Bayesian discrimination between embedded models

Jean-Michel Marin

Université Montpellier 2

Joint work with Christian Robert

Plan

Introduction

Importance sampling solutions

- Regular importance sampling

- Bridge sampling

- Harmonic means

Chib's solution

The Savage–Dickey ratio

ABC method for model choice

Introduction

Model choice

Several models available for the same observation

$$\mathfrak{M}_i : \mathbf{y} \sim f_i(\mathbf{y}|\boldsymbol{\theta}_i), \qquad i \in I$$

where I can be finite or infinite.

Bayesian resolution

Probabilise the entire model/parameter space

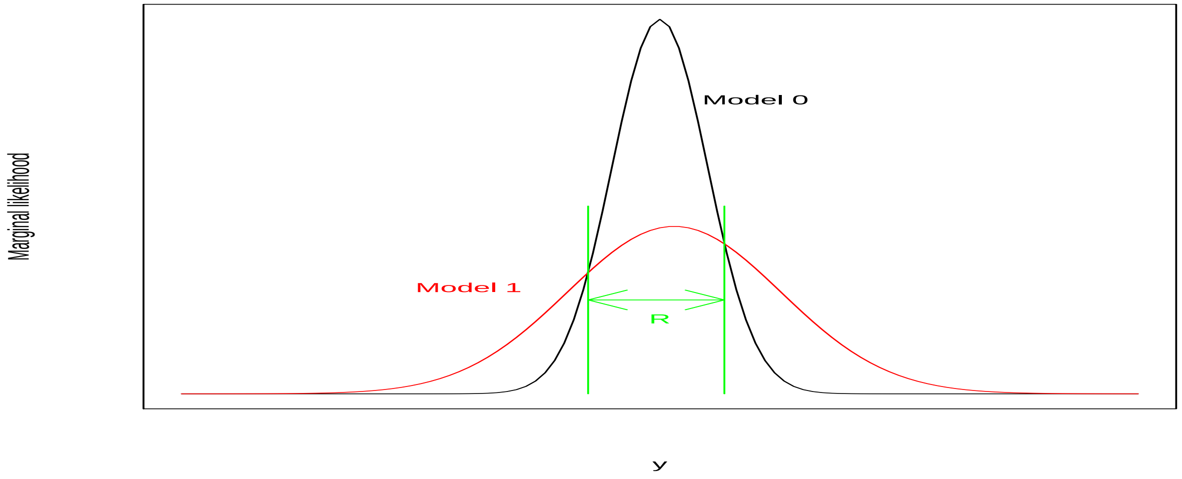
- allocate probabilities p_i to all models $\mathfrak{M}_i$,
- define priors $\pi_i(\boldsymbol{\theta}_i)$ for each parameter space Θ_i,
- compute

$$\mathbb{P}(\mathfrak{M}_i|\mathbf{y}) \propto p_i \int_{\Theta_i} f_i(\mathbf{y}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)\mathrm{d}\boldsymbol{\theta}_i\,,$$

- take largest $\mathbb{P}(\mathfrak{M}_i|\mathbf{y})$ to determine "best" model, or use averaged predictive of $\mathbf{y}'$

$$\sum_j \mathbb{P}(\mathfrak{M}_j|\mathbf{y}) \int_{\Theta_j} p_j(\mathbf{y}'|\boldsymbol{\theta}_j,\mathbf{y})\pi_j(\boldsymbol{\theta}_j|\mathbf{y})\mathrm{d}\boldsymbol{\theta}_j\,.$$

Why Bayesian inference embodies Occam's razor?



This graph gives the basic intuition for why complex models can turn out to be less probable.

The horizontal axis represents the space of possible data sets. Bayes' theorem rewards models in proportion to how much they predicted the data that occurred. These predictions are quantied by a normalized probability distribution.

A simple model, like Model 0, makes only a limited range of predictions; a more powerful model, like Model 1, that has, for example, more free parameters, is able to predict a greater variety of data sets.

Suppose that equal prior probabilities have been assigned to the two models. Then, if the data set falls in region R, the less powerful model will be the more probable model.

The marginal likelihood, which is called the evidence, corrresponds to a **penalized** likelihood!

The BIC information criterium comes from an asymptotic Laplace approximation of the evidence.

Bayes factor

For models $\mathfrak{M}_1$ and $\mathfrak{M}_0$,

$$B_{10} = \frac{\displaystyle\int_{\Theta_1} f_1(\mathbf{y}|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)\mathrm{d}\boldsymbol{\theta}_1}{\displaystyle\int_{\Theta_0} f_0(\mathbf{y}|\boldsymbol{\theta}_0)\pi_0(\boldsymbol{\theta}_0)\mathrm{d}\boldsymbol{\theta}_0}\,.$$

Outside decision-theoretic environment:

- Jeffreys' scale of evidence:
  - if $\log_{10}(B_{10})$ between 0 and 0.5, evidence against $\mathfrak{M}_0$ *weak*,
  - if $\log_{10}(B_{10})$ 0.5 and 1, evidence *substantial*,
  - if $\log_{10}(B_{10})$ 1 and 2, evidence *strong* and,
  - if $\log_{10}(B_{10})$ above 2, evidence *decisive*;

    $(\log_{10}(3) \approx 0.5$ and $\log_{10}(10) = 1$ and $\log_{10}(100) = 2)$.
- Requires the computation of the marginal/evidence under both hypotheses/models.

Evidence

All these problems end up with a similar quantity, the *evidence*, that is the marginal likelihood

$$m_k(\mathbf{y}) = \int_{\Theta_k} \pi_k(\boldsymbol{\theta}_k) f_k(\mathbf{y}|\boldsymbol{\theta}_k)\,\mathrm{d}\boldsymbol{\theta}_k\,.$$

Difficulties with the Bayesian model choice paradigm

Prior difficulties:

- When we have prior informations, how to choose the prior distributions on the parameters of each model in a compatible way? What about the prior distribution in the models's space?
- When we do not have any prior information, **we can not use improper prior distribution**. Indeed, in that case, the models's posterior probabilities are only defined up to some arbitrary constants. How to choose the various prior distributions?

Computational difficulties:

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

We will consider here the case of a limited number of models, typically two embedded models. We will not consider trans-dimensional sampling solutions, like the reversible jump algorithm.

We will concentrate on the crucial question: how to approximate the evidences, and then the Bayes factor?

Importance sampling solutions

Regular importance sampling

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

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

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

MCMC for probit models

Use of either a random walk proposal

$$\boldsymbol{\theta}' = \boldsymbol{\theta} + \boldsymbol{\epsilon}$$

in a Metropolis-Hastings algorithm (since the likelihood is available);

or of a Gibbs sampler that takes advantage of a missing variable representation: a probit model can be represented as a natural latent variable model: $z|\boldsymbol{\theta} \sim \mathcal{N}_1\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}, 1\right)$ and $y = \mathbb{I}_{z>0}$.

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 $\hat{\Sigma}_1$:

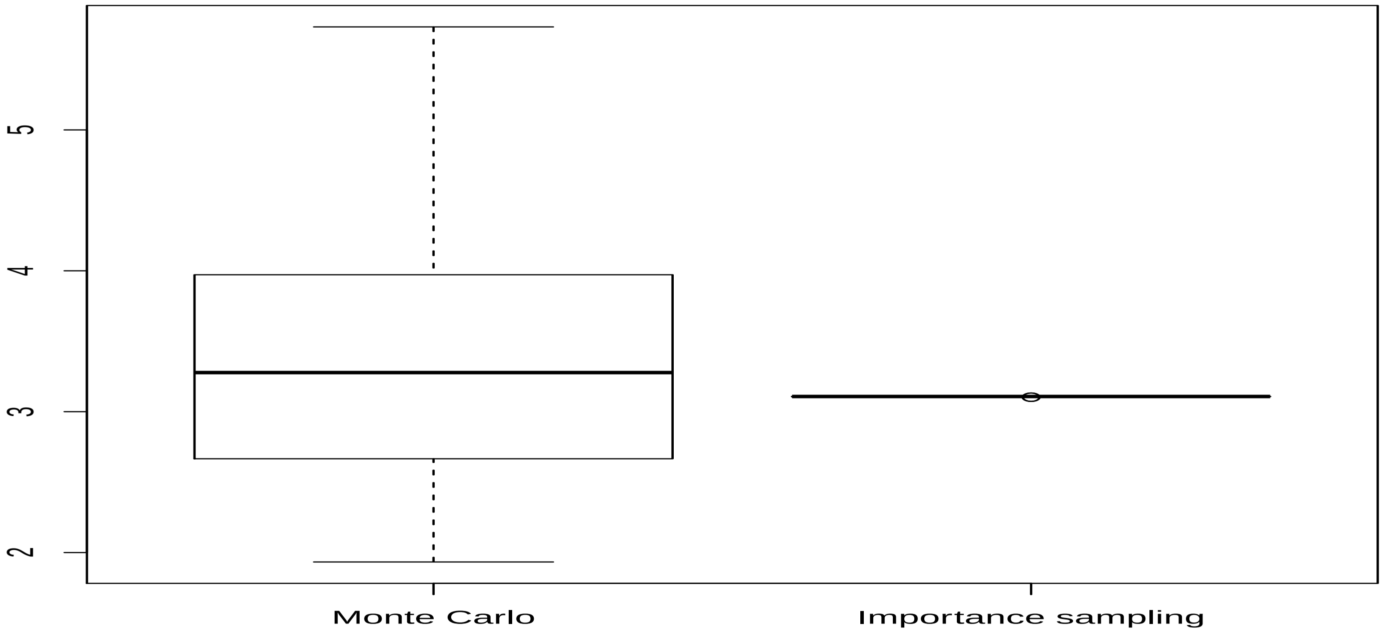
$$g_0(\cdot) \sim \mathcal{N}_2(\hat{\boldsymbol{\theta}}_0, \hat{\Sigma}_0)\,,$$

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



Bridge sampling

If

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

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

$$\begin{aligned}B_{01} &= \int_\Theta f_0(\mathbf{y}|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\bigg/\int_\Theta f_1(\mathbf{y}|\boldsymbol{\theta})\pi_1(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\\ &= \mathbb{E}_{\pi_1}\left[\left.\frac{f_0(\mathbf{y}|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})}{f_1(\mathbf{y}|\boldsymbol{\theta})\pi_1(\boldsymbol{\theta})}\right|\mathbf{y}\right].\end{aligned}$$

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

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

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

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

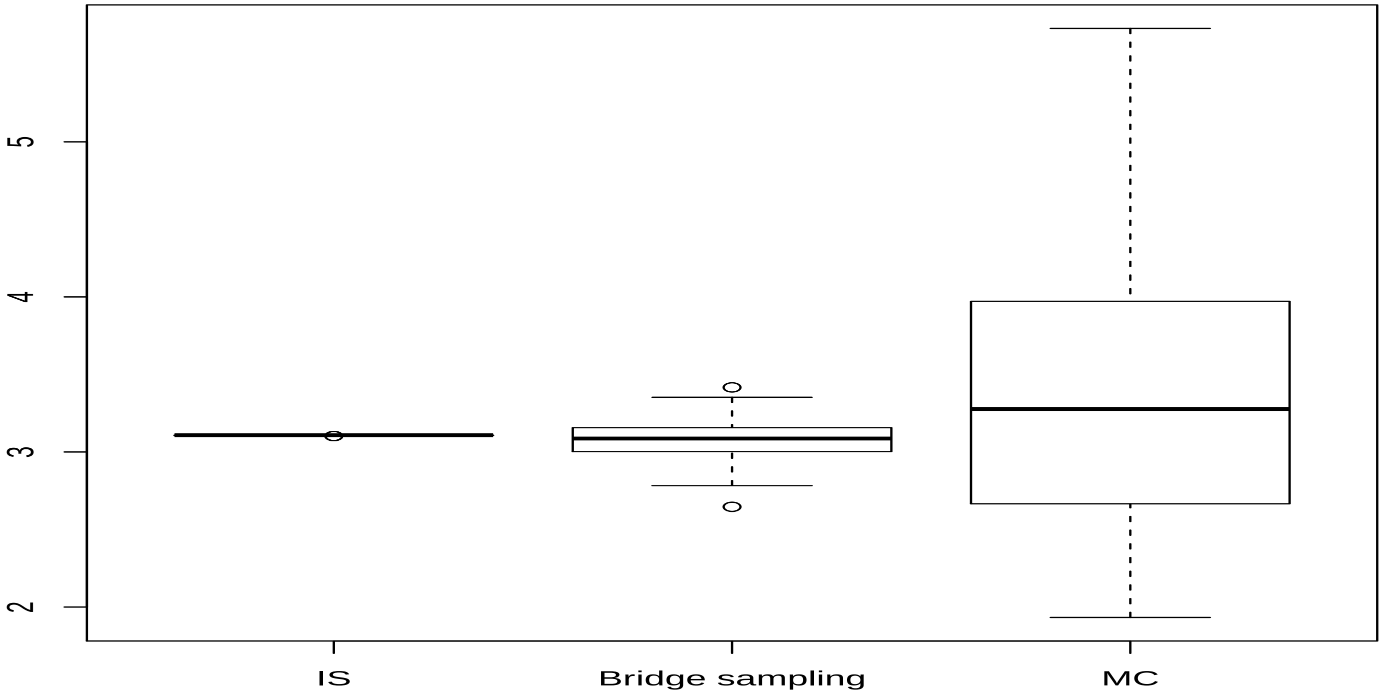
for any conditional density $\omega(\psi|\boldsymbol{\theta})$.

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



The original harmonic mean estimator

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

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

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

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

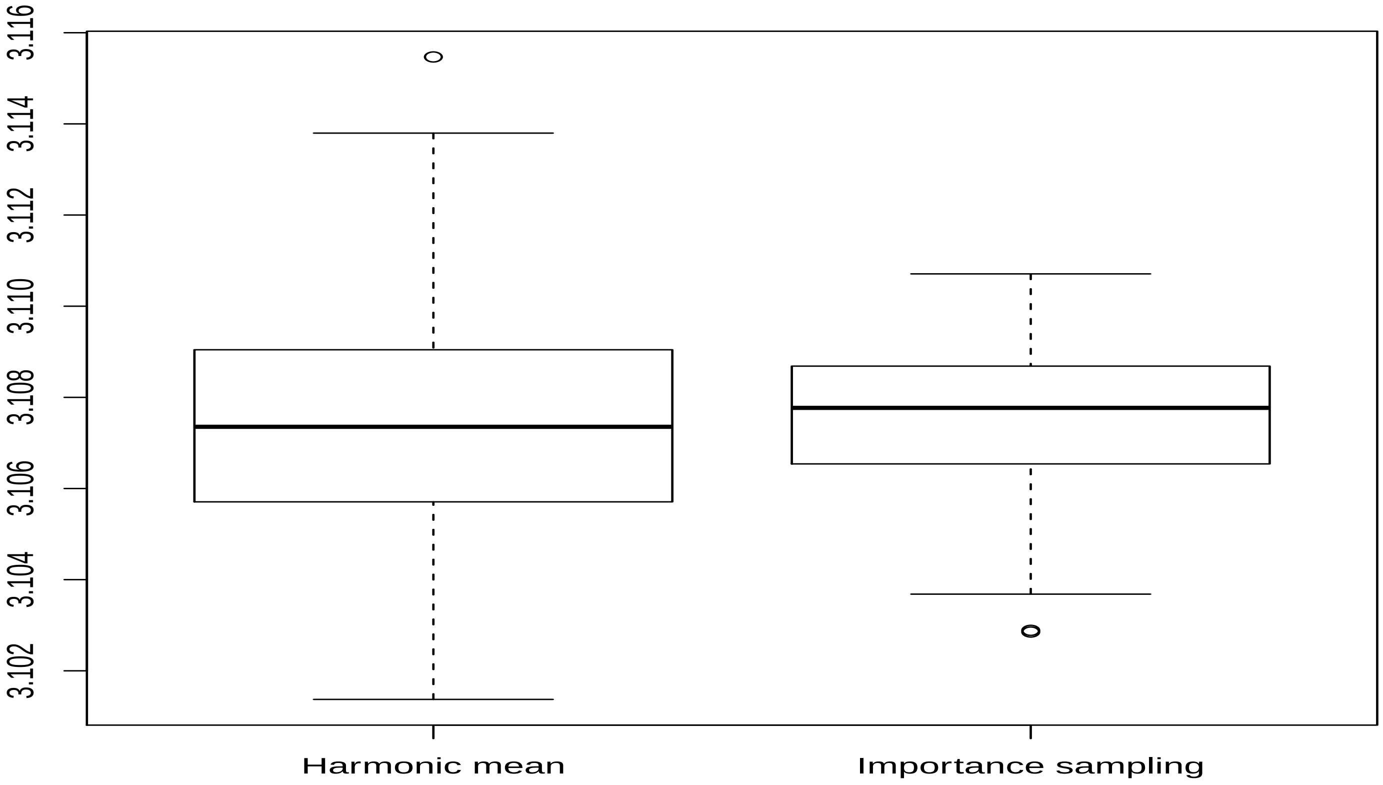
“The good news is that the Law of Large Numbers guarantees that this estimator is consistent ie, it will very likely be very close to the correct answer if you use a sufficiently large number of points from the posterior distribution.

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.”

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 (cont'd)

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



Chib's 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.

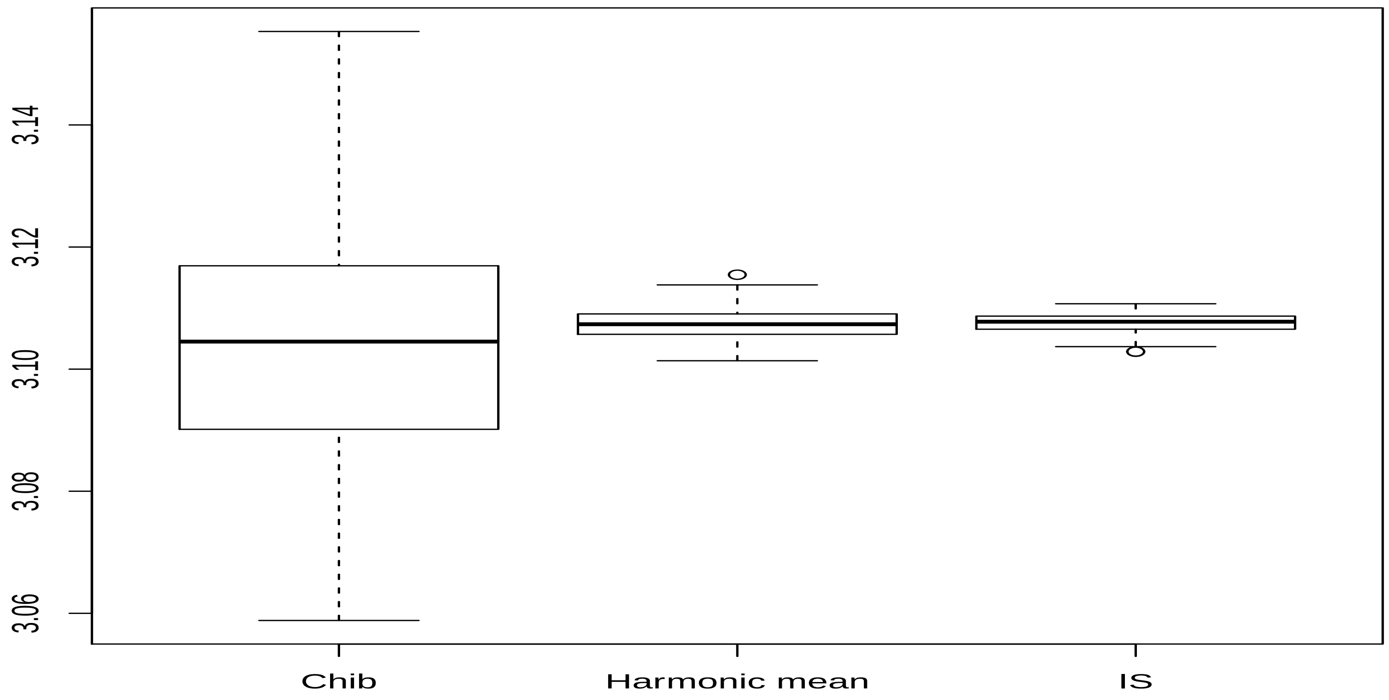
In the special setting of latent variables models, Chib's approximation is particularly attractive as there exists a natural approximation to $\pi_k(\boldsymbol{\theta}^*|\mathbf{y})$, based on the Rao-Blackwell estimate

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

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 Savage–Dickey ratio

Considering a testing problem with an embedded model, $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, and a nuisance parameter ψ, for a sampling distribution $f(\mathbf{y}|\boldsymbol{\theta}, \psi)$, the representation

$$B_{01} = \frac{\pi_1(\boldsymbol{\theta}_0|\mathbf{y})}{\pi_1(\boldsymbol{\theta}_0)}\,,$$

with the obvious notations

$$\pi_1(\boldsymbol{\theta}) = \int \pi_1(\boldsymbol{\theta}, \psi)\mathrm{d}\psi \quad \text{and} \quad \pi_1(\boldsymbol{\theta}|\mathbf{y}) = \int \pi_1(\boldsymbol{\theta}, \psi|\mathbf{y})\mathrm{d}\psi\,,$$

holds under Dickey's assumption

$$\pi_1(\psi|\boldsymbol{\theta}_0) = \pi_0(\psi)\,.$$

Measure-theoretic difficulty

$$
\begin{aligned}
B_{01} &= \frac{\int \pi_0(\psi) f(\mathbf{y}|\boldsymbol{\theta}_0,\psi)\,\mathrm{d}\psi}{\int \pi_1(\boldsymbol{\theta},\psi) f(\mathbf{y}|\boldsymbol{\theta},\psi)\,\mathrm{d}\psi\mathrm{d}\boldsymbol{\theta}} && \text{[by definition]}\\
&= \frac{\int \pi_1(\psi|\boldsymbol{\theta}_0) f(\mathbf{y}|\boldsymbol{\theta}_0,\psi)\,\mathrm{d}\psi\,\pi_1(\boldsymbol{\theta}_0)}{\int \pi_1(\boldsymbol{\theta},\psi) f(\mathbf{y}|\boldsymbol{\theta},\psi)\,\mathrm{d}\psi\mathrm{d}\boldsymbol{\theta}\,\pi_1(\boldsymbol{\theta}_0)} && \text{[using a specific version of } \pi_1(\psi|\boldsymbol{\theta}_0)\text{]}\\
&= \frac{\int \pi_1(\boldsymbol{\theta}_0,\psi) f(\mathbf{y}|\boldsymbol{\theta}_0,\psi)\,\mathrm{d}\psi}{m_1(\mathbf{y})\pi_1(\boldsymbol{\theta}_0)} && \text{[using a specific version of } \pi_1(\boldsymbol{\theta}_0,\psi)\text{]}\\
&= \frac{\pi_1(\boldsymbol{\theta}_0|\mathbf{y})}{\pi_1(\boldsymbol{\theta}_0)}\,, && \text{[using a specific version of } \pi_1(\boldsymbol{\theta}_0|\mathbf{y})\text{]}
\end{aligned}
$$

The last equality leading to the Savage–Dickey representation relies on the choice of a specific version of $\pi_1(\boldsymbol{\theta}_0|x)$ as well, namely

$$\frac{\pi_1(\boldsymbol{\theta}_0|\mathbf{y})}{\pi_1(\boldsymbol{\theta}_0)} = \frac{\int \pi_0(\psi) f(\mathbf{y}|\boldsymbol{\theta}_0,\psi)\,\mathrm{d}\psi}{m_1(\mathbf{y})}\,.$$

Similar measure-theoretic difficulty

Verdinelli-Wasserman have proposed a generalisation of the Savage-Dickey density ratio when the constraint on the prior densities is not verified.

$$B_{01} = \frac{\pi_1(\boldsymbol{\theta}_0|\mathbf{y})}{\pi_1(\boldsymbol{\theta}_0)} \mathbb{E}^{\pi_1(\psi|\mathbf{y},\boldsymbol{\theta}_0)} \left[\frac{\pi_0(\psi)}{\pi_1(\psi|\boldsymbol{\theta}_0)} \right] .$$

This representation remains valid for any choice of versions for $\pi_1(\boldsymbol{\theta}_0|\mathbf{y})$, $\pi_1(\boldsymbol{\theta}_0)$, $\pi_1(\psi|\boldsymbol{\theta}_0)$, provided the conditional density $\pi_1(\psi|\boldsymbol{\theta}_0, \mathbf{y})$ is defined by

$$\pi_1(\psi|\boldsymbol{\theta}_0, \mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta}_0, \psi)\pi_1(\psi|\boldsymbol{\theta}_0)\pi_1(\boldsymbol{\theta}_0)}{m_1(\mathbf{y})\pi_1(\boldsymbol{\theta}_0|\mathbf{y})} .$$

Given a sample $(\boldsymbol{\theta}^{(1)}, \psi^{(1)}, z^{(1)}), \ldots, (\boldsymbol{\theta}^{(T)}, \psi^{(T)}, z^{(T)})$ simulated from (or converging to) $\pi_1(\boldsymbol{\theta}, \psi, z|x)$, the sequence

$$\frac{1}{T} \sum_{t=1}^{T} \pi_1(\boldsymbol{\theta}_0|\mathbf{y}, z^{(t)}, \psi^{(t)})$$

converges to $\pi_1(\boldsymbol{\theta}_0|\mathbf{y})$ under the constraint

$$\frac{\pi_1(\boldsymbol{\theta}_0|\mathbf{y}, z, \psi)}{\pi_1(\boldsymbol{\theta}_0)} = \frac{f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}_0, \psi)}{\int f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}, \psi)\pi_1(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}}\,.$$

Moreover, if $\left(\tilde\psi^{(1)},\tilde z^{(1)}\right),\ldots,\left(\tilde\psi^{(T)},\tilde z^{(T)}\right)$ is a sample generated from (or converging to) $\pi_1(\psi,z|\mathbf{y},\boldsymbol{\theta}_0)$, the sequence

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\pi_0(\tilde\psi^{(t)})}{\pi_1(\tilde\psi^{(t)}|\boldsymbol{\theta}_0)}$$

is converging to

$$\mathbb{E}^{\pi_1(\psi|\mathbf{y},\boldsymbol{\theta}_0)}\left[\frac{\pi_0(\psi)}{\pi_1(\psi|\boldsymbol{\theta}_0)}\right]$$

under the constraint

$$\pi_1(\psi,\mathbf{z}|\boldsymbol{\theta}_0,\mathbf{y})\propto f(\mathbf{y},\mathbf{z}|\boldsymbol{\theta}_0,\psi)\pi_1(\psi|\boldsymbol{\theta}_0)\,.$$

An alternative representation

$$B_{01} = \frac{\int \pi_0(\psi) f(\mathbf{y}|\boldsymbol{\theta}_0,\psi)\,\mathrm{d}\psi}{\int \pi_1(\boldsymbol{\theta},\psi) f(\mathbf{y}|\boldsymbol{\theta},\psi)\,\mathrm{d}\psi\mathrm{d}\boldsymbol{\theta}}\,\frac{\pi_1(\boldsymbol{\theta}_0)}{\pi_1(\boldsymbol{\theta}_0)}\,,$$

the numerator can be seen as involving a specific version in $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ of the marginal posterior density

$$\tilde\pi_1(\boldsymbol{\theta}|\mathbf{y}) \propto \int \pi_0(\psi) f(\mathbf{y}|\boldsymbol{\theta},\psi)\,\mathrm{d}\psi\,\pi_1(\boldsymbol{\theta})\,,$$

which is associated with the alternative prior $\tilde\pi_1(\boldsymbol{\theta},\psi) = \pi_1(\boldsymbol{\theta})\pi_0(\psi)$.

This density $\tilde\pi_1(\boldsymbol{\theta}|\mathbf{y})$ appears as the marginal posterior density of the posterior distribution defined by the density

$$\tilde\pi_1(\boldsymbol{\theta},\psi|\mathbf{y}) = \frac{\pi_0(\psi)\pi_1(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta},\psi)}{\tilde m_1(\mathbf{y})}\,.$$

The version of the marginal posterior density in $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is obtained by imposing

$$\frac{\tilde{\pi}_1(\boldsymbol{\theta}_0|\mathbf{y})}{\pi_0(\boldsymbol{\theta}_0)} = \frac{\int \pi_0(\psi) f(\mathbf{y}|\boldsymbol{\theta}_0, \psi)\,\mathrm{d}\psi}{\tilde{m}_1(\mathbf{y})}\,,$$

where the right hand side of the equation is uniquely defined.

This constraint amounts to imposing that Bayes' theorem holds in $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ instead of almost everywhere (and thus not necessarily in $\boldsymbol{\theta} = \boldsymbol{\theta}_0$).

It then leads to the alternative representation

$$B_{01} = \frac{\tilde{\pi}_1(\boldsymbol{\theta}_0|\mathbf{y})}{\pi_1(\boldsymbol{\theta}_0)}\,\frac{\tilde{m}_1(\mathbf{y})}{m_1(\mathbf{y})}.$$

Given a sample $(\bar{\boldsymbol{\theta}}^{(1)}, \bar{\psi}^{(1)}, \bar{z}^{(1)}), \ldots, (\bar{\boldsymbol{\theta}}^{(T)}, \bar{\psi}^{(T)}, \bar{z}^{(T)})$ simulated from (or converging to) $\tilde{\pi}_1(\boldsymbol{\theta}, \psi, z|\mathbf{y})$, the sequence

$$\frac{1}{T}\sum_{t=1}^{T} \tilde{\pi}_1(\boldsymbol{\theta}_0|\mathbf{y}, \bar{z}^{(t)}, \bar{\psi}^{(t)})$$

converges to $\tilde{\pi}_1(\boldsymbol{\theta}_0|\mathbf{y})$ in T under the constraint

$$\frac{\tilde{\pi}_1(\boldsymbol{\theta}_0|\mathbf{y}, z, \psi)}{\pi_1(\boldsymbol{\theta}_0)} = \frac{f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}_0, \psi)}{\int f(y, z|\boldsymbol{\theta}, \psi)\pi_1(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}\,.$$

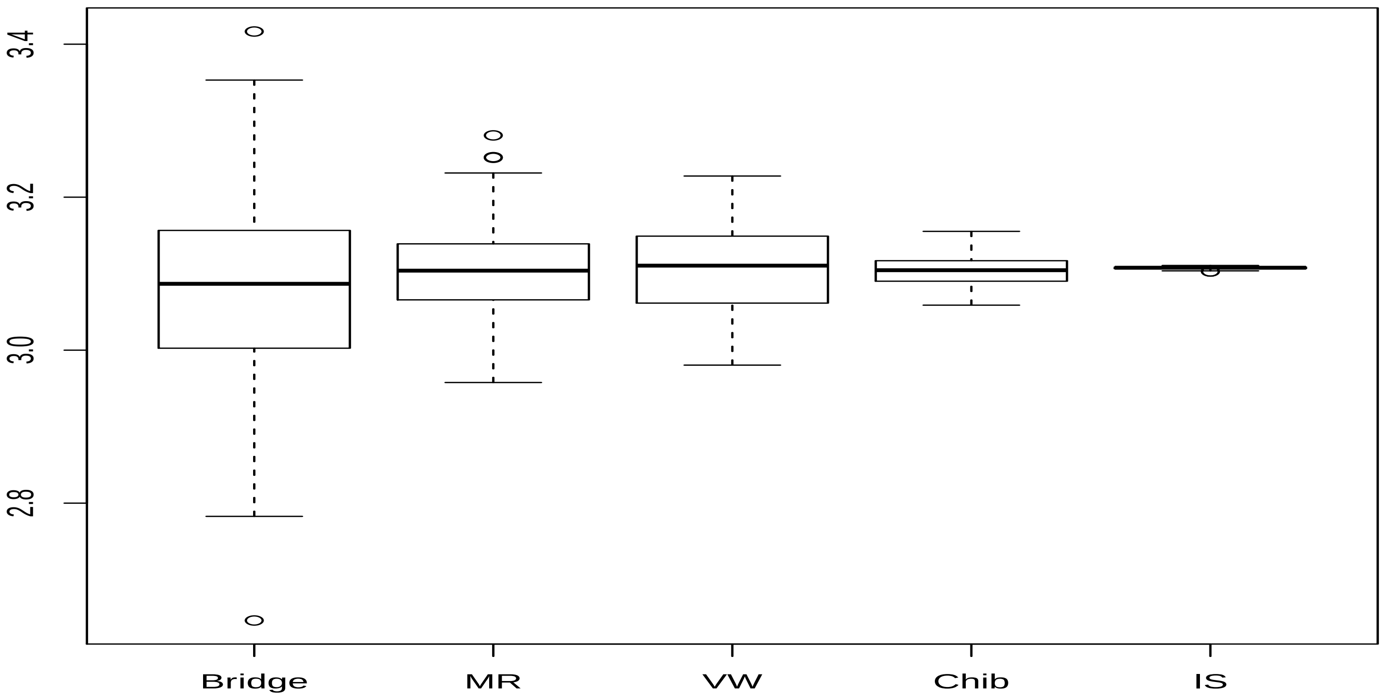
Moreover, if $(\boldsymbol{\theta}^{(1)}, \psi^{(1)}), \ldots, (\boldsymbol{\theta}^{(T)}, \psi^{(T)})$ is a sample independently simulated from (or converging to) $\pi_1(\boldsymbol{\theta}, \psi|\mathbf{y})$, then

$$\frac{1}{T} \sum_{t=1}^{T} \frac{\pi_0(\psi^{(t)})}{\pi_1(\psi^{(t)}|\boldsymbol{\theta}^{(t)})}$$

is a convergent and unbiased estimator of $\tilde{m}_1(\mathbf{y})/m_1(\mathbf{y})$.

Diabetes in Pima Indian women (cont'd)

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



ABC method for model choice

We consider here the Bayesian paradigm.

When the likelihood function $f(\mathbf{y}|\boldsymbol{\theta})$ is expensive or impossible to calculate, it is almost impossible to sample from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$.

ABC is a recent technique that only requires being able to sample from the likelihood $f(\cdot|\boldsymbol{\theta})$.

Likelihood free rejection sampling (Beaumont et al. (2002))

```
1) Set $i = 1$,
2) Generate $\boldsymbol{\theta}'$ from the prior distribution $\pi(\cdot)$,
3) Generate $\mathbf{z}$ from the likelihood $f(\cdot|\boldsymbol{\theta}')$,
4) If $\rho(\eta(\mathbf{z}), \eta(\mathbf{y})) \leq \epsilon$, set $\boldsymbol{\theta}_i = \boldsymbol{\theta}'$ and $i = i + 1$,
5) If $i \leq N$, return to 2).
```

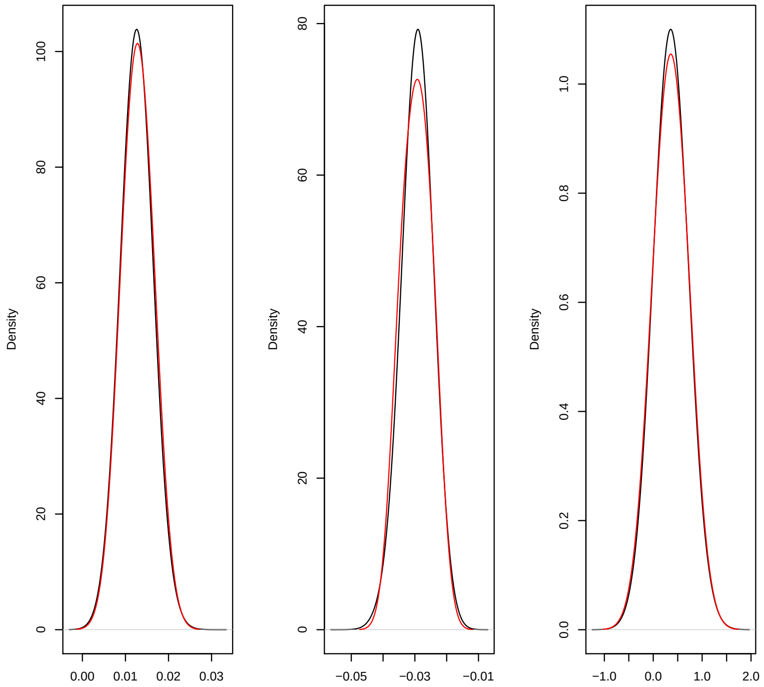
The likelihood free algorithm sample from the marginal in $\mathbf{z}$ of:

$$\pi_\epsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta}) \mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z})}{\int_{A_{\epsilon,\mathbf{y}} \times \boldsymbol{\theta}} \pi(\boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta}) \mathrm{d}\mathbf{z}\mathrm{d}\boldsymbol{\theta}},$$

- $\epsilon > 0$ a tolarance level,
- $\mathbb{I}_B(\cdot)$ the indicator function of a given set B,
- $A_{\epsilon,\mathbf{y}} = \{\mathbf{z} \in \mathcal{D} | \rho(\eta(\mathbf{z}), \eta(\mathbf{y})) < \epsilon\}$.

The idea behind ABC is that the summary statistics coupled with a small tolerance should provide a good approximation of the posterior distribution:

$$\pi_\epsilon(\boldsymbol{\theta}|\mathbf{y}) = \int \pi_\epsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})\mathrm{d}\mathbf{z} \approx \pi(\boldsymbol{\theta}|\mathbf{y})\,.$$



Indian Pima dataset: comparison between densities estimates of the marginal posterior distributions θ_1 (left), θ_2 (center) and θ_3 (right) from ABC rejection samples (in red) and MCMC samples (in black).

The tuning of the ABC algorithm is to use 10^6 simulations, with ϵ set as the 1% quantile of the distances $\rho(\eta(\mathbf{z}), \eta(\mathbf{y}))$, ρ chosen as the Euclidean distance, and $\eta(\mathbf{z})$ as the predictive distribution based on the current parameter, while $\eta(\mathbf{y})$ is the predictive distribution based on the MLE.

In this special case we are therefore avoiding the simulation of the observations themselves as predictive functions are available.

This choice reduces the variability in the divergence between $\eta(\mathbf{z})$ and $\eta(\mathbf{y})$, and explains for the very good results.

ABC-MCMC method

Likelihood free MCMC sampler (Majoram et al. (2003))

```
1) Use the likelihood free rejection sampling to get a realization $\boldsymbol{\theta}^{(0)}$ from the
   ABC target distribution $\pi_\epsilon(\boldsymbol{\theta}|\mathbf{y})$,
2) Set $t = 1$,
3) Generate $\boldsymbol{\theta}'$ from the Markov kernel $q\left(\cdot|\boldsymbol{\theta}^{(t-1)}\right)$,
4) Generate $\mathbf{z}$ from the likelihood $f(\cdot|\boldsymbol{\theta}')$,
5) Generate $u$ from $\mathcal{U}_{[0,1]}$,
6) If $u \le \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^{(t-1)}q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)}))}\mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z})$,
   set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}'$ else $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$,
7) Set $t = t + 1$,
8) If $t \le N$ return to 3).
```

Rejection sampling and MCMC methods can perform poorly if the tolerance level ϵ is small.

Consequently various sequential Monte Carlo algorithms have been constructed as an alternative to these two methods (Beaumont et al. (2009)).

The key idea is to decompose the difficult problem of sampling from $\pi_\epsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$ into a series of simpler subproblems.

The algorithm begins at time 0 sampling from $\pi_{\epsilon_0}(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$ with large ϵ_0, then simulating from an increasing difficult sequence of target distribution $\pi_{\epsilon_t}(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$, that is $\epsilon_t < \epsilon_{t-1}$.

ABC methods for model choice in Gibbs random fields

We consider a finite set of sites $\mathcal{S} = \{1, \cdots, n\}$.

At each site $i \in \mathcal{S}$, we observe $x_i \in \mathcal{X}_i$ where $\mathcal{X}_i$ is a finite set of states.

We also consider an undirected graph $\mathcal{G}$: the sites i and $i^{'}$ are said neighbours, if there is a vertex between i and $i^{'}$.

A clique c is a subset of $\mathcal{S}$ where all elements are mutual neighbours (Daroch, 1980).

We denote by $\mathcal{C}$ the set of all cliques of the undirected graph $\mathcal{G}$.

Gibbs Random Fields (GRFs) are probabilistic models associated with densities

$$f(\mathbf{x}) = \frac{1}{Z}\exp\{-U(\mathbf{x})\} = \frac{1}{Z}\exp\left\{-\sum_{c\in\mathcal{C}} U_c(\mathbf{x})\right\},$$

where $U(\mathbf{x}) = \sum_{c\in\mathcal{C}} U_c(\mathbf{x})$ is the potential and Z is the corresponding normalising constant

$$Z = \sum_{\mathbf{x}\in\mathcal{X}} \exp\left\{-\sum_{c\in\mathcal{C}} U_c(\mathbf{x})\right\}.$$

If the density f of a Markov Random Field (MRF) is everywhere positive, then the Hammersley-Clifford theorem establishes that there exists a GRF representation of this MRF (Besag, 1974).

We consider here GRF with potential $U(\mathbf{x}) = -\boldsymbol{\theta}^{\mathrm{T}} S(\mathbf{x})$ where $\boldsymbol{\theta} \in \mathbb{R}^p$ is a scale parameter, $S(\cdot)$ is a function taking values in $\mathbb{R}^p$.

$S(\mathbf{x})$ is defined on the cliques of the neighbourhood system in that $S(\mathbf{x}) = \sum_{c \in \mathcal{C}} S_c(\mathbf{x})$.

In that case, we have

$$f(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z_{\boldsymbol{\theta}}} \exp\{\boldsymbol{\theta}^{\mathrm{T}} S(\mathbf{x})\} \,,$$

the normalising constant $Z_{\boldsymbol{\theta}}$ now depends on the scale parameter $\boldsymbol{\theta}$.

GRF are used to model the dependency within spatially correlated data, with applications in epidemiology and image analysis, among others (Rue and Held, 2005).

They often use a Potts model defined by a sufficient statistic S taking values in $\mathbb{R}$ in that

$$S(\mathbf{x}) = \sum_{i' \sim i} \mathbb{I}_{\{x_i = x_{i'}\}} \,,$$

where $\sum_{i' \sim i}$ indicates that the summation is taken over all the neighbour pairs.

$\mathcal{X}_i = \{1, \cdots, K\}$, $K = 2$ corresponding to the Ising model, and $\boldsymbol{\theta}$ is a scalar.

$S(\cdot)$ monitors the number of identical neighbours over $\mathcal{X}$.

In most realistic settings, the summation

$$Z_{\boldsymbol{\theta}} = \sum_{\mathbf{x}\in\mathcal{X}} \exp\{\boldsymbol{\theta}^{\mathrm{T}} S(\mathbf{x})\}$$

involves too many terms to be manageable.

Selecting a model with sufficient statistic S_0 versus a model with sufficient statistics S_1 relies on the Bayes factor

$$BF_{m_0/m_1}(\mathbf{x}) = \int \exp\{\boldsymbol{\theta}_0^{\mathrm{T}} S_0(\mathbf{x})\}/Z_{\boldsymbol{\theta}_0,0}\pi_0(\mathrm{d}\boldsymbol{\theta}_0) \Bigg/ \int \exp\{\boldsymbol{\theta}_1^{\mathrm{T}} S_1(\mathbf{x})\}/Z_{\boldsymbol{\theta}_1,1}\pi_1(\mathrm{d}\boldsymbol{\theta}_1)$$

This quantity is not easily computable.

For a fixed neighbourhood or model, the unavailability of $Z_{\boldsymbol{\theta}}$ complicates inference on the scale parameter $\boldsymbol{\theta}$.

The difficulty is increased manifold when several neighbourhood structures are under comparison.

We propose a procedure based on an ABC algorithm aimed at selecting a model.

We consider the toy example of an iid sequence [with trivial neighbourhood structure] tested against a Markov chain model [with nearest neighbour structure].

In a model choice perspective, we face M Gibbs random fields in competition.

Each model m is associated with sufficient statistic S_m $(0 \leq m \leq M-1)$, i.e. with corresponding likelihood

$$f_m(\mathbf{x}|\boldsymbol{\theta}_m) = \exp\left\{\boldsymbol{\theta}_m^{\mathrm{T}} S_m(\mathbf{x})\right\} \Big/ Z_{\boldsymbol{\theta}_m,m}\,,$$

where $\boldsymbol{\theta}_m \in \boldsymbol{\theta}_m$ and $Z_{\boldsymbol{\theta}_m,m}$ is the unknown normalising constant.

The choice between those models is driven by the posterior probabilities of the models.

We consider an extended parameter space $\boldsymbol{\theta} = \cup_{m=0}^{M-1}\{m\} \times \boldsymbol{\theta}_m$ that includes the model index $\mathcal{M}$,

We define a prior distribution on the model index $\pi(\mathcal{M} = m)$ as well as a prior distribution on the parameter conditional on the value m of the model index, $\pi_m(\boldsymbol{\theta}_m)$, defined on the parameter space $\boldsymbol{\theta}_m$.

The computational target is thus the model posterior probability

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) \propto \int_{\boldsymbol{\theta}_m} f_m(\mathbf{x}|\boldsymbol{\theta}_m)\pi_m(\boldsymbol{\theta}_m)\,\mathrm{d}\boldsymbol{\theta}_m\,\pi(\mathcal{M} = m)\,,$$

the marginal of the posterior distribution on $(\mathcal{M}, \boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_{M-1})$ given $\mathbf{x}$.

If $S(\mathbf{x})$ is a sufficient statistic for the joint parameters $(\mathcal{M}, \boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_{M-1})$,

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) = \mathbb{P}(\mathcal{M} = m|S(\mathbf{x}))\,.$$

Each model has its own sufficient statistic $S_m(\cdot)$.

Then, for each model, the vector of statistics $S(\cdot) = (S_0(\cdot), \ldots, S_{M-1}(\cdot))$ is obviously sufficient.

We have shown that the statistic $S(\mathbf{x})$ is also sufficient for the joint parameters $(\mathcal{M}, \boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_{M-1})$.

That the concatenation of the sufficient statistics of each model is also a sufficient statistic for the joint parameters is a property that is specific to Gibbs random field models.

When we consider M models from generic exponential families, this property of the concatenated sufficient statistic rarely holds.

ABC algorithm for model choice (Grelaud et al. (2009))

```
1) Set $i = 1$,
2) Generate $m'$ from the prior $\pi(\mathcal{M} = m)$,
3) Generate $\boldsymbol{\theta}'_{m'}$ from the prior $\pi_{m'}(\cdot)$,
4) Generate $\mathbf{z}$ from the model $f_{m'}(\cdot|\boldsymbol{\theta}'_{m'})$,
5) If $\rho(S(\mathbf{z}), S(\mathbf{x})) \leq \epsilon$, set $m^i = m'$, $\boldsymbol{\theta}^i_{m^i} = \boldsymbol{\theta}'_{m'}$ and $i = i + 1$,
6) If $i \leq N$, return to 2).
```

Simulating a data set $\mathbf{z}$ from $f_{m'}(\cdot|\boldsymbol{\theta}'_{m'})$ at step 3 is non-trivial for GRFs (Møller and Waagepetersen, 2003).

It is often possible to use a Gibbs sampler updating one clique at a time conditional on the others.

This algorithm results in an approximate generation from the joint posterior distribution

$$\pi\left\{(\mathcal{M}, \boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_{M-1})|\rho(S(\mathbf{x}), S(\mathbf{z})) \leq \epsilon\right\} .$$

When it is possible to achieve $\epsilon = 0$, the algorithm is exact since S is a sufficient statistic.

Once a sample of N values of $(\boldsymbol{\theta}^i_{m^i}, m^i)$ $(1 \leq i \leq N)$ is generated from this algorithm, a standard Monte Carlo approximation of the posterior probabilities is provided by the empirical frequencies of visits to the model, namely

$$\widehat{\mathbb{P}}(\mathcal{M} = m|\mathbf{x}) = \sharp\{m^i = m\}\big/N\,,$$

where $\sharp\{m^i = m\}$ denotes the number of simulated m^i's equal to m.

$$BF_{m_0/m_1}(\mathbf{x}) = \frac{\mathbb{P}(\mathcal{M}=m_0|\mathbf{x})}{\mathbb{P}(\mathcal{M}=m_1|\mathbf{x})} \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)}$$

$$\overline{BF}_{m_0/m_1}(\mathbf{x}) = \frac{\sharp\{m^i=m_0\}}{\sharp\{m^i=m_1\}} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)}\,,$$

This estimate is only defined when $\sharp\{m_i=m_1\} \neq 0$.

To bypass this difficulty, the substitute

$$\widehat{BF}_{m_0/m_1}(\mathbf{x}) = \frac{1+\sharp\{m^i=m_0\}}{1+\sharp\{m^i=m_1\}} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)}$$

is particularly interesting because we can evaluate its bias.

We set $N_0 = \sharp\{m^i = m_0\}$ and $N_1 = \sharp\{m^i = m_1\}$.

If $\pi(\mathcal{M} = m_1) = \pi(\mathcal{M} = m_0)$, then N_1 is a binomial $\mathcal{B}(N, \rho)$ random variable with probability $\rho = (1 + BF_{m_0/m_1}(\mathbf{x}))^{-1}$ and

$$\mathbb{E}\left[\frac{N_0+1}{N_1+1}\right] = BF_{m_0/m_1}(\mathbf{x}) + \frac{1}{\rho(N+1)} - \frac{N+2}{\rho(N+1)}(1-\rho)^{N+1}\,.$$

The bias of $\widehat{BF}_{m_0/m_1}(\mathbf{x})$ is $\{1-(N+2)(1-\rho)^{N+1}\}/(N+1)\rho$, which goes to zero as N goes to infinity.

$\widehat{BF}_{m_0/m_1}(\mathbf{x})$ can be seen as the ratio of the posterior means on the model probabilities under a $\mathcal{D}ir(1, \cdots, 1)$ prior.

Results on a toy example:

Our example compares an iid Bernoulli model with a two-state first-order Markov chain.

Both models are special cases of GRF, the first one with a trivial neighbourhood structure and the other one with a nearest neighbourhood structure.

Furthermore, the normalising constant $Z_{\boldsymbol{\theta}_m,m}$ can be computed in closed form, as well as the posterior probabilities of both models.

We consider a sequence $\mathbf{x} = (x_1, .., x_n)$ of binary variables. Under model $\mathcal{M} = 0$, the GRF representation of the Bernoulli distribution $\mathcal{B}(\exp(\theta_0)/\{1+\exp(\theta_0)\})$ is

$$f_0(\mathbf{x}|\theta_0) = \exp\left(\theta_0 \sum_{i=1}^{n} \mathbb{I}_{\{x_i=1\}}\right) \Big/ \{1+\exp(\theta_0)\}^n \,.$$

For $\theta_0 \sim \mathcal{U}(-5,5)$, the posterior probability of this model is available since the marginal when $S_0(\mathbf{x}) = s_0$ ($s_0 \neq 0$) is given by

$$\frac{1}{10} \sum_{k=0}^{s_0-1} \binom{s_0-1}{k} \frac{(-1)^{s_0-1-k}}{n-1-k} \left[(1+e^5)^{k-n+1} - (1+e^{-5})^{k-n+1}\right] .$$

Model $\mathcal{M} = 1$ is chosen as a Markov chain.

We assume a uniform distribution on x_1 and

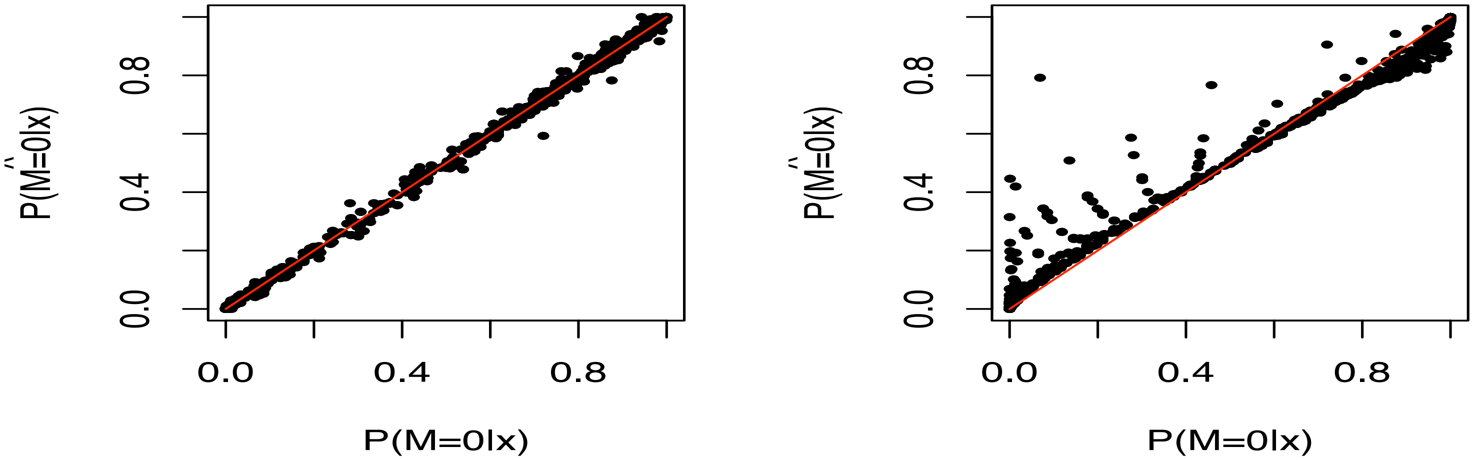
$$f_1(\mathbf{x}|\theta_1) = \frac{1}{2} \exp\left(\theta_1 \sum_{i=2}^{n} \mathbb{I}_{\{x_i = x_{i-1}\}}\right) \Big/ \{1 + \exp(\theta_1)\}^{n-1} \,.$$

For $\theta_1 \sim \mathcal{U}(0,6)$, the posterior probability of this model is once again available, the likelihood being of the same form as when $\mathcal{M} = 0$.

We simulated $2,000$ datasets $\mathbf{x} = (x_1, \cdots, x_n)$ with $n = 100$ under each model, using parameters simulated from the priors.

For each of those $2,000$ datasets $\mathbf{x}$, the ABC-MC algorithm was run for 4×10^6 loops, meaning that 4×10^6 sets $(m, \boldsymbol{\theta}_m, \mathbf{z})$ were exactly simulated from the joint distribution.

A random number of those were accepted when $S(\mathbf{z}) = S(\mathbf{x})$. (In the worst case scenario, the number of acceptances was 12!)



Comparison of the true $\mathbb{P}(\mathcal{M} = 0|\mathbf{x})$ with $\widehat{\mathbb{P}}(\mathcal{M} = 0|\mathbf{x})$ over $2,000$ simulated sequences and 4×10^6 proposals from the prior. The red line is the diagonal. *(right)* Same comparison when using a tolerance ϵ corresponding to the 1% quantile on the distances.