On the flexibility of Metropolis-Hastings acceptance probabilities in auxiliary variable proposal generation

Geir Storvik

Department of Mathematics and (sfi)[2] Statistics for Innovation, University of Oslo, Norway.
February 25, 2009

Summary. Use of auxiliary variables for generating proposal variables within a Metropolis-Hastings setting has been suggested in many different settings. This has in particular been of interest for simulation from complex distributions such as multimodal distributions or in transdimensional approaches. For many of these approaches, the acceptance probabilities that are used turn up somewhat magic and different proofs for their validity have been given in each case.

In this paper I will present a general framework for construction of acceptance probabilities in auxiliary variable proposal generation. In addition to demonstrate the similarities between many of the proposed algorithms in the literature, the framework also demonstrate that there is a great flexibility in how to construct such acceptance probabilities, in addition to the flexibility in how to construct the proposals. With this flexibility, alternative acceptance probabilities are suggested. Some numerical experiments are also reported.

1. Introduction

Many approaches for constructing efficient sampling algorithms are based on the use of auxiliary variables. This is particularly the case within the class of Markov Chain Monte Carlo (MCMC) algorithms (Metropolis et al., 1953; Hastings, 1970; Gilks et al., 1996; Robert and Casella, 2004) but has also been considered in e.g. importance sampling (IS) (Neal, 2001). The most common way of doing this is to extend the sample space with some extra variable, in order to construct simpler or more efficient algorithms in this extended space (Tanner and Wong, 1987; Edwards and Sokal, 1988; Besag and Green, 1993; Higdon, 1998; Damien et al., 1999). More recently, auxiliary variables have been used as a tool within a Metropolis-Hastings setting for either generating better proposals (e.g. Tjelmeland and Hegstad, 2001; Jennison and Sharp, 2007) or for calculation of acceptance probabilities (e.g. Beaumont, 2003; Andrieu and Roberts, 2009).

The flexibility of choosing proposal distributions has long been recognised. In this paper we take a different look at the use of auxiliary variables in that we discuss the simultaneous flexibility in choosing target distributions. Assume our aim is to sample $y^* \sim \pi(\cdot)$. A proposal is generated by first simulating $x^* \sim q_x(\cdot)$ followed by $y^* \sim q_{y|x}(\cdot|x^*)$. The flexibility follows in that the target distribution, $\bar{\pi}$ say, for the combined variables (x^*, y^*) can be any distribution with π as its marginal distribution. For a given set of proposal distributions $q_x, q_{y|x}$, different acceptance probabilities can be obtained by different choices of $\bar{\pi}$. By taking this alternative viewpoint, a common understanding of different algorithms suggested in the literature is obtained, and also serves as a toolbox for suggesting new algorithms. This toolbox will in particular be useful in cases where standard MCMC algorithms fail and more complicated versions are of need. Typical examples are algorithms

for jumping between possible modes (Tjelmeland and Hegstad, 2001; Jennison and Sharp, 2007) and reversible jump algorithms (Green, 1995, 2001).

An important special class of algorithms is where $x^* = x_{1:t}^*$ is generated sequentially, typically with x_1^* generated through a big jump (for jumping between modes or dimensions) followed by a few steps of smaller jumps. For most such algorithms suggested in the literature, acceptance ratios within an Metropolis-Hastings (MH) setting are typically either only depending on the first x_1^* (e.g. Al-Awadhi et al., 2004) or the last x_t^* (e.g. Tjelmeland and Hegstad, 2001; Jennison and Sharp, 2007). By combining different target distributions, acceptance ratios depending on averages over all generated $x_1^*, ..., x_t^*$ can be constructed, giving higher and more stable acceptance probabilities.

Other algorithms that can be considered as special cases of this general framework is annealed importance sampling (Neal, 2001), mode or model jumping (Tjelmeland and Hegstad, 2001; Jennison and Sharp, 2007; Al-Awadhi et al., 2004), multiple-try methods (Liu et al., 2000), proposals based on particle filters (Andrieu et al., 2008), pseudo-marginal algorithms (Beaumont, 2003), sampling algorithms for distributions with intractable normalising constants (Møller et al., 2006) and delayed rejection sampling (Tierney and Mira, 1999; Mira, 2001; Green and Mira, 2001).

We start in Section 2 by discussing the flexibility of weight functions in auxiliary importance sampling. This setting is of interest in itself but can also be seen as a starting point for discussing auxiliary MH algorithms because MH acceptance probabilities are directly related to the importance weights used in importance sampling. Such MCMC algorithms are considered in Section 3. In Section 4 we focus on sequential generations of auxiliary variables, while in Section 5 we apply our general results to many algorithms suggested in the literature and also consider alternative versions of these. Although the main motivation for this paper is the construction of a toolbox for construction of MCMC algorithms, in Section 6 we consider some numerical experiments, demonstrating that alternative weights and acceptance probabilities can improve the performance of an algorithm. We conclude the paper by some final remarks and discussion in Section 7.

2.   Auxiliary importance sampling

For simplicity we will assume $\pi(y)$ is a distribution in some continuous space $\mathcal{R}^m$ with full support although the results are applicable to more general situations.

For consistent notation with MCMC sampling considered in the next section, we will use x^* for the generated auxiliary variable(s) and y^* for the proposed variable. The idea of auxiliary importance sampling is to assume that y^* is generated trough first simulating x^* from some distribution $q_x(\cdot)$ and thereafter generating y^* from the conditional distribution $q_{y|x}(\cdot|x^*)$. The auxiliary variable x^* might be a single variable or a sequence of variables $x^* = (x_1^*, ..., x_t^*)$ (the latter being discussed in section 4). Most applications of this idea concentrate on the flexibility of choosing $q_x(\cdot)$ and $q_{y|x}(\cdot|x^*)$. The following result shows that there also is a flexibility in choosing importance weights in this situation:

Proposition 1
Assume $x^* \sim q_x(\cdot)$ and $y^*|x^* \sim q_{y|x}(\cdot|x^*)$. Define $\mathcal{S} = \{(x, y) : q_x(x)q_{y|x}(y|x) > 0\}$ and assume the marginal support for $q_y(y) = \int_x q_x(x)q_{y|x}(y|x)dx$ includes the support of $\pi(y)$. Let $h(x|y)$ be any distribution such that $\mathcal{T} = \{(x, y) : \pi(y)h(x|y) > 0\}$ is a subset of $\mathcal{S}$.

Then for any measurable function $g(y)$

$$E[w(x^*, y^*)g(y^*)] = \int_y \pi(y)g(y)dy \equiv E^\pi[g(y)] \tag{1}$$

where

$$w(x^*, y^*) = \frac{\pi(y^*)h(x^*|y^*)}{q_x(x^*)q_{y|x}(y^*|x^*)}. \tag{2}$$

PROOF. We have

$$\begin{aligned}
E[w(x^*, y^*)g(y^*)] &= \int_{(x,y)\in\mathcal{S}} \frac{\pi(y)h(x|y)}{q_x(x)q_{y|x}(y|x)} q_x(x)q_{y|x}(y|x)g(y)dxdy \\
&= \int_y \pi(y)g(y) \int_{x:(x,y)\in\mathcal{S}} h(x|y)dxdy \\
&= \int_y \pi(y)g(y)dy
\end{aligned}$$

where in the last equation we have used that $\mathcal{T}$ is a subset of $\mathcal{S}$. $\qquad\square$

This result shows that for $(x_i^*, y_i^*), i = 1, ..., N$ all independent and generated by $x_i^* \sim q_x(\cdot)$ and $y_i^*|x_i^* \sim q_{y|x}(\cdot|x_i^*)$,

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} w(x_i^*, y_i^*)g(y_i^*)$$

will be an unbiased estimate for θ.

For the special case $h(x|y) = q_x(x)$, the weight reduces to

$$w(x^*, y^*) = \frac{\pi(y^*)}{q_{y|x}(y^*|x^*)}, \tag{3}$$

that is only the transition probability $q_{y|x}(y^*|x^*)$ is involved. Note however that this choice is only legal if $q_{y|x}(y^*|x^*)$ has full support.

Different choices of h correspond to the fact that (x^*, y^*) can follow many different simultaneous distributions giving the same *marginal* distribution of y^* while $q_x(x^*)q_{y|x}(y^*|x^*)$ corresponds to the usual proposal distributions in importance sampling. The ability of playing around with different choices of h in addition to q_x and $q_{y|x}$ will be important in the following.

Note that while for proposal distributions, we need to choose the support large enough, for the h distribution we must be sure that the support is small enough. In many cases neither of these restrictions will cause any problems in that both the proposal and the h distribution will have "full" support, but in some situations the proposal distributions will make restrictions on S in which case some care must be taken in the construction of h.

For specific choices of g, optimal choices of proposal distributions exist. In cases where expectations with respect to many g-functions are to be calculated simultaneously, simultaneous optimal choices are not possible to obtain. In such cases, the variances of the weight functions are reasonable global measures to consider. The unconditional expectation of $w(x^*, y^*)$ is 1, as in ordinary importance sampling. More of interest is the properties of $w(x^*, y^*)$ conditional on y^*, given in the next proposition:

Proposition 2
Under the assumptions of Proposition 1,

$$\mathsf{E}[w(x^*, y^*)|y^*] = \frac{\pi(y^*)}{q_y(y^*)}, \tag{4}$$

where $q_y(y)$ is the marginal proposal distribution as given in Proposition 1, and

$$\mathsf{Var}[w(x^*, y^*)] \geq \mathsf{Var}\left[\frac{\pi(y^*)}{q_y(y^*)}\right]. \tag{5}$$

PROOF. We have

$$\mathsf{E}[w(x^*, y^*)|y^*] = \int_{x:(x,y)\in\mathcal{S}} \frac{h(x|y)\pi(y)}{q_x(x)q_{y|x}(y|x)} \frac{q_x(x)q_{y|x}(y|x)}{q_y(y)} dx$$
$$= \frac{\pi(y)}{q_y(y)} \int_{x:(x,y)\in\mathcal{S}} h(x|y) dx = \frac{\pi(y)}{q_y(y)}.$$

Further,

$$\mathsf{Var}[w(x^*, y^*)] = \mathsf{Var}[\mathsf{E}[w(x^*, y^*)|y^*]] + \mathsf{E}[\mathsf{Var}[w(x^*, y^*)|y^*]] \geq \mathsf{Var}\left[\frac{\pi(y^*)}{q_y(y^*)}\right]. \qquad \square$$

Given that small variance of the importance weights is a desirable property, using the importance weights given in (4) would have been a good strategy, if they were possible to calculate. The weight $w(x, y)$ can be seen as an unbiased estimate of that "optimal" importance weight function.

The variance of the weight function in (2) is higher than the variance of the ordinary importance weight function. The challenge is therefore to choose weight functions or equivalently the conditional distributions $h(x|y)$ such that the variability do not increase too much.

Note that given a set of weight functions, all satisfying (4), also weighted averages of these weight functions will satisfy (4). This will be useful in the following.

3.   Weight functions within Metropolis-Hastings algorithms

The flexibility of weight functions are directly transferable to a flexibility in acceptance probabilities in MH algorithms. Our aim in this case is to generate a sequence of variables having invariant distribution $\pi(y)$. Our typical situation will be that a new y^* is generated through an auxiliary x^* with proposal density $q_x(x^*|y)q_{y|x}(y^*|y, x^*)$. We will however consider a slightly more general scheme allowing for more general simulations of the pair (x^*, y^*). In the following we will for weight functions, acceptance ratios and acceptance probabilities list the variables involved in the order they are generated with semicolon separating variables that have been generated at the previous iteration,

The following results consider the situation where the auxiliary variables are stored at each iteration. In order to get full generality, we allow the proposals to depend on the previous x as well.

Proposition 3
Assume $(x^*, y^* \sim q(x^*, y^*|x, y)$. Define

$$w(y; x^*, y^*) = \frac{\pi(y^*)h(x^*|y^*)}{q(x^*, y^*|x, y)}$$

where $h(x^*|y^*)$ is an arbitrary conditional distribution. Then an acceptance probability of the form

$$\alpha(x, y; x^*, y^*) = \min\{1, r(x, y; x^*, y^*)\}\}$$

where the acceptance ratio is given by

$$r(x, y; x^*, y^*) = \frac{w(y; x^*, y^*)}{w(y^*; x, y)} \tag{6}$$

defines an MH algorithm with invariant distribution $\pi(y)h(x|y)$ and marginal distribution $\pi(y)$.

PROOF. The ordinary MH acceptance ratio for (x^*, y^*) with $\pi(y)h(x|y)$ as target distribution and $q(x^*, y^*|x, y)$ as proposal distribution is

$$r(x, y; x^*, y^*) = \frac{\pi(y^*)h(x^*|y^*)q(x, y|x^*, y^*)}{\pi(y)h(x|y)q(x^*, y^*|x, y)} = \frac{w(y; x^*, y^*)}{w(y^*; x, y)}.$$

Since $\pi(y)$ is the marginal distribution for y of $\pi(y)h(x|y)$, the result follows. ☐

In this case there is no restriction on that $\mathcal{S}(y) = \{(x^*, y^*) : q(x^*, y^*|x, y) > 0\}$ should cover $\mathcal{T}$, the support of $h(x^*|y^*)\pi(y^*)$. However, if the integral

$$E[w(y; x^*, y^*)|y] = \int_{(x^*, y^*) \in \mathcal{S}(y)} h(x^*|y^*)\pi(y^*)dx^*dy^*$$

vary heavily with y, the ratio (6) can be far from one, making moves more difficult to achieve. Note in particular that in order for the Markov chain to converge properly, also irreducibility with respect to the target distribution need to be fulfilled.

Proposition 3 can be generalised in different ways. Other types of acceptance probabilities as discussed in Hastings (1970) could also be considered. Similar to the importance weights discussed in the previous section, also combinations of weights can be inserted into the acceptance probabilities.

In the case of $q(x^*, y^*|x, y) = q_x(x^*|y^*)q_{y|x}(y^*|y, x^*)$, similar to the simplifications that leads to (3) we can assume $h(x^*|y) = q_x(x^*|y) = q_x(x^*)$ in which case the weight function reduces to

$$w(y; x^*, y^*) = \frac{\pi(y^*)}{q_{y|x}(y^*|y, x^*)}. \tag{7}$$

The setting described above assumes the auxiliary variable x generated in the previous iterations is stored. An alternative version is to assume a new x is generated in a "reverse" proposal. Andrieu and Roberts (2009) show that generating new x's at each iteration rather than reusing those generated at the previous iteration can improve the acceptance rates. The following proposition consider this setting.

Proposition 4

Assume $(x^*, y^*) \sim q(x^*, y^*|y)$. Assume further $x \sim h(x|y, x^*, y^*)$ where $h(x|y, x^*, y^*)$ is an arbitrary distribution. Then an acceptance ratio of the form

$$r(y; x^*, y^*, x) = \frac{w(y; x^*, y^*, x)}{w(y^*; x, y, x^*)} \tag{8}$$

where

$$w(y; x^*, y^*, x) = \frac{\pi(y^*)h(x^*|y^*, x, y)}{q(x^*, y^*|y)} \tag{9}$$

defines a Markov Chain Monte Carlo algorithm with invariant distribution $\pi(y)$.

PROOF. Write $q(x^*, y^*|y) = q_x(x^*|y)q_{y|x}(y^*|y, x^*)$ and define $\bar{\pi}(y, x^*) = \pi(y)q_x(x^*|y)$. Assume $y \sim \pi(y)$. Since $x^* \sim q_x(x^*|y)$, this imply $(y, x^*) \sim \bar{\pi}$. Now consider the generated (y^*, x) as a new proposal in an MH setting. Then the ordinary acceptance ratio for (y^*, x) is

$$\frac{\pi(y^*)q_x(x|y^*)q_{y|x}(y|x, y^*)h(x^*|y^*, x, y)}{\pi(y)q_x(x^*|y)q_{y|x}(y^*|x^*, y)h(x|y, x^*, y^*)} = r(y; x^*, y^*, x)$$

showing that the Markov chain is invariant with respect to $\bar{\pi}(y, x^*)$ of which $\pi(y)$ is the marginal.   □

In this case, the h function is involved both in the generation of x and in the acceptance probability, but not in the proposal of y^*. Some special cases are of general interest.

When $h(x|y, x^*, y^*) = q_x(x|y^*)$, the acceptance ratio reduces to

$$r(y; x^*, y^*, x) = \frac{\pi(y^*)q_{y|x}(y|y^*, x)}{\pi(y)q_{y|x}(y^*|y, x^*)}, \tag{10}$$

that is we obtain a similar simplification as (3) and (7). If x^* and x are low-dimensional, this choice can work reasonable well. If however x^* is generated as a sequence of auxiliary variables (Section 4), $q_{y|x}(y|y^*, x)$ can be very small giving low acceptance probability.

We also have the following result, which corresponds to Proposition 2:

Proposition 5

Under the assumptions of Proposition 4,

$$\mathsf{E}[r(y; x^*, y^*, x)|y, y^*] = \frac{\pi(y^*)q_y(y^*|y)}{\pi(y)q_y(y^*|y)} \equiv r(y; y^*), \tag{11}$$

where $q_y(y^*|y)$ is the marginal proposal density for y^* generated through x^*, and

$$\mathsf{Var}[r(y; x^*, y^*, x)] \geq \mathsf{Var}[r(y, y^*)]. \tag{12}$$

PROOF. Define $q_{x|y}(x^*|y, y^*)$ to be the conditional distribution of x^* given both y and y^*, that is

$$q_{x|y}(x^*|y, y^*) = \frac{q(x^*, y^*|y)}{q_y(y^*|y)}.$$

Then

$$E[r(y; x^*, y^*, x)|y, y^*] = \int_{x^*} \int_x r(y; x^*, y^*, x) q_{x|y}(x^*|y, y^*) h(x|y, x^*, y^*) dx dx^*$$

which by direct insertion of the definition of $r(y; x^*, y^*, x)$ reduces to

$$\int_{x^*} \int_x \frac{\pi(y^*) q_y(y|y^*)}{\pi(y) q_y(y^*|y)} q_{x|y}(x|y^*, y) h(x^*|y^*, x, y) dx dx^* = \frac{\pi(y^*) q_y(y|y^*)}{\pi(y) q_y(y^*|y)}$$

proving (11). The second part is proved similarly to the second part of Proposition 2. □

Note that since $\min\{1, x\}$ is a concave function, by Jensen's inequality,

$$E[\alpha(y; x^*, y^*, x)|y, y^*] \leq \min\{1, r(y; y^*)\}$$

showing that the optimal acceptance ratio (in the Peskun (1973) sense) would be to use $r(y; y^*)$. In practise $q_y(y^*|y)$ will not be possible to evaluate but by clever choices of h, we hopefully get close to this.

The approach in Proposition 4 assumes auxiliary variables generated both forwards and backwards. Some algorithms suggested in the literature (e.g. Al-Awadhi et al., 2004) only consider generation forwards while the same variables are used backwards. This can be obtained by choosing h to be a distribution giving probability one to a specific value of x (typically dependent on (y^*, x^*, y)). The most direct approach is to choose $h(x|y, x^*, y^*) = \delta(x - x^*)$ where δ is Dirac's delta distribution in which case the acceptance ratio becomes

$$r(y; x^*, y^*, x) = \frac{\pi(y^*) q_x(x|y^*) q_{y|x}(y|y^*, x^*)}{\pi(y) q_x(x^*|y) q_{y|x}(y^*|y, x^*)}.$$

This choice shares the same weaknesses as (10). If $x^* = x^*_{1:t} = (x^*_1, ..., x^*_t)$ is generated in sequence, an alternative is to assume $x_{1:t} = x^*_{t:1}$ with probability one. For such a choice

$$r(y; x^*, y^*, x) = \frac{\pi(y^*) q_x(x^*_{t:1}|y^*) q_{y|x}(y|y^*, x^*_{t:1})}{\pi(y) q_x(x^*_{1:t}|y) q_{y|x}(y^*|y, x^*_{1:t})}.$$

In this case both $q_x(x^*_{t:1}|y^*)$ and $q_{y|x}(y|y^*, x^*_{t:1})$ should be reasonable large. We will explore these possibilities further in Section 4 for specific choices of proposal distributions.

Note that even though a choice of h making x a deterministic function of x^* give a proposal distribution not moving around in the full state space of $\bar{\pi}(y, x^*) = \pi(y) q_x(x^*|y)$, this will typically not cause problems because this MH step is combined with the generation of x^*.

4. Sequential generation of auxiliary variables

In this section we will discuss the use of different importance weights in the case where x is generated through a sequence of steps. Such schemes have been considered e.g. by Neal (2001) in his annealed importance sampling scheme (to be discussed further in Section 4.4), by Jennison and Sharp (2007) who proposed a method for mode jumping by first applying one big jump and thereafter several smaller moves and by Al-Awadhi et al. (2004) within a reversible jump MCMC framework. The weight functions obtained in this section will both be of interest in themselves, but will also be used as building blocks for more advanced algorithms in Section 5.

4.1. A general approach

For ease of notation, let $x_0^* \equiv y$ and $x_{t+1}^* \equiv y^*$. Assume $x_i^* \sim q_i(\cdot | x_{i-1}^*)$ for $i = 1, ..., t+1$. The typical setup we will consider is where q_1 corresponds to a large jump while the following proposal densities satisfy the detailed balance criterion

$$\pi(x) q_i(y|x) = \pi(y) q_i(x|y). \tag{13}$$

A consequence of such an assumption is that

$$\frac{\prod_{i=s}^{t} q_{i+1}(x_i^* | x_{i+1}^*)}{\prod_{i=s}^{t} q_{i+1}(x_{i+1}^* | x_i^*)} = \frac{\pi(x_s^*)}{\pi(x_{t+1}^*)}, \tag{14}$$

which will be used repeatedly.

Our aim is to construct acceptance probabilities for the proposal $y^* = x_{t+1}^*$ when y is the current state. A complicating factor is that the sequential approach for generating y^* will not make the proposal distribution for y^* given y directly available. Note that if q_1 do not depend on y, we obtain an independence sampler where the proposal is generated by a sequence of internal MCMC steps.

Writing $x_{1:t+1}^* = (x_1^*, ..., x_{t+1}^*)$, the general weight function will have the form

$$w(y; x_{1:t+1}^*, x) = \frac{\pi(x_{t+1}^*) h(x_{1:t}^* | x_{t+1}^*, x, y)}{\prod_{i=1}^{t+1} q_i(x_i^* | x_{i-1}^*)}.$$

A wide variety of weight functions can be considered by different choices of h. Consider the class of distributions

$$h_s(x_{1:t} | x_{t+1}) = \prod_{i=1}^{s-1} q_i(x_i | x_{i-1}) \prod_{i=s}^{t} q_{i+1}(x_i | x_{i+1}) \tag{15}$$

and assume that for $i > s$, the detailed balance criterion (13) is satisfied. For the setup of Propositions 1 and 3 we must assume $q_1(x_1^* | y) = q_1(x_1^*)$, which is no real restriction in the importance sampling setting but *is* a restriction in the MH setting. In the latter case if the generation of x_1^* is depending on y, a $\tilde{q}_1(x_1)$ could be used in (15) above. Such proposals could for instance be used for target distributions with several modes where q_1 corresponds to a large jump while the subsequent moves are ordinary MCMC moves (Tjelmeland and Hegstad, 2001; Jennison and Sharp, 2007). This particular setting will be further discussed in Section 4.2. Another possibility is model or dimension moves in a reversible jump setting, see Section 4.3. Allowing the additional moves also to depend on i gives the possibilities for different blocks of a high-dimensional state vector to be updated at different steps.

Under the assumptions above, from (14), we obtain

$$w(y; x_{1:t+1}^*) = \frac{\pi(x_{t+1}^*) \prod_{i=s}^{t} q_{i+1}(x_i^* | x_{i+1}^*)}{q_s(x_s^* | x_{s-1}^*) \prod_{i=s+1}^{t+1} q_i(x_i^* | x_{i-1}^*)} = \frac{\pi(x_s^*)}{q_s(x_s^* | x_{s-1}^*)}. \tag{16}$$

The weights only depend on the ratio between the marginal distributions and the transition kernels for different components of the auxiliary variables. Note that the proposal distributions $q_i, i \leq s$ can in fact be arbitrary. The special case $s = 1$ gives

$$w_1(y; x_{1:t+1}) = \frac{\pi(x_1^*)}{q_1(x_1^* | y)}, \tag{17}$$

which corrects for using q_1 for drawing the first sample but then utilises that the following samples keeps the distribution invariant with respect to π. This weight function do however not account for that subsequent samples will be closer to π.

Another special case is $s = t + 1$ which gives weight

$$w_t(y; x_{1:t+1}) = \frac{\pi(y^*)}{q_{t+1}(y^*|x_t^*)}, \tag{18}$$

only depending on the density of and the transition to y^*. Note that in this case no assumptions on q_i are made for any i although $q_{t+1}(y^*, x_t^*)$ should be possible to compute. A weakness here is that when t is large we would expect the marginal distribution of y to be close to $\pi(y^*)$, but this is not accounted for. This can however be accomplished by *combining* the weights. Since weighted averages of weight functions are allowed, a proper weight function is given by

$$w(y; x_{1:t+1}^*) = \sum_{s=1}^{t+1} a_s w_s(y; x_{1:t+1}^*), \tag{19}$$

A particular interesting case is when $a_s = (t + 1 - b)^{-1} I(s \geq b)$, with b corresponding to some "burn-in" period for the proposal generation. In this case, $w(y; x_{1:t+1}^*)$ is a standard MCMC estimate of $E[w(y; x_{1:t+1}^*)]$ with the first b iterations used as burn-in. If further requirements for the Markov chain using $q_b, ..., q_{t+1}$ as transition probabilities to be ergodic is satisfied, the weight function will then converge towards its expectation which is 1, thereby implying that the acceptance probability also converges towards 1. This indicates that the weight function indeed reflects that x_{t+1}^* converges in distribution to $\pi(\cdot)$ as t grows.

In some cases, updates are made component-wise (or block-wise). Similar to the non-reversibility of the Gibbs sampler, some extra care need to be taken under systematic scan updates. The weights (16) can however still be shown to be valid.

In order to use the weight functions from the previous section, transition probabilities of the form $q_s(x_s^*|x_{s-1}^*)$ must be possible to calculate. When $\{x_{1:t+1}^*\}$ is generated through internal MH steps, such transition probabilities are not always directly available (this is in particular the case when an internal proposal is not accepted). There are different ways around this. The details on this is considered in Appendix A.

4.2. Mode jumping

Jennison and Sharp (2007) considered an approach for performing mode jumping within MCMC. Their approach was based on starting with a large jump followed by a sequence of (typically smaller) MCMC steps. To be specific, $x_1^* = y + \phi$ where $\phi \sim f(\cdot)$ is a symmetric distribution while $x_t^* \sim q(x_t^*|x_{t-1}^*)$ for $t = 2, ..., t$ and finally $y^* \sim q_{y|x}(y^*|x_t^*)$. Here q is a transition kernel leaving π invariant. A different transition at the last step is chosen for computational reasons, see below.

By choosing

$$h(x|y, x^*, y^*) = \delta(x_1 - y^* + \phi) \prod_{i=2}^{t} q(x_i|x_{i-1}),$$

we obtain through Proposition 4 (for $\phi = x_1^* - y = x_1 - y^*$) the acceptance ratio

$$
\begin{aligned}
r(x^*, y^*, x|y) =& \frac{\pi(y^*)f(-\phi)\prod_{i=2}^t q(x_i|x_{i-1})q_{y|x}(y|x_t)\prod_{i=2}^t q(x_i^*|x_{i-1}^*),}{\pi(y)f(\phi)\prod_{i=2}^t q(x_i^*|x_{i-1}^*)q_{y|x}(y^*|x_t^*)\prod_{i=2}^t q(x_i|x_{i-1}),} \\
=& \frac{\pi(y^*)q_{y|x}(y|x_t)}{\pi(y)q_{y|x}(y^*|x_t^*)},
\end{aligned}
$$

which is equal to the acceptance ratio obtained by Jennison and Sharp (2007). In order to get a computationally tractable density, Jennison and Sharp (2007) suggested choosing $q_{y|x}(y^*|x_t^*)$ as a local approximation to $\pi(\cdot)$. Note that $x_{2:t}$ is not involved neither in the generation of y^* nor in the acceptance probability making these variables unnecessary to generate.

As discussed in Appendix A, $q_{y|x}(y^*|x_t^*)$ can also be defined through a MH step. Within such a setting, alternative constructions can be considered. Assume $q_{y|x} = q$. Similar to (15), consider now averages of

$$
h_s(x|y^*, x^*, y) = \delta(x_1 - y^* + \phi)\prod_{i=2}^{s-1} q(x_i|x_{i-1})\prod_{i=s}^t q(x_i|x_{i+1}).
$$

Using (14), we obtain an acceptance ratio equal to

$$
r(x^*, y^*, x|y) = \frac{\sum_{s=2}^{t+1} \frac{\pi(x_s^*)}{q(x_s^*|x_{s-1}^*)}}{\sum_{s=2}^{t+1} \frac{\pi(x_s)}{q(x_s|x_{s-1})}}.
$$

In general, if q is capable of moving efficiently around in the whole state space, this ratio will converge to one. In more practical settings where q is only able to move around within the current mode, the ratio will converge towards the ratio between the masses of the corresponding modes.

An earlier approach suggested by Tjelmeland and Hegstad (2001) is quite similar to the approach by Jennison and Sharp (2007). In this case a similar large jump was followed by a deterministic move to the nearest local mode. $\mu(x^*)$. Thereafter, a second move was performed through $y^* \sim q_{y|x}(y^*|\mu(x^*))$. Tjelmeland and Hegstad (2001) suggested to choose $q_{y|x}$ to be a Gaussian with mean at $\mu(x^*)$ and covariance matrix derived from the second derivatives of $\log \pi(\mu(x^*))$ (very much similar to the choice of local approximation in Jennison and Sharp (2007)). In order to calculate an acceptance probability, a backwards move $y^* \to x \to \mu(x) \to y$ was defined but now with $x = y^* - \phi$. Applying Proposition 4 using $h(x|y, x^*, y^*) = \delta(x - y^* - y + x^*)$, we obtain (for $\phi = x^* - y = y^* - x$)

$$
r(x^*, y^*, x|y) = \frac{\pi(y^*)f(-\phi)q_{y|x}(y|\mu(x))}{\pi(y)f(\phi)q_{y|x}(y^*|\mu(x^*))} = \frac{\pi(y^*)q_{y|x}(y|\mu(x))}{\pi(y)q_{y|x}(y^*|\mu(x^*))}.
$$

which is the same acceptance ratio obtained by Tjelmeland and Hegstad (2001). Using the general framework, combinations of these two approaches can also be considered (i.e. local optimisation followed by a few MCMC steps).

4.3. Reversible jump MCMC proposals

Reversible jump algorithms (Green, 1995, 2001) is an important class of MH algorithms for jumping between spaces of different dimensions. Obtaining reasonable acceptance rates is

however recognised to be a hard problem (Brooks et al., 2003). Denote the state in this case by $\bar{y} = (m, y)$ where m is the dimension (or model) while y is the set of parameters/variables within the model (of which state space may depend on m). Further write $\pi(\bar{y}) = \pi_M(m)\pi_m(y)$.

The problem of jumping between different subspaces/models is to construct a proposal $\bar{y}^*$ giving high support. Al-Awadhi et al. (2004) suggested a change of model through $q_M(n|m, y)q_{n|m}(x_1^*|y)$ followed by t fixed-dimension MCMC steps through states $\bar{x}_2^*, ..., \bar{x}_{t+1}^* = y^*$ where $\bar{x}_i = (m_i^*, x_i^*)$ but $m_i^* = n$. A reverse move was proposed somewhat asymmetric in that in this case t fixed-dimension MCMC steps were taken before a final change of space.

In order to put this into the general framework, assume that for each pair (m, n) if a change from m to n performs a model change first, a change from n to m performs a model change last. Using Proposition 4 with $\bar{x}_{1:t} = \bar{x}_{t:1}^*$ with probability one, we obtain the acceptance ratio

$$r(\bar{y}; \bar{x}^*, \bar{y}^*, \bar{x}) = \frac{\pi(n)\pi_n(x_{t+1}^*)q_M(m|n, x_1^*)q_{m|n}(y|x_1^*)\prod_{i=1}^t q_n(x_i^*|x_{i+1}^*)}{\pi(m)\pi_m(y)q_M(n|m, y)q_{n|m}(x_1^*|y)\prod_{i=2}^{t+1} q_n(x_i^*|x_{i-1}^*)}$$

Al-Awadhi et al. (2004) assumed $\pi_n^*(y)q_n(y^*|y) = \pi_n^*(y^*)q_n(y|y^*)$ for some distribution π_n^* that might differ from π. Using (14) this reduces to

$$r(\bar{y}; \bar{x}^*, \bar{y}^*, \bar{x}) = \frac{\pi(n)\pi_n(x_{t+1}^*)q_M(m|n, x_1^*)q_{m|n}(y|x_1^*)\pi_n^*(x_1^*)}{\pi(m)\pi_m(y)q_M(n|m, y)q_{n|m}(x_1^*|y)\pi_n^*(x_{t+1}^*)}$$

which is equal to the acceptance rate obtained by Al-Awadhi et al. (2004). They suggested choosing π_n^* to be some intermediate distribution between π_n and q_{mn}. Note however that the number of steps t does not affect the acceptance probability.

Consider now an alternative scheme where both the forward and the reverse generation of proposals starts with a change of model followed by t MCMC steps within the chosen model. We further assume q_n is invariant with respect to π_n, avoiding the extra burden of defining π_n^*. Using Proposition 4 with

$$h_s(\bar{x}|\bar{y}^*, \bar{x}^*, \bar{y}) = \delta(m_1 - m)q_{m|n}(x_1|x_{t+1}^*)\prod_{i=2}^{s-1} q_m(x_i|x_{i-1})\prod_{i=s}^t q_m(x_i|x_{i+1})$$

results in weights

$$w_s(\bar{y}; \bar{x}^*, \bar{y}^*, \bar{x}) = \frac{\pi(n)\pi_n(x_{t+1}^*)\prod_{i=s}^t q_n(x_i^*|x_{i+1}^*)}{q_M(n|m, y)\prod_{i=s}^{t+1} q_n(x_i^*|x_{i-1})} = \frac{\pi(n)\pi(x_s^*)}{q_M(n|m, y)q_n(x_s^*|x_{s-1}^*)}$$

who can be combined to give an acceptance ratio of the form

$$r(\bar{y}; \bar{x}^*, \bar{y}^*, \bar{x}) = \frac{\pi(n)q_M(m|n, y^*)\sum_{s=2}^{t+1} a_s \frac{\pi_n(x_s^*)}{q_n(x_s^*|x_{s-1}^*)}}{\pi(m)q_M(n|m, y)\sum_{s=2}^{t+1} a_s \frac{\pi_m(x_s)}{q_m(x_s|x_{s-1})}}.$$

If q_m and q_n define ergodic Markov chains within their respective models and we choose $a_s = 1$, this acceptance ratio converges towards $\pi(n)q_M(m|n, y^*)/\pi(m)q_M(n|m, y)$ as t increases.

4.4. Annealed importance sampling

Neal (2001) proposed a sequential method for generating variables using different transition distributions at each step. Also this approach can be considered as a special case of our general framework, although a small generalisation of the discussion in Section 4.1.

Define a sequence of distributions $\pi_1, ..., \pi_{t+1} = \pi$ where π_1 is some distribution which is easy to sample from while π is the distribution of interest. Now generate $x_1^*, ..., x_{t+1}^* = y^*$ through the sequence $x_i^* \sim q_i(x_i^*|x_{i-1}^*)$ for $i = 1, ..., t + 1$ (where $q_1(x_1^*|x_0^*) = \pi_1(x_1^*)$). Here q_i is a transition distribution satisfying detailed balance with respect to π_i,

$$\pi_i(x_{i-1})q_i(x_i|x_{i-1}) = \pi_i(x_i)q_i(x_{i-1}|x_i).$$

Similar to (15), define

$$h_s(x_{1:t}^*|x_{t+1}^*) = \prod_{i=1}^{s-1} q_i(x_i^*|x_{i-1}^*) \prod_{i=s}^{t} q_{i+1}(x_i^*|x_{i+1}^*)$$

which, using similar calculations as earlier, results in

$$w_s(x_{1:t+1}^*) = \frac{\pi_{s+1}(x_s^*)}{q_s(x_s^*|x_{s-1}^*)} \prod_{i=s+1}^{t} \frac{\pi_{i+1}(x_i^*)}{\pi_i(x_i^*)}.$$

This gives a weight function that can be used in the setting of Proposition 1 for importance sampling and in the setting of Proposition 4 for MCMC sampling.

Using $s = 1$, we obtain

$$w_1(x_{1:t+1}^*) = \prod_{i=1}^{t} \frac{\pi_{i+1}(x_i^*)}{\pi_i(x_i^*)}. \tag{20}$$

This is the weight function used in Neal (2001). If the sequence $\pi_1, ..., \pi_{t+1}$ is chosen such that $\pi_i(y) \approx \pi_{i-1}(y)$, the weights will be close to one. Note that if using these weights within an MH algorithm, π_i only needs to be known up to a proportionality constant.

Another interesting special case is when $s = t + 1$ in which case

$$w_t(x_{1:t+1}) = \frac{\pi_{t+1}(x_{t+1}^*)}{q_t(x_{t+1}^*|x_t^*)}. \tag{21}$$

This last option is also a special case of (3) (or (7) if used within an MH setting). Comparing (21) with the more "standard" choice (20), we see that a multiplication of a ratio of many densities is avoided. On the other hand a conditional density $q_t(x_{t+1}|x_t)$ is needed, which is not always available, see however the discussion in Appendix A.

4.5. Multi-try methods and particle proposals

Liu et al. (2000) suggested a multiple-try algorithm in which several proposal (or particles) where generated with one of these selected as the actual proposal. Andrieu et al. (2008) considered a generalisation of this idea where each particle was generated through sequential Monte Carlo methods (Doucet et al., 2001). We will discuss these approaches in the latter setting.

The simplest form is to generate many parallel sequences $\{x^*_{j,1:t+1}, j = 1, ..., N\}$, choose an index $K \in \{1, ..., N\}$ with some probability depending on all the sequences and then use as proposal $y^* = x^*_{K,t+1}$. Andrieu et al. (2008) considered more general schemes where at each iteration resampling is performed. We will, mainly for notational simplicity, consider the simpler case where no resampling is performed.

Assume $x^*_{j,1:t+1}$ is generated sequentially with $x^*_{j,i} \sim q_i(x^*_{j,i}|x^*_{j,i-1})$, similar to the procedure in Section 4.4. Identify $x^*_{j,0} = y$ for all j. Assume $q_{y|x}(y^*|x^*)$ is a discrete probability distribution with

$$q_{y|x}(y^* = x^*_{K,t+1}|x^*) \propto \frac{\pi(x^*_{K,t+1})}{q_{t+1}(x^*_{K,t+1}|x^*_{K,t})}. \tag{22}$$

In this case, choose

$$h(x|y, x^*, y^*) = N^{-1}\delta(x_{K,t+1} - y) \prod_{i=1}^{t} q_i(x_{K,i}|x_{K,i-1}) \prod_{j \neq K} \prod_{i=1}^{t+1} q_i(x_{j,i}|x_{j,i-1}).$$

By inserting into (2), we obtain for $y^* = x^*_{K,t+1}$

$$w(y; x^*, y^*, x) = N^{-1} \sum_{j=1}^{N} \frac{\pi(x^*_{j,t+1})}{q_{t+1}(x^*_{j,t+1}|x^*_{j,t})},$$

giving the acceptance ratio

$$r(y; x^*, y^*, x) = \frac{\sum_{j=1}^{N} \frac{\pi(x^*_{j,t+1})}{q_{t+1}(x^*_{j,t+1}|x^*_{j,t})}}{\sum_{j=1}^{N} \frac{\pi(x_{j,t+1})}{q_{t+1}(x_{j,t+1}|x_{j,t})}}.$$

Comparing the weight function to (21), we see that we now have obtained an average of similar terms which will have reduced variability. Alternative weight functions and acceptance ratios can be obtained by considering other choices of h functions.

A special case of this approach using $t = 0$ corresponds to the multiple-try method by Liu et al. (2000). They considered different types of acceptance probabilities that can be related to different choices of h functions in our setting.

5. Non-sequential approaches

In this section we will discuss non-sequential algorithms proposed in the literature that also can be seen as examples of the general scheme. Putting these into the general framework both makes it easier to see the basic ideas these approaches are based on and suggest alternative weight functions or acceptance probabilities to use.

5.1. The Pseudo-marginal algorithm
Beaumont (2003) considered a situation where $\pi(y)$ is not known specifically, but given through

$$\pi(y) = \int_x \bar{\pi}(x, y)dx$$

with $\bar{\pi}(x, y)$ known (up to perhaps a proportionality constant). The following algorithm was proposed: Given $(y, x_1, ..., x_N)$,

(1) simulate $y^* \sim q_y(y^*|y)$,
(2) simulate $x_1^*, ..., x_N^* \sim q_{x|y}(x^*|y^*)$,
(3) accept the new sample with acceptance ratio

$$r(x, y; x^*, y^*) = \frac{q_y(y|y^*)}{q_y(y^*|y)} \frac{\frac{1}{N}\sum_{j=1}^{N} \frac{\bar{\pi}(x_j^*, y^*)}{q_{x|y}(x_j^*|y^*)}}{\frac{1}{N}\sum_{j=1}^{N} \frac{\bar{\pi}(x_j, y)}{q_{x|y}(x_j|y)}} \tag{23}$$

(slightly rewritten). Using that $\frac{1}{N}\sum_{j=1}^{N} \frac{\bar{\pi}(x_j^*, y^*)}{q_{x|y}(x_j^*|y^*)}$ can be considered as a Monte Carlo estimate of $\pi(y^*)$, the acceptance ratio is a Monte Carlo estimate of

$$r(y; y^*) = \frac{q_y(y|y^*)\pi(y^*)}{q_y(y^*|y)\pi(y)},$$

the acceptance ratio for marginal simulation of y. The algorithm has therefore been termed the pseudo-marginal algorithm Andrieu and Roberts (2009).

In order to put this algorithm into our general framework, define

$$h_s(x_{1:N}|y) = \bar{\pi}(x_s|y)\prod_{j\neq s} q_{x|y}(x_j|y). \tag{24}$$

The corresponding weight function has the form

$$w_s(y; x_{1:N}^*, y^*) = \frac{\pi(y^*)\bar{\pi}(x_s^*|y^*)\prod_{j\neq s} q_{x|y}(x_j^*|y^*)}{q_y(y^*|y)\prod_{j=1}^{N} q_{x|y}(x_j^*|y^*)} = \frac{\bar{\pi}(x_s^*, y^*)}{q_y(y^*|y)q_{x|y}(x_s^*|y^*)}$$

A weighted average of these weight functions is also a proper weight function resulting in the acceptance ratio (23). As an alternative to store $x_1, ..., x_N$, new x's can be generated according to the mixture distribution the h_s distributions given in (24). It then follows from the general theory that applying these weights in a MH setting makes (under suitable regularity conditions) the sequence of generated y's converge towards $\pi(y)$. Beaumont (2003) showed this in the special setting of independent x's while Andrieu and Roberts (2009) generalised this to the more general setting, although through a different route.

Andrieu and Roberts (2009) generalised this idea to allow for non-independent x's and also considered the theoretical properties of this algorithm. A possible framework in this case is to divide the set $\{1, ..., N\}$ into three disjoint subsets $\{s\}, v_1, v_2$ and define

$$h_s(x_{1:t}^*|y^*) = \bar{\pi}(x_s^*|y^*)q_{x|y}(x_{v_1}^*|y^*)q_{x|y}(x_{v_2}^*|x_s^*, x_{v_1}^*, y^*)$$

where $x_v = \{x_i, i \in v\}$. Then the corresponding weight function has the form

$$w_s(y; x_{1:t}^*, y^*) = \frac{\bar{\pi}(x_s^*, y^*)}{q_y(y^*|y)q_{x|y}(x_s^*|x_{v_1}^*, y^*)}.$$

A weighted average of these weight functions is also a proper weight function. It then follows from the general theory that applying these weights in a MH setting makes (under suitable regularity conditions) the sequence of generated y's converge towards $\pi(y)$. Beaumont (2003) showed this in the special setting of independent x's while Andrieu and Roberts (2009) generalised this to the more general setting, although through a different route.

5.2.   Distributions with intractable normalising constants

Møller et al. (2006) considered a problem of drawing from a posterior distribution

$$\pi(y) = p(y|z) = C^{-1}p(y)p(z|y)$$

where the likelihood for data z is

$$p(z|y) = Z^{-1}(y)\tilde{p}(z|y).$$

Here both the normalisation constant involved in the posterior C and the normalisation constant defining the likelihood $Z(y)$, which depend on y, are unknown (a problem often encountered in spatial modelling). The data z is dropped in $\pi(y)$ since we are considering this as a given constant. Møller et al. (2006) proposed an MCMC algorithm that we now will describe in the setting of Proposition 3.

Assume (x, y) is the current state and generate a proposal (x^*, y^*) through $q(x^*, y^*|x, y) = q_y(y^*|x, y)q_{x|y}(x^*|x, y, y^*)$ where $q_{x|y}(x^*|x, y, y^*) = Z^{-1}(y^*)\tilde{p}(x^*|y^*)$. Consider a weight function of the form

$$w(x, y; x^*, y^*) = \frac{C^{-1}p(y^*)Z^{-1}(y^*)\tilde{p}(z|y^*)h(x^*|y^*)}{q_y(y^*|x, y)Z^{-1}(y^*)\tilde{p}(x^*|y^*)} = C^{-1}\frac{p(y^*)\tilde{p}(z|y^*)h(x^*|y^*)}{q_y(y^*|x, y)\tilde{p}(x^*|y^*)}$$

and an acceptance ratio

$$r(x, y; x^*, y^*) = \frac{p(y^*)\tilde{p}(z|y^*)h(x^*|y^*)q_y(y|x^*, y^*)\tilde{p}(x|y)}{p(y)\tilde{p}(z|y)h(x|y)q_y(y^*|x, y)\tilde{p}(x^*|y^*)}$$

where $h(x|y)$ has a state space similar to z but may otherwise be arbitrary (and even depend on z). The given algorithm will according to Proposition 3 have $\pi(y)h(x|y)$ as invariant distribution. This is the same ratio obtained by Møller et al. (2006). They also discussed several options for choosing the h distribution.

Consider now the case where $q_{y|x}(y^*|y, x) = q_{y|x}(y^*|y)$, a special case only considered further by Møller et al. (2006) after their general description of the algorithm. Through Proposition 4 we then are able to construct an alternative algorithm where x do not need to be stored from one iteration to another. Using the same q function as above but now simulating $x \sim h(x|y, x^*, y^*)$ in addition to x^*, y^*, we obtain

$$r(x^*, y^*|x, y) = \frac{h(x^*|y^*, x, y)p(y^*)\tilde{p}(z|y^*)q_{y|x}(y^*|y)\tilde{p}(x|y)}{h(x|y, x^*, y^*)p(y)\tilde{p}(z|y)q_{y|x}(y|y^*)\tilde{p}(x^*|y^*)}.$$

In this case, more general choices of the h distribution can be considered. One option is to choose $h(x|y, x^*, y^*) = \delta(x - x^*)$, making generation of an extra x unnecessary. In which case the acceptance ratio reduces to

$$r(x^*, y^*|x, y) = \frac{p(y^*)\tilde{p}(z|y^*)q_{y|x}(y^*|y)\tilde{p}(x^*|y)}{p(y)\tilde{p}(z|y)q_{y|x}(y|y^*)\tilde{p}(x^*|y^*)}.$$

5.3.   Delayed rejection sampling

Tierney and Mira (1999) suggested a method for composing a new (different) proposal in an MH setting when the first proposal was rejected. This approach was further considered and extended in Mira (2001) and Green and Mira (2001).

Consider a situation where given the current state y, a first proposal x_1^* is generated by $q_1(x_1^*|y)$ and accepted with a probability $\alpha_1(y; x_1^*)$. If rejected, a new proposal x_2^* is generated by $q_2(x_2^*|y, x_1^*)$ and this new proposal is accepted with probability $\alpha_2(y, x_1^*; x_2^*)$. We will see how this approach can be considered as a special case of Proposition 4.

In this case it is reasonable to consider $q_{y|x}(y^*|y, x_1^*, x_2^*)$ to be a discrete distribution putting $y^* = x_1^*$ with probability $\alpha_1(y; x_1^*)$ and $= x_2^*$ otherwise, so

$$q_{y|x}(y^*|y, x_1^*, x_2^*) = \alpha_1(y; x_1^*)^{I(y^*=x_1^*)}[1 - \alpha_1(y; x_1^*)]^{I(y^*=x_2^*)}.$$

Consider first a situation where in addition to (x_1^*, x_2^*) also (x_1, x_2) are generated using the proposal distribution $h(x_1, x_2|y, x_1^*, x_2^*, y^*)$. In order to get the variables to work within the same spaces, we make $h(x_1, x_2|y, x_1^*, x_2^*, y^*)$ degenerate in the sense that either x_1 or x_2 is equal to y. Consider the choice

$$h(x_1, x_2|y, x_1^*, x_2^*, y^*) = \begin{cases} \delta(x_1 - y)h_2(x_2|y, x_2^*, y^*) & \text{if } y^* = x_1^*; \\ \delta(x_2 - y)h_1(x_1|y, x_1^*, y^*) & \text{if } y^* = x_2^*. \end{cases}$$

We consider the two possibilities separately:

$y^* = x_1^*$: In that case $x_1 = y$ and by applying Proposition 4, we obtain

$$\begin{aligned} r(y; x^*, y^*, x) &= \frac{\pi(y^*)q_1(y|y^*)q_2(x_2|y^*, x_1)\alpha_1(y^*; y)h_2(x_2^*|y^*, x_2, y)}{\pi(y)q_1(y^*|y)q_2(x_2^*|y, y^*)\alpha_1(y; y^*)h_2(x_2|y, x_2^*, y^*)} \\ &= \frac{q_2(x_2|y^*, x_1)h_2(x_2^*|y^*, x_2, y)}{q_2(x_2^*|y, y^*)h_2(x_2|y, x_2^*, y^*)} \end{aligned}$$

where we have used that $\pi(y^*)q_1(y|y^*)\alpha(y^*; y) = \pi(y)q_1(y^*|y)\alpha_1(y; y^*)$ for a standard MH choice of $\alpha_1(y; y^*)$. For $h_2(x_2^*|y^*, x_1, y) = q_2(x_2^*|y, x_1^*)$, this reduces to 1, which corresponds nicely to accepting the first proposal x_1^* with probability $\alpha(y; x_1^*)$.

$y^* = x_2^*$: In this case $x_2 = y$ and we obtain

$$r(y; x^*, y^*, x) = \frac{\pi(y^*)q_1(x_1|y^*)q_2(y|y^*, x_1)[1 - \alpha_1(y^*; x_1)]h_1(x_1^*|y^*, x_1, y)}{\pi(y)q_1(x_1^*|y)q_2(y^*|y, x_1^*)[1 - \alpha_1(y; x_1^*)]h_1(x_1|y, x_1^*, y^*)}$$

Choosing now $h_1(x_1|y, x_1^*, y^*)$ to be degenerate in that $x_1 = x_1^*$ with probability one, this reduces to

$$r(y; x^*, y^*, x) = \frac{\pi(y^*)q_1(x_1^*|y^*)q_2(y|y^*, x_1^*)[1 - \alpha_1(y^*; x_1^*)]}{\pi(y)q_1(x_1^*|y)q_2(y^*|y, x_1^*)[1 - \alpha_1(y; x_1^*)]} \tag{25}$$

which is the same acceptance ratio as that given by Tierney and Mira (1999). An alternative in this case is to choose $h_1(x_1|y, x_1^*, y^*) = q_1(x_1|y^*)$ in which case the acceptance ratio reduces to

$$r(y; x^*, y^*, x) = \frac{\pi(y^*)q_2(y|y^*, x_1)[1 - \alpha_1(y^*; x_1)]}{\pi(y)q_2(y^*|y, x_1^*)[1 - \alpha_1(y; x_1^*)]}. \tag{26}$$

This acceptance probability might be a reasonable choice in the case where you have the choice between using a (computationally) cheap proposal q_1 and a more expensive q_2 which is close to π. Note that in the special case where $q_2 = \pi$, this acceptance ratio reduces to

$$r(y; x^*, y^*, x) = \frac{1 - \alpha_1(y^*; x_1)}{1 - \alpha_1(y; x_1^*)}.$$

With q_1 being a bad approximation to π, both the acceptance probabilities $\alpha_1(y; x_1^*)$ and $\alpha_1(y^*; x_1)$ will be small, making $r(y; x^*, y^*, x)$ almost equal to 1, as it should!

6. Experiments

The usefulness of most of the different algorithms discussed in Section 5 have already been presented in the literature. In this section we will concentrate on the flexibility of choosing different weight functions/acceptance probabilities.

6.1. *Sequential updating and Gaussian mixtures*

Consider a model in $\mathcal{R}^p$

$$\pi(y) = \beta N(y; \mu_1, I) + (1 - \beta)N(y; \mu_2, I)$$

where $\mu_{1,1} = -\mu_{2,1} = -10$ while $\mu_{i,j} = 0, i = 1, 2, j = 2, ..., p$. This corresponds to a model with multiple modes so separated that ordinary MCMC methods will get stuck in one of the modes.

We will consider an MH algorithm where proposals are generated sequentially as described in Section 4 with starting value generated from $N(\mu_1, I)$.

Proposals y^* are generated by simulating a sequence $x_1^*, ..., x_{t+1}^*$ with $y^* = x_{t+1}^*$. Here $x_1^* \sim N(0, \sigma_{large}^2 I)$ followed by a set of t "inner" MH steps using a discrete version Langevin diffusion (Besag, 1994; Roberts and Tweedie, 1996) where proposals z_j^* are generated by

$$z_j^* \sim N(x_{j-1}^* + \tfrac{1}{2}\sigma_{small}^2 \Delta \log \pi(x_{j-1}^*), \sigma_{small}^2 I)$$

and proposals are accepted using ordinary MH acceptance rates. The number of "outer" MH steps, i.e. the number of generated y^*'s will be denoted by M.

The final proposals y^* are accepted with probabilities given by Proposition 3 with three different weight functions, all based on the general weight functions (16):

(1) The special case $s = 1$ corresponding to weight function (17).
(2) The special case $s = t + 1$ corresponding to weight function (18).
(3) A weighted average of (16) as given in (19) with $a_s = 0$ for $s < (t+1)/2$ and $= 2/(t+1)$ for $s \geq (t + 1)/2$.

The full algorithm was restarted N times and the last sample was stored in each case, giving samples $y_1, ..., y_N$. As a measure for the performance of the different weight functions,

$$d(\widehat{F}_1, F_1) = \sup_{y_1} |\widehat{F}_1(y_1) - F_1(y_1)| \tag{27}$$

was used where $\widehat{F}_1$ is the empirical distribution function based on the the first component of the samples $y_1, ..., y_N$ while F_1 is the true marginal cumulative distribution function of the corresponding variable.

Figure 1 (panels $(a)-(e)$) shows boxplots of $d(\widehat{F}, F)$ with $N = 10000, M = 20, t+1 = 10$ and with $\sigma_{large} = 15, \sigma_{small} = 1.5$. The different panels correspond to different dimensions p. The boxplots are obtained by repeating the experiments 100 times. We see in all cases that the first weight function do not perform that good, an expected property given the previous discussion about this particular choice. Both the other choices give significant improvements with the third weight function performing best in all cases. The difference between these seem to depend on dimension though. In particular it seems like the difference between them are decreasing with dimension, but then increasing for $p = 10$. This is probably due to that for $p = 10$, 20 iterations in the outer MCMC iterations is not that much. Increasing M to 40 (panel (f)), the similarities between the second and the third weight functions are again obtained.

Figure 2 shows the true cumulative distribution for the first component and typical examples of empirical distributions based on the three different weight functions for $p = 10$ and $M = 20$. The differences between the choices of weight functions shown in Figure 1 is clearly seen. Note in particular that the third weight function produces estimates that are indistinguishable from the truth, indicating that convergence has been reached even with this small number of MCMC iterations.

6.2.   *Delayed rejection and multiple precision models*

Consider a model where observations $z \in \mathcal{R}^p$ follow a model

$$z_i = \mu + \eta_i + \varepsilon_i$$

where $\varepsilon_1, ..., \varepsilon_n$ are iid zero-mean Gaussian variables with precision τ_1 while $\eta_1, ..., \eta_n$ are spatially correlated variables with a conditional autoregressive (CAR) structure (Besag, 1974; Besag and Kooperberg, 1995) with conditional distributions given by

$$\eta_i | \eta_{-i} \sim N(\beta n_i^{-1} \sum_{j \sim i} \eta_j, n_i^{-1} \kappa)$$

with n_i being the number of neighbors for observation i. Our interest will be in the posterior distribution of $\tau = (\tau_1, \tau_2)$ where $\tau_2 = \kappa^{-1}$. Assuming independent Gamma distributions for τ_1, τ_2), both with shape parameter α and rate parameter β, the posterior distribution has the form

$$\pi(\tau) \propto \tau_1^{\alpha-1} e^{-\beta \tau_1} \tau_2^{\alpha-1} e^{-\beta \tau_2} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}.$$

He et al. (2007) state that such distributions in many cases are "l-shaped with two long arms pressed tightly along one or both coordinate axes", which is demonstrated in the left panel of Figure 3 showing the posterior distribution obtained from simulated values of $z_1, ..., z_n$ on a 5×5 grid using $\mu = 0, \beta = 0.25, \tau_1 = 1$ and $\tau_2 = 1/3$.

We will consider a delayed rejection sampling algorithm where at the first stage proposals are generated through scale-proposals suggested by Knorr-Held and Rue (2002) where

$$\tau_j^1 = \tau_j f_j$$

and $f_j \in [F^{-1}, F]$ with density proportional to $1 + f^{-1}$. At the second stage, a proposal is generated through the reparametrisations

$$\tau = \frac{\tau_1 \tau_2}{\tau_1 + \tau_2}, \quad r = \frac{\tau_2}{\tau_1 + \tau_2} \tag{28}$$
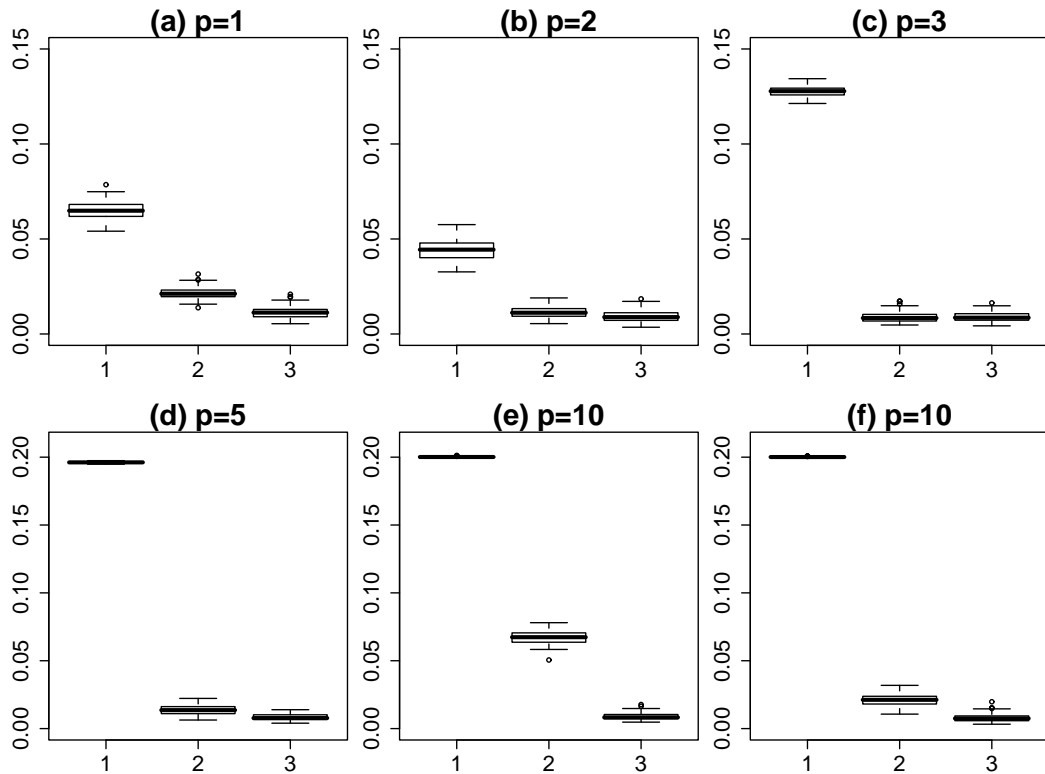
Fig. 1. Boxplot of $d(\widehat{F}, F)$ for the Gaussian mixture distribution. The different panels correspond to results on different dimensions $p = 1, 2, 3, 5, 10$. The experimental setup is described in the text. Panels (a)-(e) correspond to $M = 20$ while panel (f) corresponds to $M = 40$. The numbers on the x-axis correspond to the different types of weight functions.

where a new r^2 is generated uniformly on $[0, 1]$ while τ^2 is generated through the posterior of τ given r which can be shown to be a Gamma distribution with shape parameter $2\alpha + n$ and rate parameter $[\beta + 0.5ssq]/[r(1 - r)]$ where ssq is the sum of squares using the covariance matrix divided by τ. New proposals τ_1^2, τ_2^2 are then given by the inverse transformations of (28).

Our aim will be to compare the two alternative acceptance probabilities (25) and (26), referring to the references in Section 5.3 for demonstration of the power in using delayed rejection sampling algorithms in general. The right panel of Figure 3 shows the distance measure (27) as function of number of MCMC iterations based on the "standard" delayed rejection sampling acceptance probability (25) (solid line) and the alternative acceptance probability (26) (dashed line). We clearly see the improvements made using the alternative acceptance probability.
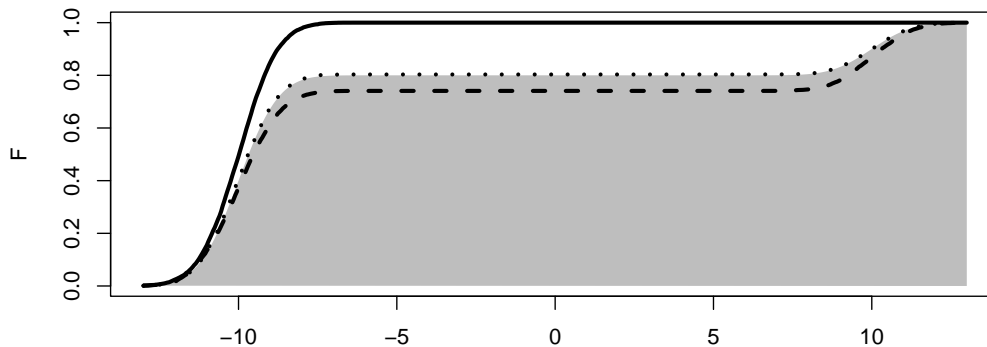
Fig. 2. Mixture Gaussian distribution with dimension $p = 10$ and number of outer MCMC iterations $M = 20$. True (shaded) and examples of estimated cumulative distributions of the first component. Solid line corresponds to the first weight function, dashed line to the second weight function and dotted line to the third weight function.

7. Summary and discussion

In this paper we have presented a framework for construction of auxiliary variable proposal generation. Many algorithms proposed in the literature are shown to fit within the framework and in several cases, alternative acceptance probabilities and/or importance weights are suggested. Numerical experiments demonstrate that in some cases significant improvements can be obtain by using these alternatives.

The variety of options for acceptance probabilities have also been suggested elsewhere. In this paper we have shown that these options can be related to choices of extended state spaces for the variables to be generated. These choices comes in addition to the flexibility in how proposals are generated.

Although the added flexibility makes it possible to define alternative and hopefully better algorithms, it also extends the number of choices to make in order to construct an efficient algorithm. Theoretical results guiding the practitioner in these choices would be valuable additions to the framework but are so far lacking.

Acknowledgement

A. Transition densities for internal Metropolis-Hastings moves

In this section we discuss the use of weight functions when proposals are generated by a fixed number of internal MH steps. The general difficulty in this case is that the transition probabilities $q_s(x_s^*|x_{s-1}^*)$ not always are directly available

One possibility, explored by both Tjelmeland and Hegstad (2001) and Jennison and Sharp (2007) is to only consider transition probabilities for the last proposal and in this case use a local approximation to π (typically a multivariate Gaussian) in which case the transition probabilities involved can easily be calculated, see Section 4.2.
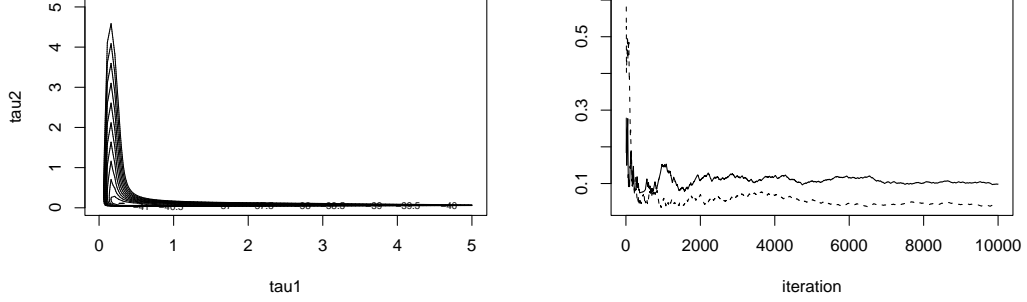
Fig. 3. Left panel: Posterior distribution of (τ_1, τ_2) in the CAR model based on 50 simulated observations on a 5×5 grid using $\mu = 0, \beta = 0.25$ and $\tau_1 = 1, \tau_2 = 1/3$. Right panel: Distance, as defined through (27), between true and empirical distribution as function of the number of MCMC iterations for the "standard" delayed rejection sampling acceptance probability (25) (solid line) and the alternative acceptance probability (26) (dashed line).

In the case where only parts of the state vector is changed at the time and at least one of the blocks can be moved through a Gibbs sampler update, using $a_s > 0$ only for those steps at which a Gibbs sampler is used, avoids the need for calculating more complicated transition kernels.

In order to utilise all the generated auxiliary variables, an alternative is to include also the proposed values at each iteration into the set of auxiliary variables. Assume that $x_{1:t+1}^*$ (with proposals $z_{1:t}^*$) are generated by MH steps, i.e.

$$
\begin{aligned}
z_i^* &\sim q_i^z(\cdot | x_{i-1}^*) \\
x_i^* &= \begin{cases} z_i^* & \text{with probability } \alpha_i(x_{i-1}^*; z_i^*); \\ x_{i-1}^* & \text{otherwise.} \end{cases}
\end{aligned}
$$

The generating distribution can be written as

$$
\prod_{i=1}^{t+1} q_i(x_i^* | x_{i-1}^*) q_i^{z|x}(z_i^* | x_{i-1}^*, x_i^*)
$$

where $q_i(x_i^* | x_{i-1}^*), i = 1, ..., t+1$ is the transition kernel for the Markov chain $\{x_{1:t+1}^*\}$ while

$$
q_i^{z|x}(z_i^* | x_{i-1}, x_i^*) = \frac{q_i^z(z_i^* | x_{i-1}^*) \alpha_i(x_{i-1}^*; z_i^*)^{I(x_i^* = z_i^*)} [1 - \alpha_i(x_{i-1}^*; z_i^*)]^{I(x_i^* = x_{i-1}^*)}}{q_i(x_i^* | x_{i-1}^*)}
$$

is the conditional distribution for z_i^* given (x_{i-1}^*, x_i^*). Now consider the choice

$$
\begin{aligned}
h_s(z_{1:t+1}^*&,x_{1:t}^*|y,x,y^*) \\
&= \prod_{i=1}^{s-1} q_i(x_i^*|x_{i-1}^*)q_i^{z|x}(z_i^*|x_{i-1}^*,x_i^*) \times \\
&\quad h_s^z(z_s^*|x_{s-1}^*,x_s^*) \prod_{i=s}^{t} q_{i+1}(x_i^*|x_{i+1}^*)q_{i+1}^{z|x}(z_{i+1}^*|x_i^*,x_{i+1}^*).
\end{aligned}
$$

Using (14), we obtain

$$
\begin{aligned}
w_s(y;z_{1:t}^*,x_{1:t+1}^*) &= \frac{\pi(y^*)h_s^z(z_s^*|x_{s-1}^*,x_s^*)\prod_{i=s}^{t}q_{i+1}(x_i^*|x_{i+1}^*)}{q_s(x_s^*|x_{s-1}^*)q_s^{z|x}(z_s^*|x_{s-1}^*,x_s^*)\prod_{i=s}^{t}q_{i+1}(x_{i+1}^*|x_i^*)} \\
&= \frac{\pi(x_s^*)h_s^z(z_s^*|x_{s-1}^*,x_s^*)}{q_s^z(z_s^*|x_{s-1}^*)\alpha_{z,s}(x_{s-1}^*;z_s^*)^{I(x_s^*=z_s^*)}[1-\alpha_{z,s}(x_{s-1}^*;z_s^*)]^{I(x_s^*=x_{s-1}^*)}}.
\end{aligned}
$$

For the choice of $h_s^z(z_s^*|x_{s-1}^*,x_s^*)$, some care should be taken in order to increase the probability for nonzero acceptance probabilities. In particular the distribution should reflect that $x_s^* \neq x_{s-1} \Rightarrow z_s^* = x_s^*$. Therefore, assume

$$
h_s^z(z_s^*|x_{s-1}^*,x_s^*) = \begin{cases} \delta(z_s^* - x_s^*) & \text{if } x_s^* \neq x_{s-1}^*; \\ \tilde{q}_s^{z|x}(z_s^*|x_{s-1}^*,x_s^*) & \text{if } x_s^* = x_{s-1}^*, \end{cases}
$$

where $\tilde{q}_s^{z|x}(\cdot|x_{s-1}^*,x_s^*)$ is a general density with support equal to $q_s^z(\cdot|x_{s-1}^*)$. The obvious option is to choose $\tilde{q}_s^{z|x}(z_s^*|x_{s-1}^*,x_s^*) = q_s^z(z_s^*|x_{s-1}^*)$ in which case the weight function reduces to

$$
w_s(y;x_{1:t+1}^*) = \begin{cases} \frac{\pi(x_s^*)}{q_s^z(z_s^*|x_{s-1}^*)\alpha_{z,s}(x_{s-1}^*;x_s^*)} & \text{if } x_s^* \neq x_{s-1}^*; \\ \frac{\pi(x_s^*)}{1-\alpha_{z,s}(x_{s-1}^*;z_s^*)} & \text{if } x_s^* = x_{s-1}^*. \end{cases}
$$

Another option is to choose

$$
h_s^z(z_s^*|x_{s-1}^*,x_s^*) = \begin{cases} \delta(z_s^* - x_s^*) & \text{with probability } 1-\alpha_{z,s}(x_{s-1}^*,x_s^*); \\ q_s(z_s^*|x_{s-1}^*) & \text{with probability } \alpha_{z,s}(x_{s-1}^*,x_s^*) \end{cases}
$$

In that case, the weight function reduces to

$$
w_s(y;z_{1:t}^*,x_{1:t+1}^*) = \begin{cases} \frac{\pi(x_s^*)}{q_s(x_s^*|x_{s-1}^*)} & \text{if } x_s^* \neq x_{s-1}^*; \\ 0 & \text{otherwise.} \end{cases}
$$

In this case, only those iterations corresponding to acceptance are considered.

References

Al-Awadhi, F., M. Hurn, and C. Jennison (2004). Improving the acceptance rate of reversible jump mcmc proposals. *Statistics & Probability Letters 69*, 189–198.

Andrieu, C., A. Doucet, and R. Holenstein (2008). Particle Markov chain Monte Carlo. Preprint.

Andrieu, C. and G. O. Roberts (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*. To be published.

Beaumont, M. (2003). Estimation of Population Growth or Decline in Genetically Monitored Populations. *Genetics 164*(3), 1139–1160.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B 36*(2), 192–236.

Besag, J. (1994). Comments on Representations of knowledge in complex systems by U. Grenander and MI Miller. *J. Roy. Statist. Soc. Ser. B 56*, 591–592.

Besag, J. and P. J. Green (1993). Spatial statistics and Bayesian computation (with discussion). *Journal of Royal Statistical Society, Series B 55*(1), 25–37.

Besag, J. and C. Kooperberg (1995). On conditional and intrinsic autoregressions. *Biometrika 82*(4), 733–746.

Brooks, S., P. Giudici, and G. Roberts (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society Series B(Statistical Methodology) 65*(1), 3–39.

Damien, P., J. Wakefield, and S. Walker (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of Royal Statistical Society, Series B 61*(2), 331–344.

Doucet, A., N. de Freitas, and N. Gordon (2001). *Sequential Monte Carlo in Practice.* Springer-Verlag.

Edwards, R. G. and A. D. Sokal (1988). Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Phys. Rev. 38*, 2009–2012.

Gilks, W., S. Richardson, and D. J. Spiegelhalter (1996). *Markov chain Monte Carlo in practice.* London: Chapman & Hall.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*(4), 711–732.

Green, P. (2001). A primer on Markov chain Monte Carlo. In O. E. Barndorff-Nielsen, D. R. Cox, and C. Kluppelberg (Eds.), *Complex Stochastic Systems*, pp. 1–62. Chapman and Hall, London.

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

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

He, Y., J. Hodges, and B. Carlin (2007). Re-considering the variance parameterization in multiple precision models. *Bayesian Analysis 2*(3), 529–556.

Higdon, D. (1998). Auxiliary variable methods for Markov chain Monte Carlo with applications. *Journal of the American Statistical Association 93*(442), 585–595.

Jennison, C. and R. Sharp (2007). Mode jumping in MCMC: Adapting proposals to the local environment. Talk at Conference to honour Allan Seheult, Durham, March 2007. Available at http://people.bath.ac.uk/mascj/.

Knorr-Held, L. and H. Rue (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics 29*(4), 597–614.

Liu, J., F. Liang, and W. Wong (2000). The multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association 95*(449), 121–134.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*(6), 1087.

Mira, A. (2001). On Metropolis-Hastings algorithms with delayed rejection. *Metron 59*(3-4), 231–241.

Møller, J., A. Pettitt, R. Reeves, and K. Berthelsen (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika 93*(2), 451–458.

Neal, R. M. (2001). Annealing importance sampling. *Statistics and Computing 2*, 125–139.

Peskun, P. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika 60*(3), 607–612.

Robert, C. P. and G. Casella (2004). *Monte Carlo Statistical Methods* (Second ed.). New York: Springer.

Roberts, G. and R. Tweedie (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli 2*, 341–364.

Tanner, M. and W. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 82*, 528–550.

Tierney, L. and A. Mira (1999). Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine 18*(1718), 2507–2515.

Tjelmeland, H. and B. Hegstad (2001). Mode jumping proposals in MCMC. *Scand. J. Statist. 28*(1), 205–223.