INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2010-33

Free Energy Methods for Efficient Exploration of Mixture Posterior Densities

N. CHOPIN[1] – T. LELIÈVRE[2]
G. STOLTZ[3]

1 ENSAE-CREST, 3 avenue Pierre Larousse, 92245 Malakoff, France.
Corresponding author : nicolas.chopin@ensae.fr

2 Université Paris Est, CERMICS, Projet MICMAC Ecole des Ponts ParisTech-INRIA. 6 & 8 Avenue Pascal, 77455 Marne-la-Vallée Cédex 2, France.

3 Université Paris Est, CERMICS, Projet MICMAC Ecole des Ponts ParisTech-INRIA. 6 & 8 Avenue Pascal, 77455 Marne-la-Vallée Cédex 2, France.

Free energy methods for efficient exploration of mixture posterior densities

Nicolas Chopin[1*], Tony Lelièvre[2] and Gabriel Stoltz[2]

1: ENSAE-CREST, 3, Avenue Pierre Larousse, 92245 Malakoff, France.

2: Université Paris Est, CERMICS, Projet MICMAC Ecole des Ponts ParisTech - INRIA

6 & 8 Av. Pascal, 77455 Marne-la-Vallé Cedex 2, France.

March 29, 2010

Abstract

Because of their multimodality, mixture posterior densities are difficult to sample with standard Markov chain Monte Carlo (MCMC) methods. We propose a strategy to enhance the sampling of MCMC in this context, using a biasing procedure which originates from computational statistical physics. The principle is first to choose a "reaction coordinate", that is, a direction in which the target density is multimodal. In a second step, the marginal log-density of the reaction coordinate is estimated; this quantity is called "free energy" in the computational statistical physics literature. To this end, we use adaptive biasing Markov chain algorithms which adapt their invariant distribution on the fly, in order to overcome sampling barriers along the chosen reaction coordinate. Finally, we perform an importance sampling step in order to remove the bias and recover the true posterior. The efficiency factor can easily be estimated *a priori* once the bias is known, and is large enough for the test cases we considered.

A crucial point is the choice of the reaction coordinate. One standard choice (used for example in the classical Wang-Landau algorithm) is the opposite of the log-posterior density. We show that another convenient and efficient reaction coordinate is the hyper-parameter that determines the order of magnitude of the variance of each component. We also show how to adapt the method to perform model choice between different number of components. We illustrate our approach by analyzing two real data sets.

Keywords: Adaptive Biasing Force; Adaptive Biasing Potential; Adaptive Markov chain Monte Carlo; Importance sampling; Mixture models.

1 Introduction

Mixture modelling is presumably the most popular and the most flexible way to model heterogeneous data; see e.g. Titterington et al. (1986), McLachlan and Peel (2000) and Frühwirth-Schnatter (2006) for an overview of applications of mixture models. Due to the emergence of Markov chain Monte Carlo (MCMC), interest in the Bayesian analysis of such models has sharply increased in recent years, starting with Diebolt and Robert (1994). However, MCMC analysis of mixtures poses several problems (Celeux et al., 2000; Jasra et al., 2005). Here, we focus on the difficulties arising from the multimodality of the posterior distribution.

For the sake of exposition, we concentrate our discussion on Gaussian mixtures, but we explain in the conclusion (Section 6) how our ideas may be extended to other mixture models. The data vector $y = (y_1, \ldots, y_n)$ contains independent and identically distributed (i.i.d.) real random

*Corresponding author: nicolas.chopin@ensae.fr

variables with density

$$p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta), \qquad p(y_i|\theta) = \sum_{k=1}^{K} q_k \, \varphi(y_i; \mu_k, \lambda_k^{-1}), \tag{1}$$

where the vector θ contains all the unknown parameters, i.e. the mixture weights $q_1, ..., q_{K-1}$, the means $\mu_1, \ldots, \mu_K$, the precisions $\lambda_1, \ldots, \lambda_K$, and possibly hyper-parameters; and $\varphi(\cdot; \mu, \lambda^{-1})$ denotes the Gaussian density with mean μ and variance λ^{-1}.

This model is not identifiable, because both the likelihood and the posterior density are invariant by permutation of components, provided the prior is symmetric. This is one reason for the aforementioned problems. By construction, any local mode in the posterior density admits $K! - 1$ symmetric replicates, while a typical MCMC sampler recovers only one of these $K!$ copies. A possible remedy is to introduce steps that permute randomly the components (Frühwirth-Schnatter, 2001). However, mixture posterior densities are often "genuinely multimodal", following the terminology of Jasra et al. (2005): the number of sets of $K!$ equivalent modes is often larger than one; see also Marin and Robert (2007, Chap. 6) for an example of a multimodal posterior density generated by an identifiable mixture model, and Section 2.2 of this paper for an example based on real data. A standard MCMC method rarely escapes from one multimodal region before entering another one: the MCMC dynamics is said to be "metastable". Following Celeux et al. (2000) and Jasra et al. (2005), we consider that a minimum requirement of convergence for a MCMC or similar sampler is to visit all possible labelling of the parameter, without resorting to random permutations. Our aim is to develop samplers that meet this requirement, using an adaptive importance sampling strategy (Darve and Pohorille, 2001; Hénin and Chipot, 2004; Marsili et al., 2006; Lelièvre et al., 2007).

The method we propose is inspired from algorithms developed in the molecular dynamics community. The principle is (i) to choose a "direction of metastability", namely a low-dimensional function of the parameters θ (called the "reaction coordinate" in the molecular dynamics literature); (ii) to compute the marginal log-density of this reaction coordinate (the opposite of this log-density is called the "free energy" in the molecular dynamics literature); and (iii) to use the free energy to bias the MCMC sampler, and hopefully get rid of the metastabilities of the original MCMC dynamics. The normalizing constant (the marginal likelihood) of, and expectations with respect to, the posterior distribution may then be computed using a simple importance sampling step from the biased posterior density to the actual posterior density. Compared to classical importance sampling strategies, the difference here is that the importance function is not given *a priori*, but computed automatically once a reaction coordinate has been chosen.

If the reaction coordinate is well-chosen, it is likely that the free-energy biased dynamics converges to equilibrium much faster than the original dynamics. Indeed, when the free energy is added to the posterior log-density, the marginal distribution in the reaction coordinate is uniform (within chosen bounds). Two questions are then in order: how to choose the reaction coordinate? and how to compute the free energy?

Concerning the choice of the reaction coordinate, the application of adaptive biasing methods to mixture models is not straightforward. We show that the degree of metastability of a mixture posterior density is strongly determined by certain hyper-parameters, typically those that calibrate in the prior the spread of each Gaussian component. We show that such hyper-parameters are good reaction coordinates, in the sense that the corresponding adaptive algorithm computes efficiently the associated free energy, and the corresponding biased dynamics explores quickly the (biased) posterior distribution. Other reaction coordinates are discussed, such as the posterior log-density, which is also a very good reaction coordinate (with the problem however that an appropriate range of variation for this reaction coordinate, which is needed for the method, is difficult to determine *a priori*). This reaction coordinate is the standard choice for the Wang-Landau algorithm, see for instance Liang (2005) and Atchadé and Liu (2010) for related works in Statistics.

To compute the free energy, we resort to adaptive biasing algorithms. Note that, in contrast with the adaptive MCMC algorithms considered in the statistical literature (see the review of Andrieu and Thoms (2008) and references therein), these adaptive algorithms modify sequentially

the invariant distribution of the Markov chain, instead of the parameters of the Markov kernel. Such algorithms were initially designed to compute the free energy in molecular dynamics. Note that several other standard techniques exist in molecular dynamics for computing the free energy, such as e.g. thermodynamic integration, see Lelièvre et al. (2010) for a recent review.

We measure the efficiency of the overall algorithm using two criteria: the efficiency of the sampling of the biased dynamics and the relevance of the samples generated by the biased dynamics for the sampling of the original posterior distribution. For the first criterion, and as mentioned above, the basic requirement is to check that the dynamics does not remain stuck in a local minimum, and that permutations of the labelling of the parameters are visited. For the second criterion, we compute the efficiency factor (or the effective sample size) of the biased sample.

The aim of this paper is thus twofold: (i) to demonstrate the interest of importance sampling methods based on marginal laws or free energies associated to well chosen reaction coordinates for the sampling of posterior distributions; (ii) to propose an efficient implementation within MCMC algorithms of adaptive biasing methods which have been proposed in the molecular dynamics community to efficiently compute the free energy associated with a given reaction coordinate.

Since our method is some importance sampling technique, it is of course possible to use it in addition to other strategies aimed at overcoming metastability issues, such as tempering methods; see for instance Iba (2001); Neal (2001) and references therein, as well as Celeux et al. (2000) for an application to mixture posteriors.

The paper is organized as follows. Section 2 presents the Gaussian mixture model, and the difficulties encountered with classical MCMC algorithms (metastable dynamics). Section 3 describes the method we propose, which is based on adaptive algorithms to compute free energies. Section 4 explains how to perform Bayesian inference on mixture models using the free energy, with in particular a discussion of the choice of the reaction coordinate and of the importance sampling step. Section 5 illustrates our approach with two real data-sets. Section 6 explains how to generalize our method to other mixture models, and discusses further research directions.

2 The model and the metastabilities of naive algorithms

2.1 Posterior distribution

As explained in the introduction, we focus on the Gaussian mixture model (1), associated with the following prior, taken from Richardson and Green (1997), which is symmetric with respect to the components $k = 1, \ldots, K$:

$$\begin{aligned} \mu_k &\sim \mathrm{N}(M, \kappa^{-1}), \\ \lambda_k &\sim \mathrm{Gamma}(\alpha, \beta), \\ \beta &\sim \mathrm{Gamma}(g, h), \\ (q_1, \ldots, q_{K-1}) &\sim \mathrm{Dirichlet}_K(1, \ldots, 1). \end{aligned}$$

In our examples, we take $\kappa = 4/R^2$, $\alpha = 2$, $g = 0.2$, $h = 100g/\alpha R^2$, where R and M are respectively the range and the mean of the observed data, as in Jasra et al. (2005). The posterior density reads

$$\begin{aligned} p(\theta|y) &= \frac{1}{Z_K} p(\theta) p(y|\theta) \qquad (2) \\ &= \frac{\kappa^{K/2} g^h \beta^{K\alpha+g-1}}{Z_K \Gamma(\alpha)^K \Gamma(g) (2\pi)^{\frac{n+K}{2}}} \left(\prod_{k=1}^K \lambda_k \right)^{\alpha-1} \exp\left\{ -\frac{\kappa}{2} \sum_{k=1}^K (\mu_k - M)^2 - \beta \left(h + \sum_{k=1}^K \lambda_k \right) \right\} \\ &\quad \times \prod_{i=1}^n \left[\sum_{k=1}^K q_k \lambda_k^{1/2} \exp\left\{ -\frac{\lambda_k}{2} (y_i - \mu_k)^2 \right\} \right]. \end{aligned}$$

In this expression, θ is the vector

$$\theta = (q_1, \dots, q_{K-1}, \mu_1, \dots, \mu_K, \lambda_1, \dots, \lambda_K, \beta) \in \Omega = \mathcal{S}_K \times \mathbb{R}^K \times (\mathbb{R}_+)^{K+1}, \tag{3}$$

the set

$$\mathcal{S}_K = \left\{ (q_1, \dots, q_{K-1}) \in (\mathbb{R}_+)^{K-1}, \quad \sum_{i=0}^{K-1} q_i \leq 1 \right\}$$

is the probability simplex, and

$$Z_K = \int_\Omega p(\theta) p(y|\theta) \, d\theta \tag{4}$$

is the normalizing constant (namely the marginal likelihood in y), which depends on K and y.

The sampling of the posterior distribution (3) is the focus of the paper. Such problems arise in other contexts, such as computational statistical physics (see for instance (Balian, 2007)), where Boltzmann-Gibbs measures

$$\pi(x) = \frac{1}{Z} \exp\left\{-V(x)\right\}, \qquad Z = \int_\Omega \exp\left\{-V(x)\right\} \, dx, \tag{5}$$

(with $x \in \Omega \subset \mathbb{R}^d$) must be sampled in order to compute average thermodynamic properties of the system under consideration.

2.2 Metastability

Probability densities such as the posterior density (2) for mixture models, or the Boltzmann-Gibbs density (5) for many systems in statistical physics, are often multimodal. Standard sampling strategies, based on the Hastings-Metropolis algorithm (Metropolis et al., 1953; Hastings, 1970) for instance, are therefore typically metastable. The trajectory remains stuck in some local minima of the potential V (namely the opposite of the log-posterior density in a Bayesian context), which prevents an efficient sampling of the target measure. We refer to Figure 1 for an illustration of this phenomenon for the sampling of the posterior (3), for the Fishery data (see Section 5.1) and for the Hidalgo stamps data (see Section 5.2), for $K = 3$ components. Very few mode switchings (if any) are observed, which is symptomatic of a bad sampling of the parameter space. A simple random walk Hastings-Metropolis was used to produce these plots, but we obtained the same type of plots (not shown) with Gibbs sampling.

The top right plot of Figure 1 also represents a Gaussian mixture probability density (in dashed line) which corresponds to a very unlikely local mode of the posterior density for the Hidalgo dataset. In numerical tests not reported here, when this local mode is used as a starting point, the Hastings-Metropolis sampler used needs about $T = 10^5$ iterations to leave the attraction of this meaningless mode. It is easy to make this problem worse by slightly changing the data. For instance, T was increased to 10^7 by adding 2 to the three largest y_i's (while this local mode remained remained of very small probability). This local mode illustrates the typical "genuine multimodality" of mixture posteriors mentioned in the introduction – multimodality which cannot be cured by label permutation.

3 Free-energy biased sampling

3.1 Principle of the method

Consider a generic probability density $\pi(x)$, with $x \in \Omega \subset \mathbb{R}^d$ as defined in (5). Without any additional assumptions, it is difficult to find a remedy to the metastability of π. Consider the conditional probability measures

$$\pi^\xi(dx \mid \xi(x) = z) = \frac{\exp\{-V(x)\} \, \delta_{\xi(x)-z}(dx)}{\displaystyle\int_{\Sigma(z)} \exp\{-V(x')\} \, \delta_{\xi(x')-z}(dx')},$$

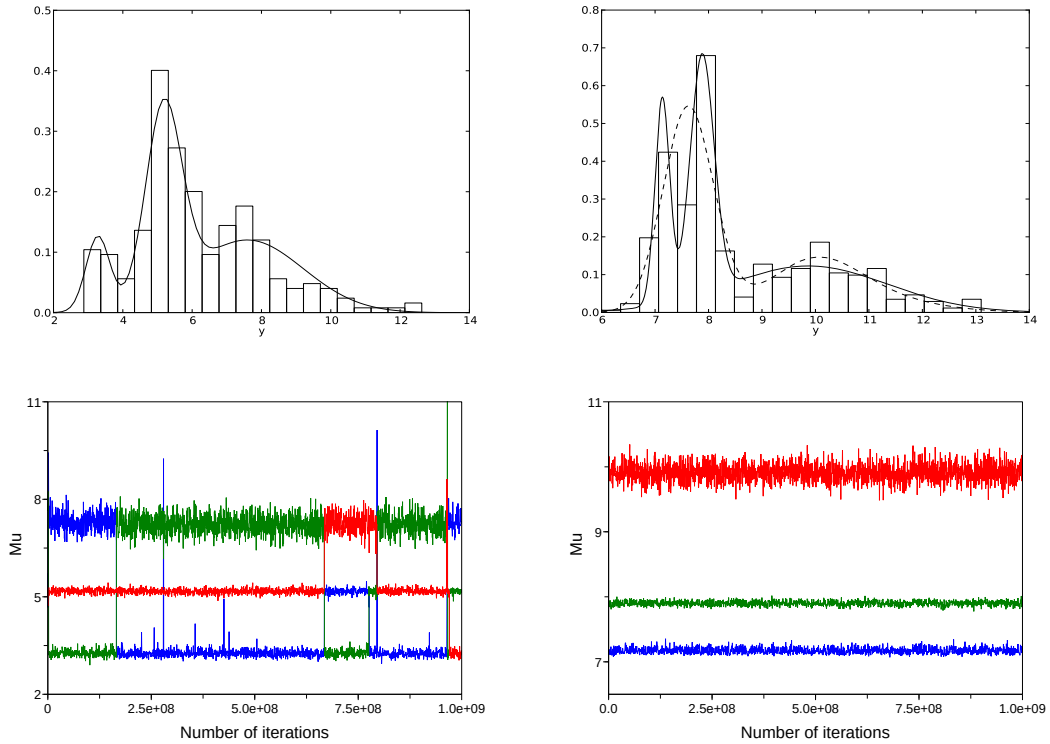

Figure 1: Left (resp. right) hand side corresponds to Fishery (resp. Hidalgo stamps) dataset; see Section 5 for details. Top row: histograms of the data and an estimated 3-component Gaussian mixture probability density function (solid line). The dashed line corresponds to a local mode of the posterior density. Bottom row: MCMC trajectories of (μ_1, μ_2, μ_3) as a function of the number of iterations. A simple random walk Hastings-Metropolis algorithm is used, see Section 5 for more details.

where $\delta_{\xi(x)-z}(dx)$ is a measure with support

$$\Sigma(z) = \Big\{ x \in \Omega \;\Big|\; \xi(x) = z \Big\},$$

defined by the formula: for all smooth test functions φ,

$$\int_\Omega \varphi(x)\, dx = \int \int_{\Sigma(z)} \varphi(x)\, \delta_{\xi(x)-z}(dx)\, dz.$$

The main assumption behind free-energy biased methods is that there exists a function $\xi : \Omega \to \omega \subset \mathbb{R}$, such that the sampling of $\pi^\xi(dx \mid \xi(x) = z)$ is significantly easier than the sampling of $\pi(x)$, at least for some values of z, see (Lelièvre et al., 2008; Lelièvre and Minoukadeh, 2010) for more details on the mathematical context, and a precise quantification of this statement in a functional analysis framework. The function ξ is called the *reaction coordinate* in the physics literature, since the transitions from one value of ξ to another are associated with the progress of some chemical event at a coarser scale (chemical reaction or change of conformation for example), which happens much slower than the typical microscopic fluctuations of the system around its average configuration. Extensions to reaction coordinates with values in $\mathbb{R}^m$ with $m \geq 2$ are possible (Lelièvre et al., 2010).

Let us now introduce the free energy $A(z)$ defined as

$$\mathrm{e}^{-A(z)} = \int_{\Sigma(z)} \exp\left\{-V(x)\right\}\, \delta_{\xi(x)-z}(dx), \tag{6}$$

see for instance Section 5.6 in Balian (2007). The free energy is the log-marginal density of the reaction coordinate. When the reaction coordinate is some coordinate of the system, for instance $\xi(x) = x_1$, and the domain Ω is a product $\Omega = \omega^d$, the free-energy is simply, up to an additive constant, equal to

$$A(x_1) = -\log\left(\int_{\omega^{d-1}} \mathrm{e}^{-V(x)}\, dx_2 \ldots dx_d \right).$$

The function $A(z)$ may be used to define a biased distribution according to

$$\pi_A(x) = \frac{1}{Z_A} \exp\left\{-V(x) + (A \circ \xi)(x)\right\}, \quad Z_A = \int_\Omega \exp\left\{-V(x) + (A \circ \xi)(x)\right\} dx,$$

where $(A \circ \xi)(x) = A(\xi(x))$. Typically, the sampling of π_A is easier than the sampling of π since (i) we assumed that the sampling of the conditional probability measures $\pi_A^\xi(dx \mid \xi(x) = z) = \pi^\xi(dx \mid \xi(x) = z)$ is easy at least for some values of z, meaning that there is no metastability in the direction orthogonal to ξ at those values of z (see again (Lelièvre et al., 2008; Lelièvre and Minoukadeh, 2010) for more mathematical precisions). This allows to switch from one mode to another; (ii) the marginal of π_A along ξ is uniform, which ensures that the visits to these regions without metastability are numerous enough.

3.2 Computing the free energy by adaptive methods

In most cases, the free energy A defined in (6) does not admit a closed-form expression, and must be estimated. We present in this section some possible techniques to this end.

3.2.1 General structure of adaptive methods

In adaptive biasing Markov chain Monte Carlo methods, a time-varying biasing potential $A_t(z)$ is considered. The biasing potential A_t is sequentially updated in order to converge to the free energy A in the limit. As already mentioned in the introduction, the term "adaptive" refers in this paper to the dynamic adaptation of the targeted probability measure, and not of the parameters of a Markov kernel used in the simulations.

Adaptive methods can be classified into two categories, depending on whether the potential $A_t(z)$ or its derivative $A'_t(z)$ with respect to z are updated. Instances of the first strategy, called adaptive biasing potential (ABP) method, include non-equilibrium metadynamics (Bussi et al., 2006; Raiteri et al., 2006), the Wang-Landau algorithm (Wang and Landau, 2001a,b) or the self-healing umbrella sampling (Marsili et al., 2006), a version of which is used in this work. The adaptive biasing force (ABF) methodology (Darve and Pohorille, 2001; Hénin and Chipot, 2004; Lelièvre et al., 2007), which is also used here, is an instance of the second.

Let us now describe how to implement these methods within a MCMC algorithm. At iteration t (where t is an integer), a current estimate A_t of the free energy A associated with the reaction coordinate ξ is obtained. The time-varying target density is therefore

$$\pi_{A_t}(x) = \frac{1}{Z_{A_t}} \exp\left\{-V(x) + (A_t \circ \xi)(x)\right\}. \tag{7}$$

This probability density defines the Boltzmann-Gibbs measure associated to the modified potential $V - A_t \circ \xi$. An adaptive MCMC algorithm simulates a non-homogeneous Markov chain (x_t), $t = 1, 2, \ldots$, using the two following steps at iteration t:

(1) a MCMC move according to the current target π_{A_t} defined in (7),

$$x_{t+1} \sim K_t(x_t, \cdot),$$

where K_t is a Markov kernel leaving π_{A_t} invariant;

(2) the update of the bias A_{t+1}, using a trajectorial average, see below.

The first step may be done with a Hastings-Metropolis scheme for instance, see Figure 2 below.

Before explaining the second step, we make two basic remarks. First, the biasing potential cannot be estimated on the whole space ω, but only on a bounded subset $\widetilde{\omega}$, typically some interval $\widetilde{\omega} = [z_{\min}, z_{\max}] \subset \omega \subset \mathbb{R}$. (We discuss the choice of $\widetilde{\omega}$ for various reaction coordinates for the specific model at hand in Section 4.2.) Second, a proper discretization of the biasing potential should be chosen. A simple strategy, which we adopt in this paper, is to use predefined bins, and approximate the biasing potential and its derivative by piecewise constant functions. Specifically, we consider N_z bins of equal sizes Δz,

$$\widetilde{\omega} = [z_{\min}, z_{\max}] = \bigcup_{i=0}^{N_z-1} [z_i, z_{i+1}], \quad z_i = z_{\min} + i\Delta z, \quad \Delta z = \frac{z_{\max} - z_{\min}}{N_z}.$$

3.2.2 Two important strategies to update the bias

There are basically two strategies for the update step. The ABP strategy based on (Marsili et al., 2006) consists in updating A_{t+1} as follows. The biasing potential for $z \in (z_i, z_{i+1})$ is initially set to $\exp\{-A_0(z)\} = 1/N_z$, and then updated for $t \geq 1$ as

$$\forall z \in (z_i, z_{i+1}), \qquad \exp\{-A_t(z)\} = \frac{1}{Z_t} \left(1 + \sum_{j=1}^{t-1} \mathbb{1}\{z_i \leq \xi(x_j) < z_{i+1}\} \exp\left[- A_j \circ \xi(x_j)\right] \right), \tag{8}$$

the normalization factor Z_t being such that

$$\sum_{i=0}^{N_z-1} \exp\left[-A_t\left(\frac{z_i + z_{i+1}}{2} \right) \right] = 1.$$

Notice that $\exp\{-A_t(z)\}$ is indeed an estimator of the probability that $\xi(x) \in (z_i, z_{i+1})$ when x is distributed according to the *unbiased* density π, since the weight $\exp\left[- A_j(\xi(x_j))\right]$, proportional to $\pi(x_j)/\pi_{A_t}(x_j)$, corrects for the bias introduced at iteration t. It is easy to check that, provided

the method converges, the only possible limit for the biasing potential A_t is the free energy A (or more precisely a discretized version of A).

The ABF strategy (Darve and Pohorille, 2001; Hénin and Chipot, 2004; Lelièvre et al., 2007) is based on the following formula for the derivative of A (called the *mean force*):

$$A'(z) = F(z) = \mathbb{E}^{\pi}\Big(f(x)\,\Big|\,\xi(x) = z\Big), \tag{9}$$

where f admits an analytic expression in terms of ξ and V:

$$f = \frac{\nabla V \cdot \nabla \xi}{|\nabla \xi|^2} - \operatorname{div}\left(\frac{\nabla \xi}{|\nabla \xi|^2}\right), \tag{10}$$

where ∇ is the gradient operator, and div is the divergence operator. In the simple case when $\xi(x) = x_1$, this expression simplifies to $f = \partial V/\partial x_1$. As mentioned above, the conditional measure of π with respect to ξ is the same as the conditional measure of π_{A_t} with respect to ξ. Thus, a natural ABF updating strategy is to compute at iteration t the following approximation of the mean force:

$$\forall z \in (z_i, z_{i+1}), \qquad F_t(z) = \frac{\displaystyle\sum_{j=1}^{t} f(x_j)\,\mathbb{1}\{z_i \le \xi(x_j) \le z_{i+1}\}}{\displaystyle\sum_{j=1}^{t} \mathbb{1}\{z_i \le \xi(x_j) \le z_{i+1}\}}. \tag{11}$$

From this approximation of F, an approximation A_t of the bias can be recovered by integrating F_t. As above, it can be shown that, provided the method converges, the biasing potential A_t converges to (a discretized version of) the free energy A, up to an unimportant additive constant.

The interest of ABP compared to ABF is that it does not require computing f, see (10), which is sometimes cumbersome. On the other hand, it is observed that the ABF method yields very good results since the derivative of A_t is approximated, so that after integration, the adaptive biasing potential is smoother in z for ABF than for ABP. In the following, we use the ABF method, except when we consider as a reaction coordinate the potential V itself (minus the log posterior density in the mixture context), for which the computation of f is cumbersome. In this case, the ABP method is used.

The long-time convergence of the adaptive biasing force method (using conditional expectations rather than trajectorial averages), has been studied in (Lelièvre et al., 2008), and its associated discretization using many replica of the simulated Markov chain has been considered in Jourdain et al. (2009). For refinements concerning the implementation of such a strategy, we refer to Lelièvre et al. (2007).

3.2.3 Practical implementation of adaptive algorithms

To summarize the method, we give in Figure 2 the details of the ABF algorithm. A similar algorithm is used in the ABP case. In practice, we stop the algorithm when the bias A_t is no longer significantly modified from $t = n$ to $t = n + N_{\text{cvg}}$, where N_{cvg} is the number of iterations between two convergence checks. See Section 5.1.1 for an illustration of this strategy.

The so-obtained bias is then considered as a good approximation of the free energy, and used in the subsequent importance sampling step, as explained below. Note that a perfect convergence is not required, in the sense that the bias does not need to estimated very accurately, as it is removed in the final importance sampling step, as described in the next section. Moreover, note that the biasing potential A_t needs to be computed only up to an unimportant additive constant which does not play any role in the overall procedure.

Figure 2: The Markov chain Monte-Carlo adaptive biasing force algorithm.

Algorithm 1. *Consider a reaction coordinate ξ. Starting from some initial configuration x_0 and the biasing potential $A_0 = 0$, and for $t \geq 0$,*

(a) Propose a move from x_t to y_{t+1} according to the transition density $T(x_t, y_{t+1})$;

(b) Compute the acceptance rate

$$\alpha_t = \min\left(\frac{\pi_{A_t}(y_{t+1})\,T(y_{t+1}, x_t)}{\pi_{A_t}(x_t)\,T(x_t, y_{t+1})},\, 1\right),$$

where the biased probability density π_{A_t} is defined as

$$\pi_{A_t}(x) \propto \pi(x) \exp\left[A_t \circ \xi(x)\right];$$

(c) Draw a random variable U_t uniformly distributed in $[0,1]$ ($U_t \sim \mathcal{U}[0,1]$);

(i) if $U_t \leq \alpha_t$, accept the move and set $x_{t+1} = y_{t+1}$;

(ii) if $U_t > \alpha_t$, reject the move and set $x_{t+1} = x_t$.

(d) Following (11), update the biasing force, hence the biasing potential, to get A_{t+1}.

(e) Go to Step (a).

3.3 Reweighting free-energy biased simulations

Upon stabilization of the adaptive algorithm at iteration T, an estimate $\hat{A} = A_T$ of the biasing potential A is obtained, which is such that the biased posterior density

$$\begin{aligned} \widetilde{\pi}(x) &= \frac{1}{\widetilde{Z}}\,\pi(x) \exp\left\{\widehat{A} \circ \xi(x)\right\} \\ &\propto \pi(x) \exp\left\{\widehat{A} \circ \xi(x)\right\} \end{aligned} \tag{12}$$

generates an approximately uniform marginal distribution over $[z_{\min}, z_{\max}]$ for the random variable $\xi = \xi(x)$. To recover the true distribution π, we use the following simple strategy: we simulate a standard MCMC algorithm, e.g. a random walk Hastings-Metropolis algorithm, targeted at the biased posterior density $\widetilde{\pi}$, and perform an importance sampling step from $\widetilde{\pi}$ to π, based on the importance sampling weights:

$$w(x) = \exp\{-\widehat{A} \circ \xi(x)\}. \tag{13}$$

From the MCMC chain (x_t), $t = 1, \ldots, T$, the expectation of a function h can be estimated as

$$\frac{\displaystyle\sum_{t=1}^{T} w(x_t) h(x_t)}{\displaystyle\sum_{t=1}^{T} w(x_t)}. \tag{14}$$

In this second stage, there is no need to truncate the sampling space, upon extending the definition of the bias function as follows: $\hat{A}(z) = \hat{A}(z_{\max})$ for $z > z_{\max}$, $\hat{A}(z) = \hat{A}(z_{\min})$ for $z < z_{\min}$. In this way, $\widetilde{\pi}$ is defined over the whole parameter space Ω.

4 Bayesian inference from free energy biased dynamics

In this section, we explain how to perform Bayesian inference for the Gaussian mixture model described Section 2.1, that is, how to compute quantities such as posterior expectations and ratios of marginal likelihoods (equal to ratios of normalizing constants Z_K defined in (4)), using the free energy associated to a given reaction coordinate as an importance function. The Gaussian mixture model corresponds, in the notation of Section 3, to $x = \theta \in \Omega$ and $\pi(x) = p(\theta|y) \propto p(\theta)p(y|\theta)$, hence $V(\theta) = -\log\{p(\theta)p(y|\theta)\}$. The reaction coordinate is now a function $\xi = \xi(\theta)$. The free-energy biased probability distribution is $\widetilde{p}(\theta|y) \propto p(\theta|y)/w(\theta) \propto p(\theta)p(y|\theta)/w(\theta)$, where w is defined in (13).

We start by listing the criteria we use to assess the quality of the importance sampling procedure in Section 4.1. As mentioned in the introduction, the strategy to sample the posterior distribution (2) consists in three steps: choosing a reaction coordinate, computing (an approximation of) the free energy associated to this reaction coordinate, and using the free energy to build an importance sampling proposal distribution. The previous section was devoted to the second and third steps. We discuss the first step in Section 4.2 for the mixture model at hand. Section 4.3 presents an extension of the method to the computation of the ratio of normalizing constants associated to different values of the number of components K, in order to perform model choices between models corresponding to different number of components. Notice that we discuss in this section the theoretical efficiency of the whole approach. These discussions are supported by numerical experiments in Section 5.

4.1 Criteria for choosing the reaction coordinate

4.1.1 General criteria

We consider the following criteria for choosing ξ:

(a) Does the approximate free energy A_t converge rapidly to its equilibrium value A, using the ABF or ABP algorithm?

(b) How efficient is the importance sampling step from the biased density to the actual posterior density? Actually, this criterion is twofold:

  (b1) How efficient is the sampling of the biased density?

  (b2) How representative are the biased samples with respect to the target posterior density?

(c) An additional, more practical, criterion is: how difficult is it to determine, a priori, an interval $[z_{\min}, z_{\max}]$ for the reaction coordinate values, which ensures good performance with respect to Criteria (a) and (b)?

Criterion (b2) is discussed in the next section. Criteria (a) and (b1) can actually be shown to be equivalent; see Lelièvre et al. (2008). Roughly speaking, an adaptive algorithm yields quickly an estimate of the free energy, if and only if the free energy is indeed a good biasing potential, in the sense that the dynamics driven by the biased potential converges quickly to equilibrium. To obtain quick convergence, the basic requirement is that the multimodality of the target measure conditional on $\xi(\theta) = z$ (at least for some values of z, possibly far away from the posterior mass) is much less noticeable than the multimodality of the unrestricted target measure, and the modes are not strongly separated. Thus, ensuring a uniform exploration along ξ enables mode switchings when ξ reaches those regions where the modes are entangled. Two basic checks should therefore be performed: (i) that the output of the adaptive algorithm has explored the full range $[z_{\min}, z_{\max}]$; and (ii) using the criterion mentioned in the introduction, and specifically for mixture posterior densities, that the algorithm has visited the $K!$ symmetric replicates of any significant local mode.

We discuss the practical choice of ξ in the mixture context with respect to the above criteria in Section 4.2, and we illustrate numerically this idea in Section 5.1.2.

4.1.2 Efficiency of the importance sampling step

We detail here Criterion (b2). To evaluate the performance of the importance sampling step, we compute the following efficiency factor

$$\mathrm{EF} = \frac{\left(\sum_{t=1}^{T} w(\theta_t)\right)^2}{T\sum_{t=1}^{T} w(\theta_t)^2}$$

where $w(\theta)$ is defined in (13), and where $\{\theta_t\}_{t\geq 0}$ denotes a sample distributed according to the biased probability $\widetilde{p}$, typically obtained by a Hastings-Metropolis algorithm. The efficiency factor is the Effective Sample Size of Kong et al. (1994) divided by the number of sampled values. This quantity lies in $[0,1]$. It is close to one (resp. to zero) when the random variable $w(\theta)$ has a small (resp. a large) variance. Indeed, it is easy to check that

$$\mathrm{EF} = \left(\frac{\mathrm{Var}_T(w)}{(\mathrm{E}_T(w))^2} + 1\right)^{-1}, \qquad \mathrm{Var}_T(w) = \frac{1}{T}\sum_{t=1}^{T} w(\theta_t)^2 - \Big[\mathrm{E}_T(w)\Big]^2,$$

where the latter quantity is the empirical variance of the sample $\{w(\theta_t)\}_{1\leq t\leq T}$, and $\mathrm{E}_T(w) = \sum_{t=1}^{T} w(\theta_t)/T$ its empirical average.

We now propose an estimate of the efficiency factor in terms of the converged bias $\hat{A}$ only, which may therefore be computed *before* the MCMC algorithm is run, and the importance sampling step is performed. This estimate is based on the fact that, with respect to $\widetilde{p}$, the marginal distribution of ξ is approximately uniform over $[z_{\min}, z_{\max}]$. For $z_{\min}$ and $z_{\max}$ well chosen, $\xi(\theta_t)$ hardly visits values out of interval $[z_{\min}, z_{\max}]$ and thus

$$\frac{\mathrm{Var}_T(w)}{(\mathrm{E}_T(w)^2} \simeq \frac{\dfrac{1}{z_{\max}-z_{\min}}\displaystyle\int_{z_{\min}}^{z_{\max}} \left(\exp\{-\widehat{A}(z)\} - \frac{1}{z_{\max}-z_{\min}}\int_{z_{\min}}^{z_{\max}} \exp\{-\widehat{A}(z)\}\right)^2 dz}{\left(\dfrac{1}{z_{\max}-z_{\min}}\displaystyle\int_{z_{\min}}^{z_{\max}} \exp\{-\widehat{A}(z)\}\, dz\right)^2},$$

which provides a justification for the following "theoretical" efficiency factor:

$$\mathrm{EF}_{\mathrm{theoretical}} = \frac{\left(\displaystyle\int_{z_{\min}}^{z_{\max}} \exp\{-\hat{A}(z)\}\, dz\right)^2}{(z_{\max}-z_{\min})\displaystyle\int_{z_{\min}}^{z_{\max}} \exp\{-2\hat{A}(z)\}\, dz}. \tag{15}$$

The agreement between theoretical and numerically computed efficiency factors in our simulations is very good, see Tables 1, 2 and 4 in Section 5.1.3. Thus, the theoretical efficiency factor allows for a quick check that the subsequent importance sampling is reasonably efficient.

From the expression (15), it is seen that the efficiency factor is close to one when A is close to a constant. Thus, Criterion (b2) mentioned in the previous section is likely to be satisfied if the free energy associated to ξ has a small amplitude, i.e. $A = \max A - \min A$ is as small as possible.

4.2 Practical choice of the reaction coordinate

We now discuss the choice of the reaction coordinate in the mixture context, i.e. the choice of the scalar function $\xi : \theta \to \mathbb{R}$. We focus on the following four choices: $\xi(\theta) = q_1$, $\xi(\theta) = \mu_1$, $\xi(\theta) = \beta$, and the opposite of the (unnormalized) posterior log-density, $\xi(\theta) = -\log\{p(\theta)p(y|\theta)\}$.

The requirement that the multimodality of the target measure conditional on $\xi(\theta) = z$ is much less noticeable than the multimodality of the unrestricted target measure rules out the choice of

μ_1 as a good reaction coordinate since, conditionally on μ_1, the posterior density still has at least $(K-1)!$ modes, as the components 2 to K remain exchangeable. Numerical tests indeed support these considerations, see below.

A more natural reaction coordinate is the opposite of the posterior log-density $\xi(\theta) = -\log\{p(\theta)p(y|\theta)\}$, in the spirit of the original Wang-Landau algorithm (Wang and Landau, 2001a,b). Indeed, exploring regions of low posterior density should help to escape more easily from local modes. Unfortunately, determining a range $[z_{\min}, z_{\max}]$ of "likely values" (with respect to the posterior distribution) for such functions of θ is not straightforward; see Criterion (c) above. Moreover, since the posterior density is expected to be multimodal and difficult to explore, there seems to be little point in performing MCMC pilot runs in order to determine $[z_{\min}, z_{\max}]$. A conservative approach is to choose a very wide interval $[z_{\min}, z_{\max}]$, but this makes the subsequent importance sampling step quite inefficient. In our simulations, we report satisfactory results for this reaction coordinate, but with the caveat that our choice for $[z_{\min}, z_{\max}]$ was facilitated by our different simulation exercises, based on several reaction coordinates. Another practical difficulty we observed in one case is that the estimated bias was quite inaccurate in the immediate vicinity of the posterior mode, because the free energy tends to increase very sharply in this region, see Section 5.2 for more details.

The choice $\xi(\theta) = q_1$ is satisfactory with respect to Criterion (c): the range on which it can vary, namely $(0, 1)$, is clearly known. With respect to (a), this choice looks appealing as well, since forcing q_1 to get close to one should empty the $K-1$ other components, which then may swap more easily. Unfortunately, we observe in some of our experiments that the dynamics biased by the free-energy associated with this reaction coordinate is not very successful in terms of mode switchings, see Figure 6.

Finally, $\xi(\theta) = \beta$ is a good trade-off with respect to our criteria, at least in the examples we treat. Concerning the determination of the interval $[z_{\min}, z_{\max}]$ (criterion (c)), since β determines the order of magnitude of the component variances $\sigma_k^2 = \lambda_k^{-1}$, there should be high posterior probability that β is a small fraction of R^2, where R is the range of the data. For instance, we obtain satisfactory results in all our experiments with $[z_{\min}, z_{\max}] = [R^2/2000, R^2/20]$. Concerning Criterion (a), we observe that the choice $\xi = \beta$ performs well (see the numerical results below). We propose the following explanation. Since the λ_k's have a Gamma(α, β) prior, large values of β penalize large values for the component precisions λ_k, or equivalently penalize small values for the component standard deviations $\sigma_k = \lambda_k^{-1/2}$. If β is large enough, the Gaussian components are forced to cover the complete range of the data, and thus can switch easily. This interesting phenomenon is illustrated by Figure 8 and the discussion surrounding these plots.

4.3 Computing normalizing constants and model choice

In this section, we discuss an extension of the method to perform model choice between models with different numbers of components. The principle is to compute the normalizing constant Z_K of the posterior density for different values of K, see (4) for a definition of Z_K. More precisely, it is sufficient to evaluate Z_K/Z_{K-1} for a given range of K (see (Robert and Casella, 2004, Chap. 7) for a review on Bayesian model choice).

We propose the following strategy. The estimation of Z_K/Z_{K-1} can be performed by first estimating $Z_K/\widetilde{Z}$, then estimating $Z_{K-1}/\widetilde{Z}$, and finally dividing the two quantities. A simple estimator of $Z_K/\widetilde{Z}$ (where $\widetilde{Z}$ is the normalizing constant in (12)) is given by

$$\widehat{I}_K = \frac{1}{T}\sum_{t=1}^{T} w(\theta_t),$$

where $\{\theta_t\}_{t\geq 0}$ is a sample distributed according to the biased probability $\widetilde{p}$ (with K normal components). This formula is based on the fact that the expectation of $w(\theta) = \exp\{-\widehat{A} \circ \xi(\theta)\}$ with respect to $\widetilde{p}$ is $Z_K/\widetilde{Z}$.

Let θ_{-k} denote the parameter vector obtained by removing in θ the parameters attached to a given component k, and replacing the probabilities q_l (for $l \neq k$) by $\widetilde{q}_l = q_l/(1-q_k)$. Let $p(y|\theta_{-k})$

denote the likelihood of the model with $K-1$ components, and parameter θ_{-k}. Then the following quantity

$$\widehat{I}_{K-1} = \frac{1}{K}\sum_{k=1}^{K} \widehat{I}_{K-1,k}, \qquad \widehat{I}_{K-1,k} = \frac{1}{T}\sum_{t=1}^{T} w_{-k}(\theta_t),$$

where

$$w_{-k}(\theta) = \frac{p(y|\theta_{-k})}{p(y|\theta)} \exp\left\{-\widehat{A} \circ \xi(\theta)\right\}, \tag{16}$$

is an estimator of $Z_{K-1}/\widetilde{Z}$.

The estimators $\widehat{I}_K$ and $\widehat{I}_{K-1}$ are reminiscent of the birth and death moves of the reversible jump algorithm of Richardson and Green (1997), where a new model is proposed by adding or removing a component chosen at random. The difference is that the biased posterior $\widetilde{p}$ acts as an intermediate state between the posterior with K components, $p(\theta|y)$ and the posterior with $K-1$ components (or more precisely, the posterior with $K-1$ components times the prior of a K-th "non-acting" component, in order to match the dimensionality of both $p(\theta|y)$ and $\widetilde{p}(\theta)$).

In our numerical experiments, the estimator of Z_K/Z_{K-1} obtained from this strategy performs well, see Section 5 (in particular Table 3).

5 Numerical examples

In our experiments and as explained above, we use the following approach. First, we run an adaptive algorithm (ABF, except for $\xi(\theta) = V(\theta) = -\log\{p(\theta)p(y|\theta)\}$, in which case we use ABP), for a given choice of the reaction coordinate ξ, and a given interval $[z_{\min}, z_{\max}]$, until a converged bias $\hat{A}$ is obtained. Second, we run a MCMC algorithm, with $\widetilde{p}$ given by (12) as invariant density. Third, we perform an importance sampling step from $\widetilde{p}$ to p, the unbiased posterior density. See the introduction of Section 4 for the notation.

The quality of the biasing procedure is assessed using the criteria mentioned in Sections 4.1. This consists in: (i) checking that the output is symmetric with respect to labellings, and many switchings between the modes are observed; (ii) computing the efficiency factor (a good indicator being the estimator (15) defined in terms of $\hat{A}$).

In the first step of the method (approximation of the free energy using adaptive algorithm), we deliberately use the simplest exploration strategy, namely a Gaussian random walk Hastings-Metropolis update, with small scales (see below for the precise values). This is to illustrate that the ability of adaptive algorithms to approximate the free energy does not crucially depend on a finely tuned updating strategy.

In the second step, we run a Hastings-Metropolis algorithm targeted at the biased posterior, using Cauchy random walk proposals, and the following scales: $\tau_\mu = R/1000$, $\tau_v = 2/R^2$, $\tau_\beta = 2 \times 10^{-5}\alpha R^2$, where R is the range of the data, which leads to acceptance rates between 10% and 30% in all cases.

5.1 A first example : the Fishery data

We first consider the Fishery data of Titterington et al. (1986), see also Frühwirth-Schnatter (2006), which consist of the lengths of 256 snappers, and a Gaussian mixture model with $K = 3$ components; see Figure 1 for a histogram.

5.1.1 Convergence of the adaptive algorithms

In the adaptive algorithm, we use Gaussian random walk proposals with scales $\tau_q = 5 \times 10^{-4}$, $\tau_\mu = 0.025$, $\tau_v = 0.05$ and $\tau_\beta = 5 \times 10^{-3}$. These parameters were also used to produce the unbiased trajectory in Figure 1. We illustrate here the convergence process in the case $\xi(\theta) = \beta$, using the ABF algorithm described in Section 3.2, with $z_{\min} = 0.05$, $z_{\max} = 4.0$ and $\Delta z = 0.01$.

First, we plot on Figure 3 the trajectory of (μ_1, μ_2, μ_3) and β for $T = 10^8$ iterations. With the ABF algorithm, the values visited by β cover the whole interval $[z_{\min}, z_{\max}]$, and the applied bias enables a frequent switching of the modes (observed here on the parameters (μ_1, μ_2, μ_3)). The trajectories for (μ_1, μ_2, μ_3) should be compared with the ones given on Figure 1, where no adaptive biasing force is applied (notice that the x-axis scale is not the same on both plots).

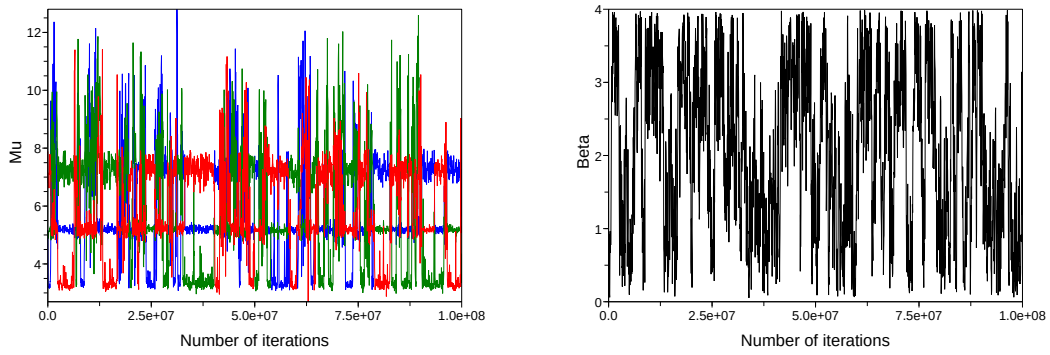

Figure 3: Trajectories over the 10^8 first iterations of the ABF algorithm for the choice $\xi(\theta) = \beta$, for (μ_1, μ_2, μ_3) (left) and for the β variable (right).

Second, we monitor the convergence of the biasing potential. To this end, we run a simulation for a total number of iterations $T = 10^9$, and store the biasing potential every $N_{\text{cvg}} = 10^6$ iterations. The distance between the current bias and the bias at iteration $t - N_{\text{cvg}}$ is measured by

$$\delta_t = \sqrt{\inf_{c \in \mathbb{R}} \sum_{i=0}^{N_z - 1} \left(A_{t,i} - A_{t - N_{\text{cvg}}, i} - c \right)^2}, \tag{17}$$

where $A_{t,i}$ denotes the value of the bias in bin i, i.e. $A_t(z) = A_{t,i}$ if $z \in (z_i, z_{i+1})$. Since the potential is defined only up to an additive constant, we consider the optimal shift constant c which minimizes the mean-squared distance between the two profiles. An elementary computation shows that this constant is equal to the difference between the averages of A_t and $A_{t - N_{\text{cvg}}}$. We finally renormalize this distance as

$$\varepsilon_t = \frac{\delta_t}{\sqrt{\sum_{i=0}^{N_z - 1} A_{t,i}^2}}.$$

The relative distance ε_t as a function of the iteration index t is plotted in Figure 4. Correct approximations of the bias are obtained after a few multiples of N_{cvg} iterations (the relative error being already lower than 0.1 at the first convergence check). We emphasize again that we did not optimize the proposal moves in order to reach the fastest convergence of the bias. It is very likely that better convergence results could be obtained by carefully tuning the parameters of the proposal function, or resorting to proposals of a different type.

On Figure 5, we plot the free energies associated to the four reaction coordinates mentioned above, as estimated by adaptive algorithms. For $\xi(\theta) = -\log\{p(\theta)p(y|\theta)\}$, we used an ABP algorithm, with $z_{\min} = 500$, $z_{\max} = 540$ and $\Delta z = 0.1$. For $\xi(\theta) = q_1$ and $\xi(\theta) = \mu_1$, we used ABF, with respectively $z_{\min} = 0$, $z_{\max} = 1$, $\Delta z = 0.005$ and $z_{\min} = \min y_i = 2.5$, $z_{\max} = \max y_i = 13$, $\Delta z = 0.05$. Recall that the so-obtained bias is the opposite of the log-posterior density of the marginal law. This is why the three important modes in the μ parameter can be read from the corresponding bias in Figure 5. Notice also that there is a lower bound on the admissible values of the opposite of the log-posterior density, hence the plateau value of the corresponding bias for low values of the reaction coordinate.

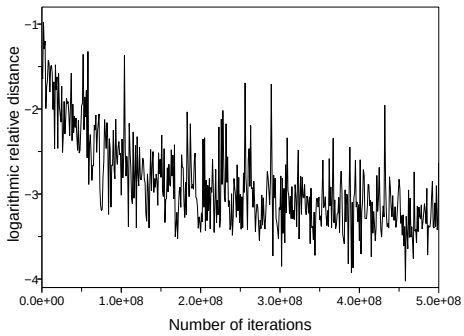

Figure 4: Convergence of the logarithmic relative distance $\log(\varepsilon_t)/\log(10)$ (see (17)), as a function of the number of iterations.

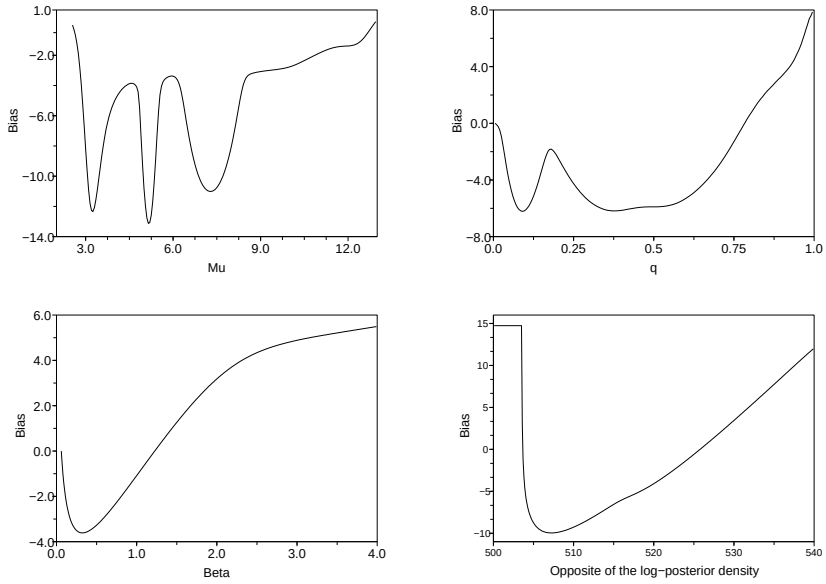

Figure 5: The fishery data. Free energies obtained for the reaction coordinates: $\xi(\theta) = \mu_1$ (top left), $\xi(\theta) = q_1$ (top right), $\xi(\theta) = \beta$ (bottom left) and $\xi(\theta) = -\log\{p(\theta)p(y|\theta)\}$ (bottom right).

Reaction coordinate	β	$-\log\{p(\theta)p(y\|\theta)\}$	q_1	μ_1
EF (numerical)	0.17	0.16	0.48	0.04
EF (theoretical)	0.179	0.178	0.454	0.079

Table 1: Efficiency factor for various choices of reaction coordinates, in the case $K = 3$.

5.1.2 Efficiency of the biasing procedure

We now discuss the results of the MCMC algorithm targeted at the biased posterior density, using the free energies computed above. In Figure 6, we observe that all the biased dynamics are much more successful in terms of mode switchings than the unbiased dynamics (see Figure 1). More precisely, the dynamics biased by the free energy associated with $\xi(\theta) = -\log\{p(\theta)p(y|\theta)\}$ is the most successful in terms of switchings, but the dynamics with $\xi(\theta) = \beta$ performs correctly as well. The dynamics with $\xi(\theta) = q_1$ seems to be less successful. In the case $\xi(\theta) = \mu_1$, one value of the parameters μ is forced to visit the whole range of values. The lowest mode (around $\mu = 3$) is not very well visited here.

The efficiency factors are presented in Table 1. They are rather large, which shows that the importance sampling procedure does not generate a degenerate sample. The choice $\xi(\theta) = q_1$ is the best one, but $\xi(\theta) = \beta$ and $\xi(\theta) = -\log\{p(\theta)p(y|\theta)\}$ give comparable and satisfactory results as well. The choice $\xi(\theta) = \mu_1$ on the other hand is a poor choice in this case.

In view of these results, it seems that $\xi(\theta) = -\log\{p(\theta)p(y|\theta)\}$ is the best choice, with the problem however that we needed to slightly modify the bias for the lowest values of the reaction coordinate because of the too sharp variations of the bias in this region. [1] On the other hand, the procedure is more automatic for $\xi(\theta) = q_1$ and β, the latter reaction coordinate being a much better choice when it comes to mode switchings.

We now focus on $\xi(\theta) = \beta$. As explained in the introduction, a good sampler should visit all the possible labellings of the parameter space. This implies in particular that the marginal posterior distributions of the simulated component parameters should be nearly identical. This is clearly the case here, see the scatter plots of the 1-in-10^4 sub-sample of the simulated pairs $(\mu_i, \log\lambda_i)$, $i = 1, \ldots, 3$ in Figure 7. The top left picture in Figure 7 also demonstrates that the biased dynamics indeed samples uniformly the values of β over the chosen interval $[z_{\min}, z_{\max}]$.

Finally, Figure 8 illustrates why the reaction coordinate $\xi(\theta) = \beta$ allows for escaping from local modes; see the discussion in Section 4.2. Each plot represents a sub-sample of the simulated pairs $(\mu_{k,t}, \log\lambda_{k,t})$ (where the subscript t denotes the iteration index while the subscript k labels the modes), restricted to β_t being in intervals, from left to right, $[0, 0.5]$, $[1.5, 2]$ and $[3.5, 4]$. Since the bias function depends only on β, these plots are rough approximations of the posterior distribution conditional on $\beta = 0.25$ (its posterior expectation), $\beta = 1.75$ and $\beta = 3$. In the leftmost plot of Figure 8, β is fixed to its posterior expectation and the three modes are well separated. As β is forced to take artificially large values (in the sense that the posterior probability density of such values is very small), the three modes get closer and eventually merge.

5.1.3 Larger values of K and model choice

We applied our approach to other values of K, namely $K = 4$ to 6, in the case $\xi(\theta) = \beta$. Table 2 reports the efficiency factor as a function of K. These factors remain quite satisfactory, which is related to the fact that the amplitude of the free energy (difference between the maximum and the minimum values) associated to this reaction coordinate is not too large over the chosen interval $[z_{\min}, z_{\max}]$, and does not change dramatically, see the profiles obtained for different values of K in Figure 9.

Figure 10 represents the marginal posterior distribution of $(\mu_1, \log\lambda_1)$, for $K = 4, 5, 6$. These plots were obtained by resampling 2000 points from the output of the MCMC targeting the biased

[1] We prefer not to spend too much time on these issues, but our numerical experience is that the bias obtained from the opposite of the log-posterior density is sometimes difficult to use in practice.

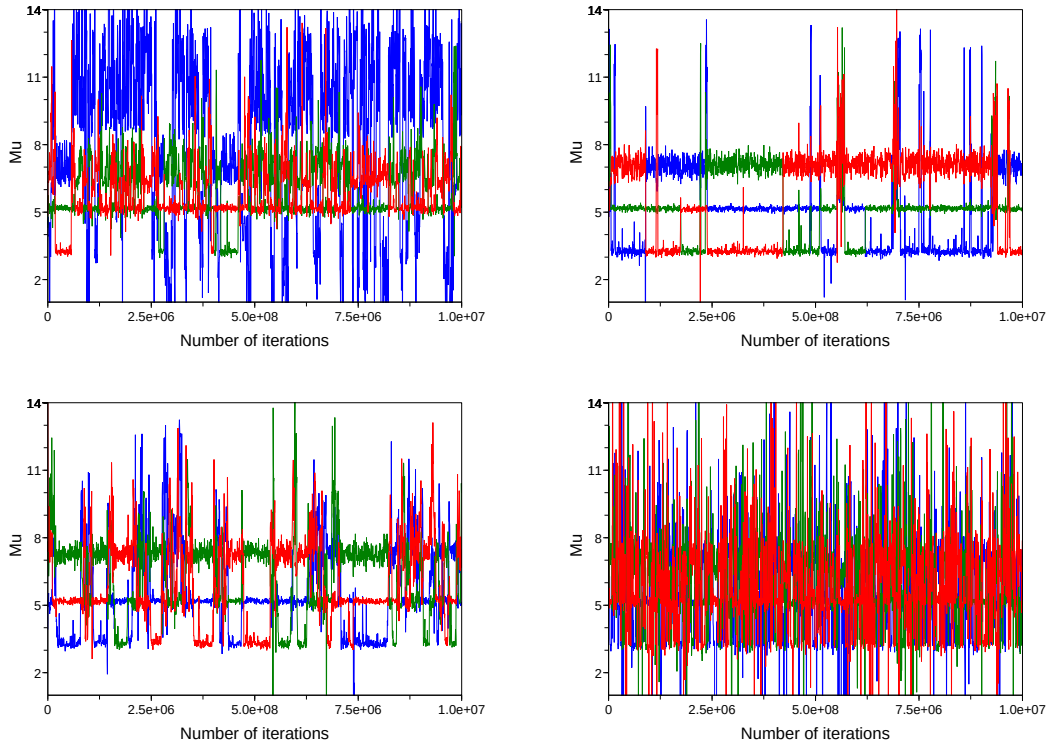

Figure 6: The fishery data. Trajectories of (μ_1, μ_2, μ_3) for the biased dynamics for different reaction coordinates. Top left: $\xi(\theta) = \mu_1$. Top right: $\xi(\theta) = q_1$. Bottom left: $\xi(\theta) = \beta$. Bottom right: $\xi(\theta) = -\log\{p(\theta)p(y|\theta)\}$.

K	3	4	5	6
EF (numerical)	0.17	0.18	0.17	0.16
EF (theoretical)	0.179	0.195	0.180	0.171

Table 2: Efficiency factors, for various values of the number of components considered in the mixture, with the choice $\xi(\theta) = \beta$.

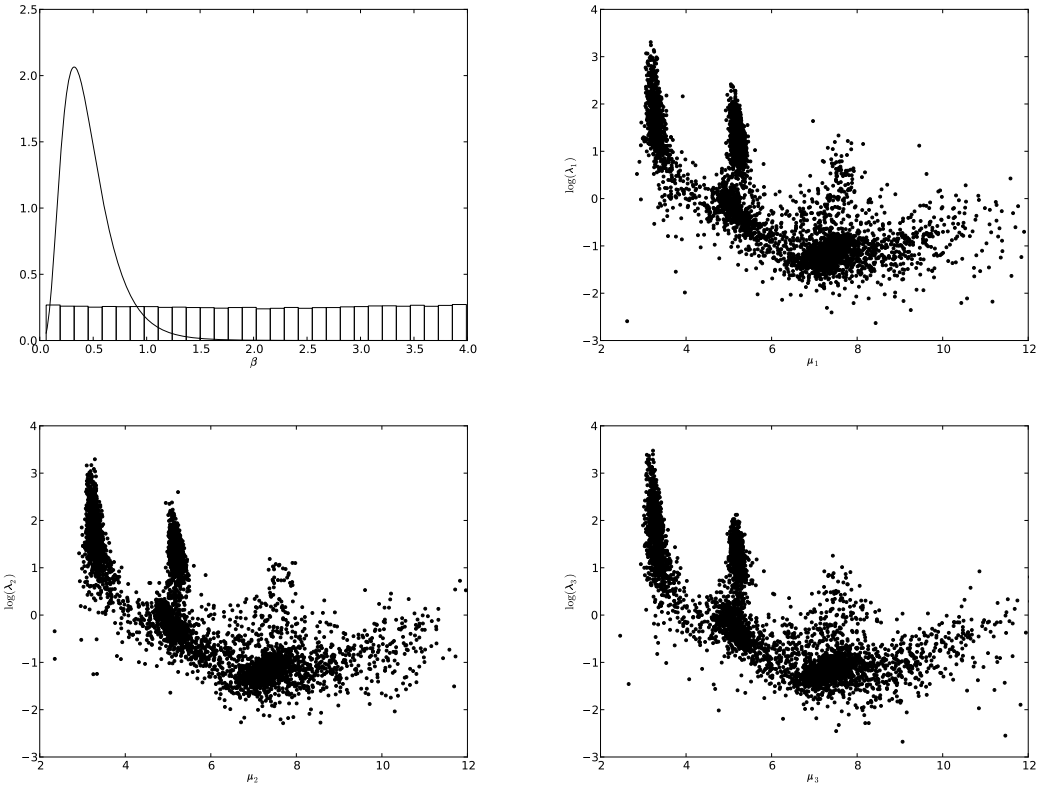

Figure 7: Top left: Histogram of simulated β's and estimated marginal posterior density of β. Remaining pictures: Scatter plots of simulated values for $(\mu_i, \log \lambda_i)$, for $i = 1, 2, 3$ when β is used as a reaction coordinate.

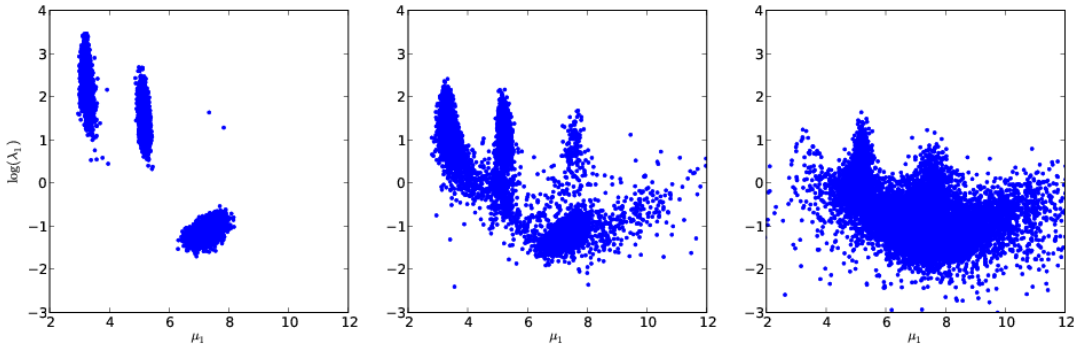

Figure 8: Simulated pairs $(\mu_1, \log \lambda_1)$ conditional on, from left to right, $\beta \in [0, 0.5]$, $\beta \in [1.5, 2]$ and $\beta \in [3.5, 4]$, see Section 5.1.3 for more details.

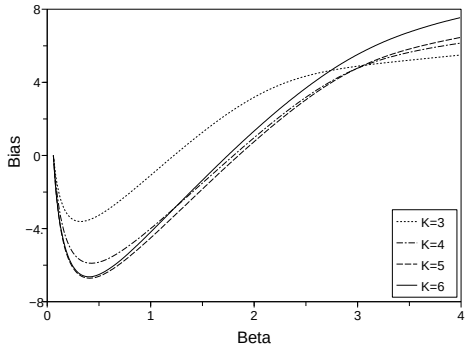

Figure 9: Estimated bias (free energy), for $K = 3, \ldots, 6$, and $\xi(\theta) = \beta$.

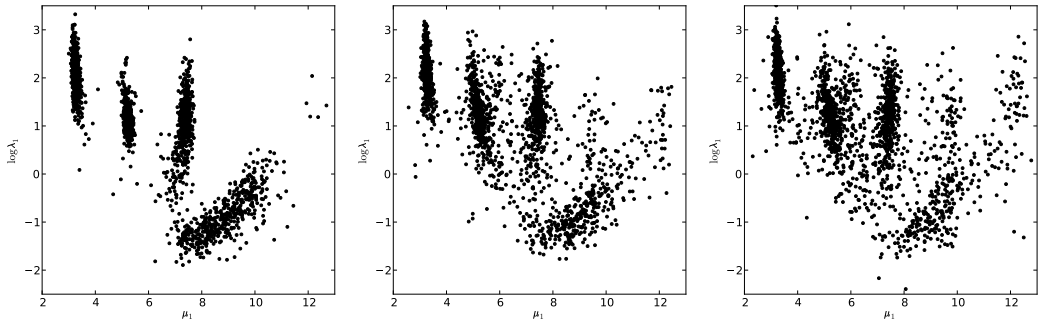

Figure 10: Marginal posterior distribution of $(\mu_1, \log \lambda_1)$, from left to right, for $K = 4$, 5 and 6, as represented by 2000 points resampled from the MCMC output.

posterior, with probability proportional to the importance sampling weight defined in (13). In each case, we checked that the MCMC output was symmetric with respect to label permutations.

Table 3 reports, for $K = 3, \ldots, 6$, the estimated log-Bayes factor for choosing a mixture model with K components against a mixture model with $K-1$ components, which equals $\log Z_K/Z_{K-1}$, assuming equal prior probability for different values of K. The reported error levels in Table 3 correspond to 90% confidence intervals, which were deduced from repeated runs of $T = 10^7$ iterations from the same MCMC algorithm targeting the biased posterior. The estimation error is quite small, despite being based on importance sampling steps in high dimensional spaces.

K	$\log Z_K/Z_{K-1}$	error
3	7.1	$\pm$ 0.2
4	4.2	$\pm$ 0.1
5	1.5	$\pm$ 0.1
6	0.9	$\pm$ 0.1

Table 3: Log Bayes factors for comparing model with K components against $K-1$ components, for $K = 3, ..., 6$; estimation error as evaluated from repeated MCMC runs.

5.2 A second example: the Hidalgo stamps data

Another well-known benchmark for mixtures is the Hidalgo stamps dataset, first studied by Izenman and Sommer (1988) (see also e.g. Basford et al. (1997)), which consists of the thickness (in mm) of $n = 485$ stamps from a given Mexican stamp issue; see Figure 1 for a histogram. (For convenience we multiplied the observations by 100.) We focus our presentation on the case $K = 3$, as this appears in our simulations to be the most challenging value compared to $K \geq 4$, in the sense that the sampling barriers in the posterior density are stronger in this case. For other values of K between 4 and 7 our approach performs better than for $K = 3$. For the sake of space the corresponding results are not reported.

This example is more challenging than the previous one, presumably because the number of observations is higher, which makes the likelihood more peaked. A clear sign of the increasing metastability is the increase in the free-energy barriers. For the reaction coordinates $\xi(\theta) = q_1$, $\xi(\theta) = \beta$, $\xi(\theta) = -\log\{p(\theta)p(y|\theta)\}$, we had to run the adaptive algorithm for $T = 10^9$ iterations in order to obtain a converged bias and recover a biased posterior sample which is symmetric by labelling. Again, a more elaborate proposal strategy for the Hastings-Metropolis step in the adaptive algorithm would be likely to stabilize the bias faster. As an illustration, Figure 11 presents the trajectories of (μ_1, μ_2, μ_3) sampled by the adaptive algorithm, with random walk scales: $\tau_q = 0.001$, $\tau_\mu = 0.05$, $\tau_v = 0.1$ and $\tau_\beta = 0.005$. The trajectories should be compared to the ones depicted in Figure 1, which were obtained with the same proposal, but without any biasing procedure.

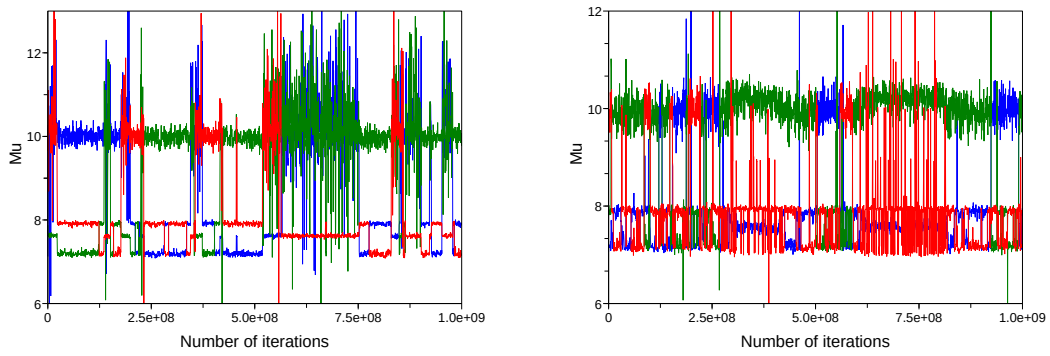

Figure 11: Sampled trajectories for (μ_1, μ_2, μ_3) during the adaptive biasing procedure. Left: ABF trajectory when the reaction coordinate is β. Right: ABP trajectory when the reaction coordinate is the opposite of the log-posterior density.

In Figure 12, we represent the biases obtained with various choices of the reaction coordinate. In the case $\xi(\theta) = \beta$, we set $z_{\min} = 0.005$, $z_{\max} = 2.5$ and $\Delta z = 0.005$. For $\xi(\theta) = q_1$, we consider $z_{\min} = 0$, $z_{\max} = 1$ and $\Delta z = 0.005$. Finally, for $\xi(\theta) = -\log\{p(\theta)p(y|\theta)\}$, we choose $z_{\min} = 720$, $z_{\max} = 780$ and $\Delta z = 0.1$.

Biased trajectories are presented in Figure 13 for $\xi(\theta) = q_1$, $\xi(\theta) = \beta$, and $\xi(\theta) = -\log\{p(\theta)p(y|\theta)\}$. Efficiency factors are reported in Table 4. The results show that, in terms of mode switching, $\xi(\theta) = -\log\{p(\theta)p(y|\theta)\}$ is the best choice. The choice $\xi(\theta) = q_1$ is a poor choice since very few switchings are observed; in particular, the mode starting around 7.5 does not change during the first 2.5×10^7 iterations. Such transitions are observed in the case $\xi(\theta) = \beta$.

6 Conclusion

We showed in this paper how to sample efficiently mixtures of univariate Gaussian distributions, using the free energy as a biasing potential (equivalently, using the marginal law as an importance sampling function).

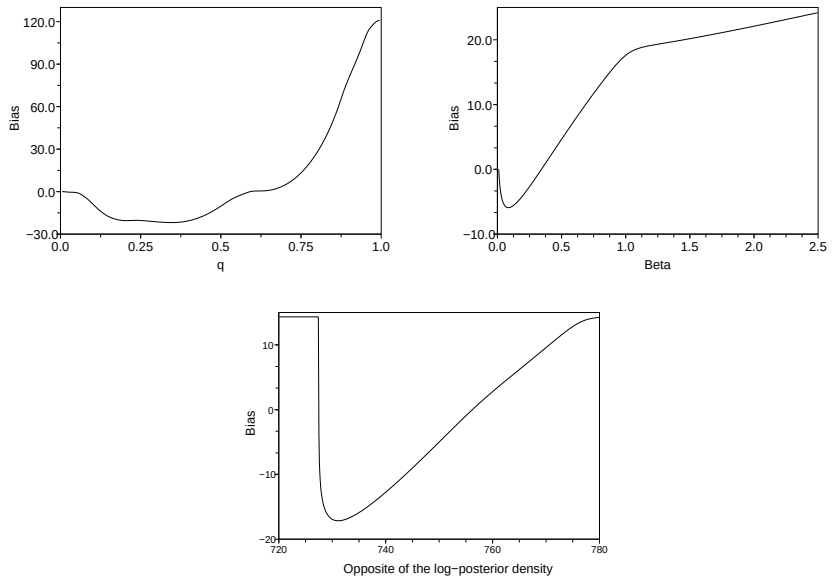

Figure 12: Hidalgo stamps problem. Free energies obtained for the reaction coordinates: $\xi(\theta) = q_1$ (top left), $\xi(\theta) = \beta$ (top right) and $\xi(\theta) = -\log\{p(\theta)p(y|\theta)\}$ (bottom).

Reaction coordinate	β	$-\log\{p(\theta)p(y\|\theta)\}$	q_1
EF (numerical)	0.02	0.24	0.23
EF (theoretical)	0.06	0.13	0.18

Table 4: Efficiency factor for various choices of reaction coordinates, in the case $K = 3$.

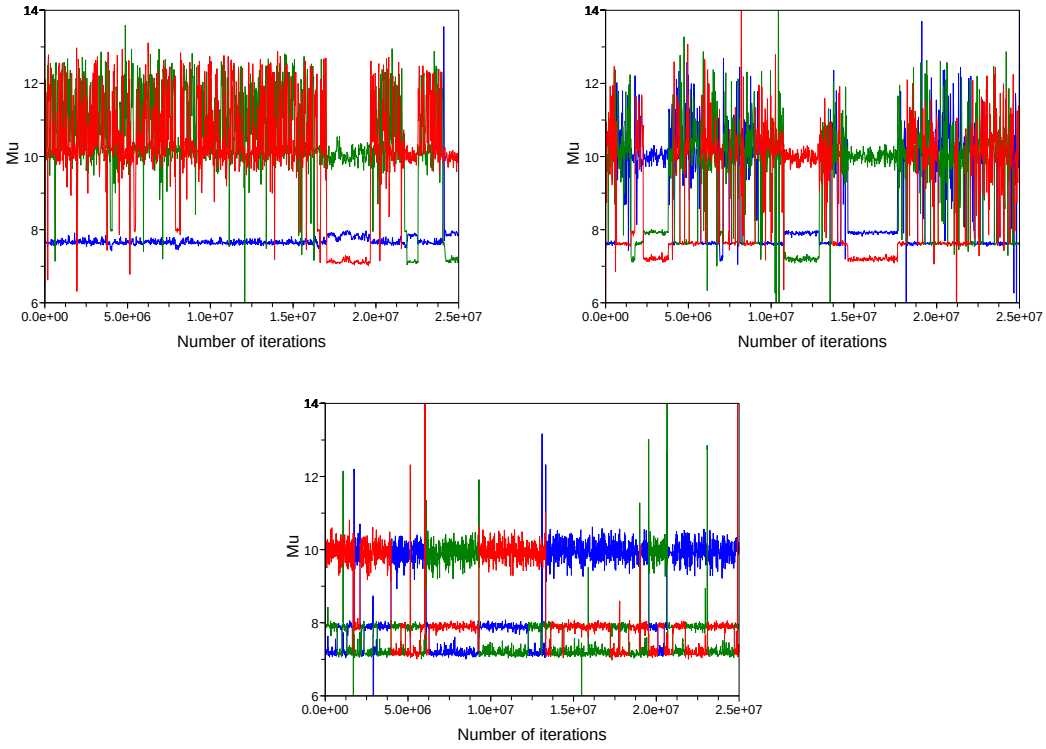

Figure 13: Hidalgo stamps problem. Trajectories of (μ_1, μ_2, μ_3) of the biased dynamics. Top left: reaction coordinate $\xi(\theta) = q_1$. Top right: $\xi(\theta) = \beta$. Bottom: $\xi(\theta) = -\log\{p(\theta)p(y|\theta)\}$.

To summarize our findings, a good approach for the Gaussian mixture model seems to be as follows: (i) choose β or $-\log\{p(\theta)p(y|\theta)\}$ as the reaction coordinate. When β is chosen, estimate the bias on the interval $[c_1R^2, c_2R^2]$ (with R the range of the data, and c_1, c_2 small constants, say $c_1 = 1/2000$ and $c_2 = 1/20$), while the determination of the interval for $-\log\{p(\theta)p(y|\theta)\}$ requires some preliminary tests; (ii) run an adaptive algorithm to obtain the corresponding free energy, (iii) run a MCMC algorithm targeting the biased posterior density, and perform importance sampling to remove the bias and recover the true posterior.

The same ideas may be applied to other mixture models. For instance, Figure 14 plots the posterior density of a two-component Poisson mixture model, conditional on different values for the hyper-parameters. Specifically, $p(y_i|\theta) = q_1\text{Poisson}(y_i;\lambda_1) + (1-q_1)\text{Poisson}(y_i;\lambda_2)$ for $i = 1,\ldots,n$, where $\text{Poisson}(\cdot;\lambda)$ denotes the probability density function of a Poisson distribution of parameter λ. Here, $\theta = (q_1, \lambda_1, \lambda_2)$. We use a $\text{Gamma}(\beta\bar{y}, \beta)$ prior for the λ_k's, and a uniform prior for q_1. The $n = 100$ observations are simulated from this model with parameter $\theta = (0.7, 3, 10)$. It can be seen again that biasing the posterior distribution towards larger values of β makes it possible to reduce the distance between the different modes. In the same spirit, we are currently working on applying our approach to multivariate Gaussian mixtures.

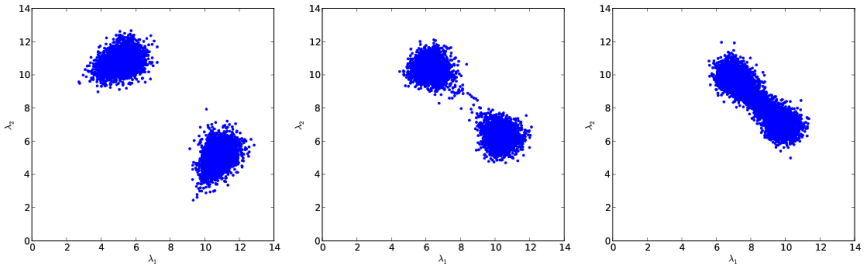

Figure 14: Scatter plots of 1000 simulated pairs (λ_1, λ_2) from the posterior distribution of a two-component Poisson mixture model, and $n = 100$ simulated data points, with a $\text{Gamma}(\alpha, \beta)$ prior for the λ_k, $\alpha = \beta\bar{y}$, and, from left to right, $\beta = 1$, 10, 20.

Finally, we would like to highlight some practical advantages of our approach. First, it requires little tuning: the only tuning parameters are the scales of the random walks in both algorithms (adaptive, and MCMC), and we obtained satisfactory results without trying to optimize these scales. Second, it is easy to check that the final results are correct: if the free energy has been well estimated, and the MCMC algorithm for the biased posterior has converged, then a nearly uniform marginal distribution for the reaction coordinate is observed, and the marginal distributions for the (μ_k, λ_k, q_k) are nearly identical, because of the symmetry of the true posterior, and the numerous mode switchings in the MCMC trajectories.

Acknowledgements

Part of this work was done while the two last authors were participating to the program "Computational Mathematics" at the Haussdorff Institute for Mathematics in Bonn, Germany. Support from the ANR grants ANR-008-BLAN-0218 and ANR-09-BLAN-0216 of the French Ministry of Research is acknowledged. The authors thank Julien Cornebise, Arnaud Doucet, Pierre Jacob, Christian P. Robert and Gareth Roberts for insightful remarks

References

C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4): 343–373, 2008.

Y. F. Atchadé and J. S. Liu. The Wang-Landau algorithm for Monte-Carlo computation in general state spaces. *Stat. Sinica*, 20(1):209–233, 2010.

R. Balian. *From Microphysics to Macrophysics. Methods and Applications of Statistical Physics*, volume I - II. Springer, 2007.

K.E. Basford, G.J. Mclachlan, and M.G. York. Modelling the distribution of stamp paper thickness via finite normal mixtures: The 1872 Hidalgo stamp issue of Mexico revisited. *J. Appl. Stat.*, 24(2):169–180, 1997.

G. Bussi, A. Laio, and M. Parrinello. Equilibrium free energies from nonequilibrium metadynamics. *Phys. Rev. Lett.*, 96(9):090601, 2006.

G. Celeux, M. Hurn, and C. P. Robert. Computational and inferential difficulties with mixture posterior distributions. *J. Am. Statist. Assoc.*, 95:957–970, 2000.

E. Darve and A. Pohorille. Calculating free energies using average force. *J. Chem. Phys.*, 115(20): 9169–9183, 2001.

J. Diebolt and C. P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *J. R. Statist. Soc. B*, 56:363–375, 1994.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer, 2006.

S. Frühwirth-Schnatter. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *J. Am. Statist. Assoc.*, 96(453):194–209, 2001.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

J. Hénin and C. Chipot. Overcoming free energy barriers using unconstrained molecular dynamics simulations. *J. Chem. Phys.*, 121(7):2904–2914, 2004.

Y. Iba. Extended ensemble Monte Carlo. *Int. J. Modern Phys. C*, 12(5):623–656, 2001.

A.J. Izenman and C.J. Sommer. Philatelic mixtures and multimodal densities. *J. Am. Stat. Assoc.*, 83(404):941–953, 1988.

A. Jasra, CC Holmes, and DA Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, pages 50–67, 2005.

B. Jourdain, T. Lelièvre, and R. Roux. Existence, uniqueness and convergence of a particle approximation for the adaptive biasing force process, 2009. in preparation.

A. Kong, J. S. Liu, and W. H. Wong. Sequential imputation and bayesian missing data problems. *J. Am. Statist. Assoc.*, 89:278–288, 1994.

T. Lelièvre and K. Minoukadeh. Long-time convergence of an adaptive biasing force method: the bi-channel case. *in preparation*, 2010.

T. Lelièvre, M. Rousset, and G. Stoltz. Computation of free energy profiles with parallel adaptive dynamics. *J. Chem. Phys*, 126:134111, 2007.

T. Lelièvre, M. Rousset, and G. Stoltz. Long-time convergence of an adaptive biasing force method. *Nonlinearity*, 21:1155–1181, 2008.

T. Lelièvre, M. Rousset, and G. Stoltz. *Free-energy computations: a mathematical perspective*. Imperial College Press, 2010.

F. Liang. A generalized Wang-Landau algorithm for Monte-Carlo computation. *J. Am. Stat. Assoc.*, 100(472):1311–1327, 2005.

J.M. Marin and C.P. Robert. *Bayesian core: a practical approach to computational Bayesian statistics*. Springer Verlag, 2007.

S. Marsili, A. Barducci, R. Chelli, P. Procacci, and V. Schettino. Self-healing Umbrella Sampling: A non-equilibrium approach for quantitative free energy calculations. *J. Phys. Chem. B*, 110 (29):14011–14013, 2006.

G.J. McLachlan and D. Peel. *Finite mixture models*. Wiley New York, 2000.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1091, 1953.

R. M. Neal. Annealed importance sampling. *Statist. Comput.*, 11:125–139, 2001.

P. Raiteri, A. Laio, F. L. Gervasio, C. Micheletti, and M. Parrinello. Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *J. Phys. Chem. B*, 110(8): 3533–3539, 2006.

S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *J. R. Statist. Soc. B*, 59(4):731–792, 1997.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods, 2nd ed.* Springer-Verlag, New York, 2004.

D. M. Titterington, A. F. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, 1986.

F. G. Wang and D. P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, 86(10):2050–2053, 2001a.

F.G. Wang and D.P. Landau. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Phys. Rev. E*, 64(5):056101, 2001b.