

## Chapter 5

# The Metropolis-Hastings Algorithm

### 5.1 Algorithm

In the previous chapter we have studied the Gibbs sampler, a special case of a Monte Carlo Markov Chain (MCMC) method: the target distribution is the invariant distribution of the Markov chain generated by the algorithm, to which it (hopefully) converges.

This chapter will introduce another MCMC method: the Metropolis-Hastings algorithm, which goes back to Metropolis et al. (1953) and Hastings (1970). Like the rejection sampling algorithm 3.1, the Metropolis-Hastings algorithm is based on proposing values sampled from an instrumental distribution, which are then accepted with a certain probability that reflects how likely it is that they are from the target distribution  $f$ .

The main drawback of the rejection sampling algorithm 3.1 is that it is often very difficult to come up with a suitable proposal distribution that leads to an efficient algorithm. One way around this problem is to allow for “local updates”, i.e. let the proposed value depend on the last accepted value. This makes it easier to come up with a suitable (conditional) proposal, however at the price of yielding a Markov chain instead of a sequence of independent realisations.

**Algorithm 5.1 (Metropolis-Hastings).** Starting with  $\mathbf{X}^{(0)} := (X_1^{(0)}, \dots, X_p^{(0)})$  iterate for  $t = 1, 2, \dots$

1. Draw  $\mathbf{X} \sim q(\cdot | \mathbf{X}^{(t-1)})$ .
2. Compute

$$\alpha(\mathbf{X} | \mathbf{X}^{(t-1)}) = \min \left\{ 1, \frac{f(\mathbf{X}) \cdot q(\mathbf{X}^{(t-1)} | \mathbf{X})}{f(\mathbf{X}^{(t-1)}) \cdot q(\mathbf{X} | \mathbf{X}^{(t-1)})} \right\}. \quad (5.1)$$

3. With probability  $\alpha(\mathbf{X} | \mathbf{X}^{(t-1)})$  set  $\mathbf{X}^{(t)} = \mathbf{X}$ , otherwise set  $\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)}$ .

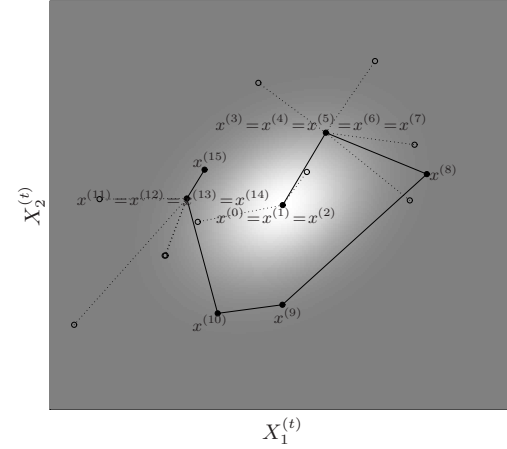
Figure 5.1 illustrates the Metropolis-Hastings algorithm. Note that if the algorithm rejects the newly proposed value (open disks joined by dotted lines in figure 5.1) it stays at its current value  $\mathbf{X}^{(t-1)}$ . The probability that the Metropolis-Hastings algorithm accepts the newly proposed state  $\mathbf{X}$  given that it currently is in state  $\mathbf{X}^{(t-1)}$  is

$$a(\mathbf{x}^{(t-1)}) = \int \alpha(\mathbf{x} | \mathbf{x}^{(t-1)}) q(\mathbf{x} | \mathbf{x}^{(t-1)}) d\mathbf{x}. \quad (5.2)$$

Just like the Gibbs sampler, the Metropolis-Hastings algorithm generates a Markov chain, whose properties will be discussed in the next section.

**Remark 5.1.** The probability of acceptance (5.1) does not depend on the normalisation constant, i.e. if  $f(\mathbf{x}) = C \cdot \pi(\mathbf{x})$ , then

$$\frac{f(\mathbf{x}) \cdot q(\mathbf{x}^{(t-1)} | \mathbf{x})}{f(\mathbf{x}^{(t-1)}) \cdot q(\mathbf{x} | \mathbf{x}^{(t-1)})} = \frac{C\pi(\mathbf{x}) \cdot q(\mathbf{x}^{(t-1)} | \mathbf{x})}{C\pi(\mathbf{x}^{(t-1)}) \cdot q(\mathbf{x} | \mathbf{x}^{(t-1)})} = \frac{\pi(\mathbf{x}) \cdot q(\mathbf{x}^{(t-1)} | \mathbf{x})}{\pi(\mathbf{x}^{(t-1)}) \cdot q(\mathbf{x} | \mathbf{x}^{(t-1)})}$$



**Figure 5.1.** Illustration of the Metropolis-Hastings algorithm. Filled dots denote accepted states, open circles rejected values.

Thus  $f$  only needs to be known up to normalisation constant.<sup>1</sup>

### 5.2 Convergence results

**Lemma 5.2.** The transition kernel of the Metropolis-Hastings algorithm is

$$K(\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}) = \alpha(\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)}) q(\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)}) + (1 - a(\mathbf{x}^{(t-1)})) \delta_{\mathbf{x}^{(t-1)}}(\mathbf{x}^{(t)}), \quad (5.3)$$

where  $\delta_{\mathbf{x}^{(t-1)}}(\cdot)$  denotes Dirac-mass on  $\{\mathbf{x}^{(t-1)}\}$ .

Note that the transition kernel (5.3) is *not* continuous with respect to the Lebesgue measure.

*Proof.* We have

$$\begin{aligned} \mathbb{P}(\mathbf{X}^{(t)} \in \mathcal{X} | \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}) &= \mathbb{P}(\mathbf{X}^{(t)} \in \mathcal{X}, \text{ new value accepted} | \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}) \\ &\quad + \mathbb{P}(\mathbf{X}^{(t)} \in \mathcal{X}, \text{ new value rejected} | \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}) \\ &= \int_{\mathcal{X}} \alpha(\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)}) q(\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)}) d\mathbf{x}^{(t)} \\ &\quad + \underbrace{\frac{\mathbb{I}_{\mathcal{X}}(\mathbf{x}^{(t-1)})}{\int_{\mathcal{X}} \delta_{\mathbf{x}^{(t-1)}}(d\mathbf{x}^{(t)})}}_{= f_{\mathcal{X}}(1 - a(\mathbf{x}^{(t-1)})) \delta_{\mathbf{x}^{(t-1)}}(d\mathbf{x}^{(t)})} \underbrace{\mathbb{P}(\text{new value rejected} | \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)})}_{= 1 - a(\mathbf{x}^{(t-1)})} \\ &= \int_{\mathcal{X}} \alpha(\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)}) q(\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)}) d\mathbf{x}^{(t)} + \int_{\mathcal{X}} (1 - a(\mathbf{x}^{(t-1)})) \delta_{\mathbf{x}^{(t-1)}}(d\mathbf{x}^{(t)}) \quad \square \end{aligned}$$

**Proposition 5.3.** The Metropolis-Hastings kernel (5.3) satisfies the detailed balance condition

$$K(\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}) f(\mathbf{x}^{(t-1)}) = K(\mathbf{x}^{(t)}, \mathbf{x}^{(t-1)}) f(\mathbf{x}^{(t)})$$

and thus  $f(\mathbf{x})$  is the invariant distribution of the Markov chain  $(\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \dots)$  generated by the Metropolis-Hastings sampler. Furthermore the Markov chain is reversible.

<sup>1</sup> On a similar note, it is enough to know  $q(\mathbf{x}^{(t-1)} | \mathbf{x})$  up to a multiplicative constant independent of  $\mathbf{x}^{(t-1)}$  and  $\mathbf{x}$ .

*Proof.* We have that

$$\begin{aligned} \alpha(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})f(\mathbf{x}^{(t-1)}) &= \min \left\{ 1, \frac{f(\mathbf{x}^{(t)})q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{f(\mathbf{x}^{(t-1)})q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})} \right\} q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})f(\mathbf{x}^{(t-1)}) \\ &= \min \left\{ f(\mathbf{x}^{(t-1)})q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}), f(\mathbf{x}^{(t)})q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) \right\} \\ &= \min \left\{ \frac{f(\mathbf{x}^{(t-1)})q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})}{f(\mathbf{x}^{(t)})q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}, 1 \right\} q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})f(\mathbf{x}^{(t)}) = \alpha(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})f(\mathbf{x}^{(t)}) \end{aligned}$$

and thus

$$\begin{aligned} K(\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)})f(\mathbf{x}^{(t-1)}) &= \underbrace{\alpha(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})f(\mathbf{x}^{(t-1)})}_{=\alpha(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})f(\mathbf{x}^{(t)})} \\ &\quad + \underbrace{(1 - \alpha(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})) \delta_{\mathbf{x}^{(t-1)}}(\mathbf{x}^{(t)})}_{=0 \text{ if } \mathbf{x}^{(t)} \neq \mathbf{x}^{(t-1)}} f(\mathbf{x}^{(t-1)}) \\ &= \underbrace{(1 - \alpha(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})) \delta_{\mathbf{x}^{(t)}}(\mathbf{x}^{(t-1)})}_{(1 - \alpha(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})) \delta_{\mathbf{x}^{(t)}}(\mathbf{x}^{(t-1)})} f(\mathbf{x}^{(t-1)}) \\ &= K(\mathbf{x}^{(t)}, \mathbf{x}^{(t-1)})f(\mathbf{x}^{(t)}) \end{aligned}$$

The other conclusions follow by theorem 1.22, which also applies in the continuous case (see page 21).  $\square$

Next we need to examine whether the Metropolis-Hastings algorithm yields an irreducible chain. As with the Gibbs sampler, this does not need to be the case, as the following example shows.

**Example 5.1 (Reducible Metropolis-Hastings).** Consider using a Metropolis-Hastings algorithm for sampling from a uniform distribution on  $[0, 1] \cup [2, 3]$  and a  $U(x^{(t-1)} - \delta, x^{(t-1)} + \delta)$  distribution as proposal distribution  $q(\cdot|x^{(t-1)})$ . Figure 5.2 illustrates this example. It is easy to see that the resulting Markov chain is *not* irreducible if  $\delta \leq 1$ : in this case the chain either stays in  $[0, 1]$  or  $[2, 3]$ .  $\triangleleft$

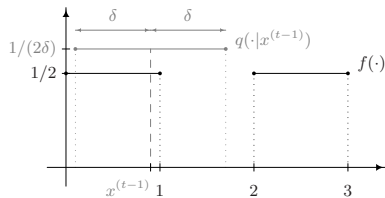


Figure 5.2. Illustration of example 5.1

Under mild assumptions on the proposal  $q(\cdot|x^{(t-1)})$  one can however establish the irreducibility of the resulting Markov chain:

- If  $q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})$  is positive for all  $\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)} \in \text{supp}(f)$ , then it is easy to see that we can reach any set of non-zero probability under  $f$  within a single step. The resulting Markov chain is thus strongly irreducible. Even though this condition seems rather restrictive, many popular choices of  $q(\cdot|x^{(t-1)})$  like multivariate Gaussians or  $t$ -distributions fulfil this condition.
- Roberts and Tweedie (1996) give a more general condition for the irreducibility of the resulting Markov chain: they only require that

$$\forall \epsilon \exists \delta : q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) > \epsilon \text{ if } \|\mathbf{x}^{(t-1)} - \mathbf{x}^{(t)}\| < \delta$$

together with the boundedness of  $f$  on any compact subset of its support.

The Markov chain  $(\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \dots)$  is further aperiodic, if there is positive probability that the chain remains in the current state, i.e.  $\mathbb{P}(\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)}) > 0$ , which is the case if

$$\mathbb{P} \left( f(\mathbf{X}^{(t-1)})q(\mathbf{X}|\mathbf{X}^{(t-1)}) > f(\mathbf{X})q(\mathbf{X}^{(t-1)}|\mathbf{X}) \right) > 0.$$

Note that this condition is *not* met if we use a “perfect” proposal which has  $f$  as invariant distribution: in this case we accept every proposed value with probability 1.

**Proposition 5.4.** *The Markov chain generated by the Metropolis-Hastings algorithm is Harris-recurrent if it is irreducible.*

*Proof.* Recurrence follows from the irreducibility and the fact that  $f$  is the unique invariant distribution (using proposition 1.28). For a proof of Harris recurrence see (Tierney, 1994).  $\square$

As we have now established (Harris-)recurrence, we are now ready to state an ergodic theorem (using theorem 1.30).

**Theorem 5.5.** *If the Markov chain generated by the Metropolis-Hastings algorithm is irreducible, then for any integrable function  $h : E \rightarrow \mathbb{R}$*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n h(\mathbf{X}^{(t)}) \rightarrow \mathbb{E}_f(h(\mathbf{X}))$$

for every starting value  $\mathbf{X}^{(0)}$ .

As with the Gibbs sampler the above ergodic theorem allows for inference using a single Markov chain.

### 5.3 The random walk Metropolis algorithm

In this section we will focus on an important special case of the Metropolis-Hastings algorithm: the random walk Metropolis-Hastings algorithm. Assume that we generate the newly proposed state  $\mathbf{X}$  not using the fairly general

$$\mathbf{X} \sim q(\cdot|\mathbf{X}^{(t-1)}), \quad (5.4)$$

from algorithm 5.1, but rather

$$\mathbf{X} = \mathbf{X}^{(t-1)} + \epsilon, \quad \epsilon \sim g, \quad (5.5)$$

with  $g$  being a *symmetric* distribution. It is easy to see that (5.5) is a special case of (5.4) using  $q(\mathbf{x}|\mathbf{x}^{(t-1)}) = g(\mathbf{x} - \mathbf{x}^{(t-1)})$ . When using (5.5) the probability of acceptance simplifies to

$$\min \left\{ 1, \frac{f(\mathbf{X}) \cdot q(\mathbf{X}^{(t-1)}|\mathbf{X})}{f(\mathbf{X}^{(t-1)}) \cdot q(\mathbf{X}|\mathbf{X}^{(t-1)})} \right\} = \min \left\{ 1, \frac{f(\mathbf{X})}{f(\mathbf{X}^{(t-1)})} \right\},$$

as  $q(\mathbf{X}|\mathbf{X}^{(t-1)}) = g(\mathbf{X} - \mathbf{X}^{(t-1)}) = g(\mathbf{X}^{(t-1)} - \mathbf{X}) = q(\mathbf{X}^{(t-1)}|\mathbf{X})$  using the symmetry of  $g$ . This yields the following algorithm which is a special case of algorithm 5.1, which is actually the original algorithm proposed by Metropolis et al. (1953).

**Algorithm 5.2 (Random walk Metropolis).** Starting with  $\mathbf{X}^{(0)} := (X_1^{(0)}, \dots, X_p^{(0)})$  and using a symmetric distribution  $g$ , iterate for  $t = 1, 2, \dots$

1. Draw  $\epsilon \sim g$  and set  $\mathbf{X} = \mathbf{X}^{(t-1)} + \epsilon$ .
2. Compute

$$\alpha(\mathbf{X}|\mathbf{X}^{(t-1)}) = \min \left\{ 1, \frac{f(\mathbf{X})}{f(\mathbf{X}^{(t-1)})} \right\}. \quad (5.6)$$

3. With probability  $\alpha(\mathbf{X}|\mathbf{X}^{(t-1)})$  set  $\mathbf{X}^{(t)} = \mathbf{X}$ , otherwise set  $\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)}$ .

*Example 5.2 (Bayesian probit model).* In a medical study on infections resulting from birth by Cesarean section (taken from Fahrmeir and Tutz, 2001) three influence factors have been studied: an indicator whether the Cesarean was planned or not ( $z_{i1}$ ), an indicator of whether additional risk factors were present at the time of birth ( $z_{i2}$ ), and an indicator of whether antibiotics were given as a prophylaxis ( $z_{i3}$ ). The response  $Y_i$  is the number of infections that were observed amongst  $n_i$  patients having the same influence factors (covariates). The data is given in table 5.1.

Number of births with infection		planned	risk factors	antibiotics
$y_i$	$n_i$	$z_{i1}$	$z_{i2}$	$z_{i3}$
11	98	1	1	1
1	18	0	1	1
0	2	0	0	1
23	26	1	1	0
28	58	0	1	0
0	9	1	0	0
8	40	0	0	0

**Table 5.1.** Data used in example 5.2

The data can be modeled by assuming that

$$Y_i \sim \text{Bin}(n_i, \pi_i), \quad \pi = \Phi(\mathbf{z}'_i \boldsymbol{\beta}),$$

where  $\mathbf{z}_i = (1, z_{i1}, z_{i2}, z_{i3})$  and  $\Phi(\cdot)$  being the CDF of the  $N(0, 1)$  distribution. Note that  $\Phi(t) \in [0, 1]$  for all  $t \in \mathbb{R}$ .

A suitable prior distribution for the parameter of interest  $\boldsymbol{\beta}$  is  $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbb{I}/\lambda)$ . The posterior density of  $\boldsymbol{\beta}$  is

$$f(\boldsymbol{\beta}|y_1, \dots, y_n) \propto \left( \prod_{i=1}^n \Phi(\mathbf{z}'_i \boldsymbol{\beta})^{y_i} \cdot (1 - \Phi(\mathbf{z}'_i \boldsymbol{\beta}))^{n_i - y_i} \right) \cdot \exp\left(-\frac{\lambda}{2} \sum_{j=0}^3 \beta_j^2\right)$$

We can sample from the above posterior distribution using the following random walk Metropolis algorithm. Starting with any  $\boldsymbol{\beta}^{(0)}$  iterate for  $t = 1, 2, \dots$ :

1. Draw  $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$  and set  $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t-1)} + \boldsymbol{\varepsilon}$ .
2. Compute

$$\alpha(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t-1)}) = \min\left\{1, \frac{f(\boldsymbol{\beta}|Y_1, \dots, Y_n)}{f(\boldsymbol{\beta}^{(t-1)}|Y_1, \dots, Y_n)}\right\}.$$

3. With probability  $\alpha(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t-1)})$  set  $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}$ , otherwise set  $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)}$ .

To keep things simple, we choose the covariance  $\boldsymbol{\Sigma}$  of the proposal to be  $0.08 \cdot \mathbb{I}$ .

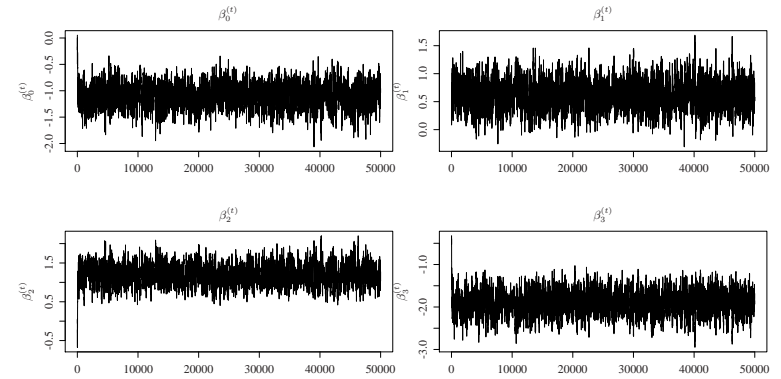
Figure 5.3 and table 5.2 show the results obtained using 50,000 samples<sup>2</sup>. Note that the convergence of the  $\beta_j^{(t)}$

		Posterior mean	95% credible interval	
intercept	$\beta_0$	-1.0952	-1.4646	-0.7333
planned	$\beta_1$	0.6201	0.2029	1.0413
risk factors	$\beta_2$	1.2000	0.7783	1.6296
antibiotics	$\beta_3$	-1.8993	-2.3636	-1.471

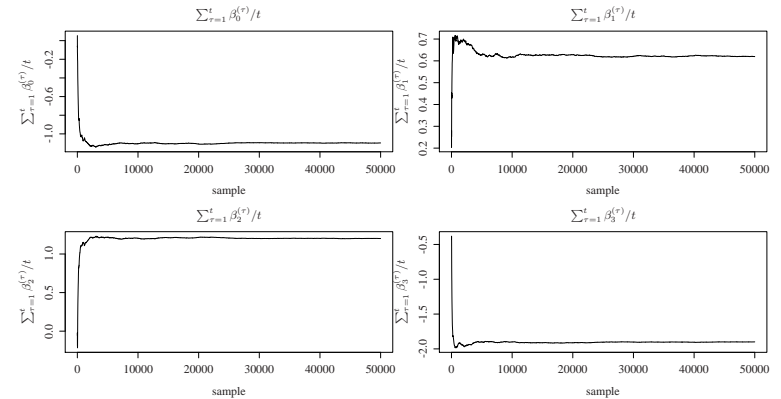
**Table 5.2.** Parameter estimates obtained for the Bayesian probit model from example 5.2

is to a distribution, whereas the cumulative averages  $\sum_{\tau=1}^t \beta_j^{(\tau)}/t$  converge, as the ergodic theorem implies, to a value. For figure 5.3 and table 5.2 the first 10,000 samples have been discarded (“burn-in”). ◁

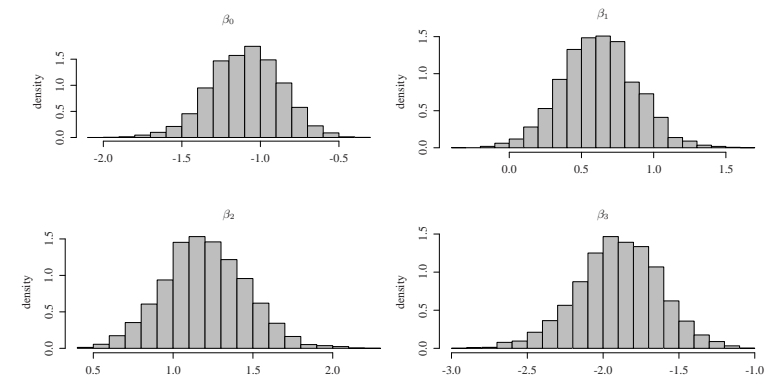
<sup>2</sup> You might want to consider a longer chain in practise.



(a) Sample paths of the  $\beta_j^{(t)}$



(b) Cumulative averages  $\sum_{\tau=1}^t \beta_j^{(\tau)}/t$



(c) Posterior distributions of the  $\beta_j$

**Figure 5.3.** Results obtained for the Bayesian probit model from example 5.2

## 5.4 Choosing the proposal distribution

The efficiency of a Metropolis-Hastings sampler depends on the choice of the proposal distribution  $q(\cdot|\mathbf{x}^{(t-1)})$ . An ideal choice of proposal would lead to a small correlation of subsequent realisations  $\mathbf{X}^{(t-1)}$  and  $\mathbf{X}^{(t)}$ . This correlation has two sources:

- the correlation between the current state  $\mathbf{X}^{(t-1)}$  and the newly proposed value  $\mathbf{X} \sim q(\cdot|\mathbf{X}^{(t-1)})$ , and
- the correlation introduced by retaining a value  $\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)}$  because the newly generated value  $\mathbf{X}$  has been rejected.

Thus we would ideally want a proposal distribution that both allows for fast changes in the  $\mathbf{X}^{(t)}$  and yields a high probability of acceptance. Unfortunately these are two competing goals. If we choose a proposal distribution with a small variance, the probability of acceptance will be high, however the resulting Markov chain will be highly correlated, as the  $X^{(t)}$  change only very slowly. If, on the other hand, we choose a proposal distribution with a large variance, the  $X^{(t)}$  can potentially move very fast, however the probability of acceptance will be rather low.

*Example 5.3.* Assume we want to sample from a  $N(0, 1)$  distribution using a random walk Metropolis-Hastings algorithm with  $\varepsilon \sim N(0, \sigma^2)$ . At first sight, we might think that setting  $\sigma^2 = 1$  is the optimal choice, this is however not the case. In this example we examine the choices:  $\sigma^2 = 0.1$ ,  $\sigma^2 = 1$ ,  $\sigma^2 = 2.38^2$ , and  $\sigma^2 = 10^2$ . Figure 5.4 shows the sample paths of a single run of the corresponding random walk Metropolis-Hastings algorithm. Rejected values are drawn as grey open circles. Table 5.3 shows the average correlation  $\rho(X^{(t-1)}, X^{(t)})$  as well as the average probability of acceptance  $\alpha(X|X^{(t-1)})$  averaged over 100 runs of the algorithm. Choosing  $\sigma^2$  too small yields a very high probability of acceptance, however at the price of a chain that is hardly moving. Choosing  $\sigma^2$  too large allows the chain to make large jumps, however most of the proposed values are rejected, so the chain remains for a long time at each accepted value. The results suggest that  $\sigma^2 = 2.38^2$  is the optimal choice. This corresponds to the theoretical results of Gelman et al. (1995).  $\triangleleft$

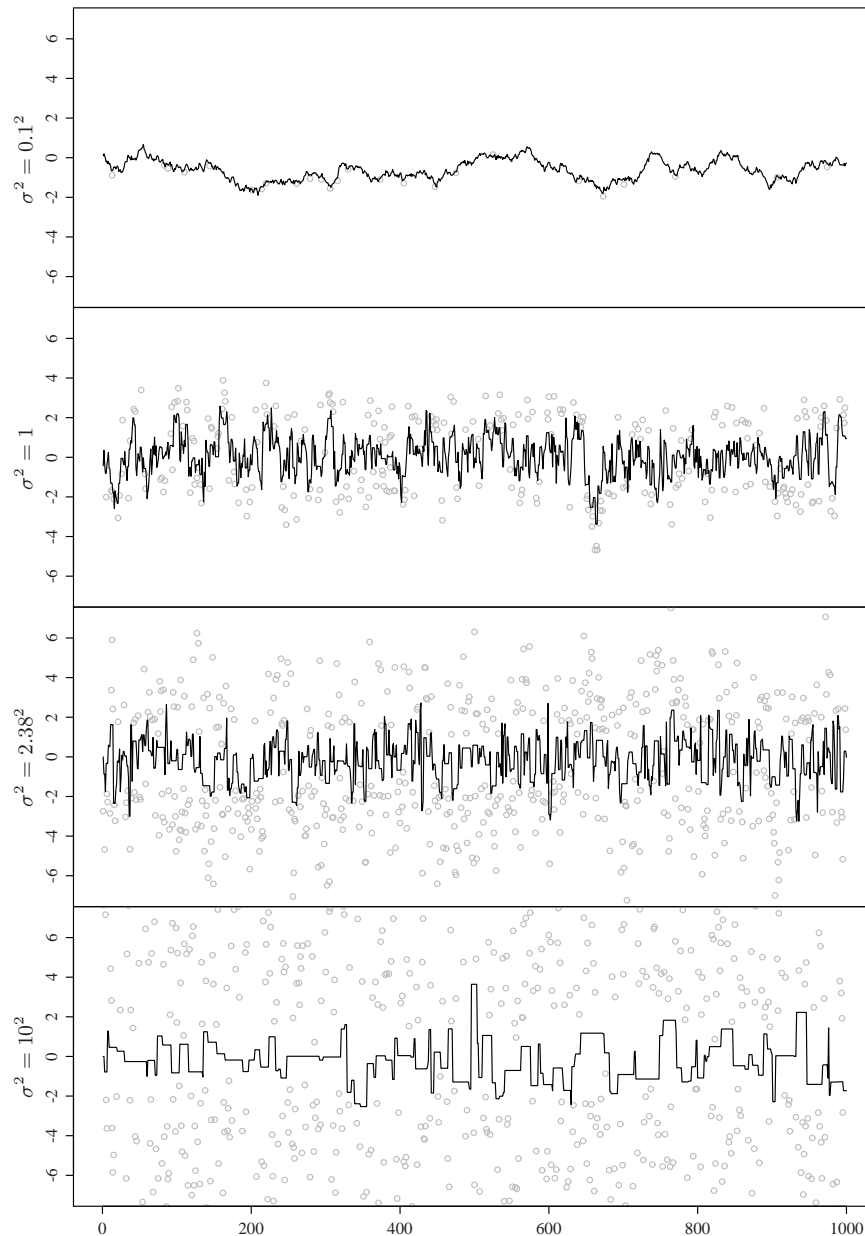
	Autocorrelation $\rho(X^{(t-1)}, X^{(t)})$		Probability of acceptance $\alpha(X, X^{(t-1)})$	
	Mean	95% CI	Mean	95% CI
$\sigma^2 = 0.1^2$	0.9901	(0.9891, 0.9910)	0.9694	(0.9677, 0.9710)
$\sigma^2 = 1$	0.7733	(0.7676, 0.7791)	0.7038	(0.7014, 0.7061)
$\sigma^2 = 2.38^2$	0.6225	(0.6162, 0.6289)	0.4426	(0.4401, 0.4452)
$\sigma^2 = 10^2$	0.8360	(0.8303, 0.8418)	0.1255	(0.1237, 0.1274)

**Table 5.3.** Average correlation  $\rho(X^{(t-1)}, X^{(t)})$  and average probability of acceptance  $\alpha(X|X^{(t-1)})$  found in example 5.3 for different choices of the proposal variance  $\sigma^2$ .

Finding the ideal proposal distribution  $q(\cdot|\mathbf{x}^{(t-1)})$  is an art.<sup>3</sup> This is the price we have to pay for the generality of the Metropolis-Hastings algorithm. Popular choices for random walk proposals are multivariate Gaussians or t-distributions. The latter have heavier tails, making them a safer choice. The covariance structure of the proposal distribution should ideally reflect the expected covariance of the  $(X_1, \dots, X_p)$ . Gelman et al. (1997) propose to adjust the proposal such that the acceptance rate is around 1/2 for one- or two dimensional target distributions, and around 1/4 for larger dimensions, which is in line with the results we obtained in the above simple example and the guidelines which motivate them. Note however that these are just rough guidelines.

*Example 5.4 (Bayesian probit model (continued)).* In the Bayesian probit model we studied in example 5.2 we drew

<sup>3</sup> The optimal proposal would be sampling directly from the target distribution. The very reason for using a Metropolis-Hastings algorithm is however that we cannot sample directly from the target!



**Figure 5.4.** Sample paths for example 5.3 for different choices of the proposal variance  $\sigma^2$ . Open grey discs represent rejected values.

$$\varepsilon \sim N(\mathbf{0}, \Sigma)$$

with  $\Sigma = 0.08 \cdot \mathbf{I}$ , i.e. we modeled the components of  $\varepsilon$  to be independent. The proportion of accepted values we obtained in example 5.2 was 13.9%. Table 5.4 (a) shows the corresponding autocorrelation. The resulting Markov chain can be made faster mixing by using a proposal distribution that represents the covariance structure of the posterior distribution of  $\beta$ .

This can be done by resorting to the frequentist theory of generalised linear models (GLM): it suggests that the asymptotic covariance of the maximum likelihood estimate  $\hat{\beta}$  is  $(\mathbf{Z}'\mathbf{D}\mathbf{Z})^{-1}$ , where  $\mathbf{Z}$  is the matrix of the covariates, and  $\mathbf{D}$  is a suitable diagonal matrix. When using  $\Sigma = 2 \cdot (\mathbf{Z}'\mathbf{D}\mathbf{Z})^{-1}$  in the algorithm presented in section 5.2 we can obtain better mixing performance: the autocorrelation is reduced (see table 5.4 (b)), and the proportion of accepted values obtained increases to 20.0%. Note that the determinant of both choices of  $\Sigma$  was chosen to be the same, so the improvement of the mixing behaviour is entirely due to a difference in the structure of the the covariance.  $\triangleleft$

(a) $\Sigma = 0.08 \cdot \mathbf{I}$				
	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$
Autocorrelation $\rho(\beta_j^{(t-1)}, \beta_j^{(t)})$	0.9496	0.9503	0.9562	0.9532

(b) $\Sigma = 2 \cdot (\mathbf{Z}'\mathbf{D}\mathbf{Z})^{-1}$				
	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$
Autocorrelation $\rho(\beta_j^{(t-1)}, \beta_j^{(t)})$	0.8726	0.8765	0.8741	0.8792

**Table 5.4.** Autocorrelation  $\rho(\beta_j^{(t-1)}, \beta_j^{(t)})$  between subsequent samples for the two choices of the covariance  $\Sigma$ .