Approximating marginal likelihood using a Markov chain Monte Carlo simulation

Matija Piškorec
Ruđer Bošković institute
Zagreb, Croatia
matija.piskorec@irb.hr

January 11, 2017

Abstract

Marginal likelihood is a key quantity in Bayesian statistical modeling where it is used as a normalizing constant in a posterior distribution of model's parameters. More generally, it is used to evaluate model's uncertainty given data, a characteristic which is used for model selection - selecting best model out of several possible candidate models, and model averaging - aggregating predictions given by several different models. However, direct calculation of marginal likelihood involves marginalization over the whole parameter space, which can be prohibitively expensive to compute, so various approximations are used, many of which use Markov chain Monte Carlo (MCMC) simulations. MCMC simulations are widely used in Bayesian statistical modeling because they allow efficient estimation of model parameters and their confidences by sampling from a posterior distribution, even for high-dimensional models. They are also used in approximating the marginal likelihood, and this seminar reviews some of the commonly used methods.

1 Introduction

Why would we even be interested in computing the marginal likelihood[1]? In general, a goal of statistical inference is to find a posterior probability of parameters θ given a particular model $\mathcal{M}_k$ and data $\mathcal{D}$ [4, 5]. This is given by Bayes theorem:

$$p(\theta|\mathcal{D}, \mathcal{M}_k) = \frac{p(\mathcal{D}|\theta, \mathcal{M}_k)p(\theta|\mathcal{M}_k)}{p(\mathcal{D}|\mathcal{M}_k)} = \frac{p(\mathcal{D}|\theta, \mathcal{M}_k)p(\theta|\mathcal{M}_k)}{\int p(\mathcal{D}|\theta, \mathcal{M}_k)p(\theta|\mathcal{M}_k)d\theta} \tag{1}$$

Here, marginal likelihood $p(\mathcal{D}|\mathcal{M}_k)$ appears in denominator and serves as a normalizing constant. It can be expressed as a marginalization of likelihood over all possible parameter values. As such it is not needed for parameter inference because it does not depend on the parameters themselves. However, there are two important use cases for which marginal likelihood is crucial. First of these is *model selection* - choosing between two or more competing models by evaluating their fit to data [6, 7]. A quantity which measures this is a posterior probability of a model $\mathcal{M}_k$ given the data $\mathcal{D}$, which we can again express using the Bayes theorem:

$$p(\mathcal{M}_k|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_k)p(\mathcal{M}_k)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\mathcal{M}_k)p(\mathcal{M}_k)}{\sum_{\mathcal{M}_i \in \{\mathcal{M}_k\}} p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)} \tag{2}$$

Here, $p(\mathcal{D})$ is marginalization of data over all possible models $\{\mathcal{M}_k\}$[2], so we can simplify our posterior:

$$p(\mathcal{M}_k|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{M}_k)p(\mathcal{M}_k) \tag{3}$$

We can compare the fit of two models by the ratio of their posterior probabilities, which can be separated into ratios of prior probabilities and ratios of marginal likelihoods, also called *Bayes factor* [8]:

$$\underbrace{\frac{p(\mathcal{M}_i|\mathcal{D})}{p(\mathcal{M}_j|\mathcal{D})}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathcal{D}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_j)}}_{\text{Bayes factor}} \underbrace{\frac{p(\mathcal{M}_i)}{p(\mathcal{M}_j)}}_{\text{prior odds}} \tag{4}$$

[1] Other commonly used names are *integrated likelihood* [1], *model evidence* [2] and *normalizing constant* [3].

[2] This is a discrete marginalization because it involves a finite number of specific models. In contrasts, marginalization over the parameter space (equation 1) is a continuous marginalization.

Quantitative interpretation, in terms of the strength of evidence of model $\mathcal{M}_i$ over model $\mathcal{M}_j$, is usually the following [9, 8]: from 0 to 0.5 not worth a mention, from 0.5 to 1 substantial evidence, from 1 to 2 strong evidence and over 2 a decisive evidence in favor of model $\mathcal{M}_i$.

Second use case is *model averaging* - averaging predictions made by multiple models, weighted proportional to their posterior model probabilities, in that way incorporating model uncertainty into the prediction [10]. Posterior probability of future observations $\mathcal{D}^*$ given current observations $\mathcal{D}$ is the average over posterior predictive distributions given each model $p(\mathcal{D}^*|\mathcal{M}_k, \mathcal{D})$ weighted by their respective posterior model probabilities $p(\mathcal{M}_k|\mathcal{D})$:

$$p(\mathcal{D}^*|\mathcal{D}) = \sum_{k=1}^{l} p(\mathcal{D}^*|\mathcal{M}_k, \mathcal{D}) p(\mathcal{M}_k|\mathcal{D}) \tag{5}$$

Again, in order to obtain posterior model probability $p(\mathcal{M}_k|\mathcal{D})$ we have to calculate marginal likelihood $p(\mathcal{M}_k|\mathcal{D})$, as the two quantities are proportional (equation 3).

So we established that marginal likelihood is important and that it is worthwhile to compute it. Question remains - how actually to compute it? As I will show in the next section, direct calculation is usually infeasible but, luckily, there are many methods for approximating it sufficiently well for practical applications.

2 Approximations to marginal likelihood

Marginal likelihood $p(\mathcal{D}|\mathcal{M}_k)$ is computationally demanding to compute because it requires marginalization of data $\mathcal{D}$ over the whole parameter space θ:

$$p(\mathcal{D}|\mathcal{M}_k) = \int p(\mathcal{D}|\theta, \mathcal{M}_k) p(\theta|\mathcal{M}_k) d\theta \tag{6}$$

In some specific cases, such as when using *conjugate priors*, it is possible to express marginal likelihood in closed form. In other cases, it is possible to derive approximate analytical estimations or to use numerical integration which, however, becomes infeasible for high-dimensional models. However, it is possible to use Markov chain Monte Carlo (MCMC) methods to sample directly from the posterior distribution, and to use these samples for estimation of the parameters of a model [11, 12, 13]. These samples can also be used for estimation of marginal likelihood. Methods where marginal likelihood is estimated for each model separately are referred to as *within-model methods* [2]. I will review several of them in the next subsections - direct estimate via *mean* or a *harmonic mean* of the likelihood values, and Chib's method [14, 15]. Other within-model methods include annealed importance sampling [16] and power posterior method [17]. Another category of methods, which I will not cover in this seminar, are *between-model methods* where multiple models are combined into a single MCMC simulation. One example of these is reversible jump (trans-dimensional) MCMC [18].

2.1 Direct estimates of marginal likelihood

The most direct estimate of marginal likelihood using MCMC traces can be achieved in two ways [18]: (1) sampling from the posterior and taking the harmonic mean of the likelihood values, which is called a *harmonic mean estimator* [19]:

$$\hat{p}_1(\mathcal{D}|\mathcal{M}_k) = \left(\frac{1}{N} \sum_{t=1}^{N} \{ p(\mathcal{D}|\mathcal{M}_k, \theta_k^{(t)}) \}^{-1} \right)^{-1} \tag{7}$$

where $\theta_k^{(1)}, \theta_k^{(2)} \dots$ is a MCMC sample from the posterior $p(\theta_k|\mathcal{D}, \mathcal{M}_k)$, and (2) sampling from the prior and taking a mean of the likelihood values:

$$\hat{p}_2(\mathcal{D}|\mathcal{M}_k) = \frac{1}{N} \sum_{t=1}^{N} p(\mathcal{D}|\mathcal{M}_k, \theta_k^{(t)}) \tag{8}$$

where $p(\theta_k|\mathcal{M}_k)$ is a sample from the prior. These estimates are unbiased and simulation-consistent. For example, it is easy to see that the harmonic mean estimate is the expectation of the likelihood with respect to the posterior distribution [19, 20, 4]:

$$\begin{aligned} E\left[\frac{1}{p(\mathcal{D}|\theta)} |\mathcal{D} \right] &= \int \frac{1}{p(\mathcal{D}|\theta)} p(\theta|\mathcal{D}) d\theta = \int \frac{1}{p(\mathcal{D}|\theta)} \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} d\theta \\ &= \frac{1}{p(\mathcal{D})} \int p(\theta) d\theta = \frac{1}{p(\mathcal{D})} \approx \hat{p}_1(\mathcal{D}|\mathcal{M}_k) \end{aligned} \tag{9}$$

However, these estimates do not have finite variance in general and as such are often unstable, requiring large number of samples to converge to a meaningful value [1, 20]. This is because, in general, neither $\hat{p}_1(\mathcal{D}|\mathcal{M}_k)$ nor $\hat{p}_2(\mathcal{D}|\mathcal{M}_k)$ satisfy a Gaussian central limit theorem [19]. In the case of $\hat{p}_1(\mathcal{D}|\mathcal{M}_k)$, $p(\mathcal{D}|\theta)^{-1}$ is in general not square integrable with respect to the posterior distribution, and so there may be a value $\theta^{(r)}$ with very small likelihood which produces large effect on the final result. On the other hand, in the case of $\hat{p}_2(\mathcal{D}|\mathcal{M}_k)$, most of the $\theta^{(r)}$ will have small likelihoods and convergence will be quite slow, but, in the same time, dominated by few large values of likelihood, increasing the variance of the estimator. Another interpretation (see [4], p. 872, as well as [10, 21]) is that marginal likelihood is sensitive both to the prior and the posterior, while in estimates 7 and 8 we are sampling from just one of them.

One way of increasing the stability is to somehow combine the two estimates. For example, Newton and Raftery [19] propose simulating from a mixture of prior and posterior $\widetilde{p}(\theta_k;\mathcal{D},\mathcal{M}_k) = \delta p(\theta_k) + (1-\delta)p(\theta_k;\mathcal{D},\mathcal{M}_k)$ to stabilize the estimate:

$$\hat{p}_3(\mathcal{D}|\mathcal{M}_k) = \frac{\sum_{t=1}^{N} p(\mathcal{D}|\mathcal{M}_k, \theta_k^{(t)}) w(\theta_k^{(t)})}{\sum_{t=1}^{N} w(\theta_k^{(t)})} \tag{10}$$

where $w(\theta_k) = p(\theta_k|\mathcal{M}_k)/\widetilde{p}(\theta_k;\mathcal{D},\mathcal{M}_k)$. Equations 7, 8, and their combination 10 could all be derived from a more general *importance sampling* estimate of integrals of the form $I = \int g(\theta)p(\theta)d\theta$ [19]:

$$\hat{I} = \sum_{i=1}^{m} w_i g(\theta^{(i)}) / \sum_{i=1}^{m} w_i \tag{11}$$

where we take $g(\theta) = p(x|\theta)$. Also, weights $w_i = p(\theta^{(i)})/p^*(\theta^{(i)})$ with $p^*(\theta)$ being the *importance sampling function.* In the case of harmonic mean estimate (equation 7) the importance sampling function is the posterior distribution $p^*(\theta) = p(\theta|x)$ while in the case of normal estimate (equation 8) it is the prior distribution $p^*(\theta) = p(\theta)$.

2.2 Chib's method

Chib's method (also known as the *candidate method*) [14] follows from rearranging of the Bayes theorem for parameter inference (equation 1):

$$p(\mathcal{D}|\mathcal{M}_k) = \frac{p(\mathcal{D}|\theta,\mathcal{M}_k)p(\theta|\mathcal{M}_k)}{p(\theta|\mathcal{D},\mathcal{M}_k)} \tag{12}$$

For brevity, we will temporarily drop a conditioning on the model $\mathcal{M}_k$ from the equations. The logarithmic version of the equation above is:

$$\log(p(\mathcal{D})) = \log(p(\mathcal{D}|\theta^*)) + \log(p(\theta^*)) - \log(\hat{p}(\theta^*|\mathcal{D})) \tag{13}$$

Where $\hat{p}(\theta^*|\mathcal{D})$ is an estimate of posterior density at the parameter θ^*. In theory, the expression is valid for any value of θ^*, but in practice a parameter of high posterior probability (mode, mean or median) is used in order to maximize the accuracy of approximation [10, 6]. Usually, both likelihood $p(\mathcal{D}|\theta^*)$ and prior $p(\theta^*)$ terms can be evaluated in closed form. In contrast, posterior $\hat{p}(\theta^*|\mathcal{D})$ usually has to be estimated from a MCMC simulation. In the case of Gibbs sampler this can be done by first partitioning parameter space into blocks for which full conditional distributions are available in closed form. In the case of two blocks $\theta = (\theta_1, \theta_2)$ where full conditional probabilities are $p(\theta_1|\mathcal{D}, \theta_2)$ and $p(\theta_2|\mathcal{D}, \theta_1)$ we can estimate $\hat{p}(\theta_2|\mathcal{D})$ with ([6], page 175):

$$\hat{p}(\theta_2|\mathcal{D}) = \frac{1}{L}\sum_{j=1}^{L} p(\theta_2|\mathcal{D}, \theta_1^{(j)}) \tag{14}$$

Where $\theta_1^j; j = 1, \ldots, L$ is a set of posterior samples. The joint posterior $\hat{p}(\theta^*|\mathcal{D})$ can now be calculated as:

$$p(\theta|\mathcal{D}) = p(\theta_1|\mathcal{D}, \theta_2)p(\theta_2|\mathcal{D}) \tag{15}$$

And the estimator of marginal likelihood can now be written as:

$$\log(p(\mathcal{D})) \approx \log(p(\mathcal{D}|\theta_1^*, \theta_2^*)) + \log(p(\theta_1^*, \theta_2^*)) - \log(p(\theta_1^*|\mathcal{D}, \theta_2^*)) - \log(\hat{p}(\theta_2^*|\mathcal{D})) \tag{16}$$

This expression generalizes to cases with any number of blocks. In general, for B blocks we have:

$$\log(p(\mathcal{D})) \approx \log(p(\mathcal{D}|\theta^*) + \log(p(\theta^*)) - \sum_{k=1}^{B} \log(\hat{p}(\theta_k^*|\mathcal{D}, \theta_{k+1}^*, \ldots, \theta_B^*)) \tag{17}$$

Chib's method was originally developed only for outputs from Gibbs sampler, where conditional probabilities of each parameter considering all others are readily available, but was later also extended to accommodate outputs from a Metropolis-Hastings samplers [15]. An assumption of Chib's method which is often violated in practice is that all modes of a posterior should be sufficiently sampled [4].

3 Conclusion

In this seminar I reviewed several methods for approximation of marginal likelihood using MCMC simulations. Marginal likelihood is a crucial quantity in Bayesian statistical modeling needed to asses model uncertainty, which is used in model selection and as well as prediction through the use of model averaging. Using MCMC and similar sampling methods in this context is crucial for high-dimensional models for which direct marginalization of the whole parameter space needed for calculation of marginal likelihood is unfeasible. The most direct ways of approximating marginal likelihood from a MCMC simulation is through the mean and arithmetic mean of the likelihood values obtained from a simulation. These estimators, although simple and straightforward to implement, are nonetheless practically useless as they do not have finite variance, making estimates extremely unstable and requiring large number of samples for estimation. A slightly more advanced method, Chib's candidate method, does not suffer from these pitfalls, although it does depend on several assumptions which are usually hard to satisfy in practice. There are many more methods developed in recent years which were not covered in this seminar, or were only mentioned in passing.

Acknowledgment

This seminar has been written for the course: “Discrete Stochastic Processes” during the academic year 2016/2017 at the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia.

References

[1] Adrian E. Raftery, Michael A. Newton, Jaya M. Satagopan, and Pavel N. Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics*, 2007.

[2] X. Didelot, R.G. Everitt, A.M. Johansen, and D.J. Lawson. Likelihood-free estimation of model evidence. *Bayesian Analysis*, 6(1):49–76, 2011.

[3] A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 1998.

[4] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective.* The MIT Press, 2012.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Springer-Verlag, 2007.

[6] Ando T. *Bayesian Model Selection and Statistical Modeling.* Chapman and Hall/CRC, 2010.

[7] Burnham K. P. and D. R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach.* Springer-Verlag, 2002.

[8] Raftery A. E. Kass R. E. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[9] H. Jeffreys. *Theory of probability, 3rd edition.* Clarendon Press, Oxford, 1961.

[10] Nial Friel and Jason Wyse. Estimating the evidence – a review. *Statistica Neerlandica*, 66(3):288–308, 2012.

[11] Christian Robert and George Casella. *Monte Carlo Statistical Methods, 2nd edition.* Springer-Verlag, 2004.

[12] C. Andrieu, N. De Freitas, A. Doucet, and M.I. Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50(1-2):5–43, 2003.

[13] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, 3rd Edition.* Chapman and Hall/CRC, 2013.

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

[15] S. Chib and I. Jeliazkov. Marginal likelihood from the metropolis-hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.

[16] R.M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

[17] N. Friel and A.N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 70(3):589–607, 2008.

[18] P.J. Green. *Trans-dimensional Markov chain Monte Carlo.* Oxford University Press, 2003.

[19] M.A. Newton and Raftery A.E. Approximate bayesian inference with the weighted likelihood bootstrap. 1994.

[20] Radford Neal. The harmonic mean of the likelihood: Worst monte carlo method ever. Radford Neal's blog, 2008.

[21] C.P. Robert and D. Wraith. Computational methods for bayesian model choice. volume 1193, pages 251–262, 2009.