Adaptive Importance Sampling in General Mixture Classes

Olivier Cappé, Randal Douc, Arnaud Guillin, Jean-Michel Marin, Christian P. Robert

Adaptive Importance Sampling in General Mixture Classes*

Olivier Cappé,
LTCI, ENST & CNRS, Paris
Randal Douc,
INT, Evry
Arnaud Guillin,
École Centrale & LATP, CNRS, Marseille
Jean-Michel Marin†,
INRIA Futurs, Project select*, Laboratoire de Mathématiques, Université Paris-Sud*
& Christian P. Robert
CEREMADE, Université Paris Dauphine & CREST, INSEE

October 23, 2007

Abstract

In this paper, we propose an adaptive algorithm that iteratively updates both the weights and component parameters of a mixture importance sampling density so as to optimise the importance sampling performances, as measured by an entropy criterion. The method is shown to be applicable to a wide class of importance sampling densities, which includes in particular mixtures of multivariate Student t distributions. The performances of the proposed scheme are studied on both artificial and real examples, highlighting in particular the benefit of a novel Rao-Blackwellisation device which can be easily incorporated in the updating scheme.
Keywords: Importance Sampling mixtures, adaptive Monte Carlo, Population Monte Carlo, entropy.

1 Introduction

In recent years, there has been a renewed interest in using Monte Carlo procedures based on Importance Sampling (abbreviated to IS in the following) for inference tasks. Compared to alternatives such as Markov Chain Monte Carlo methods, the main appeal of IS procedures lies in the possibility of developing parallel implementations, which becomes more and more important with the generalisation of multiple core machines and computer clusters. Importance sampling procedures are also attractive in that they allow for an easy assessment of the Monte Carlo error and, correlatively, for the development of learning mechanisms. In many applications, the fact that IS procedures may be tuned—by choosing an appropriate IS density—to minimise the approximation error for a specific function of interest is also crucial. On the other hand, the shortcomings of IS approaches are also well-known, including a poor scaling to highly multidimensional problems and an acute sensitivity to

*This work has been supported by the Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75012 Paris) through the 2006-2008 project Adap'MC. Both last authors are grateful to the participants to the BIRS 07w5079 meeting on "Bioinformatics, Genetics and Stochastic Computation: Bridging the Gap", Banff, for their helpful comments. The last author also acknowledges an helpful discussion with Geoff McLachlan.
†Corresponding author

the choice of the IS density combined with the fact that it is impossible to come up with a universally efficient IS density.

Adaptive Monte Carlo is a natural solution to remedy for the latter class of difficulties by gradually improving the IS density based on some form of Monte Carlo approximation. While there exist a wide variety of solutions in the literature (see, e.g. Robert and Casella, 2004, Chapter 14), this paper concentrates on the construction of iterated importance sampling schemes or *population Monte Carlo*. Population Monte Carlo (or PMC) was introduced by Cappé et al. (2004) as an repeated Sampling Importance Resampling (SIR) procedure: once a sample $(X_1, \ldots, X_N)$ is produced by SIR, it provides an approximation to the target distribution π and can be used as a stepping stone towards a better approximation to π. More precisely, if $(X_1, \ldots, X_N)$ is a sample approximately distributed from π, it may be perturbed stochastically using an arbitrary Markov transition kernel $q(x, x')$ so as to produce new sample $(X'_1, \ldots, X'_N)$. Conducting a resampling step based on the IS weights $\omega_i = \pi(X'_i)/q(X_i, X'_i)$, we will then produce a new sample $(\tilde{X}_1, \ldots, \tilde{X}_N)$ that also constitutes an approximation to the target distribution π. Repeating this scheme in an iterative manner is however only of interest if samples that have been previously simulated are used to update (or adapt) the kernel $q(x, x')$, in the sense that keeping the same kernel q over iterations does not modify the statistical properties of the sample produced at each iteration and, therefore, reduces the efficiency of the approximation by introducing extra Monte Carlo variance. Failing to improve upon the choice of the kernel q thus cancels the appeal of using several iterations, when compared with one single IS draw with the same total sample size (see Douc et al., 2007a).

Population Monte Carlo therefore requires an updating scheme that takes advantage of previously generated samples so that it improves the choice of the IS transition kernel against a given measure of efficiency. In the approach advocated by Douc et al. (2007a), one considers a transition kernel q consisting of a mixture of *fixed transition kernels*

$$q_\alpha(x, x') = \sum_{d=1}^{D} \alpha_d q_d(x, x')\,, \qquad \sum_{d=1}^{D} \alpha_d = 1\,, \tag{1}$$

whose weights $\alpha_1, \ldots, \alpha_D$ are tuned adaptively. The proposed adaptation procedure aims at minimising the deviance or entropy criterion between the kernel q_α and the target π,

$$\mathfrak{E}(\pi, q_\alpha) = \mathbb{E}_\pi^X \left[D(\pi \| q_\alpha(X, \cdot)) \right]\,, \tag{2}$$

where $D(p\|q) = \int \log\{p(x)/q(x)\}\, p(x)\mathrm{d}x$ denotes the Kullback-Leibler divergence (also called relative entropy), and where the expectation is taken under the target distribution $X \sim \pi$ since the kernels $q_d(x, x')$ depend on the starting value x. In the sequel, we refer to the criterion in (2) as the *entropy criterion* as it is obviously related to the performance measure used in the cross-entropy method of Rubinstein and Kroese (2004). In Douc et al. (2007b), a version of this algorithm was developed to minimise the asymptotic variance of the IS procedure, *for a specific function of interest*, in lieu of the entropy criterion.

A major limitation in the approaches of both Douc et al. (2007a,b) is that the proposal kernels q_d remain fixed over the iterative process while only the mixture weights α_d get improved. In the present contribution, we remove this limitation by extending the framework of Douc et al. (2007a) to allow for the adaption of IS densities of the form

$$q_{(\alpha,\theta)}(x) = \sum_{d=1}^{D} \alpha_d q_d(x; \theta_d)\,, \tag{3}$$

with respect to *both* the weights α_d and the internal parameters θ_d of the component densities. In theory, as explained through the example considered in Section 4, the proposed adaptive scheme,

which relies on an integrated EM update mechanism, is applicable to more general families of latent-data IS densities. This proposed extension and, in particular, the introduction of (multidimensional) scaling parameters raises challenging robustness issues for which we propose a Rao-Blackwellisation scheme that empirically appears to be very efficient while inducing a modest additional algorithmic complexity.

Note that we consider here the generic entropy criterion of Douc et al. (2007a) rather than the function-specific variance minimisation objective of Douc et al. (2007b). This choice is motivated by the recognition that in most applications, the IS density is expected to perform well for a range of typical functions of interest rather than for a specific target function h. In addition, the generalisation of the approach of Douc et al. (2007b) to a class of mixture IS densities that are parameterised by more than the weights remains an open question (see also Section 5).

A second remark is that in contrast to the previously cited works, we consider in this paper only "global" independent IS densities of the form given in (3). Thus the proposed scheme really is an iterated importance sampling scheme, contrary to what happens when using more general IS transition kernels as in (1). Obviously, resorting to moves that depend on the current sample is initially attractive because it allows for some local moves as opposed to the global exploration provided by independent IS densities. However, the fact that the entropy criterion in (2) is a global measure of fit tends to modify the parameters of each transition kernel depending on its average performance over the whole sample, rather than locally. In addition, structurally imposing a dependence on the points sampled at the previous iteration induces some extra-variability which can be detrimental when more parameters are to be estimated.

The paper is organised as follows: In Section 2, we develop a generic updating scheme for independent IS mixtures (3), establishing that the integrated EM argument of Douc et al. (2007a) remains valid in our setting. Note once again that the integrated EM update mechanism we uncover in this paper is applicable to all missing data representations of the proposal kernel, and not only to finite mixtures. In Section 3, we consider the case of Gaussian mixtures which naturally extend the case of mixtures of Gaussian random walks with fixed covariance structure considered in Douc et al. (2007a,b). In Section 4, we show that the algorithm also applies to mixture of multivariate t distributions with the continuous scale mixing representation used in Peel and McLachlan (2000). Section 5 provides some conclusive remarks about the performances of this approach as well as possible extensions.

2 Updating the IS density

2.1 Entropies and perplexity

When considering independent mixture IS densities of the form (3), the entropy criterion $\mathfrak{E}$ defined in (2) reduces to the Kullback-Leibler divergence between the target density π and the mixture $q_{(\alpha,\theta)}$:

$$\mathfrak{E}(\pi, q_{(\alpha,\theta)}) = D(\pi \| q_{(\alpha,\theta)}) = \int \log \left(\frac{\pi(x)}{\sum_{d=1}^{D} \alpha_d q_d(x;\theta_d)} \right) \pi(x) \mathrm{d}x \,. \tag{4}$$

As usual in applications of the IS approach to Bayesian inference, the target density π is known up to a normalisation constant only and we will focus on the self-normalised version of IS which only requires the knowledge of an unnormalised version π_{unn} of π (Geweke, 1989). As a side comment, note that while $\mathfrak{E}(\pi, q_{(\alpha,\theta)})$ is a convex function of the weights $\alpha_1, \ldots, \alpha_D$ (Douc et al., 2007a), it is generally not so when also optimising with respect to the component parameters $\theta_1, \ldots, \theta_D$.

It is well known that if one considers a function h of interest, the self-normalised IS estimation

of its expectation $\widehat{\pi(h)} = \sum_{i=1}^N \bar{\omega}_i h(X_i)$, where $\bar{\omega}_i = (\pi(X_i)/q_{(\alpha,\theta)}(X_i))/(\sum_{j=1}^N \pi(X_j)/q_{(\alpha,\theta)}(X_j))$ and $X_i \sim q_{(\alpha,\theta)}$, has an asymptotic variance of

$$\upsilon(h) = \int \left\{h(x) - \pi(h)\right\}^2 \pi^2(x)/q_{\alpha,\theta}(x)\mathrm{d}x\,,$$

assuming that $\int(1 + h^2(x))\pi^2(x)/q_{\alpha,\theta}(x)\mathrm{d}x < \infty$. In addition, this asymptotic variance may be consistently estimated using the IS sample itself as $N\sum_{i=1}^N \bar{\omega}_i^2\{h(X_i) - \widehat{\pi(h)}\}^2$ (Geweke, 1989).

Obviously, for a *given* function h, there is no direct link between $\upsilon(h)$ and the entropy criterion in (4), a fact that motivated the work of Douc et al. (2007b). However it is easily shown that

$$\sup_{\{h:\|h-\pi(h)\|_\infty \leq M\}} \upsilon(h) = M^2 \int \pi^2(x)/q_{(\alpha,\theta)}(x)\mathrm{d}x,$$

where the latter integral term is lower and upper bounded by 1 and $\exp\left[\mathfrak{E}(\pi, q_{(\alpha,\theta)})\right]$ respectively, by direct applications of Jensen's inequality. Hence minimising $\mathfrak{E}(\pi, q_{(\alpha,\theta)})$ indeed reduces the worst case performance of the IS approach, at least for bounded functions. In addition, rewriting

$$\exp\left[-\mathfrak{E}(\pi, q_{(\alpha,\theta)})\right] = \exp\left(\int -\log\frac{\pi_{\mathrm{unn}}(x)}{q_{(\alpha,\theta)}(x)}\pi(x)\mathrm{d}x\right)\left(\int \pi_{\mathrm{unn}}(x)\mathrm{d}x\right)$$

and estimating the first integral by self-normalised IS as

$$-\sum_{i=1}^N \bar{\omega}_i \log\frac{\pi_{\mathrm{nn}}(X_i)}{q_{(\alpha,\theta)}(X_i)}$$

and the second one by classical IS, as

$$1/N\sum_{i=1}^N \pi_{\mathrm{nn}}(X_i)/q_{(\alpha,\theta)}(X_i),$$

shows that $\exp(H_N)/N$, where $H_N = -\sum_{i=1}^N \bar{\omega}_i \log\bar{\omega}_i$ is the Shannon entropy of the normalised IS weights, is an estimator of the inverse of the term $\exp\left[\mathfrak{E}(\pi, q_{(\alpha,\theta)})\right]$. Thus, minimisation of the entropy criterion is connected with the maximisation of $\exp(H_N)/N$, were H_N is the entropy of the IS weights, a frequently used criterion for assessing the quality of an IS sample—together with the so-called Effective Sample Size (ESS) criterion (Chen and Liu, 1996, Doucet et al., 2001, Cappé et al., 2005). In the following, we refer to $\exp(H_N)/N$ as the *normalised perplexity* of the IS weights, following the terminology in use in the field of natural language processing.

2.2 Integrated updates

Let $\alpha^t = \left(\alpha_1^t, \ldots, \alpha_D^t\right)$ and $\theta^t = \left(\theta_1^t, \ldots, \theta_D^t\right)$ denote, respectively, the mixture weights and the component parameters at the t-th iteration of the algorithm (where $t = 1, \ldots, T$). In order to update the parameters (α^t, θ^t) of the independent IS density (3), we will take advantage of the latent variable structure that underlines the objective function (4). The resulting algorithm—still theoretical at this stage as it involves integration with respect to π—may be interpreted as an integrated EM (Expectation-Maximisation) scheme that we now describe.

Given that minimising (4) in (α, θ) is equivalent to maximising

$$\int \log\left(\sum_{d=1}^D \alpha_d q_d(x;\theta_d)\right)\pi(x)\,\mathrm{d}x\,,$$

we are facing a task that formally resembles standard mixture maximum likelihood estimation but with an integration with respect to π replacing the empirical sum over observations. As usual in mixtures, the latent variable Z is the component indicator, with values in $\{1,\ldots,D\}$ such that the joint density f of x and z satisfies

$$f(z)=\alpha_z \quad \text{and} \quad f(x|z)=q_z(x;\theta_z)\,,$$

which produces (3) as the marginal in x. At iteration t of our algorithm, we can therefore take advantage of this latent variable representation by considering the expected complete log-likelihood

$$\mathbb{E}_\pi^X\left[\mathbb{E}^Z_{(\alpha^t,\theta^t)}\left\{\log\left(\alpha_Z q_Z(X;\theta_Z)\right)|X\right\}\right]\,,$$

where the inner expectation is computed under the conditional distribution of Z for the current value of the parameters, (α^t,θ^t), i.e.

$$f(z|x)=\alpha_z^t q_z(x;\theta_z^t)\Big/\sum_{d=1}^D \alpha_d^t q_d(x;\theta_d^t)\,.$$

The updating mechanism in our algorithm then corresponds to setting the new parameters $(\alpha^{t+1},\theta^{t+1})$ equal to

$$(\alpha^{t+1},\theta^{t+1})=\arg\max_{(\alpha,\theta)}\mathbb{E}_\pi^X\left[\mathbb{E}^Z_{(\alpha^t,\theta^t)}\left\{\log(\alpha_Z q_Z(X;\theta_Z))|X\right\}\right]\,,$$

as in a regular EM estimation of the parameters of a mixture, except for the extra expectation over X. The convexity argument that shows that EM increases the objective function at each step also apply in this setup. Solving the maximisation program, we have

$$(\alpha^{t+1},\theta^{t+1})=\arg\max_{(\alpha,\theta)}\left(\mathbb{E}_\pi^X\left[\mathbb{E}^Z_{(\alpha^t,\theta^t)}\left\{\log(\alpha_Z)|X\right\}\right]+\mathbb{E}_\pi^X\left[\mathbb{E}^Z_{(\alpha^t,\theta^t)}\left\{\log(q_Z(X;\theta_Z))|X\right\}\right]\right)\,.$$

If we define $g_1(\alpha)=\mathbb{E}_\pi^X\left[\mathbb{E}^Z_{(\alpha^t,\theta^t)}\left\{\log(\alpha_Z)|X\right\}\right]$ and $g_2(\theta)=\mathbb{E}_\pi^X\left[\mathbb{E}^Z_{(\alpha^t,\theta^t)}\left\{\log(q_Z(X;\theta_Z))|X\right\}\right]$, we get

$$(\alpha^{t+1},\theta^{t+1})=\arg\max_{(\alpha,\theta)}\left(g_1(\alpha)+g_2(\theta)\right)=\left(\arg\max_\alpha g_1(\alpha),\arg\max_\theta g_2(\theta)\right)\,.$$

Therefore, setting

$$\rho_d(X;\alpha,\theta)=\alpha_d q_d(X;\theta_d)\Big/\sum_{\ell=1}^D \alpha_\ell q_\ell(X;\theta_\ell)\,, \tag{5}$$

we need to solve

$$\alpha^{t+1}=\arg\max_\alpha \mathbb{E}_\pi^X\left[\sum_{d=1}^D \rho_d(X;\alpha^t,\theta^t)\log(\alpha_d)\right]\,,$$

and, for $d\in\{1,\ldots,D\}$, we obtain

$$\alpha_d^{t+1}=\mathbb{E}_\pi^X\left[\rho_d(X;\alpha^t,\theta^t)\right]\,. \tag{6}$$

Similarly,

$$\theta^{t+1}=\arg\max_\theta \mathbb{E}_\pi^X\left[\sum_{d=1}^D \rho_d(X;\alpha^t,\theta^t)\log(q_d(X;\theta_d))\right]\,,$$

and, for $d\in\{1,\ldots,D\}$,

$$\theta_d^{t+1}=\arg\max_{\theta_d} \mathbb{E}_\pi^X\left[\rho_d(X;\alpha^t,\theta^t)\log(q_d(X;\theta_d))\right]\,. \tag{7}$$

As in the regular mixture estimation problem, the resolution of this maximisation program ultimately depends on the shape of the density q_d. If q_d belongs to an exponential family, it is easy to derive a "closed-form" solution, which however involves expectations under π. Section 3 provides an illustration of this fact in the Gaussian case, while the non-exponential Student's t case is considered in Section 4.

2.3 Approximate updates

As argued before, the adaptivity of the proposed procedure is achieved by updating the parameters based on the previously simulated sample. We thus start the PMC algorithm by arbitrarily fixing the mixture parameters (α^0, θ^0) and we then sample from the resulting proposal $\sum \alpha_d^0 q_d(x;\theta_d^0)$ to obtain our initial sample $(X_{i,0})_{1\leq i\leq N}$, associated with the latent variables $(Z_{i,0})_{1\leq i\leq N}$ that indicate from which component of the mixture the corresponding $(X_{i,0})_{1\leq i\leq N}$ have been generated. From this stage, we proceed recursively. Starting at time t from a sample $(X_{i,t})_{1\leq i\leq N}$, associated with $(Z_{i,t})_{1\leq i\leq N}$ and with $(\alpha^{t,N}, \theta^{t,N})$, we denote by $\bar{\omega}_{i,t}$ the normalised importance weights of the sample point $X_{i,t}$:

$$\bar{\omega}_{i,t} = \frac{\pi(X_{i,t})}{\sum_{d=1}^D \alpha_d^{t,N} q_d(X_{i,t};\theta_d^{t,N})} \Big/ \sum_{j=1}^N \frac{\pi(X_{j,t})}{\sum_{d=1}^D \alpha_d^{t,N} q_d(X_{j,t};\theta_d^{t,N})}\,. \tag{8}$$

To approximate (6) and (7), Douc et al. (2007a) proposed the following update rule:

$$\begin{aligned}
\alpha_d^{t+1,N} &= \sum_{i=1}^N \bar{\omega}_{i,t} \mathbb{1}\{Z_{i,t}=d\}\,,\\
\theta_d^{t+1,N} &= \arg\max_{\theta_d} \left[\sum_{i=1}^N \bar{\omega}_{i,t}\mathbb{1}\{Z_{i,t}=d\} \log\left\{q_d\left(X_{i,t};\theta_d^{t,N}\right)\right\}\right]\,.
\end{aligned} \tag{9}$$

The computational cost of this update is of order N whatever the number D of components is, since the weight and the parameter of each component are updated based only on the points that were actually generated from this component. However, this observation also suggests that (9) may be highly variable when N is small and/or D becomes larger. To make the update more robust, we propose a simple Rao-Blackwellisation step that consists in replacing $\mathbb{1}\{Z_{i,t}=d\}$ with its conditional expectation given $X_{i,t}$, that is, $\rho_d\left(X_{i,t};\alpha_d^{t,N},\theta_d^{t,N}\right)$:

$$\begin{aligned}
\alpha_d^{t+1,N} &= \sum_{i=1}^N \bar{\omega}_{i,t}\rho_d\left(X_{i,t};\alpha_d^{t,N},\theta_d^{t,N}\right)\,,\\
\theta_d^{t+1,N} &= \arg\max_{\theta_d} \left[\sum_{i=1}^N \bar{\omega}_{i,t}\rho_d\left(X_{i,t};\alpha_d^{t,N},\theta_d^{t,N}\right) \log\left\{q_d\left(X_{i,t};\theta_d^{t,N}\right)\right\}\right]\,.
\end{aligned} \tag{10}$$

Examining (5) indicates why the evaluation of the posterior probabilities $\rho_d(X_{i,t};\alpha_d^{t,N},\theta_d^{t,N})$ does not represent a significant additional computation cost in the PMC scheme, given that the denominator of this expression has already been computed when evaluating the IS weights according to (8). The most significant difference between and (9) and (10) is that, with the latter, all points contribute to the updating of the d-th component, for an overall cost proportional to $D \times N$. Note however that in many applications of interest, the most significant computational cost is associated with the evaluation of π so that the cost of the update is mostly negligible, even with the Rao-Blackwellised version.

Convergence of the estimated updated parameters as N increases can be established using the same approach as in Douc et al. (2007a,b), relying mainly on the convergence property of triangular

arrays of random variables (see Theorem A.1 in Douc et al., 2007a). For the Rao-Blackwellised version, assuming that for all θ's, $\pi(q_d(\cdot;\theta_d)=0)=0$, for all α's and θ's, $\rho_d(\cdot;\alpha,\theta)\log q_d(\cdot,\theta_d)\in L^1(\pi)$, and, some (uniform in x) regularity conditions on $q_d(x;\theta)$ viewed as a function of θ, yield

$$\alpha_d^{t+1,N} \xrightarrow{\mathbb{P}} \alpha_d^{t+1}, \qquad\qquad \theta_d^{t+1,N} \xrightarrow{\mathbb{P}} \theta_d^{t+1}$$

when N goes to infinity. Note that we do not expand on the regularity conditions imposed on q_d since, for the algorithm to be efficient, we definitely need a closed-form expression on the parameter updates. It is then easier to deal with the convergence of the approximation of these update formulas on a case-by-case basis, as will be seen for instance in the following Gaussian example.

3 The Gaussian mixture case

As a first example, we consider the case of p-dimensional Gaussian mixture IS densities of the form

$$q_d(X;\theta_d) = \left\{(2\pi)^p\left|\Sigma_d\right|\right\}^{-1/2}\exp\left\{-\frac{1}{2}(X-\mu_d)^{\mathrm{T}}\Sigma_d^{-1}(X-\mu_d)\right\},$$

where $\theta_d=(\mu_d,\Sigma_d)$ denotes the parameters of the d-th Gaussian component density. This parametrisation of the IS density provides a general framework for approximating multivariate targets π and the corresponding adaptive algorithm is a straightforward instance of the general framework discussed in the previous section.

3.1 Update formulas

The integrated update formulas are obtained as the solution of

$$\theta_d^{t+1,N} = \arg\min_{\theta}\mathbb{E}_\pi^X\left[\rho_d(X;\alpha^t,\theta^t)\left(\log\left|\Sigma_d\right|+(X-\mu_d)^{\mathrm{T}}\Sigma_d^{-1}(X-\mu_d)\right)\right].$$

It is straightforward to check that the infimum is reached when, for $d\in\{1,\ldots,D\}$,

$$\mu_d^{t+1} = \frac{\mathbb{E}_\pi^X\left[\rho_d(X;\alpha^t,\theta^t)X\right]}{\mathbb{E}_\pi^X\left[\rho_d(X;\alpha^t,\theta^t)\right]},$$

and

$$\Sigma_d^{t+1} = \frac{\mathbb{E}_\pi^X\left[\rho_d(X;\alpha^t,\theta^t)(X-\mu_d^{t+1})(X-\mu_d^{t+1})^{\mathrm{T}}\right]}{\mathbb{E}_\pi^X\left[\rho_d(X;\alpha^t,\theta^t)\right]}.$$

At iteration $t+1$ of the PMC algorithm, both the numerator and the denominator of each of the above expressions are approximated using self-normalised importance sampling, yielding the following empirical update equations for the basic updating strategy

$$\begin{aligned}
\alpha_d^{t+1,N} &= \sum_{i=1}^N \bar\omega_{i,t}\mathbb{1}\{Z_{i,t}=d\},\\
\mu_d^{t+1,N} &= \frac{\sum_{i=1}^N \bar\omega_{i,t}X_{i,t}\mathbb{1}\{Z_{i,t}=d\}}{\sum_{i=1}^N \bar\omega_{i,t}\mathbb{1}\{Z_{i,t}=d\}},\\
\Sigma_d^{t+1,N} &= \frac{\sum_{i=1}^N \bar\omega_{i,t}(X_{i,t}-\mu_d^{t+1,N})(X_{i,t}-\mu_d^{t+1,N})^{\mathrm{T}}\mathbb{1}\{Z_{i,t}=d\}}{\sum_{i=1}^N \bar\omega_{i,t}\mathbb{1}\{Z_{i,t}=d\}},
\end{aligned} \tag{11}$$

and

$$\begin{aligned}
\alpha_d^{t+1,N} &= \sum_{i=1}^{N} \bar{\omega}_{i,t} \rho_d(X_{i,t}; \alpha^{t,N}, \theta^{t,N}) \,, \\
\mu_d^{t+1,N} &= \frac{\sum_{i=1}^{N} \bar{\omega}_{i,t} X_{i,t} \rho_d(X_{i,t}; \alpha^{t,N}, \theta^{t,N})}{\sum_{i=1}^{N} \bar{\omega}_{i,t} \rho_d(X_{i,t}; \alpha^{t,N}, \theta^{t,N})} \,, \\
\Sigma_d^{t+1,N} &= \frac{\sum_{i=1}^{N} \bar{\omega}_{i,t} \left(X_{i,t} - \mu_d^{t+1,N}\right) \left(X_{i,t} - \mu_d^{t+1,N}\right)^{\mathrm{T}} \rho_d(X_{i,t}; \alpha^{t,N}, \theta^{t,N})}{\sum_{i=1}^{N} \bar{\omega}_{i,t} \rho_d(X_{i,t}; \alpha^{t,N}, \theta^{t,N})} \,,
\end{aligned} \tag{12}$$

for the Rao-Blackwellised scheme. Note that as discussed at the end of Section 2, one observes that in the Gaussian case the convergence of the parameter update can be established by assuming only that $\rho_d(x; \alpha, \theta)x^2$ is integrable with respect to π.

3.2 A simulation experiment

To illustrate the results of the algorithm presented above, we consider a toy example in which the target density consists of a mixture of two multivariate Gaussian densities. The appeal of this example is that it is sufficiently simple to allow for an explicit characterisation of the attractive points for the adaptive procedure, while still illustrating the variety of situations found in more realistic applications. In particular, the model contains an attractive point that does not correspond to the global minimum of the entropy criterion as well as some regions of attraction that can eventually lead to a failure of the algorithm. The results obtained on this example also illustrate the improvement brought by the Rao-Blackwellised update formulas in (12).

The target π is a mixture of two p-dimensional Gaussian densities such that

$$\pi(x) = 0.5\mathcal{N}(x; -s\mathbf{u}_p, \mathbf{I}_p) + 0.5\mathcal{N}(x; s\mathbf{u}_p, \mathbf{I}_p) \,,$$

when $\mathbf{u}_p$ is the p-dimensional vector whose coordinates are equal to 1 and $\mathbf{I}_p$ stands for the identity matrix. In the sequel, we focus on the case where $p = 10$ and $s = 2$. Note that one should not be misled by the image given by the marginal densities of π: in the ten dimensional space, the two components of π are indeed very far from one another. It is for instance straightforward to check that the Kullback-Leibler divergence between the two components of π, $D\{\mathcal{N}(-s\mathbf{u}_p, \mathbf{I}_p) \| \mathcal{N}(-s\mathbf{u}_p, \mathbf{I}_p)\}$, is equal to $\frac{1}{2}\|2s\mathbf{u}_p\|^2 = 2s^2p$, that is 80 in the case under consideration. In particular, if we were to use one of the components of the mixture as an IS density for the other, we know from the arguments exposed at the beginning of Section 2 that the normalised perplexity of the weight will eventually tend to $\exp(-80)$. This number is so small, that for any feasible sample size, using one of the component densities of π as an IS instrumental density for the other component or even for π itself can only provide useless biased estimates.

The initial IS density q_0 is chosen here as the isotropic ten-dimensional Gaussian density with a covariance matrix of $5\mathbf{I}_p$. The performances of q_0 as an importance sampling density, when compared to various other alternatives, are fully detailed in Table 1 below but the general comment is that it corresponds to a poor initial guess which would provide highly variable results when used with any sample size under $50,000$.

In addition to figures related to the initial IS density q_0, Table 1 also reports performance obtained with the best fitting Gaussian IS density (with respect to the entropy criterion), which is straightforwardly obtained as the centred Gaussian density whose covariance matrix matches the one of π, that is, $\mathbf{I}_p + s^2\mathbf{u}_p\mathbf{u}_p^{\mathrm{T}}$. Of course the best possible performances achievable with a mixture of two Gaussian densities, always with the entropy criterion, is obtained when using π as an IS density (second line

Proposal	N-PERP	N-ESS	$\sigma^2(x_1)$
q_0 †	6.5E-4	1.5E-4	37E3
Best fitting Gaussian †	0.31	0.27	19
Target mixture †	1 †	1 †	5 †
Best fitting Gaussian (defensive option)	0.28	0.23	22
Best fitting two Gaussian mixture (defensive option)	0.89	0.87	5.8

Table 1: Performance of various importance sampling densities in terms of N-PERP: Normalised perplexity; N-ESS: Normalised Effective Sample Size; $\sigma^2(x_1)$: Asymptotic variance of self-normalised IS estimator for the coordinate projection function $h(x) = x_1$. Quantities marked with a dagger sign are straightforward to determine, all others have been obtained using IS with a sample size of one million.

of Table 1). Finally both final lines of Table 1 report the best fit obtained with IS densities of the form $0.9 \sum_{d=1}^{D} \alpha_d \mathcal{N}(\mu_d, \Sigma_d) + 0.1 q_0(\cdot)$ when, respectively, $D = 1$ and $D = 2$ (further comments on the use of these are given below). As a general comment on Table 1, note that the variations of the perplexity of the IS weights, of the ESS and of the asymptotic variance of the IS estimate for the coordinate projection function are very correlated. This is a phenomenon that we have observed on many examples and which justifies our postulate that minimising the entropy criterion does provide very significant variance reductions for the IS estimate of "typical" functions of interest.

In this example, one may categorise the possible outcomes of adaptive IS algorithms based on mixtures of Gaussian IS densities into mostly four situations:

Disastrous (D.) After T iterations of the PMC scheme, $q_{(\alpha^T, \theta^T)}$ is not a valid IS density and may lead to inconsistent estimates. Typically, this may happen if $q_{(\alpha^T, \theta^T)}$ becomes much too peaky with light tails. As discussed above, it will also practically be the case if the algorithm only succeeds in fitting $q_{(\alpha^T, \theta^T)}$ to one of both Gaussian modes of π. Another disastrous outcome is when the direct application of the adaptation rules described above leads to numerical problems, usually due to the poor conditioning of some of the covariance matrices Σ_d. Rather than fixing these issues by ad-hoc solutions (eg. diagonal loading), which could nonetheless be useful in practical applications, we consider below more principled ways of making the algorithm more resistant to such failures.

Mediocre (M.) After adaptation, $q_{(\alpha^T, \theta^T)}$ is not significantly better than q_0 in terms of the performance criteria displayed in Table 1 and, in this case, the adaptation is useless.

Good (G.) After T iterations, $q_{(\alpha^T, \theta^T)}$ selects the best fitting Gaussian approximation (second line of Table 1) which already provides a very substantial improvement as it results in variance reductions by about four orders of magnitude for typical functions of interest.

Excellent (E.) After T iterations, $q_{(\alpha^T, \theta^T)}$ selects the best fitting mixture of two Gaussian densities, which in this somewhat artificial example corresponds to a perfect fit of π. Note, however that, the actual gain over the previous outcome is rather moderate with a reduction of variance by a factor less than four.

Of course, a very important parameter here is the IS sample size N: for a given initial IS density q_0, if N is too small, any method based on IS is bound to fail, conversely when N gets large all reasonable algorithms are expected to reach either the G. or E. result. Note that with local adaptive rules such as the ones proposed in this paper, it is not possible to guarantee that only the E. outcome will be achieved as the best fitting Gaussian IS density is indeed a stationary point (and in fact a

local minima) of the entropy criterion. So, depending on the initialisation, there always is a non zero probability that the algorithm converges to the G. situation only.

To focus on situations where algorithmic robustness is an issue, we purposely chose to select a rather small IS sample size of $N = 5,000$ points. As discussed above, direct IS estimates using q_0 as IS density would be mostly useless with such a modest sample size. We evaluated four algorithmic versions of the PMC algorithm. The first, *Plain PMC*, uses the parameter update formulas in (11) and q_0 is only used as an initialisation value, which is common to all D components of the mixture (which also initially have equal weights). Only the means of the components are slightly perturbed to make it possible for the adaptation procedure to actually provide distinct mixture components. One drawback of the plain PMC approach is that we do not ensure during the course of the algorithm that the adapted mixture IS density remains valid, in particular that it provides reliable estimates of the parameter update formulas. To guarantee that the IS weights stay well behaved, we consider a version of the PMC algorithm in which the IS density is of the form

$$(1-\alpha_0)\sum_{d=1}^{D}\alpha_d\mathcal{N}(\mu_d,\Sigma_d)+\alpha_0 q_0$$

with the difference that α_0 is a fixed parameter which is not adapted. The aim of this version, which we call *Defensive PMC* in reference to the work of Hesterberg (1995), is to guarantee that the importance function remains bounded by $\alpha_0^{-1}\pi(x)/q_0(x)$, whatever happens during the adaptation, thus guaranteeing a finite variance. Since q_0 is a poor IS density, it is preferable to keep α_0 as low as possible and we used $\alpha_0 = 0.1$ in all the following simulations. As detailed in both last lines of Table 1, this modification will typically slightly limit the performances achievable by the adaptation procedure, although this drawback could probably be avoided by allowing for a decrease of α_0 during the iterations of the PMC. The parameter update formulas for this modified mixture model are very easily deduced from (11) and are omitted here for the sake of conciseness. The third version we considered is termed *Rao-Blackwellised PMC* and consists in replacing the update equations (11) by their Rao-Blackwellised version (12). Finally, we consider a fourth option in which both the defensive mixture density and the Rao-Blackwellised update formulas are used.

All simulations were carried out using a sample size of $N = 5,000$, 20 iterations of the PMC algorithm and Gaussian mixtures with $D = 3$ components. Note that we purposely avoided to chose $D = 2$ to avoid the very artificial "perfect fit" phenomenon. This also means that for most runs of the algorithm, at least one component will disappear (by convergence of its weight to zero) or will be duplicated, with several components sharing very similar parameters.

	Disastrous	Mediocre	Good	Excellent
Plain	55	0	33	12
Defensive	13	51	30	6
R.-B.	18	1	70	11
Defensive + R.-B.	5	11	76	8

Table 2: Number of outcomes of each category for the four algorithmic versions, as recorded from 100 independent runs.

Table 2 display the performances of the four algorithms in repeated independent adaptation runs. The most significant observation about Table 2 is the large gap in robustness between the non Rao-Blackwellised versions of the algorithm, which returned disastrous or mediocre results in about 60% of the cases, a fraction that falls bellow 20% when the Rao-Blackwellised update formulas are used. Obviously the fact that the Rao-Blackwellised updates are based on all simulated values and not just on those actually simulated from a particular mixture component is a major source of robustness

of the method when the sample size N is small, given the misfit of the initial IS density q_0. The same remark also applies when the PMC algorithm is to be implemented with a large number D of components. The role of the defensive mixture component is more modest although it does improve the performances of both versions of the algorithm (non Rao-Blackwellised and Rao-Blackwellised altogether), at the price of a slight reduction of the frequency of the "Excellent" outcome. Also notice that the results obtained when the defensive mixture component is used are slightly beyond those of the unconstrained adaptation (see Table 1). The frequency of the perfect or "Excellent" match is about 10% for all methods but this is a consequence of the local nature of the adaptation rule as well as of the choice of the initialisation of the algorithm. It should be stressed however that as we are not interested in modelling π by a mixture but rather that we are seeking good IS densities, the solutions obtained in the G. or E. situations are only mildly different in this respect (see Table 1). As a final comment, recall that the results presented above have been obtained with a fairly small sample size of $N = 5,000$. Increasing N quickly reduces the failure rate of all algorithms: for $N = 20,000$ for instance, the failure rate of the plain PMC algorithm drops to 7/100 while the Rao-Blackwellised versions achieve either the G. or E. result (and mostly the G. one, given the chosen initialisation) for all runs.

4 Robustification via mixtures of multivariate t's

We now consider the setting of a proposal composed of a mixture of p-dimensional t distributions,

$$\sum_{d=1}^{D} \alpha_d \mathcal{T}(\nu_d, \mu_d, \Sigma_d)\,. \tag{13}$$

We here follow the recommendations of West (1992) and Oh and Berger (1993) who proposed using mixtures of t distributions in importance sampling. The t mixture is preferable to a normal mixture because of its heavier tails that can capture a wider range of non-Gaussian targets with a smaller number of components. This alternative setting is more challenging however and one must take advantage of the missing variable representation of the t distribution itself to achieve a closed-form updating of the parameters $(\mu_d, \Sigma_d)_d$ approximating (7), since a true closed-form cannot be derived.

4.1 The latent-data framework

Using the classical normal/chi-squared decomposition of the t distribution, a joint distribution associated with the t mixture proposal (13) is

$$\begin{aligned} f(x,y,z) &\propto \alpha_z |\Sigma_z|^{-1/2} \exp\left\{-(x-\mu_z)^{\mathrm{T}}\Sigma_z^{-1}(x-\mu_z)y/2\nu_z\right\} y^{(\nu_z+p)/2-1}e^{-y/2} \\ &\propto \alpha_z\, \varphi(x;\mu_z,\nu_z\Sigma_z/y)\, \varsigma(y;\nu_z/2,1/2)\,, \end{aligned}$$

where, as above, x corresponds to the observable in (13), z corresponds to the mixture indicator, and y corresponds to the χ^2_ν completion. The normal density is denoted by φ and the gamma density by ς. Both y and z correspond to latent variables in that the integral of the above in (y,z) returns (13).

In the associated PMC algorithm, we only update the expectations and the covariance structures of the t distributions and not the number of degrees of freedom, given that there is no closed-form solution for the later. In that case, $\theta_d = (\mu_d, \Sigma_d)$ and, for each $d = 1, \ldots, D$, the number of degrees of freedom ν_d is fixed.

At iteration t, the integrated EM update of the parameter will involve the following "E" function

$$Q\{(\alpha^t,\theta^t),(\alpha,\theta)\} = \mathbb{E}^X_\pi \left[\mathbb{E}^{Y,Z}_{(\alpha^t,\theta^t)} \left\{\log(\alpha_Z) + \log(\varphi(X;\mu_Z,\nu_Z\Sigma_z/Y))|\, X\right\}\right]\,,$$

since the χ^2 part does not involve the parameter $\theta = (\mu, \Sigma)$. Given that

$$Y, Z|X, \theta \sim f(y, z|x) \propto \alpha_z \, \varphi(x; \mu_z, \nu_z \Sigma_z / y) \, \varsigma(y; \nu_z/2, 1/2) \,,$$

we have that

$$Y|X, Z = d, \theta \sim \mathcal{G}a \left[(\nu_d + p)/2, \frac{1}{2} \left\{ 1 + (X - \mu_d)^{\mathrm{T}} \Sigma_d^{-1} (X - \mu_d)/\nu_d \right\} \right]$$

and therefore

$$\begin{aligned} Q\{(\alpha^t, \theta^t), (\alpha, \theta)\} &= \mathbb{E}_\pi^X \left[\mathbb{E}^Z_{(\alpha^t, \theta^t)} \left\{ \log(\alpha_Z) | X \right\} \right] \\ &\quad - \frac{1}{2} \mathbb{E}_\pi^X \left[\mathbb{E}^{Y,Z}_{(\alpha^t, \theta^t)} \left\{ \log |\Sigma_Z| + \frac{(X - \mu_Z)^{\mathrm{T}} {\Sigma_Z}^{-1} (X - \mu_Z) Y}{\nu_Z} \middle| X \right\} \right] \\ &= \mathbb{E}_\pi^X \left[\sum_{d=1}^D \rho_d(X; \alpha^t, \theta^t) \log(\alpha'_d) \right] \\ &\quad - \frac{1}{2} \, \mathbb{E}_\pi^X \left[\sum_{d=1}^D \rho_d(X; \alpha^t, \theta^t) \Big\{ \log |\Sigma_d| + (X - \mu_d)^{\mathrm{T}} {\Sigma_d}^{-1} (X - \mu_d) \right. \\ &\qquad \left. \times \frac{\nu_d + p}{\nu_d + (X - \mu_d^t)^{\mathrm{T}} (\Sigma_d^t)^{-1} (X - \mu_d^t)} \Big\} \right] , \end{aligned}$$

where we have used both the definition in (5),

$$\rho_d(X; \alpha^t, \theta^t) = \mathbb{P}_{\alpha^t, \theta^t}(Z = d|X) = \frac{\alpha_d^t t(x; \nu_d, \mu_d^t, \Sigma_d^t)}{\sum_{\ell=1}^D \alpha_\ell^t t(x; \nu_\ell, \mu_\ell^t, \Sigma_\ell^t)} \,,$$

with $t(x; \nu, \mu, \Sigma)$ denoting the $\mathcal{T}(\nu, \mu, \Sigma)$ density, and the fact that

$$\gamma_d(X; \theta^t) = \mathbb{E}^Y_{\theta^t} \left\{ Y/\nu_d | X, Z = d \right\} = \frac{\nu_d + p}{\nu_d + (X - \mu_d^t)^{\mathrm{T}} (\Sigma_d^t)^{-1} (X - \mu_d^t)} \,.$$

Therefore, the "M" step of the integrated EM update is

$$\begin{aligned} \alpha_d^{t+1} &= \mathbb{E}_\pi^X \left[\rho_d(X; \alpha^t, \theta^t) \right] \\ \mu_d^{t+1} &= \frac{\mathbb{E}_\pi^X \left[\rho_d(X; \alpha^t, \theta^t) \gamma_d(X; \theta^t) X \right]}{\mathbb{E}_\pi^X \left[\rho_d(X; \alpha^t, \theta^t) \gamma_d(X; \theta^t) \right]} \\ \Sigma_d^{t+1} &= \frac{\mathbb{E}_\pi^X \left[\rho_d(X; \alpha^t, \theta^t) \gamma_d(X; \theta^t) (X - \mu_d^{t+1})(X - \mu_d^{t+1})^{\mathrm{T}} \right]}{\mathbb{E}_\pi^X \left[\rho_d(X; \alpha^t, \theta^t) \right]} \,. \end{aligned}$$

While the first update is the generic weight modification (6), the latter formulae are (up to the integration with respect to X) essentially those found in Peel and McLachlan (2000) for a mixture of t distributions.

4.2 Parameter update

As in Section 3.1, the empirical update equations are obtained by using self-normalised IS with weights $\bar{\omega}_{i,t}$ given by (8) for both the numerator and the denominator of each of the above expressions. The

Rao-Blackwellised approximation based on (10) yields

$$\alpha_d^{t+1,N} = \sum_{i=1}^N \bar{\omega}_{i,t}\, \rho_d(X_{i,t};\alpha^{t,N},\theta^{t,N})\,,$$
$$\mu_d^{t+1,N} = \frac{\sum_{i=1}^N \bar{\omega}_{i,t}\, \rho_d(X_{i,t};\alpha^{t,N},\theta^{t,N})\, \gamma_d(X_{i,t};\theta^{t,N})\, X_{i,t}}{\sum_{i=1}^N \bar{\omega}_{i,t}\, \rho_d(X_{i,t};\alpha^{t,N},\theta^{t,N})\, \gamma_d(X_{i,t};\theta^{t,N})}\,,$$
$$\Sigma_d^{t+1,N} = \frac{\sum_{i=1}^N \bar{\omega}_{i,t}\, \rho_d(X_{i,t};\alpha^{t,N},\theta^{t,N})\, \gamma_d(X_{i,t};\theta^{t,N})\, (X_{i,t}-\mu_d^{t+1,N})(X_{i,t}-\mu_d^{t+1,N})^{\mathrm{T}}}{\sum_{i=1}^N \bar{\omega}_{i,t}\, \rho_d(X_{i,t};\alpha^{t,N},\theta^{t,N})}\,,$$

while the standard update equations, based on (9), are obtained by replacing $\rho_d(X_{i,t};\alpha^{t,N},\theta^{t,N})$ by $\mathbb{1}\{X_{i,t}=d\}$ in the above equations.

4.3 Pima Indian example

As a realistic if artificial illustration of the performances of the t mixture (13), we study the posterior distribution of the parameters of a probit model. The corresponding dataset is borrowed from the MASS library of R (R Development Core Team, 2006). It consists in the records of 532 Pima Indian women who were tested by the U.S. National Institute of Diabetes and Digestive and Kidney Diseases for diabetes. Four quantitative covariates were recorded, along with the presence or absence of diabetes. The corresponding probit model analyses the presence of diabetes, i.e.

$$\mathbb{P}_\beta(y=1|\mathbf{x}) = 1-\mathbb{P}_\beta(y=0|\mathbf{x}) = \Phi(\beta_0+\mathbf{x}^{\mathrm{T}}(\beta_1,\beta_2,\beta_3,\beta_4))$$

with $\beta=(\beta_0,\ldots,\beta_4)$, $\mathbf{x}$ made of four covariates, the number of pregnancies, the plasma glucose concentration, the body mass index weight in kg/(height in m)^2, and the age, and Φ corresponds to the cumulative distribution function of the standard normal. We use the flat prior distribution $\pi(\beta|\mathbf{X})\propto 1$; in that case, the 5-dimensional target posterior distribution is such that

$$\pi(\beta|\mathbf{y},\mathbf{X}) \propto \prod_{i=1}^{532} \left[\Phi\{\beta_0+(\mathbf{x}^i)^{\mathrm{T}}(\beta_1,\beta_2,\beta_3,\beta_4)\}\right]^{y_i} \left[1-\Phi\{\beta_0+(\mathbf{x}^i)^{\mathrm{T}}(\beta_1,\beta_2,\beta_3,\beta_4)\}\right]^{1-y_i}$$

where $\mathbf{x}^i$ is the value of the covariates for the i-th individuals and y_i is the response of the i-th individuals.

We first present some results for $N=10,000$ sample points and $T=500$ iterations on Figures 1—3, based on a mixture with 4 components and with the degrees of freedom chosen as $\nu=(3,6,9,18)$, respectively, when using the non Rao-Blackwellised version (9). The unrealistic value of T is chosen purposely to illustrate the lack of stability of the update strategy when not using the Rao-Blackwellised version. Indeed, as can be seen from Figure 1, which describes the evolution of the μ_d's, some components vary quite widely over iterations, but they also correspond to a rather stable overall estimate of β,

$$\sum_{i=1}^N \bar{\omega}_{i,T}\beta_{i,T}\,,$$

equal to $(-5.54, 0.051, 0.019, 0.055, 0.022)$ over most iterations. When looking at Figure 3, the quasi-constant entropy estimate after iteration 100 or so shows that, even in this situation, there is little need to perpetuate the iterations till the 500-th.

Using a Rao-Blackwellised version of the updates shows a strong stabilisation for the updates of the parameters α_d and (μ_d,Σ_d), both in the number of iterations and in the range of the parameters. The approximation to the Bayes estimate is obviously very close to the above estimation $(-5.63, 0.052, 0.019, 0.056, 0.022)$. Figures 4 and 5 show the immediate stabilisation provided by the Rao-Blackwellisation step.

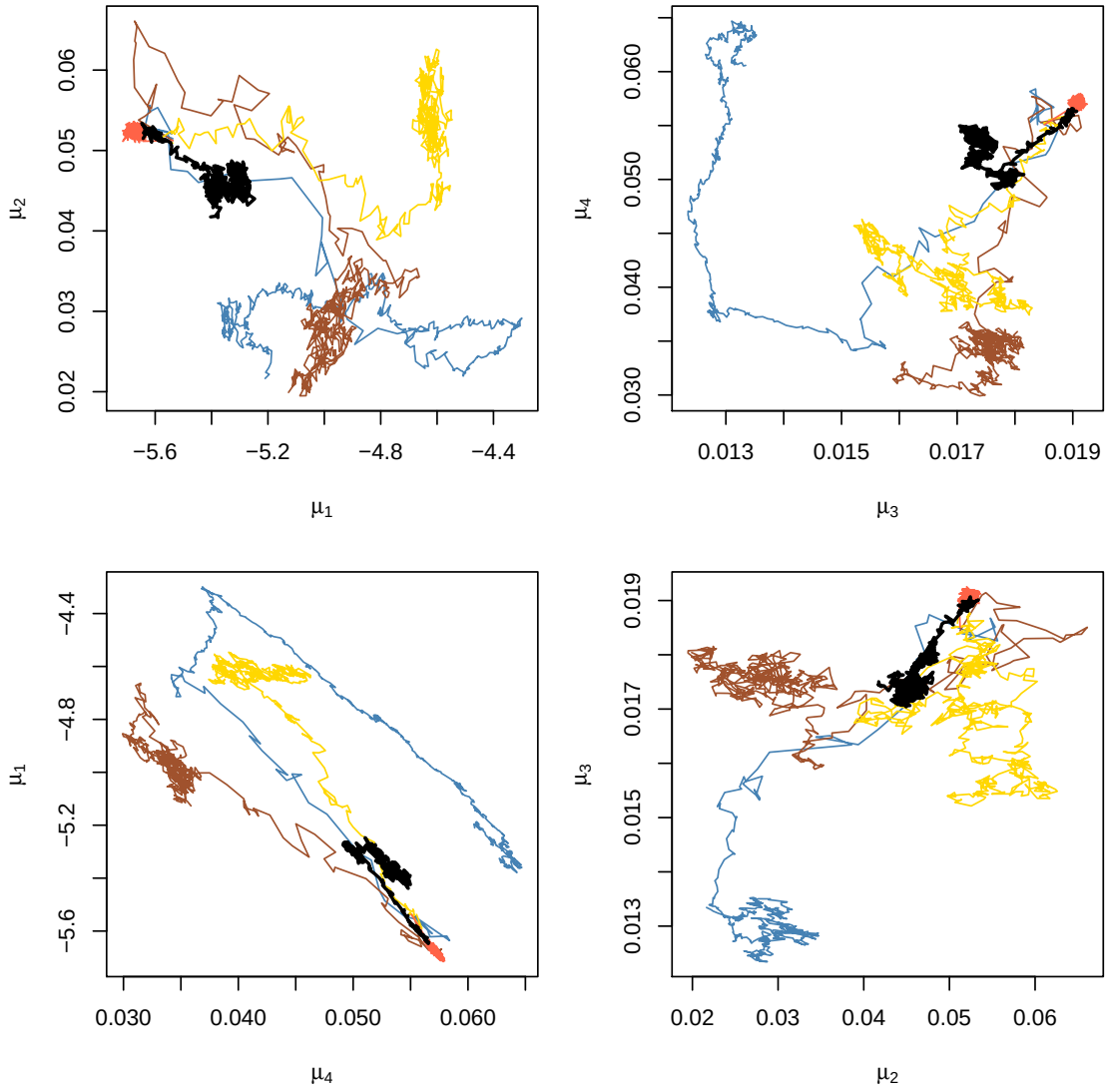

Figure 1: **Pima Indians:** Evolution of the components of the five μ_d's over 500 iterations plotted by pairs: *(clockwise from upper left side)* $(1,2)$, $(3,4)$, $(4,1)$ and $(2,3)$. The colour code is blue for μ_1, yellow for μ_2, brown for μ_3 and red for μ_4. The additional dark path corresponds to the estimate of β. All μ_d's were started in the vicinity of the MLE $\hat{\beta}$.

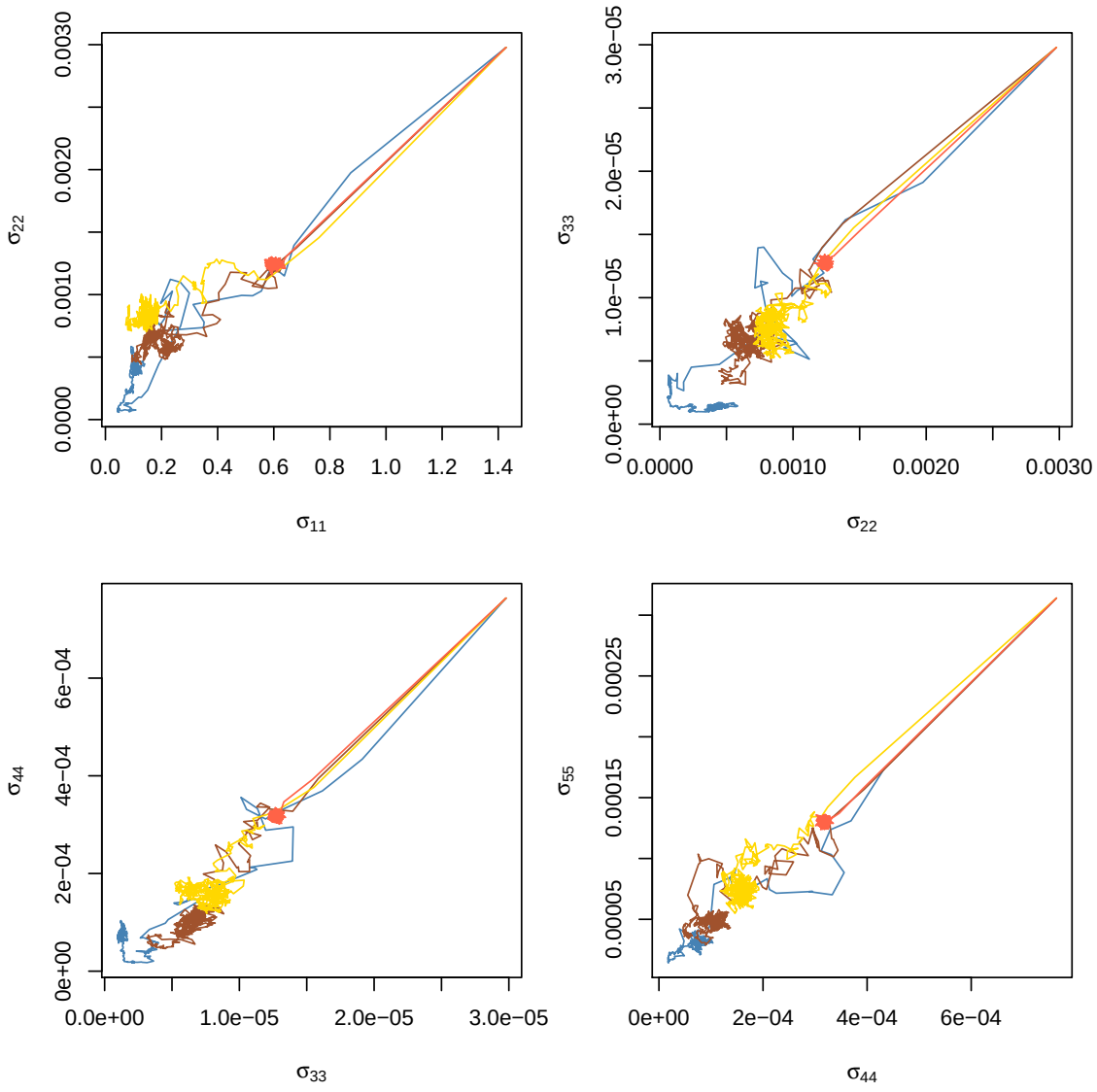

Figure 2: **Pima Indians:** Evolution of the five Σ_d's over 500 iterations plotted by pairs for the diagonal elements: *(clockwise from upper left side)* $(1,2)$, $(3,4)$, $(4,1)$ and $(2,3)$. The colour code is blue for Σ_1, yellow for Σ_2, brown for Σ_3 and red for Σ_4. All Σ_d's were started at the covariance matrix of $\hat{\beta}$ produced by `R glm()` procedure.

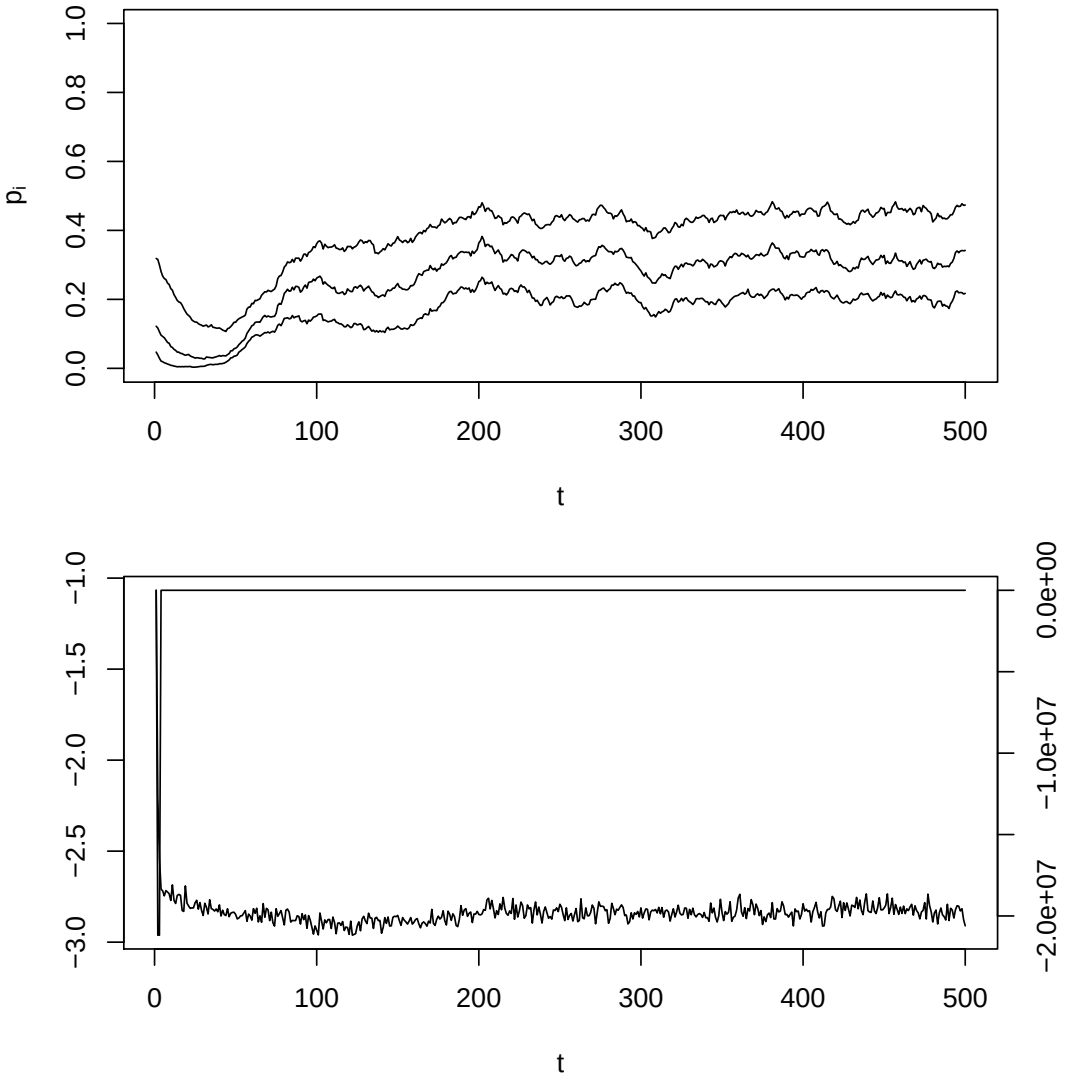

Figure 3: **Pima Indians:** Evolution of the cumulated weights *(top)* and of the estimated entropy divergence $\mathbb{E}^{\pi}[\log(q_{\alpha,\theta}(\beta))]$ *(bottom)*.

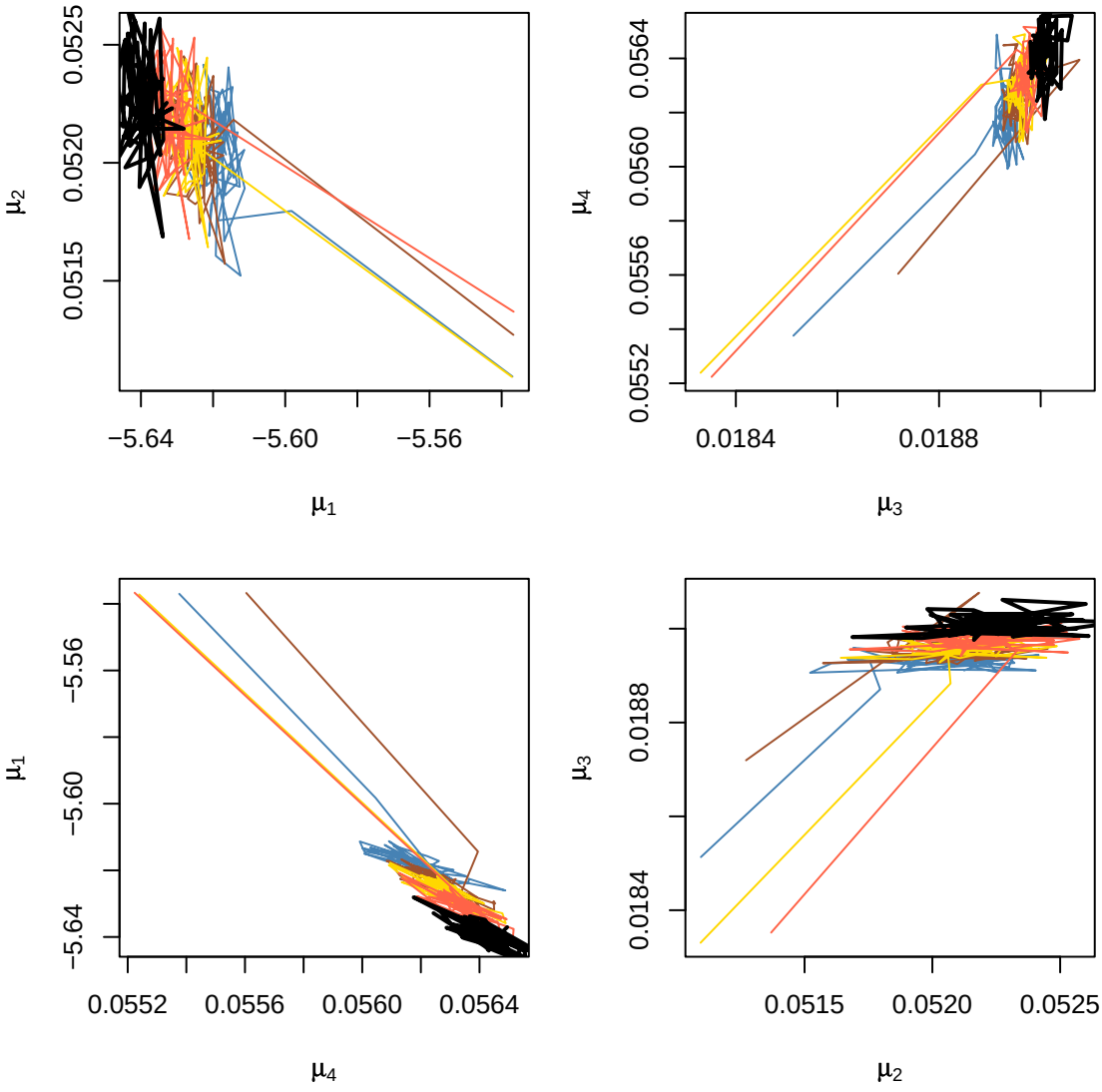

Figure 4: **Pima Indians:** Evolution of the components of the five μ_d's over 50 Rao-Blackwellised iterations plotted by pairs: *(clockwise from upper left side)* $(1,2)$, $(3,4)$, $(4,1)$ and $(2,3)$. The colour code is blue for μ_1, yellow for μ_2, brown for μ_3 and red for μ_4. The additional dark path corresponds to the estimate of β. All μ_d's were started in the vicinity of the MLE $\hat{\beta}$.

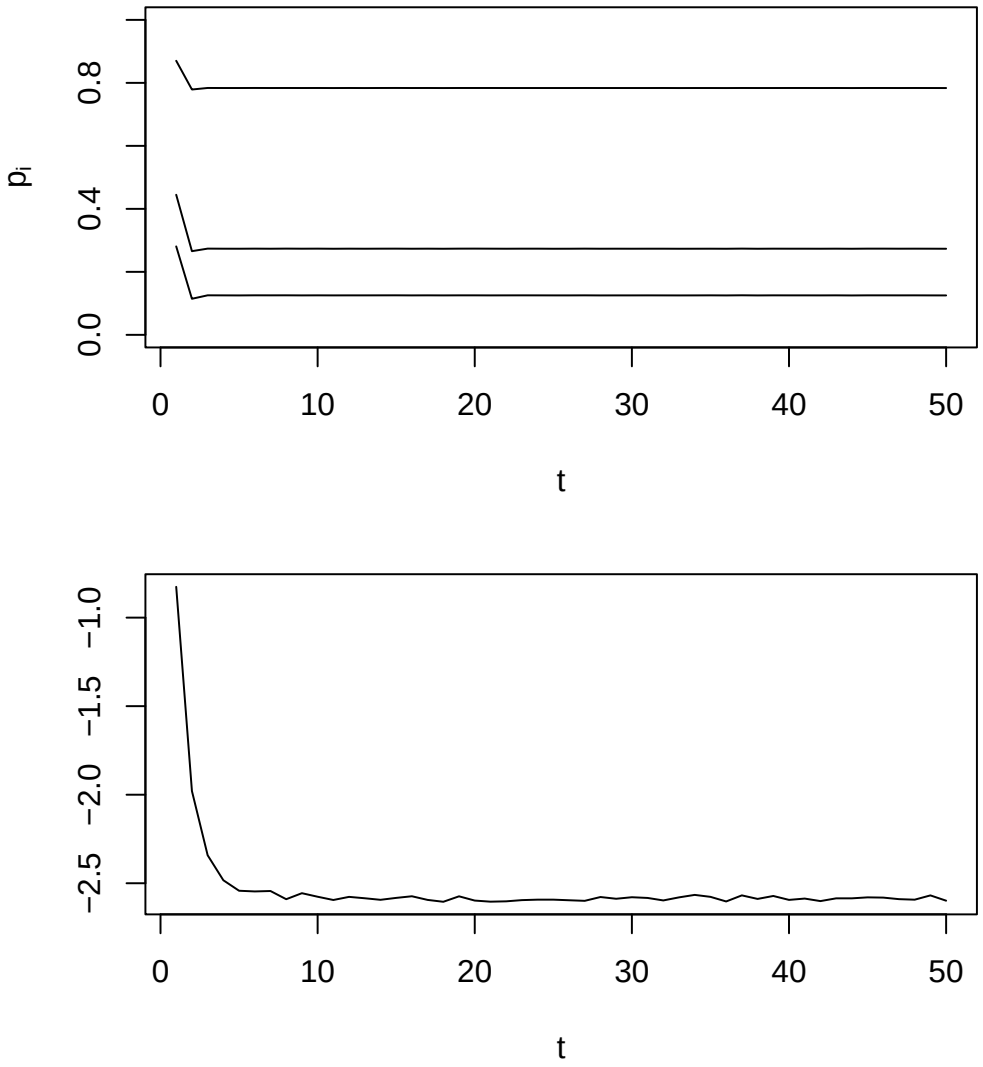

Figure 5: **Pima Indians:** Evolution of the cumulated weights *(top)* and of the estimated entropy divergence $\mathbb{E}^{\pi}[\log(q_{\alpha,\theta}(\beta))]$ *(bottom)* for the Rao-Blackwellised version.

5 Conclusions

The proposed algorithm provides a flexible and robust framework for adapting general importance sampling densities represented as mixtures. The extension to mixtures of t distribution broadens the scope of the method by allowing approximation of heavier tail targets. Moreover, we can extend here the remarks made in Douc et al. (2007a,b), namely that the update mechanism provides an early stabilisation of the parameters of the mixture. It is therefore unnecessary to rely on a large value of T: with large enough sample sizes N at each iteration—especially on the initial iteration that requires many points to counter-weight a potentially poor initial proposal—, it is quite uncommon to fail to spot a stabilisation of both the estimates and of the entropy criterion within a few iterations.

While this paper relies on the generic entropy criterion to update the mixture density, we want to stress that it is also possible to use a more focussed deviance criterion, namely the h-entropy

$$\mathfrak{E}_h(\pi, q_{(\alpha,\theta)}) = D(\pi_h \| q_{(\alpha,\theta)}) \,, \tag{14}$$

with

$$\pi_h(x) \propto |h(x) - \pi(h)|\pi(x) \,,$$

that is tuned to the estimation of a particular function h, as it is well-known that the optimal choice of the importance density for the self-normalised importance sampling estimator is exactly π_h. Since the normalising constant in π_h does not need to be known, one can derive an adaptive algorithm that resembles the method presented in this paper. It is expected that this modification will be helpful in reaching IS densities that provide a low approximation error for a specific function h, which is also an important feature of importance sampling in several applications.

References

Cappé, O., Guillin, A., Marin, J., and Robert, C. (2004). Population Monte Carlo. *J. Comput. Graph. Statist.*, 13(4):907–929.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer-Verlag, New York.

Chen, R. and Liu, J. S. (1996). Predictive updating method and Bayesian classification. *J. Royal Statist. Soc. Series B*, 58(2):397–415.

Douc, R., Guillin, A., Marin, J.-M., and Robert, C. (2007a). Convergence of adaptive mixtures of importance sampling schemes. *Ann. Statist.*, 35(1):420–448.

Douc, R., Guillin, A., Marin, J.-M., and Robert, C. (2007b). Minimum variance importance sampling via population Monte Carlo. *ESAIM: Probability and Statistics*, 11:427–447.

Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57:1317–1340.

Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194.

Oh, M. and Berger, J. (1993). Integration of multimodal functions by Monte Carlo importance sampling. *J. American Statist. Assoc.*, 88:450–456.

Peel, D. and McLachlan, G. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, 10:339–348.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, New York, second edition.

Rubinstein, R. Y. and Kroese, D. P. (2004). *The Cross-Entropy Method*. Springer-Verlag, New York.

West, M. (1992). Modelling with mixtures. In Berger, J., Bernardo, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics 4*, pages 503–525. Oxford University Press, Oxford.