Advanced Monte Carlo methods for Bayesian model choice

Jean-Michel Marin

Université de Montpellier
Institut Montpelliérain Alexander Grothendieck

Plan

Recap on Bayesian model choice
The basic Monte Carlo solution
Usual importance sampling approximations
The harmonic mean approximation
Exploiting functional equalities
The bridge sampling methodology to compared embedded models
Monte Carlo Markov Chain algorithms to explore the space of models

Recap on Bayesian model choice

J models in competition

A model is characterized by a likelihood function $f_k(\mathbf{y}|\boldsymbol{\theta}_k)$ and a prior distribution on the parameter $\boldsymbol{\theta}_k \in \Theta_k$.

Prior probabilities in the model space are defined.

The posterior distribution in the model space is such that

$$\mathbb{P}^{\pi}\left(\mathcal{M} = k|\mathbf{y}\right) \propto \mathbb{P}(\mathcal{M} = k) \int_{\Theta_k} f_k(\mathbf{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k)\,\mathbf{d}\boldsymbol{\theta}_k\,.$$

Some computational difficulties:

- How to approximate the evidences?
- When the number of models in consideration is huge, how to explore the models's space?

The basic Monte Carlo solution

The generic approach for approximating the target integral

$$\mathfrak{I}(\mathbf{y}) = \int_{\Theta} h(\boldsymbol{\theta}, \mathbf{y}) \pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

where $\pi(\cdot)$ is a pdf, is to produce via a computer program a sample from the distribution associated to $\pi(\cdot)$.

A formal Monte Carlo algorithm associated with the target $\mathfrak{I}$ goes as follows:

Basic Monte Carlo Algorithm

1) Set $i = 1$,

2) Generate $\boldsymbol{\theta}^{(i)}$ from the distribution associated to $\pi(\cdot)$,

3) Set $i = i + 1$,

4) If $i \leq N$, return to **2)**.

The corresponding crude Monte Carlo approximation of $\mathfrak{I}$ is given by:

$$\widehat{\mathfrak{I}}^{\mathrm{MC}}(\mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} h\left(\boldsymbol{\theta}^{(i)}, \mathbf{y}\right) .$$

When the computing effort N grows to infinity, the approximation $\widehat{\mathfrak{I}}^{\mathrm{MC}}(\mathbf{y})$ converges to $\mathfrak{I}(\mathbf{y})$ and the speed of convergence is $1/\sqrt{N}$ if h is square-integrable against $\pi(\cdot)$.

The standard MC approximation to $B_{01}(\mathbf{y})$ consists in using a ratio of two standard Monte Carlo approximations based on simulations from the corresponding priors.

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

$$\frac{n_0^{-1} \sum_{j=1}^{n_0} f_0(\mathbf{y}|\theta_{0,j})}{n_1^{-1} \sum_{j=1}^{n_1} f_1(\mathbf{y}|\theta_{1,j})} \tag{1}$$

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

Usual importance sampling approximations

A generalisation of the basic Monte Carlo algorithm stems from an alternative representation of the above $\mathfrak{I}(\mathbf{y})$, changing both the integrating density and the integrand:

$$\mathfrak{I}(\mathbf{y}) = \int_{\Theta} \frac{h(\boldsymbol{\theta}, \mathbf{y})\pi(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} g(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{2}$$

where the support of $\pi(\cdot)$ is included in the one of $g(\cdot)$.

Importance Sampling Scheme

1) Set $i = 1$,

2) Generate $\boldsymbol{\theta}^{(i)}$ from the importance distribution $g(\cdot)$,

3) Calculate the importance weight
$\omega^{(i)} = \pi\left(\boldsymbol{\theta}^{(i)}\right) / g\left(\boldsymbol{\theta}^{(i)}\right)$,

4) Set $i = i + 1$,

5) If $i \leq N$, return to **2)**.

The corresponding importance approximation of $\mathfrak{I}(\mathbf{y})$ is given by

$$\widehat{\mathfrak{I}}_g^{\mathrm{IS}}(\mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \omega^{(i)} h\left(\boldsymbol{\theta}^{(i)}, \mathbf{y}\right) .$$

Obviously, $g(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$ often is a possible choice for $g(\boldsymbol{\theta})$.

More surprisingly, the choice $\pi(\boldsymbol{\theta})$ is generally far from being the most efficient choice.

The only requirement is that the support of $\pi(\cdot)$ should be included in the one of $g(\cdot)$),

Poor choices of $g(\cdot)$ lead to unreliable approximations: for instance, if $\int_\Theta h^2(\boldsymbol{\theta}, \mathbf{y})\omega^2(\boldsymbol{\theta})g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ is infinite, the variance of the estimator is also infinite.

Let $g_i(\cdot)$ ($i \in \{0,1\}$) be importance functions which are strictly positive when $f_i(\cdot|\mathbf{y})\pi_i(\cdot)$ are stricly positive.

$$B_{01} = \frac{\displaystyle\int_{\Theta_0} f_0(\mathbf{y}|\boldsymbol{\theta}_0)\pi_0(\boldsymbol{\theta}_0)\mathrm{d}\boldsymbol{\theta}_0}{\displaystyle\int_{\Theta_1} f_1(\mathbf{y}|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)\mathrm{d}\boldsymbol{\theta}_1} = \frac{\mathbb{E}_{\pi_0}\left[f_0(\mathbf{y}|\boldsymbol{\theta}_0)\right]}{\mathbb{E}_{\pi_1}\left[f_1(\mathbf{y}|\boldsymbol{\theta}_1)\right]}$$

$$= \frac{\displaystyle\int_{\Theta_0} \frac{f_0(\mathbf{y}|\boldsymbol{\theta}_0)\pi_0(\boldsymbol{\theta}_0)}{g_0(\boldsymbol{\theta}_1)} g_0(\boldsymbol{\theta}_0)\mathrm{d}\boldsymbol{\theta}_0}{\displaystyle\int_{\Theta_1} \frac{f_1(\mathbf{y}|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)}{g_1(\boldsymbol{\theta}_1)} g_1(\boldsymbol{\theta}_1)\mathrm{d}\boldsymbol{\theta}_1} = \frac{\mathbb{E}_{g_0}\left[\dfrac{f_0(\mathbf{y}|\boldsymbol{\theta}_0)\pi_0(\boldsymbol{\theta}_0)}{g_0(\boldsymbol{\theta}_0)}\right]}{\mathbb{E}_{g_1}\left[\dfrac{f_1(\mathbf{y}|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)}{g_1(\boldsymbol{\theta}_1)}\right]}.$$

The regular importance approximation of B_{01} is given by

$$\widehat{B}_{01} = \frac{n_0^{-1} \sum_{i=1}^{n_0} f_0(\mathbf{y}|\boldsymbol{\theta}_0^i)\pi_0(\boldsymbol{\theta}_0^i)/g_0(\boldsymbol{\theta}_0^i)}{n_1^{-1} \sum_{i=1}^{n_1} f_1(\mathbf{y}|\boldsymbol{\theta}_1^i)\pi_1(\boldsymbol{\theta}_1^i)/g_1(\boldsymbol{\theta}_1^i)}$$

where $\boldsymbol{\theta}_0^1, \ldots, \boldsymbol{\theta}_0^{n_0}$ is an n_0-sample from $g_0(\cdot)$ and $\boldsymbol{\theta}_1^1, \ldots, \boldsymbol{\theta}_1^{n_1}$ is an n_1-sample from $g_1(\cdot)$.

Diabetes in Pima Indian women benchmark example

"A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix (AZ), was tested for diabetes according to WHO criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases."

332 Pima Indian women with observed variables

- plasma glucose concentration (x_1),
- diastolic blood pressure (x_2),
- diabetes pedigree function (x_3),
- presence/absence of diabetes (y).

Probit modelling on Pima Indian women

We suppose that

$$\mathbb{P}(y = 1|\mathbf{x}) = \Phi(x_1\theta_1 + x_2\theta_2 + x_3\theta_3)\,.$$

The goal is to test the hypothesis $H_0 : \theta_3 = 0$.

We denote by $\mathbf{X}_0$ the 332×2 matrix containing the values of x_1 and x_2 for the 332 individuals and by $\mathbf{X}_1$ the 332×3 matrix containing the values of the covariates x_1, x_2 and x_3.
Under H_0 (for model $\mathfrak{M}_0$), we use the following prior modelling

$$\boldsymbol{\theta}_0 = (\theta_{1,0}, \theta_{2,0})|\mathbf{X}_0 \sim \mathcal{N}_2\left(0_2, n(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0)^{-1}\right)\,.$$

Under H_1 (for model $\mathfrak{M}_1$), we use

$$\boldsymbol{\theta}_1 = (\theta_{1,1}, \theta_{2,1}, \theta_{3,1})|\mathbf{X}_1 \sim \mathcal{N}_3\left(0_3, n(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\right)\,.$$

The Bayes factor B_{01} is equal to

$$\frac{\mathbb{E}_{\mathcal{N}_2(0_2,n(\mathbf{x}_0^{\mathrm{T}}\mathbf{x}_0)^{-1})}\left[\prod_{i=1}^{n}\{1-\Phi\left((\mathbf{X}_0)_{i,.}\boldsymbol{\theta}\right)\}^{1-y_i}\Phi\left((\mathbf{X}_0)_{i,.}\boldsymbol{\theta}\right)^{y_i}\right]}{\mathbb{E}_{\mathcal{N}_3(0_3,n(\mathbf{x}_1^{\mathrm{T}}\mathbf{x}_1)^{-1})}\left[\prod_{i=1}^{n}\{1-\Phi\left((\mathbf{X}_1)_{i,.}\boldsymbol{\theta}\right)\}^{1-y_i}\Phi\left((\mathbf{X}_1)_{i,.}\boldsymbol{\theta}\right)^{y_i}\right]}$$

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

Importance sampling for the Pima Indian dataset

Use of the importance function inspired from the MLE estimate distributions:

gaussian distributions with means equal to the Maximum Likelihood (ML) estimates $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}_1$ and covariance matrices equal to the estimated covariance matrices of the ML estimates $\hat{\Sigma}_0$ and $\hat{\Sigma}_1$:

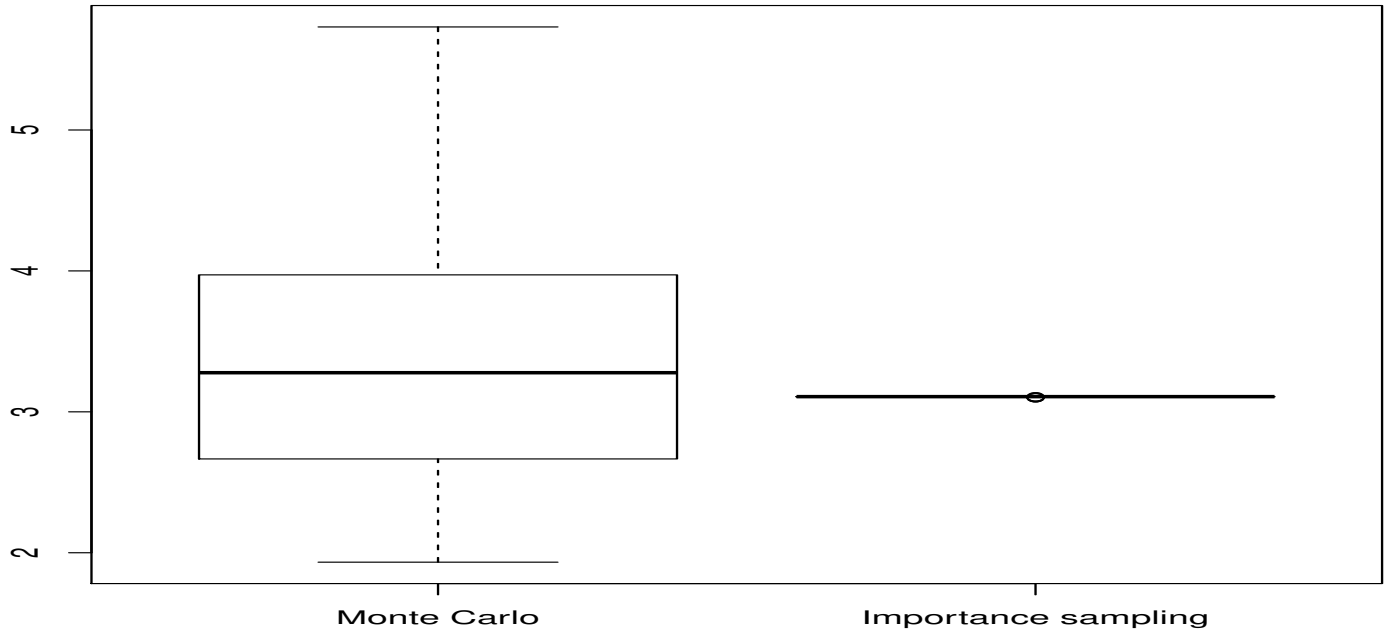
$$g_0(\cdot) \sim \mathcal{N}_2(\hat{\boldsymbol{\theta}}_0, \hat{\Sigma}_0)\,,$$

and

$$g_1(\cdot) \sim \mathcal{N}_3(\hat{\boldsymbol{\theta}}_1, \hat{\Sigma}_1)\,.$$

Diabetes in Pima Indian women

Comparison of the variation of the Bayes factor approximations based on 100 replicas for $20,000$ simulations from the prior and the above MLE importance sampler



The harmonic mean approximation

$$\mathbb{E}_{\pi_k}\left[\left.\frac{\varphi_k(\boldsymbol{\theta})}{\pi_k(\boldsymbol{\theta})f_k(\mathbf{y}|\boldsymbol{\theta})}\right|\mathbf{y}\right] = \int \frac{\varphi_k(\boldsymbol{\theta})}{\pi_k(\boldsymbol{\theta})f_k(\mathbf{y}|\boldsymbol{\theta})}\,\frac{\pi_k(\boldsymbol{\theta})f_k(\mathbf{y}|\boldsymbol{\theta})}{m_k(\mathbf{y})}\,\mathrm{d}\boldsymbol{\theta} = \frac{1}{m_k(\mathbf{y})}$$

holds, no matter what the density $\varphi_k(\boldsymbol{\theta})$ is, provided $\varphi_k(\boldsymbol{\theta}) = 0$ when $\pi_k(\boldsymbol{\theta})f_k(\mathbf{y}|\boldsymbol{\theta}) = 0$.

As opposed to usual importance sampling constraints, the density $\varphi_k(\boldsymbol{\theta})$ must have lighter—rather than fatter—tails than $\pi_k(\boldsymbol{\theta})f_k(\mathbf{y}|\boldsymbol{\theta})$ for the approximation of the Bayes factor

$$1 \Big/ N^{-1} \sum_{i=1}^{N} \frac{\varphi_k(\boldsymbol{\theta}_k^i)}{\pi_k(\boldsymbol{\theta}_k^i) f_k(\mathbf{y}|\boldsymbol{\theta}_k^i)}$$

to enjoy finite variance.

Using $\varphi_k(\boldsymbol{\theta}) = \pi_k(\boldsymbol{\theta})$ as in the original harmonic mean approximation will most usually result in an infinite variance estimator.

"The Worst Monte Carlo Method Ever"

Radford Neal's blog, Aug. 23, 2008

"The good news is that the Law of Large Numbers guarantees that this estimator is consistent ie, it will very likely be very close to the correct answer if you use a sufficiently large number of points from the posterior distribution.

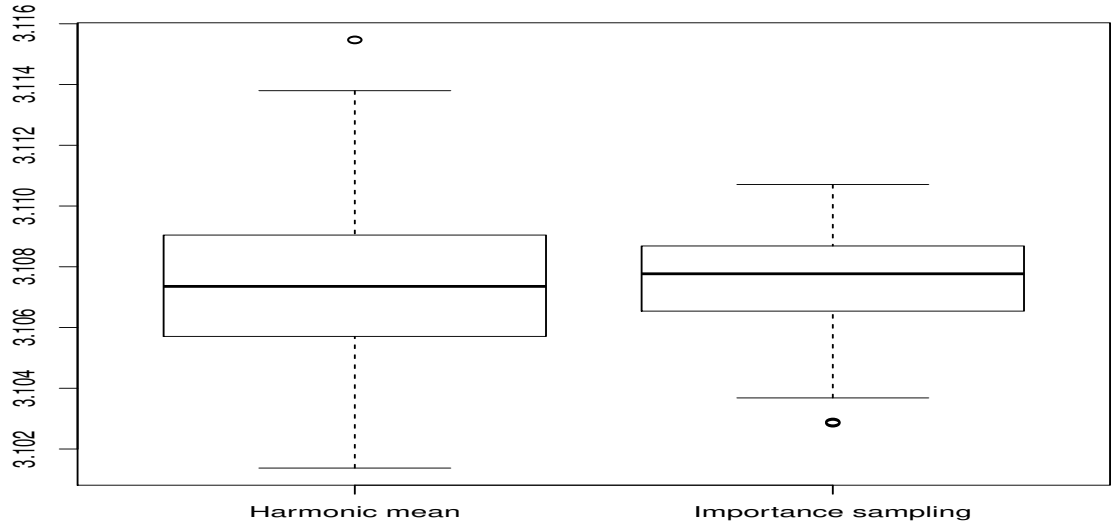
The bad news is that the number of points required for this estimator to get close to the right answer will often be greater than the number of atoms in the observable universe. The even worse news is that it's easy for people to not realize this, and to naïvely accept estimates that are nowhere close to the correct value of the marginal likelihood."

The Pima Indian dataset

For the Pima Indian benchmark, we propose to use instead as our distributions $\varphi_k(\boldsymbol{\theta})$ 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.

Diabetes in Pima Indian women

Comparison of the variation of the Bayes factor approximations based on 100 replicas for $20,000$ simulations



Exploiting functional equalities

Chib (1995) solution

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

for all $\boldsymbol{\theta}$.

Therefore, if an arbitrary value of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^*$, is selected, the Chib's approximation to the evidence is

$$\hat{m}_k(\mathbf{y}) = \frac{f_k(\mathbf{y}|\boldsymbol{\theta}^*)\,\pi_k(\boldsymbol{\theta}^*)}{\hat{\pi}_k(\boldsymbol{\theta}^*|\mathbf{y})}\,.$$

$\hat{\pi}_k(\boldsymbol{\theta}|\mathbf{y})$ may be the Gaussian approximation based on the MLE.

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.

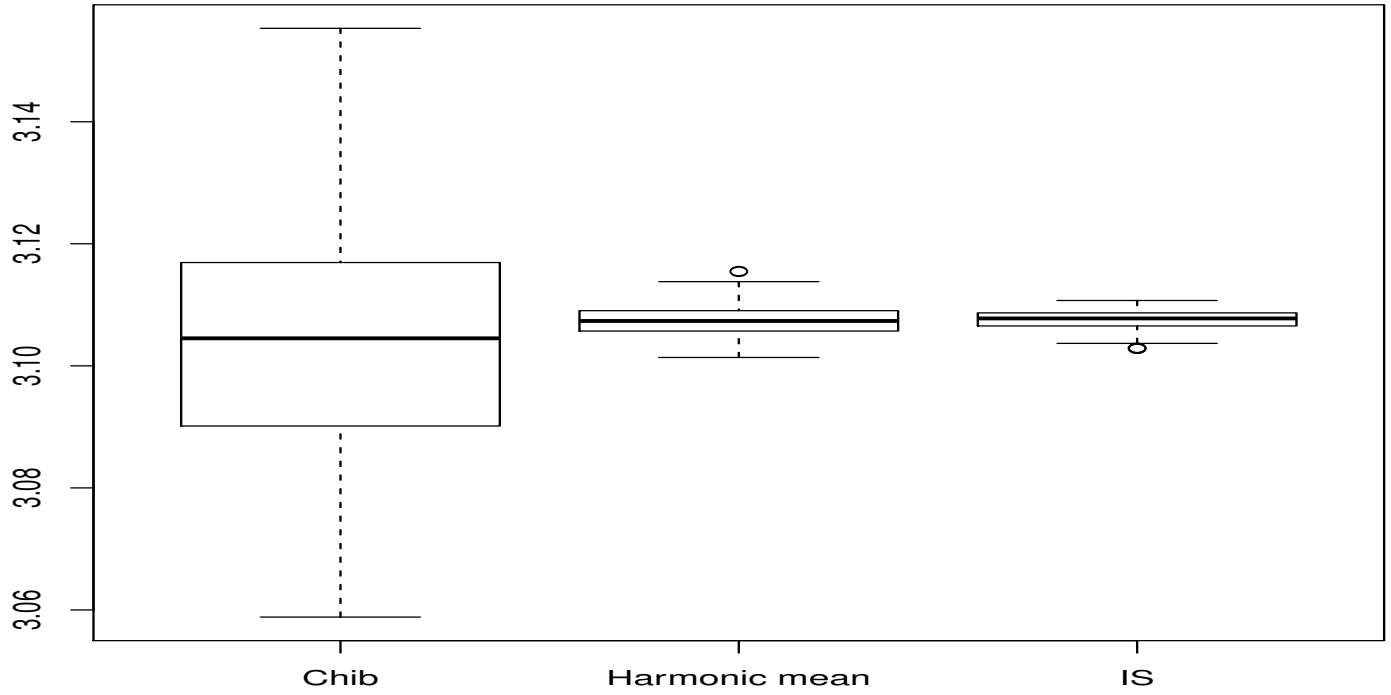
In the special setting of latent variables models, Chib's approximation is particularly attractive as there exists a natural approximation to $\pi_k(\boldsymbol{\theta}^*|\mathbf{y})$, based on the Rao-Blackwell estimate

$$\hat{\pi}_k(\boldsymbol{\theta}^*|\mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T} \pi_k(\boldsymbol{\theta}^*|\mathbf{y}, \mathbf{z}^{(t)})\,,$$

where the $\mathbf{z}^{(t)}$'s are the latent variables simulated by the MCMC sampler.

Diabetes in Pima Indian women (cont'd)

Comparison of the variation of the Bayes factor approximations based on 100 replicas for $20,000$ simulations



The bridge sampling methodology to compared embedded models

When

$$\begin{aligned}\pi_0(\boldsymbol{\theta}_0|\mathbf{y}) &\propto \tilde{\pi}_0(\boldsymbol{\theta}_0|\mathbf{y})\\ \pi_1(\boldsymbol{\theta}_1|\mathbf{y}) &\propto \tilde{\pi}_1(\boldsymbol{\theta}_1|\mathbf{y})\end{aligned}$$

live on the same space ($\Theta_0 = \Theta_1 = \Theta$), then

$$B_{01} = \int_\Theta f_0(\mathbf{y}|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \Big/ \int_\Theta f_1(\mathbf{y}|\boldsymbol{\theta})\pi_1(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$$

$$= \mathbb{E}_{\pi_1}\left[\frac{f_0(\mathbf{y}|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})}{f_1(\mathbf{y}|\boldsymbol{\theta})\pi_1(\boldsymbol{\theta})}\middle|\mathbf{y}\right] .$$

In that case, the bridge sampling approximation of B_{01} is given by

$$\hat{B}_{01} = N^{-1} \sum_{j=1}^{N} \frac{f_0(\mathbf{y}|\boldsymbol{\theta}_j)\pi_0(\boldsymbol{\theta}_j)}{f_1(\mathbf{y}|\boldsymbol{\theta}_j)\pi_1(\boldsymbol{\theta}_j)} = N^{-1} \sum_{j=1}^{N} \frac{\tilde{\pi}_0(\boldsymbol{\theta}_j|\mathbf{y})}{\tilde{\pi}_1(\boldsymbol{\theta}_j|\mathbf{y})}$$

where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ is an N-sample from $\pi_1(\cdot|\mathbf{y})$.

For all $\alpha(\cdot)$, if $\Theta_0 = \Theta_1 = \Theta$, we have

$$B_{01} = \frac{\displaystyle\int_\Theta \tilde{\pi}_0(\boldsymbol{\theta}|\mathbf{y})\alpha(\boldsymbol{\theta})\pi_1(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}}{\displaystyle\int_\Theta \tilde{\pi}_1(\boldsymbol{\theta}|\mathbf{y})\alpha(\boldsymbol{\theta})\pi_0(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}} \,.$$

Using this equality, the bridge sampling estimator of B_{01} is given by

$$\hat{B}_{01} = \frac{\dfrac{1}{n_0}\displaystyle\sum_{i=1}^{n_0} \tilde{\pi}_0(\boldsymbol{\theta}_1^i|\mathbf{y})\alpha(\boldsymbol{\theta}_1^i)}{\dfrac{1}{n_1}\displaystyle\sum_{i=1}^{n_1} \tilde{\pi}_1(\boldsymbol{\theta}_0^i|\mathbf{y})\alpha(\boldsymbol{\theta}_0^i)}$$

where $\boldsymbol{\theta}_0^1, \ldots, \boldsymbol{\theta}_0^{n_0}$ is an n_0-sample from $\pi_0(\cdot|\mathbf{y})$ and $\boldsymbol{\theta}_1^1, \ldots, \boldsymbol{\theta}_1^{n_1}$ is an n_1-sample from $\pi_1(\cdot|\mathbf{y})$.

Optimal bridge sampling

The optimal choice of auxiliary function is

$$\alpha^\star(\boldsymbol{\theta}) = \frac{n_0 + n_1}{n_0\pi_0(\boldsymbol{\theta}|\mathbf{y}) + n_1\pi_1(\boldsymbol{\theta}|\mathbf{y})}\,.$$

The dependence on the unknown normalizing constants can be solved iteratively.

Extension to varying dimensions

When $\dim(\Theta_0) \neq \dim(\Theta_1)$, typically $\boldsymbol{\theta}_1 = (\boldsymbol{\theta}, \psi)$ and $f_0(\mathbf{y}|\boldsymbol{\theta}) = f_1(\mathbf{y}|\boldsymbol{\theta}, \psi_0)$ introduction of a pseudo-posterior density, $\omega(\psi|\boldsymbol{\theta}, \mathbf{y})$, augmenting $\pi_0(\boldsymbol{\theta}|\mathbf{y})$ into joint distribution

$$\pi_0(\boldsymbol{\theta}|\mathbf{y})\omega(\psi|\boldsymbol{\theta}, \mathbf{y})$$

on Θ_1 so that

$$B_{01} = \frac{\displaystyle\int_{\Theta_1} f_1(\mathbf{y}|\boldsymbol{\theta}, \psi_0)\pi_0(\boldsymbol{\theta})\alpha(\boldsymbol{\theta}, \psi)\pi_1(\boldsymbol{\theta}, \psi|\mathbf{y})\omega(\psi|\boldsymbol{\theta}, \mathbf{y})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\psi}{\displaystyle\int_{\Theta_1} f_1(\mathbf{y}|\boldsymbol{\theta}, \psi)\pi_1(\boldsymbol{\theta}, \psi)\alpha(\boldsymbol{\theta}, \psi)\pi_0(\boldsymbol{\theta}|\mathbf{y})\omega(\psi|\boldsymbol{\theta}, \mathbf{y})\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\psi}\,,$$

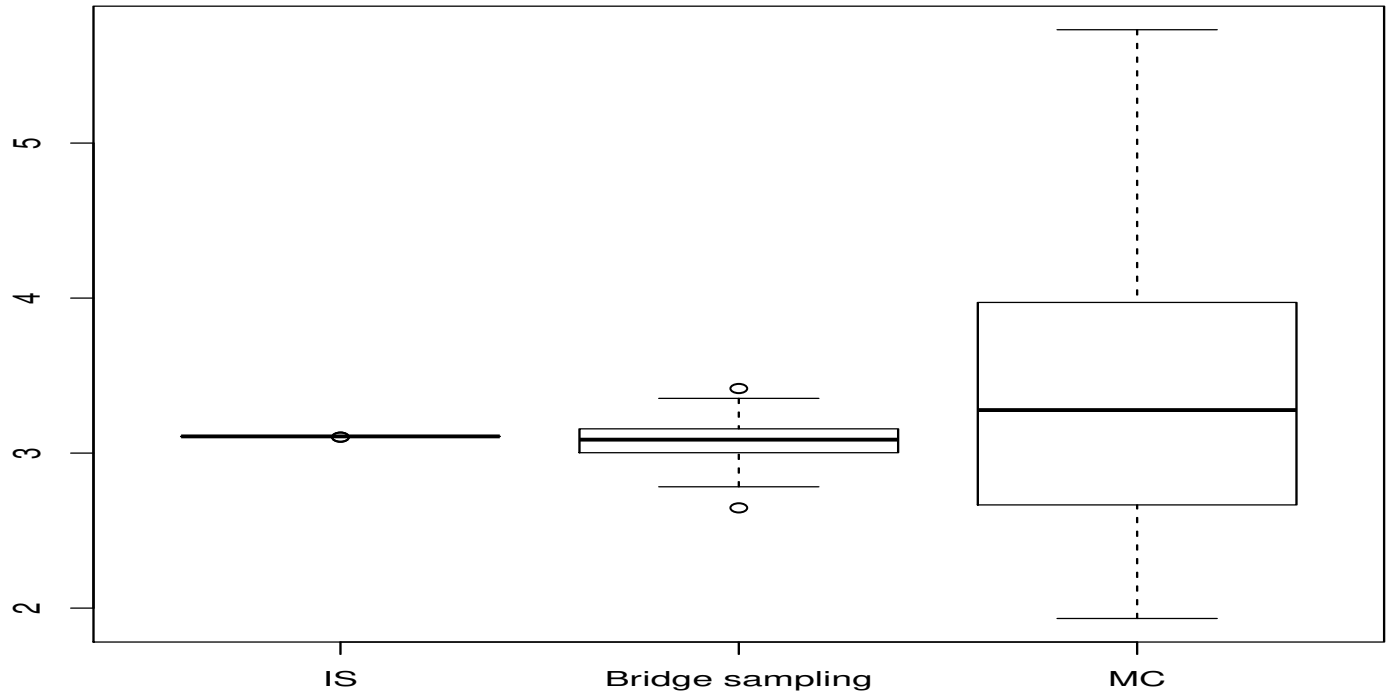
for any conditional density $\omega(\psi|\boldsymbol{\theta})$.

Illustration for the Pima Indian dataset

Use of the MLE induced conditional of θ_3 given (θ_1, θ_2) as a pseudo-posterior and mixture of both MLE approximations on θ_3 in bridge sampling estimate

Diabetes in Pima Indian women (cont'd)

Comparison of the variation of the Bayes factor approximations based on 100 replicas for $20,000$ simulations



Monte Carlo Markov Chain algorithms to explore the space of models

There exists Monte Carlo Markov Chain algorithms, like the reversible jump proposal of Green (1995) and the saturation scheme of Carlin and Chib (1995), that does not require to calculate/approximate the integrated likelihood.

There exists also MCMC algorithms which explore the space of models and for which the acceptance probability depends on approximation or exact value of the integrated likelihood.