Learn From Thy Neighbor: Parallel-Chain and Regional Adaptive MCMC

Radu V. Craiu
Department of Statistics
University of Toronto
Toronto, ON, M5S 3G3, Canada
craiu@utstat.toronto.edu

Jeffrey Rosenthal
Department of Statistics
University of Toronto
Toronto, ON M5S 3G3, Canada
jeff@math.toronto.edu

Chao Yang
Department of Statistics
University of Toronto
Toronto, M5S 3G3, Canada
chaoyang@math.toronto.edu

July 2008; last revised April 2009

Abstract

Starting with the seminal paper of Haario, Saksman and Tamminen (Haario et al. (2001)), a substantial amount of work has been done to validate adaptive Markov chain Monte Carlo algorithms. In this paper we focus on two practical aspects of adaptive Metropolis samplers. First, we draw attention to the deficient performance of standard adaptation when the target distribution is multi-modal. We propose a parallel chain adaptation strategy that incorporates multiple Markov chains which are run in parallel. Second, we note that the current adaptive MCMC paradigm implicitly assumes that the adaptation is uniformly efficient on all regions of the state space. However, in many practical instances, different "optimal" kernels are needed in different regions of the state space. We propose here a regional adaptation algorithm in which we account for possible errors made in defining the adaptation regions. This corresponds to the more realistic case in which one does not know exactly the optimal regions for adaptation. The methods focus on the random walk Metropolis sampling algorithm but their scope is much wider. We provide theoretical justification for the two adaptive approaches using the existent theory build for adaptive Markov chain Monte Carlo. We illustrate the performance of the methods using simulations and analyze a mixture model for real data using an algorithm that combines the two approaches.

Keywords: Adaptive Markov chain Monte Carlo, Metropolis sampling, random walk Metropolis sampling , parallel chains, regional adaptation.

1 Introduction

Markov chain Monte Carlo (MCMC) techniques have become an important tool in the statistician's arsenal for solving complex analyses. One of the most widely used algorithms is the Metropolis (Metropolis et al., 1953) and its generalization, the Metropolis-Hastings (MH) (Hastings, 1970) sampler. If the goal is to sample from a distribution π with support $\mathcal{S}$, the MH sampler is started with a random value $X_0 \sim \mu$ and, at each iteration t, a proposal Y is drawn from a proposal distribution $Q(y|X_t)$ with density $q(y|X_t)$ and is retained as the next state of the chain with probability $\alpha(X_t, Y) = \min\left\{1, \frac{\pi(Y)q(X_t|Y)}{\pi(X_t)q(Y|X_t)}\right\}$. If $q(y|x)$ is the density of $y = x + \epsilon$ where ϵ has a symmetric distribution, we obtain the *random walk Metropolis* algorithm.

In order to design an efficient Metropolis algorithm it is necessary to carefully adapt the parameters of the proposal distribution Q so that the performance of the algorithm is optimal (note that there are multiple definitions of "optimal" available). On one hand one can argue that many modern MCMC algorithms incorporate a certain notion of local adaptation in their design, e.g. Gilks et al. (1998), Liu et al. (2000) and Craiu and Lemieux (2007), Green and Mira (2001), Eidsvik and Tjelmeland (2006). In this paper, we refer to a more global version of adaptation which is based on learning the geography of π "on the fly" from all the samples available up to the current time t. Such an approach violates the Markovian property as the subsequent realizations of the chain depend not only on the current state but also on all past realizations. This implies that one can validate theoretically this approach only if one is able to prove from first principles that the adaptive algorithm is indeed sampling from π. In Haario et al. (2001) the authors provide such a theoretical justification for adapting the covariance matrix Σ of the Gaussian proposal density used in a random walk Metropolis. They continually adapt Σ using the empirical distribution of the available samples. Their choice of adaptation is motivated by the optimal results proved by Roberts et al. (1997) and Roberts and Rosenthal (2001). Subsequently, the convergence results of adaptive algorithms have been made more general in Andrieu and Robert (2001), Andrieu et al. (2005), Andrieu and Moulines (2006), Atchade and Rosenthal (2005), and Roberts and Rosenthal (2007). An adaptive algorithm for the independent Metropolis sampler was proposed by Gasemyr (2003) and Haario et al. (2005) extended their previous work to Metropolis-within-Gibbs sampling. A class of quasi-perfect adaptive MCMC algorithms is introduced by Andrieu and Atchade (2006) and a nice tutorial on adaptive methods is given by Andrieu and Thoms (2008). Alternative approaches to adaptation within MCMC can be found in Brockwell and Kadane (2005), Nott and Kohn (2005), Giordani and Kohn (2006). We quote from Giordani and Kohn (2006):

> Although more theoretical work can be expected, the existing body of results provides sufficient justification and guidelines to build adaptive MH samplers for challenging problems. The main theoretical obstacles having been solved, research is now needed to design efficient and reliable adaptive samplers for broad classes of problems.

In the present paper we try to close some of the gap between theory and practice by focusing on the practical aspects of adaptive MCMC (AMCMC). More precisely, we discuss complications arising when using AMCMC, especially adaptive random walk Metropolis, for sampling from multi-modal targets and also when the optimal proposal distribution is regional, i.e. the optimal proposal should change across regions of the state space. In the next section we discuss the inter-chain adaptation. In Section 3 we discuss the regional adaptation. The theoretical challenge is to show that the algorithms proposed here fall within the scope of general theorems that are used to validate adaptive MCMC. These results are presented in Section 4 while simulation examples and a real data analysis are shown in Section 5. We close with discussion of further research.

2 Inter-chain Adaptation (INCA)

To begin, consider a simulation setting where the target distribution is a mixture of two ten-dimensional Gaussian distributions. More precisely, the target distribution is

$$\pi(x|\mu_1, \mu_2, \Sigma_1, \Sigma_2) = 0.5n_{10}(x; \mu_1, \Sigma_1) + 0.5n_{10}(x; \mu_2, \Sigma_2),$$

with $n_d(x; \mu, \Sigma)$ denoting the density of a d-dimensional Gaussian random variable with mean μ and covariance matrix Σ and where $\mu_1 = (0.03, -0.06, -0.24, -1.39, 0.52, 0.61, 1.26, -0.71, -1.38, -1.53)^T$, $\mu_{1i} - \mu_{2i} = 6$, $\forall 1 \le i \le 10$, $\Sigma_1 = \mathbf{I}_{10}$ and $\Sigma_2 = 4\mathbf{I}_{10}$. In Figure 1 we present the results of a simulation in which we applied the adaptive Metropolis sampler of Haario et al. (2001) with an initialisation period of 10,000 samples. The chain is started in one of the target's modes (the one corresponding to μ_1). Although the final sample size is $N = 250,000$, we can see that the chain does not visit the second mode. In this case, the adaptation can not improve much on the unadapted version of the Metropolis sampler as the second mode "is invisible" in the initialization period and it will likely take a long time for a chain incorrectly adapted to a unimodal distribution to discover the second high probability region.

In the classic MCMC literature difficulties related to sampling from a multi-modal

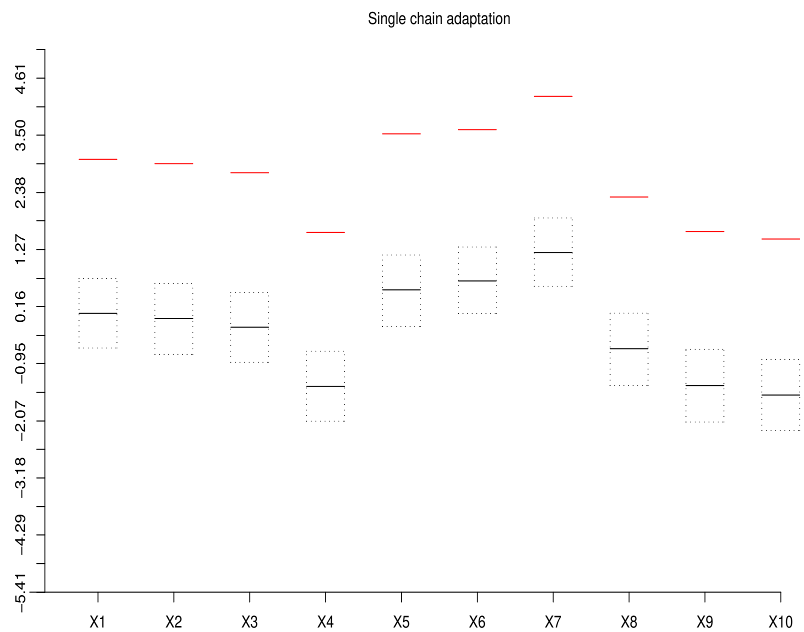

Figure 1: *Boxplots of N=250,000 samples obtained using a single-chain adaptive Metropolis; each boxplot correspond to one component of the 10-dimensional random vector. The red lines represent the entries of the target's mean vector. The chain does not visit the second mode of the target.*

distribution are tackled using multiple parallel chains as in Gelman and Rubin (1992) and tempering as in Neal (1994) and Geyer and Thompson (1994). Both ideas influence our approach.

The parallel chains implementation has been proven helpful for a more systematic exploration of the sample space as in Craiu and Meng (2005). In the present setting we use it to detect the different regions of significant mass under the posterior distribution and our starting example shows that such detection is extremely important for adaptive MCMC. We thus propose running in parallel a number, say K, of Markov chains. We can further robustify the performance of the algorithm if the chains are started from a distribution μ that is *overdispersed* with respect to π. It should be noted that finding μ can be quite challenging. The problem of finding good starting points for parallel chains is also discussed by Jennison (1993), Applegate et al. (1990), Gelman and Rubin (1992) and Brooks and Gelman (1998). We would like to add a word of caution following Gill (2008) which states that a bad choice for μ can be deleterious and may dramatically alter the simulation results.

A question of interest in adaptive MCMC is whether one should wait a short or a long time before starting the adaptation. In Gasemyr (2003), the burn-in time is random but bounded below, while Giordani and Kohn (2006) propose a different strategy in which adaption starts early and is performed frequently in what they call *intensive adaptation*. However, they also warn that one should make sure that enough distinct samples are obtained in order to avoid singularity problems. In the multi-modal situation considered here we adopt a longer burn-in, partly because the multi-modality of π makes it difficult to have a good idea about its geography when only a few draws are available. A longer burn-in increases the stability of the inferences obtained and reduces the risk of missing one important mode.

We thus propose a new strategy, *inter-chain adaptive MCMC (INCA)*, as follows. We run K different chains in parallel, each started independently from the same overdispersed starting distribution. After the burn-in period the K kernels are simultaneously adapted using *all the samples* provided by the K chains so far. In the case of a random walk Metropolis with Gaussian proposals we do this by setting the proposal covariance to the sample covariance matrix of all the available samples. Denote γ_m the adaptation parameter, e.g. the variance of the random walk Metropolis proposal distribution, used at step m in each marginal transition kernel. We run the chains independently conditional on the $\{\gamma_m\}$, so the joint transition kernel, $\tilde{T}_{\gamma_m}$ is obtained as the product of K identical copies of the marginal transition kernel T_{γ_m} such that

$$\tilde{T}_{\gamma_m}(\tilde{x}, \tilde{A}) = T_{\gamma_m}(x_1; A_1) \otimes T_{\gamma_m}(x_2; A_2) \otimes \ldots \otimes T_{\gamma_m}(x_K; A_K),$$

where $\tilde{A} = A_1 \times \ldots \times A_K$ and $\tilde{x} = (x_1, \ldots, x_K)$.

The motivation for using multiple chains lies in our attempt to discover as early as possible all the modal regions of π (or at least all the important ones). After the chains have explored the regions of interest and the simulation parameters are updated one may wish to return to a single chain. A question of interest is then how to decide when the exchange of information between chains has stopped. The criterion we use is the well-known Brooks-Gelman-Rubin (BGR) diagnostic, R, as developed in Gelman and Rubin (1992) and Brooks and Gelman (1998). Given a number, say K, of parallel chains, the *potential scale reduction* R is a normalized ratio of the between-chain and within-chain variances computed from the available samples (Gelman and Rubin (1992), page 465). While R was originally designed as a convergence indicator, here it is used to determine whether the chains contribute different information about π.

In Figure 2 we show for the mixture of Gaussian distributions the evolution of the R statistics. One can see that the exchange of information between chains is gradually decreasing along with the adaptation. An astute reader may wonder whether the learning process can be accelerated using tempering strategies in order to learn the geography of π more quickly.

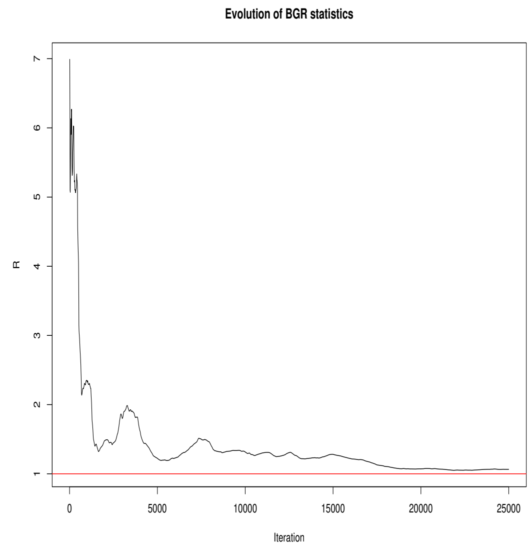

Figure 2: *The evolution of BGR's R statistic. It takes approximately 18,000 iterations to reach below 1.1.*

2.1 Tempered INCA (TINCA)

Tempering in MCMC relies on a series of "canonical" distributions, π_T, each of which are obtained by varying a "temperature" parameter T in a set $\{t_1, t_2, \ldots, t_{max}\}$ such that $\pi_{t_1} = \pi$ and while π_{t_j} is not too different from $\pi_{t_{j+1}}$, there is a substantial difference between π and $\pi_{t_{max}}$ in that the latter has less isolated modes (or is "flatter") so that it is considerably easier to sample using MCMC algorithms. One generic procedure (although not the only one) defines $\pi_T = \pi^{1/T}$ for $T \geq 1$. In Figure 3 we illustrate the effect of tempering on a bivariate mixture of Gaussian distributions.

One expects that for large values of T (or hot temperatures), adaptive algorithms designed for π_T will be more efficient. For instance, if INCA is implemented we expect the running time needed to stabilize $R \approx 1$ to be much shorter than at the "cool" temperature $T = 1$. One could possibly envision a gradual temperature-driven adaptation following Meng (2007). Start with $T = t_{max}$ and at each temperature perform the following steps:

Step I For $T = t_j$ perform INCA for target density π_T until R is below a pre-specified

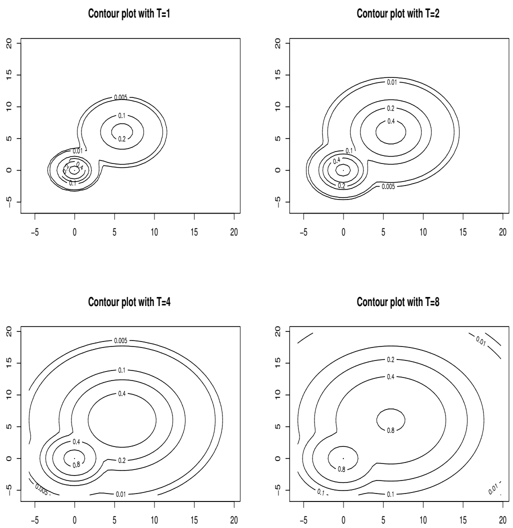

Figure 3: *Tempered distributions with* $T = 1$ *(original target),* $T = 2$*,* $T = 4$ *and* $T = 8$*.*

threshold $1 + \xi$.

Step II Keep the simulation parameters obtained and perform Step I with the next colder temperature $T = t_{j-1}$. Stop after $T = 1$.

The implementation assumes that the kernel learned/adapted at temperature t_j is a reasonable starting choice for the kernel used at temperature t_{j-1}. In addition to speeding up the adaptation process, this *tempered INCA (TINCA)* is aimed at solving the difficult task of producing a reasonable starting proposal in a high dimensional problem. We implemented TINCA with $T \in \{1, 2, 4, 8, 16\}$ for the example discussed in this section and the total number of iterations, including those produced at temperatures $T > 1$, required to reach $R \leq 1.1$ at $T = 1$ was 10,000, compared to the 18,000 reported without tempering. Additional simulations using TINCA are discussed in Section 5.

It should be noted that INCA and/or TINCA can be implemented along many other adaptive MCMC strategies. As mentioned by many authors working in the field, the performance of the algorithm during the initialization (or burn-in) period, when no adaption is taking place, is crucial. We believe that INCA is most useful in the initial stage of the simulation since it accelerates the "data gathering" about the geography of π and improves the overall performance of the adaptive process.

3 Regional Adaptation (RAPT)

In the previous section, we considered a simple example in which the target distribution had its mass equally divided between the two modes. However, examples abound where the modes of the distribution have different relative mass and in these situations a simple remedy such as INCA may be ineffective. One can easily see that in such cases there is no "universal" good proposal, i.e. the learning must be adapted to different regions of the state space. Regional adaptation has been suggested in a different form by Andrieu and Robert (2001) and Roberts and Rosenthal (2009). For our discussion assume that there is a partition of the space $\mathcal{S}$ made of two regions $\mathcal{S}_{01}, \mathcal{S}_{02}$ such that adaptation should be carried over independently in the two regions. In other words, in the case of a Metropolis algorithm, in region $\mathcal{S}_{0i}$ we would use proposals from distribution Q_i while only samples from this region will be used to adapt Q_i. Such an algorithm is valid as long as one carefully computes acceptance ratios for proposed moves that switch regions, as was also noted by Roberts and Rosenthal (2009). In the case of two regions the acceptance ratio is then

$$r(x, x_{new}) = \begin{cases} \frac{\pi(x_{new})}{\pi(x)}, & \text{if } x, x_{new} \in \mathcal{S}_{0i} \\ \frac{\pi(x_{new}) q_1(x|x_{new})}{\pi(x) q_2(x_{new}|x)}, & \text{if } x \in \mathcal{S}_{02}, x_{new} \in \mathcal{S}_{01} \\ \frac{\pi(x_{new}) q_2(x|x_{new})}{\pi(x) q_1(x_{new}|x)}, & \text{if } x \in \mathcal{S}_{01}, x_{new} \in \mathcal{S}_{02}. \end{cases},$$

where q_i is the density of Q_i.

While there exist sophisticated methods to detect the modes of a multimodal distribution (see Sminchisescu and Triggs, 2001, 2002; Sminchisescu et al., 2003; Neal, 2001), it is not always clear how to use such techniques since defining a good partition of the sample space may need more than just the location of the modes. In Craiu and Di Narzo (2009) we follow the methods of Andrieu and Moulines (2006) and Cappé

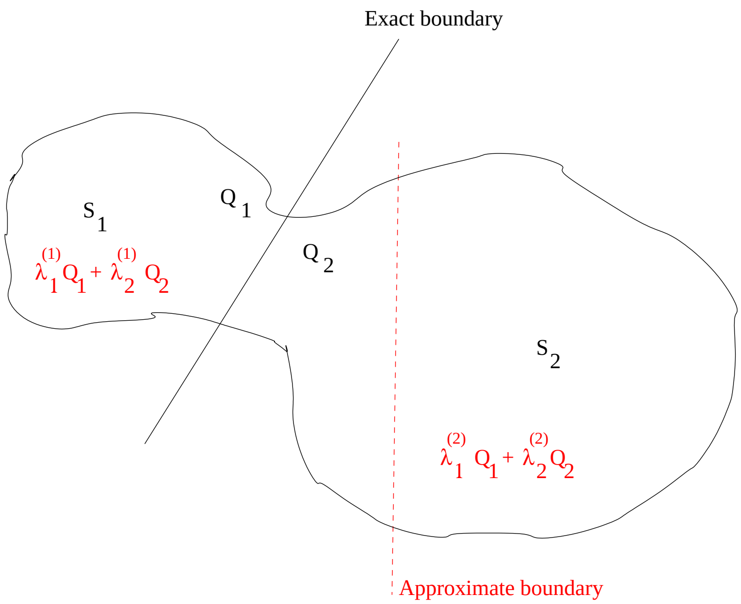

Figure 4: *Illustration of the regional adaptive MCMC sampler. The dashed black line represents the exact boundary (unknown), between regions $\mathcal{S}_{01}$ and $\mathcal{S}_{02}$. The dashed red line delimitates the regions $\mathcal{S}_1$ and $\mathcal{S}_2$ used for the regional adaptation.*

and Moulines (2009) to propose a mixture-based approach for adaptively determining the boundary between high probability regions. Suppose we approximate the target distribution using the mixture of Gaussians

$$\tilde{Q}(x) = \beta n(x; \mu_1, \Sigma_1) + (1 - \beta) n(x; \mu_2, \Sigma_2). \tag{1}$$

Then Craiu and Di Narzo (2009) define the regions $\mathcal{S}_k$ as the set in which the k-th component of the mixture density $\tilde{Q}$ dominates the other one., i.e.

$$\mathcal{S}_k = \{x : \ \arg\max_{k'} n(x; \mu_{k'}, \Sigma_{k'}) = k\}. \tag{2}$$

Regardless of the method used, in most cases we do not have enough knowledge to choose the partition made exactly of regions $\mathcal{S}_{01}$ and $\mathcal{S}_{02}$. Instead, suppose we define a partition made of regions $\mathcal{S}_1$ and $\mathcal{S}_2$. An illustration of this idea is shown in Figure 4. The solid black line represents the exact boundary (unknown), between regions $\mathcal{S}_{01}$ and $\mathcal{S}_{02}$. The dashed red line delimitates the regions $\mathcal{S}_1$ and $\mathcal{S}_2$ used for the regional adaptation. If we were to apply the simple regional adaptive algorithm described above, when the chain is in one of the states situated between the two dashed lines the wrong proposal would be used. Therefore, in order to account for the error made when specifying the boundary between regions we propose to sample our proposals from a mixture that includes both Q_1 and Q_2. However, the mixture proportions are different in each region $\mathcal{S}_i$ and are adaptively modified. The resulting Regional Adaptive (RAPT) algorithm has proposal distribution

$$Q_\gamma(x, dy) = \sum_{i=1}^{2} 1_{\mathcal{S}_i}(x)[\lambda_1^{(i)} Q_1(x, dy) + \lambda_2^{(i)} Q_2(x, dy)], \tag{3}$$

where $\lambda_1^{(i)} + \lambda_2^{(i)} = 1$. In this case, we use the index γ on Q to emphasize the fact that the proposal is adapted with the adaption parameter $\gamma = (\lambda_1^{(1)}, \lambda_1^{(2)}) \in \mathcal{Y} = [0, 1]^2$.

The mixture proportions $\lambda_j^{(i)}$, $1 \leq i, j \leq 2$ are chosen to reflect which of the two proposals is more "appropriate" to use in the given region. Evidently, one has some freedom over what can be considered a good proposal in this setup. For instance, one could choose

$$\lambda_j^{(i)} = \frac{n_j^{(i)}(t)}{\sum_{h=1}^2 n_h^{(i)}(t)},$$

where $n_j^{(i)}(t)$ is the number of accepted moves up to time t computed when the accepted proposals are distributed with Q_j and the current state of the chain lies in $\mathcal{S}_i$. However, this choice would favor small steps of the chain since these have higher acceptance rates. To counterbalance, we take into account the average squared jump distance so that

$$\lambda_j^{(i)} = \begin{cases} \frac{d_j^{(i)}(t)}{\sum_{h=1}^2 d_h^{(i)}(t)}, & \text{if } \sum_{h=1}^2 d_h^{(i)}(t) > 0 \\ 1/2, & \text{otherwise} \end{cases}, \tag{4}$$

where $d_j^{(i)}(t)$ is the average squared jump distance up to time t computed when the proposals were sampled from Q_j and the current state of the chain lies in $\mathcal{S}_i$. More precisely, suppose $\{x_j\}_{j=0}^t$ are the samples obtained until time t and $N_i(t)$ is the number of elements in the set $\{x_{t_g}^i\}_{g=1}^{N_i(t)}$ which contains all the samples generated up to time t that are lying in $\mathcal{S}_i$. We also define the set of time points at which the proposal is generated from Q_j and the current state is in $\mathcal{S}_i$, $W_j^{(i)}(t) = \{0 \le s \le t : x_s \in \mathcal{S}_i$ and proposal at time s is generated from $Q_j\}$. Then

$$d_j^{(i)}(t) = \frac{\sum_{s \in W_j^{(i)}(t)} |x_{s+1} - x_s|^2}{|W_j^{(i)}(t)|},$$

where $|W_j^{(i)}(t)|$ denotes the number of elements in the set $W_j^{(i)}(t)$. If $W_j^{(i)}(t) = \emptyset$ then $d_j^{(i)}(t) = 0$. If we implement RAPT within INCA/TINCA with K parallel chains then in the calculation of $d_j^{(i)}(t)$ we need to consider *all* the samples obtained up to time t by *all* the K chains.

Better performance can be achieved using the algorithm (3) for which both the mixture weights and the proposals, Q_1, Q_2, are adapted, which is called *Dual RAPT*. We suggest here to adapt the covariance matrix of each proposal distribution in the same vein as Haario et al. (2001).

When the current state X_{t-1} lies in $\mathcal{S}_i$, the components of the mixture (3) are the Gaussian distributions with densities $q_i^{(t)}$ and with mean at the current point X_{t-1} and covariance $C_i(t)$, where $C_i(t)$ is defined below.

$$C_i(t) = \begin{cases} C_{0i}, \quad t \le t_0 \\ s_d \ \mathrm{Cov}(X_{t_1}^i, X_{t_2}^i, \cdots, X_{t_{N_i(t)}}^i) + s_d \epsilon \ \mathbf{I}_d, \quad t > t_0 \end{cases}, \ i = 1, 2, \tag{5}$$

where $s_d = (2.4)^2/d$. This form of adaption follows (separately within each region) the Adaptive Metropolis algorithm of Haario et al. (2001), and is based on the results of Gelman et al. (1996), Roberts et al. (1997), and Roberts and Rosenthal (2001) who showed that this choice optimizes the mixing of random walk Metropolis at least in the case of Gaussian targets and Gaussian proposals. The implicit premise is that in each region the Gaussian approximation of the target is reasonable. The addition of $s_d \epsilon \ \mathbf{I}_d$, where $\epsilon > 0$ is a small constant, guarantees that the matrices $C_i(t)$ are all in $\mathbb{M}(c_1, c_2)$ for some fixed constants $0 < c_1 \le c_2 < \infty$, where $\mathbb{M}(c_1, c_2)$ is the set of all $k \times k$ positive definite matrices M such that $c_1 \mathbf{I}_k \le M \le c_2 \mathbf{I}_k$, i.e. such that both $M - c_1 \mathbf{I}_k$ and $c_2 \mathbf{I}_k - M$ are non-negative definite.

The adaption parameter is then

$$\gamma = (\lambda_1^{(1)}, \lambda_1^{(2)}, C_1, C_2) \in \mathcal{Y} = [0,1] \times [0,1] \times \mathbb{M}(c_1, c_2) \times \mathbb{M}(c_1, c_2).$$

An observant reader may notice that while the algorithm may perform well in each region, there is no guarantee that there will be a good flow *between* regions. For this reason, in practice we consider the *Mixed RAPT* algorithm in which we add a third adaptive component to the mixture (3). In this variant,

$$Q_\gamma(x, dy) = (1-\beta)\sum_{i=1}^{2} 1_{\mathcal{S}_i}(x)[\lambda_1^{(i)} Q_1(x, dy) + \lambda_2^{(i)} Q_2(x, dy)] + \beta Q_{whole}(x, dy), \quad (6)$$

where Q_{whole} is adapted using all the samples in $\mathcal{S}$ and β is constant throughout the simulation. Once more we adapt the ideas in Haario et al. (2001) and use the covariance of all the simulations available at time t to adapt the covariance of the Gaussian proposal density q_{whole} in (6). We shall use

$$C(t) = \begin{cases} C_0, & t \le t_0 \\ s_d \operatorname{Cov}(X_0, X_1, \cdots, X_t) + s_d \epsilon\, \mathbf{I}_d, & t > t_0 \text{ and } \operatorname{Tr}(C(t)) \le M \end{cases} . \quad (7)$$

Given that all the distributions and parameters (except β) in (6) are evolving, the adaption parameter is

$$\gamma = (\lambda_1^{(1)}, \lambda_1^{(2)}, C_1, C_2, C) \in \mathcal{Y} = [0,1] \times [0,1] \times \mathbb{M}(c_1, c_2) \times \mathbb{M}(c_1, c_2) \times \mathbb{M}(c_1, c_2).$$

3.1 INCA/TINCA Versions of RAPT

The descriptions so far of the various RAPT, Dual RAPT, and Mixed RAPT algorithms have all been for a single chain. However, it is also possible to combine these algorithms with the INCA approach of Section 2.

Indeed, for RAPT, all that is required is to compute the quantities $d_j^{(i)}(t)$ in equation (4) using *all* of the proposals from *all* of the K parallel chains.

For Dual RAPT, it is required in addition that the covariance matrix adaptions of equation (5) use the appropriate samples X from *all* of the K parallel chains.

And, for Mixed RAPT, it is required in addition that the covariance matrix adaptions of equation (7) also use the appropriate samples X from *all* of the K parallel chains.

Similarly, it is possible to combine all of this with the tempered (TINCA) approach of Section 2.1. Indeed, all that is required is to run each of the chains on the distribution $\pi_{T_j} = \pi^{1/T_j}$ until $R < 1 + \epsilon$, and then to replace j by $j-1$ and continue, until such time as we reach $T_j = 1$ corresponding to $\pi_{T_j} = \pi$.

4 Theoretical Results

In this section, we prove that each of our previously-defined adaptive algorithms is "ergodic to π", i.e. that

$$\lim_{n\to\infty} \sup_{A\subseteq S} |\mathbf{P}(X_n \in A) - \pi(A)| \;=\; 0\,,$$

assuming the following compactness condition:

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

We believe that it is possible to remove the assumption that $\mathcal{S}$ is compact, but the resulting arguments are more technical, so we will pursue them elsewhere (Yang et al., 2009). Of course, even compact sets can be arbitrarily large, so in practice (A1) does not impose any significant limitation.

We shall first prove ergodicity of the RAPT algorithm, where only the weights $\lambda_j^{(i)}$ are adapted, as in (4). In this case, since the proposal densities q_i are arbitrary, we also need to assume that they are continuous and positive throughout $\mathcal{S}$.

Theorem 4.1. *Assuming (A1), and that the proposal densities q_i are continuous and positive throughout $\mathcal{S} \times \mathcal{S}$, the RAPT algorithm is ergodic to π.*

We shall then prove ergodicity of the Dual RAPT algorithm. In this case, since the proposal distributions are assumed to be Gaussian, no further assumptions are necessary.

Theorem 4.2. *Assuming (A1), the Dual RAPT algorithm is ergodic to π.*

Finally, we shall prove ergodicity of the full Mixed RAPT algorithm, again with no further assumptions required since the proposals are Gaussian.

Theorem 4.3. *Assuming (A1), the Mixed RAPT algorithm is ergodic to π.*

Note that Theorems 4.1, 4.2, and 4.3 apply both to the single-chain versions of RAPT / Dual RAPT / Mixed RAPT as described in Section 3, and to the INCA/TINCA modifications as described in Section 3.1.

4.1 Theorem Proofs

For notational simplicity, we prove the theorems for the case of a single adaptive chain, but the proofs go through virtually without change for the INCA versions of these algorithms as described in Section 3.1, and also (by iterating) for the TINCA versions as described in Sections 2.1 and 3.1.

To facilitate our proofs, we introduce some notation. Let γ be shorthand for all of the parameters being adapted, e.g.

$$\gamma = (\lambda_1^{(1)}, \lambda_1^{(2)}, \lambda_2^{(1)}, \lambda_2^{(2)})$$

for the RAPT algorithm, while

$$\gamma = (\lambda_1^{(1)}, \lambda_1^{(2)}, \lambda_2^{(1)}, \lambda_2^{(2)}, C_1, C_2)$$

for the Dual RAPT algorithm, etc. Let Γ_n be the actual (random) adaptive parameters in use at time n, so that P_{Γ_n} is the (random) Markov chain kernel used to update the state at time n. Write P_γ for the Markov chain kernel corresponding to a particular fixed choice γ, so that

$$P_\gamma(x, A) \;=\; \mathbf{P}(X_{n+1} \in A \,|\, X_n = x,\, \Gamma_n = \gamma)\,.$$

A basic assumption of adaptive MCMC is that each *individual* kernel P_γ preserves the stationarity of π, i.e. that

$$\int P_\gamma(x, A)\, \pi(dx) \;=\; \pi(A)\,, \qquad A \subseteq \mathcal{S} \tag{8}$$

for *fixed* γ, which is certainly true for the adaptive algorithms introduced here. However, when the parameters $\{\gamma_n\}$ are modified during the run, then stationarity of π no longer holds, and the resulting ergodicity is much more subtle. For a simple graphical illustration of this, see Rosenthal (2004).

Our proofs shall make use of Theorem 5 of Roberts and Rosenthal (2007), which implies that an adaptive algorithm is ergodic to π if it satisfies (a) the *diminishing adaption* property that

$$\lim_{n\to\infty} \sup_{x\in\mathcal{S}} \sup_{A\subseteq\mathcal{S}} |P_{\Gamma_{n+1}}(x,A) - P_{\Gamma_n}(x,A)| \;=\; 0 \tag{9}$$

in probability, i.e. that the amount of adaptive change from time n to time $n+1$ goes to zero as $n \to \infty$, and (b) the *simultaneous uniform ergodicity* property that there is $\rho < 1$ with

$$|P^n_\gamma(x,A) - \pi(A)| \;\leq\; \rho^n\,, \qquad n \in \mathbf{N},\; \gamma \in \mathcal{Y},\; x \in \mathcal{S},\; A \subseteq \mathcal{S}\,. \tag{10}$$

So, to prove Theorem 4.1, it suffices to establish (9) and (10), which we do in the following two lemmas.

Lemma 4.1. *Under the conditions of the Theorem 4.1, the simultaneous uniform ergodicity property (10) holds.*

Proof. Since $\mathcal{S}$ is compact, by positivity and continuity we have $d \equiv \sup_{x\in\mathcal{S}} \pi(x) < \infty$ and $\epsilon \equiv \min\{\inf_{x,y\in\mathcal{S}} q_1(x,y), \inf_{x,y\in\mathcal{S}} q_2(x,y)\} > 0$. From (3), it follows that

$$q_\gamma(x,y) \;\equiv\; \sum_{i=1}^{2} 1_{\mathcal{S}_i}(x)[\lambda_1^{(i)} q_1(x,y) + (1-\lambda_1^{(i)}) q_2(x,y)] \;\geq\; \epsilon\,, \qquad x, y \in \mathcal{S}\,.$$

For $x \in \mathcal{S}$ and $B \subseteq \mathcal{S}$, denote

$$R_{x,\gamma}(B) = \left\{ y \in B : \frac{\pi(y) q_\gamma(y,x)}{\pi(x) q_\gamma(x,y)} < 1 \right\}$$

and $A_{x,\gamma}(B) = B \setminus R_{x,\gamma}(B)$. Then we have

$$\begin{aligned}
P_\gamma(x,B) &\geq \int_{R_{x,\gamma}(B)} q_\gamma(x,y) \min\left\{\frac{\pi(y)q_\gamma(y,x)}{\pi(x)q_\gamma(x,y)}, 1\right\} \mu^{Leb}(dy) \\
&\quad + \int_{A_{x,\gamma}(B)} q_\gamma(x,y) \min\left\{\frac{\pi(y)q_\gamma(y,x)}{\pi(x)q_\gamma(x,y)}, 1\right\} \mu^{Leb}(dy) \\
&= \int_{R_{x,\gamma}(B)} \frac{\pi(y)q_\gamma(y,x)}{\pi(x)} \mu^{Leb}(dy) + \int_{A_{x,\gamma}(B)} q_\gamma(x,y) \mu^{Leb}(dy) \\
&\geq \frac{\epsilon}{d} \int_{R_{x,\gamma}(B)} \pi(y)\mu^{Leb}(dy) + \frac{\epsilon}{d} \int_{A_{x,\gamma}(B)} \pi(y)\mu^{Leb}(dy) = \frac{\epsilon}{d}\pi(B).
\end{aligned}$$

Thus $\mathcal{S}$ is small since

$$P_\gamma(x,B) \geq \nu(B)\,, \qquad x \in \mathcal{S},\ \gamma \in \mathcal{Y},\ B \subseteq \mathcal{S},$$

where $\nu(B) = \frac{\epsilon}{d}\pi(B)$ is a non-trivial measure on $\mathcal{S}$. Condition (10) then follows from Theorem 16.0.2 of Meyn and Tweedie (1993), with $\rho = 1 - \nu(\mathcal{S}) = 1 - \frac{\epsilon}{d}$. $\square$

Lemma 4.2. *Under the conditions of the Theorem 4.1, the diminishing adaption condition (9) holds.*

Proof. Let $f_\lambda(x,y) = \lambda q_1(x,y) + (1-\lambda)q_2(x,y)$. Since $\mathcal{S}$ is compact, we have that $M \equiv \max\{\sup_{x,y\in\mathcal{S}} q_1(x,y), \sup_{x,y\in\mathcal{S}} q_2(x,y)\} < \infty$. For any $x \in S_1$ and $A \in \mathcal{B}(\mathcal{S})$, we have:

$$\begin{aligned}
P_{\gamma_k}(x,A) &= \int_{A\cap S_1} f_{\lambda_1^{(1)}(k)}(x,y) \cdot \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} dy \\
&+ \int_{A\cap S_2} f_{\lambda_1^{(1)}(k)}(x,y) \min\left\{1, \frac{\pi(y) f_{\lambda_1^{(2)}(k)}(x,y)}{\pi(x) f_{\lambda_1^{(1)}(k)}(x,y)}\right\} dy \\
&+ \delta_x(A) \int_{S_1} f_{\lambda_1^{(i)}(k)}(x,y) \cdot \left[1 - \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}\right] dy \\
&+ \delta_x(A) \int_{S_2} f_{\lambda_1^{(1)}(k)}(x,y) \left[1 - \min\left\{1, \frac{\pi(y) f_{\lambda_1^{(2)}(k)}(x,y)}{\pi(x) f_{\lambda_1^{(1)}(k)}(x,y)}\right\}\right] dy.
\end{aligned}$$

Denote the first term $I_k(x,A)$, the second term $II_k(x,A)$, the third term $III_k(x,A)$

and the fourth term $IV_k(x,A)$. Then we have:

$$\begin{aligned}|P_{\gamma_{k+1}}(x,A)-P_{\gamma_k}(x,A)| &\le |I_{\gamma_{k+1}}(x,A)-I_{\gamma_k}(x,A)|+|II_{\gamma_{k+1}}(x,A)-II_{\gamma_k}(x,A)| \\ &+ |III_{\gamma_{k+1}}(x,A)-III_{\gamma_k}(x,A)|+|IV_{\gamma_{k+1}}(x,A)-IV_{\gamma_k}(x,A)|.\end{aligned}$$

Let

$$\alpha_{k_1^{(i)}}(x,y)=\min\left\{1,\frac{\pi(y)[\lambda_1^{(i)}(k)q_1(y,x)+(1-\lambda_1^{(i)}(k))q_2(y,x)]}{\pi(x)[\lambda_1^{(1)}(k)q_1(x,y)+(1-\lambda_1^{(1)}(k))q_2(x,y)]}\right\}.$$

Then

$$\begin{aligned}|II_{\gamma_{k+1}}(x,A)-II_{\gamma_k}(x,A)| &\le \int_{A\cap S_2}|f_{\lambda_1^{(1)}(k+1)}(x,y)\alpha_{(k+1)_1^{(2)}}(x,y)-f_{\lambda_1^{(1)}(k)}(x,y)\alpha_{k_1^{(2)}}(x,y)|dy \\ &\le \int_{A\cap S_2}|f_{\lambda_1^{(1)}(k+1)}(x,y)\alpha_{(k+1)_1^{(2)}}(x,y)-f_{\lambda_1^{(1)}(k+1)}(x,y)\alpha_{k_1^{(2)}}(x,y) \\ &+ f_{\lambda_1^{(1)}(k+1)}(x,y)\alpha_{k_1^{(2)}}(x,y)-f_{\lambda_1^{(1)}(k)}(x,y)\alpha_{k_1^{(2)}}(x,y)|dy \\ &\le \int_{A\cap S_2}f_{\lambda_1^{(1)}(k+1)}(x,y)|\alpha_{(k+1)_1^{(1)}}(x,y)-\alpha_{k_1^{(1)}}(x,y))|dy \\ &+ \int_{A\cap S_2}\alpha_{k_1^{(1)}}(x,y))|f_{\lambda_1^{(1)}(k+1)}(x,y)-f_{\lambda_1^{(1)}(k)}(x,y)|dy \\ &\le M\int_{A\cap S_2}|\alpha_{(k+1)_1^{(1)}}(x,y)-\alpha_{k_1^{(1)}}(x,y))|dy \\ &+ \int_{A\cap S_2}|f_{\lambda_1^{(1)}(k+1)}(x,y)-f_{\lambda_1^{(1)}(k)}(x,y)|dy\,.\end{aligned}$$

Now,

$$\begin{aligned}&M\int_{A\cap S_2}|\alpha_{(k+1)_1^{(1)}}(x,y)-\alpha_{k_1^{(1)}}(x,y))|dy \\ =\ & M\int_{A\cap S_2}\frac{\pi(y)}{\pi(x)}\left|\frac{f_{\lambda_{k+1}^{(2)}}(x,y)}{f_{\lambda_{k+1}^{(1)}}(x,y)}-\frac{f_{\lambda_k^{(2)}}(x,y)}{f_{\lambda_k^{(1)}}(x,y)}\right|dy \\ \le\ & \frac{Md}{\pi(x)}\int_{A\cap S_2}\left|\frac{f_{\lambda_{k+1}^{(2)}}(x,y)}{f_{\lambda_{k+1}^{(1)}}(x,y)}-\frac{f_{\lambda_k^{(2)}}(x,y)}{f_{\lambda_k^{(1)}}(x,y)}\right|dy\,,\end{aligned}$$

and $|f_{\lambda_1^{(1)}(k+1)}(x,y)-f_{\lambda_1^{(1)}(k)}(x,y)|\le 2M|\lambda_1^{(1)}(k+1)-\lambda_1^{(1)}(k)|$. We shall prove that $\lim_{k\to\infty}|\lambda_1^{(i)}(k+1)-\lambda_1^{(i)}(k)|=0$; it will then follow that $\lim_{k\to\infty}|f_{\lambda_{k+1}^{(2)}}-f_{\lambda_{k+1}^{(1)}}|=0$, and hence (again by compactness) that $|II_{\gamma_{k+1}}(x,A)-II_{\gamma_k}(x,A)|\to 0$.

To that end, recall that $\lambda_j^{(i)}(k) = \frac{d_j^{(i)}(k)}{d_1^{(i)}(k)+d_2^{(i)}(k)}$, $i = 1, 2;\ j = 1, 2$. Therefore,

$$
\begin{aligned}
& |\lambda_1^{(1)}(k+1) - \lambda_1^{(1)}(k)| \\
= \ & |\frac{d_1^{(1)}(k+1)}{d_1^{(1)}(k+1) + d_2^{(1)}(k+1)} - \frac{d_1^{(1)}(k)}{d_1^{(1)}(k) + d_2^{(1)}(k)}| \\
= \ & \left| \frac{d_1^{(1)}(k+1)d_2^{(1)}(k) - d_1^{(1)}(k)d_2^{(1)}(k+1)}{[d_1^{(1)}(k+1) + d_2^{(1)}(k+1)][d_1^{(1)}(k) + d_2^{(1)}(k)]} \right| \\
\leq \ & \left| \frac{(k+1)^{-1}\{[kd_1^{(1)}(k) + (x_{k+1} - x_k)^2]d_2^{(1)}(k) - d_1^{(1)}(k)[kd_2^{(1)}(k) + (x_{k+1} - x_k)^2]\}}{[d_1^{(1)}(k+1) + d_2^{(1)}(k+1)][d_1^{(1)}(k) + d_2^{(1)}(k)]} \right| \\
\leq \ & \left| \frac{(k+1)^{-1}\{[kd_1^{(1)}(k) + (x_{k+1} - x_k)^2]d_2^{(1)}(k) + d_1^{(1)}(k)[kd_2^{(1)}(k) + (x_{k+1} - x_k)^2]\}}{[d_1^{(1)}(k+1) + d_2^{(1)}(k+1)][d_1^{(1)}(k) + d_2^{(1)}(k)]} \right| \\
\leq \ & \frac{R^2}{(k+1)(d_1^{(1)}(k+1) + d_2^{(1)}(k+1))} \ = \ \frac{R^2}{\sum_{i=1}^{k+1}(x_i - x_{i-1})^2} .
\end{aligned}
$$

Now, since $\mathcal{S}$ is compact, there are $\delta, \epsilon > 0$ such that $\mathbf{P}\left[(x_i - x_{i-1})^2 > \epsilon \,|\, \gamma_{i-1}\right] \geq \delta$ for all x_{i-1} and γ_{i-1}. It follows that $\lim_{k\to\infty} \sum_{i=1}^{k+1}(x_i - x_{i-1})^2 = \infty$ with probability 1, hence that $|\lambda_1^{(1)}(k+1) - \lambda_1^{(1)}(k)| \to 0$, and hence that $|II_{\gamma_{k+1}}(x, A) - II_{\gamma_k}(x, A)| \to 0$. Similarly we can prove that $|I_{\gamma_{k+1}}(x, A) - I_{\gamma_k}(x, A)| \to 0$, $|III_{\gamma_{k+1}}(x, A) - III_{\gamma_k}(x, A)| \to 0$, and $|IV_{\gamma_{k+1}}(x, A) - IV_{\gamma_k}(x, A)| \to 0$. Therefore, diminishing adaptation holds. □

Proof of Theorem 4.1. In light of Lemmas 4.1 and 4.2, the result follows immediately from Theorem 5 of Roberts and Rosenthal (2007). □

Proof of Theorem 4.2. Recall that $\mathbb{M}(c_1, c_2)$ is the set of all the $k \times k$ positive definite matrices M such that $c_1\mathbf{I}_k \leq M \leq c_2\mathbf{I}_k$. It follows from the proof of Theorem 1 in Haario et al. (2001) that there are $c_1, c_2 > 0$ such that all the covariances $C = C_i^{(t)}$ are in $\mathbb{M}(c_1, c_2)$.

Since $\mathcal{S}$ is compact, $\inf_{x,y\in\mathcal{S},\ M\in\mathbb{M}(c_1,c_2)} q_M(x, y) > 0$ (where q_M denotes the density function of Gaussian distribution with covariance matrix M). Hence, we have $\inf_{x,y\in\mathcal{S},\gamma\in\mathcal{Y}} q_\gamma(x, y) > 0$. Then following a similar proof to that of Lemma 4.1, one can show that the simultaneous uniform ergodicity condition (10) holds. Similarly to the proof of Lemma 4.2, we can prove that the diminishing adaptation condition (9) holds for Dual RAPT. The result then follows as in the proof of Theorem 4.1. □

Proof of Theorem 4.3. It follows as in the previous proof that $\inf_{x,y\in\mathcal{S},\gamma\in\mathcal{Y}} q_\gamma(x,y) > 0$. Then, similar to Lemma 4.1, it follows that the simultaneous uniform ergodicity condition (10) holds. Diminishing adaptation (9) follows similarly to Lemma 4.2. The result then follows as in the proof of Theorem 4.1. □

5 Examples

5.1 Simulated Examples

We study the performance of the methods proposed using a bimodal target distribution which is a mixture of two Gaussians. By varying the means and variances of the mixture components we try to cover a wider variety of situations. Let us consider the target distribution

$$\pi(x) = 0.5 \times N(\mu_1, \Sigma_1) + 0.5 \times N(\mu_2, \Sigma_2),$$

where μ_i are ten-dimensional vectors and $\Sigma_i = (\sigma_i - \rho_i)\mathbf{I}_{10} + \rho_i \mathbf{1}_{10},\ i = 1, 2$, where $\mathbf{1_d}$ is the $d \times d$ matrix of 1's. The considered scenarios are:

Scenario A: $\rho_1 = 0.2,\ \rho_2 = 0.3,\ \frac{\sigma_1}{\sigma_2} = \frac{1}{3},\ \mu_{1j} = 3,\ \mu_{2j} = -3,\ 1 \le j \le 10.$

Scenario B: $\rho_1 = 0.2,\ \rho_2 = 0.3,\ \frac{\sigma_1}{\sigma_2} = \frac{1}{3},\ \mu_{1j} = 0.5,\ \mu_{2j} = -0.5,\ 1 \le j \le 10.$

Scenario C: $\rho_1 = -0.1,\ \rho_2 = 0.1,\ \frac{\sigma_1}{\sigma_2} = \frac{1}{3},\ \mu_{1j} = 3,\ \mu_{2j} = -3,\ 1 \le j \le 10.$

Scenario D: $\rho_1 = 0.1,\ \rho_2 = -0.1,\ \frac{\sigma_1}{\sigma_2} = \frac{1}{3},\ \mu_{1j} = 3,\ \mu_{2j} = -3,\ 1 \le j \le 10.$

Scenario E: $\rho_1 = -0.1,\ \rho_2 = 0.1,\ \frac{\sigma_1}{\sigma_2} = \frac{1}{3},\ \mu_{1j} = 1,\ \mu_{2j} = -1,\ 1 \le j \le 10.$

Scenario F: $\rho_1 = 0.1,\ \rho_2 = -0.1,\ \frac{\sigma_1}{\sigma_2} = \frac{1}{3},\ \mu_{1j} = 1.5,\ \mu_{2j} = -1.5,\ 1 \le j \le 10.$

It should be noted that scenarios C and D and E and F are different due to the different standard deviations. In our study we chose to implement the HST algorithm (Haario et al., 2001), the Dual RAPT and the Mixed RAPT either for only one chain or, within the paradigm of INCA or TINCA, for five chains in parallel.

The starting value for the $i-th$ chain is $x_{i,0} = (3-i, 3-i, \ldots, 3-i)^T$ for $1 \le i \le 5$ and in the case we implement any of the above algorithms using a single chain, the starting value is $x_0 = (0, \ldots, 0)$. The initial values for the covariance matrices are $\Sigma_1 = \Sigma_2 = \mathbf{I}_{10}$ and $\Sigma_{whole} = 25\mathbf{I}_{10}$. The HST algorithm has initial value $\Sigma = \mathbf{I}_{10}$. The ϵ used in (5) and (7) is set to 0.01. The initialization period contains a total of 10,000 samples which means that in the case of five parallel chains each has an initialization period of 2000 simulations. Throughout the simulation, in the case of Mixed RAPT, we set $\beta = 0.2$. Under all scenarios the partition is defined using $\mathcal{S}_1 = \sum_{i=1}^{10} x_i \le 0$ and $\mathcal{S}_2 = \sum_{i=1}^{10} x_i > 0$. This choice produces a partition that is, in all examples, relatively far from the optimal one.

In order to assess the performance of the algorithm we show the histograms of the first two and last two coordinates, i.e. x_1, x_2, x_9, x_{10}. In a unimodal setting one could compare the covariance of the proposal with the optimal covariance. Unfortunately, when the target is a mixture of unimodal distributions the optimal proposal is not known. One can still compare the number of inter-mode transitions (switches) which is roughly the same as the number of times the chain has crossed from $\mathcal{S}_1$ to $\mathcal{S}_2$ and vice-versa.

Under Scenario A, after 100,000 iterations the mixture parameters of the proposal (6) are $\lambda_1^{(1)} = 0.681$ and $\lambda_1^{(2)} = 0.353$.

The histograms show that a single mixed RAPT chain does a much better job at finding both modes, see Figure 6, compared to a single chain constructed using the simpler dual RAPT algorithm, Figure 5, or the HST algorithm, Figure 7. These results reinforce the intuitive idea that when the modes are far apart neither the HST nor the dual RAPT are efficiently exploring the space. We had similar findings in all scenarios in which the distance between the modes was large, i.e. Scenarios A, C and

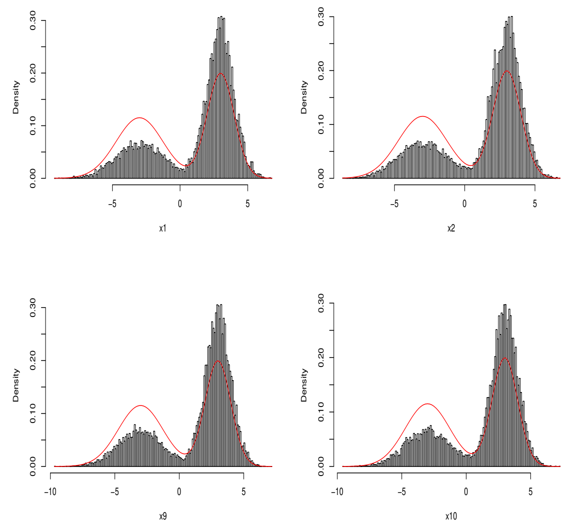

Figure 5: *Scenario A: Histograms of 100,000 samples obtained for* X_1, X_2, X_9, X_{10} *using the dual RAPT algorithm.*

D.

It is important for the initial variances of Q_{whole} to be large enough so that during the initialization period, both modes are visited. For instance, under scenario D running a single mixed RAPT algorithm with starting value $x_0 = (0, \ldots, 0)^T$, $\beta = 0.3$ and $\Sigma_{whole} = diag(10, \ldots, 10)$ the algorithm does not detect both modes even after an initialization period of 10,000 samples. If we use the initial $\Sigma_{whole} = diag(25, \ldots, 25)$ then the performance of mixed RAPT is quite good. In real applications, one does not always have this information and in that case we recommend using INCA or TINCA to reduce the risk of missing regions with high probability under π.

For the same scenario D, we ran five parallel chains, each of them for 20,000 iterations. To test the robustness of INCA we used $\Sigma_{whole} = \mathbf{I}_{10}$. The histograms of the samples corresponding to the first two coordinates and the last two coordinates are as shown in Figure 9.

The results confirm that, although the initial variances are small, the process is mixing well after the initialization period. We also used TINCA with four temperature levels $T = \{1, 2, 4, 8\}$ and once again the algorithm yields the correct samples as can

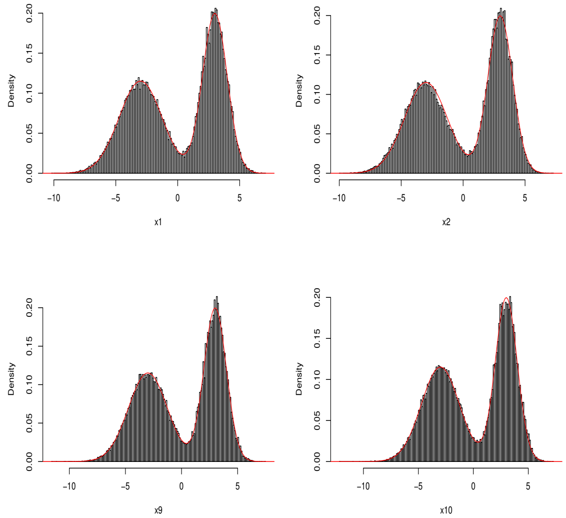

Figure 6: *Scenario A: Histograms of 100,000 samples obtained for* X_1, X_2, X_9, X_{10} *with mixed RAPT.*

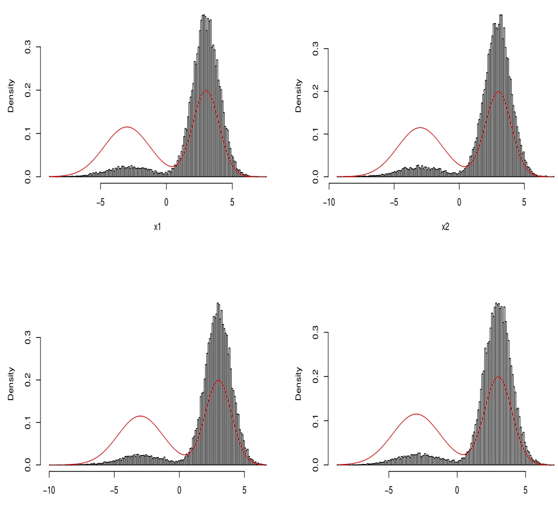

Figure 7: *Scenario A: Histograms of 100,000 samples obtained for* X_1, X_2, X_9, X_{10} *with the HST algorithm.*

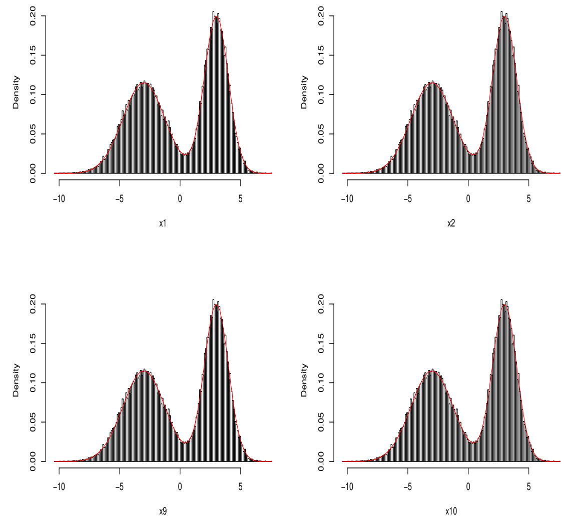

Figure 8: *Scenario D: Histogram of 100,000 samples obtained for* X_1, X_2, X_9, X_{10} *using TINCA with temperatures* $T = \{1, 2, 4, 8\}$ *for five mixed RAPT chains.*

be seen from Figure 8. In the case in which the modes are close, as specified in Scenario B the performance of the HST algorithm is similar to that of mixed RAPT. Our simulations also show that the number of mode switches are comparable for both algorithms. Not surprisingly, the pattern changes when the distance between the modes is increased, as illustrated by Figure 10.

5.2 Real Data Example: Genetic Instability of Esophageal Cancers

Cancer cells undergo a number of genetic changes during neoplastic progression, including loss of entire chromosome sections. We call the loss of a chromosome section containing one allele by abnormal cells by the term “Loss of Heterozygosity” (LOH). When an individual patient has two different alleles, LOH can be detected using laboratory assays. Chromosome regions with high rates of LOH are hypothesized to contain genes which regulate cell behavior so that loss of these regions disables important cellular controls.

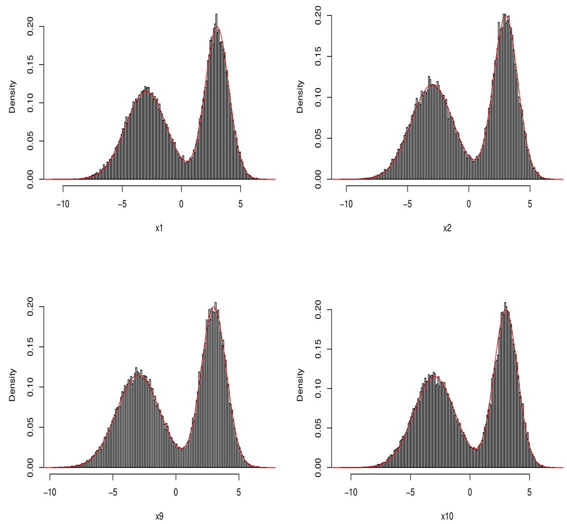

Figure 9: *Scenario D: Histogram of 100,000 samples obtained for* X_1, X_2, X_9, X_{10} *using five parallel mixed RAPT chains.*

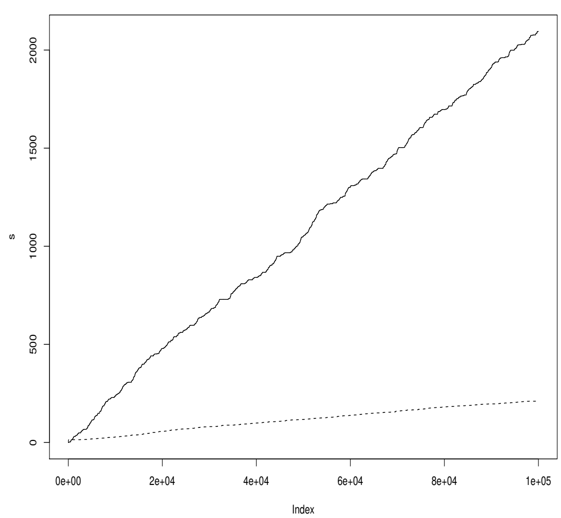

Figure 10: *Scenario E: Number of switches for the HST algorithm (dotted line) and for the mixed RAPT (solid line).*

To locate "Tumor Suppressor Genes"(TSGs), the Seattle Barrett's Esophagus research project (Barrett et al., 1996) has collected LOH rates from esophageal cancers for 40 regions, each on a distinct chromosome arm. A hierarchical mixture model has been constructed by Warnes (2001) in order to determine the probability of LOH for both the "background" and TSG groups. The labeling of the two groups is unknown so we model the LOH frequency using a mixture model, as described by Desai (2000). We obtain the hierarchical Binomial-BetaBinomial mixture model

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

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. Here we parameterize the Beta-Binomial so that γ is a variance parameter defined on the range $-\infty \leq \gamma \leq \infty$. As $\gamma \to -\infty$ the beta-binomial becomes a binomial and as $\gamma \to \infty$ the beta-binomial becomes a uniform distribution on $[0,1]$. This results in the unnormalized posterior density

$$\pi(\eta, \pi_1, \pi_2, \gamma | x) \propto \Pi_{i=1}^{N} f(x_i, n_i | \eta, \pi_1, \pi_2, \omega_2)$$

on the prior range, where

$$\begin{aligned} f(x, n | \eta, \pi_1, \pi_2, \omega_2) &= \eta \begin{pmatrix} n \\ x \end{pmatrix} \pi_1^x (1-\pi_1)^{n-x} + \\ &+ (1-\eta) \begin{pmatrix} n \\ x \end{pmatrix} \frac{\Gamma(\frac{1}{\omega_2})}{\Gamma(\frac{\pi_2}{\omega_2})\Gamma(\frac{1-\pi_2}{\omega_2})} \frac{\Gamma(x+\frac{\pi_2}{\omega_2})}{\Gamma(n-x+\frac{1-\pi_2}{\omega_2})\Gamma(n+\frac{1}{\omega_2})} \end{aligned}$$

Mean in	Region 1	Region 2	Whole space
η	0.897	0.079	0.838
π_1	0.229	0.863	0.275
π_2	0.714	0.237	0.679
γ	15.661	-14.796	13.435

Table 1: *Simulation results for the LOH data.*

and $\omega_2 = \frac{e^\gamma}{2(1+e^\gamma)}$. In order to use the random walk Metropolis we have used the logistic transformation on all the parameters with range $[0,1]$. However, all our conclusions are presented on the original scale for an easier interpretation.

Using the optimization procedures used by Warnes (2001) we determine that the two modes of π are reasonably well separated by the partition made of $S_1 = \{(\eta, \pi_1, \pi_2, \gamma) \in [0,1] \times [0,1] \times [0,1] \times [-30, 30] | \pi_2 \geq \pi_1\}$ and $S_2 = \{(\eta, \pi_1, \pi_2, \gamma) \in [0,1] \times [0,1] \times [0,1] \times [-30, 30] | \pi_2 \leq \pi_1\}$.

5.2.1 Simulation results

We have run five parallel mixed RAPT algorithms to simulate from π using the partition $S_1 \cup S_2$. The initialization period contained 5,000 iterations for each chain. The covariance matrices were initialized as $\Sigma_1 = \Sigma_2 = 0.1\mathbf{I}_4$ and $\Sigma_{whole} = 20\mathbf{I}_4$. After $50,000$ iterations from each chain, we obtain $\lambda_1^{(1)} = 0.923$ and $\lambda_1^{(2)} = 0.412$. The estimates for the parameters of interest are shown in Table 5.2.1.

Figure 11 gives a two dimensional scatterplot of the (π_1, π_2) samples. This is similar to the findings of Warnes (2001) (Figure 8). To illustrate the exchange of information between the parallel the chains, we use the BGR diagnostic statistic, R. When the BGR R statistics is close to to 1, we can assume all chains have the same information regarding π. For this example, after 20,000 iterations the BGR's R statistics stabilizes below 1.1 as one can see in Figure 12.

To compare the performance of the mixed RAPT with and without INCA we

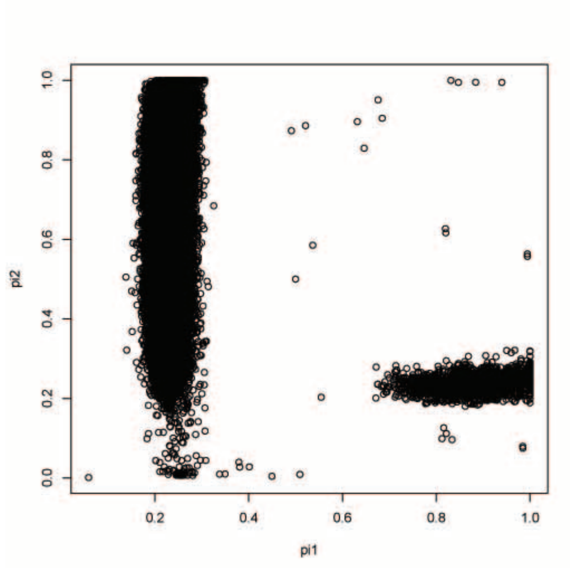

Figure 11: *Scatterplot of the 250,000 samples for* (π_1, π_2).

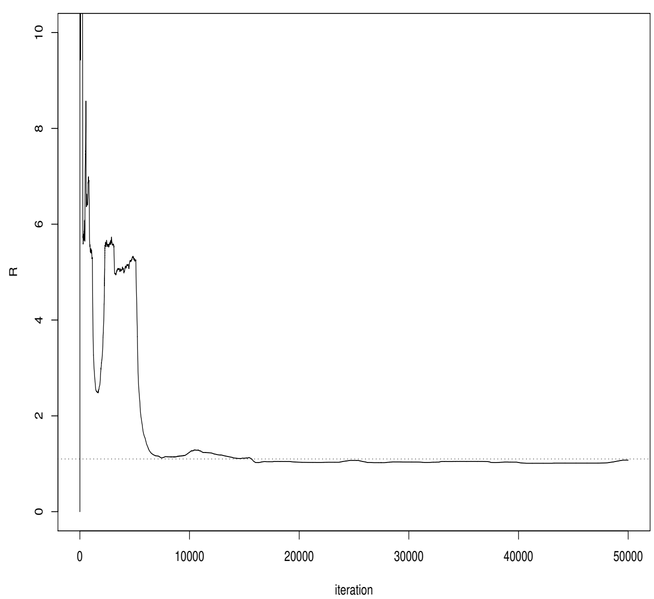

Figure 12: *LOH Data Example: The evolution of BGR's R statistics for 5 mixed RAPT chain; the dotted line represents the threshold 1.1.*

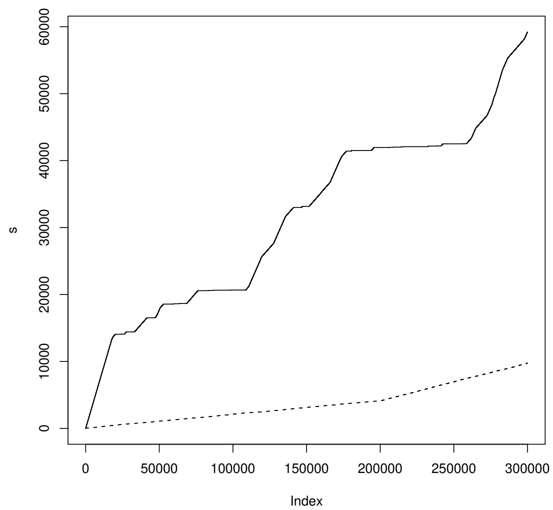

Figure 13: *The total number of switches times for the five parallel Mixed RAPT chains (run for 60,000 iterations each) vs the number of switch times of a single Mixed RAPT (run for 300,000 iterations).*

monitor the number of switches between $\mathcal{S}_1$ and $\mathcal{S}_2$. We run a single Mixed RAPT algorithm for 300,000 iterations, and independently five parallel Mixed RAPT algorithms for 60,000 iterations each. In Figure 13 we plot the total number of switches for the five parallel processes up to time t and the switch time for the single run up to time $5t$ for a fair comparison. One can see that the Mixed RAPT performs better together with INCA than by itself.

6 Conclusions and Further Work

This work is concerned with the practical aspects of adaptive MCMC, particularly related to sampling from multimodal distributions. The aim for most of our theoretical results is the adaptive random walk Metropolis since it is one of the most used algorithms in practice. The inter-chain adaptation strategy is widely applicable and could be used for a large number of adaptive MCMC algorithms with significant potential gains. The regional adaptation algorithm proposed here has been discussed in the context of two separate regions. Evidently, the construction can be generalized but one has to keep in mind that besides good sampling properties within each region the sampler should be also required to visit all regions often enough. In the case of many regions this could present complications.

Acknowledgment

The authors acknowledge the support of the Natural Sciences and Engineering Research Council of Canada. We thank Gregory Warnes for discussion and for generously providing the data set used in section 5, and Xiao-Li Meng and Eric Moulines for helpful discussions and suggestions. We are grateful to two referees, one Associate Editor and the Editor for a set of thorough and very helpful reviews.

References

Andrieu, C., and Atchade, Y. F. (2006), On the efficiency of adaptive MCMC algorithms,, in *Proceedings of the 1st International conference on Performance evaluation methodologies and tools*, Vol. ACM International Conference Proceeding Series; Vol 180.

Andrieu, C., and Moulines, E. (2006), "On the ergodicity properties of some adaptive MCMC algorithms," *Annals of Applied Probability*, 16, 1462–1505.

Andrieu, C., Moulines, E., and Priouret, P. (2005), "Stability of stochastic approximation under verifiable conditions," *Siam Journal On Control and Optimization*, 44, 283–312.

Andrieu, C., and Robert, C. P. (2001), Controlled MCMC for optimal sampling,, Technical report, Université Paris Dauphine.

Andrieu, C., and Thoms, J. (2008), "A tutorial on adaptive MCMC," *Statist. Comput.*, 18, 343–373.

Applegate, D., Kannan, R., and Polson, N. (1990), Random Polynomial time algorithms for sampling from joint distributions,, Technical Report 500, Carnegie-Mellon University.

Atchade, Y., and Rosenthal, J. (2005), "On adaptive Markov chain Monte Carlo algorithms," *Bernoulli*, 11, 815–828.

Barrett, M., Galipeau, P., Sanchez, C., Emond, M., and Reid, B. (1996), "Determination of the frequency of loss of heterozygosity in esophageal adeno-carcinoma nu cell sorting, whole genome amplification and microsatellite polymorphisms," *Oncogene*, 12(1873-1878).

Brockwell, A., and Kadane, J. (2005), "Identification of regeneration times in MCMC simulation, with application to adaptive schemes," *Journal of Computational and Graphical Statistics*, 14, 436–458.

Brooks, S. P., and Gelman, A. (1998), "General methods for monitoring convergence of iterative simulations," *J. Comput. Graph. Statist.*, 7(4), 434–455.

Cappé, O., and Moulines, E. (2009), "Online EM algorithm for latent data models," *J. Roy. Statist. Soc. Ser. B*, p. To appear.

Craiu, R. V., and Di Narzo, A. (2009), A mixture-based approach to regional adaptation for MCMC,, Technical report, University of Toronto.

Craiu, R. V., and Lemieux, C. (2007), "Acceleration of the Multiple-try Metropolis Algorithm using Antithetic and Stratified sampling," *Statistics and Computing*, 17(2), 109–120.

Desai, M. (2000), Mixture Models for Genetic changes in cancer cells, PhD thesis, University of Washington.

Eidsvik, J., and Tjelmeland, H. (2006), "On directional Metropolis-Hastings algorithms," *Statistics and Computing*, 16, 93–106.

Gasemyr, J. (2003), “On an adaptive version of the Metropolis-Hastings algorithm with independent proposal distribution,” *Scand. J. Statist.*, 30, 159–173.

Gelman, A., Roberts, G. O., and Gilks, W. R. (1996), “Efficient Metropolis jumping rules,” in *Bayesian statistics, 5 (Alicante, 1994)*, Oxford Sci. Publ., New York: Oxford Univ. Press, pp. 599–607.

Gelman, A., and Rubin, D. B. (1992), “Inference from iterative simulation using multiple sequences (with discussion),” *Statist. Sci.*, pp. 457–511.

Geyer, C. J., and Thompson, E. A. (1994), Annealing Markov chain Monte Carlo with applications to ancestral inference,, Technical Report 589, University of Minnesota.

Gilks, W., Roberts, G., and Suhu, S. (1998), “Adaptive Markov chain Monte Carlo through regeneration,” *Journal of the American Statistical Association*, 93, 1045–1054.

Gill, J. (2008), “Is Partial-Dimension Convergence a problem for Inferences from MCMC Algorithms?,” *Political Analysis*, 16, 153–178.

Giordani, P., and Kohn, R. (2006), Adaptive independent Metropolis-Hastings by fast estimation of mixtures of normals,. Preprint.

Green, P., and Mira, A. (2001), “Delayed rejection in reversible jump Metropolis-Hastings,” *Biometrika*, 88, 1035–1053.

Haario, H., Saksman, E., and Tamminen, J. (2001), “An adaptive Metropolis algorithm,” *Bernoulli*, 7, 223–242.

Haario, H., Saksman, E., and Tamminen, J. (2005), “Componentwise adaptation for high dimensional MCMC,” *Computational Statistics*, 20, 265–273.

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

Jennison, C. (1993), “Discussion of ”Bayesian computation via the Gibbs sampler and Related Markov chain Monte Carlo Methods,” by Smith and Roberts,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 55, 54–56.

Liu, J., Liang, F., and Wong, W. (2000), “The multiple-try method and local optimization in Metropolis sampling,” *Journal of the American Statistical Association*, 95, 121–134.

Meng, X. L. (2007), Parallel evolution in MCMC,. Personal communication.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953), “Equations of state calculations by fast computing machines,” *J. Chem. Ph.*, 21, 1087–1092.

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

Neal, R. M. (1994), Sampling from multimodal distributions using tempered transitions,, Technical Report 9421, University of Toronto.

Neal, R. M. (2001), “Annealed importance sampling,” *Stat. Comput.*, 11(2), 125–139.

Nott, D., and Kohn, R. (2005), “Adaptive sampling for Bayesian variable selection,” *Biometrika*, 92, 747–763.

Roberts, G. O., Gelman, A., and Wilks, W. (1997), “Weak convergence and optimal scaling of random walk Metropolis algorithms,” *Ann. Appl. Probab.*, 7, 110–120.

Roberts, G. O., and Rosenthal, J. S. (2001), “Optimal scaling for various Metropolis-Hastings algorithms,” *Statist. Sci.*, 16(4), 351–367.

Roberts, G. O., and Rosenthal, J. S. (2007), “Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms,” *J. Appl. Probab.*, 44(2), 458–475.

Roberts, G. O., and Rosenthal, J. S. (2009), “Examples of adaptive MCMC,” *Journal of Computational and Graphical Statistics*, p. To appear.

Rosenthal, J. S. (2004), Adaptive MCMC Java Applet,. On the web at:.
URL: *http://probability.ca/jeff/java/adapt.html*

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, pp. 447–454.

Sminchisescu, C., and Triggs, B. (2002), Hyperdynamics importance sampling,, in *European Conference on Computer vision*, Vol. 1, Copenhagen, pp. 769–783.

Sminchisescu, C., Welling, M., and Hinton, G. (2003), A mode-hopping MCMC sampler,, Technical report, University of Toronto.

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.

Yang, C., Craiu, R. V., and Rosenthal, J. S. (2009), The ergodicity of mixed regional adaptive MCMC,. In progress.