Bayesian Analysis (2017) **12**, Number 1, pp. 261–287

Estimating the Marginal Likelihood Using the Arithmetic Mean Identity

Anna Pajor*

Abstract. In this paper we propose a conceptually straightforward method to estimate the marginal data density value (also called the marginal likelihood). We show that the marginal likelihood is equal to the prior mean of the conditional density of the data given the vector of parameters restricted to a certain subset of the parameter space, A, times the reciprocal of the posterior probability of the subset A. This identity motivates one to use Arithmetic Mean estimator based on simulation from the prior distribution restricted to any (but reasonable) subset of the space of parameters. By trimming this space, regions of relatively low likelihood are removed, and thereby the efficiency of the Arithmetic Mean estimator is improved. We show that the adjusted Arithmetic Mean estimator is unbiased and consistent.

Keywords: Bayesian inference, Bayesian model selection, marginal likelihood.

1 Introduction

The marginal data densities (i.e. the normalizing constants of the posterior distributions of the model parameters, also called the marginal likelihoods or the integrated likelihoods) are key quantities needed for formal Bayesian model selection and for model averaging; see, e.g. Zellner (1971). The posterior odds ratio, used for comparing two competing models, is equal to the product of the prior odds and the Bayes' factor. In turn, the Bayes' factor is defined to be the ratio of marginal likelihoods. The marginal likelihoods and the prior model probabilities are used to form the posterior model probabilities, which are necessary for Bayesian model averaging and for testing statistical hypotheses. Therefore the marginal likelihoods are essential in the Bayesian approach.

Let us consider a model in which (i) the space of parameters is denoted by Θ, (ii) $p(\theta)$ is a prior density function of the parameters[1] collected in $\theta \in \Theta$, and (iii) y is a vector of observations. The marginal data density, $p(y)$, is defined as an integral (calculated over the whole parameters' space) of the conditional data density given the vector of parameters, $p(y|\theta)$, with respect to the prior distribution:

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

Even in simple models, correct assessment of the marginal likelihood is computationally challenging. In more complicated models, high-dimensional integration is required to

*Department of Econometrics and Operations Research, Cracow University of Economics, ul. Rakowicka 27, 31-510 Kraków, Poland, pajora@uek.krakow.pl

[1]The vector of parameters can include also latent variables (if there are any in the model).

 DOI: 10.1214/16-BA1001

estimate marginal data densities. In majority of models, it is not possible to analytically integrate out parameters from the joint distribution for y and θ, $p(y, \theta)$, and a certain Monte Carlo approximation of $p(y)$ for the observed vector y is needed.

In this paper we propose a conceptually straightforward method to estimate the marginal data density value. Our proposition is motivated by the problems with the Harmonic Mean estimator and their solutions proposed by Lenk (2009). By analogy to the adjusted Harmonic Mean estimator, we propose corrected Arithmetic Mean estimator. Therefore, we start with reviewing ways of estimating the marginal data density by means of the Harmonic Mean (proposed by Newton and Raftery (1994)). Estimation of the marginal data density by Harmonic Mean has become one of the most popular methods due to its simplicity. Under the assumption that the prior distribution of the vector of parameters is proper, we have

$$1 = \int_{\Theta} p(\theta)\, d\theta = p(y) \int_{\Theta} \frac{1}{p(y|\theta)} p(\theta|y)\, d\theta. \tag{2}$$

Consequently,

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

where $E_{\theta|y}(\cdot)$ denotes the expected value with respect to the posterior distribution of θ. In (3), the reciprocal of $p(y)$ is expressed as an expected value of the inverse of the conditional density of the data y given θ with respect to the posterior distribution of the parameters. In other words, the marginal data density is equal to the posterior Harmonic Mean of the conditional density of the data. This equation suggests using the sample Harmonic Mean of the conditional density of y given θ based on draws from the posterior distribution, $p(\theta|y)$. The Harmonic Mean (HM) estimator, proposed by Newton and Raftery (1994), is given by

$$\hat{p}_{HME}(y) = \left[\frac{1}{k}\sum_{q=1}^{k} \frac{1}{p(y|\theta_{(q)})}\right]^{-1}, \tag{4}$$

where $\{\theta_{(q)}\}_{q=1}^{k}$ is drawn from the posterior distribution of the parameters by means of a Markov Chain Monte Carlo (MCMC) method.

Even though the HM estimator is consistent (see Newton and Raftery (1994)), it has some serious shortcomings. Namely, it can be unstable (see Raftery et al. (2007)), and it overestimates the marginal data density. Moreover, as pointed by Lenk (2009), it is characterised by so-called "simulation pseudo-bias". Lenk (2009) has proposed several methods for correcting the "pseudo-bias" provided that we can draw from the posterior distribution restricted to a subset of the space of parameters, of which posterior probability is close to one. The adjusted HM estimator, proposed by Lenk (2009), is given by

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

where $\hat{P}(A)$ is an assessment of the prior probability of subset $A \subseteq \Theta$, of which the posterior probability is greater than $1-\varepsilon$ for small $\varepsilon > 0$ (i.e. $P(A|y) > 1-\varepsilon$), see Lenk (2009). Based on the identity[2]

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

which is true for any subset $A \subseteq \Theta$ such that $0 < P(A) < \infty$ and $0 < P(A|y) < +\infty$, Pajor and Osiewalski (2013) have shown that Lenk's correction can be used regardless of the posterior probability accumulated in the chosen subset of the parameters' space, and that some estimators from the adjusted HM estimator class (depending on the choice of A) can be used in the case of an improper prior. Their analytical result makes it possible to select A, and consequently to improve numerical properties of the adjusted Harmonic Mean estimator. It is easy to show that the identity (6) is equivalent to

$$p(y) = \frac{P(A)}{P(A|y)} \left[\int_{\Theta} \frac{1}{p(y|\theta)} p(\theta|y, A)\, d\theta \right]^{-1}. \tag{7}$$

Given the subset $A \subseteq \Theta$, the identities (6) and (7) naturally lead to the following estimators of the marginal data density value (further called as Corrected Harmonic Mean):

$$\hat{p}_{CHME}(y) = \hat{P}(A) \left[\frac{1}{k} \sum_{q=1}^{k} \frac{I_A(\theta_{(q)})}{p(y|\theta_{(q)})} \right]^{-1} \tag{8}$$

and

$$\hat{p}_{A,CHME}(y) = \frac{\hat{P}(A)}{\hat{P}(A|y)} \left[\frac{1}{k} \sum_{q=1}^{k} \frac{1}{p(y|\theta^A_{(q)})} \right]^{-1}, \tag{9}$$

where $\{\theta_{(q)}\}_{q=1}^{k}$ and $\{\theta^A_{(q)}\}_{q=1}^{k}$ are drawn from $p(\theta|y)$ and $p(\theta|y, A)$, respectively.

The Corrected Harmonic Mean estimator (given by (8)) is still biased, and it overestimates the marginal data density. Using the method presented in Xie et al. (2011), we have showed that even if $P(A)$ is known, the expected value of the Corrected Harmonic Mean times $P(A)$ is greater than $p(y)$, i.e.

$$p(y) < E_{\theta|y} \left(P(A) \left[\frac{1}{k} \sum_{q=1}^{k} \frac{I_A(\theta_{(q)})}{p(y|\theta_{(q)})} \right]^{-1} \right) \tag{10}$$

(see Supplementary Appendix A in Pajor (2016)). The bias results from the presence of the reciprocal function in (8), which is convex for positive arguments. In order to overcome the problem of biasedness of the estimator, we propose using the Arithmetic

[2] $I_A(\theta)$ is the indicator function which takes on the value 1 when its argument belongs to the set A and is 0 otherwise.

Mean instead of the Harmonic Mean. In turn, to reduce an unacceptably high variance of the Arithmetic Mean estimator, we trim the prior sample to eliminate problematic regions of the parameter space. Our method belongs to the class of methods which are based on importance sampling and used to estimate $p(y)$ separately for each model. Estimation of marginal likelihoods instead of the Bayes factor allows a new model to be directly compared with other models for which the marginal likelihood has been calculated.

Obviously, various other numerical methods have been proposed in the literature to estimate the marginal data density value directly or to compute ratios of two marginal likelihoods (i.e. Bayes factors); for a review see, e.g. Ardia et al. (2012), Friel and Wyse (2012). Very popular methods for computing Bayes factors or marginal posterior probabilities are based on the reversible jump algorithm of Green (1995) and on bridge sampling methods (see Meng and Wong (1996)). Bartolucci et al. (2006) proposed a class of estimators of the Bayes factor based on an extension of the bridge sampling identity of Meng and Wong (1996) and combined with reversible jump sampler.

A widely used method of approximating the marginal likelihood is Laplace–Metropolis approximation (see Kass and Raftery (1995), Raftery (1996)). The Laplace–Metropolis method is based on a numerical approximation of the marginal likelihood, obtained by substituting the posterior density by the Normal density with mean equal to the posterior mode and covariance matrix equal to the posterior variance matrix, obtained from posterior simulation output. Unfortunately, the approximation may not be valid for some models, e.g. mixture models, where asymptotic normality does not hold (see, e.g. Frühwirth-Schnatter (2006)).

Among the numerical methods used to estimate the marginal likelihood, a very popular one is the method of Chib (1995). The approach of Chib (1995) is aimed at estimation of the marginal likelihood from Gibbs sampling output (with the use of Rao–Blackwellization). The idea was extended by Chib and Jeliazkov (2001) to deal with cases where the Metropolis–Hastings algorithm is used to generate posterior samples. The methods require an additional sample using reduced Gibbs or reduced Metropolis–Hastings sampling, which may be time-consuming when the parameter space is split into a large number of blocks. Moreover, the numerical accuracy of the estimate depends on the choice of a specific point of the parameter space.

Friel and Pettitt (2008) proposed the power posterior method, inspired by ideas from the path sampling, which had been proposed by Gelman and Meng (1998), and based on samples from a distribution proportional to the likelihood raised to a power t times the prior density function. A difficulty with this method is that of choosing the optimal temperature schedule. Recently, Weinberg (2012) proposed a novel approach based on Lebesgue integration. All estimation methods present advantages and disadvantages. Unfortunately, no perfect solution exists.

In the following section, we propose a new class of estimators of the marginal likelihood. They are based on the Arithmetic Mean of likelihoods calculated only over an arbitrary subset of the space of model parameters corrected by the reciprocal of the posterior probability of the subset. We also show that, under some assumptions, new

estimators (which depend on the choice of a subset of the parameter space) are unbiased and consistent (see Supplementary Appendix B in Pajor (2016)). In Sections 3 and 4, we present simulation results and data examples showing that the new estimator performs very well in comparison with the Corrected Harmonic Mean, Chib and Laplace–Metropolis estimators. Finally, we conclude that preliminary experience with estimators introduced here is very promising.

2 Corrected Arithmetic Mean estimators: methodology

Equation (1) suggests using the sample Arithmetic Mean of the conditional density of y given θ based on draws from the prior distribution, $p(\theta)$. The Arithmetic Mean (AM) estimator is given by

$$\hat{p}_{AME}(y) = \frac{1}{k}\sum_{q=1}^{k} p(y|\theta_{(q)}), \tag{11}$$

where $\{\theta_{(q)}\}_{q=1}^{k}$ is now drawn from the prior distribution of the parameters.

This estimator was mentioned by, e.g. Hammersley and Handscomb (1964) and Raftery and Banfield (1991), and it was used by McCulloch and Rossi (1992) for logistic regression models as well as by Lewis and Raftery (1997) for comparing alternative hierarchical (i.e. random-effects) models. Although this AM estimator is unbiased, it can have a very high variance (and thus it can be quite inefficient). If the posterior distribution is much more concentrated than that of the prior, then while sampling from the prior distribution, we obtain most points in the area, in which the sampling density of the data (or the likelihood) is close to zero. Hence, the estimate of $p(y)$ depends on only few points from the area of high value of $p(y|\theta)$. Consequently, an incredibly large simulation sample would be required to obtain adequate result (see Lewis and Raftery (1997), Raftery (1996)). Moreover, the AM estimator cannot be used for improper priors. A natural remedy for the inefficiency is to trim the prior sample to eliminate regions of the parameter space with very low likelihood, similarly as for the Corrected Harmonic Mean estimator. This motivates us to propose a modification of the Arithmetic Mean estimator.

Let us assume that $A \subseteq \Theta$, $0 < P(A) < +\infty$, and $0 < P(A|y) < +\infty$. Starting from the identity

$$P(A|y) = \int_A p(\theta|y)\, d\theta, \tag{12}$$

we obtain

$$P(A|y) = \int_\Theta I_A(\theta)\frac{p(y|\theta)p(\theta)}{p(y)}\, d\theta, \tag{13}$$

and consequently,

$$p(y) = \frac{1}{P(A|y)}\int_\Theta p(y|\theta)I_A(\theta)p(\theta)\, d\theta = \frac{E_\theta(p(y|\theta)I_A(\theta))}{P(A|y)} \tag{14}$$

or equivalently,

$$p(y) = \frac{P(A)}{P(A|y)} \int_{\Theta} p(y|\theta)p(\theta|A)\, d\theta = \frac{P(A)}{P(A|y)} E_{\theta}(p(y|\theta)|A), \tag{15}$$

where $E_{\theta}(\cdot)$ denotes the expected value with respect to the prior distribution of θ, and $E_{\theta}(\cdot|A)$ denotes the conditional expected value of θ given A.

Equation (14) says that the marginal density of the data can be expressed as a product of the reciprocal of the posterior probability of the subset A, $P(A|y)$, and the expected value of the conditional density of the data times the indicator function of subset A, $p(y|\theta)I_A(\theta)$. This expected value is calculated with respect to the prior distribution of the model parameters. Identity (14) naturally leads to the following estimator of the marginal data density value (further called Corrected Arithmetic Mean estimator):

$$\hat{p}_{CAME}(y) = \frac{1}{\hat{P}(A|y)} \frac{1}{k} \sum_{q=1}^{k} p(y|\theta_{(q)}) I_A(\theta_{(q)}), \tag{16}$$

where $\{\theta_{(q)}\}_{q=1}^{k}$ is drawn from the prior distribution, $p(\theta)$. The assessment of the posterior probability of the subset A, $\hat{P}(A|y)$, requires also sampling from the posterior distribution.

In turn, identity (15) suggests that the marginal likelihood can be approximated by the product of the ratio of prior to posterior probabilities of the subset A and the sample Arithmetic Mean of the conditional data density,

$$\hat{p}_{A,CAME}(y) = \frac{\hat{P}(A)}{\hat{P}(A|y)} \frac{1}{k} \sum_{q=1}^{k} p(y|\theta_{(q)}^{A}), \tag{17}$$

based on $\{\theta_{(q)}^{A}\}_{q=1}^{k}$ drawn from the prior distribution, restricted to the subset A, $p(\theta|A)$. Note that we have just defined a new class of estimators indexed by the subset $A \subseteq \Theta$. Under additional assumptions (first, the subset A is compact; second, the likelihood is bounded on A; third, $P(A|y)$ is known; fourth, it is possible to generate samples from the prior distribution), it is easy to show that the Corrected Arithmetic Mean (CAM) estimator is unbiased and consistent (see Theorem 2 in Supplementary Appendix B in Pajor (2016) for details). It stems from the fact that the CAM estimator is just a simple Monte Carlo estimator. In practice, $P(A|y)$ needs to be estimated via posterior simulation. If it is possible to generate samples forming an ergodic Markov chain whose equilibrium distribution is the posterior distribution, then the CAM estimator remains consistent (see Theorem 3 in Supplementary Appendix B in Pajor (2016)). Because of the opportunity to arbitrarily select the subset A in (16) and (17), the problem with inefficiency of the Arithmetic Mean (11) can be overcome. Indeed, subset A should be chosen in the area of large values of $p(y|\theta)$. By doing so, the variance of the Arithmetic Mean can be reduced. Note that, from the numerical point of view, an optimal choice of A is to have $\hat{P}(A|y) = 1$ because then $\hat{p}_{CAME}(y) = \frac{1}{k} \sum_{q=1}^{k} p(y|\theta_{(q)}) I_A(\theta_{(q)})$, and the

CAM estimator needs additional simulation only from the prior distribution. Moreover, to evaluate the expected value in (14), importance sampling can be used. Suppose that there exists an algorithm for generating sample according to the probability density function $s(\cdot)$ referred as an importance function. Then

$$E_\theta(p(y|\theta)I_A(\theta)) = \int\limits_\Theta p(y|\theta)I_A(\theta)p(\theta)d\theta = \int\limits_\Theta \frac{p(y|\theta)I_A(\theta)p(\theta)}{s(\theta)}s(\theta)d\theta. \tag{18}$$

This yields the importance sampling estimator of $p(y)$ with importance function $s(\cdot)$, namely,

$$\hat{p}_{CAME}(y) = \frac{1}{\hat{P}(A|y)}\frac{1}{k}\sum_{q=1}^{k}\frac{p(y|\theta_{(q)})p(\theta_{(q)})I_A(\theta_{(q)})}{s(\theta_{(q)})}, \tag{19}$$

where $\{\theta_{(q)}\}_{q=1}^{k}$ is drawn from the importance sampling distribution.

The simplicity of the Corrected Arithmetic Mean estimator and its good properties are its main advantages over other techniques based on importance sampling, e.g. stabilized version of the Harmonic Mean (see Raftery et al. (2007)), annealed importance sampling (see Neal (2001)), important-weighted marginal density estimators (see Chen (1994), Chen (2005)), and others (see, e.g. Meng and Wong (1996), Raftery (1996), Lewis and Raftery (1997), Han and Carlin (2001), Raftery et al. (2007), Friel and Pettitt (2008), Lenk (2009), Xie et al. (2011), Weinberg (2012)). Moreover, our new estimators can be used in the case of an improper prior (under the assumption that $A \subseteq \Theta$ is a set with non-zero and finite prior measure), in which $p(\theta)$ is a density function of some σ-finite measure such that $p(y)$ is non-zero and finite, i.e. the posterior distribution exists.

3 Simulation study

In this section we present simulation studies for three classes of models in which, under some assumptions, the true values of the marginal likelihoods are known (i.e. can be calculated analytically). We consider two examples presented by Lenk (2009) (the conjugate normal model and linear regression models) and autoregressive (AR) models. As regards AR models, we elaborate two cases: first, autoregressive models with conjugate inverse Gamma–Normal prior distribution; second, autoregressive models with covariance stationarity conditions and independent inverse Gamma–Normal prior distributions. In the second class of AR models, unlike the conjugate AR models, close-form expression for the marginal likelihood does not exist. Thus we calculated the natural logarithm of Bayes factors in favour of the true (assumed) model compared with other ones. In addition to the new estimator, the HM, corrected HM, AM, Laplace–Metropolis (LM), and Chib's estimators are used. The Laplace–Metropolis estimator is defined as follows (see Raftery (1996)):

$$\hat{p}_{LM}(y) = (2\pi)^{d/2}\det(\Psi)^{1/2}p(y|\tilde{\theta})p(\tilde{\theta}), \tag{20}$$

where $\tilde{\theta}$ is the posterior mode of $\ln[p(y|\theta)p(\theta)]$, Ψ is the negative inverse Hessian of $\ln[p(y|\theta)p(\theta)]$ evaluated at $\theta = \tilde{\theta}$, and d is the dimension of θ. In our examples $\tilde{\theta}$

was estimated from the posterior simulation output. Function $\ln[p(y|\theta_{(q)})p(\theta_{(q)})]$ was computed for each $q = 1, \ldots, k$, and then the value of $\theta_{(q)}$ for which it is largest was taken. Matrix Ψ was approximated by the estimated posterior covariance matrix from the posterior simulation output (as suggested by Raftery (1996)).

As a "gold standard" for calculating the marginal likelihood, we also consider the method proposed by Chib (1995), and extended by Chib and Jeliazkov (2001). The Chib's method is based on the identity

$$p(y) = \frac{p(y|\theta^*)p(\theta^*)}{p(\theta^*|y)}, \tag{21}$$

leading to the following equation:

$$\ln p(y) = \ln p(y|\theta^*) + \ln p(\theta^*) - \ln p(\theta^*|y), \tag{22}$$

which holds for any $\theta^* \in \Theta$. In practice (for the sake of estimation efficiency), the vector θ^* is chosen from the area of high posterior density values. Usually, both the prior $p(\theta^*)$ and likelihood terms $p(y|\theta^*)$ can be easily calculated. The estimation of the posterior probability density function $p(\theta^*|y)$ is difficult, but it can be approximated based on output from an MCMC sampler. This method requires performing reduced Gibbs steps (or reduced Metropolis–Hastings steps) in addition to the Gibbs sampler (or to the Metropolis–Hastings algorithm).

3.1 Conjugate normal model

As a very simple example let us consider the conjugate normal model, similar to Lenk (2009). Let us assume that $y_t|\mu, \sigma^2 \sim N(\mu, \sigma^2)$, $t = 1, 2, \ldots, N$, and $\mu|\sigma^2 \sim N(m_0, \sigma^2/w_0)$, $\sigma^2 \sim IG(r_0/2, s_0/2)$, where $N(m_0, \sigma^2/w_0)$ denotes the Normal distribution with mean m_0 and variance σ^2/w_0, in turn, $IG(r_0/2, s_0/2)$ stands for the inverse Gamma distribution with mean $\frac{s_0/2}{r_0/2-1}$ for $r_0 > 2$, and variance $\frac{(s_0/2)^2}{(r_0/2-1)^2(r_0/2-2)}$ for $r_0 > 4$. Whereas the conditional posterior distribution for μ is a Normal. i.e.

$$\mu|\sigma^2, y \sim N(m_N, \sigma^2/w_N), \quad \text{where } m_N = \frac{N\bar{y} + w_0 m_0}{N + w_0},\ w_N = N + w_0,$$

the marginal posterior distribution for σ^2 is an inverse Gamma:

$$\sigma^2|y \sim IG(r_N/2, s_N/2), \quad \text{where } r_N = r_0 + N,\ s_N = \sum_{t=1}^{N}(y_t - \bar{y})^2 + \frac{Nw_0}{N + w_0}(\bar{y} - m_0)^2 + s_0,$$

$\bar{y}$ is the Arithmetic Mean for sample y.

By integrating out parameters from the joint distribution of parameters and data, $p(y|\mu, \sigma^2)p(\mu, \sigma^2)$, the marginal density of the data can be presented in closed form (see Lenk (2009)) as

$$p(y) = \pi^{-N/2}\left(\frac{w_0}{w_N}\right)^{1/2}\frac{\Gamma(r_N/2)s_0^{r_0/2}}{\Gamma(r_0/2)s_N^{r_N/2}}. \tag{23}$$

In Figure 1, we present natural logarithms of consecutive estimates obtained for HM, CHM and CAM estimators, in the framework of the conjugate normal model for 100

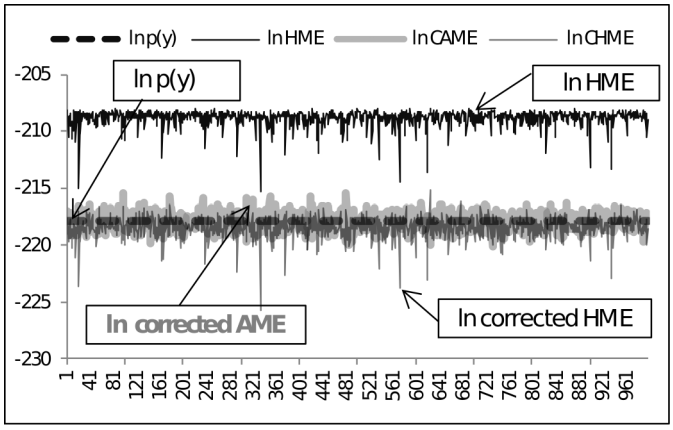

Figure 1: The Harmonic and Arithmetic Mean estimators in the conjugate normal model. 10000 iterations in each Monte Carlo procedure were used. 1000 reestimations for $N = 100$ observations from a Normal distribution with a mean of 30 and a variance of 4 were carried out. The prior hyperparameters were: $m_0 = 0, w_0 = 0.05, r_0 = 3, s_0 = 3$ (see Lenk (2009)). The true value of logarithm of likelihood is equal to -217.893, $A = [\min\{\mu_{(q)}\}, \max\{\mu_{(q)}\}] \times [\min\{\sigma^2_{(q)}\}, \max\{\sigma^2_{(q)}\}]$.

observations generated from a normal distribution with a mean of 30 and a variance of 4. Similar to Lenk (2009), the prior parameters are as follows: $m_0 = 0, w_0 = 0.05, r_0 = 3, s_0 = 3$. True value of logarithm of likelihood is equal to -217.893 (the horizontal, dashed line in Figure 1). We report results based on 10000 Monte Carlo simulations, which were repeated 1000 times. As regards subset A, in this model we assume that it is a rectangle limited by the range of the Monte Carlo sampler output $(\{\theta_{(q)}\}_{q=1}^{k}$ drawn from the posterior distribution, where $\theta_{(q)} = (\mu_{(q)}, \sigma^2_{(q)})'$), i.e. $A = [\min\{\mu_{(q)}\}, \max\{\mu_{(q)}\}] \times [\min\{\sigma^2_{(q)}\}, \max\{\sigma^2_{(q)}\}]$. The prior probability of subset A, $P(A)$, was approximated, using importance sampling. As a sampling distribution, the inverse Gamma–Normal truncated to A was used. The mean and variance of the sampling distribution were set at the mean and variance of the posterior distribution of $(\mu, \sigma^2)'$. The evaluation of $p(y)$ using (16) requires estimation of $P(A|y)$. We put $\hat{P}(A|y) = 1$ and applied MCIS method with the importance function being an independent inverse Gamma–Normal with the mean and variance of the posterior distribution (estimated from the MC draws). We can see from Figure 1 that the natural logarithms of the corrected HM and AM estimators cover the true value of $\ln p(y)$. The spread of estimates is clearly smaller for the CAM estimator. The prior probability of A, $P(A)$, is approximately $1.17 \cdot 10^{-4}$.

In Table 1, average errors (AE) and root mean squared errors (RMSE) in the conjugate normal model are presented. In this model $\ln p(y)$ can be analytically calculated, and thus we can compare the true value of $\ln p(y)$ with its estimates. As mentioned above, we consider a few very popular estimators of the marginal likelihood: the Harmonic Mean, Corrected Harmonic Mean, Arithmetic Mean, Corrected Arithmetic Mean (newly proposed), Chib, and Laplace–Metropolis estimators. All of the realizations of

		true ln $p(y)$	ln HME	ln CHME	ln AME	**ln CAME**	ln ChibE	ln LME
$N =$ 50	Mean	–116.144	–108.407	–116.69	–116.144	**–116.158**	–116.144	–116.009
	AE	–	–7.737	0.547	0	**0.014**	0	–0.135
	RMSE	–	7.769	0.888	0	**0.423**	0	0.136
$N =$ 100	Mean	–222.402	–213.560	–222.895	–222.402	**–222.421**	–222.402	–222.334
	AE	–	–8.842	0.492	0	**0.018**	0	–0.069
	RMSE	–	8.872	0.996	0	**0.783**	0	0.069

Table 1: Mean, average error (AE, true – estimated) and root mean squared error (RMSE) in the conjugate normal model. Results obtained for $N = 50$ and $N = 100$ observations from a normal distribution with a mean of 30 and a variance of 4. 1000 datasets were generated. The ln $p(y)$ was estimated with Monte Carlo sampler based on 10000 iterations. The ln HME denotes the natural logarithm of the Harmonic Mean estimator, ln CHME – the natural logarithm of the Corrected Harmonic Mean estimator, ln AME – the natural logarithm of the Arithmetic Mean estimator, ln ChibE – the natural logarithm of the Chib's estimator, ln CAME – the natural logarithm of the Corrected Arithmetic Mean estimator, ln LME – the natural logarithm of the Laplace–Metropolis estimator.

these estimators are computed on the log scale. The Arithmetic Mean estimator and the Chib's method perform best of all considered methods, but it is natural in such a simple model. Normalizing constants of the conditional posterior distribution for μ and of the marginal posterior distribution for σ^2 are known, therefore the use of the Chib's method or the Arithmetic Mean leads to the true value of the marginal likelihood. Consequently, average errors and RMSE are equal to zero. Because of "simulation pseudo-bias", the HM estimator performs worst. The HM estimator has the highest average and root mean squared errors from all estimators under consideration. It is natural because the RMSE can be written as a root of a sum of the variance term and the bias term: $RMSE(\ln \hat{p}(y)) = \sqrt{E[\ln \hat{p}(y) - \ln p(y)]^2} = \sqrt{Var(\ln \hat{p}(y)) + (E[\ln \hat{p}(y)] - \ln p(y))^2}$. Thus, more biased estimators have higher values of the RMSE. The Corrected Arithmetic Mean estimator has smaller absolute values of average error than the Laplace–Metropolis estimator, but the latter has smaller root mean squared errors. The results demonstrate that our proposed estimator can be better than the CHM estimator.

3.2 Linear regression models

Simulation properties of our new estimator can be easily checked in linear regression models. Such a model can be written in the following standard notation:

$$y = X\beta + \varepsilon, \tag{24}$$

where X is an $N \times K$ matrix of regressors, β is a $K \times 1$ vector of parameters. Moreover, we assume that $\varepsilon \sim N(0, \sigma^2 I_N)$ with a conjugate family of distributions, $\beta|\sigma^2 \sim N_K(b_0, \sigma^2 V_0)$, $\sigma^2 \sim IG(r_0/2, s_0/2)$, and with prior hyperparameters, $b_0 = 0$, $V_0 = 7I_K$, $r_0 = 2$, $s_0 = 1$. Values of regressors are generated from a standard normal distribution. Moreover, to simulate datasets we generated samples of size $N = 25, 100, 200$, data points from model (24) with $K = 3, 20, 40, 100$, $\varepsilon \sim N(0, I_N)$, and $\beta = (5, 1, -2, 1, 1, \ldots,$

1)′. It has been shown that the joint posterior density for β and σ^2 also has inverse Gamma–Normal form. The posterior distribution for β conditional on σ^2 and the marginal posterior distribution of σ^2 are as follows:

$$\beta|\sigma^2, y \sim N_K(b_1, \sigma^2 V_1), \tag{25}$$

$$\sigma^2|y \sim IG\left(\frac{N+r_0}{2}, \frac{y'y - b_1'V_1^{-1}b_1 + b_0'V_0^{-1}b_0 + s_0}{2}\right), \tag{26}$$

where $V_1 = (X'X + V_0^{-1})^{-1}$ and $b_1 = V_1(X'y + V_0^{-1}b_0)$. The natural logarithm of the marginal data density value can be expressed as

$$\ln p(y) = \ln a - \frac{r_0+N}{2} \ln\left(1 + \frac{1}{s_0}(y - Xb_0)'(I_N + XV_0X')^{-1}(y - Xb_0)\right), \tag{27}$$

where $\ln a = -\frac{N}{2}\ln(\pi s_0) - \frac{1}{2}\ln\det(I_N + XV_0X') + \ln\Gamma\left(\frac{r_0+N}{2}\right) - \ln\Gamma\left(\frac{r_0}{2}\right)$ (see, e.g. Appendix to Lenk (2009)).

Results obtained using various estimators of the log-marginal likelihood are presented in Table 2. First of all, for the CHM estimator we assume that the subset A is an intersection of the parameter space where the conditional density function $p(y|\theta)$ exceeds the smallest value of $p(y|\theta)$ evaluated at pseudo-random sample $\{\theta_{(q)}\}_{q=1}^k$ from the posterior distribution and the hypercuboid limited by the range of the sampler output: $A = \bigotimes_i [\min\{\theta_{i(q)}\}, \max\{\theta_{i(q)}\}] \cap \{\theta : p(y|\theta) \geq L\}$, where $\{\theta_{i(q)}\}$ is the ith component of the vector $\{\theta_{(q)}\}$, $i = 1, \ldots, K+1$, $L = \min\{p(y|\theta_{(q)}), q = 1, \ldots, k\}$, $\theta_{(q)} = (\beta_{(q)}', \sigma^2_{(q)})'$.

Moreover, the CHM estimator is computed using three methods of "simulation pseudo-bias" assessments. Namely, to approximate $P(A)$, similar to Lenk (2009), we use Monte Carlo Importance Sampling: Importance Sampling with Uniform distribution as an importance one, MCIS with Normal distribution, and MCIS with independent inverse Gamma–Normal distribution. The first two are based on Lenk's proposition and his fragments of code (available in supplemental materials for Lenk (2009)). In the latter method, specifications of mean and variance of importance distribution are evaluated with MC draws from posterior distributions of β and σ^2. In turn, for the CAM estimator we assume that the subset A is a hypercuboid limited by the range of the posterior sampler output, $A = \bigotimes_i [\min\{\theta_{i(q)}\}, \max\{\theta_{i(q)}\}]$. Consequently, $\hat{P}(A|y) = 1$ and the prior simulation support is in regions where the likelihood is significant. To simulate datasets, samples of size $N = 100$ data points from the model (24) were generated. The Monte Carlo sampler was run for 10000 iterations. The true and estimated marginal data density values are presented in Table 2 and Figure 2.

In Table 2, we also present average errors and root mean squared errors. The Corrected Arithmetic Mean estimator performs best. For this estimator the average errors and RMSE are the smallest. Although three importance sampling distributions generate different estimates produced by the same CHM estimator, each of them is worse than the estimates produced by the CAM estimator with the inverse Gamma–Normal density as the importance function. Moreover, it seems that the CAM estimator has (as compared to the Harmonic Mean) a clearly less spread distribution (see Figure 2). Note

	$K = 3$ $N = 25$	$K = 3$ $N = 100$	$K = 20$ $N = 100$	$K = 40$ $N = 100$	$K = 100$ $N = 200$
average of the true $\ln p(y)$	–44.94	–154.18	–200.59	–251.47	–579.63
average error (true-estimated)					
ln HME	–6.555	–9.428	–51.045	–100.925	–290.307
ln CHME; MCIS: Uniform	0.656	0.634	1.241	–84.385	–229.304
ln CHME; MCIS: Normal	0.042	1.757	2.241	1.219	4.235
ln CHME; MCIS: inverse Gamma–Normal	0.674	0.635	1.161	1.453	1.773
ln LME	–1.509	–1.155	–5.701	–9.198	–12.260
ln AME	1.248	9.600	—	—	—
ln CAME; MCIS: inverse Gamma–Normal	**0.001**	**0.001**	**0.003**	**0.012**	**0.116**
RMSE					
ln HME	6.602	9.458	51.070	100.951	290.326
ln CHME; MCIS: Uniform	0.905	0.884	1.584	84.871	229.368
ln CHME; MCIS: Normal	1.183	1.848	2.712	2.624	5.338
ln CHME; MCIS: inverse Gamma–Normal	0.941	0.888	1.494	1.791	2.084
ln LME	1.509	1.155	5.712	9.233	12.377
ln AME	2.329	12.247	—	—	—
ln CAME; MCIS: inverse Gamma–Normal	**0.009**	**0.008**	**0.025**	**0.073**	**0.395**

Table 2: Average error and root mean squared error in the linear regression model. Results obtained for simulated data from linear regression models. 1000 datasets were generated. The ln $p(y)$ was estimated with Monte Carlo sampler based on 10000 iterations. The true value of ln $p(y)$ was analytically calculated. The ln HME denotes the natural logarithm of the Harmonic Mean estimator, ln CHME – the natural logarithm of the Corrected Harmonic Mean estimator, ln AME – the natural logarithm of the Arithmetic Mean estimator, ln CAME – the natural logarithm of the Corrected Arithmetic Mean estimator, ln LME – the natural logarithm of the Laplace-Metropolis estimator.

that for $K \geq 20$ the AM estimator is inefficient (an overflow or an underflow occur). It is so because the posterior distribution is much more concentrated than that of the prior. By trimming the space of parameters to the subset A, we eliminate very low likelihood regions, and consequently we reduce the variance of the estimator.

As regards the Chib's method, for any vector $\theta^* = (\beta^*, \sigma^{2*})$ the posterior density can be decomposed as $p(\theta^*|y) = p(\beta^*|\sigma^{2*}, y)p(\sigma^{2*}|y)$, where the first term is a density function of a Normal distribution, and the second is a density function of an inverse Gamma one, see (24) and (26). Consequently, the Chib's method reduces itself to the use of (27).

3.3 Autoregressive models

While the previous examples illustrated good properties of the new estimator in models with closed-form marginal density of the data, the following examples pertain to

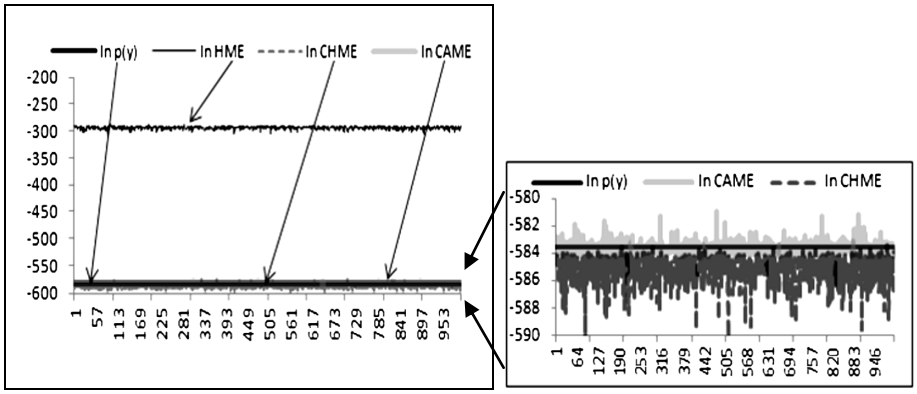

Figure 2: Estimates of the log-marginal likelihood ($\ln p(y)$) in linear model with $K = 100$. Simulation study of linear model using 200 simulated data points, $N = 200$. 10000 Monte Carlo iterations for estimation were used. The true value $\ln p(y)$ is analytically calculated. The ln HME denotes the natural logarithm of the Harmonic Mean estimator, ln CHME – the natural logarithm of the Corrected Harmonic Mean estimator, ln CAME – the natural logarithm of the Corrected Arithmetic Mean estimator.

models in which analytical derivation of marginal posterior densities by integrating out parameters from a joint distribution (which determines the normalizing constant of the posterior distribution) may not be possible. We consider the autoregressive models of order p,

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varepsilon_t, \quad t = 1, \ldots, N, \tag{28}$$

with $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)' \sim N(0, \sigma^2 I_N)$, independent inverse Gamma–Normal prior distribution (i.e. $\beta \sim N_{p+1}(b_0, V_0)$, $\sigma^2 \sim IG(r_0/2, s_0/2)$), and with stationarity conditions for $\{y_t\}$. In AR(p) models with independent prior distributions, the joint posterior density for β and σ^2 does not take a standard form, but the full conditional distributions are standard:

$$\beta|\sigma^2, y \sim N_K(b_N,\ V_N) I_S(\beta), \tag{29}$$

$$\sigma^2|\beta, y \sim IG\left(\frac{N + r_0}{2},\ \frac{(y - X\beta)'(y - X\beta) + s_0}{2}\right), \tag{30}$$

where $V_N = (X'X/\sigma^2 + V_0^{-1})^{-1}$, $b_N = V_N(X'y/\sigma^2 + V_0^{-1}b_0)$, S is the subset of parameter space for β, where stationarity conditions of $\{y_t\}$ are satisfied. Thus, the Gibbs sampler can be used to simulate from the joint posterior distribution (by simulation from these Normal and inverse Gamma distributions). Similar to the linear regression model for the CAM estimator, we assumed that the subset A constitutes an intersection of the parameter space and the hypercuboid limited by the range of the posterior sampler output, $A = \underset{i}{\otimes}[\min\{\theta_{i(q)}\}, \max\{\theta_{i(q)}\}] \cap S$, whereas for the CHM estimators: $A = \underset{i}{\otimes}[\min\{\theta_{i(q)}\}, \max\{\theta_{i(q)}\}] \cap \{\theta : p(y|\theta) \geq L\} \cap S$, $L = \min\{p(y|\theta_{(q)}), q = 1, \ldots, k\}$, $i = 1, \ldots, K + 1$, $\theta_{(q)} = (\beta'_{(q)}, \sigma^2_{(q)})'$. A sample of size $N = 100$ data points was generated from model (28) with $\varepsilon \sim N(0, 0.01 I_N)$, $p = 2, 5, 15$, $\beta = (0.1, 0.4, -0.3)'$ or

AR(p) model		$p=2$	$p=5$	$p=15$		$p=2$	$p=5$	$p=15$
TRUE $\ln p(y)$	Mean	93.295	80.009	87.494	RMSE	0	0	0
	SD	0	0	0	AE	0	0	0
ln HME	Mean	95.042	81.318	89.068	RMSE	1.830	1.398	1.656
	SD	0.544	0.488	0.515	AE	1.747	1.309	1.574
ln CHME	Mean	92.903	79.666	86.579	RMSE	0.599	0.545	1.301
MCIS: Uniform	SD	0.453	0.424	0.925	AE	–0.393	–0.343	–0.915
ln CHME	Mean	92.397	80.021	89.317	RMSE	1.040	0.599	2.127
MCIS: Normal	SD	0.524	0.599	1.094	AE	–0.898	0.012	1.823
ln CHME	Mean	92.901	79.664	87.126	RMSE	0.603	0.546	0.564
MCIS: IG–N	SD	0.544	0.423	0.428	AE	–0.395	–0.345	–0.368
ln LME	Mean	94.448	82.192	92.327	RMSE	1.153	2.184	4.840
	SD	0.016	0.049	0.261	AE	1.153	2.183	4.833
ln AME	Mean	93.295	80.007	87.493	RMSE	0.077	0.051	0.060
	SD	0.077	0.051	0.060	AE	–0.0005	–0.002	–0.003
ln CAME	**Mean**	**93.294**	**80.008**	**87.490**	**RMSE**	**0.008**	**0.010**	**0.020**
MCIS: IG–N	**SD**	**0.008**	**0.010**	**0.019**	**AE**	**–0.001**	**–0.001**	**–0.003**

Table 3: Results obtained in AR(p) models with conjugate inverse Gamma–Normal priors, y held fixed for each p. SD – standard deviation, RMSE – root mean squared error, AE – average error (true – estimate). $\ln p(y)$ was calculated via its analytical expression, given in (23), IG–N – inverse Gamma–Normal.

$\beta = (0.1, 0.4, 0.03, \ldots, 0.03, -0.3)'$ (depending on p). Prior hyperparameters were assumed to be $r_0 = 4$, $s_0 = 0.1$, $V_0 = diag(v_{0,ii})$, $v_{0,ii} = 0.1i^{-2}$, and $b_0 = (0.1, 0.4, -0.3)'$ or $b_0 = (0.1, 0.4, 0.03, \ldots, 0.03, -0.3)'$. The prior hyperparameters are chosen so that the prior probability of stationarity conditions is not too small. Consequently, we can directly draw from the prior distribution.

It is worth noting that for a conjugate prior distribution, i.e. $\beta|\sigma^2 \sim N_{p+1}(b_0, \sigma^2 V_0)$ and $\sigma^2 \sim IG(r_0/2, s_0/2)$, this model is a special case of the normal linear model with

$$y = (y_1, y_2, \ldots, y_N)' \quad \text{and} \quad X = \begin{bmatrix} 1 & y_0 & \cdots & y_{1-p} \\ 1 & y_1 & \cdots & y_{2-p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & y_{N-1} & \cdots & y_{N-p} \end{bmatrix}.$$

Therefore, in the conjugate AR models, the joint posterior density for $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ and σ^2 takes a well known form: a product of an inverse Gamma and a (conditional) Normal. Thus, the natural logarithm of the marginal data density value can be calculated from (27). In Table 3, we present the main characteristics of estimates of $\ln p(y)$, obtained with the use of the proposed estimator and a collection of popular methods. In normal autoregressive models with conjugate prior distributions (where the marginal data density value is available analytically) among considered estimators, the CAM estimator performs best because it is characterised by smallest root mean squared errors and average errors. Note that the CAM estimator has substantially smaller average errors than the CHM (with three different methods of estimation $P(A)$). Thus, the newly proposed estimator has better simulation properties than that proposed by Lenk (2009). Moreover, the CAM estimator is good for even small samples.

AR(p) models		$p=2$	$p=5$	$p=15$		$p=2$	$p=5$	$p=15$
TRUE $\ln p(y)$		–	–	–		–	–	–
ln HME	Mean	95.918	80.891	87.133	RMSE	4.243	3.554	3.775
	SD	0.727	0.797	0.623	AE	4.180	3.464	3.723
ln CHME	Mean	91.141	76.880	82.608	RMSE	0.832	0.853	1.450
MCIS: Uniform	SD	0.580	0.655	1.209	AE	–0.597	–0.547	–0.802
ln CHME	Mean	89.728	75.837	90.683	RMSE	2.104	1.725	8.428
MCIS: Normal	SD	0.620	0.669	4.257	AE	–2.010	–1.590	7.273
ln CHME	Mean	91.138	76.858	82.897	RMSE	0.844	0.880	0.717
MCIS: IG–N	SD	0.594	0.671	0.501	AE	–0.600	–0.569	–0.513
ln LM	Mean	92.858	79.472	87.572	RMSE	1.120	2.046	4.168
	SD	0.018	0.049	0.236	AE	1.120	2.045	4.162
ln AME	Mean	91.564	77.628	85.517	RMSE	0.624	0.378	2.113
	SD	0.599	0.321	0.169	AE	–0.174	0.201	2.106
ln CAME	**Mean**	**91.722**	**77.398**	**83.374**	**RMSE**	**0.018**	**0.031**	**0.044**
MCIS: IG–N	**SD**	**0.008**	**0.01**	**0.019**	**AE**	**–0.016**	**–0.029**	**–0.037**
ln ChibE(1)	Mean	91.738	77.427	83.410	RMSE	0	0	0
posterior maximum	SD	0.002	0.006	0.014	AE	0	0	0
ln ChibE(2)	Mean	91.738	77.428	83.422	RMSE	0.001	0.005	0.018
posterior mean	SD	0.002	0.003	0.004	AE	0.000	0.001	0.012

Table 4: Mean, standard deviation, root mean squared error and average error of the natural logarithm of estimates obtained in AR(p) models with independent inverse Gamma–Normal priors and stationarity conditions, y held fixed for each p. SD – standard deviation, RMSE – root mean squared error relative to ln ChibE(1), AE – average error relative to ln ChibE(1) (ln ChibE(1) – estimate), IG–N – inverse Gamma–Normal. $\ln p(y)$ is not available analytically.

Table 4 contains means, standard deviations, root mean squared errors and average errors (relative to ln ChibE(1)) of the natural logarithm of estimates obtained in AR(p) models with independent inverse Gamma–Normal priors and under stationarity conditions. Unlike the conjugate AR model a closed-form expression for the marginal likelihood does not exist. Therefore, the root mean squared errors and average errors are determined relative to the Chib's method calculated at the posterior maximum. The Gibbs sampler was run for 10000 iterations. Results of the initial draw were adopted as equal to maximum likelihood estimates. We assume $b_0 = (0.1, 0.4, 0.03, \ldots, 0.03, -0.3)'$ in the case of the AR(p) structure used, and $b_0 = (0.1, 0.4, 0.03, \ldots, 0.03)'$ for AR($p-1$) as well as for AR($p+1$) structures. The vector of the data used, y, in stationarity AR models and in the conjugate autoregressive model remains the same. For each $p \in \{2, 5, 15\}$ we estimate three models: AR(p) (which is true) and AR($p-1$) as well as AR($p+1$) (which are false). We use 1000 replications for each model.

As regards of the Chib's method, the joint posterior density of (β^*, σ^{2*}) can be decomposed as follows: $p(\theta^*|y) = p(\sigma^{2*}|y)p(\beta^*|\sigma^{2*}, y)$, where (β^*, σ^{2*}) is chosen to be the posterior mean of (β, σ^2) or to be the maximum of the posterior probability density function. Because the full-conditional densities, including their normalizing constants, are known (see (29) and (30)), the conditional density $p(\beta^*|\sigma^{2*}, y)$ is available directly, while the ordinate $p(\sigma^{2*}|y)$ is estimated from the draws of the Gibbs run (by averaging full-conditional densities).

	$p=2$ vs $p=3$	$p=2$ vs $p=1$	$p=5$ vs $p=6$	$p=5$ vs $p=4$	$p=15$ vs $p=16$	$p=15$ vs $p=14$
AR(p) models with conjugate inverse Gamma–Normal prior						
TRUE	1	1	1	1	1	1
ln HME	0.989	0.989	0.993	0.984	0.994	0.987
ln CHME; MCIS: Uniform	0.992	0.992	0.998	0.994	0.993	0.989
ln CHME; MCIS: Normal	0.994	0.994	0.990	0.993	0.996	0.988
ln CHME; MCIS: IG–N	0.993	0.993	0.997	0.994	0.999	0.995
ln LME	**1.000**	**1.000**	0.995	**1.000**	0.998	0.999
ln AME	**1.000**	**1.000**	0.998	0.998	0.999	0.999
ln CAME; MCIS: IG–N	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
AR(p) models with independent inverse Gamma–Normal prior and covariance stationarity						
ln HME	0.957	0.968	0.961	0.967	0.976	0.964
ln CHME; MCIS: Uniform	0.971	0.973	0.963	0.968	0.962	0.911
ln CHME; MCIS: Normal	0.969	0.967	0.941	0.973	0.102	0.463
ln CHME; MCIS: IG–N	0.971	0.972	0.964	0.967	0.979	0.972
ln LME	0.969	**0.996**	0.955	**0.992**	0.979	**0.979**
ln AME	**0.982**	0.973	0.965	0.977	**0.992**	0.977
ln CAME; MCIS: IG–N	**0.982**	0.987	**0.982**	0.982	0.978	0.975
ln ChibE(1); posterior maximum	**0.982**	0.987	**0.982**	0.982	0.978	0.974
ln ChibE(2); posterior mean	**0.982**	0.987	**0.982**	0.982	0.978	0.975

Table 5: Frequency of indications of the correct AR(p) structure based on logarithm of Bayes factors for the true AR(p) structure against others. Datasets for which analytically calculated Bayes factors indicate true specification are considered. Results obtained for simulated data from autoregressive models. 1000 datasets were generated. In AR(p) models with independent inverse Gamma–Normal (IG–N) prior, $\ln p(y)$ was estimated with Gibbs sampler based on 10000 iterations. The Chib's method was calculated at posterior means and at the numerical approximations of the maximum of the posterior probability density function. In the case of AR(p) models with the conjugate inverse Gamma–Normal prior, the true value of $\ln p(y)$ was analytically calculated; moreover, direct MC samples from the posterior distributions were used.

We can see from Table 4 that the Chib's estimators have the smallest dispersion measured by standard deviation. The average values of the natural logarithm of the CAM estimates are closer to those of the Chib's.

Due to the fact that in models with independent prior distributions the $p(y)$ cannot be derived analytically, we calculated Bayes factors for the true (assumed for the purpose of simulation) AR(p) model against neighbouring (false) models, AR($p \pm 1$). The results are summarized in Table 5 where frequencies of indications of the correct structure are presented. In the AR(p) models with conjugate inverse Gamma–Normal prior distributions, the CAM estimator favours the true specifications in all cases. The LM and AM estimators perform as the second best methods. In AR(p) models with independent inverse Gamma–Normal prior distributions, once again the LM, AM, CAM, and Chib's estimators perform best. Even though the number of iterations is fairly small and the HM estimator is biased (i.e. the marginal likelihoods are overestimated), this leads to the correct conclusion about model selection in over 95.7% of cases. The HM and CHM estimators provide worse estimates of the Bayes factors than does the CAM

estimator. Once again, the CAM estimator performs very well, being as good as the Chib's method.

4 Empirical examples

4.1 Non-nested linear regression models

We now present a classical example for Bayesian model selection considered among others in Carlin and Chib (1995), Han and Carlin (2001), Bartolucci et al. (2006), Friel and Pettitt (2008), where different methods for estimating the Bayes factor between two non-nested competing linear regression models have been compared. Based on $n = 42$ specimens of radiata pine, the maximum compressive strength parallel to the grain (y_t) is described related to specimen's density (x_t) or to its density adjusted for resin content (z_t). The data set is taken from Williams (1959). Two competing models are the following:

$$M_1\text{: } y_t = \alpha + \beta(x_t - \bar{x}) + \varepsilon_t, t = 1, 2, \ldots, n, \{\varepsilon_t\} \sim iiN(0, \sigma_\varepsilon^2),$$

$$M_2\text{: } y_t = \gamma + \delta(z_t - \bar{z}) + \eta_t, t = 1, 2, \ldots, n, \{\eta_t\} \sim iiN(0, \sigma_\eta^2).$$

In order to facilitate comparisons with other methods, we use the same prior specification as in articles cited above, i.e. $N(185,\ 10^4)$ for β and δ, $N(3000,\ 10^6)$ for α and γ, and $IG(6/2,\ 300^2/2)$ for σ_ε^2 and σ_η^2. We also assume prior independence among all the parameters. As regards subset A, used for both CAM and CHM estimators, it is assumed to be an intersection of the parameter space, where the conditional density function $p(y|\theta)$ exceeds the smallest value of $p(y|\theta)$ evaluated at pseudo-random sample $\{\theta_{(q)}\}_{q=1}^k$ from the posterior distribution and the cuboid limited by the range of the Gibbs sampler output.

We estimated the marginal likelihoods for M_1 and M_2 models, and we computed the Bayes factor in favour of model M_2, i.e. B_{21}. Each estimate was calculated 100 times, so the bias, standard and relative errors of the estimates could be computed. The Chib's method was calculated at posterior means and at the numerical approximations of the maximum of the posterior probability density function – analogously to the example shown in Section 3.3.

The aim of this example is to compare estimates obtained by using the Corrected Arithmetic Mean with those obtained by the methods proposed by Bartolucci et al. (2006), Friel and Pettitt (2008). The Gibbs sampler was run for 40000 iterations, of which the first 10000 are treated as burn-in and are discarded. The results are displayed in Table 6, where the mean, average errors, standard errors, root mean squared errors, and relative RMSE for the Bayes factor are presented. To compute RMSE and relative RMSE, following Han and Carlin (2001), Friel and Pettitt (2008) and Bartolucci et al. (2006), we used $B_{21} = 4862$ computed by Green and O'Hagan (1998).

As can be seen from Table 6, the CAM estimator performs very well, having an efficiency very similar to (even quite better than) the Chib's marginal likelihood method. Of course, in these models the Chib's method is not difficult to use because full conditional

Estimator	Mean	Average error	Standard deviation	RMSE	Relative error (%)	median
HME	5511.66	–649.66	3142.15	3208.61	65.99	4838.11
CHME; MCIS: Uniform	5139.91	–277.91	2241.80	2258.96	46.46	4738.65
CHME; MCIS: Normal	5555.32	–693.32	4359.21	4414.00	90.79	4420.21
CHME; MCIS: inverse Gamma–Normal	4962.57	–100.57	2262.99	2265.22	46.59	4528.62
LME	4857.72	4.28	55.58	55.74	1.15	4853.49
AME; prior	4869.86	–7.86	612.69	612.74	12.60	4858.44
CAME; MCIS: inverse Gamma–Normal	**4861.45**	**0.55**	**7.98**	**8.00**	**0.16**	**4861.89**
ChibE (posterior maximum)	4861.02	0.98	8.89	8.94	0.18	4861.78
ChibE (posterior mean)	4860.89	1.14	9.55	9.62	0.20	4861.17
Friel and Pettitt (2008), p. 601:						
Serial MCMC	4852.00	10	132	–	–	–
Population MCMC	4840.00	22	154	–	–	–
Bartolucci et al. (2006), p. 49:						
Rao-Blackwellized Rev. Jump:	4864.8	–2.8	204.5	–	4.21	–
optimal bridge function (1):	4864.3	–2.3	204.4	–	4.20	–
optimal bridge function (2):	4848.9	13.10	246.3	–	5.07	–

Table 6: Comparison of the Bayes factor estimators for non-nested linear regression models. Average error, RMSE and relative error are calculated relative to $B_{21} = 4862$. The Chib's method was evaluated at posterior means and at the numerical approximations of the maximum of the posterior probability density function.

distributions are standard and easy to sample from. Performance of the Harmonic and Corrected Harmonic Mean estimators is the poorest, but it is possible that the CHM estimator could be improved by selecting better importance functions. It is very important to stress that our method leads to estimate of B_{21} with better efficiency to that of the power posterior methods, presented by Friel and Pettitt (2008). Biases and standard errors for estimates B_{21}, obtained by using power posterior methods, are several times greater than those obtained by using our Corrected Arithmetic Mean. The bias and the standard error for estimates of the Bayes factor, obtained by using the serial MCMC approach, are equal to 10 and 132, respectively. Also, we can compare our results with those presented by Bartolucci et al. (2006). According to their results, the estimators of B_{21} based on an extension of the bridge sampling identity perform worse than the CAM estimator. The standard errors for these estimates range from 204.5 to 246.3 (being over 25-fold higher than those for the CAM estimator), and the relative errors range from 4.20% to 5.07%, being over 26-fold higher than for the CAM estimator.

4.2 Mixture models for galaxy data

Now we will show that using our new estimator can help one to handle difficult problems with multimodal posterior distributions. Thus, in this section our method of approximation of the marginal likelihood is applied to one-dimensional Gaussian mixture models for so-called "galaxy data". The dataset consists of the velocities (in 10^3 km/s) of 82 distant galaxies, diverging from our own galaxy, presented by Postman et al. (1986) and analysed with the use of different mixture models by a number of researches, including,

for example, Roeder (1990), Carlin and Chib (1995), Chib (1995), Phillips and Smith (1996), Raftery (1996), Richardson and Green (1997), Neal (1999), Liang and Wong (2001), Steele et al. (2006), and many others.[3] A Gaussian mixture model that we apply in our examples is considered by Chib (1995). In the model with d components, the conditional probability density function of independent and identically distributed observations, y_t, $t = 1, 2, \ldots, n$, is as follows:

$$p(y|\theta) = \prod_{t=1}^{n} \sum_{i=1}^{d} \omega_i \phi(y_t|\mu_i, \sigma_i^2), \tag{31}$$

where $y = (y_1, \ldots, y_n)$ is a vector of observations, $\omega = (\omega_1, \ldots, \omega_d)$ is a probability vector, elements of which add up to one, $\phi(y_t|\mu_i, \sigma_i^2)$ is a probability density of normal distribution with mean μ_i and variance σ_i^2, $\theta = (\mu_1, \ldots, \mu_d, \sigma_1^2, \ldots, \sigma_d^2, \omega_1, \ldots, \omega_d)'$ summarizes all unknown model parameters. Similar to Chib (1995) it is assumed that all elements of θ are mutually independent and have the following prior distributions: $\mu_i \sim N(\mu_0, \sigma_0^2)$, $\sigma_i^2 \sim IG(r_0/2, s_0/2)$, $\omega \sim Dirichlet(\alpha_1, \ldots, \alpha_d)$, where $\mu_0 = 20$, $\sigma_0^2 = 100$, $r_0 = 6$, $s_0 = 40$, $\alpha_i = 1$, $i = 1, \ldots, d$. As pointed out in Chib (1995), the prior hyperparameters are chosen so that these prior distributions reflect weak prior information about the parameters. The posterior distribution is given by

$$p(\theta|y) \propto p(\theta) \prod_{t=1}^{n} \sum_{i=1}^{d} \omega_i \phi(y_t|\mu_i, \sigma_i^2), \tag{32}$$

where $p(\theta)$ is the density function of the vector of parameters.

The posterior distribution has $d!$ symmetric modes because the prior distribution is exchangeable and the likelihood function is invariant to the permutation of the component labels (see Frühwirth-Schnatter (2001)). The symmetry of the posterior distributions of the parameters causes several numerical problems (e.g. lack of label-switching). To simulate from the posterior distribution, Chib (1995) used the augmented Gibbs sampler. However, the Gibbs sampler can be inefficient as a numerical tool to visit all $d!$ symmetric modes in the posterior distribution. There exist different Bayesian solutions for the problem (see, e.g. Celeux et al. (2000), Stephens (2000), Berkhof et al. (2003), Marin et al. (2005), Chung et al. (2004), Jasra et al. (2005), Frühwirth-Schnatter (2006)). In this paper we used a simple and effective algorithm proposed by Frühwirth-Schnatter (2001), the permutation sampling combined with the Gibbs sampler.

Similar to Chib (1995) the mixture model can be expressed in terms of independent and identically distributed latent variables associated with observations, z_t, $t = 1, \ldots, n$, that take values in a discrete space $\{1, 2, \ldots, d\}$, such that $Pr(z_t = i|\theta) = \omega_i$ and $p(y_t|z_t = i, \theta) = \phi(y_t|\mu_i, \sigma_i^2)$. It implies that

$$p(y_t|\theta) = \sum_{i=1}^{d} p(y_t, z_t = i|\theta) = \sum_{i=1}^{d} p(y_t|z_t = i, \theta) \Pr(z_t = i|\theta) = \sum_{i=1}^{d} \phi(y_t|\mu_i, \sigma_i^2)\omega_i,$$

which leads to the mixture model in (32).

[3] The original dataset consists of 83 observations, but one of them (a velocity of 5607 km/s) does not appear in the dataset analysed by Roeder (1990) and others.

As pointed out in Chib (1995), data-augmentation simplifies posterior simulation via Gibbs sampling from the full conditional distributions of $\omega = (\omega_1, \ldots, \omega_d)$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_d^2)$, $\mu = (\mu_1, \ldots, \mu_d)$ and $z = (z_1, \ldots, z_n)$. Let $T_i = \{t : z_t = i\}$ be the set of indices of observations which are classified into the ith population and let $n_i = \sum_{t=1}^{n} I_{\{i\}}(z_t)$ denote the number of observations assigned to the ith population. Then, the full conditional distributions are standard:

$$z_t|\theta, y \sim \Pr(z_t = i|\theta, y) = \frac{\omega_i \phi(y_t|\mu_i, \sigma_i^2)}{\sum_{i=1}^{d} \omega_i \phi(y_t|\mu_i, \sigma_i^2)}, i = 1, \ldots, d, t = 1, \ldots, n, \tag{33}$$

$$\mu|z, \sigma^2, \omega, y \sim N_d(\tilde{\mu},\ \tilde{\Sigma}_\mu), \tag{34}$$

where $\tilde{\mu} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ with $\tilde{\mu}_i = (\sigma_0^{-2} + \sigma_i^{-2} n_i)^{-1}(\sigma_0^{-2}\mu_0 + \sigma_i^{-2} n_i \sum_{t \in T_i} y_t)$, $\tilde{\Sigma}_\mu = diag(\tilde{\sigma}_{\mu,1}^2, \ldots, \tilde{\sigma}_{\mu,d}^2)$, $\tilde{\sigma}_{\mu,i}^2 = (\sigma_0^{-2} + \sigma_i^{-2} n_i)^{-1}$, $diag(\tilde{\sigma}_{\mu,1}^2, \ldots, \tilde{\sigma}_{\mu,d}^2)$ denotes a diagonal matrix with $\tilde{\sigma}_{\mu,1}^2, \ldots, \tilde{\sigma}_{\mu,d}^2$ on the diagonal;

$$\sigma_i^2|\sigma_{-i}^2, z, \mu, \omega, y \sim IG\left(\frac{n_i + r_0}{2},\ \frac{\delta_i + s_0}{2}\right), i = 1, \ldots, d, \tag{35}$$

where $\delta_i = \sum_{t \in T_i} (y_t - \mu_i)^2$, $\sigma_{-i}^2 = (\sigma_1^2, \ldots, \sigma_{i-1}^2, \sigma_{i+1}^2, \ldots, \sigma_d^2)$ is the vector σ^2 with its ith component removed;

$$\omega|z, \mu, \sigma^2, y \sim Dirichlet(\alpha_1 + n_1, \ldots, \alpha_d + n_d). \tag{36}$$

In models with an additional assumption that the variance σ_i^2 remains constant across components, the full conditional distribution of σ^2 is

$$\sigma^2|z, \mu, \omega, y \sim IG\left(\frac{n + r_0}{2},\ \frac{\delta + s_0}{2}\right), \tag{37}$$

where $\delta = \sum_{i=1}^{d} \sum_{t \in T_i} (y_t - \mu_i)^2$.

Because the Gibbs sampler may not explore the whole unconstrained parameter space, but tends to stick to one posterior mode with occasional switches among other posterior modes (as a consequence, some of the posterior modes are hardly ever or just never "visited"), the random permutation sampler (introduced by Frühwirth-Schnatter (2001)) is applied.

Similar to Raftery (1996), models with two, three, and four components are considered. In addition, mixture models with equal and unequal variances are taken into account. Results obtained using various simple estimators of the log-marginal likelihood are presented in Tables 7 and 8. Each method was repeated 100 times, thus 100 estimates was obtained for each procedure. As in the previous example, for the CHM and the CAM estimators, we assume that the subset A is an intersection of the parameter space, where

		unequal variance models			equal variance models		
number of components		$d=2$	$d=3$	$d=4$	$d=2$	$d=3$	$d=4$
ln HME	Mean	–229.842	–221.481	–221.615	–236.056	–220.213	–219.841
	SD	2.055	1.101	0.734	0.961	0.817	0.786
ln CHME;	Mean	–234.963	–227.321	–227.079	–240.461	–227.594	–226.531
MCIS:IG–N–D	SD	1.546	0.795	0.538	0.903	0.552	0.590
ln LME	Mean	–226.382	–211.976	–210.820	–236.133	–216.273	–213.497
	SD	0.561	0.108	0.145	0.106	0.021	0.097
ln AME	Mean	–233.315	–227.723	–227.454	–239.825	–228.111	–227.115
	SD	1.005	1.296	1.191	0.284	1.768	1.569
ln CAME; MCIS:IG–N–D	**Mean**	**–232.890**	**–227.098**	**–227.000**	**–239.765**	**–226.824**	**–226.428**
	SD	**0.192**	**0.649**	**0.962**	**0.051**	**0.260**	**0.805**
ln ChibE	Mean	–231.894	–225.659	–225.209	–238.566	–224.837	–224.036
max lnL	SD	0.737	0.212	0.271	1.115	1.119	1.054
ln ChibE	Mean	–231.940	–225.692	–225.141	–238.565	–224.904	–223.352
post. mode	SD	0.094	0.027	0.091	1.117	1.192	1.095
ln AME*:		–232.891	–226.775	–226.608	–239.768	–226.812	–226.069
Neal (1999):		–	–226.791	–	–239.764	–226.803	–

Table 7: Results for Galaxy Data. $\ln p(y)$ was estimated with the Gibbs sampler based on 30000 iterations. The Chib's method was calculated at one of the posterior modes and at the numerical approximations of the maximum of the posterior probability density function. ln AME* was computed based on 10^8 points drawn from the prior distribution. Results reported by Neal (1999) are also included.

the conditional density function $p(y|\theta)$ exceeds the smallest value of $p(y|\theta)$ evaluated at pseudo-random sample $\{\theta_{(q)}\}_{q=1}^{k}$ from the posterior distribution, and the hypercuboid limited by the range of the sampler output, $A = \bigotimes_i [\min\{\theta_{i(q)}\}, \max\{\theta_{i(q)}\}] \cap \{\theta : p(y|\theta) \geq \min\{p(y|\theta_{(q)}), q = 1, \ldots, k\}\}$. Consequently, $\hat{P}(A|y) = 1$, and the prior simulation support is in regions where the likelihood is significant. To estimate $p(y)$ using (16), we applied the MCIS method with the importance function being the independent inverse Gamma–Normal and Dirichlet with the mean and variance of the posterior distribution (estimated from the MCMC draws). Similarly, to approximate $P(A)$ we use the Monte Carlo Importance Sampling with independent inverse Gamma–Normal and Dirichlet distributions. The Gibbs sampler with random permutations was run for 31000 iterations, the first 1000 being discarded as burn-in.[4] In case of the Chib's method, the natural logarithm of the posterior density function, $\ln p(\theta^*|y)$, was computed from the decomposition

$$\ln p(\theta^*|y) = \ln p(\omega^*|\mu^*, {\sigma^2}^*, y) + \ln p({\sigma^2}^*|\mu^*, y) + \ln p(\mu^*|y), \tag{38}$$

where the vector θ^* was calculated, based on posterior simulation output as (i) a posterior mode of $\ln[p(y|\theta)p(\theta)]$, and (ii) a numerical approximate maximum likelihood value. Unsatisfactory results were obtained when the Chib's estimator was applied at the posterior means. In order to calculate components of (38), reduced Gibbs runs were used. The value of the posterior density function at θ^* was estimated through Rao–Blackwellization using reduced samples of size 30000. Since $p(y)$ is unknown, we

[4]The numerical procedure is much more time consuming than for other examples, thus the number of burn-in iterations was reduced to 1000.

		unequal variance models			equal variance models		
number of components		$d = 2$	$d = 3$	$d = 4$	$d = 2$	$d = 3$	$d = 4$
ln HME	RMSE	3.677	5.407	5.047	3.833	6.650	6.278
	AE	–3.049	–5.294	–4.994	–3.711	–6.599	–6.228
ln CHME;	RMSE	2.584	0.964	0.715	1.139	0.957	0.749
MCIS:IG–N–D	AE	2.071	0.545	0.470	0.694	0.782	0.461
ln LME	RMSE	6.533	14.800	15.790	3.636	10.540	12.573
	AE	–6.509	–14.799	–15.789	–3.635	–10.540	–12.572
ln AME	RMSE	1.090	1.606	1.460	0.290	2.194	1.886
	AE	0.423	0.948	0.845	0.058	1.299	1.046
ln CAME; MCIS:IG–N–D	**RMSE**	**0.210**	**0.724**	**1.038**	**0.050**	**0.260**	**0.882**
	AE	**–0.012**	**0.322**	**0.392**	**–0.003**	**0.011**	**0.358**
ln ChibE	RMSE	1.240	1.139	1.425	2.201	2.271	2.290
max lnL	AE	–0.997	–1.116	–1.399	–1.201	–1.976	–2.033
ln ChibE	RMSE	0.958	1.083	1.470	1.641	2.250	2.930
post. mode	AE	–0.953	–1.083	–1.467	–1.203	–1.908	–2.718

Table 8: Performance of different estimators of the marginal likelihood in Gaussian mixture models. Results for Galaxy Data. The Chib's method was calculated at one of the posterior modes and at the numerical approximations of the maximum of the posterior probability density function.

estimated it by the Arithmetic Mean estimator using 10^8 draws from the prior distribution. For the sake of comparison, results reported by Neal (1999) are also included in Table 7.

In Table 8, average errors (AE) and root mean squared errors (RMSE) obtained in equal and unequal variance models are presented. Closed-form expression for the marginal likelihood does not exist, therefore root mean squared errors and average errors are calculated relative to ln AME*. As can be seen from the results in Table 8, our new estimator performs best. The CAM estimator appears to work very well, being better than the Chib's method. It has the smallest average errors and RMSE. Due to multimodality of the posterior distributions, the Laplace–Metropolis estimator (here used in a naive manner) is the least accurate across all models. Despite the simulation pseudo-bias, the Harmonic Mean estimator works better than the LM estimator. Finally, it is important to stress that the CAM estimator can be easily implemented because it requires only additional draws from inverse the Gamma–Normal and Dirichlet distributions, centered at the posterior means.

5 Discussion

In the paper a new class of estimators of the marginal data density is proposed. The idea of the estimators is based on correction of the arithmetic mean estimator by trimming the prior sample to certain subset $A \subseteq \Theta$. We show that under following assumptions:

1. The subset A is compact;
2. A is a set with non-zero and finite prior as well as posterior measures;
3. The likelihood is bounded on A;

4. It is possible to generate samples forming an ergodic Markov chain whose equilibrium distribution is the posterior distribution;

5. Random samples can be generated from the prior distribution (or from the importance sampling distribution);

the new estimators are consistent. Moreover, the CAM estimators can be used in the case of an improper prior when the posterior distribution is proper. In simulation we focus on simple models where the marginal data density, $p(y)$, is known. Comparison of the true value of the marginal likelihood with its estimates obtained with the use of different estimators points to higher accuracy of the method proposed in this paper. The operational characteristics of our estimator were also illustrated using two very popular data examples. The first involved a choice between two non-nested regression models for specimens of radiata pine. The second example involved mixture Gaussian models for galaxy data. These results demonstrate that the AM estimators are promising. The properties of the new estimators depend on the choice of the subset A. At present there is no answer to the question: "What is the most appropriate subset A?". The challenge will be to choose the subset A so as to minimize variance of the estimator. Simulation studies and empirical examples considered in the paper demonstrate that the choice of A as an intersection of the parameter space (with the additional restriction that the likelihood function exceeds the smallest value of the likelihood evaluated at pseudo-random sample from the posterior distribution) and the hypercuboid limited by the range of the posterior sampler output is effective in the sense of a relatively small simulation sample being sufficient to obtain acceptable results.

Supplementary Material

Supplementary Material of "Estimating the Marginal Likelihood Using the Arithmetic Mean Identity" (DOI: 10.1214/16-BA1001SUPP; .pdf).

References

Ardia, D., Bastürk, N., Hoogerheide, L., and van Dijk, H. K. (2012). "A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood." *Computational Statistics and Data Analysis*, 56: 3398–3414. MR2943902. doi: http://dx.doi.org/10.1016/j.csda.2010.09.001. 264

Bartolucci, F., Scaccia, L., and Mira, A. (2006). "Efficient Bayes factor estimation from the Reversible Jump output." *Biometrika*, 93: 41–52. MR2277738. doi: http://dx.doi.org/10.1093/biomet/93.1.41. 264, 277, 278

Berkhof, J., van Mechelen, I., and Gelman, A. (2003). "A Bayesian approach to the selection and testing of mixture models." *Statistica Sinica*, 13:423–442. MR1977735. 279

Carlin, B. P. and Chib, S. (1995). "Bayesian Model Choice via Markov Chain Monte Carlo." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57: 473–484. 277, 279

Chen, M.-H. (1994). "Importance-weighted marginal Bayesian posterior density estimation." *Journal of the American Statistical Association*, 89: 818–824. MR1294727. 267

Chen, M.-H. (2005). "Computing marginal likelihoods from a single MCMC output." *Statistica Neerlandica*, 59: 16–29. MR2137379. doi: http://dx.doi.org/10.1111/j.1467-9574.2005.00276.x. 267

Chib, S. (1995). "Marginal likelihood from the Gibbs output." *Journal of the American Statistical Association*, 90: 1313–1321. MR1379473. 264, 279, 280

Chib, S. and Jeliazkov, I. (2001). "Marginal likelihood from the Metropolis–Hastings output." *Journal of the American Statistical Association*, 96: 270–281. MR1952737. doi: http://dx.doi.org/10.1198/016214501750332848. 264

Chung, H., Loken, E., and Schafer, J. L. (2004). "Difficulties in Drawing Inferences With Finite–Mixture Models: A Simple Example with a Simple Solution." *The American Statistician*, 58: 152–158. MR2109393. doi: http://dx.doi.org/10.1198/0003130043286. 279

Celeux, G., Hurn, M., and Robert, C. (2000). "Computational and inferential difficulties with mixtures posterior distribution." *Journal of the American Statistical Association*, 95(3):957–979. MR1804450. doi: http://dx.doi.org/10.2307/2669477. 279

Friel, N. and Pettitt, A. N. (2008). "Marginal likelihood estimation via power posteriors." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70: 589–607. MR2420416. doi: http://dx.doi.org/10.1111/j.1467-9868.2007.00650.x. 264, 267, 277, 278

Friel, N. and Wyse, J. (2012). "Estimating the evidence: a review." *Statistica Neerlandica*, 66(3): 288–308. MR2955421. doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00515.x. 264

Frühwirth-Schnatter, S. (2001). " Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models." *Journal of the American Statistical Association*, 96: 194–209. MR1952732. doi: http://dx.doi.org/10.1198/016214501750333063. 279, 280

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York: Springer. MR2265601. 264, 279

Gelman, A. and Meng, X.-L. (1998). "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling." *Statistical Science*, 13: 163–185. MR1647507. doi: http://dx.doi.org/10.1214/ss/1028905934. 264

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82: 711–732. MR1380810. doi: http://dx.doi.org/10.1093/biomet/82.4.711. 264

Green, P. J. and O'Hagan, A. (1998). "Carlin and Chib do not need to sample from pseudopriors." *Research Report 98-1*, University of Nottingham, Nottingham. 277

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods.* London: Methuen and Co LTD. MR0223065. 265

Han, C. and Carlin, B. P. (2001). "MCMC methods for computing Bayes factors: A comparative review." *Journal of the American Statistical Association*, 96: 1122–1132. 267, 277

Jasra, A., Holmes, C., and Stephens, D. (2005). "Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling." *Statistical Science*, 20(1):50–67. MR2182987. doi: http://dx.doi.org/10.1214/088342305000000016. 279

Kass, R. and Raftery, A. (1995) "Bayes factors." *Journal of the American Statistical Association*, 90: 773–795. MR3363402. doi: http://dx.doi.org/10.1080/01621459.1995.10476572. 264

Lenk P. (2009) "Simulation pseudo-bias correction to the harmonic mean estimator of integrated likelihoods." *Journal of Computational and Graphical Statistics*, 18: 941–960. MR2750446. doi: http://dx.doi.org/10.1198/jcgs.2009.08022. 262, 263, 267, 268, 269, 271, 274

Lewis, S. M. and Raftery, A. E. (1997). "Estimating Bayes factors via posterior simulation with the Laplace–Metropolis estimator." *Journal of the American Statistical Association*, 92: 648–655. MR1467855. doi: http://dx.doi.org/10.2307/2965712. 265, 267

Liang, F. and Wong, W. H. (2001). "Real-Parameter Evolutionary Monte Carlo with Applications to Bayesian Mixture Models." *Journal of the American Statistical Association*, 96: 653–666. MR1946432. doi: http://dx.doi.org/10.1198/016214501753168325. 279

Marin, J.-M., Mengersen, K., and Robert, C. (2005). "Bayesian modelling and inference on mixtures of distributions." In: Rao, C. and Dey, D. (eds.), *Handbook of Statistics*, 25: 459–507, Springer-Verlag, New York. MR2490536. doi: http://dx.doi.org/10.1016/S0169-7161(05)25016-2. 279

McCulloch, R. E. and Rossi, P. E. (1992). "Bayes factors for nonlinear hypotheses and likelihood distributions." *Biometrika*, 49(4): 663–676. MR1209468. doi: http://dx.doi.org/10.1093/biomet/79.4.663. 265

Meng, X.-L. and Wong, W.-H. (1996). "Simulating ratios of normalizing constants: a theoretical exploration." *Statistica Sinica*, 6: 831–860. MR1422406. 264, 267

Neal, R. M. (1999). "Erroneous Results in "Marginal Likelihood from the Gibbs Output"." Available online at http://www.cs.toronto.edu/~radford/ftp/chib-letter.pdf. 279, 281, 282

Neal, R. M. (2001). "Annealed importance sampling." *Statistics and Computing*, 11: 125–139. MR1837132. doi: http://dx.doi.org/10.1023/A:1008923215028. 267

Newton, M. A. and Raftery, A. E. (1994). "Approximate Bayesian inference by the weighted likelihood bootstrap [with discussion]." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 56(1): 3–48. MR1257793. 262

Pajor, A. (2016). "Supplementary Material of "Estimating the Marginal Likelihood Using the Arithmetic Mean Identity"." *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/16-BA1001SUPP. 263, 265, 266

Pajor, A. and Osiewalski, J. (2013). "A Note on Lenk's Correction of the Harmonic Mean Estimator." *Central European Journal of Economic Modelling and Econometrics*, 5(4): 271–275. 263

Phillips, D. B. and Smith, A. F. M. (1996). "Bayesian model comparison via jump diffusions." In: Gilks, W. R., Spiegelhalter, D. J. and Richardson S. (eds.), *Markov Chain Monte Carlo in Practice*, 215–239, London, Chapman and Hall. MR1397970. 279

Postman, M., Huchra, J., and Geller, M. (1986). "Probes of Large-Scale Structure in the Corona Borealis Region." *The Astronomical Journal*, 92: 1238–1247. 278

Raftery, A. E. (1996). "Hypothesis testing and model selection." In: Gilks, W. R., Spiegelhalter, D. J. and Richardson S. (eds.), *Markov Chain Monte Carlo in Practice*, 163–188, London, Chapman and Hall. MR1397966. doi: http://dx.doi.org/10.1007/978-1-4899-4485-6. 264, 265, 267, 268, 279, 280

Raftery, A. E. and Banfield, J. D. (1991). "Stopping the Gibbs sampler, the use of morphology and other issues in spatial statistics." *Annals of the Institute of Statistical Mathematics*, 43: 32–43. 265

Raftery, A. E., Newton, M. A., Satagopan, J. M., and Krivitsky, P. N. (2007). "Estimating the Integral Likelihood via Posterior Simulation Using the Harmonic Mean Identity." In: Bernardo, M., Bayarri, M. J., Berger,J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West M. (eds.), *Bayesian Statistics*, 8, 1–45, Oxford University Press. MR2433201. 262, 267

Richardson, S., and Green, P. J. (1997). "On Bayesian Analysis of Mixtures with an Unknown Number of Components." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59: 731–758. MR1483213. doi: http://dx.doi.org/10.1111/1467-9868.00095. 279

Roeder, K. (1990). "Density Estimation With Confidence Sets Exemplified by Superclusters and Voids in the Galaxies." *Journal of the American Statistical Association*, 85: 617–624. 279

Steele, R. J., Raftery, A. E., and Emond, M. J. (2006). "Computing Normalizing Constants for Finite Mixture Models via Incremental Mixture Importance Sampling (IMIS)." *Journal of Computational and Graphical Statistics*, 15: 712–734. MR2291269. doi: http://dx.doi.org/10.1198/106186006X132358. 279

Stephens, M. (2000). "Dealing with label switching in mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62: 795–809. MR1796293. doi: http://dx.doi.org/10.1111/1467-9868.00265. 279

Weinberg, M.D. (2012). "Computing the Bayes Factor from a Markov Chain Monte Carlo Simulation of the Posterior Distribution." *Bayesian Analysis*, 7(3): 737–770. MR2981634. doi: http://dx.doi.org/10.1214/12-BA725. 264, 267

Williams, E. (1959). "Regression Analysis." New York: John Wiley and Sons, INC. MR0112212. 277

Xie, W., Lewis, O. P., Fan, Y., Kuo, L., and Chen, M-H. (2011). "Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection." *Systematic Biology*, 60(20): 150–160. 263, 267

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* New York: John Wiley and Sons, INC. MR0433791. 261

Acknowledgments

Publication was financed from the funds granted to the Faculty of Management at Cracow University of Economics, within the framework of the subsidy for the maintenance of research potential.