DRAM: Efficient adaptive MCMC

Heikki Haario,[†] Marko Laine,[†] Antonietta Mira,[*] Eero Saksman,[††]

[†] University of Helsinki, Helsinki, Finland

[*] University of Insubria, Varese, Italy

[††] University of Jyväskylä, Jyväskylä, Finland

Abstract

We propose to combine two quite powerful ideas that have recently appeared in the Markov chain Monte Carlo literature: adaptive Metropolis samplers and delaying rejection. The ergodicity of the approach is proved, and the efficiency of the combination is demonstrated with various test examples. We present situations where the combination outperforms the original methods: the adaptation clearly enhances the efficiency of the delayed rejection algorithm in cases where good candidates for the proposal distributions are not available. Similarly, the delayed rejection provides a systematic remedy for cases where the adaptation has difficulties to get started.

Keywords: Adaptive Markov chain Monte Carlo, Delaying rejection, Efficiency ordering, Adaptive Metropolis-Hastings

1 Introduction and motivation

Markov chain Monte Carlo (MCMC) methods allow to estimate $E_\pi f$, the expectation of a function f with respect to a distribution π, possibly known up to a normalizing constant. A Markov chain that has π as it unique stationary and limiting distribution is constructed and simulated. The mean of f along a realized path of the chain, $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$, is the MCMC estimator. Typically the mean is computed after a burn-in to allow the chain to reach its stationary regime. Under mild regularity condition [12] the MCMC sampler is asymptotically unbiased and normally distributed.

In this paper we propose various strategies to combine two quite powerful ideas that have recently appeared in the MCMC literature: adaptive Metropolis samplers [5, 6] and delaying rejection [14, 4, 9].

Delaying rejection (DR) is a way of modifying the standard Metropolis-Hastings algorithm (MH) [12] to improve efficiency of the resulting MCMC estimates relative to Peskun [10, 13] asymptotic variance ordering. The basic idea is that, upon rejection in a MH, instead of advancing time and retaining the current position, a second stage move is proposed. The acceptance probability of the second stage candidate is computed so that reversibility of the Markov chain relative to the distribution of interest is preserved. The process of delaying rejection can be iterated for a fixed or random number of stages. The higher stage proposals are allowed to depend on the candidates so far proposed and rejected. Thus DR allows partial adaptation of the proposal within each step of the Markov chain.

The DR can be also considered as a way of combining different proposals for MH or different kernels for MCMC. There are other strategies suggested in the MCMC literature to combine kernels all having the proper stationary distribution, namely mixing and cycling [14]. The advantage of DR over these alternatives is that a hierarchy between kernels can be exploited so that kernels that are easier to compute (in terms of CPU time) are tried first for example, thus saving in terms of simulation time. Or moves that are more "bold" (bigger variance of the proposal for example) are tried at earlier stages thus allowing the sampler to explore the state space more efficiently. Similarly, again to allow better exploration of the stage space, global moves (i.e. updating all coordinates at once) and be tried first and local moves (updating single or groups of coordinates) can be attempted later.

The global adaptive strategy we will combine with the local adaptive strategy provided by the DR, is the Adaptive Metropolis (AM) algorithm [5, 6]. The intuition behind the AM is that, on-line tuning the proposal distribution in a MH can be based on the past sample path of the sampled chain. Due to this form of adaptation sampler is neither Markovian nor reversible. In [6] the Authors prove, from first principles, that, under some regularity conditions on the way adaptation is performed, the AM retains the stationary distribution desired.

In Sections 2 and 3 we give the details of the DR and of the AM strategies respectively.

We then propose different ways of combining DR with AM (Section 4) and prove the ergodicity of the resulting algorithms (Section 5).

In Section 6, various test examples will be used to compare the proposed strategies in terms of their efficiency measured both by the asymptotic variance of the resulting MCMC estimators and in terms of CPU simulation time.

2 Delaying rejection

In this section we give the details of DR. Suppose the current position of the Markov chain is $X_t = x$. As in a regular MH a candidate move Y_1 is generated from a proposal $q_1(x, \cdot)$ and accepted with the usual probability

$$\begin{aligned} \alpha_1(x, y_1) &= 1 \wedge \frac{\pi(y_1)q_1(y_1, x)}{\pi(x)q_1(x, y_1)} \\ &= 1 \wedge \frac{N_1}{D_1}. \end{aligned} \tag{1}$$

Upon rejection, instead of retaining the same position, $X_{t+1} = x$, as we would do in a standard MH, a second stage move Y_2 is proposed. The second stage proposal is allowed to depend not only on the current position of the chain but also on what we have just proposed and rejected: $q_2(x, y_1, \cdot)$. The second stage proposal is accepted with probability

$$\begin{aligned}\alpha_2(x, y_1, y_2) &= 1 \wedge \frac{\pi(y_2)q_1(y_2, y_1)q_2(y_2, y_1, x)[1 - \alpha_1(y_2, y_1)]}{\pi(x)q_1(x, y_1)q_2(x, y_1, y_2)[1 - \alpha_1(x, y_1)]} \\ &= 1 \wedge \frac{N_2}{D_2}.\end{aligned} \tag{2}$$

This process of delaying rejection can be iterated and the i-th stage acceptance probability is, following [8],:

$$\begin{aligned}\alpha_i(x, y_1, \cdots, y_i) = 1 \wedge \Big\{ &\frac{\pi(y_i)q_1(y_i, y_{i-1})q_2(y_i, y_{i-1}, y_{i-2}) \cdots q_i(y_i, y_{i-1}, \cdots, x)}{\pi(x)q_1(x, y_1)q_2(x, y_1, y_2) \cdots q_i(x, y_1, \cdots, y_i)} \\ &\frac{[1 - \alpha_1(y_i, y_{i-1})][1 - \alpha_2(y_i, y_{i-1}, y_{i-2})] \cdots [1 - \alpha_{i-1}(y_i, \cdots, y_1)]}{[1 - \alpha_1(x, y_1)][1 - \alpha_2(x, y_1, y_2)] \cdots [1 - \alpha_{i-1}(x, y_1, \cdots, y_{i-1})]} \Big\} \\ = 1 \wedge &\frac{N_i}{D_i}.\end{aligned} \tag{3}$$

If i-th stage is reached, it means that $N_j < D_j$ for $j = 1, \cdots, i-1$, therefore $\alpha_j(x, y_1, \cdots, y_j)$ can be rewritten as $N_j/D_j,\ j = 1, \cdots, i-1$ and we obtain the recursive formula

$$D_i = q_i(x, \cdots, y_i)(D_{i-1} - N_{i-1})$$

which leads to

$$\begin{aligned}D_i = q_i(x, \cdots, y_i)[&q_{i-1}(x, \cdots, y_{i-1})[q_{i-2}(x, \cdots, y_{i-2}) \cdots \\ &[q_2(x, y_1, y_2)[q_1(x, y_1)\pi(x) - N_1] - N_2] - N_3] \cdots - N_{i-1}].\end{aligned} \tag{4}$$

Since all acceptance probabilities are computed so that reversibility with respect to π is preserved separately at each stage, the process of delaying rejection can be interrupted at any stage that is, we can, in advance, decide to try at most, say, 3 times to move away from the current position, otherwise we let the chain stay where it is. Alternatively, upon each rejection, we can toss a p-coin (i.e. a coin with head probability equal to p), and if the outcome is head we move to a higher stage proposal, otherwise we stay put.

In [14] the DR strategy is proved to outperform the standard MH in the Peskun absolute efficiency ordering. This means that, using the DR, we obtain MCMC estimators that have a smaller asymptotic variance for every function f whose expectation relative to π we want to estimate (provided f has finite variance under π).

3 Adaptive MCMC

In this section we briefly introduce the AM strategy, for more details and theory see [5],[6]. The basic idea is to create a Gaussian proposal distribution from the points of the MCMC chain. This achieved by computing the covariance matrix of the chain. The crucial point regarding the AM adaptation is how the covariance of the proposal distribution depends on the history of the chain. We take, possibly after an initial non-adaptation period, the proposal to be centered at the current position of the Markov chain, X_t, and set the covariance to be: $C_t = s_d \mathrm{Cov}(X_0, \ldots, X_{t-1}) + s_d \varepsilon I_d$, where s_d is a parameter that depends only on the dimension d of the state space where π is defined and $\varepsilon > 0$ is a constant that we may choose very small compared to the size of S. Here I_d denotes the d-dimensional identity matrix. In order to start the adaptation procedure an arbitrary strictly positive definite initial covariance, C_0, is chosen according to a priori knowledge (which may be quite poor). A time index, $t_0 > 0$, defines the length the initial non-adaptation period and we let

$$C_t = \begin{cases} C_0, & t \leq t_0 \\ s_d \mathrm{Cov}(X_0, \ldots, X_{t-1}) + s_d \varepsilon I_d, & t > t_0. \end{cases} \tag{5}$$

Recall the definition of the empirical covariance matrix determined by points $X_0, \ldots, X_k \in \mathbb{R}^d$:

$$\mathrm{Cov}(X_0, \ldots, X_k) = \frac{1}{k} \left(\sum_{i=0}^{k} X_i X_i^T \; - (k+1) \overline{X}_k \, \overline{X}_k^T \right), \tag{6}$$

where $\overline{X}_k = \frac{1}{k+1} \sum_{i=0}^{k} X_i$ and the elements $X_i \in \mathbb{R}^d$ are considered as column vectors. Substituting (6) in definition (5) for $t \geq t_0 + 1$ the covariance C_t satisfies the recursive formula:

$$C_{t+1} = \frac{t-1}{t} C_t + \frac{s_d}{t} \left(t \overline{X}_{t-1} \overline{X}_{t-1}^T - (t+1) \overline{X}_t \overline{X}_t^T + X_t X_t^T + \varepsilon I_d \right). \tag{7}$$

This permits the calculation of C_t without excessive computational cost since the mean, $\overline{X}_t$, also satisfies an obvious recursive formula.

This adaptation was proved to be ergodic in [6]. In numerical applications, some helpful observations have emerged. The choice for the length of the initial non-adaptive portion of the simulation, $t_0 > 0$ is free, but the bigger it is, the longer it takes for the effect of adaptation to take place. In the earlier, non–ergodic version of the algorithm ([5]) it was found that the adaptation should not be done at each time step, but only at given time intervals. This way of adaptation has shown to improve the mixing properties even with AM. So the index t_0, in fact, can be used define the length of non–adaptation during the whole chain. The role of the parameter ε is just to ensure that, theoretically, C_t will not become singular, but in practice it can be safely set to zero. Following [3], we take the scaling parameter to be $s_d = (2.4)^2/d$. In [3] the Authors show that, in a certain sense, this choice optimizes the mixing properties of the Metropolis search in the case of Gaussian targets and Gaussian proposals.

4 Combining DR and AM

The success of MCMC methods, in general, depends on how well the proposal distribution fits the target distribution. In its basic formulation, DR employs a given number of fixed proposals that are used at the different stages. Therefore, the success of the DR strategy depends largely on the fact that at least one of the proposals is successfully chosen. The intuition behind adaptive strategies is to learn from the information obtained during the run of the chain, and to tune the proposals to work more efficiently. There are, in principle, numerous ways of combining AM or MH within the DR framework. One could use AM only at the first stage and employ fixed MH proposals at higher stages of the delaying rejection process. The choice of the fixed MH proposals could be based on separate pilot runs. Or one might adapt the proposals at the different stages separately, with the aim of attempting 'global' moves at the first stage (update all coordinates at once) and 'local' moves at higher stages (update single coordinates of groups of them). As an alternative, at different stages of the delaying rejection different values of t_0 and s_d could be used.

We shall follow here a rather direct way of combining AM adaptation and DR. The proposal of the first stage of DR is adapted just as in AM: the covariance for AM is computed from the points of the sampled chain, no matter at which stage these points in the sample path have been accepted. The proposal for higher stages are always computed simply as scaled versions of the proposal of the first stage. The scale factor can be freely chosen: the proposals of the higher stages can have a smaller or larger variance than the proposal at earlier stages. The simulation results in [4] suggest that, it is more beneficial, in terms of asymptotic variance reduction of the resulting MCMC estimators, to have larger variance at earlier stages and then reduce the variance upon rejection.

From the DR strategy point of view, the rational of the approach is to adapt, via AM, the first stage proposal to better fit the target distribution. If the variance of the first stage proposal is too large or small, the points obtained from the higher stages will transform the variance in the right direction.

From the AM point of view, clear benefits are expected, too. It sometimes may be difficult to get the AM adaptation started. This happens if the initial guess for the proposal distribution is far from a correct one. This occurs, e.g., if the variance of the proposal is too large, or the covariance for the proposal is nearly singular. Now the DR framework provides a natural remedy for these situations: by scaling down the size of the proposals at higher DR stages we ensure that some points will be accepted. Once this happens, the above AM adaptation usually starts working properly.

Below, we shall present the discussed merits of the DRAM combination in light of concrete examples.

5 Ergodicity of DR+AM

In order to approach properties of the simulation provided by the non-Markovian DRAM algorithm we first fix some notation and define the stochastic process corresponding to the algorithm. We follow mainly the approach and notation of [6], to which we refer for unexplained concepts.

In this section we focus on two-stages DR algorithms but the theory can be generalized at DR strategies with more than two attempts to move.

To start with, denote by $q_C(x, y)$ the density of a Gaussian proposal with covariance matrix C. Thus

$$q_C(x,y) = \frac{1}{(2\pi)^{n/2}\sqrt{|C|}} e^{-(x-y)^{\mathrm{T}}C^{-1}(x-y)/2}. \tag{8}$$

We shall assume that $D \subset \mathbb{R}^d$ is a Borel-measurable subset of the Euclidean space, and the target $\pi : D \to [0,\infty)$ is a probability density on D (actually, we shall also denote by π the associated measure). As explained in Section 2, given two proposals one may always define a corresponding delayed rejection transition probability function (DR-t.p.f.). We formalize this into a definition

Definition 1 *Let π and D be as above and fix C_1, C_2 be given covariance matrices. The corresponding two-stages DR-t.p.f. is denoted by Q_{C_1,C_2}.*

In order to give an explicit formula for Q_{C_1,C_2} we write (compare with Section 2)

$$\alpha_1(x,y) = 1 \wedge \frac{\pi(y)}{\pi(x)}, \tag{9}$$

where we understand that $\pi(x) = 0$ for $x \notin D$ and α_1 takes the value 1 if both $\pi(x) = \pi(y) = 0$. Moreover,

$$\alpha_2(x,y',y) = 1 \wedge \frac{\pi(y)q_{C_1}(y,y')(1-\alpha_1(y,y'))}{\pi(x)q_{C_1}(x,y')(1-\alpha_1(x,y'))}. \tag{10}$$

Comparing the above formula with (2) one should notice the cancellation of the second stage proposals. We are now able to define for any Borel-measurable subset $A \subset D$ such that $x \notin A$

$$Q_{C_1,C_2}(x;A) = \int_A q_{C_1}(x,y')\alpha_1(x,y')dy' \tag{11}$$

$$+ \quad \int_A \left(\int_{\mathbb{R}^n} q_{C_1}(y,y')(1-\alpha_1(y,y'))q_{C_2}(x,y)\alpha_2(x,y',y)\, dy' \right) dy.$$

The definition of the t.p.f. is completed by setting

$$Q_{C_1,C_2}(x;\{x\}) = 1 - Q_{C_1,C_2}(x; D \setminus \{x\}). \tag{12}$$

For later need we estimate quantitatively the dependence of Q_{C_1,C_2} on the covariances. The following technical lemmata serve this purpose. The derivative D^k_{ij} in the first lemma are taken with respect to the (i,j):th element $(i,j = 1,\ldots,d)$ of the covariance matrix C_k $(k = 1,2)$. The easy proof of the first lemma is left to the reader.

Lemma 1 *Let $D \subset \mathbb{R}^d$ be bounded. Assume that the covariances C_1, C_2 satisfy the matrix inequality*

$$a_1 I_d \le C_1, C_2 \le a_2 I_d, \tag{13}$$

where $0 < a_1 < a_2 < \infty$ are constants and I_d is the d-dimensional identity matrix. Then there are finite constants a_3, a_4 that depend only on D, a_1, a_2 such that the inequalities

$$\frac{|D^k_{ij} q_{C_k}(x,y')|}{q_{C_k}(x,y')} \leq a_3(1+|y'|^2) \tag{14}$$

and

$$\frac{|D^1_{ij}\alpha_2(x,y',y)|}{\alpha_2(x,y',y)} \leq a_4(1+|y'|^2) \tag{15}$$

hold for all $y' \in \mathbb{R}^n$ and $x, y \in D$. Here $1 \leq k \leq 2$ and $1 \leq i, j \leq d$ are arbitrary.

(We should remark here that α_2 is not necessarily differentiable in the strict sense, but (15) should be interpreted as an estimate for the local Lipschitz constant).

Lemma 2 *Let $D \subset \mathbb{R}^d$ be bounded and assume that all the covariances C_1, C_1', C_2, C_2' satisfy the matrix inequality (13). Then there is a constant a_5 such that*

$$|Q_{C_1,C_2}(x,A) - Q_{C_1',C_2'}(x,A)| \leq a_5(\|C_1 - C_1'\| + \|C_2 - C_2'\|) \tag{16}$$

for all $x \in D$ and measurable $A \subset D$.

Proof. In order to prove (16) we first consider the case $C_2 = C_2'$. By (12) we may also assume that $x \notin A$. We obtain by (11) that

$$|Q_{C_1,C_2}(x,A) - Q_{C_1',C_2}(x,A)| \leq \int_0^1 |\frac{d}{ds}h_1(s)|ds + \int_0^1 |\frac{d}{ds}h_2(s)|ds,$$

where

$$h_1(s) = \int_A q_{C(s)}(x,y)\alpha_1(x,y)dy$$

with $C(s) = sC_1' + (1-s)C_1 = C_1 + s(C_1' - C_1)$, and

$$h_2(s) = \int_A \left(\int_{\mathbb{R}^n} q_{C(s)}(y,y')(1-\alpha_1(y,y'))q_{C_2}(x,y)\alpha_2(x,y',y)\ dy' \right) dy.$$

Observe that α_2 depends on $C(s)$. The matrix $C(s)$ clearly satisfies the inequalities (13) for all $s \in [0,1]$. Hence the previous lemma applies and we obtain the estimate

$$|\frac{d}{ds}h_1(s)| \leq a_6 a_3 \|C_1 - C_1'\| \sup_{y \in D}(1+|y|^2)h_1(s) \leq a_7\|C_1 - C_1'\|$$

since $h_1 \leq 1$ and D is bounded.

Similarly we compute

$$\begin{aligned} |\frac{d}{ds}h_2(s)| \quad &\leq \quad a_6\|C_1 - C_1'\| \int_A \Bigg(\int_{\mathbb{R}^n} (a_3+a_4)(1+|y'|^2)q_{C(s)}(y,y') \\ &\qquad\qquad (1-\alpha_1(y,y'))q_{C_2}(x,y)\alpha_2(x,y',y)\ dy') \,dy \\ &\leq \quad a_8\|C_1 - C_1'\| \int_{\mathbb{R}^n} (1+|y'|^2)q_{C(s)}(y,y')dy' \int_A q_{C_2}(x,y)dy \\ &\leq \quad a_9\|C_1 - C_1'\| \sup_{y \in D} \int_{\mathbb{R}^n} (1+|y'|^2)q_{C(s)}(y,y')dy' \leq a_{10}\|C_1 - C_1'\|. \end{aligned}$$

Above the last written supremum is clearly uniformly bounded as $C(s)$ satisfies (13).

By combining the obtained estimates the claim follows in the case $C_2 = C_2'$. The case $C_1 = C_1'$ is similar, although easier since α_2 does not depend on C_2. By combining the two cases the general statement is proved. □

The sequence (K_n) of generalized transition probability functions defining the DRAM algorithm (with second covariance proportional to the first one) is obviously given by

$$K_n(x_0, \dots, x_{n-1}; A) = Q_{C_n, \gamma C_n}, \tag{17}$$

where C_n is the covariance obtained from the history of the algorithm, defined by the formula (5) in Section 3. The constant $\gamma > 0$ is fixed (usually one chooses $\gamma \in (0,1)$, see Section 4). Our proof for the exactness of the simulation provided by DRAM is based on Theorem 2 in [6]. For readers convenience we recall this result here (Theorem 3 below). In order to do this we need to define a "freezed" transition probability. Given a generalized transition probability K_n (where $n \geq 2$) and a *fixed* $(n-1)$-tuple, $(y_0, y_1, \dots y_{n-2}) \in S^{n-1}$, we denote $\widetilde{y}_{n-2} = (y_0, y_1, \dots y_{n-2})$ and define the transition probability $K_{n,\widetilde{y}_{n-2}}$ by

$$K_{n,\widetilde{y}_{n-2}}(x; A) = K_n(y_0, y_1, \dots y_{n-2}, x; A) \tag{18}$$

for $x \in D$ and $A \subset D$. For the definition of the (Dobrushin) coefficient of ergodicity, $\delta(K)$, we refer to [6] (p. 228).

Theorem 3 *Assume that* (K_n) *satisfies the following three conditions* (i) – (iii) :

(i) *There is an integer* k_0 *and a constant* $\lambda \in (0,1)$ *such that*

$$\delta((K_{n,\widetilde{y}_{n-2}})^{k_0}) \leq \lambda < 1 \quad \text{for all } \ \widetilde{y}_{n-2} \in S^{n-1} \ \text{ and } \ n \geq 2.$$

(ii) *There is a probability measure* π *on* S *and a constant* $c_0 > 0$ *so that*

$$\|\pi K_{n,\widetilde{y}_{n-2}} - \pi\| \leq \frac{c_0}{n} \quad \text{for all } \quad \widetilde{y}_{n-2} \in S^{n-1} \ \text{ and } \ n \geq 2.$$

(iii) *The estimate for the operator norm*

$$\|K_{n,\widetilde{y}_{n-2}} - K_{n+k,\widetilde{y}_{n+k-2}}\|_{\mathcal{M}(D)\to\mathcal{M}(D)} \leq c_1 \frac{k}{n},$$

holds, where c_1 *is a positive constant,* $n, k \geq 1$ *and one assumes that the* $(n+k-1)$*-tuple* $\widetilde{y}_{n+k-2}$ *is a direct continuation of the* $(n-1)$*-tuple* $\widetilde{y}_{n-2}$.

Then, if $f : D \to \mathbb{R}$ *is bounded and measurable, it holds almost surely that*

$$\lim_{n\to\infty} \frac{1}{n+1}(f(X_0) + f(X_1) + \dots + f(X_n)) = \int_S f(x)\pi(dx). \tag{19}$$

We are ready to verify that the DRAM algorithm yields unbiased simulation of the target distribution.

Theorem 4 *Let π be the density of a target distribution supported on a bounded measurable subset $D \subset \mathbb{R}^d$ and assume that π is bounded from above. Then the DRAM algorithm, as described in Section* 4 (*see also* (17)) *is ergodic in the sense of* (19).

Proof. We are to show that the transition probabilities (17) fulfill the conditions (i)–(iii) of Theorem 3. Observe first that by (5) and boundedness of D the covariances C_n satisfy a uniform estimate (13) with constants depending only on D, d and ε. Hence the corresponding densities $q_{C_n}(x, y)$ are uniformly bounded from below for $x, y \in D$ and the first term in the formula (11) easily yields the estimate

$$K_{n,\widetilde{y}_{n-2}}(x\ A) \geq a_3 \pi(A)$$

since π is bounded from above. This is well known to yield condition (i) with $k_0 = 1$ (compare [6, p. 230]).

In order to check condition (ii) we fix $\widetilde{y}_{n-2} \in D^{n-1}$ and denote $C^* = C_{n-1}(y_0, \dots y_{n-2})$. By the very definitions (5)–(6) it follows that

$$\|C^* - C_n(y_0, \dots y_{n-2}, y)\| \leq a_{10}/n, \tag{20}$$

where a_{10} does not depend on $y \in S$. We may hence apply Lemma 2 to to deduce for all measurable $A \subset D$ that $|K_{n,\widetilde{y}_{n-2}}(y; A) - Q_{C^*,\gamma C^*}(y; A)| \leq a_{11}/n$, which in turn implies that $\|K_{n,\widetilde{y}_{n-2}} - Q_{C^*,\gamma C^*}\|_{\mathcal{M}(D)\to\mathcal{M}(D)} \leq 2a_{11}/n$. By [14] the delayed rejection kernel satisfies $\pi Q_{C^*,\gamma C^*} = \pi$, and we obtain

$$\|\pi - \pi K_{n,\widetilde{y}_{n-2}}\| = \|\pi(M_{C^*} - K_{n,\widetilde{y}_{n-2}})\| \leq \frac{2a_{11}}{n},$$

as desired.

Finally, the verification of condition (iii) is based on Lemma 2, which yields that

$$\begin{aligned} \|K_{n,\widetilde{y}_{n-2}} - K_{n+k,\widetilde{y}_{n+k-2}}\|_{\mathcal{M}(D)\to\mathcal{M}(D)} &\leq 2 \sup_{y\in S, A\in\mathcal{B}(D)} |K_{n,\widetilde{y}_{n-2}}(y; A) - K_{n+k,\widetilde{y}_{n+k-2}}(y; A)| \\ &\leq 2a_5(1+\gamma) \sup_{y_1,\dots,y_{n+k-2}\in S} \|C_n - C_{n+k}\| \leq a_{12}k/n, \end{aligned}$$

where the last estimate follows from the definition (5). □

Remark. One could easily modify the proof of [6, Thm 2] to obtain a considerably stronger result with less restrictive assumptions. However, the above result is enough for our purposes here. Some interesting generalizations (for a slightly modified algorithm, though) are obtained in [1],[2]. We expect the result to hold under quite minimal assumptions, especially without the extra smoothness and decay of π assumed in [2]).

The above proof works without changes to modifications of the DRAM algorithm, where one, e.g., keeps the second covariance fixed all the time, or adapts only after prescribed periods.

6 Test examples

The examples presented below are artificially constructed to show that when either one of the two building blocks of DRAM, namely DR and AM, are badly designed, the combination of them almost automatically solves the problems that would appear running each one of them separately.

In the test examples we consider unimodal target distributions, i.e., Gaussian type distributions in the case of linear models and 'banana' shaped distributions in cases of nonlinear models.

We conclude with a more realist application where neither DR nor AM alone works properly but the combination of the two seems to work quite well.

6.1 Test example 1

We shall first employ the targets already used in [5] and [6] as test cases. More specifically, we use the correlated Gaussian distribution (π_2) and the 'strongly nonlinear banana-shaped' distribution (π_4). The distributions allow an exact computation for the, e.g., 50 % and 90 % probability regions, so the correctness of the MCMC runs can be easily verified.

In all the test runs we have compared the results obtained from the basic Metropolis–Hastings (MH), Adaptive Metropolis (AM), basic Delayed Rejection (DR) and the combination DR+AM (DRAM). In the first set of runs, we use correlated Gaussian target distributions π_2 in various dimensions.

We first want to test situations where the proposal distributions are selected to have too small variance with respect to the target distribution. The initial proposal distribution in all cases was a sphere multiplied by the scaling factor $s_d = (2.4)^2/d$, and additionally scaled down by a constant factor. In DR, we used one higher stage, whose proposal was obtained by scaling the proposal of the first stage by a shrinking factor 0.1.

As it is well known, in this setting the MH algorithm tends to "walk around" the target distribution with small steps, without effectively exploring the state space. The same naturally is true for DR, if all the proposals are too small . The point here is to see how the AM adaptation is able to fix this problem.

Figures 1 and 2 present typical outcomes of the runs in a two dimensional setting. Figure 1 gives the results of MH and AM, while Figure 2 exhibits the results produced by DR and DRAM, respectively. The additional scaling factor of the proposal here was 0.01. The lower parts of the figures give the proportion of the chain points within the 50% and 90% probability regions during the runs. We can see that the adaptation indeed seems to remove the problem caused by too small variance in proposal distributions for both the MH and the DR.

For more reliable statistics, we performed the above runs repeatedly, with increasing dimensions of the target distribution, $dim = 2, 5, 10, 15, ...50$. The chain length was kept constant, 20000 for all dimensions. Otherwise the settings are the same as above, the variance of the basic proposal distribution was again scaled down by a factor of 0.01. In each dimension, the runs were repeated 100 times. The mean values over the repetitions

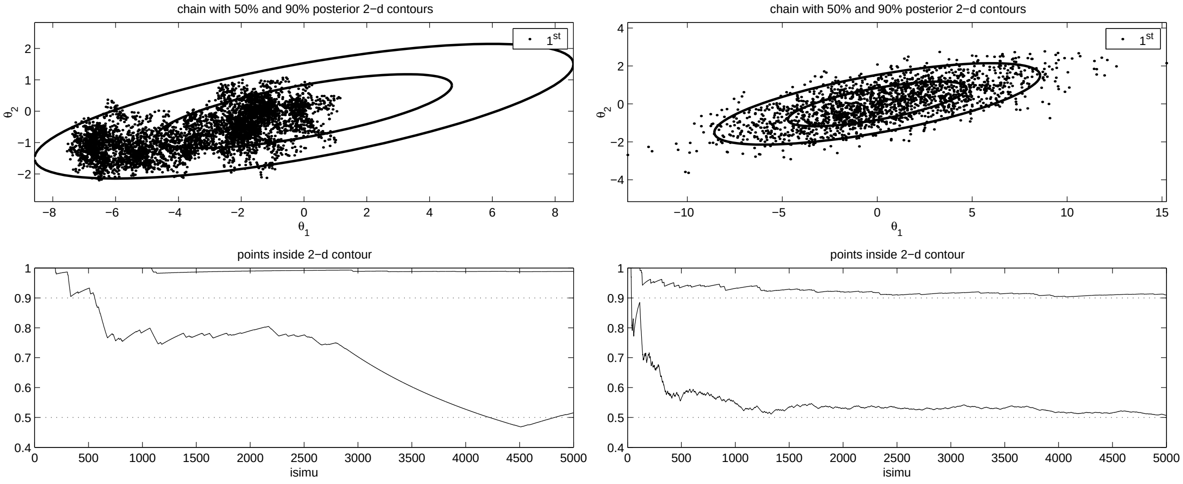

Figure 1: Left figures: results by MH, with too small variance for proposal. Right figures: results by AM, started with the same proposal distribution as with MH

were computed for the proportion of the chain points in the 50% and 90% probability regions, as well as for the center point of the distribution as computed by the chain. The center point of the target distribution was always at the origin, so the norm of the average value of the chain can be used as a measure of the error of the estimate for the expectation.

Figure 3 shows the mean errors for the center point of the distributions. We see that the adaptive algorithms clearly outperform the MH and DR runs, where the center point of the chain parameters may get strongly biased estimates.

Figure 4 shows the results for the 50% and 90% regions. We can see that for moderate dimensions, up to around $dim = 30$, the performance of all versions are comparable. For higher dimensions ($dim = 40$ and $dim = 50$), the adaptation seems to concentrate too much points in the central part of the target distribution. This is a known difficulty with the basic form of the AM adaptation. The combination, DRAM, might slightly improve the situation, but it does not remove this problem with AM in case of higher dimensions. Methods for adaptation for high dimensional problems are studied elsewhere, e.g., in [7], therefore here we will focus on simulations in moderate dimensions.

The same set of tests was also run for the strongly nonlinear (π_4) target distribution. The results were quite similar to the above ones, and are not reported here.

6.2 Test example 2

Here we run basically the same tests as in Example 1, but take the proposals so that the AM adaptation has difficulties to get started. The target distributions are the same as those in Example 1. But the basic proposal distribution – the fixed proposal for MH, initial for AM, first stage proposal for DR, and initial first stage for DRAM – is now scaled up by a factor or 4. The size of this factor was mainly chosen to get the AM adaptation

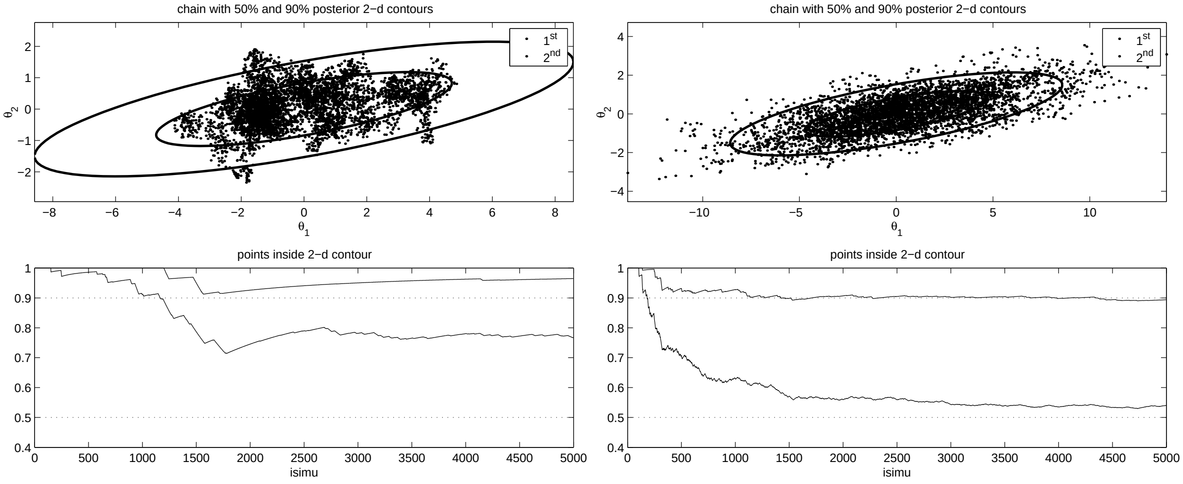

Figure 2: Left figures: results by DR, with too small variance for proposals. Right figures: results by DRAM, started with the same proposal distributions as with DR

at least started. In fact, with an essentially larger initial proposal, practically no new points would be accepted, and no adaptation would take place. The shrinking factor for DR was kept in the value 0.1.

As above, we run repeated simulations for increasing dimensions. Again, we see that the adaptive algorithms clearly outperform the MH and DR runs in computing the expected value of the distribution, see Figure 5. Figures 6 give the proportions of sampled points in the 50% and 90% probability regions. We can observe how the results of both MH and AM are clearly improved by combining them with DR. Note that now DRAM properly works in all dimensions tested.

6.3 Example 3

Our last example presents a situation where neither AM nor DR works properly alone, but the combination DRAM has no difficulties. Consider a simple chemical reaction $A \rightleftarrows_{k_2}^{k_1} B$, where a component A goes to B in a reversible manner, with reaction rate coefficients k_1, k_2. So the dynamics is given by the ODE system

$$\frac{dA}{dt} = -k_1 A + k_2 B, \quad \frac{dB}{dt} = k_1 A - k_2 B$$

with some initial values A_0, B_0 at $t = 0$. The parameter estimation task would be to find values for k_1, k_2 when data for, e.g., $A(t) = k_2/(k_1 + k_2) + (A_0 - k_2/(k_1 + k_2))e^{-(k_1+k_2)t}$ has been obtained at given sampling times of t. Suppose now that the data has been sampled too late, in the sense that the reaction already has reached a steady–state equilibrium at the sampling times, cfr. Figure 7. It is clear that from such data the values of the parameters can not be separately determined, only the ratio k_1/k_2 may be identified, as well as lower bounds for k_1 and k_2. The posterior distribution for k_1, k_2 would be a

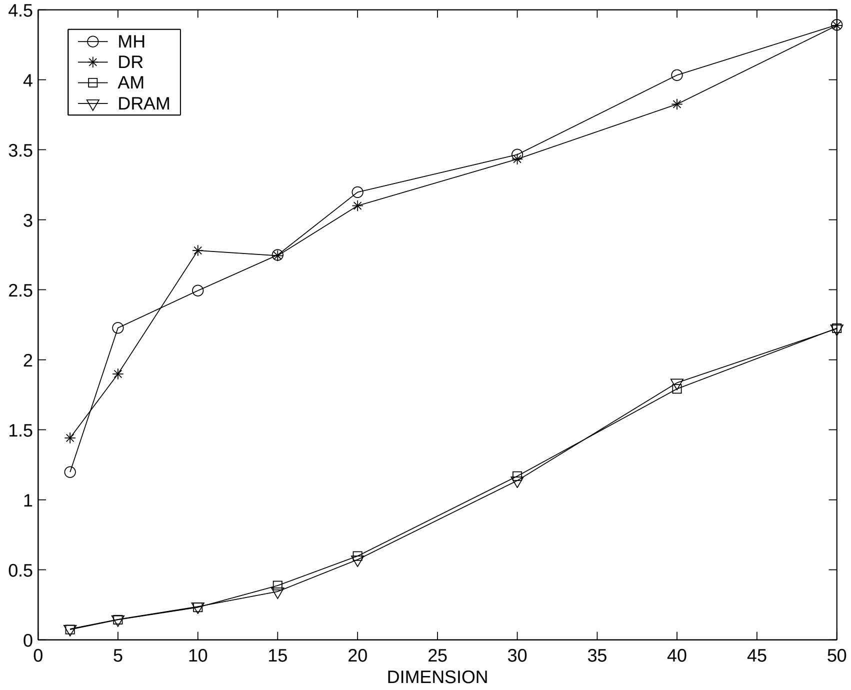

Figure 3: Errors in the estimates of the center point of the distribution. Average results for cases with too small proposal variance.

practically infinite 'zone' in a direction where $k_1/k_2 = \text{const}$. As a test case, we try to find this posterior with MH,DR,AM and DRAM.

While given here in a simplistic setting, situation of this type is, in fact, rather often faced in the parameter estimation of dynamical systems. Some parts of the dynamics is very fast, or internal structural characteristics of the model lead to strongly correlated parameter combinations. In more complex situations, it may not be easy to observe the correlations beforehand. MCMC methods should work in these situations, too. Indeed, they can provide a good tool for analyzing the identifiability of the parameters.

A standard procedure would be to estimate the parameters by least squares fitting, compute the covariance matrix of the parameters by the approximative Hessian matrix, and use it to construct the proposal distribution for MH. However, in the setting of our example there is a problem. There is no unique minimum for the least squares function, and the covariance matrix is singular.

Figure 7 shows a typical run with DR. The computed approximative covariance does not provide a good proposal, and the efficiency remains very low. The acceptance rate with the first stage proposal – that is, with the MH proposal – is around 0.4%, with the second stage proposal around 4–5%. For the second stage we used scaled version of the first one, both with smaller and larger variances. We may conclude that while DR is better than MH here, the sampled parameter values do not provide a proper coverage of the posterior of the parameters.

One could expect that AM would find, possibly after some initial trials, a well cali-

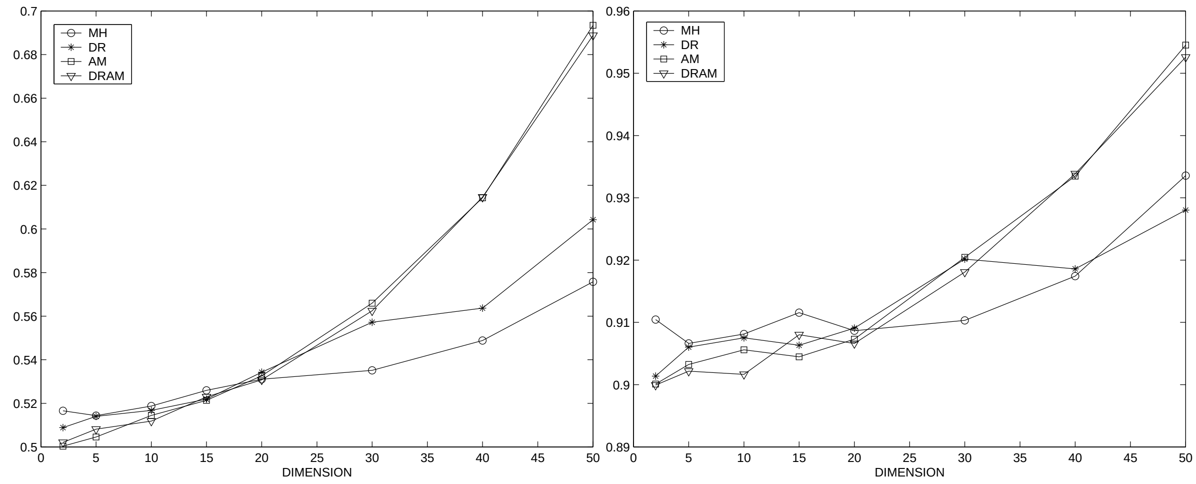

Figure 4: Proportions of sampled points in the 50 % and 90 % probability regions. Average results for cases with too small proposal variance.

brated proposal distribution. However, since the initial proposal is so poor, it can take a very long time for AM to start working. Figure 7 illustrates a typical case. The 'wake–up' time may be long or quite short, but the success of AM is, at best, uncertain.

The combination of AM and DR was employed as outlined before. The first stage had the proposal obtained from the covariance of the fit, for the second stage the proposal was scaled down by a factor of 0.1. The result is a dramatic improvement, see Figure 8. The second DR stage is able to find acceptable proposals right for the beginning, the AM adaptation starts immediately. The acceptance rate and the mixing of the chain are nearly optimal.

7 Conclusions

We show how two ways of modifying the standard MH sampler can be successfully combined. The first modification, AM, aims at adapting the proposal distribution based on the past history of the chain. The second modification, DR, aims at improving the efficiency of the resulting MCMC estimators. While AM allows for "global" adaptation, based on all the previously accepted proposals, DR may allow for "local" adaptation, only based on rejected proposals within each time-step. There are different ways of combining AM and DR. We tried some basic but very effective ones, as the simulation results show. We plan to further investigate different ways of combining DR and AM.

Acknowledgement We would like to thank Johanna Tamminen for helpful discussions during this work.

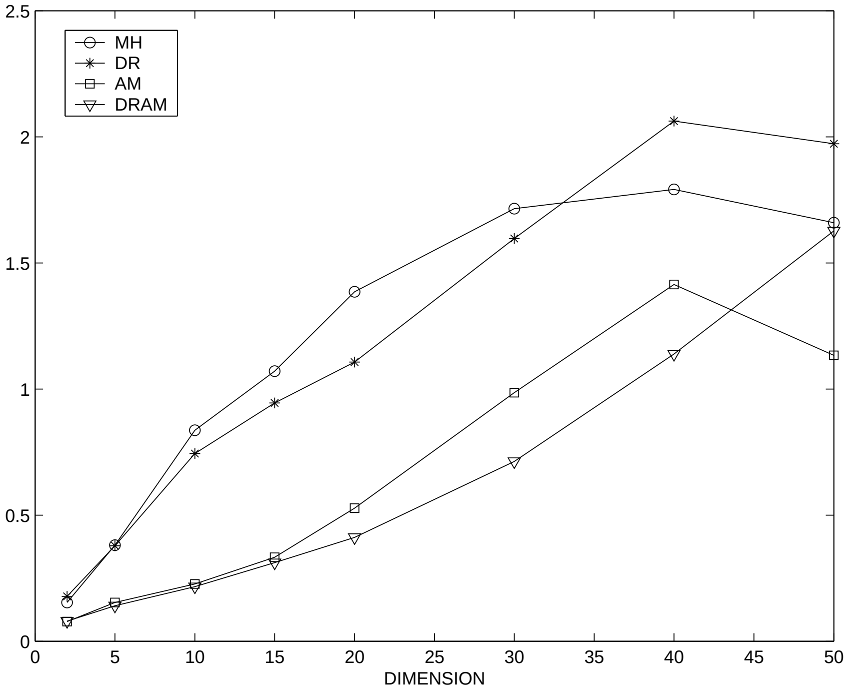

Figure 5: Errors in the estimates of the center point of the distribution. Average results for cases with too large proposal variance.

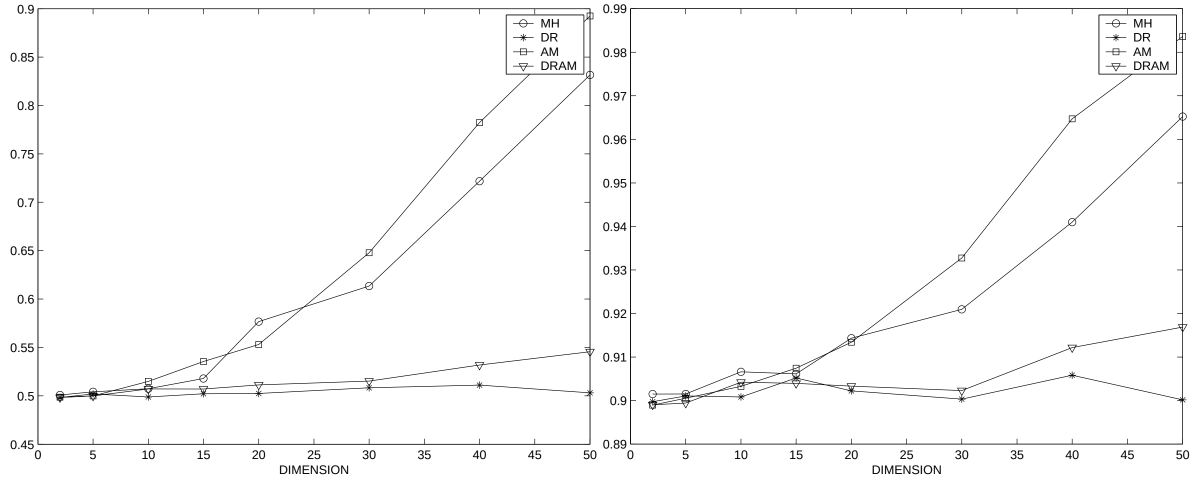

Figure 6: Proportions of sampled points in the 50 % and 90 % probability regions. Average results for cases with too large proposal variance.

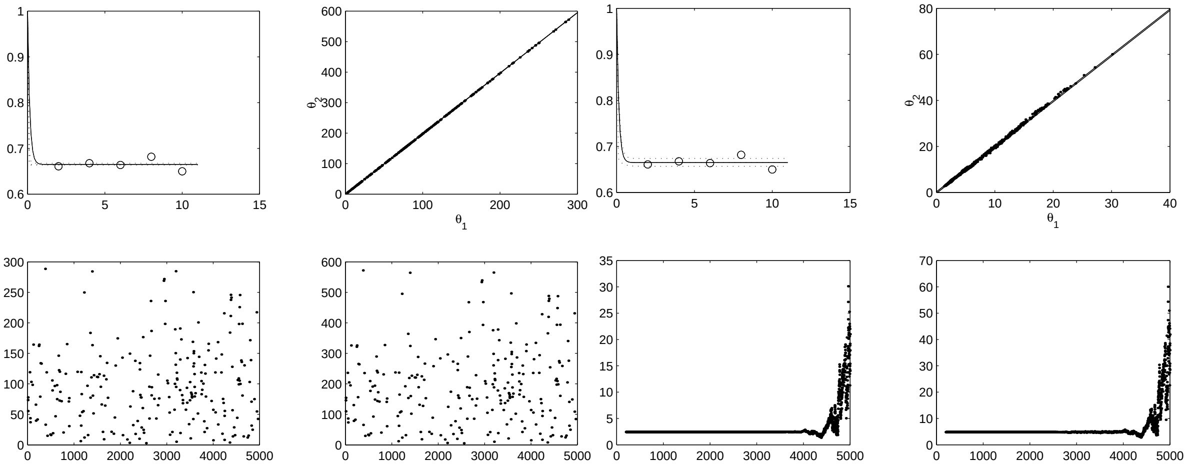

Figure 7: Left 4 figures: results by DR. Right 4 figures: results by AM. Top left: The data, the fit, the 95% probability values as computed by the MCMC chain. Top right: the computed 2D MCMC chain. The lower figures: the parameter chains.

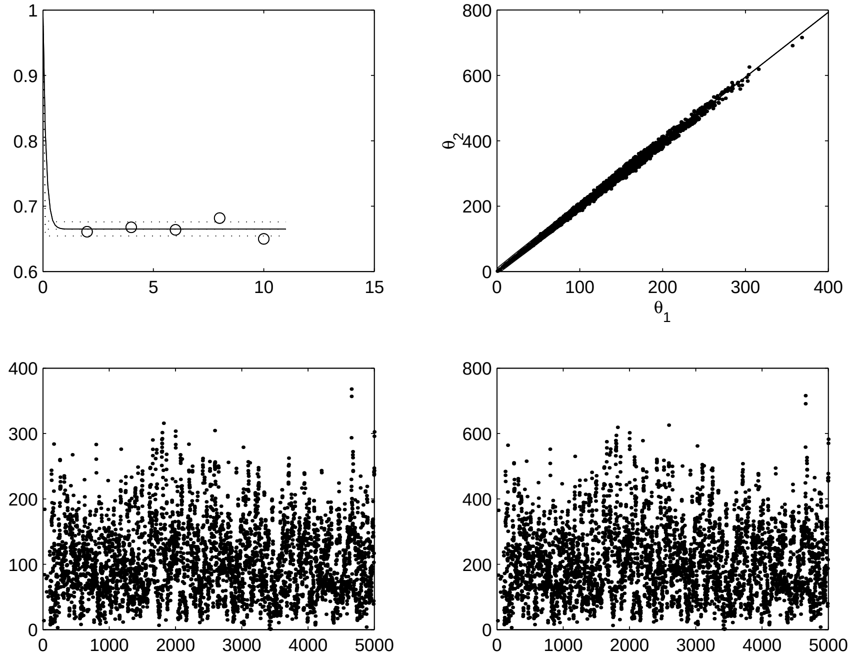

Figure 8: Results by DRAM. Top left: The data, the fit, the 95% probability values as computed by the MCMC chain. Top right: the computed 2D MCMC chain. The lower figures: the parameter chains.

References

[1] C. Andrieu and C.P. Robert: *Controlled MCMC.* Preprint 2001.

[2] C. Andrieu and E. Moulines: *On the ergodicity properties od some adaptive MCMC algorithms.* Preprint 2002.

[3] A. G. Gelman, G. O. Roberts and W. R. Gilks: *Efficient Metropolis jumping rules,* Bayesian Statistics V pp. 599–608 (eds J.M. Bernardo, J.O. Berger, A.F. David and A.F.M. Smith). Oxford University press, 1996.

[4] Green, P. J. and Mira, A. (2001). *Delayed rejection in reversible jump Metropolis-Hastings.* Biometrika, Vol. 88, pp. 1035-1053.

[5] H. Haario, E. Saksman and J. Tamminen: *Adaptive proposal distribution for random walk Metropolis algorithm,* Comp. Stat. 14 (1999), 375–395.

[6] H. Haario, E. Saksman and J. Tamminen: *An adaptive Metropolis algorithm,* Bernoulli 7 (2001), 223-242.

[7] H. Haario, E. Saksman and J. Tamminen: *Componentwise adaptation for MCMC.* Reports of the Department of Mathematics, Preprint 342, April 2003.

[8] A. Mira. On Metropolis-Hastings algorithms with delayed rejection. *Metron*, 2001, Vol. LIX, n. 3-4, pp. 231-241.

[9] A. Mira. Ordering and improving the performance of Monte Carlo Markov Chains, *Statistical Science*, 2002, Vol. 16, pp 340-350.

[10] P. H. Peskun. Optimum Monte Carlo sampling using Markov chains. *Biometrika*, 60:607–612, 1973.

[11] A. D. Sokal. Monte Carlo methods in statistical mechanics: foundations and new algorithms. *Cours de Troisième Cycle de la Physique en Suisse Romande*, Lausanne, 1998.

[12] L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1701–1762, 1994.

[13] L. Tierney. A Note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied Probability*, 8:1–9, 1998.

[14] L. Tierney and A. Mira. Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine*, 1999, 18, pp. 2507-2515.