Pitfalls of Estimating the Marginal Likelihood Using the Modified Harmonic Mean

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

Angelia L. Grant
Centre for Applied Macroeconomic Analysis,
Australian National University

March 2015

Abstract

The modified harmonic mean is widely used for estimating the marginal likelihood. We investigate the empirical performance of two versions of this estimator: one based on the observed-data likelihood and the other on the complete-data likelihood. Through an empirical example using US and UK inflation, we show that the version based on the complete-data likelihood has a substantial finite sample bias and tends to select the wrong model, whereas the version based on the observed-data likelihood works well.

Keywords: Bayesian model comparison, state space, unobserved components, inflation

JEL classification: : C11, C15, C32, C52

*Joshua Chan would like to acknowledge financial support by the Australian Research Council via a Discovery Early Career Researcher Award (DE150100795).

1 Introduction

The marginal likelihood or marginal data density is a widely used Bayesian model selection criterion and its estimation has generated a large literature. One popular method for its estimation is the modified harmonic mean estimator of Gelfand and Dey (1994) (for recent applications in economics, see, e.g., Koop and Potter, 2010; Liu, Waggoner, and Zha, 2011; Lanne, Luoma, and Luoto, 2012; Bianchi, 2013). For latent variable models such as state space and regime-switching models, this estimator is often used in conjunction with the complete-data likelihood—i.e., the joint density of the data and the latent variables given the parameters. Recent examples include Berg, Meyer, and Yu (2004), Justiniano and Primiceri (2008) and Jochmann, Koop, and Potter (2010).

This paper first introduces a new variant of the unobserved components model where the marginal likelihood can be computed analytically. Then, through a real data example we show that the Gelfand-Dey estimator based on the complete-data likelihood has a substantial finite sample bias and tends to select the wrong model, whereas the estimator based on the observed-data likelihood works fine. This finding is perhaps not surprising as the complete-data likelihood is typically very high-dimensional, which makes the corresponding estimator unstable. Our results complement findings in Li, Zeng, and Yu (2012) and Chan and Grant (2014a), who argue against the use of the complete-data likelihood in a related context of computing the deviance information criterion.

The rest of this paper is organized as follows. Section 2 discusses the Bayes factor and the Gelfand-Dey estimators. Section 3 introduces the unobserved components model and outlines the analytical computation of the marginal likelihood. Then, using US and UK CPI inflation data, we compare the two Gelfand-Dey estimators with the analytical results.

2 Bayes Factor and Marginal Likelihood

In this section we give an overview of Bayesian model comparison and discuss the method of Gelfand and Dey (1994) for estimating the marginal likelihood. To set the stage, suppose we wish to compare a collection of models $\{M_1, \ldots, M_K\}$. 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 $p(\boldsymbol{\theta}_k \mid M_k)$. One popular Bayesian model comparison criterion is the *Bayes factor* in favor of M_i against M_j, defined as

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

where $p(\mathbf{y} \mid M_k) = \int p(\mathbf{y} \mid \boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k \mid M_k) \mathrm{d}\boldsymbol{\theta}_k$ is the *marginal likelihood* under model M_k, $k = i, j$, which is simply the marginal data density under model M_k evaluated at the observed data $\mathbf{y}$. Hence, if the observed data are likely under the model, the associated marginal likelihood would be “large” and vice versa. It follows that $\mathrm{BF}_{ij} > 1$ indicates evidence in favor of model M_i against M_j, and the weight of evidence is proportional to the value of the Bayes factor.

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

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

where $\mathbb{P}(M_i)/\mathbb{P}(M_j)$ is the prior odds ratio. If both models are equally probable *a priori*, i.e., $p(M_i) = p(M_j)$, the posterior odds ratio between the two models is then equal to the Bayes factor. In that case, if, for example, $\mathrm{BF}_{ij} = 20$, then model M_i is 20 times more likely than model M_j given the data. For a more detailed discussion of the Bayes factor and its role in Bayesian model comparison, see Koop (2003) or Kroese and Chan (2014).

The Bayes factor therefore has a natural interpretation. Moreover, using it to compare models, we need only to obtain the marginal likelihoods of the competing models. One popular method for estimating the marginal likelihood of a given model $p(\mathbf{y})$—we suppress the model index from here onwards for notational convenience—is due to Gelfand and Dey (1994). Specifically, they realize that for any probability density function f with support contained in the support of the posterior density, we have the following identity:

$$\mathbb{E}\left(\frac{f(\boldsymbol{\theta})}{p(\boldsymbol{\theta})p(\mathbf{y} \,|\, \boldsymbol{\theta})} \,\middle|\, \mathbf{y}\right) = \int \frac{f(\boldsymbol{\theta})}{p(\boldsymbol{\theta})p(\mathbf{y} \,|\, \boldsymbol{\theta})} \frac{p(\boldsymbol{\theta})p(\mathbf{y} \,|\, \boldsymbol{\theta})}{p(\mathbf{y})} \mathrm{d}\boldsymbol{\theta} = p(\mathbf{y})^{-1}, \tag{1}$$

where the expectation is taken with respect to $p(\boldsymbol{\theta} \,|\, \mathbf{y}) = p(\boldsymbol{\theta})p(\mathbf{y} \,|\, \boldsymbol{\theta})/p(\mathbf{y})$. Therefore, one can estimate $p(\mathbf{y})$ using the following estimator:

$$\mathrm{GD_o} = \left\{\frac{1}{R}\sum_{i=1}^{R} \frac{f(\boldsymbol{\theta}_i)}{p(\boldsymbol{\theta}_i)p(\mathbf{y} \,|\, \boldsymbol{\theta}_i)}\right\}^{-1}, \tag{2}$$

where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R$ are posterior draws. Note that this estimator is simulation consistent in the sense that it converges to $p(\mathbf{y})$ in probability as R tends to infinity, but it is not unbiased—i.e., $\mathbb{E}(\mathrm{GD_o}) \neq p(\mathbf{y})$ in general.

Geweke (1999) shows that if the tuning function f has tails lighter than those of the posterior density, the estimator in (2) then has a finite variance. One such tuning function is a normal approximation of the posterior density with tail truncations determined by asymptotic arguments. Specifically, let $\widehat{\boldsymbol{\theta}}$ and $\mathbf{Q}_{\boldsymbol{\theta}}$ denote the posterior mean and covariance matrix respectively. Then, f is set to be the $\mathcal{N}(\widehat{\boldsymbol{\theta}}, \mathbf{Q}_{\boldsymbol{\theta}})$ density truncated within the region

$$\{\boldsymbol{\theta} \in \mathbb{R}^m : (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})'\mathbf{Q}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) < \chi^2_{\alpha,m}\},$$

where $\chi^2_{\alpha,m}$ is the $(1-\alpha)$ quantile of the χ^2_m distribution and m is the dimension of $\boldsymbol{\theta}$.

For many complex models where the likelihood $p(\mathbf{y} \,|\, \boldsymbol{\theta})$ cannot be evaluated analytically, estimation is often facilitated by data augmentation. Specifically, the model $p(\mathbf{y} \,|\, \boldsymbol{\theta})$ is augmented with a vector of latent variables $\mathbf{z}$ such that

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

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

One advantage of this augmented representation is that the complete-data likelihood $p(\mathbf{y}, \mathbf{z} \,|\, \boldsymbol{\theta}) = p(\mathbf{y} \,|\, \mathbf{z}, \boldsymbol{\theta})p(\mathbf{z} \,|\, \boldsymbol{\theta})$ is easy to evaluate by construction. Given this latent variable representation, one can use a similar argument as in (1) to obtain another estimator of the marginal likelihood $p(\mathbf{y})$ that avoids the evaluation of the observed-data likelihood $p(\mathbf{y} \,|\, \boldsymbol{\theta})$:

$$\mathrm{GD_c} = \left\{ \frac{1}{R} \sum_{i=1}^{R} \frac{f(\boldsymbol{\theta}_i, \mathbf{z}_i)}{p(\mathbf{y}, \mathbf{z}_i \,|\, \boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)} \right\}^{-1}, \tag{3}$$

where $(\boldsymbol{\theta}_1, \mathbf{z}_1), \ldots, (\boldsymbol{\theta}_R, \mathbf{z}_R)$ are posterior draws from the augmented model $p(\boldsymbol{\theta}, \mathbf{z} \,|\, \mathbf{y})$ and f is a tuning function.

However, the variance of $\mathrm{GD_c}$ is generally much larger than that of $\mathrm{GD_o}$. Moreover, the former estimator is expected to perform poorly in general—the key difficulty is to obtain a suitable tuning function f that is typically very high-dimensional. In fact, in the next section we give an example where we can compute the marginal likelihood analytically. We show that $\mathrm{GD_c}$ gives estimates that are quite different from the analytical results.

3 Application: Estimating Trend Inflation

In this section we consider a version of the unobserved components model where its marginal likelihood can be computed analytically. The analytical result is then compared to the estimates obtained by the method of Gelfand and Dey (1994) based on the complete-data and observed-data likelihoods. We use this example to investigate the empirical performance of the two estimators.

3.1 Unobserved Components Model

Consider the following unobserved components model:

$$y_t = \tau_t + \varepsilon_t, \tag{4}$$

$$\tau_t = \tau_{t-1} + u_t, \tag{5}$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ and $u_t \sim \mathcal{N}(0, g\sigma^2)$ are independent for a fixed g, with initial condition $\tau_1 \sim \mathcal{N}(0, \sigma^2 V_\tau)$ for a fixed V_τ. Note that here the error variance of the state equation (5) is assumed to be a fraction—controlled by g—of the error variance in the measurement equation (4). The states are the unobserved components $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_T)'$ and the only parameter is σ^2. In the next section we compute the marginal likelihood under different values of g.

Stacking $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)'$ and $\mathbf{y} = (y_1, \ldots, y_T)'$, the measurement equation (4) can be written as:

$$\mathbf{y} = \boldsymbol{\tau} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_T)$. Hence, the conditional likelihood of the model (4)–(5) is given by

$$f(\mathbf{y} \,|\, \boldsymbol{\tau}, \sigma^2) = (2\pi\sigma^2)^{-\frac{T}{2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(\mathbf{y}-\boldsymbol{\tau})'(\mathbf{y}-\boldsymbol{\tau})}. \tag{6}$$

In the Appendix we derive the joint density of the states $p(\boldsymbol{\tau} \,|\, \sigma^2)$ and show that the complete-data likelihood is

$$p(\mathbf{y}, \boldsymbol{\tau} \,|\, \sigma^2) = (2\pi\sigma^2)^{-T} |\mathbf{S_u}|^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2\sigma^2}\left((\mathbf{y}-\boldsymbol{\tau})'(\mathbf{y}-\boldsymbol{\tau}) + \boldsymbol{\tau}'\mathbf{H}'\mathbf{S_u}^{-1}\mathbf{H}\boldsymbol{\tau}\right)}, \tag{7}$$

where $\mathbf{S_u} = \mathrm{diag}(V_\tau, g, \ldots, g)$ and

$$\mathbf{H} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ -1 & 1 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & -1 & 1 \end{pmatrix}. \tag{8}$$

Note that $\mathbf{H}$ is a band matrix with unit determinant and the density in (7) can be evaluated quickly using band matrix routines—e.g., using methods similar to those in Chan and Grant (2014a). We also show in the Appendix that the observed-data likelihood has the following expression:

$$p(\mathbf{y} \,|\, \sigma^2) = \int f(\mathbf{y} \,|\, \boldsymbol{\tau}, \sigma^2) p(\boldsymbol{\tau} \,|\, \sigma^2) \mathrm{d}\boldsymbol{\tau} = (2\pi\sigma^2)^{-\frac{T}{2}} |\mathbf{S_u}|^{-\frac{1}{2}} |\mathbf{K}_{\boldsymbol{\tau}}|^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2\sigma^2}\left(\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{K}_{\boldsymbol{\tau}}^{-1}\mathbf{y}\right)},$$

where $\mathbf{K}_{\boldsymbol{\tau}} = \mathbf{I}_T + \mathbf{H}'\mathbf{S_u}^{-1}\mathbf{H}$. Lastly, if we assume an inverse-gamma prior for σ^2, i.e., $\sigma^2 \sim \mathcal{IG}(\nu_0, S_0)$, one can show that the marginal likelihood of the model (4)–(5) has the following analytical expression (see Appendix for derivations):

$$\begin{aligned} p(\mathbf{y}) &= \int f(\mathbf{y}, \boldsymbol{\tau} \,|\, \sigma^2) p(\sigma^2) \mathrm{d}(\boldsymbol{\tau}, \sigma^2) \\ &= (2\pi)^{-\frac{T}{2}} |\mathbf{S_u}|^{-\frac{1}{2}} |\mathbf{K}_{\boldsymbol{\tau}}|^{-\frac{1}{2}} \frac{{S_0}^{\nu_0} \Gamma\left(\frac{T}{2} + \nu_0\right)}{\Gamma(\nu_0)} \left(S_0 + \frac{\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{K}_{\boldsymbol{\tau}}^{-1}\mathbf{y}}{2}\right)^{-\left(\frac{T}{2}+\nu_0\right)}. \end{aligned} \tag{9}$$

Again both the observed-data and marginal likelihoods can be evaluated quickly using band matrix routines. We set the hyperparameters to be $\nu_0 = 5$ and $S_0 = 4$.

For the $\mathrm{GD_o}$ estimator based on the observed-data likelihood, the tuning function f is a single-variable function that depends only on σ^2. We use a univariate truncated normal density suggested by Geweke (1999) (see the discussion in Section 2) and denote it as f_{σ^2}. For the $\mathrm{GD_c}$ estimator based on the complete-data likelihood, the tuning function f needs to be a $(T+1)$-dimensional function—it depends on σ^2 and $\boldsymbol{\tau}$. Using a multivariate truncated normal density as suggested in Geweke (1999) might not be feasible as the covariance matrix alone typically involves tens of thousands of parameters.

Here we follow the proposal in Justiniano and Primiceri (2008). Specifically, we use $f(\sigma^2, \boldsymbol{\tau}) = f_{\sigma^2}(\sigma^2)p(\boldsymbol{\tau} \,|\, \sigma^2)$, where f_{σ^2} is a univariate truncated normal density discussed above, and $p(\boldsymbol{\tau} \,|\, \sigma^2)$ is the prior density of $\boldsymbol{\tau}$ given in (10). Using the prior density as part of the tuning function seems reasonable since it encapsulates our prior belief about $\boldsymbol{\tau}$. In addition, this choice simplifies the computation. The $\mathrm{GD_c}$ estimator in (3) now becomes

$$\mathrm{GD_c} = \left\{ \frac{1}{R} \sum_{i=1}^{R} \frac{f_{\sigma^2}(\sigma_i^2)}{p(\mathbf{y} \,|\, \sigma_i^2, \boldsymbol{\tau}_i)p(\sigma_i^2)} \right\}^{-1},$$

where $(\sigma_1^2, \boldsymbol{\tau}_1), \ldots, (\sigma_R^2, \boldsymbol{\tau}_R)$ are posterior draws.

3.2 Data and Empirical Results

Unobserved components models have been widely used to analyze inflation data (see, e.g., Stock and Watson, 2007; Chan, 2013; Clark and Doh, 2014). We fit the unobserved components model in (4)-(5) using US and UK annualized quarterly CPI inflation rates. The US sample is obtained from the Federal Reserve Bank of St. Louis economic database and runs from 1948Q1 to 2013Q4; the UK data are from 1955Q1 to 2013Q4 and are obtained from the OECD economic database.

It can be shown that $(\boldsymbol{\tau}, \sigma^2 \,|\, \mathbf{y})$ has a normal-inverse-gamma distribution and hence independent draws from the posterior distribution can be easily obtained (see Appendix for details). Posterior analysis is based on R independent draws from the posterior distribution, divided into 10 batches. For each fixed g, we compute the log marginal likelihood of the model in three ways: the analytical expression given in (9), the $\mathrm{GD_o}$ and $\mathrm{GD_c}$ estimators based on, respectively, the observed-data and the complete-data likelihoods. The results for the US are reported in Table 1.

Table 1: Log marginal likelihoods based on analytical results, the Gelfand-Dey estimators using the observed-data likelihood ($\mathrm{GD_o}$) and the complete-data likelihood ($\mathrm{GD_c}$) for US data. Numerical standard errors are in parenthesis.

	$g = 0.8$	$g = 0.9$	$g = 1$	$g = 1.1$	$g = 1.2$	$g = 1.3$
analytical result	-592.141	-591.994	-591.936	-591.940	-591.990	-592.073
$\mathrm{GD_o}$ ($R = 50000$)	-592.140	-591.995	-591.935	-591.941	-591.991	-592.070
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$\mathrm{GD_c}$ ($R = 50000$)	-503.28	-498.89	-494.62	-487.91	-485.08	-481.52
	(1.09)	(0.97)	(1.32)	(0.92)	(0.77)	(1.03)
$\mathrm{GD_c}$ ($R = 10^7$)	-510.64	-506.51	-502.70	-498.19	-493.53	-489.79
	(0.49)	(0.50)	(0.95)	(1.18)	(0.44)	(0.91)

A few observations can be drawn from this exercise. Firstly, it is clear that the $\mathrm{GD_o}$ estimates based on the observed-data likelihood are essentially identical to the analytical results, whereas the $\mathrm{GD_c}$ estimates based on the complete-data likelihood are substantially different. For example, when $g = 1$ and $R = 50000$, the $\mathrm{GD_c}$ estimate is around -494.6 with a numerical standard error of 1.32, but the true value is around -591.9—a difference of 97.3 in the logarithm scale. When we increase the number of replications to 10 million ($R = 10^7$), the difference between the estimate and the true value becomes smaller, but remains substantial (-502.7 versus -591.9).[1] In principle the Gelfand-Dey estimator based on the complete-data likelihood estimator is simulation consistent, but in practice the finite sample bias can remain huge even for tens of millions of replications.

It is also worth mentioning that when the $\mathrm{GD_c}$ estimator fails, its numerical standard error is also severely underestimated. For example, for the case $g = 1$ and $R = 50000$, the difference between the estimate and the true value is 97.3, but the estimated numerical standard error is only 1.32. A researcher using $\mathrm{GD_c}$ with this markedly underestimated numerical standard error is likely to be led astray. To guard against this problem, we suggest the following diagnostics: if the numerical standard error is correctly estimated, doubling the simulation size[2] should reduce the numerical standard error by $\sqrt{2}$. If this does not happen, it is a warning sign that the numerical standard error might be underestimated. In our example when the simulation size increases from 50000 to 10 million (two-hundredfold), the numerical standard error should be about 0.093 ($1.32/\sqrt{200}$), but instead the estimate is 0.95—this inconsistency indicates a problem.

On the other hand, the numerical standard errors for estimates based on the observed-data likelihood are between 500-1000 times smaller than those based on the complete-data likelihood. Moreover, the former are still wide enough such that the difference between the estimate and the true value is within two times the numerical standard error. Finally, $\mathrm{GD_c}$ most prefers the model with $g = 1.3$, even though the best model is $g = 1$.

Table 2: Log marginal likelihoods based on analytical results, the Gelfand-Dey estimators using the observed-data likelihood ($\mathrm{GD_o}$) and the complete-data likelihood ($\mathrm{GD_c}$) for UK data. Numerical standard errors are in parenthesis.

	$g = 0.05$	$g = 0.1$	$g = 0.15$	$g = 0.2$	$g = 0.25$	$g = 0.3$
analytical result	-673.362	-672.055	-672.515	-673.432	-674.489	-675.575
$\mathrm{GD_o}$ ($R = 50000$)	-673.361	-672.055	-672.514	-673.432	-674.490	-675.575
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$\mathrm{GD_c}$ ($R = 50000$)	-653.74	-647.37	-644.00	-640.35	-638.75	-636.24
	(0.58)	(0.76)	(0.67)	(0.68)	(0.87)	(0.67)
$\mathrm{GD_c}$ ($R = 10^7$)	-656.28	-650.60	-647.44	-645.00	-642.84	-640.32
	(0.36)	(0.45)	(0.47)	(0.36)	(0.49)	(0.29)

[1] For a logit regression with random effects, Frühwirth-Schnatter and Wagner (2008) compute its marginal likelihood using the Chib's method (Chib, 1995) based on the complete-data likelihood. They find that "this estimator is extremely inaccurate" and "an upward bias seems to be present," similar to what we find in this example.

[2] When the replications are not independent, e.g., when they are obtained from Markov chain Monte Carlo methods, one should use the effective sample size instead—i.e., the simulation size adjusted by the inefficiency factor.

We repeat the exercise using UK inflation data and the results are reported in Table 2. The general conclusions remain the same: the estimates based on the observed-data likelihood are virtually the same as the analytical results, whereas those based on the complete-data likelihood have a substantial finite sample bias, even when the number of replications is 10 million. In addition, the numerical standard errors associated with $\mathrm{GD_c}$ also seem to be severely underestimated.

4 Concluding Remarks

By fitting a new unobserved components model using US and UK inflation data, we show that the Gelfand-Dey estimates based on the complete-data likelihood can be very different from the analytical results. Given this finding, we argue against the use of the complete-data likelihood in conjunction with the Gelfand-Dey estimator.

An alternative is to use the Gelfand-Dey estimator based on the observed-data likelihood. Some recent papers, such as McCausland (2012), Chan and Grant (2014a) and Chan and Grant (2014b), have provided fast routines to evaluate the observed-data likelihood for various latent variable models. Frühwirth-Schnatter and Wagner (2008) describe a different approach—based on the auxiliary mixture sampling in conjunction with importance sampling and bridge sampling—that can be used to compute the marginal likelihood when the observed-data likelihood is not readily available.

Appendix

In this appendix we provide the details of the derivation of the complete-data, observed-data and marginal likelihoods of the unobserved components model in (4)–(5).

We first derive the prior density of the states $p(\boldsymbol{\tau}\,|\,\sigma^2)$. To that end, we stack the state equation (5) over t:

$$\mathbf{H}\boldsymbol{\tau} = \mathbf{u},$$

where $\mathbf{u} = (u_1, \ldots, u_T)' \sim \mathcal{N}(0, \sigma^2\mathbf{S_u})$, $\mathbf{S_u} = \mathrm{diag}(V_\tau, g, \ldots, g)$ and $\mathbf{H}$ is as defined in (8).

It follows that $(\boldsymbol{\tau}\,|\,\sigma^2) \sim \mathcal{N}(\mathbf{0}, \sigma^2\,(\mathbf{H}'\mathbf{S_u^{-1}}\mathbf{H})^{-1})$, and hence

$$p(\boldsymbol{\tau}\,|\,\sigma^2) = (2\pi\sigma^2)^{-\frac{T}{2}}|\mathbf{S_u}|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2\sigma^2}\boldsymbol{\tau}'\mathbf{H}'\mathbf{S_u^{-1}}\mathbf{H}\boldsymbol{\tau}}. \tag{10}$$

The complete-data likelihood is simply the product of the densities in (6) and (10):

$$p(\mathbf{y}, \boldsymbol{\tau}\,|\,\sigma^2) = p(\mathbf{y}\,|\,\boldsymbol{\tau}, \sigma^2)p(\boldsymbol{\tau}\,|\,\sigma^2) = (2\pi\sigma^2)^{-T}|\mathbf{S_u}|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2\sigma^2}\left((\mathbf{y}-\boldsymbol{\tau})'(\mathbf{y}-\boldsymbol{\tau})+\boldsymbol{\tau}'\mathbf{H}'\mathbf{S_u^{-1}}\mathbf{H}\boldsymbol{\tau}\right)}.$$

Moreover, after some computations, the observed-data likelihood is obtained as follows:

$$
\begin{aligned}
p(\mathbf{y}\,|\,\sigma^2) &= \int p(\mathbf{y},\boldsymbol{\tau}\,|\,\sigma^2)\mathrm{d}\boldsymbol{\tau}\\
&= (2\pi\sigma^2)^{-T}|\mathbf{S}_{\mathbf{u}}|^{-\frac{1}{2}}\int \mathrm{e}^{-\frac{1}{2\sigma^2}\left((\mathbf{y}-\boldsymbol{\tau})'(\mathbf{y}-\boldsymbol{\tau})+\boldsymbol{\tau}'\mathbf{H}'\mathbf{S}_{\mathbf{u}}^{-1}\mathbf{H}\boldsymbol{\tau}\right)}\mathrm{d}\boldsymbol{\tau}\\
&= (2\pi\sigma^2)^{-T}|\mathbf{S}_{\mathbf{u}}|^{-\frac{1}{2}}\int \mathrm{e}^{-\frac{1}{2\sigma^2}\left(\mathbf{y}'\mathbf{y}+\boldsymbol{\tau}'\boldsymbol{\tau}-2\boldsymbol{\tau}'\mathbf{y}+\boldsymbol{\tau}'\mathbf{H}'\mathbf{S}_{\mathbf{u}}^{-1}\mathbf{H}\boldsymbol{\tau}\right)}\mathrm{d}\boldsymbol{\tau}\\
&= (2\pi\sigma^2)^{-T}|\mathbf{S}_{\mathbf{u}}|^{-\frac{1}{2}}\int \mathrm{e}^{-\frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y}+\boldsymbol{\tau}'\mathbf{K}_{\boldsymbol{\tau}}\boldsymbol{\tau}-2\boldsymbol{\tau}'\mathbf{y})}\mathrm{d}\boldsymbol{\tau}\\
&= (2\pi\sigma^2)^{-T}|\mathbf{S}_{\mathbf{u}}|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2\sigma^2}\left(\mathbf{y}'\mathbf{y}-\mathbf{y}'\mathbf{K}_{\boldsymbol{\tau}}^{-1}\mathbf{y}\right)}\int \mathrm{e}^{-\frac{1}{2\sigma^2}(\boldsymbol{\tau}-\mathbf{K}_{\boldsymbol{\tau}}^{-1}\mathbf{y})'\mathbf{K}_{\boldsymbol{\tau}}(\boldsymbol{\tau}-\mathbf{K}_{\boldsymbol{\tau}}^{-1}\mathbf{y})}\mathrm{d}\boldsymbol{\tau}\\
&= (2\pi\sigma^2)^{-\frac{T}{2}}|\mathbf{S}_{\mathbf{u}}|^{-\frac{1}{2}}|\mathbf{K}_{\boldsymbol{\tau}}|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2\sigma^2}\left(\mathbf{y}'\mathbf{y}-\mathbf{y}'\mathbf{K}_{\boldsymbol{\tau}}^{-1}\mathbf{y}\right)},
\end{aligned}
$$

where $\mathbf{K}_{\boldsymbol{\tau}} = \mathbf{I}_T + \mathbf{H}'\mathbf{S}_{\mathbf{u}}^{-1}\mathbf{H}$. Next, we derive an analytical expression of the marginal likelihood. Recall that the prior for σ^2 is $\mathcal{IG}(\nu_0, S_0)$ with density

$$
p(\sigma^2) = \frac{{S_0}^{\nu_0}}{\Gamma(\nu_0)}(\sigma^2)^{-\nu_0-1}\mathrm{e}^{-\frac{S_0}{\sigma^2}}. \tag{11}
$$

Hence, the joint density $p(\mathbf{y},\boldsymbol{\tau},\sigma^2)$ is simply the product of the complete-data likelihood and the prior in (11):

$$
\begin{aligned}
p(\mathbf{y},\boldsymbol{\tau},\sigma^2) &= (2\pi)^{-T}\frac{{S_0}^{\nu_0}}{\Gamma(\nu_0)}|\mathbf{S}_{\mathbf{u}}|^{-\frac{1}{2}}\left(\frac{1}{\sigma^2}\right)^{T+\nu_0+1}\mathrm{e}^{-\frac{1}{\sigma^2}\left(S_0+\frac{(\mathbf{y}-\boldsymbol{\tau})'(\mathbf{y}-\boldsymbol{\tau})+\boldsymbol{\tau}'\mathbf{H}'\mathbf{S}_{\mathbf{u}}^{-1}\mathbf{H}\boldsymbol{\tau}}{2}\right)}\\
&= (2\pi)^{-T}\frac{{S_0}^{\nu_0}}{\Gamma(\nu_0)}|\mathbf{S}_{\mathbf{u}}|^{-\frac{1}{2}}\left(\frac{1}{\sigma^2}\right)^{T+\nu_0+1}\mathrm{e}^{-\frac{1}{\sigma^2}\left(S+\frac{(\boldsymbol{\tau}-\mathbf{K}_{\boldsymbol{\tau}}^{-1}\mathbf{y})'\mathbf{K}_{\boldsymbol{\tau}}(\boldsymbol{\tau}-\mathbf{K}_{\boldsymbol{\tau}}^{-1}\mathbf{y})}{2}\right)},
\end{aligned} \tag{12}
$$

where $S = S_0 + (\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{K}_{\boldsymbol{\tau}}^{-1}\mathbf{y})/2$ and $\mathbf{K}_{\boldsymbol{\tau}} = \mathbf{I}_T + \mathbf{H}'\mathbf{S}_{\mathbf{u}}^{-1}\mathbf{H}$.

It follows that the joint posterior distribution of $(\boldsymbol{\tau},\sigma^2)$ is a normal-inverse-gamma distribution $(\boldsymbol{\tau},\sigma^2\,|\,\mathbf{y}) \sim \mathcal{N}\text{-}\mathcal{IG}(\mathbf{K}_{\boldsymbol{\tau}}^{-1}\mathbf{y}, \mathbf{K}_{\boldsymbol{\tau}}, T/2+\nu_0, S)$ with density function

$$
p(\boldsymbol{\tau},\sigma^2\,|\,\mathbf{y}) \propto p(\mathbf{y},\boldsymbol{\tau},\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{T+\nu_0+1}\mathrm{e}^{-\frac{1}{\sigma^2}\left(S+\frac{(\boldsymbol{\tau}-\mathbf{K}_{\boldsymbol{\tau}}^{-1}\mathbf{y})'\mathbf{K}_{\boldsymbol{\tau}}(\boldsymbol{\tau}-\mathbf{K}_{\boldsymbol{\tau}}^{-1}\mathbf{y})}{2}\right)}.
$$

A draw from $p(\boldsymbol{\tau},\sigma^2\,|\,\mathbf{y})$ can be obtained in two steps. First, sample $(\sigma^2\,|\,\mathbf{y}) \sim \mathcal{IG}(T/2+\nu_0, S)$. Then, given the σ^2 drawn, sample $(\boldsymbol{\tau}\,|\,\mathbf{y},\sigma^2) \sim \mathcal{N}(\mathbf{K}_{\boldsymbol{\tau}}^{-1}\mathbf{y}, \sigma^2\mathbf{K}_{\boldsymbol{\tau}}^{-1})$. Since the precision matrix $\mathbf{K}_{\boldsymbol{\tau}}/\sigma^2$ is a band matrix, the precision sampler in Chan and Jeliazkov (2009) can be used to sample $\boldsymbol{\tau}$ efficiently.

Finally, the marginal likelihood can be obtained by integrating the expression in (12) with respect to $\boldsymbol{\tau}$ and σ^2. To that end, let $c_1 = (2\pi)^{-T}|\mathbf{S_u}|^{-\frac{1}{2}}{S_0}^{\nu_0}/\Gamma(\nu_0)$. Then,

$$\begin{aligned}
p(\mathbf{y}) &= \int p(\mathbf{y}, \boldsymbol{\tau}, \sigma^2)\mathrm{d}(\boldsymbol{\tau}, \sigma^2) \\
&= c_1 \int \left(\frac{1}{\sigma^2}\right)^{T+\nu_0+1} \mathrm{e}^{-\frac{1}{\sigma^2}\left(S + \frac{(\boldsymbol{\tau}-\mathbf{K}_{\boldsymbol{\tau}}^{-1}\mathbf{y})'\mathbf{K}_{\boldsymbol{\tau}}(\boldsymbol{\tau}-\mathbf{K}_{\boldsymbol{\tau}}^{-1}\mathbf{y})}{2}\right)}\mathrm{d}(\boldsymbol{\tau}, \sigma^2) \\
&= c_1(2\pi)^{\frac{T}{2}}|\mathbf{K}_{\boldsymbol{\tau}}|^{-\frac{1}{2}}\Gamma\left(\frac{T}{2} + \nu_0\right) S^{-\left(\frac{T}{2}+\nu_0\right)} \\
&= (2\pi)^{-\frac{T}{2}}|\mathbf{S_u}|^{-\frac{1}{2}}|\mathbf{K}_{\boldsymbol{\tau}}|^{-\frac{1}{2}} \frac{{S_0}^{\nu_0}\Gamma\left(\frac{T}{2}+\nu_0\right)}{\Gamma(\nu_0)} \left(S_0 + \frac{\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{K}_{\boldsymbol{\tau}}^{-1}\mathbf{y}}{2}\right)^{-\left(\frac{T}{2}+\nu_0\right)}.
\end{aligned}$$

References

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.

F. Bianchi. Regime switches, agents' beliefs, and post-World War II U.S. macroeconomic dynamics. *The Review of Economic Studies*, 80(2):463–490, 2013.

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

J. C. C. Chan and A. L. Grant. Issues in comparing stochastic volatility models using the deviance information criterion. *CAMA Working Paper*, 2014b.

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

S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

T. E. Clark and T. Doh. A Bayesian evaluation of alternative models of trend inflation. *International Journal of Forecasting*, 30(3):426–448, 2014.

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

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

J. Geweke. Using simulation methods for Bayesian econometric models: Inference, development, and communication. *Econometric Reviews*, 18(1):1–73, 1999.

M. Jochmann, G. Koop, and S. M. Potter. Modeling the dynamics of inflation compensation. *Journal of Empirical Finance*, 17(1):157–167, 2010.

A. Justiniano and G. E. Primiceri. The time-varying volatility of macroeconomic fluctuations. *American Economic Review*, 98(3):604–41, 2008.

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

G. Koop and S. Potter. A flexible approach to parametric inference in nonlinear and time varying time series models. *Journal of Econometrics*, 159(1):134–150, 2010.

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

M. Lanne, A. Luoma, and J. Luoto. Bayesian model selection and forecasting in noncausal autoregressive models. *Journal of Applied Econometrics*, 27(5):812–830, 2012.

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

Z. Liu, D. F. Waggoner, and T. Zha. Sources of macroeconomic fluctuations: A regime-switching DSGE approach. *Quantitative Economics*, 2(2):251–301, 2011.

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

J. H. Stock and M. W. Watson. Why has U.S. inflation become harder to forecast? *Journal of Money Credit and Banking*, 39(s1):3–33, 2007.