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 \( p(y \mid \theta_k, M_k) \) and a prior on the model-specific parameter vector \( \theta_k \) denoted by \( p(\theta_k \mid M_k) \). One popular Bayesian model comparison criterion is the Bayes factor in favor of \( M_i \) against \( M_j \), defined as

\[
BF_{ij} = \frac{p(y \mid M_i)}{p(y \mid M_j)},
\]

where \( p(y \mid M_k) = \int p(y \mid \theta_k, M_k)p(\theta_k \mid M_k)d\theta_k \) 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 \( y \). Hence, if the observed data are likely under the model, the associated marginal likelihood would be “large” and vice versa. It follows that \( BF_{ij} > 1 \) 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:

$$\frac{P(M_i \mid y)}{P(M_j \mid y)} = \frac{P(M_i)}{P(M_j)} \times BF_{ij},$$

where $P(M_i)/P(M_j)$ 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, $BF_{ij} = 20$, 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 $p(y)$—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:

$$\mathbb{E}\left(\frac{f(\theta)}{p(\theta)p(y \mid \theta)} \mid y\right) = \int \frac{f(\theta)}{p(\theta)p(y \mid \theta)} \frac{p(\theta)p(y \mid \theta)}{p(y)} d\theta = p(y)^{-1},$$

(1)

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

$$GD_0 = \left\{ \frac{1}{R} \sum_{i=1}^{R} \frac{f(\theta_i)}{p(\theta_i)p(y \mid \theta_i)} \right\}^{-1},$$

(2)

where $\theta_1, \ldots, \theta_R$ are posterior draws. Note that this estimator is simulation consistent in the sense that it converges to $p(y)$ in probability as $R$ tends to infinity, but it is not unbiased—i.e., $\mathbb{E}(GD_0) \neq p(y)$ 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 $\hat{\theta}$ and $Q_\theta$ denote the posterior mean and covariance matrix respectively. Then, $f$ is set to be the $N(\hat{\theta}, Q_\theta)$ density truncated within the region

$$\{ \theta \in \mathbb{R}^m : (\theta - \hat{\theta})' Q_\theta^{-1} (\theta - \hat{\theta}) < \chi^2_{\alpha,m} \},$$

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

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

$$p(y \mid \theta) = \int p(y, z \mid \theta)dz = \int p(y \mid z, \theta)p(z \mid \theta)dz,$$
where \( p(y, z | \theta) \) is the complete-data likelihood and \( p(y | z, \theta) \) denotes the conditional likelihood. To avoid ambiguity, \( p(y | \theta) \) 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
\[
p(y, z | \theta) = p(y | z, \theta)p(z | \theta)
\]
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 \( p(y | \theta) \):\[
GD_c = \left\{ \frac{1}{R} \sum_{i=1}^{R} \frac{f(\theta_i, z_i)}{p(y, z_i | \theta_i)p(\theta_i)} \right\}^{-1}, \tag{3}
\]
where \((\theta_1, z_1), \ldots, (\theta_R, z_R)\) are posterior draws from the augmented model \( p(\theta, z | y) \) and \( f \) is a tuning function.

However, the variance of \( GD_c \) is generally much larger than that of \( GD_o \). 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 \( GD_c \) 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 \( V_\tau \). 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 \( \tau = (\tau_1, \ldots, \tau_T)' \) and the only parameter is \( \sigma^2 \). In the next section we compute the marginal likelihood under different values of \( g \).
Stacking \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_T)' \) and \( y = (y_1, \ldots, y_T)' \), the measurement equation (4) can be written as:

\[
y = \tau + \varepsilon,
\]

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

\[
f(y \mid \tau, \sigma^2) = (2\pi\sigma^2)^{-\frac{T}{2}} e^{-\frac{1}{2\sigma^2} (y - \tau)'(y - \tau)}. \tag{6}
\]

In the Appendix we derive the joint density of the states \( p(\tau \mid \sigma^2) \) and show that the complete-data likelihood is

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

where \( S_u = \text{diag}(V_{\tau}, g, \ldots, g) \) and

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

Note that \( H \) 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:

\[
p(y \mid \sigma^2) = \int f(y \mid \tau, \sigma^2)p(\tau \mid \sigma^2)d\tau = (2\pi\sigma^2)^{-\frac{T}{2}} \left| S_u \right|^{-\frac{1}{2}} \left| K_{\tau} \right|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2} \left( y'y - y'K_{\tau}^{-1}y \right)},
\]

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

\[
p(y) = \int f(y, \tau \mid \sigma^2)p(\sigma^2)d(\tau, \sigma^2)
\]

\[
= (2\pi)^{-\frac{T}{2}} \left| S_u \right|^{-\frac{1}{2}} \left| K_{\tau} \right|^{-\frac{1}{2}} \frac{\Gamma(\nu + \frac{T}{2})}{\Gamma(\nu_0)} \left( \frac{S_0 + (y'y - y'K_{\tau}^{-1}y)}{2} \right)^{-\frac{\nu + \nu_0}{2}}. \tag{9}
\]

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 GD_o estimator based on the observed-data likelihood, the tuning function \( f \) is a single-variable function that depends only on \( \sigma^2 \). We use a univariate truncated normal density suggested by Geweke (1999) (see the discussion in Section 2) and denote it as \( f_{\sigma^2} \). For the GD_c estimator based on the complete-data likelihood, the tuning function \( f \) needs to be a \((T + 1)-\text{dimensional function—it depends on} \sigma^2 \) and \( \tau \). 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
\( f(\sigma^2, \tau) = f_{\sigma^2}(\sigma^2)p(\tau | \sigma^2) \), where \( f_{\sigma^2} \) is a univariate truncated normal density discussed
above, and \( p(\tau | \sigma^2) \) is the prior density of \( \tau \) given in (10). Using the prior density as part of the tuning function seems reasonable since it encapsulates our prior belief about \( \tau \). In addition, this choice simplifies the computation. The GD\(_c\) estimator in (3) now becomes
\[
GD_c = \left\{ \frac{1}{R} \sum_{i=1}^{R} f_{\sigma^2}(\sigma_i^2) \right\}^{-1} \frac{1}{R} \sum_{i=1}^{R} p(y | \sigma_i^2, \tau_i)p(\sigma_i^2),
\]
where \((\sigma_1^2, \tau_1), \ldots, (\sigma_R^2, \tau_R)\) 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 \((\tau, \sigma^2 | y)\) 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 GD\(_o\) and 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>
<thead>
<tr>
<th></th>
<th>(g = 0.8)</th>
<th>(g = 0.9)</th>
<th>(g = 1)</th>
<th>(g = 1.1)</th>
<th>(g = 1.2)</th>
<th>(g = 1.3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>analytical result</td>
<td>-592.141</td>
<td>-591.994</td>
<td>-591.936</td>
<td>-591.940</td>
<td>-591.990</td>
<td>-592.073</td>
</tr>
<tr>
<td>(GD_o (R = 50000))</td>
<td>-592.140</td>
<td>-591.995</td>
<td>-591.935</td>
<td>-591.941</td>
<td>-591.991</td>
<td>-592.070</td>
</tr>
<tr>
<td>(GD_c (R = 50000))</td>
<td>-503.28</td>
<td>-498.89</td>
<td>-494.62</td>
<td>-487.91</td>
<td>-485.08</td>
<td>-481.52</td>
</tr>
<tr>
<td>(GD_c (R = 10^7))</td>
<td>-510.64</td>
<td>-506.51</td>
<td>-502.70</td>
<td>-498.19</td>
<td>-493.53</td>
<td>-489.79</td>
</tr>
</tbody>
</table>

A few observations can be drawn from this exercise. Firstly, it is clear that the GD\(_o\)
estimates based on the observed-data likelihood are essentially identical to the analytical
results, whereas the GDₖ estimates based on the complete-data likelihood are substantially different. For example, when \( g = 1 \) and \( R = 50000 \), the GDₖ 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 \(-591.9)\). 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 GDₖ 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 GDₖ 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, GDₖ 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 (GDₒ) and the complete-data likelihood (GDₖ) for UK data. Numerical standard errors are in parenthesis.

<table>
<thead>
<tr>
<th>g</th>
<th>Analytical Result</th>
<th>GDₒ ((R = 50000))</th>
<th>GDₖ ((R = 50000))</th>
<th>GDₖ ((R = 10^7))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>-673.362</td>
<td>-673.361</td>
<td>-653.74</td>
<td>-656.28</td>
</tr>
<tr>
<td>0.1</td>
<td>-672.055</td>
<td>-672.055</td>
<td>-647.37</td>
<td>-650.60</td>
</tr>
<tr>
<td>0.15</td>
<td>-672.515</td>
<td>-672.514</td>
<td>-644.00</td>
<td>-647.44</td>
</tr>
<tr>
<td>0.2</td>
<td>-673.432</td>
<td>-673.432</td>
<td>-640.35</td>
<td>-645.00</td>
</tr>
<tr>
<td>0.25</td>
<td>-674.489</td>
<td>-674.490</td>
<td>-638.75</td>
<td>-642.84</td>
</tr>
<tr>
<td>0.3</td>
<td>-675.575</td>
<td>-675.575</td>
<td>-636.24</td>
<td>-640.32</td>
</tr>
</tbody>
</table>

\(^1\)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 GD 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 $p(\tau \mid \sigma^2)$. To that end, we stack the state equation (5) over $t$:

$$ H\tau = u, $$

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

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

$$ p(\tau \mid \sigma^2) = (2\pi \sigma^2)^{-T/2} |S_u|^{-1/2} e^{-\frac{1}{2\sigma^2} \tau'H'S_u^{-1}H\tau}. $$

(10)

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

$$ p(y, \tau \mid \sigma^2) = p(y \mid \tau, \sigma^2)p(\tau \mid \sigma^2) = (2\pi \sigma^2)^{-T} |S_u|^{-1/2} e^{-\frac{1}{2\sigma^2} (y-\tau)'(y-\tau) + \tau'H'S_u^{-1}H\tau}. $$
Moreover, after some computations, the observed-data likelihood is obtained as follows:

\[
p(y | \sigma^2) = \int p(y, \tau | \sigma^2) d\tau
\]

\[
= (2\pi \sigma^2)^{-T} |S_u|^{-\frac{1}{2}} \int e^{-\frac{1}{2\sigma^2}((y-\tau)'(y-\tau) + \tau'H'S_u^{-1}H\tau)} d\tau
\]

\[
= (2\pi \sigma^2)^{-T} |S_u|^{-\frac{1}{2}} \int e^{-\frac{1}{2\sigma^2}(y'y + \tau'\tau - 2\tau'y + \tau'H'S_u^{-1}H\tau)} d\tau
\]

\[
= (2\pi \sigma^2)^{-T} |S_u|^{-\frac{1}{2}} \int e^{-\frac{1}{2\sigma^2}(y'y + \tau'K_\tau \tau - 2\tau'y)} d\tau
\]

\[
= (2\pi \sigma^2)^{-T} |S_u|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(y'y - y'K_\tau^{-1}y)} \int e^{-\frac{1}{2\sigma^2}(\tau' - K_\tau^{-1}y)'K_\tau(\tau - K_\tau^{-1}y)} d\tau
\]

\[
= (2\pi \sigma^2)^{-\frac{T}{2}} |S_u|^{-\frac{1}{2}} |K_\tau|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(y'y - y'K_\tau^{-1}y)},
\]

where \( K_\tau = I_T + H'S_u^{-1}H \). Next, we derive an analytical expression of the marginal likelihood. Recall that the prior for \( \sigma^2 \) is \( IG(\nu_0, S_0) \) with density

\[
p(\sigma^2) = \frac{S_0^{\nu_0}}{\Gamma(\nu_0)} (\sigma^2)^{-\nu_0-1} e^{-\frac{S_0}{2\sigma^2}}.
\]

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

\[
p(y, \tau, \sigma^2) = (2\pi)^{-T} \frac{S_0^{\nu_0}}{\Gamma(\nu_0)} |S_u|^{-\frac{1}{2}} \left( \frac{1}{\sigma^2} \right)^{T+\nu_0+1} e^{-\frac{1}{2\sigma^2} \left( S_0 + (y'y)' \right) / 2} e^{-\frac{1}{\sigma^2} \left( \frac{(y'y)'}{2} + \tau'H'S_u^{-1}H\tau \right)}
\]

\[
= (2\pi)^{-T} \frac{S_0^{\nu_0}}{\Gamma(\nu_0)} |S_u|^{-\frac{1}{2}} \left( \frac{1}{\sigma^2} \right)^{T+\nu_0+1} e^{-\frac{1}{\sigma^2} \left( \frac{(y'y - y'K_\tau^{-1}y)}{2} + \tau'H'S_u^{-1}H\tau \right)}
\]

where \( S = S_0 + (y'y - y'K_\tau^{-1}y)/2 \) and \( K_\tau = I_T + H'S_u^{-1}H \).

It follows that the joint posterior distribution of \( (\tau, \sigma^2) \) is a normal-inverse-gamma distribution \( (\tau, \sigma^2 | y) \sim \mathcal{N}\mathcal{I}G(K_\tau^{-1}y, K_\tau, T/2 + \nu_0, S) \) with density function

\[
p(\tau, \sigma^2 | y) \propto p(y, \tau, \sigma^2) \propto \left( \frac{1}{\sigma^2} \right)^{T+\nu_0+1} e^{-\frac{1}{\sigma^2} \left( \frac{(y'y - y'K_\tau^{-1}y)}{2} + \tau'H'S_u^{-1}H\tau \right)}.
\]

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

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

\[
p(y) = \int p(y, \tau, \sigma^2) d(\tau, \sigma^2) \\
= c_1\int \left( \frac{1}{\sigma^2} \right)^{T+\nu_0+1} e^{-\frac{1}{2\sigma^2}\left( S_0 + \frac{(\tau - \mathbf{K}^{-1}\mathbf{y})'\mathbf{K}^{-1}(\tau - \mathbf{K}^{-1}\mathbf{y})}{2} \right)} d(\tau, \sigma^2) \\
= c_1(2\pi)^{-\frac{T}{2}}|\mathbf{K}_\tau|^{-\frac{1}{2}}\Gamma\left( \frac{T}{2} + \nu_0 \right) S_0^{\nu_0} \Gamma\left( \frac{T}{2} + \nu_0 \right) \\
= (2\pi)^{-\frac{T}{2}}|\mathbf{S}_u|^{-\frac{1}{2}}|\mathbf{K}_\tau|^{-\frac{1}{2}}\frac{S_0^{\nu_0} \Gamma\left( \frac{T}{2} + \nu_0 \right)}{\Gamma(\nu_0)} \left( S_0 + \frac{\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{K}^{-1}\mathbf{y}}{2} \right)^{-(\frac{T}{2} + \nu_0)}.
\]

References


