix A: Estimation Details

In this appendix we provide the estimation details for fitting the stochastic volatility models discussed in Section 3.1.

Standard Stochastic Volatility Model

Section 3.2 presents an outline of a Markov sampler for estimating the standard stochastic volatility model. Here we fill in the details of Step 1: sampling from the conditional density $p(\mathbf{h} \,|\, \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2)$. Following Chan (2015), we first obtain a Gaussian approximation of $p(\mathbf{h} \,|\, \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2)$ and use this approximation as a proposal density in the acceptance-rejection Metropolis-Hastings algorithm (see, e.g., Tierney, 1994), where candidate draws are obtained via the precision sampler in Chan and Jeliazkov (2009) instead of Kalman filter-based algorithms.

To approximate $p(\mathbf{h} \,|\, \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2)$ using a Gaussian density, note that

$$p(\mathbf{h} \,|\, \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2) \propto p(\mathbf{y} \,|\, \mu, \mathbf{h}) p(\mathbf{h} \,|\, \mu_h, \phi_h, \omega_h^2).$$

Hence, we first derive explicit expressions for the densities $p(\mathbf{y} \,|\, \mu, \mathbf{h})$ and $p(\mathbf{h} \,|\, \mu_h, \phi_h, \omega_h^2)$. It can be shown that the latter density is Gaussian (see, e.g. Chan, 2015). Let $\mathbf{H}_{\phi_h}$ be the following lower triangular matrix:

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

Then, we have $(\mathbf{h} \,|\, \mu_h, \phi_h, \omega_h^2) \sim \mathcal{N}(\boldsymbol{\delta}_h, (\mathbf{H}_{\phi_h}' \boldsymbol{\Sigma}_h^{-1} \mathbf{H}_{\phi_h})^{-1})$, where $\boldsymbol{\Sigma}_h = \text{diag}(\omega_h^2/(1-\phi_h^2), \omega_h^2, \ldots, \omega_h^2)$ and $\boldsymbol{\delta}_h = \mathbf{H}_{\phi_h}^{-1} \widetilde{\boldsymbol{\delta}}_h$ with $\widetilde{\boldsymbol{\delta}}_h = (\mu_h, (1-\phi_h)\mu_h, \ldots, (1-\phi_h)\mu_h)'$. Hence, its log-density is given by

$$\log p(\mathbf{h} \,|\, \mu_h, \phi_h, \omega_h^2) = -\frac{1}{2}(\mathbf{h}' \mathbf{H}_{\phi_h}' \boldsymbol{\Sigma}_h^{-1} \mathbf{H}_{\phi_h} \mathbf{h} - 2\mathbf{h}' \mathbf{H}_{\phi_h}' \boldsymbol{\Sigma}_h^{-1} \mathbf{H}_{\phi_h} \boldsymbol{\delta}_h) + c_5, \tag{19}$$

where c_5 is a constant independent of $\mathbf{h}$.

Next, we approximate $p(\mathbf{y} \,|\, \mu, \mathbf{h})$ by a Gaussian density in $\mathbf{h}$. To that end, expand $\log p(\mathbf{y} \,|\, \mu, \mathbf{h}) = \sum_{t=1}^T \log p(y_t \,|\, \mu, h_t)$ around a given point $\widetilde{\mathbf{h}} = (\widetilde{h}_1, \ldots, \widetilde{h}_T)' \in \mathbb{R}^T$ by a second-order Taylor expansion (the choice of $\widetilde{\mathbf{h}}$ is discussed below):

$$\begin{aligned} \log p(\mathbf{y} \,|\, \mu, \mathbf{h}) \approx & \log p(\mathbf{y} \,|\, \mu, \widetilde{\mathbf{h}}) + (\mathbf{h} - \widetilde{\mathbf{h}})' \mathbf{f} - \frac{1}{2}(\mathbf{h} - \widetilde{\mathbf{h}})' \mathbf{G} (\mathbf{h} - \widetilde{\mathbf{h}}) \\ = & -\frac{1}{2}(\mathbf{h}' \mathbf{G} \mathbf{h} - 2\mathbf{h}'(\mathbf{f} + \mathbf{G}\widetilde{\mathbf{h}})) + c_6, \end{aligned} \tag{20}$$

where c_6 is a constant independent of $\mathbf{h}$, $\mathbf{f} = (f_1, \ldots, f_T)'$ and $\mathbf{G} = \text{diag}(G_1, \ldots, G_T)$ with

$$f_t = \frac{\partial}{\partial h_t} \log p(y_t \,|\, \mu, h_t)|_{h_t = \widetilde{h}_t}, \qquad G_t = -\frac{\partial^2}{\partial h_t^2} \log p(y_t \,|\, \mu, h_t)|_{h_t = \widetilde{h}_t}.$$

That is, $\mathbf{G}$ is the negative Hessian of the log-density evaluated at $\widetilde{\mathbf{h}}$. For the standard stochastic volatility model, $\mathbf{G}$ is diagonal (hence a band matrix). In particular, since the log-density of y_t given μ and h_t is given by

$$\log p(y_t \,|\, \mu, h_t) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}h_t - \frac{1}{2}\mathrm{e}^{-h_t}(y_t - \mu)^2, \tag{21}$$

it is easy to check that

$$\begin{aligned}
\frac{\partial}{\partial h_t} \log p(y_t \,|\, \mu, h_t) &= -\frac{1}{2} + \frac{1}{2}\mathrm{e}^{-h_t}(y_t - \mu)^2, \\
\frac{\partial^2}{\partial h_t^2} \log p(y_t \,|\, \mu, h_t) &= -\frac{1}{2}\mathrm{e}^{-h_t}(y_t - \mu)^2.
\end{aligned}$$

Now, combining (19) and (20), we have

$$\begin{aligned}
\log p(\mathbf{h} \,|\, \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2) &= \log p(\mathbf{y} \,|\, \mu, \mathbf{h}) + \log p(\mathbf{h} \,|\, \mu_h, \phi_h, \omega_h^2) + c_7, \\
&\approx -\frac{1}{2}(\mathbf{h}'\mathbf{K}_\mathbf{h}\mathbf{h} - 2\mathbf{h}'\mathbf{k}_\mathbf{h}) + c_8,
\end{aligned} \tag{22}$$

where c_7 and c_8 are constants independent of $\mathbf{h}$, $\mathbf{K}_\mathbf{h} = \mathbf{H}_{\phi_h}'\boldsymbol{\Sigma}_h^{-1}\mathbf{H}_{\phi_h} + \mathbf{G}$ and $\mathbf{k}_\mathbf{h} = \mathbf{f} + \mathbf{G}\widetilde{\mathbf{h}} + \mathbf{H}_{\phi_h}'\boldsymbol{\Sigma}_h^{-1}\mathbf{H}_{\phi_h}\boldsymbol{\delta}_\mathbf{h}$. The expression in (22) is in fact the log-kernel of the $\mathcal{N}(\widehat{\mathbf{h}}, \mathbf{K}_\mathbf{h}^{-1})$ density, where $\widehat{\mathbf{h}} = \mathbf{K}_\mathbf{h}^{-1}\mathbf{k}_\mathbf{h}$ (see, e.g., Kroese and Chan, 2014, p. 238). Therefore, $p(\mathbf{h} \,|\, \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2)$ can be approximated by the Gaussian density with mean vector $\widehat{\mathbf{h}}$ and precision matrix $\mathbf{K}_\mathbf{h}$. It is important to note that $\mathbf{K}_\mathbf{h}$ is a band matrix; in fact, its nonzero elements appear only on the main diagonal and the diagonals above and below the main diagonal. Consequently, $\widehat{\mathbf{h}}$ can be computed quickly by solving the linear system $\mathbf{K}_\mathbf{h}\mathbf{x} = \mathbf{k}_\mathbf{h}$ for $\mathbf{x}$, and draws from $\mathcal{N}(\widehat{\mathbf{h}}, \mathbf{K}_\mathbf{h}^{-1})$ can be efficiently obtained using the precision sampler in Chan and Jeliazkov (2009). This Gaussian approximation is then used as the proposal density in the acceptance-rejection Metropolis-Hastings algorithm.

Finally, the point $\widetilde{\mathbf{h}}$ used in the Taylor expansion in (20) is chosen to be the mode of $p(\mathbf{h} \,|\, \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2)$, which can be quickly obtained by the Newton-Raphson method (see, e.g., Kroese et al., 2011, pp. 688-689). First, note that from (22) it follows that the negative Hessian of $\log p(\mathbf{h} \,|\, \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2)$ evaluated at $\mathbf{h} = \widetilde{\mathbf{h}}$ is $\mathbf{K}_\mathbf{h}$ and the gradient at $\mathbf{h} = \widetilde{\mathbf{h}}$ is $-\mathbf{K}_\mathbf{h}\widetilde{\mathbf{h}} + \mathbf{k}_\mathbf{h}$. Hence, we can implement the Newton-Raphson method as follows: initialize with $\mathbf{h} = \mathbf{h}^{(1)}$ for some constant vector $\mathbf{h}^{(1)}$. For $l = 1, 2, \ldots$, use $\widetilde{\mathbf{h}} = \mathbf{h}^{(l)}$ in the evaluation of $\mathbf{K}_\mathbf{h}$ and $\mathbf{k}_\mathbf{h}$, and compute

$$\mathbf{h}^{(l+1)} = \mathbf{h}^{(l)} + \mathbf{K}_\mathbf{h}^{-1}(-\mathbf{K}_\mathbf{h}\mathbf{h}^{(l)} + \mathbf{k}_\mathbf{h}) = \mathbf{K}_\mathbf{h}^{-1}\mathbf{k}_\mathbf{h}.$$

Repeat this procedure until some convergence criterion is reached, e.g., when $\|\mathbf{h}^{(l+1)} - \mathbf{h}^{(l)}\| < c$ for some prefixed tolerance level c.

Stochastic Volatility Model with AR(2) State Transition

Estimation of this variant with an AR(2) transition equation requires only minor modifications of the main algorithm for the standard stochastic volatility model. Specifically, let $\mathbf{H}_{\boldsymbol{\theta}_h}$ be the following lower triangular matrix:

$$\mathbf{H}_{\boldsymbol{\theta}_h} = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ -\rho_h & -\phi_h & 1 & 0 & \cdots & 0 \\ 0 & -\rho_h & -\phi_h & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -\rho_h & -\phi_h & 1 \end{pmatrix}.$$

Then, we can rewrite the state equation of h_t in (6) as:

$$\mathbf{H}_{\boldsymbol{\theta}_h}\mathbf{h} = \widetilde{\boldsymbol{\gamma}}_{\mathbf{h}} + \boldsymbol{\varepsilon}^h, \qquad \boldsymbol{\varepsilon}^h \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_h)$$

where $\boldsymbol{\varepsilon}^h = (\varepsilon_1^h, \ldots, \varepsilon_T^h)'$, $\widetilde{\boldsymbol{\gamma}}_h = (\mu_h, \mu_h, (1-\phi_h-\rho_h)\mu_h, \ldots, (1-\phi_h-\rho_h)\mu_h)'$ and $\mathbf{P}_h$ is a diagonal matrix in which the first two diagonal elements are the unconditional variance $(1-\rho_h)\omega_h^2(1+\rho_h)^{-1}((1-\rho_h)^2-\phi_h^2)^{-1}$ and the remaining $T-2$ elements equal ω_h^2. It follows that $(\mathbf{h} \,|\, \mu_h, \phi_h, \rho_h, \omega_h^2) \sim \mathcal{N}(\boldsymbol{\gamma}_h, (\mathbf{H}_{\boldsymbol{\theta}_h}'\mathbf{P}_h^{-1}\mathbf{H}_{\boldsymbol{\theta}_h})^{-1})$, where $\boldsymbol{\gamma}_h = \mathbf{H}_{\boldsymbol{\theta}_h}^{-1}\widetilde{\boldsymbol{\gamma}}_h$. Hence, we have

$$\log p(\mathbf{h} \,|\, \mu_h, \phi_h, \rho_h, \omega_h^2) = -\frac{1}{2}(\mathbf{h}'\mathbf{H}_{\boldsymbol{\theta}_h}'\mathbf{P}_h^{-1}\mathbf{H}_{\boldsymbol{\theta}_h}\mathbf{h} - 2\mathbf{h}'\mathbf{H}_{\boldsymbol{\theta}_h}'\mathbf{P}_h^{-1}\mathbf{H}_{\boldsymbol{\theta}_h}\boldsymbol{\gamma}_h) + c_8, \qquad (23)$$

where c_8 is a constant independent of $\mathbf{h}$. Therefore, we only need to replace (19) by (23), and the main algorithm for the standard stochastic volatility model can be directly applied. Minor modifications to the main algorithm are also needed to sample $\boldsymbol{\theta}_h$, μ_h and σ_h^2.

Stochastic Volatility Model with Jumps

To estimate the stochastic volatility model with jumps, a few modifications of the main algorithm are needed. Firstly, it is easy to see that the first and second derivatives of the conditional likelihood with respect to h_t are respectively

$$\begin{aligned} \frac{\partial}{\partial h_t}\log p(y_t \,|\, \mu, k_t, q_t, h_t) &= -\frac{1}{2} + \frac{1}{2}\mathrm{e}^{-h_t}(y_t - \mu - k_t q_t)^2, \\ \frac{\partial^2}{\partial h_t^2}\log p(y_t \,|\, \mu, k_t, q_t, h_t) &= -\frac{1}{2}\mathrm{e}^{-h_t}(y_t - \mu - k_t q_t)^2. \end{aligned}$$

Then, $\mathbf{h}$ can be sampled as before. Secondly, we need additional steps to sample $\mathbf{k} = (k_1, \ldots, k_T)'$, $\mathbf{q} = (q_1, \ldots, q_T)'$, κ and δ from the appropriate conditional distributions. Following Chib, Nardari, and Shephard (2006), we sample $\mathbf{k}$ and δ jointly as follows.

First, let $\zeta_t = \log(1+k_t)$ and stack $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_T)'$ over t. If k_t is small, as is the case for high frequency financial returns that are measured in decimals, $\exp(\zeta_t) \approx 1 + \zeta_t$, which implies $k_t q_t \approx \zeta_t q_t$. Recall that the prior for the jump size k_t is given by $\zeta_t = \log(1+k_t) \sim \mathcal{N}(-0.5\delta^2, \delta^2)$. Hence, we can integrate out ζ_t analytically. This allows us to sample δ marginally of $\boldsymbol{\zeta}$ from the density

$$p(\delta) \prod_{t=1}^{T} \phi(y_t \,|\, \mu - 0.5\delta^2 q_t, \delta^2 q_t^2 + \mathrm{e}^{h_t})$$

by the Metropolis-Hastings algorithm, where $p(\delta)$ is the prior density of δ and $\phi(x \,|\, a, b)$ is the Gaussian density with mean a and variance b evaluated at x. Once δ is sampled, we can draw $\zeta_1, \ldots, \zeta_T$ sequentially as follows: if q_t is zero, we sample ζ_t from the prior $\mathcal{N}(-0.5\delta^2, \delta^2)$; otherwise we sample from $\mathcal{N}(\widehat{\zeta}_t, D_{\zeta_t})$ where $D_{\zeta_t}^{-1} = \delta^{-2} + \mathrm{e}^{-h_t}$ and $\widehat{\zeta}_t = D_{\zeta_t}(-0.5 + \mathrm{e}^{-h_t}(y_t - \mu))$. Next, note that $q_1, \ldots, q_T$ are conditionally independent given the data and other parameters and they can be sampled individually. In particular, each q_t follows a Bernoulli distribution with

$$\begin{aligned}
\mathbb{P}(q_t = 1 \,|\, y_t, \zeta_t, h_t, \kappa) &\propto \kappa \phi(y_t \,|\, \mu + \mathrm{e}^{\zeta_t} - 1, \mathrm{e}^{h_t}) \\
\mathbb{P}(q_t = 0 \,|\, y_t, \zeta_t, h_t, \kappa) &\propto (1 - \kappa) \phi(y_t \,|\, \mu, \mathrm{e}^{h_t}).
\end{aligned}$$

Finally, we sample $(\kappa \,|\, \mathbf{q}) \sim \mathcal{B}(k_a + \sum_{t=1}^{T} q_t, k_b + T - \sum_{t=1}^{T} q_t)$.

Stochastic Volatility in Mean Model

To estimate the stochastic volatility in mean model, we only need to make two modifications of the main algorithm. Firstly, the first and second derivatives of the conditional likelihood with respect to h_t become

$$\begin{aligned}
\frac{\partial}{\partial h_t} \log p(y_t \,|\, \mu, \alpha, h_t) &= -\frac{1}{2} - \frac{1}{2}\alpha^2 \mathrm{e}^{h_t} + \frac{1}{2}\mathrm{e}^{-h_t}(y_t - \mu)^2, \\
\frac{\partial^2}{\partial h_t^2} \log p(y_t \,|\, \mu, \alpha, h_t) &= -\frac{1}{2}\alpha^2 \mathrm{e}^{h_t} - \frac{1}{2}\mathrm{e}^{-h_t}(y_t - \mu)^2.
\end{aligned}$$

Then, $\mathbf{h}$ can be sampled as before. Secondly, we replace Step 2 of the main algorithm with the joint sampling of (μ, α) from $p(\mu, \alpha \,|\, \mathbf{y}, \mathbf{h}, \mu_h, \phi_h, \omega_h^2) = p(\mu, \alpha \,|\, \mathbf{y}, \mathbf{h})$. To that end, let $\boldsymbol{\beta} = (\mu, \alpha)'$, $\mathbf{V}_{\boldsymbol{\beta}} = \mathrm{diag}(V_\mu, V_\alpha)$, $\boldsymbol{\beta}_0 = (\mu_0, \alpha_0)'$ and

$$\mathbf{X}_{\boldsymbol{\beta}} = \begin{pmatrix} 1 & \mathrm{e}^{h_1} \\ \vdots & \vdots \\ 1 & \mathrm{e}^{h_T} \end{pmatrix}.$$

Then, by standard results, we have $(\mu, \alpha \,|\, \mathbf{y}, \mathbf{h}) \sim \mathcal{N}(\widehat{\boldsymbol{\beta}}, \mathbf{D}_{\boldsymbol{\beta}})$, where $\mathbf{D}_{\boldsymbol{\beta}}^{-1} = \mathbf{V}_{\boldsymbol{\beta}}^{-1} + \mathbf{X}_{\boldsymbol{\beta}}' \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{X}_{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}} = \mathbf{D}_{\boldsymbol{\beta}}(\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}_0 + \mathbf{X}_{\boldsymbol{\beta}}' \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{y})$ with $\boldsymbol{\Sigma}_{\mathbf{y}} = \mathrm{diag}(\mathrm{e}^{h_1}, \ldots, \mathrm{e}^{h_T})$.

Stochastic Volatility Model with MA(1) Innovations

A few modifications of the main algorithm are needed to fit this variant with MA(1) innovations in the observation equation. Firstly, by appropriately transforming the data, we can sample $\mathbf{h}$ as before. Specifically, let

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

Then, (10) can be written as

$$\boldsymbol{\varepsilon}^{\mathbf{y}} = \mathbf{H}_\psi \mathbf{u}, \qquad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{y}}),$$

where $\boldsymbol{\Sigma}_{\mathbf{y}} = \text{diag}(\text{e}^{h_1}, \ldots, \text{e}^{h_T})$. Hence, if we transform the data $\mathbf{y}$ via $\widetilde{\mathbf{y}} = \mathbf{H}_\psi^{-1}(\mathbf{y} - \mu\mathbf{1})$, where $\mathbf{1}$ is a $T \times 1$ column of ones, then $(\widetilde{\mathbf{y}} \,|\, \mathbf{h}, \psi, \mu) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{y}})$. Therefore, by applying Step 1 to the transformed data $\widetilde{\mathbf{y}}$, we can sample $\mathbf{h}$ as before.

Next, to sample μ, observe that it follows from (9) and (10) that $(\mathbf{y} \,|\, \mu, \mathbf{h}, \psi) \sim \mathcal{N}(\mu\mathbf{1}, \boldsymbol{\Omega}_{\mathbf{y}})$, where $\boldsymbol{\Omega}_{\mathbf{y}} = \mathbf{H}_{\boldsymbol{\psi}} \boldsymbol{\Sigma}_{\mathbf{y}} \mathbf{H}_{\boldsymbol{\psi}}'$. Note that $\boldsymbol{\Omega}_{\mathbf{y}}$ is a band matrix with only a small number of non-zero elements along the main diagonal band. Consequently, computations involving $\boldsymbol{\Omega}_{\mathbf{y}}$ are fast. For computation details see Chan (2013). By standard linear regression results, we have $(\mu \,|\, \mathbf{y}, \mathbf{h}, \psi) \sim \mathcal{N}(\widehat{\mu}, D_\mu)$, where $D_\mu^{-1} = V_\mu^{-1} + \mathbf{1}'\boldsymbol{\Omega}_{\mathbf{y}}^{-1}\mathbf{1}$ and $\widehat{\mu} = (V_\mu^{-1}\mu_0 + \mathbf{1}'\boldsymbol{\Omega}_{\mathbf{y}}^{-1}\mathbf{y})$. Lastly, we sample ψ using an independence chain Metropolis-Hastings step as described in Chan (2013).

Stochastic Volatility Model with Leverage

A few modifications of the basic algorithm are needed to sample $(\mathbf{h} \,|\, \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2, \rho)$, where $\mathbf{h} = (h_1, \ldots, h_{T+1})'$ here is of length $T + 1$. Note that the conditional distribution of y_t given h_t, h_{t+1}, and the parameters is

$$(y_t \,|\, \mu, h_t, h_{t+1}, \mu_h, \phi_h, \rho, \omega_h^2) \sim \mathcal{N}\left(\mu + \frac{\rho}{\omega_h}\text{e}^{\frac{1}{2}h_t}(h_{t+1} - \phi_h h_t - \mu_h(1 - \phi_h)), \text{e}^{h_t}(1 - \rho^2)\right) \tag{24}$$

with log-density

$$\begin{aligned} \log p(y_t \,|\, \mu, h_t, h_{t+1}, \rho, \mu_h, \phi_h, \omega_h^2) = &-\frac{1}{2}\log(2\pi(1 - \rho^2)) - \frac{1}{2}h_t \\ &- \frac{1}{2(1 - \rho^2)}\text{e}^{-h_t}\left(y_t - \mu - \frac{\rho}{\omega_h}\text{e}^{\frac{1}{2}h_t}(h_{t+1} - \phi_h h_t - \mu_h(1 - \phi_h))\right)^2. \end{aligned}$$

For notational convenience, let $p_t = p(y_t \,|\, \mu, h_t, h_{t+1}, \rho, \omega_h^2)$. Then, the gradient and negative Hessian of the Gaussian approximation now take the form

$$\mathbf{f} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{T+1} \end{pmatrix}, \qquad \mathbf{G} = \begin{pmatrix} G_{11} & G_{12} & 0 & \cdots & 0 \\ G_{12} & G_{22} & G_{23} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & G_{T-1,T} & G_{TT} & G_{T,T+1} \\ 0 & \cdots & 0 & G_{T,T+1} & G_{T+1,T+1} \end{pmatrix},$$

where for $t = 2, \ldots, T+1$,

$$f_1 = \frac{\partial \log p_t}{\partial h_t}\bigg|_{\mathbf{h}=\widetilde{\mathbf{h}}}, \quad f_t = \frac{\partial}{\partial h_t}\left(\log p_t + \log p_{t-1}\right)\bigg|_{\mathbf{h}=\widetilde{\mathbf{h}}},$$

$$G_{11} = -\frac{\partial^2 \log p_t}{\partial h_t^2}\bigg|_{\mathbf{h}=\widetilde{\mathbf{h}}}, \quad G_{tt} = -\frac{\partial^2}{\partial h_t^2}\left(\log p_t + \log p_{t-1}\right)\bigg|_{\mathbf{h}=\widetilde{\mathbf{h}}}, \quad G_{t-1,t} = -\frac{\partial^2 \log p_t}{\partial h_t \partial h_{t+1}}\bigg|_{\mathbf{h}=\widetilde{\mathbf{h}}}.$$

It is easy to check that

$$\begin{aligned}
\frac{\partial \log p_t}{\partial h_t} =& -\frac{1}{2} - \frac{1}{2(1-\rho^2)}\left(-\mathrm{e}^{-h_t}(y_t-\mu)^2 - \frac{2\phi_h\rho^2}{\omega_h^2}(h_{t+1} - \phi_h h_t - \mu_h(1-\phi_h))\right. \\
&\left. + \frac{\rho}{\omega_h}\mathrm{e}^{-\frac{1}{2}h_t}(y_t-\mu)(h_{t+1} - \phi_h h_t - \mu_h(1-\phi_h) + 2\phi_h)\right), \\
\frac{\partial^2 \log p_t}{\partial h_t^2} =& -\frac{1}{2(1-\rho^2)}\left(\mathrm{e}^{-h_t}(y_t-\mu)^2 + \frac{2\phi_h^2\rho^2}{\omega_h^2}\right. \\
&\left. - \frac{\rho}{2\omega_h}\mathrm{e}^{-\frac{1}{2}h_t}(y_t-\mu)(h_{t+1} - \phi_h h_t - \mu_h(1-\phi_h) + 4\phi_h)\right), \\
\frac{\partial \log p_t}{\partial h_{t+1}} =& \frac{\rho}{\omega_h(1-\rho^2)}\mathrm{e}^{-\frac{1}{2}h_t}\left(y_t - \mu - \frac{\rho}{\omega_h}\mathrm{e}^{\frac{1}{2}h_t}(h_{t+1} - \phi_h h_t - \mu_h(1-\phi_h))\right), \\
\frac{\partial^2 \log p_t}{\partial h_{t+1}^2} =& -\frac{\rho^2}{\omega_h^2(1-\rho^2)}, \\
\frac{\partial^2 \log p_t}{\partial h_t \partial h_{t+1}} =& \frac{\rho}{\omega_h(1-\rho^2)}\left(\frac{\phi_h\rho}{\omega_h} - \frac{1}{2}\mathrm{e}^{-\frac{1}{2}h_t}(y_t-\mu)\right).
\end{aligned}$$

To sample ρ, note that the log conditional density of ρ is

$$\log p(\rho \,|\, \mathbf{y}, \mathbf{h}, \mu, \mu_h, \phi_h, \omega_h^2) \propto \log p(\rho) - \frac{T}{2}\log(1-\rho^2) - \frac{1}{2(1-\rho^2)}\left(k_1 - \frac{2\rho k_2}{\omega_h} + \frac{\rho^2 k_3}{\omega_h^2}\right),$$

where $p(\rho)$ is the prior density of ρ, $k_1 = \sum_{t=1}^T \mathrm{e}^{-h_t}(y_t-\mu)^2$, $k_2 = \sum_{t=1}^T \mathrm{e}^{-h_t/2}(y_t-\mu)\varepsilon_t^h$ and $k_3 = \sum_{t=1}^T (\varepsilon_t^h)^2$ with $\varepsilon_t^h = h_{t+1} - \phi_h h_t - \mu_h(1-\phi_h)$. Since ρ is bounded within the unit interval, one can sample ρ using the Griddy-Gibbs method.

The other parameters can be sampled similarly as the standard SV, by using the expression in (24). For example, the conditional distribution of μ is Gaussian:

$$(\mu \,|\, \mathbf{y}, \mathbf{h}, \rho, \mu_h, \phi_h, \omega_h^2) \sim \mathcal{N}(\widehat{\mu}, D_\mu),$$

where $D_\mu^{-1} = 1/V_\mu + (1-\rho^2)^{-1}\sum_{t=1}^T \mathrm{e}^{-h_t}$ and $\widehat{\mu} = D_\mu(\mu_0/V_\mu + (1-\rho^2)^{-1}\sum_{t=1}^T \mathrm{e}^{-h_t}(y_t - \rho\mathrm{e}^{\frac{1}{2}h_t}\varepsilon_t^h/\omega_h))$.

Stochastic Volatility Model with t Innovations

We only need a few modifications of the main algorithm to estimate this variant. Specifically, we replace (21) with

$$\log p(y_t \,|\, \mu, h_t, \lambda_t) = -\frac{1}{2}\log(2\pi\lambda_t) - \frac{1}{2}h_t - \frac{1}{2\lambda_t}\mathrm{e}^{-h_t}(y_t - \mu)^2.$$

Then, the same procedure can be applied to sample $\mathbf{h}$ jointly. Similarly, the full conditional density of μ now becomes

$$(\mu \,|\, \mathbf{y}, \mathbf{h}, \boldsymbol{\lambda}) \sim \mathcal{N}(\widehat{\mu}, D_\mu),$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_T)'$, $D_\mu^{-1} = V_\mu^{-1} + \sum_{t=1}^T \lambda_t^{-1}\mathrm{e}^{-h_t}$ and $\widehat{\mu} = D_\mu(V_\mu^{-1}\mu_0 + \sum_{t=1}^T \lambda_t^{-1}\mathrm{e}^{-h_t}y_t)$.

In addition, we need an extra block to sample the latent variables $\boldsymbol{\lambda}$. This can be easily done as $\lambda_1, \ldots, \lambda_T$ are conditionally independent given the parameters and data. In fact, each λ_t follows an independent inverse-gamma distribution:

$$(\lambda_t \,|\, y_t, \mu, h_t, \nu) \sim \mathcal{IG}\left(\frac{1}{2}(\nu+1), \frac{1}{2}\left(\nu + \mathrm{e}^{-h_t}(y_t - \mu)^2\right)\right).$$

Lastly, ν can be sampled by an independence-chain Metropolis-Hastings step with the proposal distribution $\mathcal{N}(\widehat{\nu}, K_\nu^{-1})$, where $\widehat{\nu}$ is the mode of $\log p(\nu \,|\, \boldsymbol{\lambda})$ and K_ν is the negative Hessian evaluated at the mode. For implementation details of this step, see Chan and Hsiao (2014).

Appendix B: Importance Sampling for Observed-Data Likelihoods

In this appendix we provide the details of the importance sampling algorithms. For the SV2, SVM, SVMA and SVL models, the only latent variables are the log-volatilties. For each of these models, we can use the Gaussian approximation of the conditional density of $\mathbf{h}$ given the data and other parameters as the importance density (see Appendix A for details). For example, under the SV2 model, we replace the prior density in (19) by (23) and the Gaussian approximation of $p(\mathbf{h} \,|\, \mathbf{y}, \mu, \mu_h, \phi_h, \rho_h, \omega_h^2)$ can be obtained following the same procedure as in Section 4. Moreover, all the Gaussian approximations can be quickly evaluated at any point as their precision matrices are all band matrices.

For the SVJ model, one needs to integrate out the log-volatilties $\mathbf{h}$, the jumps $\mathbf{q}$ and the jump sizes $\boldsymbol{\zeta}$ through importance sampling. Specifically, we simulate $\mathbf{h}$, $\mathbf{q}$ and $\boldsymbol{\zeta}$ as below. First, given the current posterior mode $\widehat{\mathbf{h}} = (\widehat{h}_1, \ldots, \widehat{h}_T)'$ and other parameters, we generate each q_t from the Bernoulli distribution with

$$\begin{aligned}
\mathbb{P}(q_t = 1) &\propto \kappa \phi(y_t \,|\, \mu - 0.5\delta^2, \delta^2 + \mathrm{e}^{\widehat{h}_t}) \\
\mathbb{P}(q_t = 0) &\propto (1-\kappa) \phi(y_t \,|\, \mu, \mathrm{e}^{\widehat{h}_t}).
\end{aligned}$$

Then, given the simulated draw $\mathbf{q}^*$, we draw $\zeta_1, \ldots, \zeta_T$ sequentially as follows: if q_t^* is zero, we sample ζ_t from the prior $\mathcal{N}(-0.5\delta^2, \delta^2)$; otherwise we sample from $\mathcal{N}(\widehat{\zeta}_t, D_{\zeta_t})$ where $D_{\zeta_t}^{-1} = \delta^{-2} + \mathrm{e}^{-\widehat{h}_t}$. Lastly, given $\mathbf{q}^*$ and $\boldsymbol{\zeta}^*$, we generate a draw from the Gaussian approximation of $p(\mathbf{h} \,|\, \mathbf{y}, \mu, \boldsymbol{\zeta}^*, \mathbf{q}^*, \mu_h, \phi_h, \omega_h^2)$, obtained as described in Appendix A. In this case, it is also easy to evaluate the importance density, which is a product of Bernoulli and Gaussian densities.

Finally, we consider the integrated likelihood evaluation for the SVt model. In this case, we need to integrate out both $\boldsymbol{\lambda}$ and $\mathbf{h}$. It turns out that we can integrate out $\boldsymbol{\lambda}$ analytically, and $\mathbf{h}$ is then integrated out by importance sampling as discussed above. It can be shown that the partial conditional likelihood (marginal of $\boldsymbol{\lambda}$) is given by

$$p(\mathbf{y} \,|\, \mathbf{h}, \mu, \nu) = (\nu\pi)^{-\frac{T}{2}} \mathrm{e}^{-\frac{1}{2}\sum_{t=1}^T h_t} \left(\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \right)^T \prod_{t=1}^T \left(1 + \frac{1}{\nu} \mathrm{e}^{-h_t} (y_t - \mu)^2 \right)^{-\frac{\nu+1}{2}}.$$

To obtain an importance sampling density, we note that in this case the ideal zero-variance importance sampling density is the conditional density $p(\mathbf{h} \,|\, \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2, \nu)$ marginal of $\boldsymbol{\lambda}$. As before, we approximate this with a Gaussian density with mean vector $\widehat{\mathbf{h}}$ and precision matrix $\mathbf{K}_{\mathbf{h}}$, where $\widehat{\mathbf{h}}$ and $\mathbf{K}_{\mathbf{h}}$ are respectively the mode and the negative Hessian evaluated at the mode of $\log p(\mathbf{h} \,|\, \mathbf{y}, \mu, \mu_h, \phi_h, \omega_h^2, \nu)$.

References

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

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

A. Berg, R. Meyer, and J. Yu. Deviance information criterion for comparing stochastic volatility models. *Journal of Business and Economic Statistics*, 22(1):107–120, 2004.

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

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

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

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

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

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

J. C. C. Chan and A. L. Grant. Fast computation of the deviance information criterion for latent variable models. *Computational Statistics and Data Analysis*, 2014. Forthcoming.

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

J. C. C. Chan and I. Jeliazkov. Efficient simulation and integrated likelihood estimation in state space models. *International Journal of Mathematical Modelling and Numerical Optimisation*, 1(1/2):101–120, 2009.

J. C. C. Chan, G. Koop, and S. M. Potter. A new model of trend inflation. *Journal of Business and Economic Statistics*, 31(1):94–106, 2013.

S. Chib, F. Nardari, and N. Shephard. Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics*, 108(2):281–316, 2002.

S. Chib, F. Nardari, and N. Shephard. Analysis of high dimensional multivariate stochastic volatility models. *Journal of Econometrics*, 134(2):341–371, 2006.

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

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

J. Durbin and S. J. Koopman. Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, 84(3):669–684, 1997.

E. Eisenstat and R. W. Strachan. Modelling inflation volatility. *CAMA Working Paper 24/2014*, 2014.

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

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

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

M. J. Jensen and J. M. Maheu. Bayesian semiparametric stochastic volatility modeling. *Journal of Econometrics*, 157(2):306–316, 2010.

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

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

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

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

Y. Li and J. Yu. Bayesian hypothesis testing in latent variable models. *Journal of Econometrics*, 166(2):237–246, 2012.

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

Y. Li, T. Zeng, and J. Yu. A new approach to Bayesian hypothesis testing. *Journal of Econometrics*, 178:602–612, 2014.

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

W. J. McCausland, S. Miller, and D. Pelletier. Simulation smoothing for state-space models: A computational efficiency analysis. *Computational Statistics and Data Analysis*, 55(1):199–212, 2011.

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

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

H. Mumtaz and F. Zanetti. The impact of the volatility of monetary policy shocks. *Journal of Money, Credit and Banking*, 45(4):535–558, 2013.

J. Nakajima and Y. Omori. Stochastic volatility model with leverage and asymmetrically heavy-tailed error using GH skew Student's t-distribution. *Computational Statistics and Data Analysis*, 56(11):3690–3704, 2012.

Y. Omori, S. Chib, N. Shephard, and J. Nakajima. Stochastic volatility with leverage: Fast and efficient likelihood inference. *Journal of Econometrics*, 140(2):425–449, 2007.

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

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

H. Rue. Fast sampling of Gaussian Markov random fields with applications. *Journal of the Royal Statistical Society Series B*, 63(2):325–338, 2001.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace. *Journal of the Royal Statistical Society Series B*, 71(2):319–392, 2009.

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

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

J. H. Stock and M. W. Watson. Forecasting with many predictors. In C.W.J. Granger G. Elliott and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 1, pages 515–554. Elsevier, 2006.

L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994.

G. Tsiotas. On generalised asymmetric stochastic volatility models. *Computational Statistics and Data Analysis*, 56(1):151–172, 2012.

M. Vo. Oil and stock market volatility: A multivariate stochastic volatility perspective. *Energy Economics*, 33(5):956–965, 2011.

J. J. J. Wang, S. T. B. Choy, and J. S. K. Chan. Modelling stochastic volatility using generalized t distribution. *Journal of Statistical Computation and Simulation*, 83(2): 340–354, 2013.

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

Jun Yu. On leverage in a stochastic volatility model. *Journal of Econometrics*, 127(2): 165–178, 2005.