Use in practice of importance sampling for repeated MCMC for Poisson models

Dorota Gajda, Chantal Guihenneuc-Jouyaux, Judith Rousseau, Kerrie L. Mengersen, Darfiana Nur

Electronic Journal of Statistics
Vol. 4 (2010) 361–383
ISSN: 1935-7524
DOI: 10.1214/09-EJS527

Use in practice of importance sampling for repeated MCMC for Poisson models

Dorota Gajda

Biostatistics, CESP Centre de recherche en Epidémiologie et Santé des Populations, U1018, Inserm, F-94807, Villejuif, France
Université Paris Sud, UMRS1018, Villejuif, F-94807, France
e-mail: dorota.gajda@inserm.fr

Chantal Guihenneuc-Jouyaux*

Laboratoire MAP5 (UMR CNRS 8145), Université Paris Descartes, 45 rue des Saints Peres, 75006 Paris, France
and Biostatistics, CESP Centre de recherche en Epidémiologie et Santé des Populations, U1018, Inserm, F-94807, Villejuif, France
e-mail: chantal.guihenneuc@parisdescartes.fr

Judith Rousseau

University Paris Dauphine, Paris, France

Kerry Mengersen

QUT, Brisbone, Australia

and

Darfiana Nur

School of Mathematical and Physical Sciences The University of Newcastle, Callaghan, NSW 2308, Australia

Abstract: The Importance Sampling method is used as an alternative approach to MCMC in repeated Bayesian estimations. In the particular context of numerous data sets, MCMC algorithms have to be called on several times which may become computationally expensive. Since Importance Sampling requires a sample from a posterior distribution, our idea is to use MCMC to generate only a certain number of Markov chains and use them later in the subsequent IS estimations. For each Importance Sampling procedure, the suitable chain is selected by one of three criteria we present here. The first and second criteria are based on the L^1 norm of the difference between two posterior distributions and their Kullback-Leibler divergence respectively. The third criterion results from minimizing the variance of IS estimate. A supplementary automatic selection procedure is also proposed to choose those posterior for which Markov chains will be generated and to avoid arbitrary choice of importance functions. The featured methods are illustrated in simulation studies on three types of Poisson model: simple Poisson model, Poisson regression model and Poisson regression model with extra Poisson variability. Different parameter settings are considered.

*Corresponding author.

AMS 2000 subject classifications: Primary 62F15, 65C05; secondary 65C60.
Keywords and phrases: MCMC, Importance Sampling, Poisson model.

Received November 2009.

Contents

1. Introduction

Bayesian approach allows the introduction of prior knowledge of the parameter via prior probability distribution. The Bayesian idea consists of combining two sources of information on the parameter, one from the data through a likelihood function, and the other implied by the prior. The result is the posterior distribution of the parameter which is a conditional distribution, given the data. In complex models, an analytical solution of the posterior distribution and its descriptive statistics are generally not available. In this case, approximation methods are used which are based on stochastic MCMC algorithms such as the Hastings-Metropolis algorithm (Hastings, 1970) or Gibbs sampler (Geman and Geman, 1984). These algorithms generate a Markov chain whose stationary distribution is the posterior distribution. The ergodic theorem guarantees that empirical averages provide good approximations but on condition that the algorithm achieves convergence, which often requires a huge number of iterations. In cases where new models are studied, performances of estimates can be evaluated in a simulation study. For example, traditionally, in order to assess qualities such as bias or quadratic errors of estimates, different data sets are simulated for different parameter settings, then relevant priors are set on these parameters

and posterior estimates computed for all data sets using an MCMC approach. Unfortunately, the time required to perform this computation is usually prohibitive. An alternative solution would be to apply the Importance Sampling (IS) method, by which we can estimate one expected probability distribution using a sample from a different distribution which is easier to simulate. The simultaneous use of IS and MCMC was previously proposed by Geyer and Thompson (1992) in likelihood computation, by Gelfand, Dey and Chang (1992) in cross-validation and model determination, and more recently in Population Monte Carlo approach (Cappé *et al.*, 2004; Douc *et al.*, 2007b,a; Cappé *et al.*, 2007). A global use of IS in repeated MCMC is described in McVinish *et al.* (2008) where IS is used among other techniques in simulation studies and asymptotic efficacy is discussed.

Importance Sampling has to be done under conditions where both densities are close and the support of the sampling distribution covers the support of the initial one. When simulation studies are performed, the posterior distribution is conditional on different data sets which have been simulated under the same model. Theoretically, these posteriors should resemble one another as should the data sets. A first idea is to run an MCMC algorithm on the first data set and then apply IS to obtain estimates for the other distributions using the first posterior distribution as the importance function. This approach can be compared with an idea of predictive distribution estimations of Gelfand, Dey and Chang (1992) where a single choice of importance sampling distribution is taken into account. Nevertheless, in a more general case, if the posterior distribution is conditional on one data set this could lead to poor approximations (when used as an importance function in the IS method) for the other data sets. This is why the choice of sampling distribution may depend on the choice of data set. We therefore propose three criteria for choosing the posterior sampling distribution.

The objective of this paper is to assess and improve the efficacy of IS in relation to the MCMC method as an easier and faster way of Bayesian estimation. The method is illustrated on Poisson models and the data sets used in this study are simulated for different values of the parameters. The parameters were estimated for each data set by both MCMC and the IS method, and the results were compared through the mean square errors.

In section two we present the concept of the IS method and define a context in which we use it. Then we propose three criteria for choosing the best sampling distribution for the IS approximation among a set of preselected posterior distributions. In the third section we present results of discussed estimation methods applied in different Poisson models. A selection procedure is presented in the fourth section to obtain automatically the set of preselected posterior distributions. Conclusions are reported in the final section.

2. Methods

Suppose data X are described by a probability model $(\pi(x|\theta), \theta \in \Theta)$. Prior distribution on the parameter θ (which can be multidimensional) is denoted

by $\pi(\theta)$ and the posterior distribution by $\pi(\theta|x)$. We are equally interested in some descriptive statistics of this posterior distribution such as posterior mean, posterior standard deviation, posterior quantiles (credibility interval), etc. Unfortunately, the closed-form solution for the posterior is very often not available and nor are its descriptive statistics. In the Bayesian context the standard solution for dealing with this problem is provided by MCMC methods (see Gilks, Richardson and Spiegelhalter, 1996). For a given data set X and a given prior distribution on θ, a MCMC algorithm generates a Markov chain $\theta_1, \ldots, \theta_N$ whose stationary distribution is $\pi(\theta|x)$. The ergodic theorem ensures that the expected value $E_{[\theta|X]}[g(\theta)]$with respect to the posterior distribution of θ of any integrable function g can be approximated almost surely by $1/N \sum_{i=1}^{N} g(\theta_i)$. According to the choice of the function g we can approximate different descriptive statistics for the posterior distribution.

In the case of many repeated Bayesian estimations (e.g. while testing a new model's properties by simulating many data sets and estimating the model's parameters afterwards), MCMC has to be used for each sample, which is often expensive and time-consuming. When carrying out an empirical study of estimates, data are generally simulated from the model $\pi(x|\theta)$ under different parameterization schemes. Parameters (or functions of the parameters) are then estimated for each simulated data set. Since we are interested in Bayesian estimation and the posterior does not have an analytical solution, we can approximate it via MCMC. Consider L parameterization schemes and for each K simulated samples, then the MCMC algorithm must be run L×K times, thus extending the time taken for the estimations. In order to speed up the computation of posterior expectations, we propose two strategies applying the Importance Sampling method. Globally, these methods require that MCMC algorithm is used on a subset of samples. As IS theoretical results are based on the assumption that MCMC samples are from the posterior distribution, we assume that the convergence of algorithms is carefully checked. In section 3, different diagnostics are proposed for our applications.

2.1. Importance Sampling: fixed strategy

Importance Sampling (IS) is a well-known estimation method. We are particularly interested in using it to approximate expected values under posterior distributions in order to reduce the number of times MCMC is used. We present the concept of IS on the example of $E_{[\theta|X^{(k)}]}[g(\theta)]$ estimation using two data sets $X^{(k)}$ and $X^{(m)}$ drawn from the same model $\pi(x|\theta)$ (with the same model parameterization). The expected value $E_{[\theta|X^{(k)}]}[g(\theta)]$ is given by the integral formula and the Importance Sampling method relies on the following transformation under the integral sign

$$E_{[\theta|X^{(k)}]}[g(\theta)] = \int g(\theta)\pi(\theta|X^{(k)})d\theta = \int g(\theta)\frac{\pi(\theta|X^{(k)})}{\pi(\theta|X^{(m)})}\pi(\theta|X^{(m)})d\theta \qquad (1)$$

with $\pi(\theta|X^{(m)}) \neq 0$ whenever $\pi(\theta|X^{(k)}) \neq 0$. The posterior distribution $\pi(\theta|X^{(m)})$ is called the importance function in IS procedure (right term of 1). Then, if we already have an MCMC sample $\theta_1^{(m)}, \ldots, \theta_N^{(m)}$ from the posterior distribution $\pi(\theta|X^{(m)})$ of length N, the approximation of the expected value $E_{[\theta|X^{(k)}]}[g(\theta)]$ of a real-valued function g such that $E_{[\theta|X^{(k)}]}[|g(\theta)|] < \infty$, by the ergodic theorem (Theorem 3 Tierney, 1994) is:

$$\frac{1}{N}\sum_{i=1}^{N} g(\theta_i^{(m)})\frac{\pi(\theta_i^{(m)}|X^{(k)})}{\pi(\theta_i^{(m)}|X^{(m)})} \to E_{[\theta|X^{(k)}]}[g(\theta)] \quad a.s.. \tag{2}$$

If there is no closed-form solution for the posterior distribution, the approximation (2) can be replaced by another one which is easily calculable with normalized weights:

$$\sum_{i=1}^{N} g(\theta_i^{(m)})\tilde{w}_i(k,m) \tag{3}$$

$$\text{where} \quad \tilde{w}_i(k,m) = \frac{w_i(k,m)}{\sum_{i=1}^{N} w_i(k,m)} \quad \text{and} \quad w_i(k,m) = \frac{\pi(X^{(k)}|\theta_i^{(m)})}{\pi(X^{(m)}|\theta_i^{(m)})}$$

The IS estimator with normalized weights given in (3) is also consistent for $E_{[\theta|X^{(k)}]}[g(\theta)]$ as a consequence of the strong law of large numbers. Geweke (1989) discusses the IS estimator with normalized weights for a case of iid samples, which can be extended to the case of Markovian samples as a result of the Markov chains theory (see, for example Tierney, 1994). Moreover, Geweke (1989) gives conditions for central limit theorem to hold for IS estimator with normalized weights which can also be extended for Markovian samples. Indeed, in discussion with Tierney (1994), Doss (1994) proves that for stationary and uniformly ergodic chain for which conditions $E_{\pi(\theta|X^{(m)})}[(g(\theta)w(\theta))^2] < \infty$ and $E_{\pi(\theta|X^{(m)})}[(w(\theta))^2] < \infty$ are satisfied, an IS estimator with normalized weights has a normal asymptotic distribution. As discussed in Geweke (1989), the limiting distribution assessment may give us some indications of $\left|\sum_{i=1}^{N} g(\theta_i^{(m)})\tilde{w}_i(k,m) - E_{\pi(\theta|X^{(k)})}[g(\theta)]\right|$. Hence, the MCMC sample $\theta_1^{(m)}, \ldots, \theta_N^{(m)}$ used in further calculations needs to be ergodic from the posterior distribution $\pi(\theta|X^{(m)})$ being its equilibrium distribution. Consequently we can easily approximate $E_{[\theta|X^{(k)}]}[g(\theta)]$ for all $k = 1, \ldots, m-1, m+1, \ldots, K$, using systematically the MCMC sample from $\pi(\theta|X^{(m)})$. We call this first strategy "fixed strategy" because m is fixed and unique over all IS estimations of $E_{[\theta|X^{(k)}]}[g(\theta)]$ for all $k \neq m$. In practice this means that it is enough to run MCMC only once to obtain K estimations of $E_{[\theta|X^{(k)}]}[g(\theta)]$ for $k = 1 \ldots, K$. Without loss of generality, m is fixed to 1.

Even if the data sets are simulated under the same model, considerable variability between samples could appear and thus posteriors may be remote. For this reason we propose a second strategy to improve the IS method by choosing a different importance function for each estimation.

2.2. Importance Sampling: modulated strategy

The basic idea is to use an MCMC algorithm on M preselected simulated data set ($M < K$) and to apply the IS procedure to remaining data where, for each data set, the importance function is one of the M preselected posterior distributions approximated by MCMC. We denote this second strategy by "modulated strategy." M has to be relatively small with regard to K in order to ensure a gain in computer time. In our examples, M is equal to 10 and K is equal to 100. The difficulty of this strategy is to establish relevant criteria for choosing the more adequate preselected posterior distributions for the IS procedure. We propose and compare three criteria, which are detailed further. More precisely, this second strategy can be formulated as follows.

Suppose that M simulated data $(X^{(m)}, m = 1, \ldots, M)$ have been preselected from all data sets and M Markov Chains $((\theta_1^{(m)}, \ldots \theta_N^{(m)}), m = 1, \ldots, M)$ have been produced by MCMC under $(\pi(\theta|X^{(m)}),\ m = 1, \ldots, M)$ respectively. We now want to approximate $E_{[\theta|X^{(k)}]}[g(\theta)]$ for $k = M + 1, \ldots, K$ via IS. In the modulated strategy, IS estimations are based on the choice of $m_k \in \{1, \ldots, M\}$, so that $\pi(\theta|X^{(m_k)})$ becomes the importance function which may change from one estimation to another within M possible choices. Following the same reasoning as before, the approximation of $E_{[\theta|X^{(k)}]}[g(\theta)]$ is then

$$\sum_{i=1}^{N} g(\theta_i^{(m_k)})\tilde{w}_i(k, m_k) \to E_{[\theta|X^{(k)}]}[g(\theta)] \tag{4}$$

but depends on the choice of m_k.

For $M > 1$, we propose three criteria for choosing $m_k \in \{1, \ldots, M\}$ for each $k = M + 1, \ldots, K$. The choice of importance function in the IS procedure is discussed in the literature with different solutions. For instance, solutions are proposed as an adaptation procedure of Cappé *et al.* (2007) where this function is based on minimization of deviance criterion or as in Douc *et al.* (2007b) via minimization of the asymptotic variance of IS. Let the IS estimation be as in equation (4). We expect that the IS method will provide a good estimation if $\pi(\theta|X^{(m_k)})$ is "close" to the target distribution $\pi(\theta|X^{(k)})$. It seems natural then to choose m_k for which the norm L^1 of the difference between these two posterior distributions $||\pi(\theta^{(m_k)}|X^{(k)}) - \pi(\theta^{(m_k)}|X^{(m_k)})||$ is the smallest. The first criterion corresponds to the smallest approximation of this norm and m_k is chosen, satisfying the following expression

$$\min_{m_k \in \{1,\ldots,M\}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\pi(\theta_i^{(m_k)}|X^{(k)})}{\pi(\theta_i^{(m_k)}|X^{(m_k)})} - 1 \right| \right\} \tag{5}$$

or, when the closed-form of posteriors is not available

$$\min_{m_k \in \{1,\ldots,M\}} \left\{ \sum_{i=1}^{N} \left| \tilde{w}_i(k, m_k) - \frac{1}{N} \right| \right\}. \tag{6}$$

A major concern in the context of our study is verifying the support condition to apply the importance sampling estimator as discussed in section 2.4. To handle this difficulty we may check a tail behavior of both target and sampling distributions, since we want the sampling distribution to be flatter. To control tail behaviors many authors suggest basing the choice of IS density on a minimization of a Kullback-Leibler divergence (see, for example, Asmussen, Kroese and Rubinstein (2005); Chen and Shao (1997); Rubinstein and Kroese (2004) and Gustafson and Wasserman (1995)). Following this concept, the second criterion may choose m_k corresponding to the IS distribution for which a Kullback-Leibler divergence $KL(\pi(\theta|X^{(K)}), \pi(\theta|X^{(m_k)})) = \int log(\frac{\pi(\theta|X^{(k)})}{\pi(\theta|X^{(m_k)})})\pi(\theta|X^{(k)})d\theta$ is the smallest:

$$\min_{m_k\in\{1,\ldots,M\}} \left\{ \frac{1}{N}\sum_{i=1}^{N} \frac{\pi(\theta_i^{(m_k)}|X^{(k)})}{\pi(\theta_i^{(m_k)}|X^{(m_k)})} \ln \frac{\pi(\theta_i^{(m_k)}|X^{(k)})}{\pi(\theta_i^{(m_k)}|X^{(m_k)})} \right\} \tag{7}$$

or, when the closed-form of posteriors is not available

$$\min_{m_k\in\{1,\ldots,M\}} \left\{ \sum_{i=1}^{N} \tilde{w}_i(k,m_k) \ln w_i(k,m_k) - \frac{\ln \frac{1}{N}\sum_{i=1}^{N} w_i(k,m_k)}{\frac{1}{N}\sum_{i=1}^{N} w_i(k,m_k)} \right\}. \tag{8}$$

The third criterion is based on the variance of $g(\theta)\frac{\pi(\theta|X^{(k)})}{\pi(\theta|X^{(m_k)})}$. We would like to choose $m_k \in \{1,\ldots,M\}$ satisfying the following inequality:

$$Var_{[\theta|X^{(m_k)}]}\left[\frac{g(\theta)\pi(\theta|X^{(k)})}{\pi(\theta|X^{(m_k)})}\right] < Var_{[\theta|X^{(k)}]}[g(\theta)] \tag{9}$$

corresponding to the variance of the IS estimate on the left and the variance of the MCMC estimate on the right.

As $E_{[\theta|X^{(m_k)}]}[\frac{g(\theta)\pi(\theta|X^{(k)})}{\pi(\theta|X^{(m_k)})}] = E_{[\theta|X^{(k)}]}[g(\theta)]$, it is then equivalent to:

$$E_{[\theta|X^{(m_k)}]}\left[\left(\frac{g(\theta)\pi(\theta|X^{(k)})}{\pi(\theta|X^{(m_k)})}\right)^2\right] < E_{[\theta|X^{(k)}]}[(g(\theta))^2] \tag{10}$$

As the choice of $\pi(\theta|X^{(m_k)})$ is limited to the M preselected data, it is not certain that the above inequalities will be satisfied for any of these densities. However, if $E_{[\theta|X^{(m_k)}]}[(\frac{g(\theta)\pi(\theta|X^{(k)})}{\pi(\theta|X^{(m_k)})})^2]$ is finite, we can find m_k for which the variance of the IS estimate is smallest. For this reason, for each $k = M+1,\ldots,K$, the second criterion selects $m_k \in \{1,\ldots,M\}$ such that the corresponding posterior density $\pi(\theta|X^{(m_k)})$ minimizes the variance of $g(\theta)\frac{\pi(\theta|X^{(k)})}{\pi(\theta|X^{(m_k)})}$.

It can be shown (see Robert, 2007, section 6.2.2) that to minimize this variance, $\pi(\theta|X^{(m_k)})$ should be proportional to $|g(\theta)| \cdot \pi(\theta|X^{(k)})$, therefore m_k is chosen to make $|g(\theta)|\frac{\pi(\theta|X^{(k)})}{\pi(\theta|X^{(m_k)})}$ the most stable. Thus, the third criterion chooses m_k which satisfies

$$\min_{m_k\in\{1,\ldots,M\}} \left\{ \frac{\max_{i=1,\ldots,N} u(\theta_i^{(m_k)}) - \min_{i=1,\ldots,N} u(\theta_i^{(m_k)})}{\sum_{i=1}^{N} u(\theta_i^{(m_k)})} \right\} \tag{11}$$

$$\text{where} \quad u(\theta_i^{(m_k)}) = |g(\theta_i^{(m_k)})| \cdot \pi(\theta_i^{(m_k)}|X^{(k)})/\pi(\theta_i^{(m_k)}|X^{(m_k)}) \quad \text{or}$$
$$u(\theta_i^{(m_k)}) = |g(\theta_i^{(m_k)})| \cdot \pi(X^{(k)}|\theta_i^{(m_k)})/\pi(X^{(m_k)}|\theta_i^{(m_k)})$$

when the posterior distributions cannot be calculated. Note that this third criterion depends on function g and thus must be calculated for each function g. This is not the case for two other previous criteria.

2.3. Comparison of different strategies

In a simulation study to judge the performance of one estimation method against another, we calculate the mean square error for each one, which by definition is the mean square difference between estimations and true expected value.

However, when there is no analytical expression of $E_{[\theta|X^{(k)}]}[g(\theta)]$, comparison with the true posterior mean is not possible. We therefore propose to compare the IS estimate directly with the MCMC estimate, as MCMC is the traditional procedure to use and can thus be considered as reference. Let us denote by $MCMC_{(k)}$ the estimate of $E_{[\theta|X^{(k)}]}[g(\theta)]$ obtained using the traditional MCMC procedure and by $IS_{(k)/(m)}$ the estimate obtained using IS, as in (3). We thus define the mean square errors as follows:

$$MSE_{fs} = \frac{1}{K-1} \sum_{k \neq m} (IS_{(k)/(m)} - MCMC_{(k)})^2 \tag{12}$$

for the "fixed strategy" (fs), and

$$MSE_{ms} = \frac{1}{K-M} \sum_{k \neq m_k} (IS_{(k)/(m_k)} - MCMC_{(k)})^2 \tag{13}$$

for the "modulated strategy" (ms). In this second case, the mean square errors will be evaluated for the three criteria, allowing us to assess the effect of the criteria on the IS estimates. In fact, this mean square error can be considered as a distance between the IS and MCMC methods.

As each procedure (MCMC or IS) corresponds to an approximation, we also calculate mean square errors of MCMC and IS procedures for the values fixed in the simulation scheme.

The choice of M preselected simulated data for the three criteria could appear arbitrary and not necessarily optimal. We therefore propose a procedure allowing "automatic" selection. This point will be described in section 4. Convergence of MCMC algorithms was checked systematically using different tools.

2.4. Importance Sampling performances

Generally speaking, the use of IS is justified if the support of the target posterior distribution is included in the support of the importance function. In our study, the target density is $\pi(\theta|X^{(k)})$ and the importance function is the posterior density $\pi(\theta|X^{(m)})$ for fixed strategy or $\pi(\theta|X^{(m_k)})$ for modulated strategy.

This problem of support is particularly delicate and crucial when the posterior distribution depends on parameter values. A simple example of this case is when data are assumed to be uniform on an interval determined by parameters which implies that parameter space of the posterior depends entirely on data. GLM models do not correspond to this situation. Apart from these cases of links between parameter space and data, the difficulty appears when, in any part of the support, the sampling distribution $\pi(\theta|X^{(m_k)})$ tends to zero faster than the target distribution $\pi(\theta|X^{(k)})$. Then, the weights involved in the IS procedure are very unstable with possible huge values. The accuracy of estimates obtained by IS depends a great deal on weight behavior. Hence, the support assumption may be ensured in a way by the control of the tail behavior of weights. Different tools exist in the literature to detect weight instability. For instance, Geweke (1989) proposed two diagnostics for computational accuracy: the first, based on order statistics of weights, is able to indicate thin tails in IS density relative to the posterior, and the second, relative numerical efficiency (RNE), is the ratio between the number of iterations used in IS and the number of iterations used in traditional MCMC to give the same numerical standard error. McVinish *et al.* (2008) proposed a statistic $\Delta(X^{(k)}, X^{(m_k)})$ to control the variance of weights when posterior distributions ($\pi(\theta|X^{(k)})$ and $\pi(\theta|X^{(m_k)})$) are approximated by Gaussian distributions. The greater the values of $\Delta(X^{(k)}, X^{(m_k)})$, the bigger the variability of weights. Under this gaussian approximation, we can show that the Kullback-Leibler divergence is proportional to this statistic $\Delta(X^{(k)}, X^{(m_k)})$. More precisely, with same notations as McVinish *et al.* (2008), $\pi(\theta|X^{(k)})$ and $\pi(\theta|X^{(m_k)})$ are approximated by Gaussian centered at $\hat{\theta}_k$ and $\hat{\theta}_{m_k}$ respectively with asymptotic covariances matrices J_k and J_{m_k}. If $J_k^{-1} = nI_0(1 + o_p(1))$ and $J_{m_k}^{-1} = nI_0(1+o_p(1))$, I_0 being a fixed positive definite matrix, the authors show that the asymptotic variance of weights is:

$$var(\tilde{w}(k, m_k)|X^{(k)}, X^{(m_k)}) = exp\{n(\hat{\theta}_{m_k} - \hat{\theta}_k)^T I_0(\hat{\theta}_{m_k} - \hat{\theta}_k) + o_p(1)\} - 1$$

The statistic $\Delta(X^{(k)}, X^{(m_k)}) = (\hat{\theta}_{m_k} - \hat{\theta}_k)^T J_{m_k}^{-1}(\hat{\theta}_{m_k} - \hat{\theta}_k)$ can thus predict the stability of the weights. The Kullback-Leibler divergence $KL(\pi(\theta|X^{(k)}), \pi(\theta|X^{(m_k)}))$ is:

$$\begin{aligned} &KL(\pi(\theta|X^{(k)}), \pi(\theta|X^{(m_k)}) \\ &\quad = 1/2\left(log\left(\frac{|J_{m_k}|}{|J_k|} \right) + tr(J_{m_k}^{-1} J_k) + (\hat{\theta}_{m_k} - \hat{\theta}_k)^t J_{m_k}^{-1}(\hat{\theta}_{m_k} - \hat{\theta}_k) + r \right) \end{aligned}$$

where r is the dimension of parameter space. If we assume that $J_k = J_{m_k}$ (sample sizes of $X^{(k)}$ and $X^{(m_k)}$ are implicitly assumed to be equal) then:

$$KL(\pi(\theta|X^{(k)}), \pi(\theta|X^{(m_k)}) = 1/2((\hat{\theta}_{m_k} - \hat{\theta}_k)^t J_{m_k}^{-1}(\hat{\theta}_{m_k} - \hat{\theta}_k))$$

$$KL(\pi(\theta|X^{(k)}), \pi(\theta|X^{(m_k)}) = 1/2\Delta(X^{(k)}, X^{(m_k)})$$

which is exactly the statistic controling the variability of weights in McVinish *et al.* (2008). If $X^{(k)}$ and $X^{(m_k)}$ are mutually independent and if $\hat{\theta}_k$ and $\hat{\theta}_{m_k}$

are maximum likelihood estimators then $\Delta(X^{(k)}, X^{(m)}) \xrightarrow{d} 2\chi^2_r$. This remark goes in favor of the choice of Kullback-Leibler divergence as a criterion for modulated strategy because it minimizes the statistic $\Delta(X^{(k)}, X^{(m_k)})$. A critical situation for modulated strategies would be when these computational accuracy diagnostics give poor indications for all M preselected data sets. Note that this situation occurs very rarely especially because, as mentioned by Asmussen, Kroese and Rubinstein (2005), the importance function is chosen from among the same parametric family as the target density, as happens in our case. If this critical situation exists for a specific $\pi(\theta|X^{(k)})$, one solution is to add the data $X^{(k)}$ in the preselected data set (containing then $M + 1$ elements); this can be done without any difficulty.

3. Applications

In this section we use both the MCMC (Gibbs sampling in our case) and IS methods to estimate parameters of three Poisson models. The first is a Poisson model with one parameter (the mean), the second is a Poisson regression on one covariate with two parameters (intercept and covariate association), and the third is a Poisson regression on one covariate with extra Poisson variability introduced by a Gaussian residual error term with three parameters (intercept, covariate association and residual variance). The first model can be seen as a toy example with explicit posterior distributions; the second corresponds to a widely used GLM model, and the third introduces over-dispersion which is essential, for example, in medical applications since association estimates would be biased if extra-Poisson variability was not modelled (see Breslow (1984) for motivations). For each model $K = 101$ data sets are simulated for different values of the parameters. All data sets contain $n = 20$ observations. Vague priors are assigned to the parameters and the posterior values are estimated via MCMC and IS as discussed above. Note that it is essential that MCMC convergence is achieved, therefore several (and not only one) diagnostics of convergence have to be checked as suggested by Brooks and Roberts (1998) and Mengersen, Robert and Guihenneuc-Jouyaux (1999). Many diagnostics are available in the BRuGS package CODA, namely convergence diagnostics of Geweke, Gelman and Rubin, Raftery-Lewis, Heidelberger and Welch. In our examples, we use MC error (MC error was less then 5% of the posterior standard deviation) and Gelman and Rubin diagnostics as well as graphical tools (history, autocorrelation). For all estimating values we ran an MCMC algorithm for $50,000$ iterations with a burn-in of $5,000$.

Firstly, we study performance of the fixed strategy with $m = 1$ over all IS estimations. Then we generalize the concept of the fixed strategy by evaluating mean square errors for all possible values of m ($m = 1, \ldots, 101$). This means that the fixed strategy is repeated 101 times, and each time only one posterior distribution is taken as an importance function in all estimations. However, any time the fixed strategy is repeated, this posterior distribution, which is used as an importance function, depends on another data set. This approach allows us

to assess the best and the worst performance of this strategy in terms of mean square error. We also draw box-plots to summarize all these mean square errors.

For the second strategy, we preselect the first 10 simulated data ($M = 10$) and use them next in criteria. For each criterion we calculate corresponding mean square errors, which are then reported on the box-plots mentioned above.

Finally, we compare the results with those obtained when: (1) the number of observations increases to $n = 1,000$ and (2) the number of coefficients increases by introducing 10 covariates into the regression. All the calculations were done using R software environment for statistical computing and graphics (R Development Core Team, 2008). The BRugs package (Thomas *et al.*, 2006) was used in MCMC simulations. Note that for storage reasons, a subset of MCMC iterations (thin equal to 20) is used for Model 3 with $n = 1,000$ for all procedures.

3.1. Simple Poisson model

As a tool example we consider that the data are described by a simple Poisson model with one parameter λ, that is $X_1, \ldots, X_n$ are iid and $X_i|\lambda \sim \mathcal{P}(\lambda)$. We use $\lambda \sim \mathcal{G}(\alpha, \beta)$ as a prior distribution and then the true posterior distribution of the parameter given the whole data set $X = (X_1, \ldots, X_n)$ is a $\mathcal{G}(\sum_{i=1}^{n} X_i + \alpha, \beta + n)$. To simulate data sets of size $n = 20$ each, we use two values for the model parameter $\lambda = 1$ and $\lambda = 20$. In both cases the parameters of the prior are $\alpha = 0.01$ and $\beta = 0.01$.

The mean square errors for the fixed strategy and the modulated strategy for $g(\lambda) = \lambda$ are presented in figure 1 with $\lambda = 1$ on the left and $\lambda = 20$ on the right. Black, white and grey diamonds correspond to the first, second and the third criteria respectively for the modulated strategy. The box-plots give results for the fixed strategy when all possibilities of fixed m are considered. Box-plots without ouliers (extreme MSEs) are presented in the top right corner of each graphic. We observe that mean square errors are greater for $\lambda = 20$ as data variability is

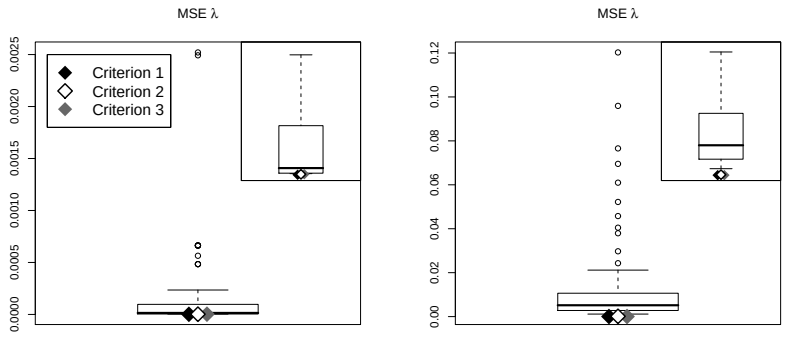

FIG 1. *Box-plot of the MSE_{fs} of all possible m in fixed strategy together with the MSE_{ms} of three criteria in Model 1 with $g(\lambda) = \lambda$ for $\lambda = 1$ (left) and $\lambda = 20$ (right); black diamond for the first criterion, white diamond for the second criterion and grey diamond for the third.*

greater. However, in both cases of lambda, modulated strategies (first to third criteria) seem to reduce the mean square errors compared with the MSE_{fs} for fixed strategy. As it is impossible to choose with certainty one good sample distribution for the fixed strategy, the results show that thanks to the criteria, we can avoid those m (fixed strategy) for which associated MSE is greatest.

The advantage of this simple example is that the true posterior mean of λ is known, equal to $(\sum_{i=1}^{n} X_i^{(k)} + \alpha)/(\beta + n)$ for simulated data $X^{(k)}$. For each approximation procedure (MCMC and IS), table 1 presents MSEs with regard to the true posterior mean for $\lambda = 1$ on the left and $\lambda = 20$ on the right.

Generally these MSEs are small for all strategies studied (notice that values reported in table 1 are multiplied by 10^6). For $\lambda = 1$ MSEs calculated for MCMC are smaller than for other strategies, however for $\lambda = 20$ MSEs calculated for the first and the third criteria are smaller than those calculated for MCMC. Three modulated strategies have MSEs sharply smaller than the biggest MSE (the worst case) of the fixed strategy. An advantage of modulated strategies is that the estimations are on average closer to the true posterior means than are estimations calculated with fixed strategy. For the second criterion based on Kullback-Leibler divergence, numerical problems can appear due to the log of weigths. For $\lambda = 20$ and second criterion, MSE is equal to 124.95 10^{-6} due to two extreme values. If these values are omitted (and then the mean is based on 89 values instead on 91 values), MSE is equal to 53.3 10^{-6}, which is comparable to the other MSEs obtained from modulated strategies. The mean square errors calculated with respect to the true values set in simulations (results not shown) show similar performances for the modulated strategies and MCMC and sometimes, as in table 1, slightly better results are obtained from modulated strategies than from the MCMC procedure.

TABLE 1
Mean square errors with regard to the true posterior mean in Model 1 for $g(\lambda) = \lambda$ with $\lambda = 1$ (left) and $\lambda = 20$ (right) and $n = 20$

strategy	$\lambda = 1$		$\lambda = 20$	
	MSE[1]	$q_{97.5\%}$[2]	MSE[1]	$q_{97.5\%}$[2]
MCMC [3]	1.050	0.035	90.327	1.194
The best fixed strategy[4]	3.481	0.071	1152.119	90.131
The worst fixed strategy[5]	2520.477	83.316	120260.480	10325.171
Modulated strategy with 1st criterion[6]	1.253	0.087	42.127	0.669
Modulated strategy with 2nd criterion[6]	2.148	0.172	124.950	3.958
Modulated strategy with 3rd criterion[6]	2.306	0.131	44.615	0.671

[1]Mean square errors with regard to the true posterior mean($\times 10^6$)
[2]Highest 97.5% of the terms involved in mean square error expression ($\times 10^6$)
[3]MCMC strategy ($\times 10^6$)
[4]IS with fixed strategy associated to the best MSE ($\times 10^6$)
[5]IS with fixed strategy associated to the worst MSE ($\times 10^6$)
[6]IS with modulated strategy ($\times 10^6$)

3.2. Poisson regression model

The second model is a Poisson regression with a covariate Z and two parameters a and b, that is $X_i|\lambda \sim \mathcal{P}(\lambda_i)$ where $log(\lambda_i) = a + bZ_i$. For both coefficients a and b we use $\mathcal{N}(0, 10^5)$ as vague priors, and data sets are simulated when $a = 0$ and $b = 0.5$. For this model, no analytical solution of posterior parameter distribution is available. Table 2 shows the mean square errors for the best and for the worst choice of the sampling distribution with fixed strategy, and also when the sampling distribution changes according to the three criteria. These results are illustrated by the box plots of all mean square errors (figure 2).

TABLE 2
Mean square errors with regard to the MCMC posterior means in Model 2 with $a = 0$, $b = 0.5$ and $n = 20$ for $g(\theta) = a$ (left)and $g(\theta) = b$ (left)

strategy	a		b	
	MSE[1]	$q_{97.5\%}$[2]	MSE[1]	$q_{97.5\%}$[2]
The best fixed strategy[3]	0.042	0.003	0.046	0.003
The worst fixed strategy[4]	15.886	2.930	37.664	2.047
Modulated strategy with 1st criterion[5]	0.007	<0.001	0.014	0.001
Modulated strategy with 2nd criterion[5]	0.058	0.070	0.155	0.020
Modulated strategy with 3rd criterion[5]	0.026	0.001	0.594	0.001

[1]Mean square errors on (K-M) samples ($\times 10^3$)
[2]Highest 97.5% of the terms involved in mean square error expression ($\times 10^3$)
[3]IS with fixed strategy associated to the best MSE ($\times 10^3$)
[4]IS with fixed strategy associated to the worst MSE ($\times 10^3$)
[5]IS with modulated strategy ($\times 10^3$)

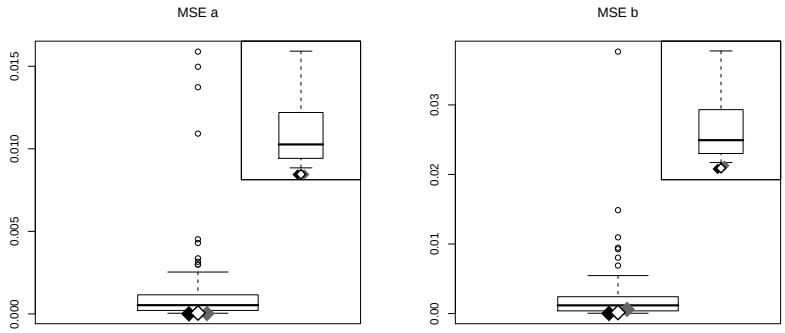

FIG 2. *Box-plot of the MSEs of all possible fixed choices together with the MSE_cs of both criteria in Model 2 with $\theta = (a = 0, b = 0.5)$ for $g(\theta) = a$ (top left) and $g(\theta) = a^2$ (top right), $g(\theta) = b$ (bottom left) and $g(\theta) = b^2$ (bottom right); black diamond for the first criterion, white diamond for the second criterion and grey diamond for the third.*

Again, both strategies provide better estimations in comparison with most of the estimations for which fixed choice of sampling distribution was used. Indeed, we can ascertain from figure 2 that the MSEs for criteria are smaller than the quasi-totality of the MSEs for the fixed choice.

3.3. Poisson regression model with extravariability

The third model gives an example of extra-variability. Poisson regression on one covariate Z with extra-variability is supposed. The model is then the following:

$$X_i|\lambda \sim \mathcal{P}(\lambda_i)$$

$$log(\lambda_i) = a + bZ_i + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

For both coefficients a and b we use $\mathcal{N}(0, 10^5)$ as vague priors, and inverse gamma distribution $\mathcal{IG}(0.01, 0.01)$ (the second coefficient being the rate) as a prior of residual variance σ^2. In the simulations we set $a = 0$ and $b = 0.5$ and we consider three parameter settings of σ^2: 1/8, 1/4, 1/2. There is no closed-form expression for the posterior parameter distribution. For each strategy and each parameter setting, mean square errors (MSE_{fs} and MSE_{ms}) with regard to MCMC posterior means are presented in table 3 as well as the highest 97.5% of the terms involved in mean square error expression.

Overall, the modulated strategy always gives better results than the worst fixed strategy for the three parameters, but sometimes the best fixed strategy is better. The impact of criterion choice for the modulated strategy is minor, with results being almost the same. Figure 3 presents box-plots for the fixed strategy when all possibilities of fixed m are considered. As previously, black, white and grey diamonds correspond to the first, second and third criteria respectively for the modulated strategy. From left to right, the results are given for the three values of σ^2 (1/8, 1/4, 1/2) respectively. These box-plots clearly confirm that the modulated strategy avoids the worst cases of fixed strategy and show that in most cases, fixed strategy leads to greater MSEs even if it is less clear that modulated strategy performances are better for the parameter σ^2.

As the MCMC procedure corresponds to an approximation, mean square errors of MCMC and of the two IS strategies are assessed with respect to the "true" parameter values (results not shown). Modulated strategies lead again to good results, sometimes better than those from the MCMC procedure. For instance, in the case of $\sigma^2 = 1/2$, MSEs of coefficient b for three criteria of modulated strategy are equal to 67×10^{-3}, 64×10^{-3} and 58×10^{-3} respectively, while the MSE of b from MCMC is 83×10^{-3}.

In all three models, there are cases where fixed strategy corresponds to smaller MSE_{fs} than MSE_{ms} and also where it is smaller than the MSE of MCMC. Nevertheless, while comparing posterior estimations with the values set in simulations, modulated strategy again shows better results than fixed strategy except in a small number of cases and avoids the worst case of the fixed strategy.

TABLE 3
Mean square errors with regard to the MCMC posterior means in Model 3 with $a = 0$, $b = 0.5$, $\sigma^2 = 1/8$ (first part) or $\sigma^2 = 1/4$ (second part) or $\sigma^2 = 1/2$ (third part) and $n = 20$ for $g(\theta) = a$ (left), $g(\theta) = b$ (middle) and $g(\theta) = \sigma^2$ (left)

strategy	a		b		σ^2	
$\sigma^2 = 1/8$	MSE[1]	$q_{97.5\%}$[2]	MSE[1]	$q_{97.5\%}$[2]	MSE[1]	$q_{97.5\%}$[2]
The best fixed strategy[3]	5.791	0.385	0.261	0.222	51.079	4.224
The worst fixed strategy[4]	66.384	3.928	31.762	2.814	132.263	8.763
Modulated strategy with 1st criterion[5]	7.058	0.579	3.674	0.338	74.385	5.527
Modulated strategy with 2nd criterion[5]	12.325	1.725	4.125	0.344	89.494	5.963
Modulated strategy with 3rd criterion[5]	12.122	1.311	5.156	0.564	95.190	6.340
	a		b		σ^2	
$\sigma^2 = 1/4$	MSE[1]	$q_{97.5\%}$[2]	MSE[1]	$q_{97.5\%}$[2]	MSE[1]	$q_{97.5\%}$[2]
The best fixed strategy[3]	8.531	1.075	3.453	0.381	72.822	4.434
The worst fixed strategy[4]	74.240	10.928	44.829	3.199	149.402	12.262
Modulated strategy with 1st criterion[5]	14.282	1.672	5.498	0.568	110.520	9.642
Modulated strategy with 2nd criterion[5]	17.309	2.562	9.282	0.861	110.328	9.902
Modulated strategy with 3rd criterion[5]	15.785	1.831	7.183	0.568	122.447	12.324
	a		b		σ^2	
$\sigma^2 = 1/2$	MSE[1]	$q_{97.5\%}$[2]	MSE[1]	$q_{97.5\%}$[2]	MSE[1]	$q_{97.5\%}$[2]
The best fixed strategy[3]	17.827	2.053	4.859	0.396	234.222	18.205
The worst fixed strategy[4]	181.103	8.557	138.765	5.907	361.212	33.731
Modulated strategy with 1st criterion[5]	29.918	2.782	6.398	0.645	287.724	31.109
Modulated strategy with 2nd criterion[5]	40.138	3.519	10.198	0.804	280.986	31.820
Modulated strategy with 3rd criterion[5]	30.869	2.389	9.941	0.761	297.203	27.430

[1]Mean square errors on (K-M) samples ($\times 10^3$)
[2]Highest 97.5% of the terms involved in mean square error expression ($\times 10^3$)
[3]IS with fixed strategy associated to the best MSE ($\times 10^3$)
[4]IS with fixed strategy associated to the worst MSE ($\times 10^3$)
[5]IS with modulated strategy ($\times 10^3$)

3.4. Sensitivity analysis

As an extension, three other cases are considered, in each of which we increase the number n of observations per data set to $n = 1,000$. The first case (case 1) is again the Poisson regression model with extra-variability as previously described in section 3.3. The same parameter values are chosen ($a = 0$, $b = 0.5$, $\sigma^2 = 1/2$) and the same priors are taken. Two other cases of Poisson

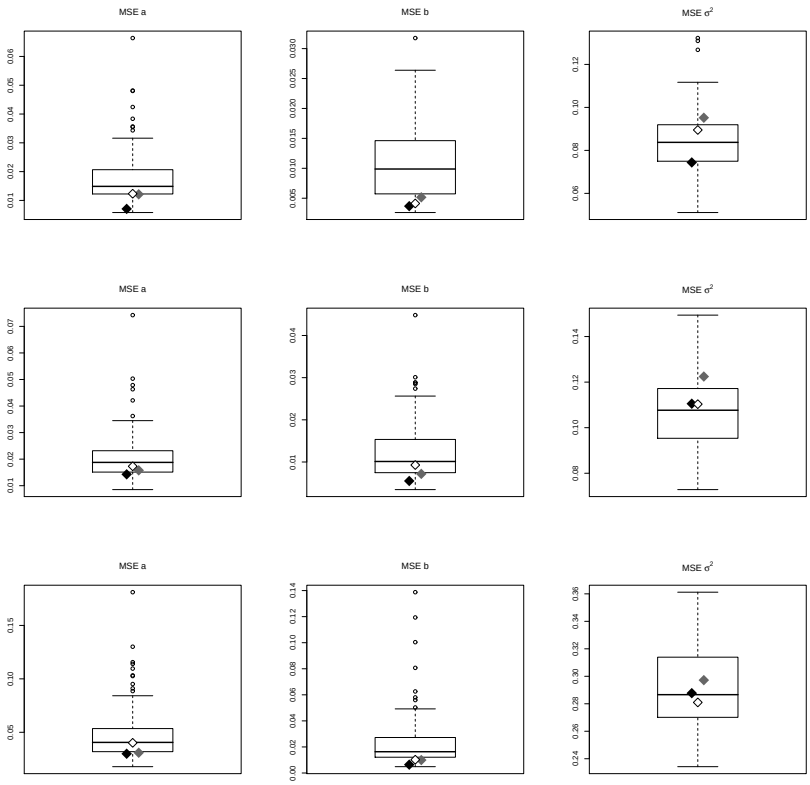

FIG 3. *Box-plot of the MSE_{fs} of all possible fixed choices together with the MSE_{ms} of both criteria in Model 3 with $a = 0$, $b = 0.5$ and $\sigma^2 = 1/8$ (1^{st} line), $\sigma^2 = 1/4$ (2^{nd} line) or $\sigma^2 = 1/2$ (3^{rd} line), for $g(\theta) = a$ (left) and $g(\theta) = b$ (middle), $g(\theta) = \sigma^2$ (right); black diamond for the first criterion, white diamond for the second criterion and grey diamond for the third criterion in modulated strategy.*

regression models are studied where the number of covariates is now 10 which are all continuous (normally distributed from $\mathcal{N}(0,1)$) (case 2) or which are continuous or binary ($Z_{ij} \sim \mathcal{N}(0,1)$ for $j = 1,\ldots,5$ and $Z_{ij} \sim Bernoulli(0.25)$ for $j = 6,\ldots,10$) (case 3). For these last two cases, the linear predictor becomes $log(\lambda_i) = a + \sum_{j=1}^{10} b_j Z_{ij} + \epsilon_i$.

We use $\mathcal{N}(0, 10^5)$ as vague priors for coefficients a and $\{b_j, j = 1,\ldots,10\}$, and inverse gamma distribution $\mathcal{IG}(0.01, 0.01)$ as prior of residual variance σ^2 of ϵ_i. In simulations we set $a = 0$, $b_j = 0.05$ for $j = 1,\ldots,10$ and $\sigma^2 = 1/2$.

For each case, mean square errors are calculated with regard to MCMC posterior means. For these three cases, similar performances are obtained for modulated strategies to those in section 3.3. The increase in the number of observations n leads to smaller MSEs, as expected. Table 4 gives these MSEs for case 1 and has to be compared with the last line of table 3 corresponding to a similar

TABLE 4
Mean square errors in Model 3 with $a = 0$, $b = 0.5$, $\sigma^2 = 1/2$ and $n = 1000$ for $g(\theta) = a$ (left), $g(\theta) = b$ (middle) and $g(\theta) = \sigma^2$ (left) calculated with regard to MCMC posterior means

strategy	a		b		σ^2	
$\sigma^2 = 1/2$	MSE[1]	$q_{97.5\%}$[2]	MSE[1]	$q_{97.5\%}$[2]	MSE[1]	$q_{97.5\%}$[2]
The best fixed strategy[3]	1.580	0.077	1.247	0.067	2.600	0.126
The worst fixed strategy[4]	23.382	0.520	9.124	0.266	52.230	1.083
Modulated strategy with 1st criterion[5]	3.486	0.144	2.131	0.097	14.806	0.506
Modulated strategy with 2nd criterion[5]	8.019	0.308	3.079	0.139	32.478	0.890
Modulated strategy with 3rd criterion[5]	2.776	0.169	2.043	0.083	16.514	0.602

[1]Mean square errors on (K-M) samples ($\times 10^3$)
[2]Highest 97.5% of the terms involved in mean square error expression ($\times 10^3$)
[3]IS with fixed strategy associated to the best MSE ($\times 10^3$)
[4]IS with fixed strategy associated to the worst MSE ($\times 10^3$)
[5]IS with modulated strategy ($\times 10^3$)

case but with $n = 20$. MSEs are approximately divided by 3 with $n = 1,000$ but the interclassification is preserved. An increase in the number of covariates gives similar results. Concerning case 2, mean of 10 MSEs corresponding to 10 covariate associations is calculated for each strategy. The three modulated strategies give very close results (2.4×10^{-3}, 2.9×10^{-3} and 2.3×10^{-3} for criteria 1, 2 and 3 respectively), MSE means for the best and the worst fixed strategy being 1.2×10^{-3} and 11.7×10^{-3} respectively. As before, the result for the worst fixed strategy is considerably less good than for modulated strategies but slightly better for the best fixed strategy. Concerning case 3, results for associations with normal covariates are separated from those with Bernoulli covariates. For the 5 normal covariates, results are completely similar to the previous case. For the 5 Bernoulli covariates, MSE means for the three modulated strategies are 13.6×10^{-3}, 15.0×10^{-3} and 12.5×10^{-3} for strategies 1, 2, and 3 and 6.1×10^{-3}, 66.4×10^{-3} for the best and worst fixed strategy respectively. Less good MSEs are approximately multiplied by 6 but conclusions are similar.

An increase in n and in the number of covariates clearly increases computation times. Improving procedures in terms of computation times is then a major challenge, particularly in high dimension. Comparisons of these times for the different procedures are presented and discussed in the last section (section 5).

4. Selection procedure

In this section, we propose a procedure to select M sampling distributions useful for the modulated strategies. The basic idea is to find M posterior distributions which are, in a certain sense, representative of the $K - M$ remaining posterior distributions. As in our examples vague priors are chosen, this is almost

equivalent to considering sampling distributions instead of posterior distributions. First, we define this selection procedure and then illustrate it on Model 3 defined previously.

4.1. Selection method

The basic idea of this automatic selection method is to build M clusters from K sampling distributions and then to select one distribution in each cluster to become the M preselected simulated data set used in section 2.2. The cluster constructions are based on a summary statistic characterizing a sampling distribution and on a distance between these statistics. Different choices of statistic and distance are possible, but for simplicity reasons, we choose the empirical mean as summary statistic and the Euclidean distance. The most suitable methods here are "k-medoids" methods which partition all elements into M clusters returning its central element for each cluster, the medoid. Since we want to select M sampling distributions representative of the K sampling distributions, the advantage of such a method is that a "medoid" of each cluster is obtained directly. Several clustering methods exist and some are implemented in R packages. To illustrate this procedure, we choose the "Partitioning Around Medoids" (PAM) method introduced by Kaufman and Rousseeuw (1990) and well-documented in Nakache and Confais (2005). The R procedure is "PAM" in the Cluster package (Maechler *et al.*, 2005). Note that other "k-medoid" methods exist, such as Clustering LARge Applications (Kaufman and Rousseeuw, 1990), Clustering LARge Applications based on RANdomized Search (Ng and Han, 1994) and Fast Intelligent Subspace Clustering Algorithm using Dimension Voting (Woo *et al.*, 2004).

4.2. Illustration

The selection procedure used in the third model was as defined in section 3.3 with the parameter settings $a = 0$, $b = 0.5$ and $\sigma^2 = 1/2$. The number of simulated data sets is $K = 101$ and of preselected data sets $M = 10$ as before. Figure 4 represents estimated marginal posterior distributions for each parameter approximated by classical MCMC, where bold densities correspond to the 10 data sets preselected by the automatic procedure PAM. This figure shows clearly that the 10 posterior distributions of selected data sets represent all the posterior existing distributions well.

Table 5 presents the mean square errors when IS with modulated strategies is used with the 10 simulated data sets selected automatically using the above procedure. The reader can check that these results are comparable to those given in table 3. These results are again always better than those obtained with the worst fixed choice of table 3. The selection procedure seems to give results which are as good as previous ones and has the great advantage of being automatic and therefore avoiding the arbitrary choice of preselected sampling distributions.

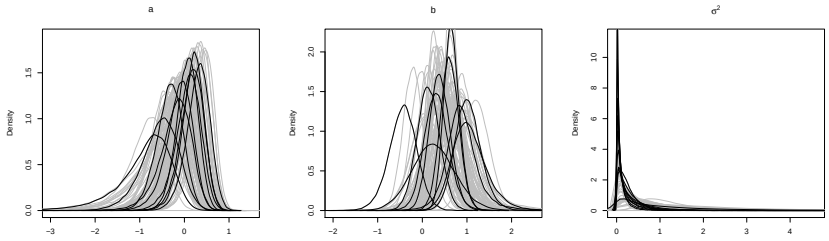

FIG 4. *Marginal posterior densities of a (left), b (middle) and σ^2 (right) obtained with MCMC in Model 3 with parameter settings $\theta = (a = 0, b = 0.5, \sigma^2 = 1/2)$. Ten posterior densities selected by PAM clustering algorithm, are drawn in bold.*

TABLE 5

Mean square errors in Model 3 with the Automatic Selection Procedure and parameter settings $\theta = (a = 0, b = 0.5, \sigma^2 = 1/2)$ for $g(\theta) = a$ (left part), $g(\theta) = b$ (middle part) and $g(\theta) = \sigma^2$ (right part)

strategy	a		b		σ^2	
	MSE[1]	$q_{97.5\%}$[2]	MSE[1]	$q_{97.5\%}$[2]	MSE[1]	$q_{97.5\%}$[2]
Modulated strategy with 1st criterion[3]	25.784	1.790	5.073	0.337	259.049	27.550
Modulated strategy with 2nd criterion[3]	40.005	3.577	13.133	0.805	279.517	35.416
Modulated strategy with 3rd criterion[3]	32.721	3.098	6.307	0.449	292.787	30.407

[1−2]See the legend of Table 4

[3]IS with the different $\pi(\theta|X^{(m)})$ selected by the 1st or the 2nd criterion ($\times 10^3$)

5. Conclusions

In this paper, we compare a classical MCMC algorithm with MCMC combined with an IS approach in repeated posterior estimations conditional on different data sets. The aim is to run MCMC on only a small number of data sets and then use generated chains in IS procedure in the following estimations conditional on the remaining data sets. The sampling distribution, namely a chain, can be the same in each IS estimation ("fixed strategy") or differ one from another according to one of the three proposed criteria ("modulated strategy"). We compare the results of the approaches using IS with those based on MCMC only, using mean square errors. When the true posterior means are available we compare estimations with them. Comparison becomes impossible since posterior expectation values are unknown. To overcome this difficulty we try to compare the estimation strategies involving IS directly with MCMC posterior mean estimations, considering MCMC estimation as a benchmark. The disadvantage of this comparison, however, is that both IS and MCMC are only approximate. In the studied examples, we chose vague priors on parameters. Hence, we also calculated mean square error with regard to the values set on parameters in the simulations.

The methods discussed were applied to the Poisson models. The results of IS estimation were quite satisfying in the case of the fixed strategy and the modulated strategy. The first model was the simplest and its true posterior characteristics were available. In this example we saw that IS estimations were very close to them and also to the MCMC estimations. For the other two models, analytical estimates were not available, and therefore the IS results were compared with the MCMC estimations showing concordance between the two methods. For Model 3, it emerges that a residual variance is more difficult to estimate and the MCMC algorithm does not handle it well either, returning estimates with large credibility intervals. In this case, larger MSEs are not surprising. For the modulated strategy, neither criterion has managed to produce mean square errors smaller than the smallest MSEs obtained for the fixed strategy, but nevertheless, they are always smaller than the largest ones and in most cases smaller than the fixed strategy MSEs. As an extension of this study, we have tested IS strategies on the same Poisson regression, but increasing the number of observations per data set to $n = 1,000$. To increase the number of parameters, we have also added more covariates. They were chosen arbitrarily to be normal or normal and Bernoulli. The main conclusions remain unchanged. The modulated strategy allows us to avoid the worst case of the fixed strategy for both types of mean square error. When comparing estimations with the values set in simulations, IS may outperform MCMC, as in the best case of the fixed strategy. Unfortunately it is impossible to indicate conditionally the data set in which a posterior distribution is the best importance function in the fixed choice strategy to give the smallest corresponding MSE_{fs}.

In terms of computation times, modulated strategies and fixed strategies always run faster than the MCMC procedure for complex models as Model 3. The gain depends on the number of observations n and the number of parameters. Generally speaking, the higher the dimension of the parameter space and/or the number of observations, the greater the time difference between the methods. All comparisons were carried out on the same computer and on the same number of iterations in the MCMC procedure. Table 6 gives the ratio of computation times between the IS and MCMC methods for Models 2 and 3 with $n = 1,000$ and ten covariates. For the Poisson model with extra-variability (Model 3), the ratio between fixed strategy and MCMC computation times is on average 0.036, the ratio between modulated strategy and MCMC is 0.278, 0.266 and 0.652 for criteria 1, 2 and 3 respectively. The benefit in terms of computation times between classic MCMC and MCMC together with IS is thus considerable. For Poisson

TABLE 6
Ratio of computation times between the IS strategies and MCMC methods with $n = 1,000$ and 10 covariates

Strategy:	fixed	criterion 1 (L_1)	criterion 2 (KL)	criterion 3
Model 2	0.057	0.957	0.773	1.156
Model 3	0.036	0.278	0.266	0.652

model without extra-variability (Model 2), the ratio between fixed strategy and MCMC computation times is on average 0.057 showing again a clear benefit of IS procedure. But, the gain between MCMC and IS with modulated strategies is less important. Even if ratio is smaller than 1 for criteria 1 and 2, note that classic MCMC is faster than IS with criterion 3 because this last criterion needs to be assessed for each function g as discussed in section 2.2. For simple models, MCMC runs very fast and so, the use of alternative approaches as IS seems less crucial than for models requiring expensive computation times as Model 3.

To conclude, modulated strategies, especially when associated with the first or second criterion, show the best compromise between estimate performances and computation times. Indeed, the estimate performances are nearly equivalent for modulated strategies and MCMC, but with computation times that are significantly smaller. Moreover, the comparison with fixed strategy revealed better performances for the modulated strategy. Interesting extensions of this work could be to study IS estimates for generalized linear mixed models (GLMM) which are widely used in practice. These models appear typically to require computation time improvement because they correspond to a high dimensional case. From a methodological point of view, IS based on a mixture of preselected posterior distributions versus only a single distribution is an important perspective. It offers the advantage that it probably proposes an importance function with greater support range, and hence more stable weights.

Acknowledgments

The authors would like to thank Dr McVinish for helpful and constructive comments which enabled us to improve this article. The authors thank the editor, associate editor and the referee for their helpful comments and suggestions which led to a significant improvement. Dr Gajda's research is supported by university Paris Sud 11 - ED420 doctoral grant.

References

Asmussen, S., Kroese, D. P. and Rubinstein, R. Y. (2005). Heavy tails, importance sampling and cross-entropy. *Stoch. Models* **21** 57–76. MR2124359

Breslow, N. E. (1984). Extra-Poisson Variation in Log-Linear Models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **33** 38–44.

Brooks, S. P. and Roberts, G. O. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing* **8** 319–335. http://dx.doi.org/10.1023/A:1008820505350

Cappé, O., Guillin, A., Marin, J. M. and Robert, C. P. (2004). Population Monte Carlo. *J. Comput. Graph. Statist.* **13** 907–929. MR2109057

Cappé, O., Douc, R., Guillin, A., Marin, J. M. and Robert, C. P. (2007). Adaptive Importance Sampling in General Mixture Classes. *Statistics and Computing (to appear)*. Available at http://www.citebase.org/abstract?id=oai:arXiv.org:0710.4242

Chen, M.-H. and Shao, Q.-M. (1997). Performance study of marginal posterior density estimation via Kullback-Leibler divergence. *Test* **6** 321–350. MR1616900

Doss, H. (1994). Discussion of the paper "Markov chains for exploring posterior distributions" by Luke Tierney. *Ann. Statist.* **22** 1728–1734.

Douc, R., Guillin, A., Marin, J.-M. and Robert, C. P. (2007a). Convergence of adaptive mixtures of importance sampling schemes. *Ann. Statist.* **35** 420–448. MR2332281

Douc, R., Guillin, A., Marin, J.-M. and Robert, C. P. (2007b). Minimum variance importance sampling via population Monte Carlo. *ESAIM Probab. Stat.* **11** 427–447 (electronic). MR2339302

Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian statistics, 4 (Peñíscola, 1991)* 147–167. Oxford Univ. Press, New York. MR1380275

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–740.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57** 1317–1339. MR1035115

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Statist. Soc. Ser. B* **54** 657–699. With discussion and a reply by the authors. MR1185217

Gilks, W., Richardson, S. and Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice. Interdisciplinary Statistics*. Chapman & Hall, London. Edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter. MR1397966

Gustafson, P. and Wasserman, L. (1995). Local sensitivity diagnostics for Bayesian inference. *Ann. Statist.* **23** 2153–2167. MR1389870

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57** 97–109.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. John Wiley & Sons Inc., New York. MR1044997

Maechler, M., Rousseeuw, P., Struyf, A. and Hubert, M. (2005). Cluster Analysis Basics and Extensions. Rousseeuw et al provided the S original which has been ported to R by Kurt Hornik and has since been enhanced by Martin Maechler: speed improvements, silhouette() functionality, bug fixes, etc. See the 'Changelog' file (in the package source).

McVinish, R., Mengersen, K., Nur, D. C., Rousseau, J. and Guihenneuc-Jouyaux, C. (2008). Use of Importance Sampling for Repeated MCMC. School of Mathematical Sciences, Queensland University of Technology.

Mengersen, K. L., Robert, C. P. and Guihenneuc-Jouyaux, C. (1999). MCMC convergence diagnostics: a reviewww. In *Bayesian statistics, 6 (Alcoceber, 1998)* 415–440. Oxford Univ. Press, New York. MR1723507

Nakache, J. P. and Confais, J. (2005). *Approche pragmatique de la classification*. Technip.

Ng, R. and Han, J. (1994). Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th Conference on VLDB, Santiago, Chili* 144–155.

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. Available at http://www.R-project.org

Robert, C. P. (2007). *The Bayesian choice*, Second ed. *Springer Texts in Statistics*. Springer-Verlag, New York. From decision-theoretic foundations to computational implementation, Translated and revised from the French original by the author. MR1835885

Rubinstein, R. Y. and Kroese, D. P. (2004). *The cross-entropy method: A unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning. Information Science and Statistics*. Springer-Verlag, New York. MR2080985

Thomas, A., O'Hara, B., Ligges, U. and Sturtz, S. (2006). Making BUGS Open. *R News* **6** 12–17. Available at http://cran.r-project.org/doc/Rnews/

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22** 1701–1762. With discussion and a rejoinder by the author. MR1329166

Woo, K. G., Lee, J. H., Kim, M. H. and Lee, Y. I. (2004). FINDIT: a fast and intelligent subspace clustering algorithm using dimension voting. *Informations & Software Technology* **46** 255–271.