Supplementary materials for this article are available online.
Please click the JCGS link at http://pubs.amstat.org.

Divide and Conquer: A Mixture-Based Approach to Regional Adaptation for MCMC

Yan BAI, Radu V. CRAIU, and Antonio F. DI NARZO

The efficiency of Markov chain Monte Carlo (MCMC) algorithms can vary dramatically with the choice of simulation parameters. Adaptive MCMC (AMCMC) algorithms allow the automatic tuning of the parameters while the simulation is in progress. A multimodal target distribution may call for regional adaptation of Metropolis–Hastings samplers so that the proposal distribution varies across regions in the sample space. Establishing such a partition is not straightforward and, in many instances, the learning required for its specification takes place gradually, as the simulation proceeds. In the case in which the target distribution is approximated by a mixture of Gaussians, we propose an adaptation process for the partition. It involves fitting the mixture using the available samples via an online EM algorithm and, based on the current mixture parameters, constructing the *regional adaptive algorithm with online recursion* (RAPTOR). The method is compared with other regional AMCMC samplers and is tested on simulated as well as real data examples.

Relevant theoretical proofs, code and datasets are posted as an online supplement.

Key Words: Adaptive MCMC; Mixture model; Online EM.

1. INTRODUCTION

In recent years, the Markov chain Monte Carlo (MCMC) class of computational algorithms has been enriched with adaptive MCMC (AMCMC). Spurred by the seminal article of Haario, Saksman, and Tamminen (2001), an increasing body of literature has been devoted to the study of AMCMC due to its attractive potential for applications. For instance, while it has long been known that the choice of the proposal distribution's parameters in a Metropolis sampler is central to the performance of the algorithm, the fine tuning required is traditionally carried out by the programmer through a lengthy and sometimes frustrating trial and error process. The recent work of Haario, Saksman, and Tamminen (2001, 2005), Andrieu and Robert (2001), Andrieu and Moulines

Yan Bai is Graduate and Radu V. Craiu is Associate Professor (E-mail: *craiu@utstat.toronto.edu*), Department of Statistics, University of Toronto, Toronto, ON M5S 3G3, Canada. Antonio F. Di Narzo is Researcher, Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland.

Journal of Computational and Graphical Statistics, ***Accepted for publication****, Pages 1–17*
DOI: 10.1198/jcgs.2010.09035

(2006), Andrieu, Moulines, and Priouret (2005), and Roberts and Rosenthal (2007) has provided the theory needed to prove that it is possible to adapt the parameters of the proposal distribution automatically and "on the fly," that is, to allow the simulation process to self-adjust at each iteration n using all the samples available by that time. However, the practical implementation of AMCMC samplers requires careful consideration in those cases when the target distribution is multimodal (Craiu, Rosenthal, and Yang 2009; Giordani and Kohn 2010). For instance, a posterior distribution may be multimodal if the sampling distribution is represented as a mixture. The latter occurs when we assume that the population of interest is heterogeneous or when such a formulation is a convenient representation of a nonstandard density. It is well known that MCMC sampling from multimodal distributions can be extremely difficult as the chain can get trapped in one region of the sample space due to areas of low probability (bottlenecks) between the modes. Therefore, a large amount of effort has been devoted to designing efficient MCMC sampling methods for multimodal target distributions (Gelman and Rubin 1992; Geyer and Thompson 1994; Neal 1994; Richardson and Green 1997; Kou, Qing, and Wong 2006). One possible approach is to approximate the multimodal posterior distribution with a mixture of Gaussians as in the article by West (1993), who also showed that such an approximation may be useful for computation even if the target is skewed and not necessarily multimodal. AMCMC algorithms based on the same natural approach have been developed by Giordani and Kohn (2010), Andrieu and Thoms (2008), and Craiu, Rosenthal, and Yang (2009).

The mixture representation implies that the geography of the target distribution may differ across various regions of the sample space. One possible way to incorporate this information in the design of a Metropolis–Hastings sampler is to consider regional adaptation in which the proposal distribution varies across regions of the sample space. In the article by Craiu, Rosenthal, and Yang (2009) the regional adaptive (RAPT) sampler is designed assuming that one has reasonable knowledge about the regions where different sampling regimes are needed. These regions remain unchanged even though one may realize, as the simulation progresses, that partitions leading to a more efficient algorithm are possible. In addition, while one could use sophisticated methods to detect the modes of a multimodal distribution (see Neal 2001; Sminchisescu and Triggs 2001), it is not obvious how to use such techniques for defining the desired partition of the sample space.

Here we consider a different framework which allows the regions to evolve as the simulation proceeds. The regional adaptive random walk Metropolis (RWM) algorithm proposed here relies on the approximation of the target distribution π with a mixture of Gaussians. The partition of the sample space used for the *regional adaptive algorithm with online recursion* (RAPTOR) is defined based on the mixture parameters which, in turn, are updated using the simulated samples. An algorithm based on a somewhat similar idea has been developed independently by Andrieu and Thoms (2008). However, their algorithm 7 (henceforth denoted AT7) differs from RAPTOR in a few important aspects that are discussed in the next section. We emphasize that the approach proposed here is aimed at continuous distributions π as a direct extension of these ideas to discrete multimodal distributions is not straightforward.

In the following sections we present: the RAPTOR algorithm and its comparison with AT7 and RAPT (Section 2), the theoretical results on RAPTOR's ergodicity (Section 3), and the simulation studies and real data applications (Sections 4 and 5). The technical proofs, data, and programs used in the article are included in the Supplemental Materials accessible from the *JCGS* website.

2. REGIONAL ADAPTATION

2.1 RAPTOR: RAPT WITH ONLINE RECURSION

Regional adaptation may be needed for a number of statistical models, but in this article we assume that it is a consequence of the target's representation as a mixture. Intuitively, if π is well approximated using a mixture of Gaussians, then it is reasonable to surmise that each component of the mixture is a good proposal in a given region of the sample space $\mathcal{S} \subset \mathbb{R}^d$. To make concepts more precise, assume that π has K modes and consider its approximation by the mixture model

$$\tilde{q}_\eta(x) = \sum_{k=1}^{K} \beta^{(k)} N_d\big(x; \mu^{(k)}, \Sigma^{(k)}\big), \tag{2.1}$$

where $\beta^{(k)} > 0$ for all $1 \le k \le K$, $\sum_{k=1}^{K} \beta^{(k)} = 1$, $N_d(x; \mu, \Sigma)$ is the probability density of a d-variate Gaussian distribution with mean μ and covariance matrix Σ, and we set $\eta = \{(\beta^{(k)}, \mu^{(k)}, \Sigma^{(k)}) : 1 \le k \le K\}$. At each time n during the simulation process one has available n samples which are used to fit the mixture (2.1). In turn, the mixture parameters decide the form of the proposal distribution. In the following sections we describe in detail the online EM algorithm used for fitting the mixture, the construction of the partition underlying the regional sampler, and finally, the proposal distribution used for RAPTOR.

2.1.1 Recursive Adaptation

The analysis of mixture models has relied for a while now on the EM algorithm (Dempster, Laird, and Rubin 1977) as discussed by Titterington, Smith, and Makov (1985) and references therein. In the setting considered here, the simulation parameters need to be updated on the fly each time new draws from the target distribution are added to the Monte Carlo sample (as opposed to refitting using the entire sample). Our approach follows that of Andrieu and Moulines (2006) who have designed the method in conjunction with an adaptive independent Metropolis algorithm. A similar online EM for independent data was studied by Cappé and Moulines (2009). Suppose that at time $n-1$ the current parameter estimates are $\eta_{n-1} = \{\beta_{n-1}^{(k)}, \mu_{n-1}^{(k)}, \Sigma_{n-1}^{(k)}\}_{1 \le k \le K}$ and the available samples are $\{x_0, x_1, \ldots, x_{n-1}\}$. If we define the mixture indicator Z_i such that $Z_i = k$ if and only if x_i has been generated from the kth component of mixture (2.1), then let $\nu_i^{(k)} = P(Z_i = k | x_i, \eta_i)$. Given the additional data x_n we have

$$\nu_i^{(k)} = \frac{\beta_{i-1}^{(k)} N_d(x_i; \mu_{i-1}^{(k)}, \Sigma_{i-1}^{(k)})}{\sum_{k'} \beta_{i-1}^{(k')} N_d(x_i; \mu_{i-1}^{(k')}, \Sigma_{i-1}^{(k')})} \qquad \forall 1 \le i \le n, 1 \le k \le K. \tag{2.2}$$

Define

$$s_n^{(k)} = \frac{1}{n+1}\sum_{i=0}^{n} v_i^{(k)} = s_{n-1}^{(k)} + \frac{1}{n+1}\big(v_n^{(k)} - s_{n-1}^{(k)}\big) \tag{2.3}$$

and $\gamma_n^{(k)} = \frac{v_n^{(k)}}{(n+1)s_n^{(k)}}$, for all $1 \le k \le K$. Then the recursive estimator $\eta_n = \{(\beta^{(k)}, \mu_n^{(k)}, \Sigma_n^{(k)}) : k = 1, \ldots, K\}$ is, for all $1 \le k \le K$,

$$\begin{aligned}
\beta_n^{(k)} &= s_n^{(k)},\\
\mu_n^{(k)} &= \mu_{n-1}^{(k)} + \rho_n \gamma_n^{(k)}\big(x_n - \mu_{n-1}^{(k)}\big),\\
\Sigma_n^{(k)} &= \Sigma_{n-1}^{(k)} + \rho_n \gamma_n^{(k)}\big(\big(1-\gamma_n^{(k)}\big)\big(x_n - \mu_{n-1}^{(k)}\big)\big(x_n - \mu_{n-1}^{(k)}\big)^{\top} - \Sigma_{n-1}^{(k)}\big),
\end{aligned} \tag{2.4}$$

where ρ_n is a user-defined nonincreasing positive sequence such that $\sum_{n=1}^{\infty} \gamma_n^{(k)} \rho_n < \infty$ a.s. (this last condition is needed for the stability of the EM algorithm—see Section 3). In practice we use $\rho_n = n^{-1.1}$ which, coupled with $\gamma_n^{(k)} \in [0, 1]$, ensures the convergence of $\sum_{n=1}^{\infty} \gamma_n^{(k)} \rho_n$.

The RAPTOR implementation described in Section 2.1.4 also requires the sample covariance matrix of $\{x_1, \ldots, x_n\}$, $\Sigma_n^{\langle w \rangle}$, which can be computed recursively using

$$\Sigma_n^{\langle w \rangle} = \Sigma_{n-1}^{\langle w \rangle} + \frac{1}{n+1}\left(\left(1 - \frac{1}{n+1}\right)\big(x_n - \mu_{n-1}^{\langle w \rangle}\big)\big(x_n - \mu_{n-1}^{\langle w \rangle}\big)^{\top} - \Sigma_{n-1}^{\langle w \rangle}\right). \tag{2.5}$$

2.1.2 Definition of Regions

Using the mixture representation (2.1), we define the partition $\mathcal{S} = \bigcup_{k=1}^{K} \mathcal{S}^{(k)}$ so that, on each set $\mathcal{S}^{(k)}$, π is more similar to $N_d(x; \mu^{(k)}, \Sigma^{(k)})$ than to any other distribution entering (2.1). More precisely, if, for simplicity, we consider the case $K = 2$, then we define the partition $\mathcal{S} = \mathcal{S}^{(1)} \cup \mathcal{S}^{(2)}$ such that we maximize the sum of differences between Kullback–Leibler (KL) divergences

$$\begin{aligned}
&\Delta\mathrm{KL}\big(\mathcal{S}^{(1)}, \mathcal{S}^{(2)}\big)\\
&\quad := \mathrm{KL}\big(\pi, N_d\big(\cdot; \mu^{(2)}, \Sigma^{(2)}\big)\big|\mathcal{S}^{(1)}\big) - \mathrm{KL}\big(\pi, N_d\big(\cdot; \mu^{(1)}, \Sigma^{(1)}\big)\big|\mathcal{S}^{(1)}\big)\\
&\qquad + \mathrm{KL}\big(\pi, N_d\big(\cdot; \mu^{(1)}, \Sigma^{(1)}\big)\big|\mathcal{S}^{(2)}\big) - \mathrm{KL}\big(\pi, N_d\big(\cdot; \mu^{(2)}, \Sigma^{(2)}\big)\big|\mathcal{S}^{(2)}\big),
\end{aligned} \tag{2.6}$$

where $\mathrm{KL}(f, g|A) = \int_A \log(f(x)/g(x)) f(x)\,dx$. Note that we expect $\Delta\mathrm{KL} > 0$ since we want the KL divergence between $N_d(x; \mu^{(k)}, \Sigma^{(k)})$ and $\pi(x)$ to be the smallest on $\mathcal{S}^{(k)}$. With this aim we define, at time n, the region $\mathcal{S}_n^{(k)}$ as the set in which the kth component of the mixture density $\tilde{q}_{\eta_n}$ dominates the other ones, that is,

$$\mathcal{S}_n^{(k)} = \Big\{x : \arg\max_{k'} N_d\big(x; \mu_n^{(k')}, \Sigma_n^{(k')}\big) = k\Big\}. \tag{2.7}$$

In the following, for simplicity we assume that π is equal to (2.1) and we drop the index n. One can see that, for $K = 2$, $N_d(x; \mu^{(i)}, \Sigma^{(i)}) \le \pi(x) \le N_d(x; \mu^{(j)}, \Sigma^{(j)})$ on

$S^{(j)}$ when $i \neq j$. Therefore,

$$\mathrm{KL}\big(\pi, N_d\big(\cdot; \mu^{(1)}, \Sigma^{(1)}\big)\big|\mathcal{S}^{(1)}\big) \leq \mathrm{KL}\big(\pi, N_d\big(\cdot; \mu^{(2)}, \Sigma^{(2)}\big)\big|\mathcal{S}^{(1)}\big), \tag{2.8}$$

$$\mathrm{KL}\big(\pi, N_d\big(\cdot; \mu^{(2)}, \Sigma^{(2)}\big)\big|\mathcal{S}^{(2)}\big) \leq \mathrm{KL}\big(\pi, N_d\big(\cdot; \mu^{(1)}, \Sigma^{(1)}\big)\big|\mathcal{S}^{(2)}\big). \tag{2.9}$$

Now consider a different partition than the one given in (2.7), say $\mathcal{S} = \tilde{\mathcal{S}}^{(1)} \cup \tilde{\mathcal{S}}^{(2)}$. Assume that $\tilde{\mathcal{S}}^{(1)} = A \cup B$ with $A \subset \mathcal{S}^{(1)}$ and $B \subset \mathcal{S}^{(2)}$. Then

$$\begin{aligned}
\Delta\mathrm{KL}\big(\tilde{\mathcal{S}}^{(1)}, \tilde{\mathcal{S}}^{(2)}\big) &= \Delta\mathrm{KL}\big(\mathcal{S}^{(1)}, \mathcal{S}^{(2)}\big) \\
&\quad + 2\big[\mathrm{KL}\big(\pi, N_d\big(\cdot; \mu^{(1)}, \Sigma^{(1)}\big)|\mathcal{S}_1 \backslash A\big) \\
&\quad - \mathrm{KL}\big(\pi, N_d\big(\cdot; \mu^{(2)}, \Sigma^{(2)}\big)|\mathcal{S}_1 \backslash A\big)\big] \\
&\quad + 2\big[\mathrm{KL}\big(\pi, N_d\big(\cdot; \mu^{(2)}, \Sigma^{(2)}\big)|B\big) - \mathrm{KL}\big(\pi, N_d\big(\cdot; \mu^{(1)}, \Sigma^{(1)}\big)|B\big)\big] \\
&\leq \Delta\mathrm{KL}\big(\mathcal{S}^{(1)}, \mathcal{S}^{(2)}\big).
\end{aligned} \tag{2.10}$$

The last inequality is true because of (2.8)–(2.9). Note that (2.10) may hold even if π is not exactly equal to the mixture (2.1).

An alternative partition of the sample space can be produced based on regions

$$\mathcal{R}_n^{(k)} = \Big\{x : \arg\max_{k'} \beta_n^{(k')} N_d\big(x; \mu_n^{(k')}, \Sigma_n^{(k')}\big) = k\Big\}. \tag{2.11}$$

However, in our experience, estimates for $\beta_n^{(k)}$ remain volatile until the simulation has been run long enough and the correct proportions of samples from each modal region are obtained. Usually, such balance requires a longer sampling time and, before it is achieved, a partition like (2.11) can be very unstable. In addition, (2.8) and (2.9) may not hold if $\mathcal{S}^{(i)}$ is replaced with $\mathcal{R}^{(i)}$, $i = 1, 2$. For all these reasons, in this article we will work throughout with regions defined using (2.7) although the theoretical justification of the algorithm can be extended to the case when the regions are defined using (2.11). Figure 1 displays some actual shapes of the boundary between the two regions specified by (2.7) which show a good level of flexibility. However, in situations such as those shown in Figure 1(c) where

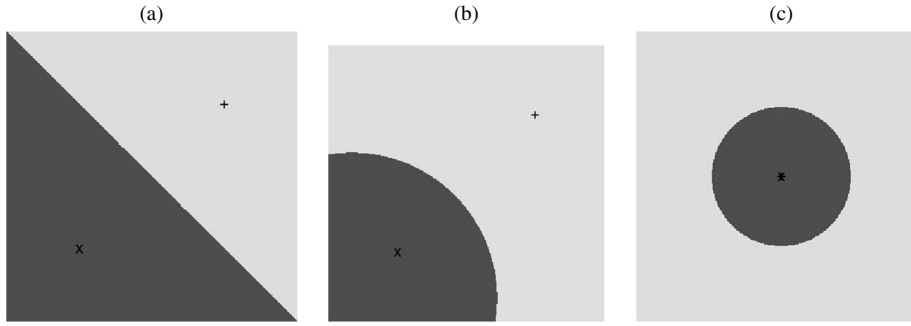

Figure 1. RAPTOR-defined regions for different relative values of the mixture components parameters. Region 1 in dark gray, region 2 in light gray. Mixture means are marked with a cross and a plus sign. (a) Different means, equal variances. (b) Different means, different variances. (c) Equal means, different variances.

the normal components have the same mean, RAPTOR performs with similar efficiency as the adaptive Metropolis algorithm of Haario, Saksman, and Tamminen (2001) (henceforth denoted AM).

2.1.3 Definition of the Proposal Distribution

The approximation (2.1) along with the online EM and the partition defined in (2.7) allow us to define the proposal distribution for RAPTOR. One may be tempted to consider, at time n, the proposal distribution

$$Q_n(x, dy) = \sum_{k=1}^{K} 1_{\mathcal{S}_n^{(k)}}(x) N_d\big(y; x, \Sigma_n^{(k)}\big)\, dy. \tag{2.12}$$

In other words, in each region $\mathcal{S}_n^{(k)}$ we would use the dominant component of the mixture as a proposal distribution. However, while such a proposal may have good local properties, it may not guarantee a good flow between different regions. A similar problem has been studied and validated theoretically in a nonadaptive setting by Guan and Krone (2007) who proved the benefit of adding long-range jumps to Metropolis–Hastings samplers for multimodal densities. Thus, at each time n during simulation, RAPTOR uses

$$Q_n(x, dy) = (1-\alpha) \sum_{k=1}^{K} 1_{\mathcal{S}_n^{(k)}}(x) N_d\big(y; x, s_d \tilde{\Sigma}_n^{(k)}\big)\, dy + \alpha N_d\big(y; x, s_d \tilde{\Sigma}_n^{\langle w \rangle}\big)\, dy, \tag{2.13}$$

where $s_d = 2.38^2/d$, a choice based on the optimality results obtained for the RWM by Roberts, Gelman, and Wilks (1997) and Roberts and Rosenthal (2001), $\tilde{\Sigma}_n^{(k)} = \Sigma_n^{(k)} + \epsilon \mathbf{I}_d$, $\tilde{\Sigma}_n^{\langle w \rangle} = \Sigma_n^{\langle w \rangle} + \epsilon \mathbf{I}_d$, $\epsilon > 0$ is a small constant, $\mathbf{I}_d$ is the $d \times d$ identity matrix, and $\alpha \in (0, 1)$ is a fixed weight which controls the flow between regions. Adding $\epsilon \mathbf{I}_d$ ensures that each covariance matrix is positive definite, an approach that follows that of Haario, Saksman, and Tamminen (2001). The theoretical derivations of Guan and Krone (2007) suggest using $\alpha = 0.3$ and this is the value used throughout the current article. We emphasize that the transition kernel (2.13) depends on the mixture parameters η_n in two ways: via the regions defined in (2.7) and, more directly, via the covariance matrices $\Sigma_n^{(k)}$ and $\Sigma_n^{\langle w \rangle}$.

2.1.4 Implementation of RAPTOR

The adaption parameter for RAPTOR is

$$\gamma = \big(\mu^{(1)}, \ldots, \mu^{(K)}, \tilde{\Sigma}^{(1)}, \ldots, \tilde{\Sigma}^{(K)}, \tilde{\Sigma}^{\langle w \rangle}\big) \in \mathcal{Y},$$

where $\mathcal{Y} = \underbrace{\mathbb{R}^d \times \cdots \times \mathbb{R}^d}_{K+1} \times \underbrace{\mathbb{R}^{d \times d} \times \cdots \times \mathbb{R}^{d \times d}}_{K+1}$ is the adaption parameter space.

In practice, γ is assigned a starting value which is kept constant during an *initialization period* after which the adaptation defined by RAPTOR is implemented. Suppose that after the initialization period the chain is in X_0. Subsequently, at every $n \geq 0$, we construct the partition (2.7) based on the value, at time n, of the adaption parameter

$$\Gamma_n = \big\{\mu_n^{(1)}, \ldots, \mu_n^{(K)}, \tilde{\Sigma}_n^{(1)}, \ldots, \tilde{\Sigma}_n^{(K)}, \tilde{\Sigma}_n^{\langle w \rangle}\big\}$$

and sample the proposal $Y_{n+1} \sim Q_n(X_n; dY)$. If $q_{\Gamma_n}(X_n, Y)$ denotes the density of $Q_n(X_n; dY)$, then the sample at time $n+1$ is

$$X_{n+1} = \begin{cases} Y_{n+1}, & \text{with probability } R_{\Gamma_n}(X_n, Y_{n+1}) \\ X_n, & \text{with probability } 1 - R_{\Gamma_n}(X_n, Y_{n+1}), \end{cases}$$

where $R_\gamma(x, y) = \min\{1, \frac{\pi(y)q_\gamma(y,x)}{\pi(x)q_\gamma(x,y)}\}$ and Γ_n is obtained from (2.2)–(2.5).

2.2 Other Regional AMCMC Algorithms

In the case in which one uses throughout the simulation a fixed partition of the space $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$, the mixed RAPT algorithm (Craiu, Rosenthal, and Yang 2009) for a random walk Metropolis (RWM) sampler uses at iteration n the following mixture as a proposal distribution:

$$\begin{aligned} Q_n(x, dy) = (1-\beta) \sum_{i=1}^{2} 1_{\mathcal{S}_i}(x) \big[& \lambda_1^{(i)}(n) N_d\big(y; x, s_d \Sigma_n^{(1)} + s_d \epsilon \mathbf{I}_d\big)\, dy \\ & + \big(1 - \lambda_1^{(i)}(n)\big) N_d\big(y; x, s_d \Sigma_n^{(2)} + s_d \epsilon \mathbf{I}_d\big)\, dy \big] \\ & + \beta N_d\big(y; x, s_d \Sigma_n^{\langle w \rangle} + s_d \epsilon \mathbf{I}_d\big)\, dy, \end{aligned} \tag{2.14}$$

where $\Sigma_n^{(i)}$ is the sample covariance matrix computed from those samples in $\mathcal{S}_i$ and $\Sigma_n^{\langle w \rangle}$ is the sample covariance matrix estimated from all n samples in $\mathcal{S}$. The mixing parameters $\lambda_1^{(i)}$, $i = 1, 2$, are adapted using the jump distances measured for each proposal, that is,

$$\lambda_j^{(i)}(n) = \begin{cases} \frac{d_j^{(i)}(n)}{\sum_{h=1}^{2} d_h^{(i)}(n)}, & \text{if } \sum_{h=1}^{2} d_h^{(i)}(n) > 0 \\ 1/2, & \text{otherwise,} \end{cases}$$

where $d_j^{(i)}(n)$ is the average square jump distance up to iteration n computed every time the accepted proposal was generated from jth regional proposal *and* the current state of the chain was in $\mathcal{S}_i$. The parameter $\beta \in (0, 1)$ is constant throughout the simulation, and is usually smaller than 0.5. For further details on RAPT we refer to the article of Craiu, Rosenthal, and Yang (2009). An important difference from RAPTOR is that the boundary between regions is fixed, essentially ignoring the information from the Monte Carlo sample. This imposes the use, in each region, of proposals from each component entering (2.14). Therefore, one has to accept a possible loss of computational efficiency each time a proposal is generated from the component of (2.14) which is not tuned for the region the chain is currently visiting. By allowing the regions to evolve, RAPTOR alleviates this problem, as in region $\mathcal{S}^{(k)}$ it considers proposals generated only from the dominant and global components of the mixture.

Algorithm AT7, developed independently by Andrieu and Thoms (2008), is also based on approximating π with (2.1). However, in their approach, instead of using a partition of the space, at time $n+1$ the proposal is generated from the mixture component most favored by the current state of the chain, x_n. More precisely, the generation involves sampling first the mixture indicator variable $z_n \sim p(Z_n | x_n, \eta)$ and then the proposal is generated from $Y \sim N_d(\cdot; x_n, \lambda^{(z_n)} \Sigma^{(z_n)})$, where $\lambda^{(z_n)}$ is a scaling factor that is also adapted based on the

available samples (for details, see Andrieu and Thoms 2008). AT7 does not have a global component in the proposal distribution so that the flow between regions may be slower than for RAPTOR. This is particularly relevant in those situations when the number of components K is misspecified so that the local approximation is good only in a subset of the sample space. In Section 4, we perform a simulation in which K is misspecified and AT7 and RAPTOR are compared.

3. THEORETICAL RESULTS

The proof of ergodicity for RAPTOR relies on the sufficient conditions established by Roberts and Rosenthal (2007) to ensure the ergodicity of general adaptive MCMC. Formally, an adaptive MCMC algorithm for the target distribution π is defined using the adaptation parameter $\gamma \in \mathcal{Y}$, that is, the transition kernel's parameter vector which is allowed to evolve as the simulation proceeds. It is assumed that for each $\gamma \in \mathcal{Y}$ the time-homogeneous Markov chain kernel P_γ has stationary distribution π (Meyn and Tweedie 1993). The adaptive paradigm assumes that given an initial point $X_0 := x_0 \in \mathcal{S} \subset \mathbb{R}^d$ and a kernel P_{Γ_0} with $\Gamma_0 := \gamma_0 \in \mathcal{Y}$, at each iteration $n+1$, X_{n+1} is generated from $P_{\Gamma_n}(X_n, \cdot)$, where Γ_{n+1} is a function of $X_0, \ldots, X_{n+1}$ and $\Gamma_0, \ldots, \Gamma_n$. We say that the adaptive MCMC process $\{X_n : n \geq 0\}$ is *ergodic* if $\lim_{n\to\infty} \|\mathcal{L}(X_n) - \pi\|_{\mathrm{TV}}$, where for any measure μ, $\|\mu\|_{\mathrm{TV}} = \sup_A |\mu(A)|$.

Roberts and Rosenthal (2007) gave two conditions, *Containment* and *Diminishing Adaptation*, which together imply ergodicity of adaptive MCMC. By definition, *Containment* holds if for all $\epsilon > 0$, the sequence $\{M_\epsilon(X_n, \Gamma_n) : n \geq 0\}$ is bounded in probability conditional on any $X_0 = x_0$ and $\Gamma_0 = \gamma_0$, that is, for all $\delta > 0$, there is $N \in \mathbb{N}$ such that $\mathbf{P}(M_\epsilon(X_n, \Gamma_n) \leq N | X_0 = x_0, \Gamma_0 = \gamma_0) \geq 1 - \delta$ for all $n \in \mathbb{N}$, where $M_\epsilon(x, \gamma) = \inf_n\{n \geq 1 : \|P^n_\gamma(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \leq \epsilon\}$ is the "ϵ-convergence" function.

The *Diminishing Adaptation* condition is defined as $\lim_{n\to\infty} D_n = 0$ in probability, where

$$D_n = \sup_{x\in\mathcal{S}} \left\| P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot) \right\|_{\mathrm{TV}} \tag{3.1}$$

represents the change between the transition kernels used at iterations n and $n+1$.

We prove the ergodicity of RAPTOR assuming the following compactness condition:

(A1) There is a compact subset $\mathcal{S} \subset \mathbb{R}^d$ such that the target density π is continuous on $\mathcal{S}$, positive on the interior of $\mathcal{S}$, and zero outside of $\mathcal{S}$.

It should be noted that compact sets can be arbitrarily large so (A1) does not impose a significant limitation in practice. The stability of the online EM depends on the properties of the sequences $\{\gamma_n^{(k)}\}_{n\geq 0}$ and $\{\rho_n\}_{n\geq 0}$. We prove the stability of the online EM using the following conditions:

(A2) The sequence $\{\rho_j : j \geq 1\}$ is positive and nonincreasing.

(A3) For all $k = 1, \ldots, K$,

$$\mathbf{P}\left(\limsup_{i\to\infty} \sup_{l\geq i} \sum_{j=i}^{l} \rho_j \gamma_j^{(k)} = 0 \right) = 1.$$

It is worth emphasizing that (2.3) implies $\gamma_j^{(k)} \in [0, 1]$ for all $1 \le j \le n$ and $1 \le k \le K$. In practice we have freedom over the choice of the sequence $\{\rho_j\}_{j\ge1}$ so we set $\rho_j = j^{-1.1}$. This choice guarantees that (A2) and (A3) are simultaneously satisfied in all simulations and real data applications considered in the next two sections.

Theorem 1. *With the above notations, the following hold*:

(a) *Assuming* (A1) *and* (A2), *the RAPTOR algorithm is ergodic to* π.

(b) *Assuming* (A2) *and* (A3), *the adaptive parameter* $\{\Gamma_n\}_{n\ge0}$ *converges in probability.*

The proof of the theorem is in the Appendix which is accessible from the *JCGS* website.

4. SIMULATIONS

In this section, we illustrate the performance of the RAPTOR algorithm using Gaussian mixtures under different scenarios which were designed to cover a range of possibilities. We compare different regional random walk Metropolis algorithms by means of the mean squared error (MSE) of the sample mean estimator of the first coordinate of the chain, evaluated by independently replicating each simulation 1000 times. Similar results hold for the other components but are not reported here due to space limitations. Relevant code is available as Supplemental Material.

We consider the target distribution equal to the d-variate Gaussian mixture restricted to a large compact set

$$\pi(x; m, s) \propto 1_{\mathbf{C}_d}(x)[0.5N_d(x; -m \times \mathbf{1}, \mathbf{I}_d) + 0.5N_d(x; m \times \mathbf{1}, s \times \mathbf{I}_d)], \tag{4.1}$$

where $\mathbf{C}_d = [-10^{10}, 10^{10}]^d$, $m \in \mathbf{R}$, $s \in [1, \infty)$.

For increasing values of m, the target distribution presents two modes which are more and more separated, while s is the ratio between the variances of the two mixture components. We will consider the cases where the state space dimension $d \in \{2, 5\}$. For all the algorithms used we set the following common starting estimates:

$$\begin{aligned} &\hat{\mu}_0^{(1)} = (-2, 0, \ldots, 0)^T, \qquad \hat{\mu}_0^{(2)} = (2, 0, \ldots, 0)^T, \\ &\hat{\beta}_0^{(k)} = 0.5, \qquad \hat{\Sigma}_0^{(1)} = 0.1 \times \mathbf{I}_d, \qquad \hat{\Sigma}_0^{(2)} = 0.1 \times s \times \mathbf{I}_d, \end{aligned} \tag{4.2}$$

and we set $\hat{\Sigma}_0^{\langle w\rangle} = 50 \times \mathbf{I}_2$ when the state space dimension is equal to 2, and $\hat{\Sigma}_0^{w} = 10 \times \mathbf{I}_5$ when it is equal to 5. These values are pessimistic, as one would normally obtain more reasonable mixture estimates using a preliminary run of the simulation. We compare the five following different regional random walk Metropolis algorithms:

- RAPTOR, with starting mixture estimates set to (4.2),
- a Regional Random Walk Metropolis (RRWM) which uses the same proposal as RAPTOR with mixture estimates *fixed to their true values* (the boundary used is defined in (2.7)),

- AT7 with starting mixture estimates set to (4.2) and proposal variance scaling factors fixed to $\epsilon_d = 2.38^2/d$,
- RAPT, with boundary $x_1 = 0$, and starting local proposal covariance matrices set to (4.2). The boundary set-up gives an approximate separation of the two modes using a $d-1$ hyperplane and is reasonable when $d = 2$ but it is less so when $d = 5$.

We run each chain from a fixed, central starting value, for a total of 1000 iterations, using the first 100 as burn-in. For each dimensionality of the state space, we considered the scenarios given by the following ten combinations of parameter values:

$$(d, m, s) \in \big\{(2,1,1), (5,0.5,1), (2,1,4), (5,0.5,4), (2,0,1), \\ (5,0,1), (2,0,4), (5,0,4), (2,2,1), (5,1,1)\big\}. \tag{4.3}$$

These scenarios cover all the boundary shapes depicted in Figure 1 and span a reasonably wide range of separation between the modes from $m = 0$ (no separation) to $m = 2$. For each simulation and for each algorithm we report the mean squared error (MSE) of the sample mean estimator of the first coordinate. These are computed based on independent replications of the simulation, knowing that the true mean of the first component is 0. All the results are reported in Table 1.

The algorithm RRWM performs best in the experimented scenarios, thus supporting the choice of partitioning the state space. When $d = 2$ all algorithms seem to perform similarly, but for $d = 5$ RAPTOR's performance comes closest to that of RRWM. Comparing RAPTOR with RAPT for $d = 5$, we notice that the strategy of adapting the boundary leads to efficiency gains. In the scenarios shown here RAPTOR does better than AT7 when the latter has been implemented with fixed variance scaling. The variable scaling is designed to preserve a certain acceptance rate but in a mixture setting one is unsure of which acceptance rate is optimal. The values used in the simulations lead to similar acceptance rates for RAPTOR and AT7 (around 40% when $d = 2$ and around 32% when $d = 5$).

Table 1. Gaussian mixture target distribution: mean squared error $\times 1000$ of the sample mean estimator of the first coordinate, for different values of the distance between modes, different ratios between target mixture variances, in state spaces with two or five dimensions.

(m, s)	RAPTOR	RRWM	AT7	RAPT
$d = 2$				
$(1, 1)$	21	21	25	22
$(1, 4)$	43	39	43	46
$(0, 1)$	10	8	12	11
$(0, 4)$	25	20	27	28
$(2, 1)$	170	170	247	136
$d = 5$				
$(0.5, 1)$	30	22	33	41
$(0.5, 4)$	72	62	90	108
$(0, 1)$	23	18	26	29
$(0, 4)$	51	48	66	62
$(1, 1)$	126	72	156	142

Table 2. Comparison of RAPTOR and AT7 when K is misspecified. Reported is the mean squared error of the sample mean of the first coordinate.

Model	RAPTOR-2	AT7-2	RAPTOR-3	AT7-3
π_1	0.167	0.240	0.157	0.221
π_2	0.033	0.068	0.045	0.092

We have also studied a number of scenarios in which the number of mixture components is misspecified as this is not uncommon in practice. Due to space restrictions we report on two representative cases:

$$\pi_1(x) \propto 1_{\mathbf{C}_5}(x)[0.4N_5(x; -3 \times \mathbf{1}, \mathbf{I}_5) + 0.2N_d(x; 0, \mathbf{I}_5) + 0.4N_d(x; 3 \times \mathbf{1}, \mathbf{I}_5)],$$

$$\pi_2(x) \propto 1_{\mathbf{C}_5}(x)[0.4N_5(x; -1.5 \times \mathbf{1}, \mathbf{I}_5) + 0.6N(x; \mathbf{1}, \mathbf{I}_5)]$$

for which we run RAPTOR and AT7 using $K = 2$ and $K = 3$. In Table 2 we report the MSE for the sample mean of first coordinate which was computed from 1000 replicates. Note that RAPTOR-k and AT7-k refer to runs of RAPTOR and AT7 algorithms when we assume the $K = k$ specification. The starting points for the two algorithms are included in the Supplemental Material available online. We suspect that the global component in RAPTOR provides an edge over AT7. However, this could be easily fixed by adding a global component in AT7 and only extensive use of the two algorithms can lead to more definitive comparisons.

5. REAL DATA EXAMPLES

5.1 GENETIC INSTABILITY OF ESOPHAGEAL CANCERS

We analyzed the "Loss of Heterozygosity" (LOH) dataset from the Seattle Barrett's Esophagus research project (Barrett et al. 1996). Cancer cells suffer a number of genetic changes during disease progression, one of which is *loss of heterozygosity* (LOH). Chromosome regions with high rates of LOH are hypothesized to contain genes which regulate cell behavior and may be of interest in cancer studies. For additional details we refer to the report by Warnes (2001). The dataset contains 40 frequencies of the event of interest (LOH) and their associated sample sizes. We model the frequencies using the mixture suggested by Desai (2000):

$$X_i \sim \eta\,\text{Binomial}(N_i, \pi_1) + (1 - \eta)\,\text{Beta-Binomial}(N_i, \pi_2, \gamma), \tag{5.1}$$

with priors $\eta \sim \text{Unif}[0, 1], \pi_1 \sim \text{Unif}[0, 1], \pi_2 \sim \text{Unif}[0, 1], \gamma \sim \text{Unif}[-30, 30]$, where η is the probability of a location being a member of the binomial group, π_1 is the probability of LOH in the binomial group, π_2 is the probability of LOH in the beta-binomial group, and γ controls the variability of the beta-binomial group. The parameterization adopted for the Beta-Binomial distribution is such that γ's range is the real line and we have used the logistic transformation on the parameters η, π_1, π_2 so that the range for each component

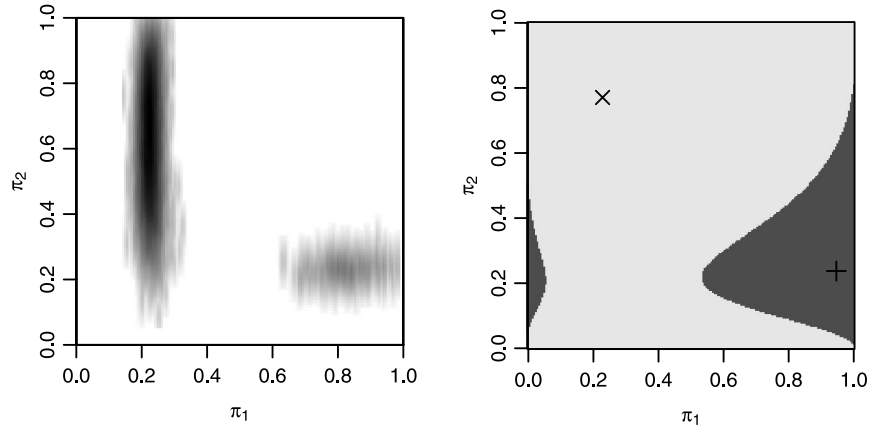

Figure 2. LOH dataset: smoothed marginal bivariate distribution (left) and a slice of the final state space partitioning (right). In the smoothed distribution plot, a darker color corresponds to a higher probability density. In the partition slice plot, parameters η and γ are fixed to their estimated posterior means. Means of mixture components are marked with crosses.

is the entire real line. In our implementation, we study the target distribution restricted to $\mathbf{C}_4 = [-10^{10}, 10^{10}]^4$ which does not have a noticeable effect on the analysis but brings it within the scope of Theorem 1. We run independently four parallel chains using the AM algorithm and RAPTOR with $\alpha = 0.3$, using starting points drawn from a quasi-random distribution uniformly covering the parameter space (in the original parameterization). We run 800,000 iterations for each chain and after dropping the first 40,000 as burn-in we retain only every 40th sample obtained. Starting values for the mixture parameters are included in the Supplemental Materials.

In Figure 2, left panel, we show the marginal scatterplot of (π_1, π_2) for all the samples obtained using the RAPTOR algorithm. In this plot we can clearly distinguish two separate modes with very different size and shape. It is difficult to visualize the partition produced by RAPTOR in the four-dimensional space so instead we show one slice of the partition. If the partition is defined according to (2.7), then, for a fixed subset I of the coordinates of interest and after fixing $x_I = (x_j : j \in I)$ at, say, $\tilde{x}_I$, we can consider the slice through $\mathcal{S}^{(k)}$ determined by $\tilde{x}_I$ as

$$\mathcal{S}^{(k)}(\tilde{x}_I) = \left\{ x_{I^c} : \arg\max_{k'} N\big(x = (\tilde{x}_I, x_{I^c}); \mu^{(k')}, \Sigma^{(k')}\big) = k \right\}, \tag{5.2}$$

where I^c is the complement of set I. We can also define $\mathcal{S}^{(k)}_{I^c}$ the projection of $\mathcal{S}^{(k)}$ on the x_{I^c}-coordinate space and then

$$\mathcal{S}^{(k)}_{I^c} = \bigcup_{\tilde{x}_I} \mathcal{S}^{(k)}(\tilde{x}_I),$$

where the union is taken over all the possible values of $\tilde{x}_I$. One must choose which slices are more informative to look at and here we choose $\tilde{x}$ equal to posterior means. In Figure 2, right panel, a bi-dimensional slice of the RAPTOR regions is plotted when η and γ are equal to their posterior means.

Table 3. Simulation results for LOH data. Region-specific and global parameters means, and mean of first 40 absolute lag-correlations for RAPTOR and AM, for each coordinate.

	RAPTOR			AM		
	Mean	sd	$\overline{\lvert \text{ACF} \rvert}$	Mean	sd	$\overline{\lvert \text{ACF} \rvert}$
η	0.828	0.155	0.15	0.820	0.171	0.27
π_1	0.248	0.106	0.19	0.254	0.121	0.32
π_2	0.614	0.174	0.05	0.608	0.179	0.13
γ	12.732	11.561	0.09	12.406	11.917	0.16

In Table 3 we summarize the RAPTOR and AM estimates on the original scales. We can see that the estimated posterior expected values and standard deviations are quite similar so the two algorithms give consistent results. However, there are important differences in the performances of the two methods. In Table 3 we also report, for each parameter, the mean of the first 40 absolute lag-correlations obtained from the unthinned chains (in the original scale), for RAPTOR and AM. While the mean absolute autocorrelation is smaller in RAPTOR for all model parameters, the increase in efficiency of RAPTOR is strengthened by its average acceptance rate, which was 19.4% compared to 4% for AM.

5.2 ACIDITY OF NORTHEASTERN U.S. LAKES

We consider an application to the acidity dataset first introduced by Small, Sutton, and Milke (1988) and later reconsidered by several authors (Crawford 1994; Richardson and Green 1997). Data are extracted from a survey started in 1983 by the U.S. Environmental Protection Agency to evaluate the acidity of lakes in the northeastern United States. One of the variables registered in the survey was the Acid-Neutralizing Capacity (ANC), which is of great interest as low values of ANC can lead to a loss of biological resources in the lake. The dataset consists of 155 measures of the ANC index on log scale, taken from different lakes. The frequency histogram of this dataset is reported in Figure 3. The data can be modeled using a mixture of two normal distributions

$$\begin{aligned} &p(y|\omega, \mu_1, \sigma_1, \mu_2, \sigma_2, w) \\ &\quad = \omega N(y; \mu_1, \sigma_1^2) + (1-\omega) N(y, \mu_2, \sigma_2^2), \end{aligned}$$

with the following improper, independent priors on the parameters defined over the restricted domain where $\mu_1 \le \mu_2$:

$$\begin{aligned} &p(\mu_k) \propto 1, \qquad p(\sigma_k) \propto \frac{1}{\sigma_k}, \qquad k = 1, 2; \\ &p(\omega) = \text{Unif}(0, 1). \end{aligned}$$

We apply the log transformation to σ_i and the logit transformation to ω so the target distribution has support $\mathbf{R}^5$. In our implementation, we study the posterior distribution on the compact space $\mathbf{C}_5 = [-10^{10}, 10^{10}]^5$.

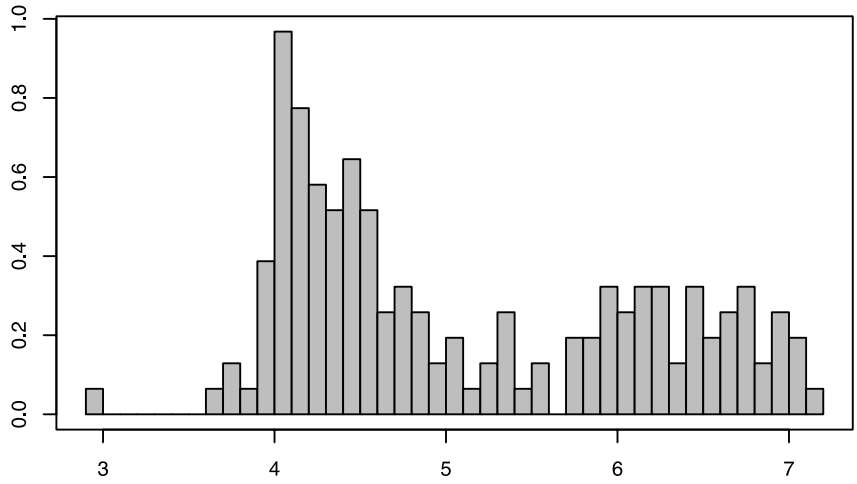

Figure 3. Histogram of the acidity dataset.

We simulate 100,000 realizations from each chain and we retain in our sample every tenth sample obtained. In order to investigate the effect of starting values we run, for each algorithm, two independent chains, starting in different regions of the sample space (the starting values are included in the Supplemental Materials). We compare the AM algorithm and the RAPTOR algorithm with $\alpha = 0.3$. In Table 4 we report the final parameter estimates (posterior mean and standard deviation) together with the mean absolute autocorrelation computed from the first 40 lags of the original (unthinned) chain. For each parameter, RAPTOR has smaller autocorrelations and its acceptance rate, 84.3%, improves significantly the average acceptance rate of AM which is 1.2%.

In Figure 4, left column, we report some of the posterior bivariate distributions of the model parameters. The distributions appear unimodal, but asymmetric and with relatively long tails. Even if in this example the target distribution is unimodal, RAPTOR produces more efficient estimates than the AM algorithm.

In particular, in Figure 4, right column, we show slices of the final partition estimated by the RAPTOR algorithm.

Table 4. Simulation results for acidity data. The standard deviations (sd) of the Monte Carlo estimators for the posterior means are calculated from 1000 independent replicates of the simulation and the chain's average absolute autocorrelations are computed from the first 40 lags.

	RAPTOR			AM		
	Mean	sd	$\overline{\lvert \text{ACF} \rvert}$	Mean	sd	$\overline{\lvert \text{ACF} \rvert}$
μ_1	4.320	0.054	0.039	4.300	0.296	0.567
μ_2	6.204	0.148	0.065	6.142	0.257	0.907
σ_1	0.369	0.052	0.054	0.360	0.168	0.572
σ_2	0.573	0.125	0.073	0.622	0.174	0.947
w_1	0.580	0.059	0.051	0.546	0.112	0.920

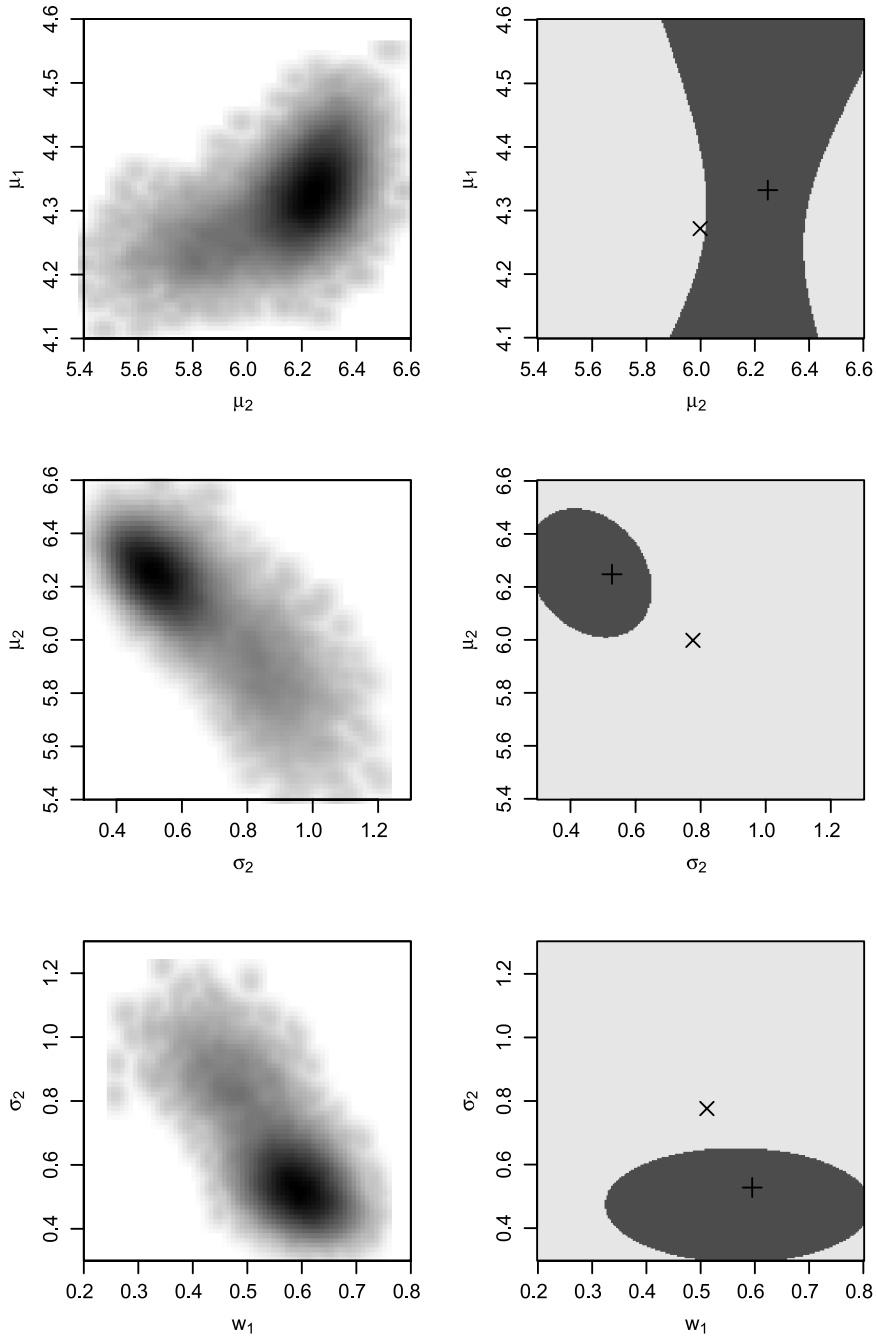

Figure 4. Acidity dataset: posterior bivariate distributions (left column) and a slice of the final state space partitioning (right column), estimated by RAPTOR for different pairs of parameters. All the nonvarying parameters are fixed to their estimated expected value in each slice. Means of mixture components are marked with crosses.

6. CONCLUSION

We propose a mixture-based approach for regional adaptation of the random walk Metropolis algorithm. We assume an approximation of the target by a mixture of Gaussians and use an online EM algorithm to estimate the mixture parameters using the stream of data produced by the MCMC algorithm. In turn, the mixture approximation is used to define an adaptive regional MCMC algorithm which compares favorably with its competitors. In the future we would like to explore the possibility of using mixtures of Student-t distributions as a way to further robustify the method against misspecification of the number of modes, K, and to broaden its applicability.

SUPPLEMENTAL MATERIALS

Theoretical proofs, code, and datasets: A zip archive containing an Appendix, C library with implementations of the algorithms, simulation scripts, LOH and Acidity datasets is available online. The file "Readme.txt" in the archive has a detailed description of the archive. (supplement.zip)

ACKNOWLEDGMENTS

The authors thank the editor, an associate editor, and two referees for a set of thorough reviews that have greatly improved the article. The research has been supported in part by the Natural Sciences and Engineering Research Council of Canada.

[Received March 2009. Revised June 2010.]

REFERENCES

Andrieu, C., and Moulines, E. (2006), "On the Ergodicity Properties of Some Adaptive MCMC Algorithms," *The Annals of Applied Probability*, 16, 1462–1505. [1-3]

Andrieu, C., and Robert, C. P. (2001), "Controlled MCMC for Optimal Sampling," technical report, Université Paris Dauphine. [1]

Andrieu, C., and Thoms, J. (2008), "A Tutorial on Adaptive MCMC," *Statistics and Computing*, 18, 343–373. [2, 7,8]

Andrieu, C., Moulines, E., and Priouret, P. (2005), "Stability of Stochastic Approximation Under Verifiable Conditions," *SIAM Journal on Control and Optimization*, 44, 283–312. [2]

Barrett, M., Galipeau, P., Sanchez, C., Emond, M., and Reid, B. (1996), "Determination of the Frequency of Loss of Heterozygosity in Esophageal Adeno-Carcinoma by Cell Sorting, Whole Genome Amplification and Microsatellite Polymorphisms," *Oncogene*, 12, 1873–1878. [11]

Cappé, O., and Moulines, E. (2009), "Online EM Algorithm for Latent Data Models," *Journal of the Royal Statistical Society, Ser. B*, 71, 593–613. [3]

Craiu, R. V., Rosenthal, J. S., and Yang, C. (2009), "Learn From Thy Neighbor: Parallel-Chain Adaptive and Regional MCMC," *Journal of the American Statistical Association*, 104, 1454–1466. [2,7]

Crawford, S. L. (1994), "An Application of the Laplace Method to Finite Mixture Distributions," *Journal of the American Statistical Association*, 89 (425), 259–267. [13]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, 39, 1–22. [3]

Desai, M. (2000), "Mixture Models for Genetic Changes in Cancer Cells," Ph.D. thesis, University of Washington. [11]

Gelman, A., and Rubin, D. B. (1992), "Inference From Iterative Simulation Using Multiple Sequences" (with discussion), *Statistical Science*, 7, 457–511. [2]

Geyer, C. J., and Thompson, E. A. (1994), "Annealing Markov Chain Monte Carlo With Applications to Ancestral Inference," Technical Report 589, University of Minnesota. [2]

Giordani, P., and Kohn, R. (2010), "Adaptive Independent Metropolis–Hastings by Fast Estimation of Mixtures of Normals," *Journal of Computational and Graphical Statistics*, 19, 243–259. [2]

Guan, Y., and Krone, S. M. (2007), "Small-World MCMC and Convergence to Multi-Modal Distributions: From Slow Mixing to Fast Mixing," *The Annals of Applied Probability*, 17, 284–304. [6]

Haario, H., Saksman, E., and Tamminen, J. (2001), "An Adaptive Metropolis Algorithm," *Bernoulli*, 7, 223–242. [1,6]

——— (2005), "Componentwise Adaptation for High Dimensional MCMC," *Computational Statistics*, 20, 265–273. [1]

Kou, S., Qing, Z., and Wong, W. (2006), "Equi-Energy Sampler With Applications in Statistical Inference and Statistical Mechanics," *The Annals of Statistics*, 34, 1581–1619. [2]

Meyn, S. P., and Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability. Communications and Control Engineering Series*, London: Springer-Verlag. [8]

Neal, R. M. (1994), "Sampling From Multimodal Distributions Using Tempered Transitions," Technical Report 9421, University of Toronto. [2]

——— (2001), "Annealed Importance Sampling," *Statistics and Computing*, 11 (2), 125–139. [2]

Richardson, S., and Green, P. J. (1997), "On Bayesian Analysis of Mixtures With an Unknown Number of Components" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 59 (4), 731–792. [2, 13]

Roberts, G. O., and Rosenthal, J. S. (2001), "Optimal Scaling for Various Metropolis–Hastings Algorithms," *Statistical Science*, 16 (4), 351–367. [6]

——— (2007), "Coupling and Ergodicity of Adaptive Markov Chain Monte Carlo Algorithms," *The Annals of Applied Probability*, 44 (2), 458–475. [2,8]

Roberts, G. O., Gelman, A., and Wilks, W. (1997), "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms," *The Annals of Applied Probability*, 7, 110–120. [6]

Small, M. J., Sutton, M. C., and Milke, M. W. (1988), "Parametric Distributions of Regional Lake Chemistry: Fitted and Derived," *Environmental Science Technology*, 22, 196–204. [13]

Sminchisescu, C., and Triggs, B. (2001), "Covariance-Scaled Sampling for Monocular 3D Body Tracking," in *IEEE International Conference on Computer Vision and Pattern Recognition*, Vol. 1, Hawaii, Washington, DC: IEEE Computer Society Press, pp. 447–454. [2]

Titterington, D. M., Smith, A. F. M., and Makov, U. (1985), *Statistical Analysis of Finite Mixture Distributions*, Chichester: Wiley. [3]

Warnes, G. (2001), "The Normal Kernel Coupler: An Adaptive Markov Chain Monte Carlo Method for Efficiently Sampling From Multi-Modal Distributions," technical report, George Washington University. [11]

West, M. (1993), "Approximating Posterior Distributions by Mixtures," *Journal of the Royal Statistical Society, Ser. B*, 55, 409–422. [2]