On the Use of Marginal Posteriors in Marginal Likelihood Estimation via Importance Sampling

Konstantinos Perrakis[1], Ioannis Ntzoufras[1], Efthymios G. Tsionas[2]

Second Bayesian Young Statisticians Meeting (BAYSM 2014)
Vienna, September 18–19, 2014

[1] Athens University of Economics and Business, Department of Statistics, Athens, Greece
<kperrakis,ntzoufras>@aueb.gr

[2]Athens University of Economics and Business, Department of Economics, Athens, Greece
tsionas@aueb.gr

Abstract

Marginal likelihood estimation based on importance sampling from the product of the marginal posterior distributions is presented. The approach is applicable to multi-block parameter vector settings, does not require further Markov Chain Monte Carlo (MCMC) sampling and is not dependent on the type of MCMC scheme used to sample from the posterior. The proposed estimator is applied to a normal regression problem and is compared to other common estimators.

Keywords: importance sampling; marginal likelihood estimation; posterior factorization; Rao-Blackwellization

1 Introduction

The marginal likelihood of a given model M_k with parameter vector $\boldsymbol{\theta}_k$ is the normalizing constant of the posterior $p(\boldsymbol{\theta}_k|\mathbf{y}, M_k)$, obtained by integrating the likelihood function $l(\mathbf{y}|\boldsymbol{\theta}_k, M_k)$ with respect to the prior $\pi(\boldsymbol{\theta}_k|M_k)$, i.e.

$$m(\mathbf{y}|M_k) = \int l(\mathbf{y}|\boldsymbol{\theta}_k, M_k)\pi(\boldsymbol{\theta}_k|M_k)\mathrm{d}\boldsymbol{\theta}_k. \tag{1}$$

Marginal likelihood estimation is essential to Bayesian model selection as it is associated with the evaluation of Bayes factors and posterior model odds [4]. A large body of literature is focused on "direct" estimation methods which utilize the posterior samples of separate models [1, 2, 3, 5, 7, 9]. In this paper we present an importance sampling approach which utilizes block factorizations of the posterior distribution [10].

2 The proposed estimator

Consider a two-block setting where $l(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi})$ is the likelihood of the data (dropping the depencency to M_k) conditional on parameter vectors $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)^T$ and $\boldsymbol{\phi} =$

$(\phi_1, \phi_2, ..., \phi_q)^T$. The idea is to use the product of the marginal posterior distributions as importance sampling density g, i.e. $g(\boldsymbol{\theta}, \boldsymbol{\phi}) \equiv p(\boldsymbol{\theta}|\mathbf{y})p(\boldsymbol{\phi}|\mathbf{y})$. Under this approach

$$m(\mathbf{y}) = \int\int \frac{l(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi})\pi(\boldsymbol{\theta}, \boldsymbol{\phi})}{p(\boldsymbol{\theta}|\mathbf{y})p(\boldsymbol{\phi}|\mathbf{y})} p(\boldsymbol{\theta}|\mathbf{y})p(\boldsymbol{\phi}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\phi}, \tag{2}$$

which can be estimated as

$$\widehat{m}(\mathbf{y}) = N^{-1}\sum_{n=1}^{N} \frac{l(\mathbf{y}|\boldsymbol{\theta}^{(n)}, \boldsymbol{\phi}^{(n)})\pi(\boldsymbol{\theta}^{(n)}, \boldsymbol{\phi}^{(n)})}{p(\boldsymbol{\theta}^{(n)}|\mathbf{y})p(\boldsymbol{\phi}^{(n)}|\mathbf{y})}. \tag{3}$$

Note that $\boldsymbol{\theta}^{(n)}, \boldsymbol{\phi}^{(n)}$, for $n = 1, 2, ..., N$, are draws from the marginal posterior distributions $p(\boldsymbol{\theta}|\mathbf{y})$ and $p(\boldsymbol{\phi}|\mathbf{y})$ and not from $p(\boldsymbol{\theta}, \boldsymbol{\phi}|y)$; nevertheless, forming a sample from the product marginal posterior can be easily implemented by systematically permuting the sampled values of $\boldsymbol{\theta}$ or of $\boldsymbol{\phi}$. Extending (3) to multi-block parameter settings is straightforward, while calculating the variance of the estimator can be handled through standard MCMC based methods.

The estimator in (3) is the optimal importance sampling density when $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are independent a-posteriori, since in this case $p(\boldsymbol{\theta}|y)p(\boldsymbol{\phi}|y) = p(\boldsymbol{\theta}, \boldsymbol{\phi}|y)$ leading to the zero-variance estimator. Although posterior independence is not frequently met in practice, the product marginal posterior can serve as a good approximation; first it has the same support as the joint posterior and second the blocking can be such that the blocks are close to orthogonal regardless whether the elements within $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are strongly correlated.

The marginal probabilities in the denominator of (3) can be calculated through moment-fitting approximations or kernel methods, while for cases of Gibbs sampling one can evaluate Rao-Blackwell estimates as

$$\widehat{p}(\boldsymbol{\theta}|\mathbf{y}) = L^{-1}\sum_{l=1}^{L} p(\boldsymbol{\theta}|\boldsymbol{\phi}^{(l)}, \mathbf{y}), \tag{4}$$

$$\widehat{p}(\boldsymbol{\phi}|\mathbf{y}) = L^{-1}\sum_{l=1}^{L} p(\boldsymbol{\phi}|\boldsymbol{\theta}^{(l)}, \mathbf{y}), \tag{5}$$

for a sufficiently large subsample of size $L < N$ from the joint posterior.

3 Example

The data concern 25 direct current (DC) electric charge measurements (volts) and wind velocity measurements (miles/hour) [6]. The models under consideration are;

i) M_0: the null model with the intercept,

ii) M_1: intercept+$(x_1 - \overline{x}_1)$,

iii) M_2: intercept+$(x_2 - \overline{x}_2)$ and

iv) M_3: intercept+$(x_1 - \overline{x}_1)$+x_1^2,

where x_1 is wind velocity and x_2 is the logarithm of wind velocity. Let j denote the model indicator, i.e. $j = 0, 1, 2, 3$. The likelihood and prior assumptions are the following

$$\begin{aligned} \mathbf{y}|\boldsymbol{\beta}_j, \sigma_j^2 &\sim \mathcal{N}(\mathbf{X}_j\boldsymbol{\beta}_j, \mathbf{I}_j\sigma_j^2) \\ \boldsymbol{\beta}_j|\sigma_j^2 &\sim \mathcal{N}(\mathbf{0}, \mathbf{V}_j\sigma_j^2) \\ \sigma_j^2 &\sim \mathcal{IG}(10^{-3}, 10^{-3}) \end{aligned}$$

where $\mathbf{y}$ is the vector of electric charge data, $\boldsymbol{\beta}_j$ and $\mathbf{X}_j$ correspond to the regression vector and design matrix of model j, respectively, and $\mathbf{V}_j = n^2(\mathbf{X}_j^T\mathbf{X}_j)^{-1}$ with $n = 25$. In relation to the context of Section 2 this is a 2-block setting where $\boldsymbol{\beta} \equiv \boldsymbol{\theta}$ and $\sigma^2 \equiv \boldsymbol{\phi}$. Under this conjugate design the marginal likelihoods can be calculated analytically.

Estimator		**Model**			
		M_0	M_1	M_2	M_3
Importance-weighted	$\log \widehat{m}(\mathbf{y})_{\text{Chen}}$	-34.8815	-13.1407	-1.5979	-2.2277
		(0.0029)	(0.0039)	(0.0031)	(0.0068)
Candidate's	$\log \widehat{m}(\mathbf{y})_{\text{Chib}}$	-34.8789	-13.1420	-1.5962	-2.2337
		(0.0020)	(0.0028)	(0.0023)	(0.0067)
Optimal bridge-sampling	$\log \widehat{m}(\mathbf{y})_{\text{obs}}$	-34.8807	-13.1412	-1.5979	-2.2294
		(0.0011)	(0.0019)	(0.0022)	(0.0030)
Proposed method					
Exact marginals	$\log \widehat{m}(\mathbf{y})_{\text{mp}}$	-34.8786	-13.1420	-1.5932	-2.2302
		(0.0023)	(0.0035)	(0.0030)	(0.0030)
Rao-Blackwellization	$\log \widehat{m}(\mathbf{y})_{\text{RB}}$	-34.8782	-13.1405	-1.5919	-2.2280
		(0.0023)	(0.0030)	(0.0030)	(0.0033)
Target value	$\log m(\mathbf{y})$	-34.8797	-13.1429	-1.5953	-2.2270

Table 1: Estimated marginal log-likelihood values compared with true values; average batch mean estimates (MC errors in parentheses) are presented using 30 batches of size 300.

Estimator (3) is calculated considering: i) the true marginals $p(\boldsymbol{\beta}_j|\mathbf{y})$, $p(\sigma_j^2|\mathbf{y})$ and ii) Rao-Blackwell estimates of $p(\boldsymbol{\beta}_j|\mathbf{y})$, $p(\sigma_j^2|\mathbf{y})$ from reduced samples of 200 draws. The two variants are denoted by $\widehat{m}(\mathbf{y})_{\text{mp}}$ and $\widehat{m}(\mathbf{y})_{\text{RB}}$, respectively. We also consider the following estimators: importance-weighted [1], candidate's [2] and optimal bridge-sampling [5]; the three additional estimators are denoted by $\widehat{m}(\mathbf{y})_{\text{Chen}}$, $\widehat{m}(\mathbf{y})_{\text{Chib}}$ and $\widehat{m}(\mathbf{y})_{\text{obs}}$, respectively. Results, based on 9000 posterior draws, are summarized in Table 1; as seen the proposed estimators yield comparable estimates to the other three methods with MC errors lower than those of $\widehat{m}(\mathbf{y})_{\text{Chen}}$ and just slightly higher than those of $\widehat{m}(\mathbf{y})_{\text{obs}}$.

4 Discussion

The performance of the proposed estimator depends on; i) the efficiency of approximating the joint posterior through block factorizations and ii) the accuracy in estimating marginal posterior densities. The first issue can be addressed with appropriate blocking or reparameterizations which will lead to near-orthogonal blocks. The second issue can be dealt with Rao-Blackwellization for Gibbs sampling settings and with moment-fitting, kernel methods or more elaborated strategies [8] for other MCMC schemes.

References

[1] Chen, M.-H. (2005). "Computing marginal likelihoods from a single MCMC output." *Statistica Neerlandica*, **59**, 16–29.

[2] Chib, S. (1995). "Marginal likelihood from the Gibbs output." *Journal of the American Statistical Association*, **90**, 1313–1321.

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

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

[5] Meng, X.-L., Wong, W.H. (1996). "Simulating ratios of normalizing constants via a simple identity: a theoretical exploration." *Statistica Sinica*, **6**, 831–860.

[6] Montgomery, D.C., Peck, E.A., Vining, G.G. (2001). *Introduction to Linear Regression Analysis*. John Wiley, New York.

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

[8] Rue, H., Martino, S., Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion)." *Journal of the Royal Statistical Society B*, **71**, 319–392.

[9] Skilling, J. (2006). "Nested sampling for general Bayesian computation." *Bayesian Analysis*, **1**, 833–860.

[10] Perrakis, K., Ntzoufras, I., Tsionas, E.G. (2014). "On the use of marginal posteriors in marginal likelihood estimation via importance sampling." *Computational Statistics and Data Analysis*, **77**, 54–69.