

Importance sampling methods for Bayesian discrimination between embedded models

J.-M. MARIN^{1,3,*} and C.P. ROBERT^{2,3,†}

¹Institut de Mathématiques et Modélisation de Montpellier,

²Université Paris Dauphine, and

³Centre de Recherche en Economie et Statistique, INSEE, Paris

October 13, 2009

Abstract

This paper surveys some well-established approaches on the approximation of Bayes factors used in Bayesian model choice, mostly as covered in Chen et al. (2000). Our focus here is on methods that are based on importance sampling strategies—rather than variable dimension techniques like reversible jump MCMC—, including: crude Monte Carlo, maximum likelihood based importance sampling, bridge and harmonic mean sampling, as well as Chib’s method based on the exploitation of a functional equality. We demonstrate in this survey how these different methods can be efficiently implemented for testing the significance of a predictive variable in a probit model. Finally, we compare their performances on a real dataset.

Keywords: Bayesian inference; model choice; Bayes factor; Monte Carlo; Importance Sampling; bridge sampling; Chib’s functional identity; supervised learning; probit model

1 Introduction

The contribution of Jim Berger to the better understanding of Bayesian testing is fundamental and wide-ranging, from establishing the fundamental difficulties with p -values in Berger and Sellke (1987) to formalising the intrinsic Bayes factors in Berger and Pericchi (1996), to solving the difficulty with improper priors in Berger et al. (1998), and beyond! While our contribution in this area is obviously much more limited, we

*Institut de Mathématiques et Modélisation de Montpellier, Université Montpellier 2, Case Courrier 51, 34095 Montpellier cedex 5, France. Email: Jean-Michel.Marin@univ-montp2.fr

†CEREMADE - Université Paris Dauphine, 75775 Paris, and CREST, INSEE, Paris, France. Email: xian@ceremade.dauphine.fr

aim at presenting here the most standard approaches to the approximation of Bayes factors.

The Bayes factor indeed is a fundamental procedure that stands at the core of the Bayesian theory of testing hypotheses, at least in the approach advocated by both Jeffreys (1939) and by Jaynes (2003). (Note that Robert et al. 2009, provides a reassessment of the crucial role of Jeffreys 1939 in setting a formal framework for Bayesian testing as well as for regular inference.) Given an hypothesis $H_0 : \theta \in \Theta_0$ on the parameter $\theta \in \Theta$ of a statistical model, with observation y and density $f(y|\theta)$, under a compatible prior of the form

$$\pi(\Theta_0)\pi_0(\theta) + \pi(\Theta_0^c)\pi_1(\theta),$$

the *Bayes factor* is defined as the posterior odds to prior odds ratio, namely

$$B_{01}(y) = \frac{\pi(\Theta_0|y)}{\pi(\Theta_0^c|y)} \bigg/ \frac{\pi(\Theta_0)}{\pi(\Theta_0^c)} = \int_{\Theta_0} f(y|\theta)\pi_0(\theta)d\theta \bigg/ \int_{\Theta_0^c} f(y|\theta)\pi_1(\theta)d\theta.$$

Model choice can be considered from a similar perspective, since, under the Bayesian paradigm (see, e.g., Robert 2001), the comparison of models

$$\mathfrak{M}_i : y \sim f_i(y|\theta_i), \quad \theta_i \sim \pi_i(\theta_i), \quad \theta_i \in \Theta_i, \quad i \in \mathfrak{I},$$

where the family $\mathfrak{I}$ can be finite or infinite, leads to posterior probabilities of the models under comparison such that

$$\mathbb{P}(\mathfrak{M} = \mathfrak{M}_i|y) \propto p_i \int_{\Theta_i} f_i(y|\theta_i)\pi_i(\theta_i)d\theta_i,$$

where $p_i = \mathbb{P}(\mathfrak{M} = \mathfrak{M}_i)$ is the prior probability of model $\mathfrak{M}_i$.

In this short survey, we consider some of the most common Monte Carlo solutions used to approximate a generic Bayes factor or its fundamental component, the *evidence*

$$m_i = \int_{\Theta_i} \pi_i(\theta_i) f_i(y|\theta_i) d\theta_i,$$

aka the marginal likelihood. Longer entries can be found in Carlin and Chib (1995), Chen et al. (2000), Robert and Casella (2004), or Friel and Pettitt (2008). Note that we only briefly mention here trans-dimensional methods issued from the revolutionary paper of Green (1995), since our goal is to demonstrate that within-model simulation methods allow for the computation of Bayes factors and thus avoids the additional complexity involved in trans-dimensional methods. While amenable to an importance sampling technique of sorts, the alternative approach of nested sampling (Skilling 2006) is discussed in Chopin and Robert (2007) and Robert and Wraith (2009).

2 The Pima Indian benchmark model

In order to compare the performances of all methods presented in this survey, we chose to evaluate the corresponding estimates of the Bayes factor in the setting of a single variable selection for a probit model and to repeat the estimation in a Monte Carlo experiment to empirically assess the variability of those estimates.

We recall that a probit model can be represented as a natural latent variable model in that, if we consider a sample $z_1, \dots, z_n$ of n independent latent variables associated with a standard regression model, i.e. such that $z_i | \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{x}_i^T \boldsymbol{\theta}, 1)$, where the $\mathbf{x}_i$'s are p -dimensional covariates and $\boldsymbol{\theta}$ is the vector of regression coefficients, then $y_1, \dots, y_n$ such that

$$y_i = \mathbb{I}_{z_i > 0}$$

is a probit sample. Indeed, given $\boldsymbol{\theta}$, the y_i 's are independent Bernoulli rv's with $\mathbb{P}(y_i = 1 | \boldsymbol{\theta}) = \Phi(\mathbf{x}_i^T \boldsymbol{\theta})$ where Φ is the standard normal cdf.

The choice of a reference prior distribution for the probit model is open to debate, but the connection with the latent regression model induced Marin and Robert (2007) to suggest a g -prior model, $\boldsymbol{\theta} \sim \mathcal{N}(0_p, n(\mathbf{X}^T \mathbf{X})^{-1})$, with n as the g factor and $\mathbf{X}$ as the regressor matrix. The corresponding posterior distribution is then associated with the density

$$\pi(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X}) \propto \prod_{i=1}^n \{1 - \Phi(\mathbf{x}_i^T \boldsymbol{\theta})\}^{1-y_i} \Phi(\mathbf{x}_i^T \boldsymbol{\theta})^{y_i} \times \exp \{ -\boldsymbol{\theta}^T (\mathbf{X}^T \mathbf{X}) \boldsymbol{\theta} / 2n \}, \quad (1)$$

where $\mathbf{y} = (y_1, \dots, y_n)$. In the completed model, i.e. when including the latent variables $\mathbf{z} = (z_1, \dots, z_n)$ into the model, the y_i 's are deterministic functions of the z_i 's and the so-called completed likelihood is

$$f(\mathbf{y}, \mathbf{z} | \boldsymbol{\theta}) = (2\pi)^{-n/2} \exp \left(- \sum_{i=1}^n (z_i - \mathbf{x}_i^T \boldsymbol{\theta})^2 / 2 \right) \prod_{i=1}^n (\mathbb{I}_{y_i=0} \mathbb{I}_{z_i \leq 0} + \mathbb{I}_{y_i=1} \mathbb{I}_{z_i > 0}).$$

The derived conditional distributions

$$z_i | y_i, \boldsymbol{\theta} \sim \begin{cases} \mathcal{N}_+(\mathbf{x}_i^T \boldsymbol{\theta}, 1, 0) & \text{if } y_i = 1, \\ \mathcal{N}_-(\mathbf{x}_i^T \boldsymbol{\theta}, 1, 0) & \text{if } y_i = 0, \end{cases} \quad (2)$$

are of interest for constructing a Gibbs sampler on the completed model, where $\mathcal{N}_+(\mathbf{x}_i^T \boldsymbol{\theta}, 1, 0)$ denotes the Gaussian distribution with mean $\mathbf{x}_i^T \boldsymbol{\theta}$ and variance 1 that is left-truncated at 0, while $\mathcal{N}_-(\mathbf{x}_i^T \boldsymbol{\theta}, 1, 0)$ denotes the symmetrical normal distribution that is right-truncated at 0. The corresponding full conditional on the parameters is given by

$$\boldsymbol{\theta} | \mathbf{y}, \mathbf{z} \sim \mathcal{N} \left(\frac{n}{n+1} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{z}, \frac{n}{n+1} (\mathbf{X}^T \mathbf{X})^{-1} \right). \quad (3)$$

Indeed, since direct simulation from the posterior distribution of $\boldsymbol{\theta}$ is intractable, Albert and Chib (1993) suggest implementing a Gibbs sampler based on the above set of full

conditionals. More precisely, given the current value of θ , one cycle of the Gibbs algorithm produces a new value for $\mathbf{z}$ as simulated from the conditional distribution (2), which, when substituted into (3), produces a new value for θ . Although it does not impact the long-term properties of the sampler, the starting value of θ may be taken as the maximum likelihood estimate to avoid burning steps in the Gibbs sampler.

Given this probit model, the dataset we consider covers a population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona. These women were tested for diabetes according to World Health Organization (WHO) criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases, and is available with the basic R package (R Development Core Team 2008). This dataset, used as a benchmark for supervised learning methods, contains information about 332 women with the following variables:

- **glu**: plasma glucose concentration in an oral glucose tolerance test;
- **bp**: diastolic blood pressure (mm Hg);
- **ped**: diabetes pedigree function;
- **type**: Yes or No, for diabetic according to WHO criteria.

For this dataset, the goal is to explain the diabetes variable **type** by using the explanatory variables **glu**, **bp** and **ped**. The following table is an illustration of a classical (maximum likelihood) analysis of this dataset, obtained using the R `glm()` function with the probit link:

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-2.1347	-0.9217	-0.6963	0.9959	2.3235

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
glu	0.012616	0.002406	5.244	1.57e-07	***
bp	-0.029050	0.004094	-7.096	1.28e-12	***
ped	0.350301	0.208806	1.678	0.0934	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 460.25 on 332 degrees of freedom
Residual deviance: 386.73 on 329 degrees of freedom
AIC: 392.73
Number of Fisher Scoring iterations: 4

This analysis sheds some doubt on the relevance of the covariate **ped** in the model and we can reproduce the study from a Bayesian perspective, computing the Bayes

factor B_{01} opposing the probit model only based on the covariates **glu** and **bp** (model 0) to the probit model based on the covariates **glu**, **bp**, and **ped** (model 1). This is equivalent to testing the hypothesis $H_0 : \theta_3 = 0$ since the models are nested, where θ_3 is the parameter of the probit model associated with covariate **ped**. (Note that there is no intercept in either model.) If we denote by $\mathbf{X}_0$ the 332×2 matrix containing the values of **glu** and **bp** for the 332 individuals and by $\mathbf{X}_1$ the 332×3 matrix containing the values of the covariates **glu**, **bp**, and **ped**, the Bayes factor B_{01} is given by

$$\begin{aligned}
& (2\pi)^{1/2} n^{1/2} \frac{|\mathbf{X}_0^T \mathbf{X}_0|^{-1/2}}{|\mathbf{X}_1^T \mathbf{X}_1|^{-1/2}} \\
& \frac{\int_{\mathbb{R}^2} \prod_{i=1}^n \{1 - \Phi((\mathbf{X}_0)_{i,\cdot}, \boldsymbol{\theta})\}^{1-y_i} \Phi((\mathbf{X}_0)_{i,\cdot}, \boldsymbol{\theta})^{y_i} \exp\{-\boldsymbol{\theta}^T (\mathbf{X}_0^T \mathbf{X}_0) \boldsymbol{\theta} / 2n\} d\boldsymbol{\theta}}{\int_{\mathbb{R}^3} \prod_{i=1}^n \{1 - \Phi((\mathbf{X}_1)_{i,\cdot}, \boldsymbol{\theta})\}^{1-y_i} \Phi((\mathbf{X}_1)_{i,\cdot}, \boldsymbol{\theta})^{y_i} \exp\{-\boldsymbol{\theta}^T (\mathbf{X}_1^T \mathbf{X}_1) \boldsymbol{\theta} / 2n\} d\boldsymbol{\theta}} \\
& = \frac{\mathbb{E}_{\mathcal{N}_2(0_2, n(\mathbf{X}_0^T \mathbf{X}_0)^{-1})} \left[\prod_{i=1}^n \{1 - \Phi((\mathbf{X}_0)_{i,\cdot}, \boldsymbol{\theta})\}^{1-y_i} \Phi((\mathbf{X}_0)_{i,\cdot}, \boldsymbol{\theta})^{y_i} \right]}{\mathbb{E}_{\mathcal{N}_3(0_3, n(\mathbf{X}_1^T \mathbf{X}_1)^{-1})} \left[\prod_{i=1}^n \{1 - \Phi((\mathbf{X}_1)_{i,\cdot}, \boldsymbol{\theta})\}^{1-y_i} \Phi((\mathbf{X}_1)_{i,\cdot}, \boldsymbol{\theta})^{y_i} \right]}
\end{aligned} \tag{4}$$

using the shortcut notation that $A_{i,\cdot}$ is the i -th line of the matrix A .

3 The basic Monte Carlo solution

As already shown above, when testing for a null hypothesis (or a model) $H_0 : \theta \in \Theta_0$ against the alternative hypothesis (or the alternative model) $H_1 : \theta \in \Theta_1$, the Bayes factor is defined by

$$B_{01}(y) = \int_{\Theta_0} f(y|\theta_0) \pi_0(\theta_0) d\theta_0 \Big/ \int_{\Theta_1} f(y|\theta_1) \pi_1(\theta_1) d\theta_1.$$

We assume in this survey that the prior distributions under both the null and the alternative hypotheses are proper, as, typically, they should be. (In the case of common nuisance parameters, a common improper prior measure can be used on those, see Berger et al. (1998), Marin and Robert (2007). This obviously complicates the computational aspect, as some methods like crude Monte Carlo cannot be used at all, while others are more prone to suffer from infinite variance.) In that setting, the most elementary approximation to $B_{01}(y)$ consists in using a ratio of two standard Monte Carlo approximations based on simulations from the corresponding priors. Indeed, for $i = 0, 1$:

$$\int_{\Theta_i} f(y|\theta) \pi_i(\theta) d\theta = \mathbb{E}_{\pi_i} [f(y|\theta)] .$$

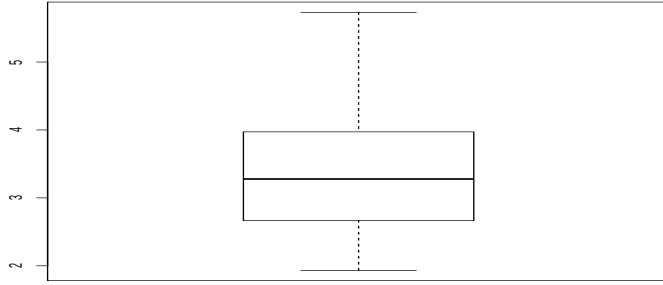


Figure 1: Pima Indian dataset: boxplot of 100 Monte Carlo estimates of $B_{01}(y)$ based on simulations from the prior distributions, for 2×10^4 simulations.

If $\theta_{0,1}, \dots, \theta_{0,n_0}$ and $\theta_{1,1}, \dots, \theta_{1,n_1}$ are two independent samples generated from the prior distributions π_0 and π_1 , respectively, then

$$\frac{n_0^{-1} \sum_{j=1}^{n_0} f(y|\theta_{0,j})}{n_1^{-1} \sum_{j=1}^{n_1} f(y|\theta_{1,j})} \quad (5)$$

is a strongly consistent estimator of $B_{01}(y)$.

In most cases, sampling from the prior distribution corresponding to either hypothesis is straightforward and fast. Therefore, the above estimator is extremely easy to derive as a brute-force evaluation of the Bayes factor. However, if any of the posterior distributions is quite different from the corresponding prior distribution—and it should be for vague priors—the Monte Carlo evaluation of the corresponding evidence is highly inefficient since the sample will be overwhelmingly producing negligible values of $f(y|\theta_{i,j})$. In addition, if $f^2(y|\theta)$ is not integrable against π_0 or π_1 , the resulting estimation has an infinite variance. Since importance sampling usually requires an equivalent computation effort, with a potentially high efficiency reward, crude Monte Carlo approaches of this type are usually disregarded.

Figure 1 and Table 1 summarize the results based on 100 replications of Monte Carlo approximations of $B_{01}(y)$, using equation (5) with $n_0 = n_1 = 20,000$ simulations. As predicted, the variability of the estimator is very high, when compared with the other estimates studied in this survey. (Obviously, the method is asymptotically unbiased and, the functions being square integrable in (4), with a finite variance. A massive simulation effort would obviously lead to a precise estimate of the Bayes factor.)

4 Usual importance sampling approximations

Defining two importance distributions with densities ϖ_0 and ϖ_1 , with the same supports as π_0 and π_1 , respectively, we have:

$$B_{01}(y) = \mathbb{E}_{\varpi_0} [f(y|\theta)\pi_0(\theta)/\varpi_0(\theta)] \Big/ \mathbb{E}_{\varpi_1} [f(y|\theta)\pi_1(\theta)/\varpi_1(\theta)] .$$

Therefore, given two independent samples generated from distributions ϖ_0 and ϖ_1 , $\theta_{0,1}, \dots, \theta_{0,n_0}$ and $\theta_{1,1}, \dots, \theta_{1,n_1}$, respectively, the corresponding importance sampling estimate of $B_{01}(y)$ is

$$\frac{n_0^{-1} \sum_{j=1}^{n_0} f_0(x|\theta_{0,j})\pi_0(\theta_{0,j})/\varpi_0(\theta_{0,j})}{n_1^{-1} \sum_{j=1}^{n_1} f_1(x|\theta_{1,j})\pi_1(\theta_{1,j})/\varpi_1(\theta_{1,j})} . \quad (6)$$

Compared with the standard Monte Carlo approximation above, this approach offers the advantage of opening the choice of the representation (6) in that it is possible to pick importance distributions ϖ_0 and ϖ_1 that lead to a significant reduction in the variance of the importance sampling estimate. This implies choosing importance functions that provide as good as possible approximations to the corresponding posterior distributions. Maximum likelihood asymptotic distributions or kernel approximations based on a sample generated from the posterior are natural candidates in this setting, even though the approximation grows harder as the dimension increases.

For the Pima Indian benchmark, we propose for instance to use as importance distributions, Gaussian distributions with means equal to the maximum likelihood (ML) estimates and covariance matrices equal to the estimated covariance matrices of the ML estimates, both of which are provided by the R `glm()` function. While, in general, those Gaussian distributions provide crude approximations to the posterior distributions, the specific case of the probit model will show this is an exceptionally good approximation to the posterior, since this leads to the best solution among all those compared here. The results, obtained over 100 replications of the methodology with $n_0 = n_1 = 20,000$ are summarized in Figure 2 and Table 1. They are clearly excellent, while requiring the same computing time as the original simulation from the prior.

5 Bridge sampling methodology

The original version of the bridge sampling approximation to the Bayes factor (Gelman and Meng 1998, Chen et al. 2000) relies on the assumption that the parameters of both models under comparison belong to the same space: $\Theta_0 = \Theta_1$. In that case, for likelihood functions f_0 and f_1 under respectively models $\mathfrak{M}_0$ and $\mathfrak{M}_1$, the bridge representation of the Bayes factor is

$$B_{01}(y) = \int_{\Theta_0} f_0(y|\theta)\pi_0(\theta)d\theta \Big/ \int_{\Theta_1} f_1(y|\theta)\pi_1(\theta)d\theta = \mathbb{E}_{\pi_1} \left[\frac{f_0(y|\theta)\pi_0(\theta)}{f_1(y|\theta)\pi_1(\theta)} \Big| y \right] . \quad (7)$$

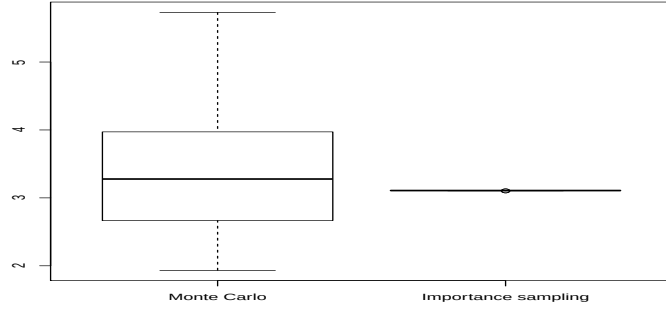


Figure 2: Pima Indian dataset: boxplots of 100 Monte Carlo and importance sampling estimates of $B_{01}(y)$, based on simulations from the prior distributions, for 2×10^4 simulations.

Given a sample from the posterior distribution of θ under model $\mathfrak{M}_1$, $\theta_{1,1}, \dots, \theta_{1,N} \sim \pi_1(\theta|y)$, a first bridge sampling approximation to $B_{01}(y)$ is

$$N^{-1} \sum_{j=1}^N \frac{f_0(y|\theta_{1,j})\pi_0(\theta_{1,j})}{f_1(y|\theta_{1,j})\pi_1(\theta_{1,j})}.$$

From a practical perspective, for the above bridge sampling approximation to be of any use, the constraint on the common parameter space for both models goes further in that, not only must both models have the same complexity, but they must also be parameterised on a common ground, i.e. in terms of some specific moments of the sampling model, so that parameters under both models have a common meaning. Otherwise, the resulting bridge sampling estimator will have very poor convergence properties, possibly with infinite variance.

Equation (7) is nothing but a very special case of the general representation (Torrìe and Valleau 1977)

$$B_{01}(y) = \mathbb{E}_{\varphi} [f_0(y|\theta)\pi_0(\theta)/\varphi(\theta)] \Big/ \mathbb{E}_{\varphi} [f_1(y|\theta)\pi_1(\theta)/\varphi(\theta)],$$

which holds for any density φ with a sufficiently large support and which only requires a single sample $\theta_1, \dots, \theta_N$ generated from φ to produce an importance sampling estimate of the ratio of the marginal likelihoods. Apart from using the *same* importance function φ for both integrals, this method is therefore a special case of importance sampling.

Another extension of this bridge sampling approach is based on the general repre-

sensation

$$\begin{aligned}
B_{01}(y) &= \int f_0(y|\theta)\pi_0(\theta)\alpha(\theta)\pi_1(\theta|y)d\theta \Big/ \int f_1(y|\theta)\pi_1(\theta)\alpha(\theta)\pi_0(\theta|y)d\theta \\
&\approx \frac{n_1^{-1} \sum_{j=1}^{n_1} f_0(y|\theta_{1,j})\pi_0(\theta_{1,j})\alpha(\theta_{1,j})}{n_0^{-1} \sum_{j=1}^{n_0} f_1(y|\theta_{0,j})\pi_1(\theta_{0,j})\alpha(\theta_{0,j})}
\end{aligned}$$

where $\theta_{0,1}, \dots, \theta_{0,n_0}$ and $\theta_{1,1}, \dots, \theta_{1,n_1}$ are two independent samples coming from the posterior distributions $\pi_0(\theta|y)$ and $\pi_1(\theta|y)$, respectively. That applies for any positive function α as long as the upper integral exists. Some choices of α lead to very poor performances of the method in connection with the harmonic mean approach (see Section 6), but there exists a quasi-optimal solution, as provided by Gelman and Meng (1998):

$$\alpha^*(y) \propto \frac{1}{n_0\pi_0(\theta|y) + n_1\pi_1(\theta|y)}.$$

This optimum cannot be used *per se*, since it requires the normalising constants of both $\pi_0(\theta|y)$ and $\pi_1(\theta|y)$. As suggested by Gelman and Meng (1998), an approximate version uses iterative versions of α^* , based on iterated approximations to the Bayes factor. Note that this solution recycles simulations from both posteriors, which is quite appropriate since one model is selected via the Bayes factor, instead of using an importance weighted sample common to both approximations. We will see below an alternative representation of the bridge factor that bypasses this difficulty (if difficulty there is!).

Those derivations are, however, restricted to the case where both models have the same complexity and thus they do not apply to embedded models, when $\Theta_0 \subset \Theta_1$ in such a way that $\theta_1 = (\theta, \psi)$, i.e., when the submodel corresponds to a specific value ψ_0 of ψ : $f_0(y|\theta) = f(y|\theta, \psi_0)$.

The extension of the most advanced bridge sampling strategies to such cases requires the introduction of a *pseudo-posterior density*, $\omega(\psi|\theta, y)$, on the parameter that does not appear in the embedded model, in order to reconstitute the equivalence between both parameter spaces. Indeed, if we augment $\pi_0(\theta|y)$ with $\omega(\psi|\theta, y)$, we obtain a joint distribution with density $\pi_0(\theta|y) \times \omega(\psi|\theta, y)$ on Θ_1 . The Bayes factor can then be expressed as

$$B_{01}(y) = \frac{\int_{\Theta_1} f(y|\theta, \psi_0)\pi_0(\theta)\alpha(\theta, \psi)\pi_1(\theta, \psi|y)d\theta\omega(\psi|\theta, y) d\psi}{\int_{\Theta_1} f(y|\theta, \psi)\pi_1(\theta, \psi)\alpha(\theta, \psi)\pi_0(\theta|y) \times \omega(\psi|\theta, y)d\theta d\psi}, \quad (8)$$

for all functions $\alpha(\theta, \psi)$, because it is clearly independent from the choice of both $\alpha(\theta, \psi)$

and $\omega(\psi|\theta, y)$. Obviously, the performances of the approximation

$$\frac{(n_1)^{-1} \sum_{j=1}^{n_1} f(y|\theta_{1,j}, \psi_0) \pi_0(\theta_{1,j}) \omega(\psi_{1,j}|\theta_{1,j}, y) \alpha(\theta_{1,j}, \psi_{1,j})}{(n_0)^{-1} \sum_{j=1}^{n_0} f(y|\theta_{0,j}, \psi_{0,j}) \pi_1(\theta_{0,j}, \psi_{0,j}) \alpha(\theta_{0,j}, \psi_{0,j})},$$

where $(\theta_{0,1}, \psi_{0,1}), \dots, (\theta_{0,n_0}, \psi_{0,n_0})$ and $(\theta_{1,1}, \psi_{1,1}), \dots, (\theta_{1,n_1}, \psi_{1,n_1})$ are two independent samples generated from distributions $\pi_0(\theta|y) \times \omega(\psi|\theta, y)$ and $\pi_1(\theta, \psi|y)$, respectively, do depend on this completion by the pseudo-posterior as well as on the function $\alpha(\theta, \psi)$. Chen et al. (2000) establish that the asymptotically optimal choice for $\omega(\psi|\theta, y)$ is the obvious one, namely

$$\omega(\psi|\theta, y) = \pi_1(\psi|\theta, y),$$

which most often is unavailable in closed form (especially when considering that the normalising constant of $\omega(\psi|\theta, y)$ is required in (8)). However, in latent variable models, approximations of the conditional posteriors often are available, as detailed in Section 7.

While this extension of the basic bridge sampling approximation is paramount for handling embedded models, its implementation suffers from the dependence on this pseudo-posterior. In addition, this technical device brings the extended bridge methodology close to the cross-model alternatives of Carlin and Chib (1995) and Green (1995), in that both those approaches rely on completing distributions, either locally (Green 1995) or globally (Carlin and Chib 1995), to link both models under comparison in a bijective relation. The density $\omega(\psi|\theta_0, y)$ is then a pseudo-posterior distribution in Chib and Carlin's (1995) sense, and it can be used as Green's (1995) proposal in the reversible jump MCMC step to move (or not) from model $\mathfrak{M}_0$ to model $\mathfrak{M}_1$. While using cross-model solutions to compare only two models does seem superfluous, given that the randomness in picking the model at each step of the simulation is not as useful as in the setting of comparing a large number or an infinity of models, the average acceptance probability for moving from model $\mathfrak{M}_0$ to model $\mathfrak{M}_1$ is related to the Bayes factor since

$$\mathbb{E}_{\pi_0 \times \omega} \left[\frac{f(y|\theta, \psi) \pi_1(\theta, \psi)}{f(y|\theta, \psi_0) \pi_0(\theta) \omega(\psi|\theta, y)} \right] = B_{01}(y)$$

even though the average

$$\mathbb{E}_{\pi_0 \times \omega} \left[\min \left\{ 1, \frac{f(y|\theta, \psi) \pi_1(\theta, \psi)}{f(y|\theta, \psi_0) \pi_0(\theta) \omega(\psi|\theta, y)} \right\} \right]$$

does not provide a closed form solution.

For the Pima Indian benchmark, we use as pseudo-posterior density $\omega(\theta_3|\theta_1, \theta_2, y)$, the conditional Gaussian density deduced from the asymptotic Gaussian distribution on $(\theta_1, \theta_2, \theta_3)$ already used in the importance sampling solution, with mean equal to the

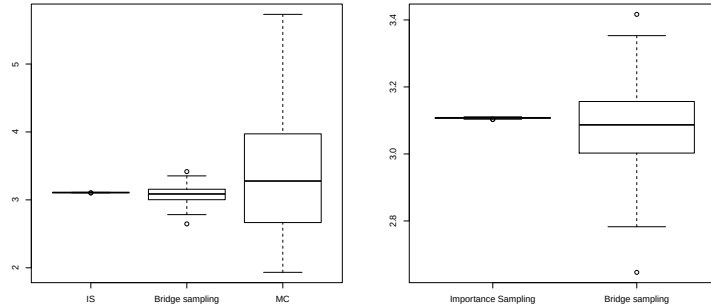


Figure 3: Pima Indian dataset: (*left*) boxplots of 100 importance sampling, bridge sampling and Monte Carlo estimates of $B_{01}(y)$, based on simulations from the prior distributions, for 2×10^4 simulations; (*right*) same comparison for the importance sampling versus bridge sampling estimates only.

ML estimate of $(\theta_1, \theta_2, \theta_3)$ and with covariance matrix equal to the estimated covariance matrix of the ML estimate. The quasi-optimal solution α^* in the bridge sampling estimate is replaced with the inverse of an average between the asymptotic Gaussian distribution in model $\mathfrak{M}_1$ and the product of the asymptotic Gaussian distribution in model $\mathfrak{M}_0$ times the above $\omega(\theta_3|\theta_1, \theta_2, y)$. This obviously is a suboptimal choice, but it offers the advantage of providing a non-iterative solution. The results, obtained over 100 replications of the methodology with $n_0 = n_1 = 20,000$ are summarized in Figure 3 and Table 1. The left-hand graph shows that this choice of bridge sampling estimator produces a solution whose variation is quite close to the (excellent) importance sampling solution, a considerable improvement upon the initial Monte Carlo estimator. However, the right-hand-side graph shows that the importance sampling solution remains far superior, especially when accounting for the computing time. (In this example, running 20,000 iterations of the Gibbs sampler for the models with both two and three variables takes approximately 32 seconds.)

6 Harmonic mean approximations

While using the generic harmonic mean approximation to the marginal likelihood is often fraught with danger (Neal 1994), the representation (Gelfand and Dey 1994) ($k = 0, 1$)

$$\mathbb{E}_{\pi_k} \left[\frac{\varphi_k(\theta)}{\pi_k(\theta)f_k(y|\theta)} \middle| y \right] = \int \frac{\varphi_k(\theta)}{\pi_k(\theta)f_k(y|\theta)} \frac{\pi_k(\theta)f_k(y|\theta)}{m_k(y)} d\theta = \frac{1}{m_k(y)} \quad (9)$$

holds, no matter what the density $\varphi_k(\theta)$ is—provided $\varphi_k(\theta) = 0$ when $\pi_k(\theta)f_k(y|\theta) = 0$ —. This representation is remarkable in that it allows for a direct processing of Monte Carlo or MCMC output from the posterior distribution $\pi_k(\theta|y)$. As with importance

sampling approximations, the variability of the corresponding estimator of $B_{01}(y)$ will be small if the distributions $\varphi_k(\theta)$ ($k = 0, 1$) are close to the corresponding posterior distributions. However, as opposed to usual importance sampling constraints, the density $\varphi_k(\theta)$ must have lighter—rather than fatter—tails than $\pi_k(\theta)f_k(y|\theta)$ for the approximation of the marginal $m_k(y)$

$$1 \Big/ N^{-1} \sum_{j=1}^N \frac{\varphi_k(\theta_{k,j})}{\pi_k(\theta_{k,j})f_k(y|\theta_{k,j})}$$

to enjoy finite variance. For instance, using $\varphi_k(\theta) = \pi_k(\theta)$ as in the original harmonic mean approximation (Newton and Raftery 1994) will most usually result in an infinite variance estimator, as discussed by Neal (1994). On the opposite, using φ_k 's with constrained supports derived from a Monte Carlo sample, like the convex hull of the simulations corresponding to the 10% or to the 25% HPD regions—that again is easily derived from the simulations—is both completely appropriate and implementable (Robert and Wraith 2009).

However, for the Pima Indian benchmark, we propose to use instead as our distributions $\varphi_k(\theta)$ the very same distributions as those used in the above importance sampling approximations, that is, Gaussian distributions with means equal to the ML estimates and covariance matrices equal to the estimated covariance matrices of the ML estimates. The results, obtained over 100 replications of the methodology with $N = 20,000$ simulations for each approximation of $m_k(y)$ ($k = 0, 1$) are summarized in Figure 4 and Table 1. They show a very clear proximity between both importance solutions in this special case and a corresponding domination of the bridge sampling estimator, even though the importance sampling estimate is much faster to compute. This remark must be toned down by considering that the computing time due to the Gibbs sampler should not necessarily be taken into account into the comparison, since samples are generated under both models.

7 Exploiting functional equalities

Chib's (1995) method for approximating a marginal (likelihood) is a direct application of Bayes' theorem: given $y \sim f_k(y|\theta)$ and $\theta \sim \pi_k(\theta)$, we have that

$$m_k = \frac{f_k(y|\theta) \pi_k(\theta)}{\pi_k(\theta|y)},$$

for all θ 's (since both the lhs and the rhs of this equation are constant in θ). Therefore, if an arbitrary value of θ , say θ_k^* , is selected and if a good approximation to $\pi_k(\theta|y)$ can be constructed, denoted $\hat{\pi}(\theta|y)$, Chib's (1995) approximation to the evidence is

$$m_k = \frac{f_k(y|\theta_k^*) \pi_k(\theta_k^*)}{\hat{\pi}_k(\theta_k^*|y)}. \quad (10)$$

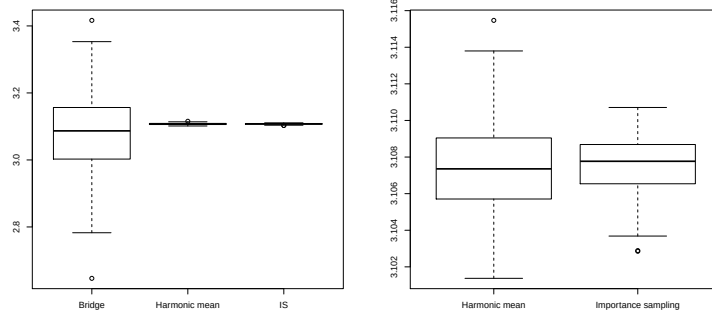


Figure 4: Pima Indian dataset: (*left*) boxplots of 100 bridge sampling, harmonic mean and importance sampling estimates of $B_{01}(y)$, based on simulations from the prior distributions, for 2×10^4 simulations; (*right*) same comparison for the harmonic mean versus importance sampling estimates only.

In a general setting, $\hat{\pi}(\theta|y)$ may be the Gaussian approximation based on the MLE, already used in the importance sampling, bridge sampling and harmonic mean solutions, but this is unlikely to be accurate in a general framework. A second solution is to use a nonparametric approximation based on a preliminary MCMC sample, even though the accuracy may also suffer in large dimensions. In the special setting of latent variables models (like mixtures of distributions but also like probit models), Chib's (1995) approximation is particularly attractive as there exists a natural approximation to $\pi_k(\theta|y)$, based on the Rao–Blackwell (Gelfand and Smith 1990) estimate

$$\hat{\pi}_k(\theta_k^*|y) = \frac{1}{T} \sum_{t=1}^T \pi_k(\theta_k^*|y, z_k^{(t)}),$$

where the $z_k^{(t)}$'s are the latent variables simulated by the MCMC sampler. The estimate $\hat{\pi}_k(\theta_k^*|y)$ is a parametric unbiased approximation of $\pi_k(\theta_k^*|y)$ that converges with rate $O(\sqrt{T})$. This Rao–Blackwell approximation obviously requires the full conditional density $\pi_k(\theta_k^*|y, z)$ to be available in closed form (constant included) but, as already explained, this is the case for the probit model.

Figure 5 and Table 1 summarize the results obtained for 100 replications of Chib's approximations of $B_{01}(y)$ with $T = 20,000$ simulations for each approximation of $m_k(y)$ ($k = 0, 1$). While Chib's method is usually very reliable and dominates importance sampling, the incredibly good approximation provided by the asymptotic Gaussian distribution implies that, in this particular case, Chib's method is dominated by both the importance sampling and the harmonic mean estimates.

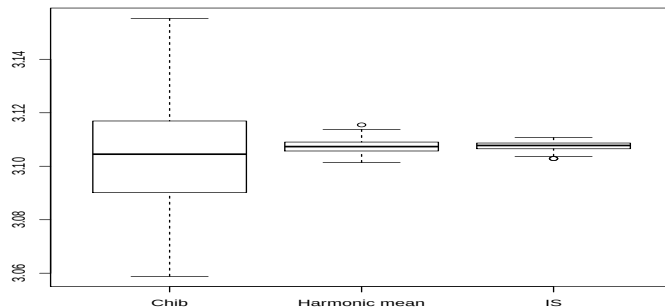


Figure 5: Pima Indian dataset: boxplots of 100 Chib’s, harmonic mean and importance estimates of $B_{01}(y)$, based on simulations from the prior distributions, for 2×10^4 simulations.

Table 1: Pima Indian dataset: Performances of the various approximation methods used in this survey.

	Monte Carlo	Importance sampling	Bridge sampling	Harmonic mean	Chib’s approximation
Median	3.277	3.108	3.087	3.107	3.104
Standard deviation	0.7987	0.0017	0.1357	0.0025	0.0195
Duration in seconds	7	7	71	70	64

8 Conclusion

In this short evaluation of the most common estimations to the Bayes factor, we have found that a particular importance sampling and its symmetric harmonic mean counterpart are both very efficient in the case of the probit model. The bridge sampling estimate is much less efficient in this example, due to the approximation error resulting from the pseudo-posterior. In most settings, the bridge sampling is actually doing better than the equivalent importance sampler (Robert and Wraith 2009), while Chib’s method is much more generic than the four alternatives. The recommendation resulting from the short experiment above is therefore to look for handy approximations to the posterior distribution, whenever available, but to fall back on Chib’s method as a backup solution providing a reference or better.

Acknowledgements

J.-M. Marin and C.P. Robert are supported by the 2009–2012 grant ANR-09-BLAN-0218 “Big’MC”.

References

- ALBERT, J. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. American Statist. Assoc.*, **88** 669–679.
- BERGER, J. and PERICCHI, L. (1996). The intrinsic Bayes factor for model selection and prediction. *J. American Statist. Assoc.*, **91** 109–122.
- BERGER, J., PERICCHI, L. and VARSHAVSKY, J. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhya A*, **60** 307–321.
- BERGER, J. and SELLKE, T. (1987). Testing a point-null hypothesis: the irreconcilability of significance levels and evidence (with discussion). *J. American Statist. Assoc.*, **82** 112–122.
- CARLIN, B. and CHIB, S. (1995). Bayesian model choice through Markov chain Monte Carlo. *J. Royal Statist. Society Series B*, **57** 473–484.
- CHEN, M., SHAO, Q. and IBRAHIM, J. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.
- CHIB, S. (1995). Marginal likelihood from the Gibbs output. *J. American Statist. Assoc.*, **90** 1313–1321.
- CHOPIN, N. and ROBERT, C. (2007). Contemplating evidence: properties, extensions of, and alternatives to nested sampling. Tech. Rep. 2007-46, CEREMADE, Université Paris Dauphine. ArXiv:0801.3887.
- FRIEL, N. and PETTITT, A. (2008). Marginal likelihood estimation via power posteriors. *J. Royal Statist. Society Series B*, **70(3)** 589–607.
- GELFAND, A. and DEY, D. (1994). Bayesian model choice: asymptotics and exact calculations. *J. Royal Statist. Society Series B*, **56** 501–514.
- GELFAND, A. and SMITH, A. (1990). Sampling based approaches to calculating marginal densities. *J. American Statist. Assoc.*, **85** 398–409.
- GELMAN, A. and MENG, X. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Science*, **13** 163–185.
- GREEN, P. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, **82** 711–732.
- JAYNES, E. (2003). *Probability Theory*. Cambridge University Press, Cambridge.
- JEFFREYS, H. (1939). *Theory of Probability*. 1st ed. The Clarendon Press, Oxford.
- MARIN, J.-M. and ROBERT, C. (2007). *Bayesian Core*. Springer-Verlag, New York.

- NEAL, R. (1994). Contribution to the discussion of “approximate bayesian inference with the weighted likelihood bootstrap” by michael a. newton and adrian e. raftery. *J. Royal Statist. Society Series B*, **56** (1) 41–42.
- NEWTON, M. and RAFTERY, A. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *J. Royal Statist. Society Series B*, **56** 1–48.
- R DEVELOPMENT CORE TEAM (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- ROBERT, C. (2001). *The Bayesian Choice*. 2nd ed. Springer-Verlag, New York.
- ROBERT, C. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*. 2nd ed. Springer-Verlag, New York.
- ROBERT, C., CHOPIN, N. and ROUSSEAU, J. (2009). Theory of Probability revisited (with discussion). *Statist. Science*. (to appear).
- ROBERT, C. and WRAITH, D. (2009). Computational methods for Bayesian model choice. In *MaxEnt 2009 proceedings* (A. I. of Physics, ed.).
- SKILLING, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis*, **1**(4) 833–860.
- TORRIE, G. and VALLEAU, J. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comp. Phys.*, **23** 187–199.