Statistics and Computing (2001) **11,** 103–115

Riemann sums for MCMC estimation and convergence monitoring

ANNE PHILIPPE* and CHRISTIAN P. ROBERT†

*Laboratoire de Statistique et Probabilités, EP CNRS 1765 UFR Mathématiques Bât M2, Université de LILLE I, 59655 Villeneuve d'Ascq, France

†Laboratoire de Statistique, CREST, Insee, Timbre J340, 92245 Malakoff cedex, France

Received July 1999 and accepted January 2000

This paper develops an extension of the Riemann sum techniques of Philippe (J. Statist. Comput. Simul. 59: 295–314) in the setup of MCMC algorithms. It shows that these techniques apply equally well to the output of these algorithms, with similar speeds of convergence which improve upon the regular estimator. The restriction on the dimension associated with Riemann sums can furthermore be overcome by Rao–Blackwellization methods. This approach can also be used as a control variate technique in convergence assessment of MCMC algorithms, either by comparing the values of alternative versions of Riemann sums, which estimate the same quantity, or by using genuine control variate, that is, functions with known expectations, which are available in full generality for constants and scores.

Keywords: simulation, numerical integration, control variate, Rao–Blackwellization, score

1. Introduction

Simulation techniques and, in particular, MCMC methods, often aim at approximating integrals of the form

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x) f(x)\, dx \tag{1}$$

for integrable functions h of interest. Once a sample $(x_1, \ldots, x_n)$ is produced, be it from f or from another distribution, independent or not, the approximation of $\mathbb{E}[h(X)]$ can proceed in many ways, some of which are vastly superior to the empirical average

$$\delta_n^E = n^{-1} \sum_{i=1}^{n} h(x_i), \tag{2}$$

which is nonetheless the favorite in most Monte Carlo studies. While a genuine "Decision Theory for simulation" is difficult to conceive, as it would have to take into account many factors (like programming, debugging, running time and such), a coherent position, from a statistical point of view, is to require a use of the sample which is as optimal as possible (see Casella 1996). The notion of "optimality" is obviously difficult to define, given that, contrary to the usual statistical setting, the value of interest, $\mathbb{E}[h(X)]$, is known exactly if formally. But it does make sense to prefer the approach which, for (approximately) the same computing effort, produces estimators with smaller (mean square) errors.

For instance, when the sample is (independently) generated from a density g, the importance sampling reweighting of the average (2) leads to an unbiased estimator whose variance may, at least formally, decrease down to 0 (see Rubinstein 1981). But the variance of the importance sampling estimator can also be infinite for poor choices of g. As already demonstrated in Philippe (1997a, b), Riemann sums can produce a considerable improvement over standard averages like (2) in some cases and we will show in Section 2 that this area of improvement extends to many MCMC settings. The Riemann sum estimator is built on a merge of analytic and simulation techniques, by considering the sum

$$\delta_T^R = \sum_{i=1}^{T-1} \left(x^{[i+1]} - x^{[i]}\right) h\left(x^{[i]}\right) f\left(x^{[i]}\right), \tag{3}$$

where $x^{[1]} \leq \cdots \leq x^{[T]}$ is the ordered sample associated with $(x_1, \ldots, x_T)$, which can be generated from the density f or from a proposal density g. The finiteness of the variance of δ_T^R is ensured by the same conditions as for the empirical average, namely $\mathbb{E}[h^2(X)] < \infty$. Moreover, when both h and its derivative h' are bounded functions, the rate of convergence of the variance is in $O(T^{-2})$ (see Philippe 1997b, a), compared with the

order $O(T^{-1})$ of the usual average. As demonstrated in Philippe (1997b), as well as in the following sections, this improvement in the speed of convergence is far from formal, since it can be clearly observed on the cumulated sum graphs.

As in importance sampling estimator, requires the density f needs to be known, at least up to a constant, and this is not always the case. A stronger impediment to the use of (3) is that it is seemingly restricted to unidimensional samples: naive extensions to larger dimensions suffer from the "curse of dimensionality" plaguing numerical methods (see Robert and Casella 1999). We show in Section 3 that extensions of (3) are available for both MCMC and multidimensional settings, by using Rao–Blackwellization (see Gelfand and Smith 1990) to both approximate marginal densities and reduce the dimension.

An important issue in using MCMC methods is to ascertain the convergence of the simulated Markov chain to the distribution of interest and, if possible, to come up with stopping rules or convergence diagnostics pertaining to this goal. Convergence diagnostics are based on many features of the MCMC samples, characterising different aspects of the MCMC sample (see Cowles and Carlin 1996, Brooks and Roberts 1999, Mengersen et al. 1999) but an important role is played by *control variates* in that they provide landmarks in the exploration of the distribution f and give lower bounds on the simulation time. We show in Section 4 that Riemann sums can be used as such, by providing evaluations of the mass of the stationary distribution explored by the Markov chain at iteration T, as in Brooks (1998). They also lead to control variates based on various score functions.

2. Riemann sums for MCMC outputs

Before getting to the extension of (3), let us recall how and why the Riemann sum estimators (3) can be used in MCMC setups, mainly through a few examples, given that the theory of order statistics on Markov chains is hardly developed. We thus consider a unidimensional setting where a density f can be approximated via an MCMC algorithm. The estimator (3) can then be reproduced in terms of the Markov chain $(x^{(t)})$, that is, at a given iteration T, the sequence $x^{(1)}, \ldots, x^{(T)}$ can be ordered into $x^{[1]} \leq \cdots \leq x^{[T]}$ to produce the *Riemann estimator* of $\mathbb{E}^f[h(X)]$,

$$\delta_T^h = \sum_{t=1}^{T-1} \left(x^{[t+1]} - x^{[t]}\right) h\left(x^{[t]}\right) f\left(x^{[t]}\right). \tag{4}$$

That (4) converges to $\mathbb{E}^f[h(X)]$ follows from Philippe (1997b) by asymptotic arguments on the convergence of the spacings $(x^{[t+1]} - x^{[t]})$, when $(x^{(t)})$ is ergodic, that is, truly converging to the stationary distribution f.

The on-line computation of (4) thus requires ranking the current value $x^{(T)}$ of the Markov chain within the ordered sequence, $x^{(T)} = x^{[i]}$ say, and updating the estimator δ_{T-1}^h as

$$\delta_T^h = \delta_{T-1}^h + \left(x^{[i+1]} - x^{[i]}\right)\left\{h\left(x^{[i]}\right) f\left(x^{[i]}\right) - h\left(x^{[i-1]}\right) f\left(x^{[i-1]}\right)\right\}.$$

Alternatively, if, for convergence assessment purposes, δ_T^h is only used at iterations $T_1, \ldots, T_k, \ldots$, the update of δ_T^h can be done for these iterations only, with a computing time still of order $O(T_i)$ if the values $x^{(T_i+1)}, \ldots, x^{(T_{i+1})}$ are ranked separately.

Since, in most MCMC settings, f is only known up to a multiplicative factor, $f(x) \propto \tilde{f}(x)$, the alternative representation

$$\frac{\sum_{t=1}^{T-1} \left(x^{[t+1]} - x^{[t]}\right) h\left(x^{[t]}\right) \tilde{f}\left(x^{[t]}\right)}{\sum_{t=1}^{T-1} \left(x^{[t+1]} - x^{[t]}\right) \tilde{f}\left(x^{[t]}\right)} \tag{5}$$

will be used, rather than (4), with obviously the same convergence properties. Note that the denominator provides an estimate of the normalization constant of $\tilde{f}$, which comes as an alternative to the methods proposed in Geyer (1993) and Chen and Shao (1997).

The main difference with the iid case is that, due to the absence of theoretical results on the speed of convergence of the spacings $(x^{[t+1]} - x^{[t]})$ and on the order of $\operatorname{var}(x^{[t+1]} - x^{[t]})$ and $\operatorname{cov}(x^{[t+1]} - x^{[t]}, x^{[t'+1]} - x^{[t']})$ for general Markov chains, the convergence rate on the variance of (4) cannot be evaluated. We will however check in Example (2) that this rate is still approximately T^{-2} when h and h' are bounded. This is not surprising, given that the spacings are asymptotically independent and thus recover most of the properties of their iid counterparts.

Example 1. Consider the density

$$f_0(x) \propto \frac{e^{-x^2/2}}{(1 + (x - x_0)^2)^{\nu}}, \tag{6}$$

motivated in Robert (1995) as the posterior distribution of a Cauchy scale parameter, which can be represented as the marginal density of

$$g(x, y) \propto e^{-x^2/2} y^{\nu - 1} e^{-(1 + (x - x_0)^2) y/2},$$

with the following conditional distributions

$$X \mid y \sim \mathcal{N}(x_0 y/(1 + y), 1/(1 + y))$$
$$Y \mid x \sim \mathcal{G}a(\nu, (1 + (x - x_0)^2)/2).$$

The corresponding Gibbs sampler is therefore straightforward to implement.

Figure 1(a) illustrates the convergences of the Riemann estimator and the empirical average for the estimation of $\mathbb{E}^{f_0}[X]$ based on a given sequence of $x^{(i)}$'s. The faster convergence of the Riemann estimator, when compared with the empirical average, is clearly visible on this graph and the Riemann estimator is close to the true value after only a few hundred iterations. Since this comparison is based on a single sequence, we also present a comparison based on a Monte Carlo experiment for 2000 replications of such sequences. We construct a 95% equal-sided confidence band for both the empirical average and the Riemann estimator which is represented on Fig. 1(b). The improvement brought by (4) is magnified on this graph, with a range of variation much smaller from the start.

In order to compare the behavior of both estimators in a Markovian setting, we can also take advantage of the fact that it

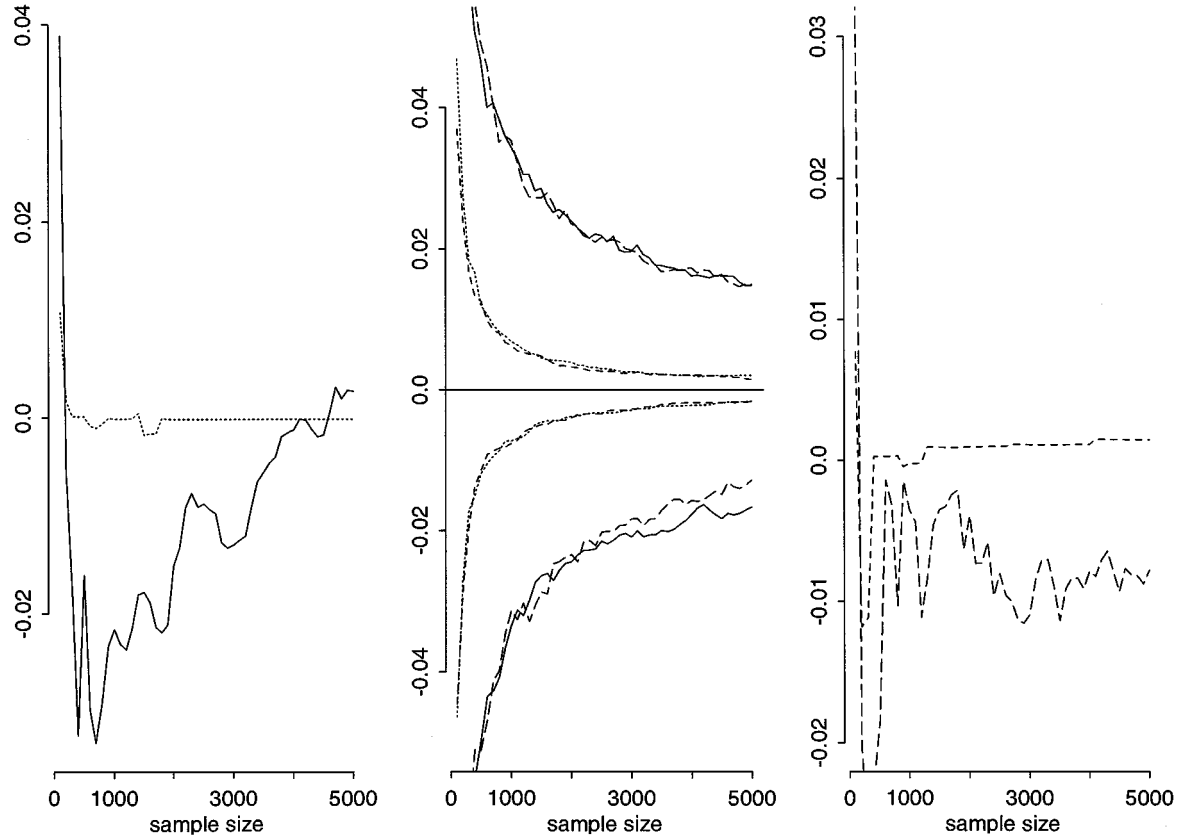

Fig. 1. *(a) Successive values of the Riemann estimator [dots] and of the empirical average [full] associated with the Gibbs sampler for the estimation of* $\mathbb{E}^{f_0}[X]$ *in Example 1 with* $x_0 = 0$ *and* $\nu = 2$*. The true value is 0 and the convergence is represented for a given MCMC sample. (b) Monte Carlo comparison of the 95% confidence bands for the Riemann estimators associated with the Gibbs sampler [dots] and with* $[A_1]$ *(see below) [dashes], and for the empirical average associated with the Gibbs sampler [long dashes] and with* $[A_1]$ *[full]. (c) Same graph as (a) for a sample simulated via* $[A_1]$ *and the Riemann estimator [dashes] and the empirical average [long dashes]*

is possible to generate directly, that is, independently, from (6). Indeed, an accept-reject algorithm based on a Gaussian proposal distribution can be derived as follows:

---

1. Generate $x \sim \mathcal{N}(0, 1)$ and $u \sim \mathcal{U}([0, 1])$
2. If $u \leq (1 + (x - x_0)^2)^{-\nu}$ take x $\quad [A_1]$
   else go to 1

Figure 1(c) plots the convergence to 0 of both the empirical average and the Riemann estimator when based on a (single) sample simulated from $[A_1]$ and it shows the considerable improvement brought by (4). Figure 1(b) also compares the performance of both estimators when based on an iid sample from (6) and it shows that both estimators have very close behavior in terms of convergence rate for both the iid and the dependent samplings. The amplitudes of the confidence regions are equal and therefore the loss of performance due to independence has no strong influence on the variance of the Riemann estimators.

Example 2. We now consider an auto-exponential model (see Besag 1974), with $Y = (Y_1, Y_2) \in \mathbb{R}_+^2$ distributed according to the density

$$g(y_1, y_2) \propto \exp(-y_1 - y_2 - y_1 y_2)\mathbb{I}_{\mathbb{R}_+^2}(y_1, y_2).$$

The conditional distributions are available and given by

$$Y_1 \mid y_2 \sim \mathcal{E}xp(1 + y_2)$$
$$Y_2 \mid y_1 \sim \mathcal{E}xp(1 + y_1),$$

and thus induce a corresponding Gibbs sampler.

Since the marginal density of Y_1 is available and equal to

$$g_1(y_1) \propto \frac{e^{-y_1}}{1 + y_1}\mathbb{I}_{\mathbb{R}_+}(y_1), \tag{7}$$

the Riemann estimator (5) can be used to estimate integrals of the form

$$\int g_1(y_1)h(y_1)\,dy_1,$$

that is, integrals that bear only on the first component of Y.

Similarly to the previous example, the alternative to generate iid variables from g_1 through an accept-reject algorithm is also available and allows for a comparison of the performances of

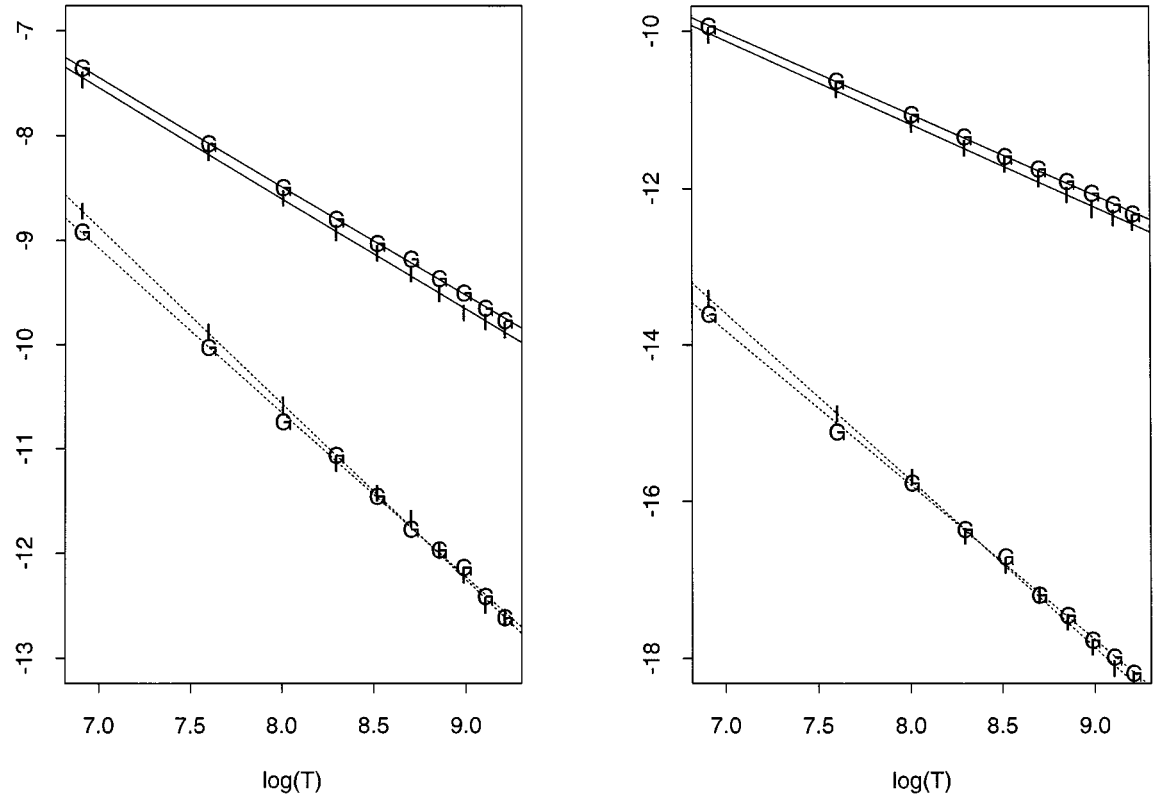

Fig. 2. *Representation of log(var(δ_T)) against log(T) for the estimation of $\mathbb{E}[h_i(Y_1)]$, by δ_T^E (full) and δ_T^R (dots) based on a Gibbs sample (***G***) and an iid sample produced by accept-reject (***I***), for the expectations of $h_1(x) = x$ (left) and $h_2(x) = 1/(1 + x)$ (right) in model (7) (based on 2000 independent replications)*

both estimators in the dependent and independent cases. (The proposal distribution is then exponential.)

Figure 2 represents the evolution of the couples $(\log(T), \log[\text{var}(\delta_T)])$ for 10 values of T (where T is the sample size), for the expectations of the functions $h_1(x) = x$ and $h_2(x) = 1/(1 + x)$; it exhibits an almost perfect linear fit. We can thus conclude that the opposition between the $O(T^{-2})$ and $O(T^{-1})$ rates derived in the independent case extends to the Markov case, given that the lines hardly differ for both cases. The estimates of the slopes α, that is, of the rate of convergence $O(T^{-\alpha})$, are given in Table 1 and indeed show very little difference between the independent and the Markov cases. For the approximation of $\mathbb{E}[h_2(X)]$, the estimate of α is close to 2, which is in line with the result of Philippe (1997b) of a convergence rate in $O(T^{-2})$ for bounded functions with bounded derivative.

Table 1. *Estimation of the decrease rate α of the variance for the empirical average (δ^E) and the Riemann estimator (δ^R), based on an iid sample (I) and a Gibbs sample (G), for the estimation of $\mathbb{E}[h_i(Y_1)]$ in model (7), when $h_1(x) = x$ and $h_2(x) = 1/(1 + x)$ (based on 2000 independent replications)*

	δ^E(I)	δ^E(G)	δ^R(I)	δ^R(G)
h_1	1.05	1.03	1.69	1.57
h_2	1.05	1.03	2.13	1.97

3. Rao–Blackwellized Riemann sums

3.1. *Data augmentation model*

We now consider the more general setup where $Y = (X, Z) \in \mathbb{R} \times \mathbb{R}^{p-1}$ is distributed from a known density g and $\mathbb{E}^f[h(X)]$ is the quantity of interest. While $\mathbb{E}^f[h(X)]$ can be formally written as (1), the marginal density f is usually unknown and the Riemann estimator (3), which depends explicitly on f is not available. We now consider an alternative, based on the Rao–Blackwell estimator of f, assuming that the full conditional densities are known in closed form (up to a constant). In particular, $\pi(x \mid z)$ denotes the full conditional density of X given $Z = z$.

Rao–Blackwellization has been promoted in Gelfand and Smith (1990) and Casella and Robert (1996) as a mean to reduce the variance of estimators of integrals, but the improvement is rarely substantial in practice (see Robert 1998). Another facet of Rao–Blackwellization is to come up with a parametric estimator of marginal densities, thus providing a significant improvement when compared with standard non-parametric estimates. For instance the Rao–Blackwell estimator of the marginal density

based on the Markov chain $(x^{(t)}, z^{(t)})$ is

$$\hat{f}(x) = T^{-1} \sum_{t=1}^{T} \pi\big(x \mid z^{(t)}\big),$$

which converges to $f(x)$. A natural extension of the Riemann estimator (4) is thus to replace the density f by its estimation $\hat{f}$. The corresponding Rao–Blackwellized Riemann sum estimator is thus

$$\begin{aligned}\delta_T^{R/\mathrm{RB}} &= \sum_{t=1}^{T-1} \big(x^{[t+1]} - x^{[t]}\big) h\big(x^{[t]}\big) \hat{f}\big(x^{[t]}\big) \\ &= T^{-1} \sum_{t=1}^{T-1} \big(x^{[t+1]} - x^{[t]}\big) h\big(x^{[t]}\big) \left(\sum_{k=1}^{T} \pi\big(x^{[t]} \mid z^{(k)}\big) \right).\end{aligned} \tag{8}$$

Since $\hat{f}$ converges to f, this estimator is also convergent. However, the finer convergence properties of the estimator are quite difficult to fathom, compared with (3), given the interdependence between $\hat{f}$ and the $x^{(t)}$'s. In setups where $\pi(x \mid z)$ is only known up to a constant, (8) is replaced by the ratio

$$\frac{\sum_{t=1}^{T-1} \big(x^{[t+1]} - x^{[t]}\big) h\big(x^{[t]}\big) \big(\sum_{k=1}^{T} \pi\big(x^{[t]} \mid z^{(k)}\big) \big)}{\sum_{t=1}^{T-1} \big(x^{[t+1]} - x^{[t]}\big) \big(\sum_{k=1}^{T} \pi\big(x^{[t]} \mid z^{(k)}\big) \big)}.$$

The update of the estimator at iteration T is then of order T. Indeed, if $(x^{(T+1)}, z^{(T+1)})$ is the output of the MCMC algorithm at iteration $T+1$, and if $x^{(T+1)} = x^{[i_0]}$,

$$\begin{aligned}\delta_{T+1} = {} & \frac{T}{T+1} \delta_T \\ & + \frac{1}{T+1} \sum_{t} \big(x^{[t+1]} - x^{[t]}\big) h\big(x^{[t]}\big) \pi\big(x^{[t]} \mid z_{i_0}\big) \\ & + \frac{x^{[i_0+1]} - x^{[i_0]}}{T+1} \sum_{t \neq i_0} \big\{ h\big(x^{[i_0]}\big) \pi\big(x^{[i_0]} \mid z^{(t)}\big) \\ & - h\big(x^{[i_0-1]}\big) \pi\big(x^{[i_0-1]} \mid z^{(t)}\big) \big\}\end{aligned}$$

To evaluate the effect of the estimation on the variance, we first consider the same examples as in the previous section, since we can compare the performance of the estimator $\delta_T^{R/RB}$ with the original Riemann estimator. Figure 3 illustrates the comparison between the Riemann estimators and $\delta_T^{R/RB}$ for Examples 1 and 2. In both cases, the strong similarity between both curves shows that the influence of replacing f with $\hat{f}$ is minimal in these cases.

3.2. *Multidimensional extensions*

So far, the Riemann sum estimator has only been defined for unidimensional chains $(x^{(t)})$. While multidimensional extensions are formally available, they suffer from the "curse of dimensionality", with larger variances than the empirical average when integrating in spaces of dimensions 4 and more

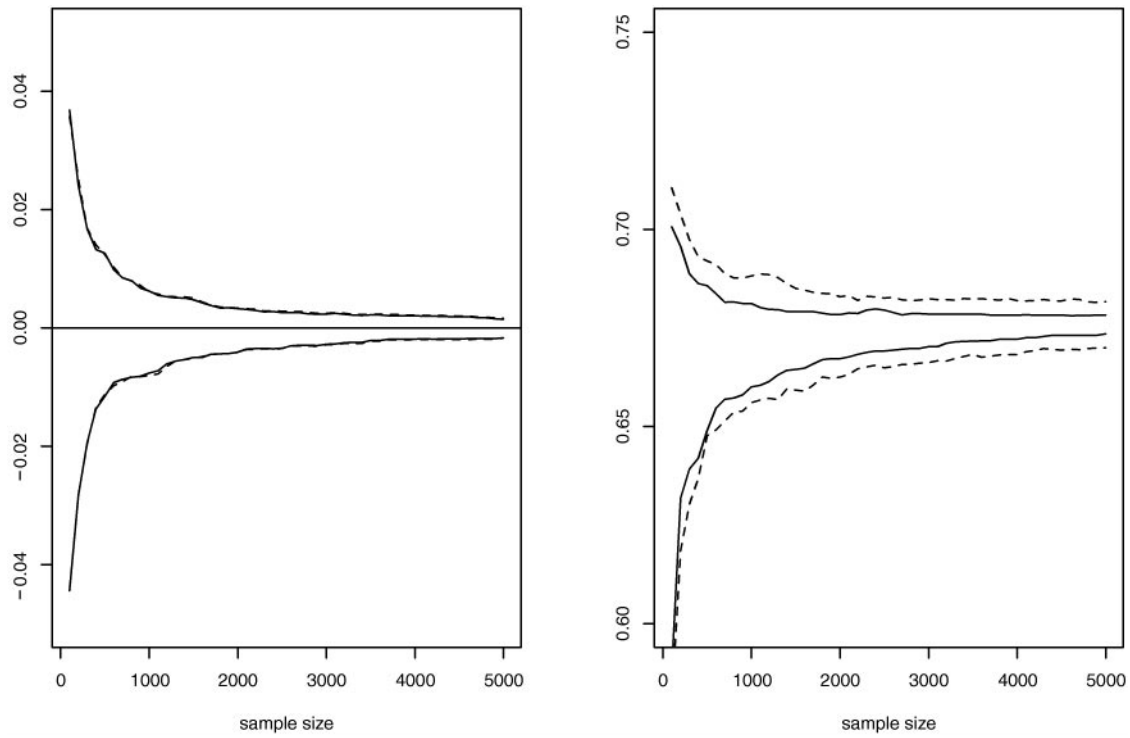

Fig. 3. *(left) The 95% confidence bands of the Riemann (full) and Rao–Blackwellized Riemann (dashes) sum estimators of $\mathbb{E}^{f_0}$ [X] in model (6) (Both curves overlap and cannot be distinguished on the graph.) and (right) The same comparison for the auto exponential model. (Both graphs are based on 2000 independent replications.)*

(see Philippe 1997b). Nonetheless, it is possible to extend the Rao–Blackwellized Riemann sum estimator to approximate multidimensional integrals. The decomposition of the integral on which this extension is based is ($l = 1, \ldots, p$)

$$\mathbb{E}^f[h(X)] = \int_{\mathbb{R}} \int_{\mathbb{R}^{p-1}} h(x)\pi_l(x_l \mid x_{-l})\pi^l(x_{-l})\, dx_{-l}dx_l, \qquad (9)$$

where $x_{-l} = (x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_p)$ and

$$f(x) = \pi_l(x_l \mid x_{-l})\pi^l(x_{-l}).$$

The expectation $\mathbb{E}^f[h(X)]$ thus appears as a unidimensional integral of the function

$$\varphi(x_l) = \int_{\mathbb{R}^{p-1}} h(x_{-l}, x_l)\pi_l(x_l \mid x_{-l})\pi^l(x_{-l})\, dx_{-l}$$

and the original Riemann sum estimator could apply if φ was known. Since it is not available in closed form in general, it can be estimated by a Rao–Blackwellized approximation, that is,

$$\hat{\varphi}(x_l) = \frac{1}{T}\sum_{k=1}^{T} h\big(x_{-l}^{(k)}, x_l\big)\pi_l\big(x_l \mid x_{-l}^{(k)}\big),$$

which converges to $\varphi(x_l)$.

This decomposition thus allows for the elimination of the multidimensional integration problem and reduces the integration to an unidimensional problem, namely the integration of $\varphi(x_l)$ on $\mathbb{R}$. The Riemann sum approximation (4) applies and leads to

$$\delta_T^l = T^{-1}\sum_{t=1}^{T-1}\big(x_l^{[t+1]} - x_l^{[t]}\big)\left\{\sum_{k=1}^{T} h\big(x_l^{[t]}, x_{-l}^{(k)}\big)\pi_l\big(x_l^{[t]} \mid x_{-l}^{(k)}\big)\right\} \qquad (10)$$

as an estimator of $\mathbb{E}^f[h(X)]$. Note that the fact that the $x_l^{[t]}$'s are marginally distributed from

$$f^l(x_l) = \int f(x)\, dx_{-l}$$

in (10) does not appear in the estimator δ_T^l, because, contrary to the standard importance sampling estimator, the importance ratio does not appear in a Riemann sum estimator (see Philippe 1997a).

Obviously, the superior performances of the original Riemann sum estimator (3) are not preserved, because the speed of convergence is now dictated by the speed of convergence of the Rao–Blackwellized estimator of the conditional density, $\hat{\varphi}$. Once again, finer convergence properties of (10) are difficult to assess, that is, further than the mere convergence of δ_T^l to $\mathbb{E}^f[h(X)]$.

3.3. *Multiple estimates*

A feature worth noticing is that (10) depends on the choice of the component l in the decomposition (9). Therefore, when all the conditional densities $\pi_l(x_l \mid x_{-l})$ ($l = 1, \ldots, p$) are available, this approach produces p convergent estimators of the integral $\mathbb{E}^f[h(X)]$, which are all based on the same Markov chain. A first implication of this multiplicity of estimates is to identify the fastest component, that is, the one with the highest rate of convergence, directly by comparison of the convergence graphs, in order to reduce the computing time. In fact, the convergence rate of the variance of δ_T^l clearly depends on the choice of the coordinate l. More particularly, Philippe (1997a) shows that the convergence properties of the Riemann sums strongly depends on the function h in the iid case. For a fixed l, we have

$$\begin{aligned}\int_{\mathbb{R}} h(x)f(x)\, dx &= \int_{\mathbb{R}} \int_{\mathbb{R}^{p-1}} h(x)\pi_l(x_l \mid x_{-l})\pi^l(x_{-l})\, dx_{-l}\, dx_l \\ &= \int_{\mathbb{R}} \frac{\int_{\mathbb{R}^{p-1}} h(x)\pi_l(x_l \mid x_{-l})\pi^l(x_{-l})\, dx_{-l}}{f^l(x_l)} \\ &\quad \times f^l(x_l)\, dx_l,\end{aligned}$$

which shows that δ_T^l is also the evaluation of the expectation of

$$\begin{aligned}\psi_l(x_l) &= \frac{\int_{\mathbb{R}^{p-1}} h(x)\pi_l(x_l \mid x_{-l})\pi^l(x_{-l})\, dx_{-l}}{f^l(x_l)} \\ &= \int_{\mathbb{R}^{p-1}} h(x)\pi(x_{-l} \mid x_l)\, dx_{-l}\end{aligned}$$

against the marginal f^l. Therefore if we can choose the component l in such a way that the function ψ_l is bounded, we are under conditions which ensure a faster convergence of the Riemann estimator. This choice of the coordinate is then related to the selection of the conditional $\pi(x_{-l} \mid x_l)$ with the lighter tails. Note that, for $h(x) = h(x_1)$ and $l = 1$, the function ψ_1 is equal to h and δ_T^l is the estimator $\delta_T^{R/RB}$.

Example 3. Consider the model introduced by Gaver and O'Muircheartaigh (1987) which was originally proposed for the analysis of failures of nuclear pumps, with the dataset given in Table 2, and is now used as a benchmark in the MCMC literature (see Robert and Casella 1999). The failures of the i-th pump are modeled according to a Poisson process with parameter λ_i ($1 \le i \le 10$). For an observation time t_i, the number of failures p_i is a Poisson $\mathcal{P}(\lambda_i t_i)$ random variable. For the prior

Table 2. *Number of failures and observation times for ten nuclear pumps (Source: Gaver and O'Muircheartaigh (1987)*

Pump	1	2	3	4	5	6	7	8	9	10
Failures	5	1	5	14	3	19	1	1	4	22
Time	94.32	15.72	62.88	125.76	5.24	31.44	1.05	1.05	2.10	10.48

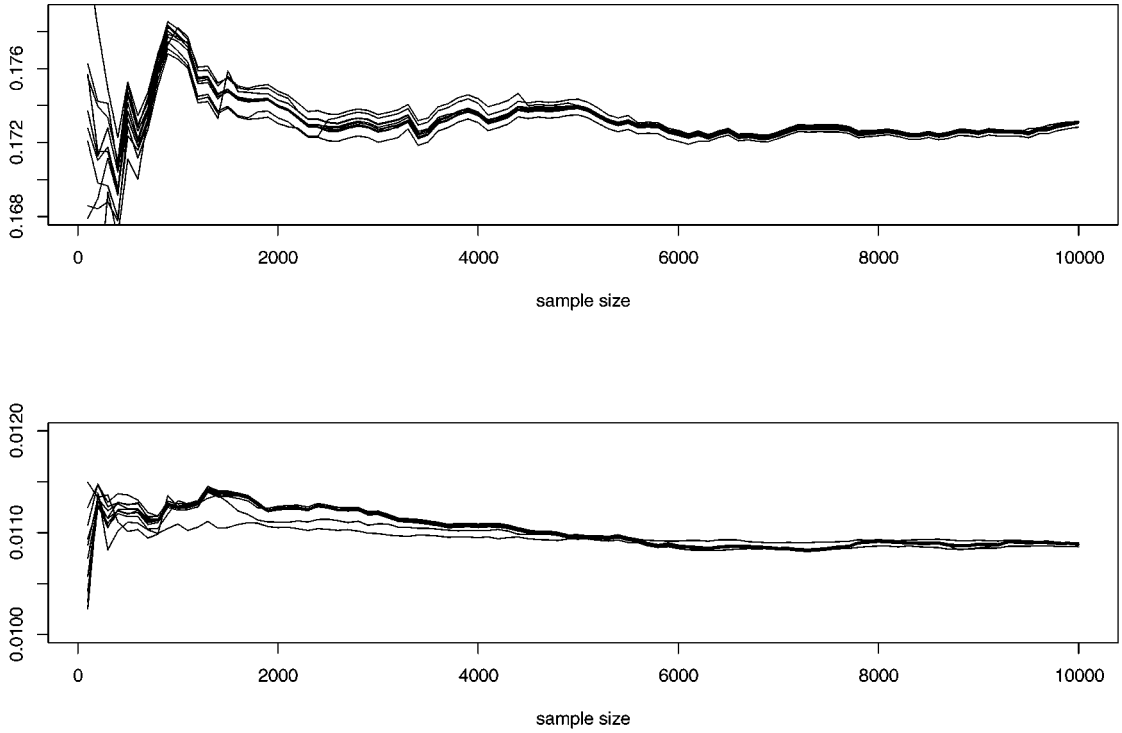

Fig. 4. *Convergence paths for the different eleven Rao–Blackwellized Riemann sum estimates of $\mathbb{E}[\beta\lambda_1]$ (top) and of $\mathbb{E}[\lambda_1\lambda_2]$ (bottom) for the pump failure example*

distributions

$$\lambda_i \overset{iid}{\sim} \mathcal{G}a(\alpha, \beta), \qquad \beta \sim \mathcal{G}a(\gamma, \delta) \quad (1 \le i \le 10),$$

with $\alpha = 1.8$, $\gamma = 0.01$ and $\delta = 1$, the joint distribution is

$$\pi(\lambda_1, \ldots, \lambda_{10}, \beta \mid t_1, \ldots, t_{10}, p_1, \ldots, p_{10}) \propto \prod_{i=1}^{10} \left\{\lambda_i^{p_i+\alpha-1} e^{-(t_i+\beta)\lambda_i}\right\} \beta^{10\alpha+\gamma-1} e^{-\delta\beta}$$

and the corresponding full conditionals are

$$\lambda_i \mid \beta, t_i, p_i \sim \mathcal{G}a(p_i + \alpha, t_i + \beta), \quad (1 \le i \le 10)$$

$$\beta \mid \lambda_1, \ldots, \lambda_{10} \sim \mathcal{G}a\left(\gamma + 10\alpha, \delta + \sum_{t=1}^{10} \lambda_i\right).$$

Each of these eleven posterior conditional distributions correspond to a different Rao–Blackwellized Riemann sum estimate (10). Figure 4 represents the convergence of these different estimates for the approximation of $\mathbb{E}[\beta\lambda_1]$ and $\mathbb{E}[\lambda_1\lambda_2]$.

Since the full posterior distributions are completely known, we can now construct eleven different Rao–Blackwellized Riemann sum estimates of $\mathbb{E}[\beta\lambda_1]$ and $\mathbb{E}[\lambda_1\lambda_2]$. Figure 4 illustrates the convergences of these different estimates. It shows similar features for the eleven estimates, although some require longer convergence times (see in particular the estimation of $\mathbb{E}[\lambda_1\lambda_2]$ where the slowest curve corresponds to λ_6).

Besides providing simultaneously different convergent estimators of quantities of interest, these Rao–Blackwellized Riemann sum estimators can also significantly contribute to convergence monitoring. In fact, they first provide a straightforward stopping rule which is that all the available estimates have converged to a similar value. This method was also proposed in Robert (1995) with different estimators. While it is not sufficient to guarantee convergence or even stationarity, especially when the monitoring is based on a single path of the Markov chain, the comparison of the various estimates increases the confidence in the evaluation of $\mathbb{E}[h(X)]$. However, another major incentive for using Rao–Blackwellized Riemann sums in convergence monitoring is that they lead to simple control variates, as shown in the next section.

4. Control variates for convergence assessment

We show in this section how a single path of the Markov chain can be used to produce a valid, i.e. convergent, evaluation of the "missing mass", that is, of the weight of the part of the support of f which has not yet been explored by the chain. Although the method is markedly different, the goal is similar to Brooks (1998), where the author also evaluates the probability of the part of the space not yet explored by the Markov chain.

4.1. *Control via estimation of 1*

When the integration problem is of dimension 1 or when a univariate marginal density f is available in closed form, the estimator (4) can be used with the constant function $h(x) = 1$,

leading to

$$\Delta_T^1 = \sum_{t=1}^{T-1} \left(x^{[t+1]} - x^{[t]}\right) f\left(x^{[t]}\right) \tag{11}$$

as an "estimator of 1". In this special case, Δ_T^1 thus works as a control variate in the sense that it must converge to 1 for the chain to converge. The important feature of (11) is, however, that it provides us with an "on-line" evaluation of the probability of the region yet unexplored by the chain and is thus a clear convergence diagnostic for stationarity issues.

Example 4. Consider the case of a bivariate normal mixture,

$$(X, Y) \sim p\mathcal{N}_2(\mu, \Sigma) + (1-p)\mathcal{N}_2(\nu, \Sigma'), \tag{12}$$

where $\mu = (\mu_1, \mu_2)$, $\nu = (\nu_1, \nu_2) \in \mathbb{R}^2$ and the covariance matrices are

$$\Sigma = \begin{pmatrix} a & c \\ c & b \end{pmatrix}, \qquad \Sigma' = \begin{pmatrix} a' & c' \\ c' & b' \end{pmatrix}.$$

In this case, the conditional distributions are also normal mixtures,

$$X \mid y \sim \omega_y \mathcal{N}\left(\mu_1 + \frac{(y-\mu_2)c}{b}, \frac{\det \Sigma}{b}\right) + (1-\omega_y)\mathcal{N}\left(\nu_1 + \frac{(y-\nu_2)c'}{b'}, \frac{\det \Sigma'}{b'}\right)$$

$$Y \mid x \sim \omega_x \mathcal{N}\left(\mu_2 + \frac{(x-\mu_1)c}{a}, \frac{\det \Sigma}{a}\right) + (1-\omega_x)\mathcal{N}\left(\nu_2 + \frac{(y-\nu_1)c'}{a'}, \frac{\det \Sigma'}{a'}\right),$$

where

$$\omega_x = \frac{p^{-1/2}\exp(-(x-\mu_1)^2/(2a))}{pa^{-1/2}\exp(-(x-\mu_1)^2/(2a)) + pa'^{-1/2}\exp(-(y-\nu_1)^2/(2a'))}$$

$$\omega_y = \frac{pb^{-1/2}\exp(-(y-\mu_2)^2/(2b))}{pb^{-1/2}\exp(-(y-\mu_2)^2/(2b)) + pb'^{-1/2}\exp(-(y-\nu_2)^2/(2b'))}.$$

They thus provide a straightforward Gibbs sampler, while the marginal distributions of X and Y are again normal mixtures,

$$X \sim p\mathcal{N}(\mu_1, a) + (1-p)\mathcal{N}(\nu_1, a')$$

$$Y \sim p\mathcal{N}(\mu_2, b) + (1-p)\mathcal{N}(\nu_2, b').$$

It is easy to see that, when both components of the normal mixture (12) are far apart, the Gibbs sampler may take a large number of iterations to jump from one component to the other. This feature is thus ideal to study the properties of the convergence diagnostic (11). As shown by Fig. 5, for the numerical values $\mu = (0, 0)$, $\nu = (15, 15)$, $p = 0.5$, $\Sigma = \Sigma' = \left(\begin{smallmatrix} 3 & 1 \\ 1 & 3 \end{smallmatrix}\right)$, the chain

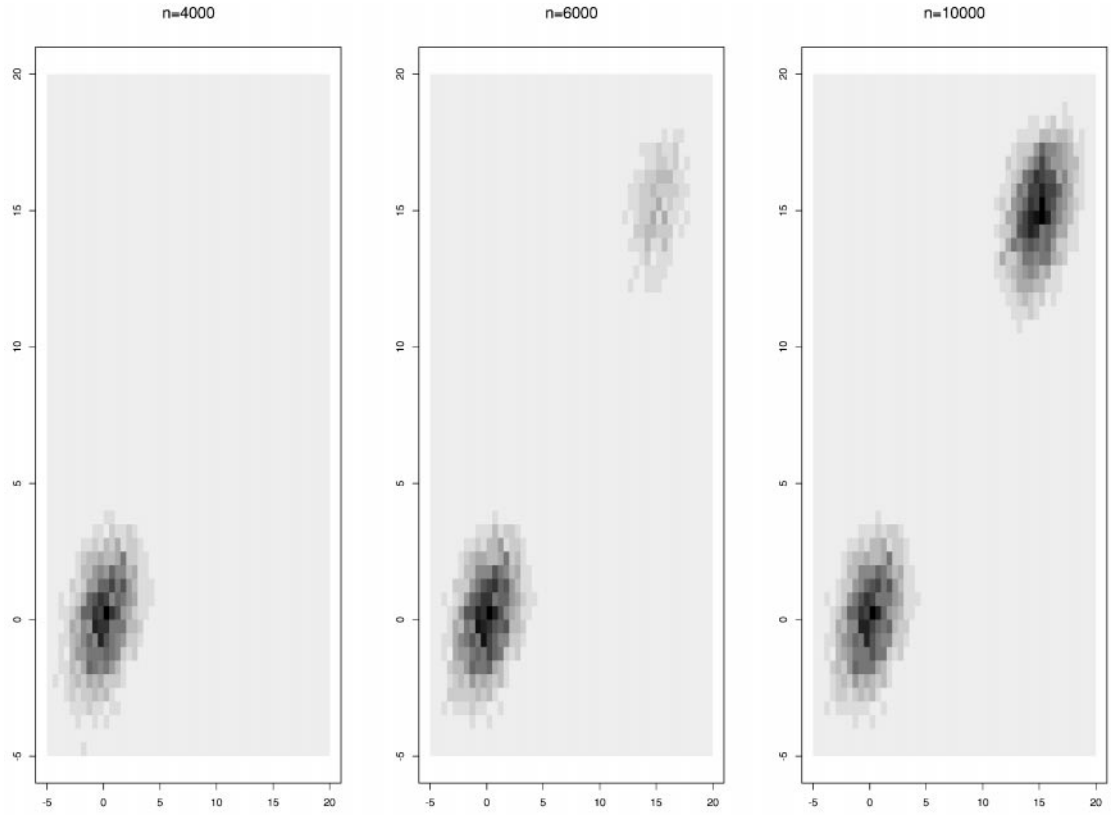

Fig. 5. *(top) Histogram of the Markov chain after 4000, 6000 and 10,000 iterations (middle) Path of the Markov chain for the first coordinate x (bottom) Control curves for the bivariate mixture model, for the parameters μ = (0, 0), ν = (15, 15), p = 0.5, $\Sigma = \Sigma' = \left(\begin{smallmatrix} 3 & 1 \\ 1 & 3 \end{smallmatrix}\right)$*
(Continued on next page).

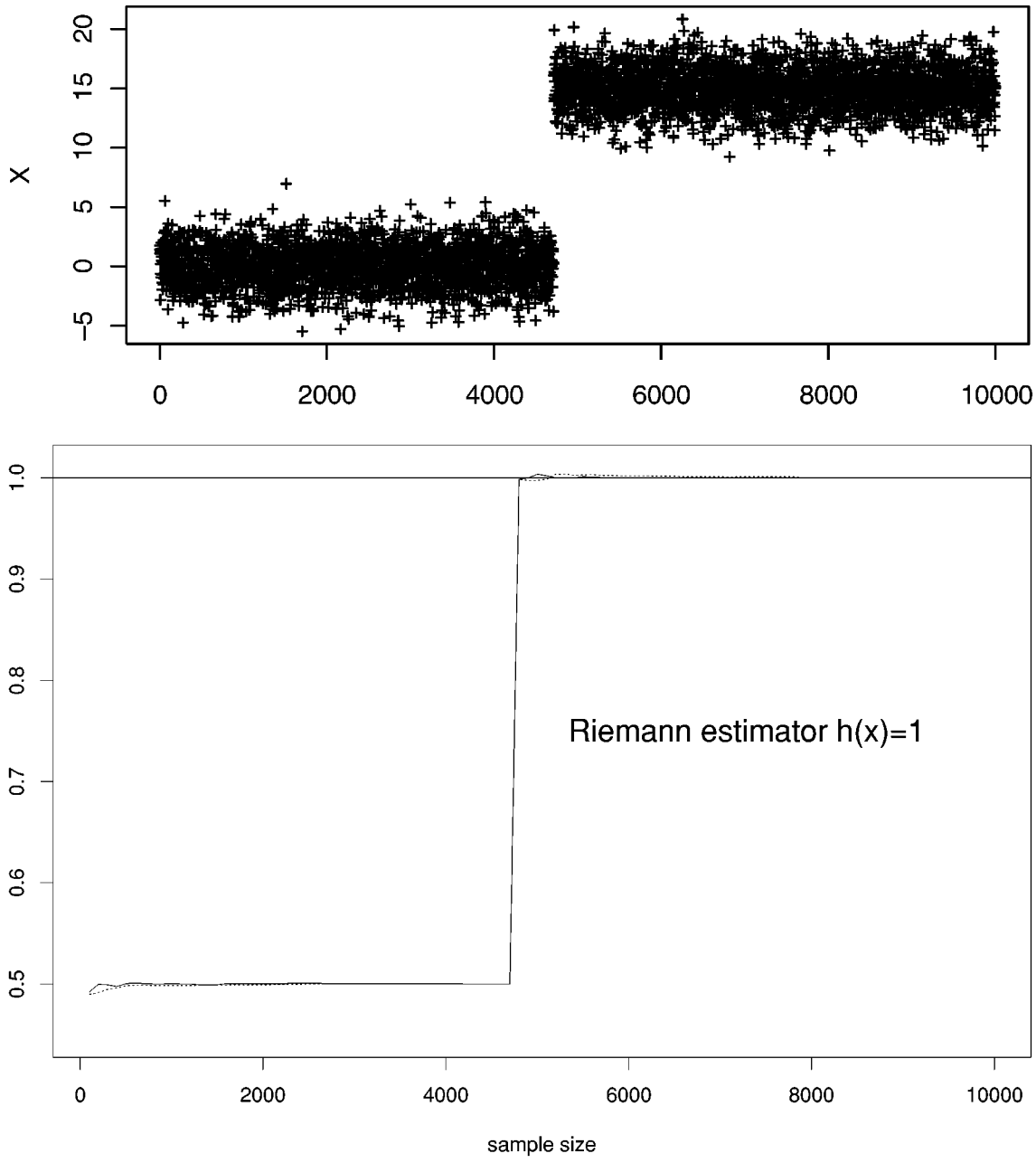

Fig. 5. *(Continued.)*

takes almost 5000 iterations to jump to the second component and this is exactly diagnosed by (11), where both indicators converge quickly to $p = 0.5$ at first and then to 1 when the second mode is being visited. This illustrates how (11) is truly an online evaluation of the probability mass of the region visited by the Markov chain.

While the previous examples all involve the Gibbs sampler, the appeal of using Riemann and Rao–Blackwellized Riemann sum estimators is not restricted to the Gibbs sampler. For instance, in the case of a unidimensional Metropolis–Hastings algorithm, the Riemann sum estimators apply as well and the estimator Δ_T^1 can be used to calibrate the instrumental density. Indeed, a monitoring of the convergence of Δ_T^1 to 1 for different instrumental densities (or different parameters of an instrumental density) can identify a fast mixing algorithm.

Example 5. Consider simulating from the inverse Gaussian density

$$g(x) \propto x^{-3/2} \exp(-\theta_1 x - \theta_2/x)\mathbb{I}_{\mathbb{R}_+},$$

using a gamma distribution $\mathcal{G}a(\beta\sqrt{\theta_2/\theta_1}, \beta)$ as proposal. This parameterization is chosen in order to preserve the first moment

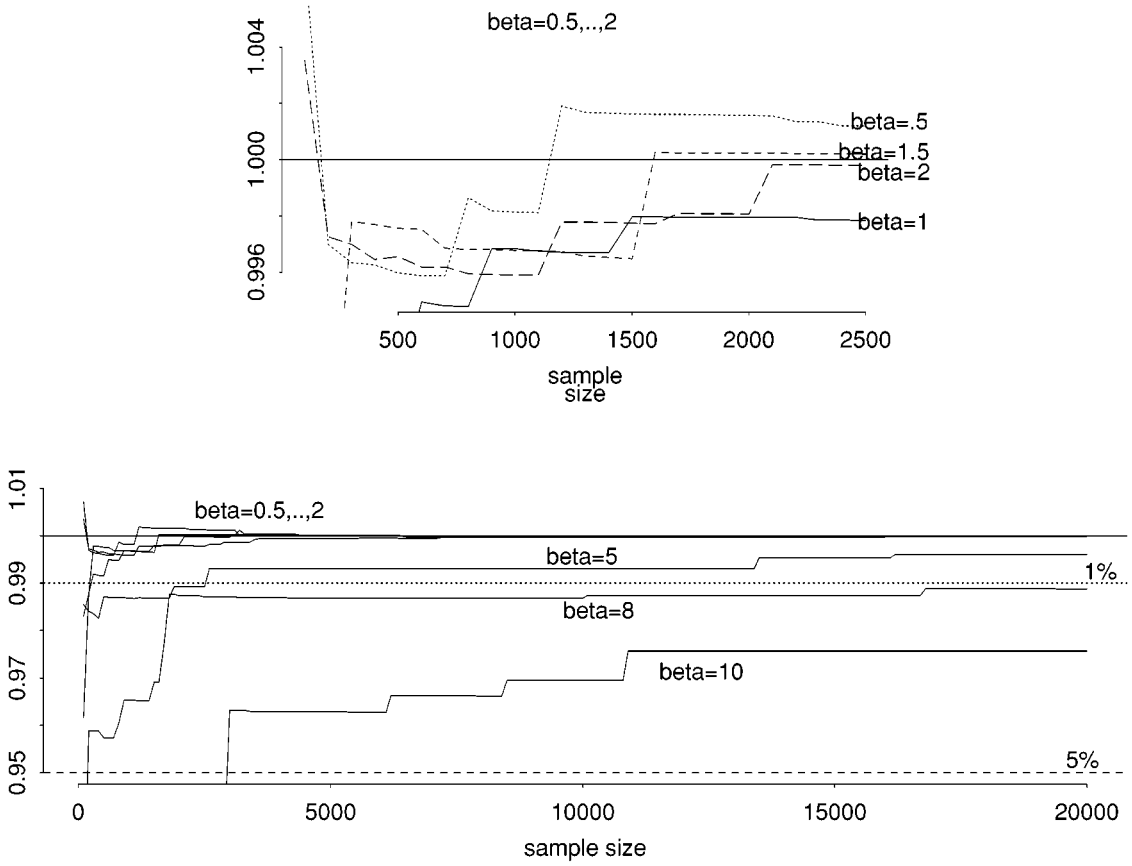

Fig. 6. *(bottom) Control variate for the simulation of the inverse Gaussian distribution using a Metropolis–Hastings algorithm associated with the instrumental densities $\mathcal{G}(\beta\ \theta_2/\theta_1, \beta)$ for different values of $\beta(\theta_1 = 1.5, \theta_2 = 2)$ (top) Magnification of the control curves for $0.5 \le \beta \le 2$ and the first 2,500 iterations*

of the inverse Gaussian distribution. There is therefore only one free parameter β. While this parameter can be calibrated on the acceptance rate of the Metropolis–Hastings algorithm, as suggested in the literature (see Robert and Casella (1999), Note 6.7.4), monitoring the convergence to 1 of the control variate (11) gives a better view of the convergence properties, that is, of the mixing speed, of the corresponding algorithm. In particular, the acceptance rates vary between 0.69 and 0.79 for the values of β under study, being quite above the suggested 0.5. Figure 6 shows convergence paths for various values of β and a preference for the value $\beta = 1.5$.

4.2. *Estimating the missing mass*

In the general case where f is not available in closed form, the Rao–Blackwellized Riemann sum extension (10) can replace (4) for the constant function $h(x) = 1$, leading to the family of control variates $(1 \le l \le p)$

$$\Delta_T^1(l) = T^{-1} \sum_{t=1}^{T-1} \left(x_l^{[t+1]} - x_l^{[t]}\right) \left(\sum_{k=1}^{T} \pi_l\left(x_l^{[t]} \mid x_{-l}^{(k)}\right) \right). \quad (13)$$

In this estimate, the average

$$T^{-1} \sum_{k=1}^{T} \pi_l\left(x_l^{[t]} \mid x_{-l}^{(k)}\right) \quad (14)$$

corresponds to the Rao–Blackwellized estimation of the marginal density of the l-th component. Therefore, when the whole support of the density has been visited, the quantity (13) converges to 1. Most unfortunately, the reverse is not true, namely the fact that (13) is close to 1 is not a sufficient condition for the exploration of the whole support of the joint density by the Markov chain.

The reason for this difficulty is that the Rao–Blackwell estimation of the marginal density is based on the Markov chain and thus on the part of the support explored so far. In the case of well separated modes, as in Example 4, the full conditional distributions will not detect the missing part of the support, given that they are concentrated on one of the modes. More precisely, at a given iteration T, let $C_l \subset \mathbb{R}$ denote the region visited by the l-th component, $C_{-l} \subset \mathbb{R}^{p-1}$ denote the region visited by the $x_{-l}^{(t)}$'s and let $C = C_l \times C_{-l}$ be the Cartesian product of the two. Since the x_{-l} subchain is concentrated on C_{-l}, the Rao–Blackwell estimate of the marginal density of x_l is biased: its

expectation is

$$\frac{\int_{C_{-l}} \pi_l(x_l \mid x_{-l})\pi^l(x_{-l})\,dx_{-l}}{\int_{C_{-l}} \pi^l(x_{-l})\,dx_{-l}},$$

rather than the marginal density f_l. The corresponding Rao–Blackwellized Riemann sum estimator is thus approximating

$$\frac{\int_C f(x)\,dx}{\int_{C_{-l}} \pi^l(x_{-l})\,dx_{-l}} = P(X \in C \mid X_{-l} \in C_{-l})\,.$$

This quantity may then be close to 1 for every l if

$$P\left(X_l \in C_l, X_{-l} \in C_{-l}^{\mathrm{c}}\right) \approx 0$$

for every l, which is the case for well separated modes as in Example 4.

4.3. *Control via score functions*

This difficulty with the control variate associated with constant functions leads us to propose a reinforcement of the convergence diagnostic by using other functions h with known expectations and different features. Brooks and Gelman (1998) consider score functions

$$h_l(x) = \frac{\partial}{\partial x_l}\log(f(x)) = \frac{\partial}{\partial x_l}\log(\pi_l(x_l \mid x_{-l})),$$

whose expectation is equal to zero when the support of the conditional distribution π_l is unbounded. More general score functions can be proposed for convergence purposes, namely $(l, m = 1, \ldots, p)$

$$h_m^l(x) = \frac{\partial}{\partial x_m}\log(\pi_l(x_l \mid x_{-l}))\,,$$

since they all have zero expectations under the stationary distribution when $l \neq m$ under fairly general conditions. We thus propose to monitor the corresponding Rao–Blackwellized Riemann sum estimators of $\mathbb{E}^{\pi}[h_l^m(X)]$

$$\Delta_T^S(l,m) = T^{-1}\sum_{t=1}^{T-1}\left(x_l^{[t+1]} - x_l^{[t]}\right)\left(\sum_{k=1}^{T}\frac{\partial}{\partial x_m}\pi_l\left(x_l^{[t]} \mid x_{-l}^{(k)}\right)\right), \tag{15}$$

till they all converge to 0.

Example 3 continued. Figure 7 compares the behavior of the control variates (13) and (15). Note that

$$\frac{\partial}{\partial \lambda_i}\log(\pi(\beta \mid \lambda))$$

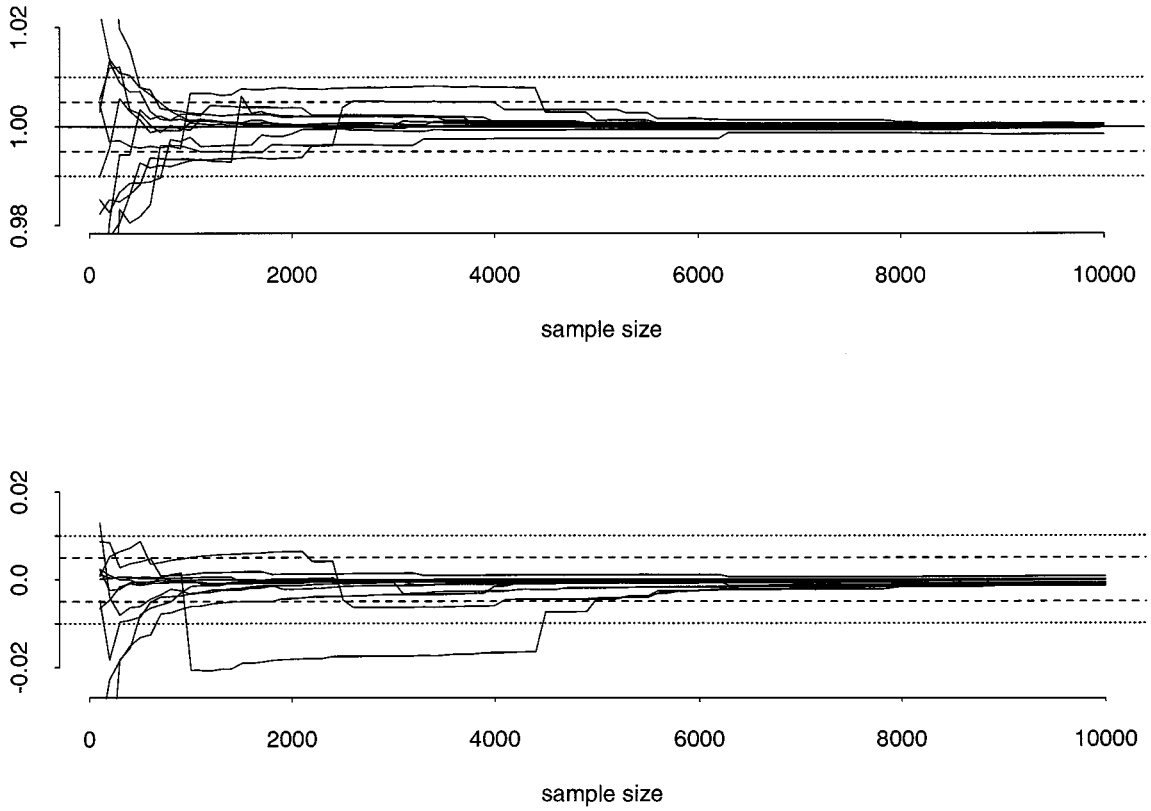

Fig. 7. *Control curves for the pump failure example: Rao–Blackwellized Riemann sum control variates for the constant function $h = 1$ (top) and for the functions h_m^l (bottom)*

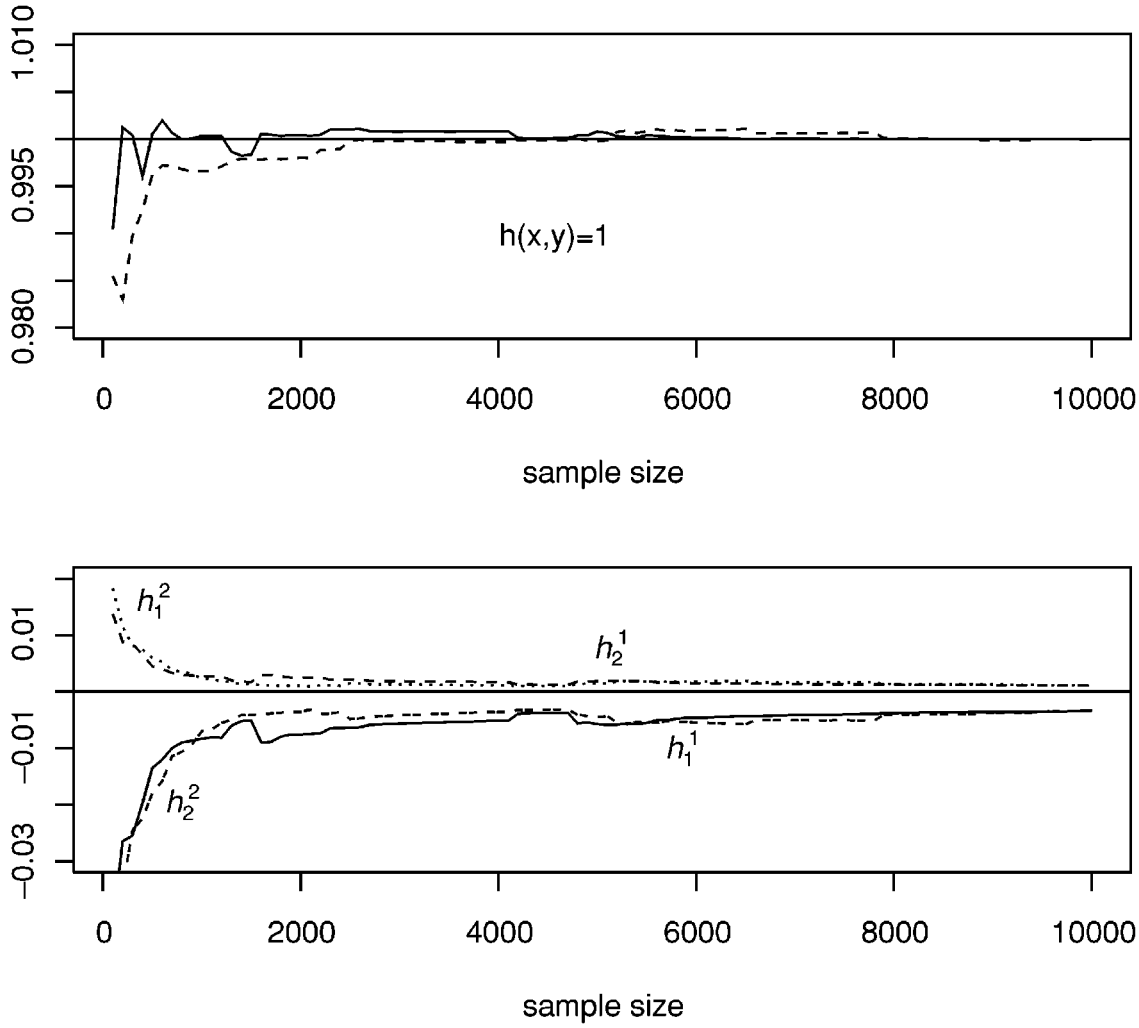

Fig. 8. *Control curves for the bivariate mixture model, for the parameters $\mu = (0, 0)$, $\nu = (15, 15)$, $p = 0.5$, and $\Sigma = \Sigma' = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$. (top) Control variates with expectation 1 (bottom) Control variates with expectation 0. (Same simulations as in Fig. 5.)*

is independent of i, while, for $i \neq j$,

$$\frac{\partial}{\partial \lambda_i} \log(\pi(\lambda_j \mid \lambda_{-j}, \beta)) = 0.$$

The number of score control variates is thus 11 in this setup. For this model, where the posterior is actually unimodal, both criteria lead to the same convergence times, in the sense that it takes about 6000 iterations to get to the expected value. Note that the control variates of the top graph sometimes converge to 1 from above, which shows that the error due to the Rao–Blackwell estimation of the marginal densities is larger than the missing mass in the explored support.

Example 4 continued. When considering the same sequence of iterations as in Fig. 5. Figure 8 illustrates the improvement brought by the score functions h_m^l, when used in complement to the control variates for 1. While the chain has only explored half of the support at iteration 6000, the control variates associated with 1 both give a positive signal very early. On the contrary, the score functions h_m^l somehow capture the fact that the whole support has not been explored at this stage.

5. Conclusion

This paper has established that Riemann sum estimation is a high performance alternative to the empirical average, and that it can be used in MCMC settings at little implementation cost since it is an on-line processing of the simulation output like Rao–Blackwellization. While providing more accurate estimations of the quantities of interest, often by an order of magnitude in the variance, it also allows for the calibration of Metropolis–Hastings algorithms and for convergence monitoring. In particular, it offers a very specific and attractive feature of estimating "on-line" the probability mass of the part of the support explored so far. The availability of many simultaneous control variates is another point of interest since, while the criterion is unidimensional on an individual basis, the possibility of multiplying the

number of control variates allows to capture most of the features of the stationary distribution.

Acknowledgment

This work has been supported by EU TMR network ERB–FMRX–CT96–0095 on *"Computational and statistical methods for the analysis of spatial data"*. The authors are grateful to Steve Brooks for helpful discussions about the score method.

References

Besag J. 1974. Spatial interaction and the statistical analysis of lattice system. J. Royal Statist. Soc B 36: 192–326.

Brooks S.P. 1998. MCMC convergence diagnosis via multivariate bounds on log-concave densities. Ann. Statist. 26: 398–433.

Brooks S.P. and Gelman A. 1998. Some issues in monitoring convergence of iterative simulations. In: Proceedings of the Section on Statistical Computing, ASA.

Casella G. 1996. Statistical inference and Monte Carlo algorithms (with discussion). TEST 5: 249–344.

Casella G. and Robert C.P. 1996. Rao-Blackwellization of sampling schemes. Biometrika 83: 81–94.

Chen M.M. and Shao Q.M. 1997. On Monte Carlo methods for estimating ratios of normalizing constants. Ann. Statist. 25: 1563–1594.

Cowles M.K. and Carlin B.P. 1996. Markov chain Monte Carlo convergence diagnostics: A comparison study. J. Amer. Statist. Assoc. 91: 883–904.

Gaver D.P. and O'Muircheartaigh I.G. 1987. Robust empirical Bayes analysis of event rate. Technometrics 29(1): 1–15.

Gelfand A.E. and Smith A.F.M. 1990. Sampling based approaches to calculating marginal densities. J. Amer. Statist. Assoc. 85: 398-409.

Geyer C.J. 1993. Estimating normalizing constants and reweighting mixtures in Markov chain Monte-Carlo. Technical Report 568, School of Statistics, Univ. of Minnesota.

Mengersen K.L., Robert C.P., and Guihenneuc-Jouyaux C. 1999. MCMC convergence diagnostics: A "rewiewww". In: Berger J.O., Bernardo J.M., Dawid A.P., Lindley D.V., and Smith A.F.M. (Eds.), Bayesian Statistics, Vol. 6. Oxford University Press, Oxford, pp. 415–440.

Philippe A. 1997b. Simulation output by Riemann sums. J. Statist. Comput. Simul. 59: 295–314.

Philippe A. 1997a. Importance sampling and Riemann Sums. Pub. I.R.MA. Lille 43(VI).

Robert C.P. and Casella G. 1999. Monte-Carlo Statistical Methods. Springer-Verlag, New-York.

Robert C.P. 1995. Convergence control techniques for MCMC algorithms. Statis. Science 10: 231–253.

Robert C.P. 1998. Discretization and MCMC Convergence Assessment. Springer-Verlag, New York. Lecture Notes in Statistics, Vol. 135.

Rubinstein B. 1981. Simulation and the Monte-Carlo Method. Wiley, New York.