Central European Journal of Economic Modelling and Econometrics

A Note on Lenk's Correction of the Harmonic Mean Estimator

Anna Pajor[*], Jacek Osiewalski[†]

Submitted: 15.12.2013, Accepted: 30.01.2014

Abstract

The paper refines Lenk's concept of improving the performance of the computed harmonic mean estimator (HME) in three directions. First, the adjusted HME is derived from an exact analytical identity. Second, Lenk's assumption concerning the appropriate subset A of the parameter space is significantly weakened. Third, it is shown that, under certain restrictions imposed on A, a fundamental identity underlying the HME also holds for improper prior densities, which substantially extends applicability of the adjusted HME.

Keywords: Bayesian inference, marginal data density, MCMC methods

JEL Classification: C11, C15.

[*]Cracow University of Economics (Department of Econometrics and Operations Research); Kraków, Poland. E-mail: `eopajor@cyf-kr.edu.pl`

[†]Cracow University of Economics (Department of Econometrics and Operations Research); Kraków, Poland. E-mail: `eeosiewa@cyf-kr.edu.pl`

1 The HME and Lenk's correction

The marginal data density values (also called the marginal likelihoods of the data, or the normalizing constants, or the integrated likelihoods) are essential in formal Bayesian model selection and model averaging.[1] Let us consider a model, in which a) the parameter space[2] is denoted by Θ; b) $p(\theta)$ is a prior density function of parameters collected in $\theta \in \Theta$; c) y is a vector of observations. The marginal data density, $p(y)$, is defined as an integral (calculated over the whole parameter space) of the conditional data density given the vector of parameters, $p(y|\theta)$, with respect to the prior distribution:

$$p(y) = \int_{\Theta} p(y|\theta)p(\theta)\, d\theta. \tag{1}$$

For the majority of models a closed-form expression for the marginal data density is not available, and a certain Monte Carlo approximation of $p(y)$ for the observed vector y is needed. For the case of MCMC simulations from the posterior distribution, Newton and Raftery (1994) proposed the simple (and hence popular) harmonic mean estimator (HME) of the marginal data density value, given by:

$$\hat{p}_{HME}(y) = \left[\frac{1}{k} \sum_{q=1}^{k} \frac{1}{p(y|\theta_{(q)})} \right]^{-1}, \tag{2}$$

where the elements of the series $\{\theta_{(q)}\}_{q=1}^{k}$, are drawn from the posterior by means of some MCMC algorithm. The form of HME is based on a simple identity:

$$p(y) = \left[\int_{\Theta} \frac{1}{p(y|\theta)} p(\theta|y)\, d\theta \right]^{-1}, \tag{3}$$

easily derived from a fundamental relation:

$$1 = \int_{\Theta} p(\theta)\, d\theta, \tag{4}$$

holding for all densities of proper (i.e. probabilistic) prior measures.
Even though the HME is consistent (see Newton and Raftery 1994), it has serious shortcomings. Namely, it lacks finite asymptotic variance and, in small samples (i.e., for finite MCMC chains), it overestimates the marginal data density. As pointed

[1]Bayesian model comparison and inference pooling - described by, e.g., Osiewalski and Steel (1993) - are rooted in testing statistical hypotheses through their posterior probabilities; see, e.g., Zellner (1971).

[2]Following the Bayesian approach, we assume that the vector of parameters includes also latent variables (if there are any in the model).

by Lenk (2009), the latter results from the "simulation pseudo-bias" of the HME, which calls for an appropriate "pseudo-bias" correction. Apart from identifying the source of the "simulation pseudo-bias", Lenk (2009) also delivers several methods of estimating the "pseudo-bias" adjustment factor. The adjusted HME for the marginal data density, proposed by Lenk (2009), is given by the formula:

$$\hat{p}_{AHME}(y) = \hat{P}(A)\left[\frac{1}{k}\sum_{q=1}^{k}\frac{1}{p(y|\theta_{(q)})}\right]^{-1} = \hat{P}(A)\hat{p}_{HME}(y), \tag{5}$$

where $\hat{P}(A)$ is an estimate of the prior probability of $A \subseteq \Theta$ (i.e. $P(A)$).
In formula (5) the elements of the series $\{\theta_{(q)}\}_{q=1}^{k}$ are the MCMC draws from the posterior distribution restricted to the subset A. It is assumed by Lenk (2009) that the posterior probability of A is greater than $1-\varepsilon$ for some small $\varepsilon > 0$ (i.e., $P(A|y) > 1-\varepsilon$), and it is presented that, due to this assumption, the adjusted HME can be used for estimating $p(y)$.
In the next section we substantially generalize the basic identity (3) that underlies the original HME. We thus firmly justify the construction of the adjusted HME and extend its applicability to different subsets A and even to improper priors.

2 A formal justification and generalization of Lenk's correction

In this section it will be shown that the adjusted HME results from the following exact identity:

$$p(y) = P(A)\left[\int_A \frac{1}{p(y|\theta)}p(\theta|y)\,d\theta\right]^{-1}, \tag{6}$$

holding for any subset $A \subseteq \Theta$ of a non-zero prior probability mass. Thus, Lenk's adjusted HME can be used not only for the subset A, for which the posterior probability is close to 1, but also for any (reasonable) subset of the parameter space.
Let us assume that $p(y) > 0$ and A is any subset of the parameter space such that $P(A) > 0$ and $p(\theta|y) > 0$ almost everywhere in A. Then we have:

$$P(A) = \int_A p(\theta)\,d\theta = \int_A \frac{p(\theta)}{p(\theta|y)}p(\theta|y)\,d\theta.$$

From Bayes' rule:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)},$$

thus

$$P(A) = \int_A \frac{p(\theta)p(y)}{p(y|\theta)p(\theta)} p(\theta|y)\, d\theta.$$

Consequently,

$$P(A) = p(y) \int_A \frac{1}{p(y|\theta)} p(\theta|y)\, d\theta,$$

which immediately leads to (6).

The fundamental identity (6) suggests that the marginal data density value, $p(y)$, can indeed be always approximated by (5), i.e. the product of 1) the harmonic mean of the likelihood values, calculated using draws $\{\theta_{(q)}\}_{q=1}^{k}$ from the posterior distribution of θ restricted to the subset A, and 2) Lenk's correction (an estimate of $P(A)$). As shown above, the subset A can be chosen arbitrarily.

Note that the identity underlying the original HME estimator heavily relies on the properness of the prior distribution. However, the derivation presented in this section for the adjusted HME does not require the prior measure to be a probabilistic one. The crucial assumption is that $P(A)$ is finite.

Let $p(\theta)$ be a density function of some σ-finite measure such that $p(y)$, defined by (1), is non-zero and finite (possibly except for measure-zero sets in the sample space), i.e. the posterior distribution exists. Let $A \subseteq \Theta$ be a set of non-zero and finite prior measure. Then our derivation remains valid and thus (6) still holds. For example, if A is a bounded subset of the parameter space R^d, then $P(A)$ is finite irrespective of whether the prior is improper uniform or proper, and the adjusted HME can be applied. An "advantage" of restricting the analysis to the proper priors only would reside in the possibility to use the original HME, which is not a reliable estimator. Of course, since we estimate marginal data density values in competing models in order to compute their Bayes factors and conduct formal model comparison, we can use improper priors only for the common parameters; see, e.g. Osiewalski and Steel (1993).

3 Concluding remarks

In this note it has been shown that Lenk's correction can be used regardless of the posterior probability accumulated in the chosen subset of the parameter space. This analytical result makes it possible to select A (the subset of the parameter space), used in estimating the marginal data density with MCMC methods, and consequently to improve numerical properties of the adjusted harmonic mean estimator. In fact, each A defines a new estimator of the marginal data density value. The properties of such estimator may heavily depend on the choice of A, and examining them should be a subject of future research. Finally, it is important to note that some estimators from the adjusted HME class can be used in the case of an improper prior.

Acknowledgements

We would like to thank Łukasz Kwiatkowski and Błażej Mazur for drawing our attention to the problems and methods described in Lenk's paper. Useful comments by a referee are gratefully acknowledged.

References

[1] Lenk P., (2009), Simulation pseudo-bias correction to the harmonic mean estimator of integrated likelihoods, Journal of Computational and Graphical Statistics, 18, 941-960.

[2] Newton M.A., Raftery A.E., (1994), Approximate Bayesian inference by the weighted likelihood bootstrap [with discussion], Journal of the Royal Statistical Society B, 56, 3-48.

[3] Osiewalski J., Steel M.F.J., (1993), A Bayesian perspective on model selection, manuscript, published in Spanish as: Una perspectiva bayesiana en selección de modelos, Cuadernos Economicos 55/3, 327-351.

[4] Zellner A., (1971), An Introduction to Bayesian Inference in Econometrics, J. Wiley, New York.