Economics Letters 131 (2015) 29–33

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Pitfalls of estimating the marginal likelihood using the modified harmonic mean

Joshua C.C. Chan [a,*], Angelia L. Grant [b]

[a] Research School of Economics, Australian National University, Australia

[b] Centre for Applied Macroeconomic Analysis, Australian National University, Australia

HIGHLIGHTS

- We investigate the performance of two versions of the modified harmonic mean.
- An unobserved components model is fitted using US and UK inflation data.
- The one based on the complete-data likelihood has a substantial finite sample bias.
- The version based on the observed-data likelihood works well.

ARTICLE INFO

Article history:
Received 28 January 2015
Received in revised form
16 March 2015
Accepted 26 March 2015
Available online 2 April 2015

JEL classification:
C11
C15
C32
C52

Keywords:
Bayesian model comparison
State space
Unobserved components
Inflation

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 bias and tends to select the wrong model, whereas the version based on the observed-data likelihood works well.

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 et al., 2011; Lanne et al., 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 et al. (2004), Justiniano and Primiceri (2008) and Jochmann et al. (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 et al. (2012) and Chan and Grant (forthcoming), who argue against the use of the complete-data likelihood in a related context of computing the deviance information criterion.

* Correspondence to: Research School of Economics, ANU College of Business and Economics, LF Crisp Building 26, The Australian National University, Canberra ACT 2601, Australia. Tel.: +61 2 612 57358; fax: +61 2 612 50182.
E-mail address: joshua.chan@anu.edu.au (J.C.C. Chan).

http://dx.doi.org/10.1016/j.econlet.2015.03.036
0165-1765/

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. Section 4 concludes.

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:

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

where $\mathbb{P}(M_i)/\mathbb{P}(M_j)$ is the prior odds ratio. 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:

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

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

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

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

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

where $p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$ is the *complete-data likelihood* and $p(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta})$ denotes the *conditional likelihood*. To avoid ambiguity, $p(\mathbf{y} \mid \boldsymbol{\theta})$ 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 $p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) = p(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta})p(\mathbf{z} \mid \boldsymbol{\theta})$ 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 $p(\mathbf{y} \mid \boldsymbol{\theta})$:

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

where $(\boldsymbol{\theta}_1, \mathbf{z}_1), \ldots, (\boldsymbol{\theta}_R, \mathbf{z}_R)$ are posterior draws from the augmented model $p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{y})$ 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:

$$y_t = \tau_t + \varepsilon_t, \quad (4)$$

$$\tau_t = \tau_{t-1} + u_t, \quad (5)$$

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