NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Mode jumping proposals in MCMC

by

Håkon Tjelmeland and Bjørn Kåre Hegstad

PREPRINT
STATISTICS NO. 1/1999

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL `http://www.stat.ntnu.no/preprint/statistics/1999/S1-1999.ps`
Håkon Tjelmeland has homepage: `http://www.stat.ntnu.no/~haakont`
E-mail: `haakont@stat.ntnu.no`
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology,
N-7034 Trondheim, Norway.

Mode jumping proposals in MCMC

Håkon Tjelmeland
Department of Mathematical Sciences
Norwegian University of Science and Technology
Trondheim, Norway

Bjørn Kåre Hegstad
Statoil
Stavanger, Norway

Abstract

Markov chain Monte Carlo algorithms generate samples from a target distribution by simulating a Markov chain. Large flexibility exists in the specification of transition matrix of the chain. In practice, however, most algorithms used only allow small changes in the state vector in each iteration. This choice typically causes problems for multi-modal distributions as moves between modes become rare and, in turn, results in slow convergence to the target distribution.

In this paper we consider continuous distributions on $\Re^n$ and specify how optimisation for local maxima of the target distribution can be incorporated in the specification of the Markov chain. Thereby, we obtain a chain with frequent jumps between modes. We demonstrate the effectiveness of the approach in two examples. The first considers a simple mixture of bivariate normal distributions. In the second example we reconsider sampling from a posterior distribution appearing in the analysis of reaction times for schizophrenic individuals (Gelman and Rubin, 1992; Gelman et al., 1995). This density is known to be multi-modal and we demonstrate how our approach prevent the sampler from being trapped in minor modes.

Key words: multi-modal distributions, Markov chain Monte Carlo, Metropolis-Hastings, optimisation

1 Introduction

Markov chain Monte Carlo (MCMC) algorithms allow generation of samples from a specified target distribution, see, for example, Besag et al. (1995) and references therein. The Metropolis-Hastings setup gives a general framework for specification of the Markov chain and each iteration consists of two parts. First, with x denoting the current state, a potential new state, x', is generated according to a specified proposal kernel, $q(x'|x)$, and, second, x' is accepted with a given probability, $\alpha(x'|x)$.

In practice, the use of MCMC to sample from multi-modal distributions causes problems, the reason being that most MCMC algorithms used only propose small changes of the state vector in each iteration. Thus, moves between modes become rare and, in turn, results in slow convergence of the chain and high correlations between successive states. Different remedies have been proposed to obtain Markov chains with faster convergence. Simulated tempering (Marinari and Parisi, 1992; Geyer and Thompson, 1995) and use of tempered transitions (Neal, 1996) are specifically designed to cope with multi-modal distributions. Both approaches define a series of distributions between the "cold" one, equal to the distribution of interest, and a "warm" distribution without any isolated modes. Letting the chain move between the different distributions, either stochastically (simulated tempering) or systematically (tempered transitions), one enables the chain to move from one mode to another by going via the warmer distributions. Another variant is Metropolis-coupled Markov chains (Geyer, 1991), where one state variable is used for each of the defined distributions and occasionally one proposes to swap values in neighbouring states.

In the present paper, we limit our attention to continuous probability distributions on $\Re^n$. An "obvious" solution is then to use local optimisation to locate modes. However, the problem is to include this in a framework that assures that modes are visited according to their probability. The solution is not to run several chains with initial states given by optimisations from different starting points, as then the frequencies of which the modes are represented are given by their "basin of attractions", not by the total probability within each mode.

The paper is organised as follows. Section 2 briefly review some Markov chain theory and the Metropolis-Hastings algorithm, followed by a new generalised scheme for construction of the transition kernel. In the next section, we introduce two specific choices within this setup. In Section 3.1, optimisation is employed to design a sampler which enables direct jumps between modes, whereas in Section 3.2 we discuss how to include proposals to an area of prior interest. Section 4 presents two examples. We first consider a simple mixture of Gaussian distributions on $\Re^2$. Thereafter, we revisit a dataset of reaction times for six schizophrenic and eleven non-schizophrenic individuals (Gelman and Rubin, 1992; Gelman et al., 1995), where the resulting posterior distribution is known to be multi-modal. Finally, Section 5 provides conclusions.

2 Markov chain Monte Carlo

For a general introduction to MCMC, see the references given above. Here, we first introduce the basic concepts of Markov chain theory and briefly review the Metropolis-Hastings algorithm. Thereafter we introduce a generalised version of the Metropolis-Hastings setup. For definiteness, we restrict the attention to continuous target distributions on $\Re^n$ and let $\pi(x)$ denote the density with respect to the Lebesgue measure.

2.1 Detailed balance condition

The basic idea in MCMC is to construct a stationary Markov chain, with transition kernel P, say, on $\Re^n$ with limiting distribution equal to the target distribution $\pi(x)$ and simulate the chain for a long but fixed period, after which it has essentially converged. Thus, P must be chosen so that

$$\int_A \pi(x)\mathrm{d}x = \int_{\Re^n} \mathrm{P}(A|x)\pi(x)\mathrm{d}x \qquad \forall\ A \in \mathcal{F}, \tag{1}$$

where $\mathcal{F}$ is the Borel σ-field on $\Re^n$. Imposing the chain to be time reversible, one obtain the more restricting detailed balance condition,

$$\int_A \pi(x)\mathrm{P}(B|x)\mathrm{d}x = \int_B \pi(x)\mathrm{P}(A|x)\mathrm{d}x \qquad \forall\ A, B \in \mathcal{F}. \tag{2}$$

If two transition kernels P_0 and P_1, both fulfil equation (1), it follows directly that also $\mathrm{P} = a\mathrm{P}_0 + (1-a)\mathrm{P}_1$ has the same property for all $a \in [0,1]$. Alternatively, a mixture of kernels indexed with a continuous parameter, $\varphi \in \Re^d$, may be used. More precisely, let P^φ define a class of kernels which for each $\varphi \in \Re^d$ fulfil equation (1) and let $f(\varphi)$ be a density on $\Re^d$. Again it follows from direct calculations that also

$$\mathrm{P}(A|x) = \int_{\Re^d} \mathrm{P}^{\varphi}(A|x) f(\varphi) \mathrm{d}\varphi \tag{3}$$

fulfils the detailed balance condition. Algorithms based on mixtures of transition kernels of the types discussed above are routinely used in MCMC.

2.2 Metropolis-Hastings

The Metropolis-Hastings algorithm provides a general design of P for which relation (2) is fulfilled. The construction is based on a proposal kernel Q on $\Re^n$. Assuming Q to have a density, $q(y|x)$, with respect to the Lebesgue measure, the corresponding proposal kernel P in the Metropolis-Hastings algorithm is given by

$$\mathrm{P}(A|x) = \int_A q(y|x)\alpha(y|x)\mathrm{d}y + I(x \in A)r(x), \tag{4}$$

where $\alpha(y|x)$ is the acceptance probability,

$$\alpha(y|x) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\frac{q(x|y)}{q(y|x)}\right\}, \tag{5}$$

$I(\cdot)$ is the indicator function and $r(x)$ is the probability for remaining at x, given by

$$r(x) = \int_{\Re^n} q(y|x)(1 - \alpha(y|x))\mathrm{d}y. \tag{6}$$

If two proposal kernels, Q_0 and Q_1, are available, different schemes exist for combining them into one kernel P that fulfils the detailed balance conditions. Two obvious possibilities are 1) to use the Metropolis-Hastings construction for each of the two proposal kernels, giving P_0 and P_1, which in turn are combined into $\mathrm{P} = a\mathrm{P}_0 + (1-a)\mathrm{P}_1$, and 2) define $\mathrm{Q} = a\mathrm{Q}_0 + (1-a)\mathrm{Q}_1$ for some $a \in [0,1]$ and use this as a proposal kernel in the Metropolis-Hastings setup. In the next section, we focus on a third possibility which in fact defines a generalisation of Metropolis-Hastings.

2.3 Generalised Metropolis-Hastings

Let Q_0 and Q_1 denote two proposal kernels on $\Re^n$ with corresponding densities $q_0(y|x)$ and $q_1(y|x)$, respectively. To obtain a transition kernel with the required invariant distribution, scheme 1) above associates $q_i(y|x)$ with $q_i(x|y)$ for each $i = 0, 1$ separately. A more flexible construction is obtained by associating $q_i(y|x)$ with $q_{1-i}(x|y)$ for each $i = 0, 1$. More precisely, let the Markov kernel P be defined by

$$P(A|x) = \frac{1}{2}\int_A q_0(y|x)\alpha_{0,1}(y|x)\mathrm{d}y + \frac{1}{2}\int_A q_1(y|x)\alpha_{1,0}(y|x)\mathrm{d}y + I(x \in A)r(x), \tag{7}$$

where the acceptance probabilities $\alpha_{0,1}(y|x)$ and $\alpha_{1,0}(y|x)$ are given by

$$\alpha_{i,1-i}(y|x) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\frac{q_{1-i}(x|y)}{q_i(y|x)}\right\} \quad \text{for} \quad i = 0, 1 \tag{8}$$

and the rejection probability $r(x)$ now becomes

$$r(x) = \frac{1}{2}\int_{\Re^n} q_0(y|x)(1-\alpha_{0,1}(y|x))\mathrm{d}y + \frac{1}{2}\int_{\Re^n} q_1(y|x)(1-\alpha_{1,0}(y|x))\mathrm{d}y. \tag{9}$$

The resulting transition kernel, P, will indeed fulfil the required detailed balance conditions, which can easily be verified by direct calculations. One may alternatively confirm the validity of the approach by observing that the resulting procedure is a Metropolis-Hastings algorithm on a larger state space, namely for the target distribution

$$\pi(x,i) = \pi(x)\cdot\frac{1}{2} \quad \text{for} \quad x \in \Re^n, i = 0, 1. \tag{10}$$

The corresponding Metropolis-Hastings algorithm alternates between two types of updates. Denoting the current and potential new states by (x,i) and (y,j), respectively, the first type of update sets $y = x$ and draw j to 0 or 1 with probability a half for each. The acceptance probability is one. The second update type proposes y from $q_i(y|x)$ and set $j = 1 - i$. The corresponding Metropolis-Hastings acceptance probability becomes $\alpha_{i,j}(y|x)$ given by equation (8). In the next section, we discuss two applications of generalised Metropolis-Hastings moves.

3 Mode jumping proposals

In this section we assume the target distribution to be multi-modal and consider how to construct transition kernels that give frequent jumps between modes. To handle this problem we define two new kernels, both of which are based on a set of transition kernels $\{\mathrm{P}^\varphi; \varphi \in \Re^n\}$ as given by equation (3). Each P^φ is constructed from two proposal kernels, Q_0^φ and Q_1^φ, according to the generalised Metropolis-Hastings setup. To obtain direct jumps between modes, the proposal kernels must be chosen so that 1) large changes in the state vector are proposed, 2) the potential new states must be located in high-probability areas, and 3) the acceptance probabilities must be reasonably high. Our strategy is to design Q_0^φ so that 1) and 2) is fulfilled and focus on 3) in the choice of Q_1^φ.

3.1 Proposals via local optimisation

We first discuss how local optimisation can be utilised in the specification of Q_0^φ and Q_1^φ. Our choice of Q_0^φ is illustrated in Figure 1 and consists of three parts; First we go from x to $T_0(x,\varphi) = x + \varphi$. Thus, if φ is large, requirement 1) above is fulfilled. Next, to accomplish 2), we locate a high probability area by a deterministic local minimisation of the energy function, $U(x) = -\ln(\pi(x))$, starting at $T_0(x,\varphi)$. Let $\mu(x)$ denote the location of the minimum found when starting at x and let $\Sigma(x)$ be the inverse of the Hessian at the minimum, i.e. $\Sigma(x) = [(\nabla^2 U)(\mu(x))]^{-1}$. Thus, our minimisation returns $\mu(T_0(x,\varphi))$ and a natural choice of $q_0^\varphi(y|x)$ is

$$q_0^\varphi(y|x) = \mathrm{N}_n(\mu(T_0(x,\varphi)), \Sigma(T_0(x,\varphi)))(y), \tag{11}$$

where $\mathrm{N}_n(\mu,\Sigma)(\cdot)$ denotes the density function of a n-variate Gaussian distribution with mean μ and covariance Σ.

To achieve 3) above, the Q_1^φ must be chosen so that the reverse jump, from y to x, has high probability under this kernel. If $\mu(T_0(x,\varphi))$ is relatively close to $T_0(x,\varphi)$, there should

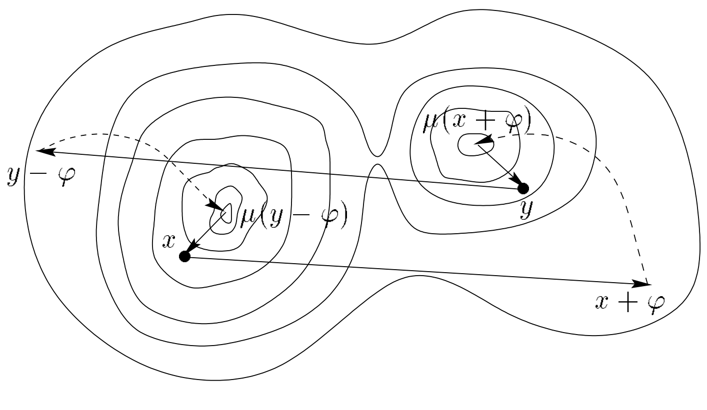


Figure 1: Illustration of proposals via local optimisation. The current state is denoted x. A deterministic local optimisation is started at $x + \varphi$ and returns $\mu(x + \varphi)$. Based on local properties of the target distribution at $\mu(x + \varphi)$, the proposal density $q_0(y|x)$ is defined and the proposal y is generated. The proposal density $q_1(x|y)$ for the reverse step is given via the optimisation started at $y - \varphi$, which ends at $\mu(y - \varphi)$.

be a good chance that $T_1(y, \varphi) = y - \varphi$ is located within the basin of attraction of the mode in which x is located. Thus, a reasonable choice of $q_1^{\varphi}(x|y)$ becomes

$$q_1^{\varphi}(x|y) = \mathrm{N}_n(\mu(T_1(y, \varphi)), \Sigma(T_1(y, \varphi)))(x), \tag{12}$$

again see Figure 1 for an illustration. Obviously one risk that $\mu(T_1(y, \varphi))$ misses the mode of x and a negligible acceptance probability results, but this causes no problem unless it happens too often.

It should be noted that even if we argue differently for Q_0^{φ} and Q_1^{φ}, they are symmetrically defined and the same reasoning is also valid with the two interchanged. Also note that the proposed algorithm may be altered in several respects without violating the validity of the procedure. We discuss this in the following remarks.

Remark 1: Obviously, the transition kernel discussed above should not be used alone, but be combined with standard Metropolis-Hastings transition kernels doing local changes within each mode more efficiently.

Remark 2: Not all elements of the state vector need to be updated in each iteration. This becomes important if the dimension of the state vector, n, is large as then optimisation in the full dimensional space will require too much computer resources. Instead, the above approach can be included in a multi-grid algorithm (Goodman and Sokal, 1989; Besag and Green, 1993) so that the optimisations are performed in lower dimensional spaces.

Remark 3: Any type of deterministic optimisation algorithm may be used in the definition of $\mu(x)$. In fact $\mu(x)$ and $\Sigma(x)$ may be any functions of x (where $\Sigma(x)$ is positive definite for all x). For example, if second order derivatives of $U(x)$ are unavailable or expensive to compute, one may use a quasi-Newton algorithm which, inherent in the optimisation, builds up an approximation to the Hessian matrix that can be used to define $\Sigma(x)$.

Remark 4: It is not necessary that the optimisation is iterated until a local optimum is found. In practice, a few iterations may be sufficient. However, then the Hessian may not be positive definite. Thus, a rule for what to do in this situation must be established. Of course, this rule may simply be to propose an unchanged state vector or to obtain a positive definite matrix by adding a sufficiently large number to the diagonal elements of $(\nabla^2 U)(\mu(x))$.

Remark 5: The deterministic transformations $T_0(x, \varphi)$ and $T_1(x, \varphi)$ may not be additive. For example, one may use $T_0(x, \varphi) = x \odot \varphi$ and $T_1(x, \varphi) = x \oslash \varphi$, where $\odot$ and $\oslash$ denote componentwise multiplication and division, respectively. To reduce the risk that $\mu(T_1(y, \varphi))$ misses the mode of x, it seems reasonable to require $T_1(T_0(x, \varphi), \varphi) = T_0(T_1(x, \varphi), \varphi) = x$, but even this is not necessary for the validity of the approach.

Remark 6: With $\varphi = (\varphi_0, \varphi_1) \in \Re^{2n}$, $f(\varphi) = f(\varphi_0) \cdot f(\varphi_1)$ and $T_i(x, \varphi) = \varphi_i$ for $i = 0, 1$ one obtains an independent proposal sampler.

Remark 7: The proposal distribution does not need to be Gaussian, any continuous multivariate distribution will do. For example, if one suspects modes with heavy tails, a multivariate t- or Cauchy-distribution (Johnson and Kotz, 1972) may be chosen. Alternatively one may at each iteration draw at random which type of proposal distribution to use. The same type may be used for both Q_0^φ and Q_1^φ, or the two may be drawn independently of each other. Formally, the parameter vector φ is now increased with one or two (discrete) elements which specify the class of distributions to use.

3.2 Proposals to area of prior interest

When constructing a Markov chain sampler for a specific target distribution $\pi(x)$ there may exist one or several areas in the sample space which one suspects to be of large importance. This believe may be based on understanding of the problem from which $\pi(x)$ is defined, results from similar studies, or, as in our example in Section 4.2, from studies of the target distribution itself. In the construction of the MCMC sampler, one would then like to ensure that moves to and from these specific areas are possible.

A reasonable approach in this situation is to construct a proposal density $g(x); x \in \Re^n$ having most of its probability mass in the area of prior interest and from which direct sampling is possible. However, to use $g(x)$ as a proposal density in a standard Metropolis-Hastings setup is of limited practical value. Unless the current state is already within the area of interest, the acceptance probability of such a move becomes very small. With the added flexibility of generalised Metropolis-Hastings, the situation change. Again, define a set of proposal kernels indexed by $\varphi \in \Re^n$, where each P^φ is defined via two proposal kernels, Q_0^φ and Q_1^φ. As the first proposal distribution we use $g(\cdot)$, i.e.

$$q_0^\varphi(y|x) = g(y). \tag{13}$$

Thus, the proposal is independent of the current state x and does not depend on the index φ. For the acceptance probability to become reasonably high, Q_1^φ must again be chosen to give high probability for the reverse transition. A reasonable choice would be

$$q_1^\varphi(x|y) = \mathrm{N}_n(\mu(\varphi), \Sigma(\varphi))(x), \tag{14}$$

where $\mu(\cdot)$ and $\Sigma(\cdot)$ are defined via optimisation as in Section 3.1. If φ is within the basin of attraction of the mode in which the current state is located, the acceptance probability $\alpha_{10}(y|x)$ becomes quite high and, thus, enables direct jumps to the area of prior interest. Correspondingly, if the current state is within the area of focus, a proposal from Q_1^φ allows direct jumps to other modes.

4 Examples

In the first of our two examples in this section we use proposals via local optimisation to sample from a mixture of Gaussian distributions on $\Re^2$. In the second example we revisit a data set analysed in Gelman and Rubin (1992), for which a Bayesian model is adopted. The resulting posterior distribution is defined on $\Re^{23}$ and is multi-modal. We specify a sampler which is able to jump between the different modes and find, as also concluded in previous studies, that essential all probability mass is contained in one of the modes.

4.1 Mixture of Gaussian distributions

Let the target distribution be given by

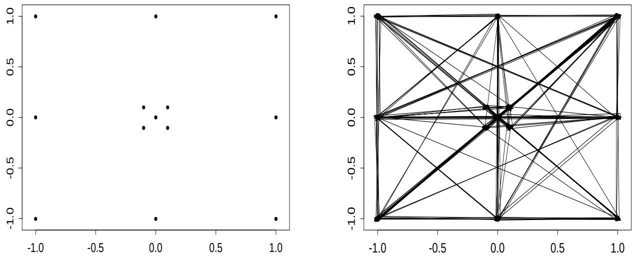
$$\pi(x) = \sum_{k=1}^{K} \omega_k \mathrm{N}_2(\mu_k, \Sigma_k)(x) \quad \text{for} \quad x \in \Re^2, \tag{15}$$

where $\omega_k > 0$ for $k = 1, \ldots, K$, $\omega_1 + \ldots + \omega_K = 1$ and μ_k and Σ_k for $k = 1, \ldots, K$ are locations and covariance matrices for each of the modes. Below, we consider two choices for K, μ_k's and Σ_k's. In both cases the same simulation algorithm is used; we alternate between 250 Metropolis-Hastings steps where local changes are proposed and one proposal via local optimisation. The total of these 251 steps we call one iteration. In the Metropolis-Hastings steps potential new states are generated from a biGaussian distribution centered at the current state and with a diagonal covariance matrix with both diagonal elements equal to 0.001^2. In the mode jumping proposals we use $T_0(x, \varphi) = x + \varphi$, $T_1(x, \varphi) = x - \varphi$ and the distribution $f(\varphi)$ is a biGaussian distribution with zero mean vector and diagonal covariance matrix with diagonal elements 2^2. For optimisation we use a Newton algorithm with line search, see, for example, Luenberger (1984). As our interest here is on the ability of the algorithm to jump between modes and not rate of convergence, we draw the initial state from the target distribution itself.

Case 1: Let $K = 13$, $\omega_1 = \ldots = \omega_{13} = 1/13$, $\mu_1 = (0, 0)$, $\mu_2 \ldots, \mu_5$ be located in $(0.1, 0.1)$, $(-0.1, 0.1)$, $(0.1, -0.1)$ and $(-0.1, -0.1)$, respectively, and let μ_6 to μ_{13} be situated at $(1, 1)$, $(1, 0)$, $(1, -1)$, $(0, 1)$, $(0, -1)$, $(-1, 1)$, $(-1, 0)$ and $(-1, -1)$, respectively, see Figure 2(a). Finally, let $\Sigma_1 = \ldots = \Sigma_{13} = \mathrm{diag}(0.01^2, 0.01^2)$, i.e. $\pi(x)$ has 13 well separated modes and transitions between modes via the local Metropolis-Hastings moves will essentially never occur. One should note that the optimisation actually defines a partition of $\Re^2$ into 13 basins of attractions, $A_1, \ldots, A_{13}$, where $x \in A_i$ if an optimisation started at x ends at μ_i. To map these sets we started an optimisation at every node of a grid covering $[-2, 2]^2$ with mesh size 0.02, see Figure 3 for A_1, A_2 and A_{13}.

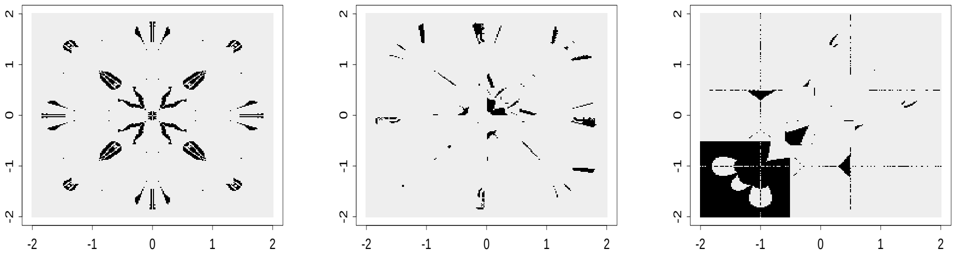
We ran $1,000,000$ iterations of the chain, the first $1,000$ of which is shown in Figure 2(b). Despite the very complex structure of $A_1, \ldots, A_{13}$, the average acceptance rate of the mode jumping proposals was 18.5%. Obviously, some of these moves were to a new location within the same mode as the current state, so the frequency of jumps between modes is somewhat lower than this figure. Defining z_i to be mode number of the mode in which the state vector, x_i, is located after iteration i, i.e.

$$z_i = \operatorname*{argmin}_{k=1,\ldots,13} \{||x_i - \mu_k||\}, \tag{16}$$



(a) (b)

Figure 2: (a) The location of the 13 modes in the case 1 model. (b) Trajectory of the first $1,000$ iterations for the case 1 model.



(a) (b) (c)

Figure 3: Basins of attractions for the modes located at (a) $(0,0)$, (b) $(0.1,0.1)$ and (c) $(-1,-1)$, i.e A_1, A_2 and A_{13}. An optimisation started within the black areas ends in the specified mode.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	.906	.002	.002	.002	.002	.009	.012	.010	.012	.012	.009	.012	.010
2	.002	.950	.010	.002	.002	.004	.003	.005	.004	.004	.005	.004	.005
3	.002	.011	.950	.002	.002	.005	.004	.005	.004	.004	.005	.004	.004
4	.002	.002	.002	.949	.010	.005	.004	.005	.003	.004	.004	.004	.005
5	.002	.002	.002	.010	.949	.005	.003	.004	.004	.003	.004	.004	.005
6	.010	.004	.005	.005	.005	.745	.024	.037	.023	.028	.038	.028	.050
7	.011	.003	.004	.004	.003	.024	.794	.023	.019	.018	.028	.038	.029
8	.010	.005	.005	.005	.004	.037	.023	.743	.029	.024	.050	.030	.037
9	.011	.003	.004	.003	.004	.024	.017	.029	.793	.040	.024	.019	.028
10	.011	.004	.004	.004	.004	.029	.018	.024	.039	.792	.027	.019	.024
11	.010	.005	.005	.004	.005	.036	.029	.049	.023	.028	.745	.024	.037
12	.012	.004	.004	.004	.004	.027	.040	.029	.018	.019	.024	.792	.024
13	.009	.005	.004	.005	.005	.050	.029	.037	.029	.024	.037	.024	.743

Table 1: Estimated transition frequencies between modes for the case 1 model.

	1	2
1	.987	.013
2	.013	.987

Table 2: Estimated transition frequencies between modes for the case 2 model.

where $||\cdot||$ denotes Euclidean distance, we estimate the transition frequency from mode k to mode l by

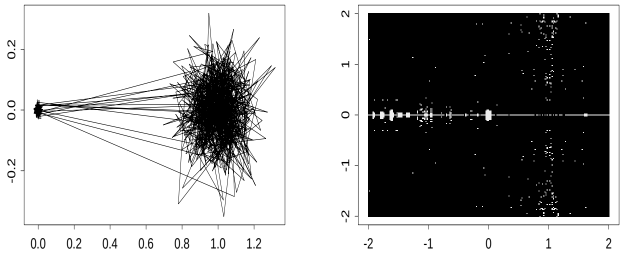
$$\widehat{p}_{k,l} = \frac{\sum_{i=2}^{N} I[z_{i-1} = k \cap z_i = l]}{\sum_{i=2}^{N} I[z_{i-1} = k]}, \tag{17}$$

see Table 1.

Case 2: Next, we consider a target distribution with $K = 2$ modes with equal weights $\omega_1 = \omega_2 = 1/2$ and centered at $\mu_1 = (0,0)$ and $\mu_2 = (1,0)$, respectively. To study how difference in width of the modes affects the performance of the algorithm, we let $\Sigma_1 = \text{diag}(0.01^2, 0.01^2)$ and $\Sigma_2 = 10^2 \cdot \Sigma_1$. Again we perform $N = 1,000,000$ iterations, see Figure 4(a) for the first $1,000$ iterations, and estimate the mean acceptance probability for mode jumping proposals, which is now 0.50. However, most of the accepted mode jumping moves are internal in mode 2, see also the estimated transition frequencies in Table 2. The low number of transitions between modes can be understood from the size of A_1 and A_2, see Figure 4(b). Obviously, if one mode has a very small basin of attraction, the chance of discovering that mode is small and, in turn, transitions to and from that mode become rare.

4.2 Reaction times and schizophrenia

In this section we reconsider sampling from a posterior distribution based on experimental measurements of reaction times for seventeen subjects; six schizophrenic and eleven non-



(a) (b)

Figure 4: (a) The trajectory of the first $1,000$ iterations in the simulation of the case 2 model. (b) Basins of attractions for the two modes in the case 2 model, A_1 in white and A_2 in black.

schizophrenic individuals. The resulting posterior distribution is known to be multi-modal (Gelman and Rubin, 1992). For discussion of the data set, see also Belin and Rubin (1995) and Gelman et al. (1995). We start with a brief introduction to the data and the Bayesian model.

Data and Bayesian model

Thirty reaction times have been measured for each subject in the study, see Figure 5. Let y_{ij} denote the logarithm of response time number j for person number i, where $i = 1, \ldots, 11$ is the non-schizophrenic subjects. Following the assumptions in Gelman and Rubin (1992), we specify a model with 23 parameters, $\alpha = (\alpha_1, \ldots, \alpha_{17})$ and $\theta = (\lambda, \sigma_\alpha^2, \sigma^2, \tau, \mu, \beta)$. The log-response times for non-schizophrenic persons are assumed to be independent and Gaussian distributed with distinct means for each person and a common variance, i.e.

$$y_{ij}|\ \alpha, \theta \sim \mathrm{N}_1(\alpha_i, \sigma^2) \quad \text{for} \quad i = 1, \ldots, 11. \tag{18}$$

To include in the model an assumed tendency of attentional deficiency of the schizophrenic individuals, y_{ij} for those subjects are modelled as a mixture of two Gaussian distributions,

$$y_{ij}|\ \alpha, \theta \sim (1-\lambda)\mathrm{N}_1(\alpha_i, \sigma^2) + \lambda \mathrm{N}_1(\alpha_i + \tau, \sigma^2) \quad \text{for} \quad i = 12, \ldots, 17, \tag{19}$$

where λ and τ, which are assumed to be common for all schizophrenic individuals, are the probability of delay in response and the size of the delay, respectively. The parameters $\alpha_1, \ldots, \alpha_{17}$ are assumed independent, Gaussian distributed with different means for the schizophrenic and non-schizophrenic subjects, but with a common variance, i.e.

$$\alpha_i|\ \theta \sim \mathrm{N}_1(\mu + \beta \cdot S_i, \sigma_\alpha^2) \quad \text{for} \quad i = 1, \ldots, 17, \tag{20}$$

where the indicator S_i is one if individual i is schizophrenic and zero otherwise. Thus, μ is the overall mean response time in the non-schizophrenic group and β measures the difference in mean response time between the two groups. Still following the assumptions in Gelman and Rubin (1992), we restrict $\lambda \in (0.001, 0.999)$, $\tau > 0$ to identify the model and $\sigma^2, \sigma_\alpha^2 > 0$. For λ, σ_α^2, σ^2, μ and β we adopt independent uniform prior distributions. Gelman and Rubin (1992) also assigns an (improper) uniform prior distribution for τ, but, as also pointed out

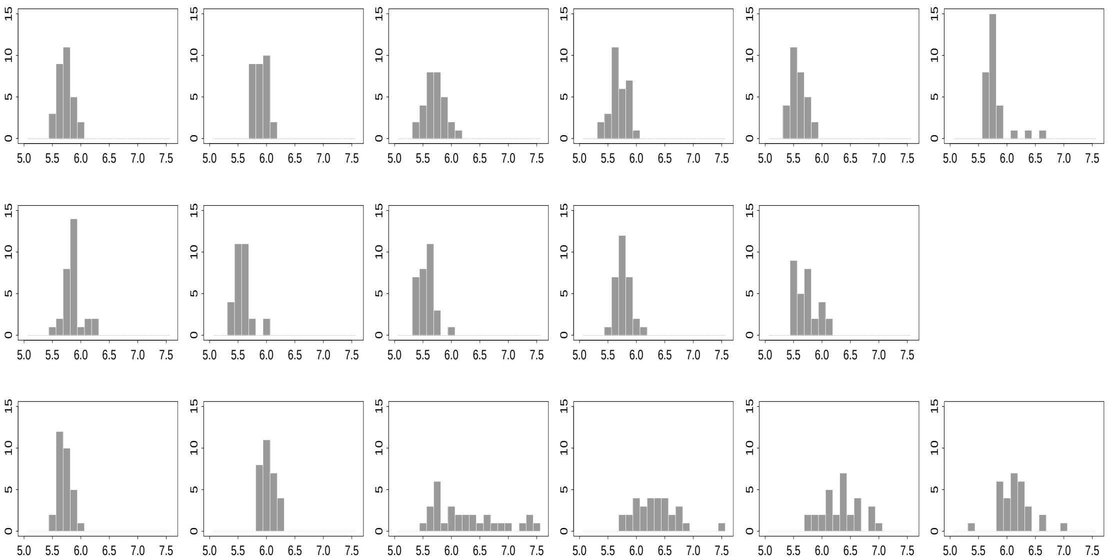


Figure 5: Histograms of log-response times for the eleven non-schizophrenic (to upper rows) and six schizophrenic (lower row) subjects.

in Gelman et al. (1995), this leads to an improper posterior. To avoid this difficulty, we take an exponential prior with mean value ten for τ.

It should be noted that the resulting posterior distribution, $\pi(\alpha, \theta|y)$, contains a singularity. As $\sigma_\alpha^2 \to 0$ and $\alpha_i \to \mu + \beta \cdot S_i$ for $i = 1, \ldots, 17$, the posterior density goes to infinity because $\pi(\alpha_i|\ \theta)$ do so. In the construction of an MCMC sampler this fact requires special attention. In the study of Gelman and Rubin (1992), the singularity is eliminated by a reparameterisation where $\gamma = \ln(\sigma_\alpha^2)$ is used as parameter. However, they seem to have missed that this leads to a mode for large negative values of γ. In fact, the maximum density of this mode is many orders of magnitude higher than the density in any of the three modes reported by them. However, as seen from our simulation results reported below, the probability mass contained in the mode is negligible. This demonstrates the fact also pointed out in Geyer (1992); one can not, as argued in Gelman and Rubin (1992), decide the relative importance of different modes by comparing densities. In stead one must construct an MCMC sampler which is able to jump between modes and observe (after the burn in period) the relative time spent in each mode.

To obtain a target distribution with positive density in all $\Re^{23}$, we reparameterise the model before defining a sampler. We denote the new parameters by (ξ, ε) where $\xi = (\xi_1, \ldots, \xi_6) \in \Re^6$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{17}) \in \Re^{17}$. The relation to the original parameters is given by

$$\lambda = 0.001 + 0.998 \cdot \frac{\exp(\xi_1)}{1 + \exp(\xi_1)}, \tag{21}$$

$\sigma_\alpha^2 = |\xi_2|$, $\sigma^2 = |\xi_3|$, $\tau = |\xi_4|$, $\mu = \xi_5$, $\beta = \xi_6$ and $\alpha_i = \varepsilon_i + \xi_5 + \xi_6 \cdot S_i$ for $i = 1, \ldots, 17$. The logit transformation for λ is the same as adopted by Gelman and Rubin (1992), whereas our transformations for σ_α^2 and σ^2 differ from the log-transformations used there. We also first tried log-transformations for these two parameters, but ran into large numerical problems. With log-transformation for σ_α^2, the singularity at $\sigma_\alpha^2 \to 0$ is transformed to a mode with maximum for a large negative value of $\gamma = \ln(\sigma_\alpha^2)$. Even with computation in double precision, the corresponding $\exp(\gamma)$ was not representable as a number larger than zero. Thus, this parameterisation causes problems when combined with the optimisation in our mode jumping proposals. It should be noted that the absolute values in the transformations for σ_α^2, σ^2 and τ, make each mode of the original $\pi(\theta, \alpha|y)$ to split into 8 modes in $\pi(\xi, \varepsilon|y)$. The transformation of α is to avoid numerical problems close to the singularity. For computation of the density $\pi(\alpha_i|\theta)$ one needs $\alpha_i - (\mu + \beta \cdot S_i)$. When $\alpha_i \approx \mu + \beta \cdot S_i$ this gives severe round off errors.

The Markov chain sampler

To sample from $\pi(\xi, \varepsilon|y)$ we simulate a Markov chain with transition kernel

$$\mathrm{P} = (0.5\mathrm{P}_1 + 0.5\mathrm{P}_2)^{50\cdot 23} \cdot (0.8\mathrm{P}_3 + 0.2\mathrm{P}_4)\,, \tag{22}$$

where P_1 and P_2 are standard Metropolis-Hastings kernels doing local changes, whereas P_3 and P_4 are generalised Metropolis-Hastings kernels of the types discussed in Section 3. Thus, we alternate between $1,150$ local and one global steps, the total of which we call one iteration. To specify the details of $\mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3$ and P_4 we discuss the corresponding proposal kernels.

$\mathbf{P_1}$: Here, we propose to change one component only, randomly chosen among the 23 present. To this component we add a $\mathrm{N}_1(0, 0.01^2)$ term.

$\mathbf{P_2}$: Again only one randomly chosen component is altered. Let ψ denote the old value of this component. The potential new value is sampled from a uniform distribution between $\psi \cdot 0.9$ and $\psi/0.9$. We found it necessary to include such a step to get the chain to mix properly close to the singularity.

$\mathbf{P_3}$: Here, we use proposals via local optimisation as discussed in Section 3.1. The φ is sampled from a Gaussian distribution centered at the origin and a diagonal covariance matrix with all elements equal to 20^2. This very wide distribution for φ is chosen to reduce the risk for missing some modes. We use $T_0(x, \varphi) = x + \varphi$, $T_1(x, \varphi) = x - \varphi$ and, according to the last remark in Section 3.1, we propose the new state from Gaussian or t-distribution with probabilities 0.75 and 0.25, respectively, independently for Q_0^φ and Q_1^φ. For optimisation we again adopt the Newton algorithm with line search, but now with a maximum of 50 iterations to be performed. Thus, the Hessian may not be positive definite. To handle this situation, we use the second approach mentioned in Remark 4, Section 3.1. Obviously, some optimisations end in the singular mode. Then our optimisation algorithm run into problems, a location close to the singularity is returned and the corresponding Hessian matrix is far from positive definite. Even with our fix, all corresponding proposals is in practice rejected. Thus, we define P_4 to handle moves to and from the singular mode.

$\mathbf{P_4}$: This is a transition kernel of the type discussed in Section 3.2. For Q_1^φ we use the same distribution for φ, optimisation procedure and randomly chosen type of proposal distribution as for P_3. The Q_0^φ is given by a density $g(\cdot)$ which we decompose into

$$g(\xi, \varepsilon) = g(\xi_2) \cdot \left[\prod_{i=1}^{17} g(\varepsilon_i|\xi_2)\right] \cdot g(\xi_1, \xi_3, \xi_4, \xi_5, \xi_6|\xi_2, \varepsilon). \tag{23}$$

For $g(\xi_2)$ we use a transformed Cauchy distribution, more specifically we set $\xi_2 = \kappa \cdot V$, where κ is a constant and V is Cauchy distributed. To determine a reasonable value for κ

we explored the singular mode by starting a local-move Markov chain close to the singularity. Based on results from this preliminary run we fixed $\kappa = 10^{-9}$. Conditional to ξ_2 we draw the potential new ε_i's according to $g(\varepsilon_i|\xi_2) = \pi(\varepsilon_i|\xi_2) = \mathrm{N}_1(0, |\xi_2|)(\varepsilon_i)$. To obtain a reasonable proposal distribution for the remaining five parameters, $g(\xi_1, \xi_3, \xi_4, \xi_5, \xi_6|\xi_2, \varepsilon)$, we use an optimisation scheme similar to what is used for Q_0^{φ}, i.e. draw initial values for the five parameters independently from a $\mathrm{N}_1(0, 20^2)$, optimise $\pi(\xi, \varepsilon|y)$ with respect to the five parameters and use the result to define a five-dimensional expectation vector and covariance matrix. Again we generate proposals from a Gaussian or a t-distribution with probabilities 0.75 and 0.25, respectively.

Simulation results

We explore the posterior distribution by simulating the above defined Markov chain. To monitor the convergence, standard output analysis (e.g. Ripley, 1987, Ch. 6) is used. We plot values of the 23 parameters and the corresponding energy, $U(\xi, \varepsilon)$, against the number of iterations and use stability of the graphs as an indication of convergence. We do several runs and, to check that the sampler does not risk being trapped in local modes, we use three strategies for assigning initial values to the parameters: 1) Draw the values independently from $\mathrm{N}_1(0, 20^2)$. 2) Locate a mode by first drawing values independently from $\mathrm{N}_1(0, 20^2)$ for each parameter and, starting at this location, carry out an optimisation. Thereafter perform $5,000 \cdot 23$ local updates with $0.5\mathrm{P}_1 + 0.5\mathrm{P}_2$ to get initial values distributed approximately according to the mode found. 3) To get initial values in the singular mode, draw values for ξ_2 and ε and perform optimisation for the remaining five parameters as in the definition for $g(\cdot)$ above. Again perform $5,000 \cdot 23$ local updates to obtain values approximately as given by the singular mode.

We performed 50 runs of 500 iterations, each time drawing initial values from 1), 2) or 3) above with probability 1/3 for each. Figure 6 shows the energy and values for ξ_1 and ξ_4 for three of the runs. The remaining 47 runs had a similar behaviour. In all 50 runs, the energy settled on approximately the same level after less than 100 iterations. A closer inspection of the simulated parameter values revealed also that all runs settled in the same mode of $\pi(\theta, \alpha|y)$. But remember that this corresponds to eight modes in $\pi(\xi, \varepsilon|y)$, from which we actually sample. Jumps between these modes continue to occur throughout our runs as they should with a well-mixing sampler, see the displayed ξ_4-values in Figure 6. The behaviour of the displayed run started in the singular mode (lower row in Figure 6) should also be noted. The chain leaves the singular mode after a few iterations and in none of the simulated chains a move back to this mode ever occurs. In the displayed run the chain leaves the singular mode via another mode of negligible importance, but in other runs also direct jumps to the major mode occur. A better understanding of how the sampler behave in the 23 dimensional parameter space can be gained by studying 2D projections of the trajectories. For the same three runs displayed in Figure 6, Figure 7 shows the simulated paths in the ξ_1—ξ_2, ξ_1—ε_2 and ξ_2—ε_2 planes. In the upper row of Figure 7, we see that mode jumping proposals enable large direct jumps from tails to a high density mode. A closer inspection of the proposal densities used reveal that such steps are, in practice, only accepted when $q_{1-i}^{\varphi}(y|x)$ is chosen to be a t-distribution, as otherwise the acceptance probability becomes extremely small. In contrast, direct jumps from one mode to another most frequently occur if both $q_0^{\varphi}(y|x)$ and $q_1^{\varphi}(y|x)$ are Gaussian distributions. Studying the left and right columns in Figure 7, one must remember that our target distribution is symmetric around $\xi_2 = 0$. From the plots we observe that the

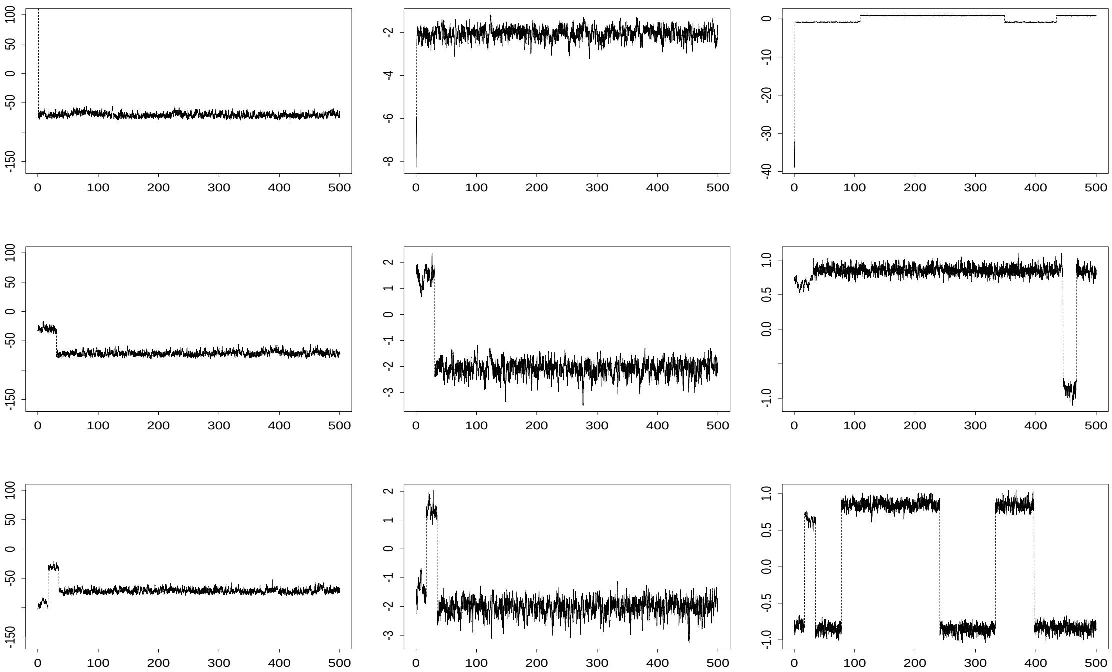


Figure 6: Plots of the energy (left), ξ_1 (middle) and ξ_4 (right) as a function of the number of iterations. The three runs used initial values generated according to 1) (upper), 2) (middle) and 3) (lower), respectively. Dashed lines are used to indicate global moves by P_3 or P_4. Note the difference in vertical scale.

major mode of $\pi(\theta, \alpha|y)$ is sufficiently close to $\xi_2 = 0$ to allow also local jumps across this line of symmetry. This is in contrast to the situation for ξ_4, see the left column in Figure 6. The run showed in the lower row of Figure 7 starts in the singular mode. This mode is extremely narrow in the ξ_2 and ε_i's directions, so in the plots we see this only as horizontal lines in the left and middle plots, whereas in the right plot it appears as just a point at the origin.

In addition to the shorter runs discussed above, we made one long run of $50,000$ iterations, from which we estimated posterior quantiles for each of the (original) parameters. The results were as expected from the preliminary runs. The major mode (of $\pi(\theta, \alpha|y)$) was found after a few iterations and thereafter never left. The estimated quantiles were in accordance with the figures in Table 2 in Gelman and Rubin (1992).

5 Closing Remarks

The paper defines a generalised Metropolis-Hastings algorithm and demonstrates how this allow local optimisation be to incorporated in MCMC procedures. Recipes for handling multi-modal distributions are given and their potential is illustrated in two examples.

In the present paper we have limited the attention to continuous distributions because the mode jumping proposals are not directly applicable for discrete distributions. First, local optimisation is less efficient in a discrete space. Second, the approach requires a class of suitable proposal distributions, from which direct sampling should be possible and for which

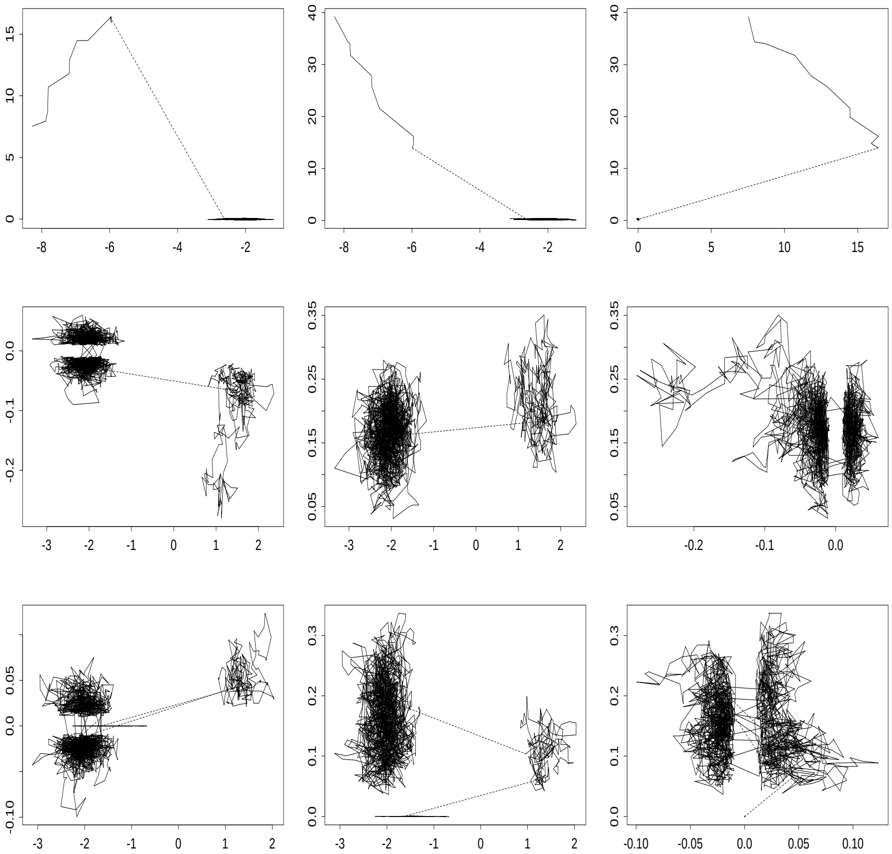


Figure 7: Simulated trajectories (first 200 iterations) projected into the ξ_1—ξ_2 (left), ξ_1—ε_2 (middle) and ξ_2—ε_2 (right) planes, respectively. The three runs (one in each row) are the same as displayed in Figure 6. Dashed lines are used to indicate global moves by P_3 or P_4. Note the difference in scale.

the normalizing constant should be available. Several satisfying continuous distributions can be found, but no clear discrete candidate exists. However, for a distribution with a mixture of many continuous and a few discrete components the natural choice is to propose values for the discrete components from a very simple, may be uniform, distribution and use optimisation to generate candidate values for the remaining variables.

Until recently, a mare in the application of Markov chain simulations has been the question of when the runs have attained convergence. For a steadily increasing class of algorithms the solution has been found in the use of coupling from the past, see Propp and Wilson (1996, 1998) and references therein. To achieve the full potential of our algorithm, also mode jumping proposals should be fit into a perfect simulation scheme. Murdoch and Green (1998) discuss coupling from the past for continuous target distributions, but their setup is not directly applicable in our case.

We have motivated the introduction of mode jumping proposals by the need in multi-modal distributions. However, if the target density is computationally intensive and contain no Markov property that can be exploited in the calculation of acceptance probabilities, large moves in the parameter space become essential for fast convergence and we expect a combination of multi-grid ideas and mode jumping proposals to prove valuable even for unimodal distributions. Work along this course is currently in progress for the so called history matching problem encountered in evaluation of ground water and petroleum reservoirs (Sun, 1994; Hegstad, 1997) and shows promising results.

Acknowledgements

We would like to thank Henning Omre, Geoff Nicholls and Håvard Rue for fruitful discussions.

References

Belin, T.R. and Rubin, D.B. (1995). "The analysis of repeated-measures data on schizophrenic reaction times using mixture models", *Statistics in Medicine,* **14**, 747-768.

Besag, J. and Green, P.J. (1993). "Spatial statistics and Bayesian computation", *J. Royal Statist. Soc. B,* **55**, 3-23.

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995). "Bayesian computation and stochastic systems (with discussion)", *Statistical Science*, **10**, 3-66.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian data analysis*, Chapman Hall, London.

Gelman, A. and Rubin, D.B. (1992). "Inference from iterative simulation using multiple sequences", *Statistical Science,* **7**, 457-511.

Geyer, C.J. (1991). "Markov chain Monte Carlo maximum likelihood", in Keramidas, E.M. (ed.) *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, Interface Foundation, 156-163.

Geyer, C.J. (1992). "Practical Markov chain Monte Carlo", *Statistical Science,* **7**, 473-511.

Geyer, C.J. and Thompson, E.A. (1995). "Annealing Markov chain Monte Carlo with applications to ancestral inference", *J. Amer. Statist. Assoc.,* **90**, 909-920.

Goodman, J. and Sokal, A.D. (1989). "Multigrid Monte Carlo method. Conceptual foundations", *Phys. Rev. D*, **40**, 2035-2071.

Hegstad, B.K. (1997). "Sampling from stochastic reservoir models constrained by production data", Dr.ing. thesis, 1997:65, Norwegian University of Science and Technology, Trondheim, Norway.

Johnson, N.L. and Kotz, S. (1972). *Distributions in statistics: Continuous multivariate distributions*, Wiley, New York.

Luenberger, D.G. (1984). *Linear and nonlinear programming*, Addison-Wesley, Reading, Massachusetts.

Marinari, E. and Parisi, G. (1992). "Simulated tempering: a new Monte Carlo scheme", *Europhys. Lett.*, **19**, 451-458.

Murdoch, D.J. and Green, P.J. (1998). "Exact sampling from a continuous state space", *Scand. J. of Statistics,* **25**, 483-502.

Neal, R. M. (1996). "Sampling from multimodal distributions using tempered transitions", *Statistics and Computing,* **6**, 353-366.

Propp, J.G. and Wilson, D.B. (1996). "Exact sampling with coupled Markov chains and applications to statistical mechanics", *Random Structures and Algorithms,* **9**, 223-252.

Propp, J.G. and Wilson, D.B. (1998). "How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph", *Journal of Algorithms,* **27**, 170-217.

Ripley, B.D. (1987). *Stochastic Simulation.* Wiley, New York.

Sun, N.Z. (1994). *Inverse problems in groundwater modeling*, volume 6 of *Theory and applications of transport in porous media*, Kluwer, Dordrecht, The Netherlands.

Håkon Tjelmeland, Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7034 TRONDHEIM, Norway.