

# Lecture 7: The Metropolis-Hastings Algorithm

Nick Whiteley

## What we have seen last time: Gibbs sampler

- Key idea: Generate a Markov chain by updating the component of  $(X_1, \dots, X_p)$  in turn by drawing from the full conditionals:

$$X_j^{(t)} \sim f_{X_j|X_{-j}}(\cdot | X_1^{(t)}, \dots, X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, \dots, X_p^{(t-1)})$$

- Two drawbacks:
  - Requires that it is possible / easy to sample from the full conditionals.
  - Can yields a slowly mixing chain if (some of) the components of  $(X_1, \dots, X_p)$  are highly correlated.

## What we will see today: Metropolis-Hastings algorithm

- Key idea: Use rejection mechanism, with a “local proposal”:  
We let the newly proposed  $\mathbf{X}$  depend on the previous state of the chain  $\mathbf{X}^{(t-1)}$ .
- Samples  $(\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \dots)$  form a Markov chain (like the Gibbs sampler).

# 5.1 Algorithm

# The Metropolis-Hastings algorithm

## Algorithm 5.1: Metropolis-Hastings

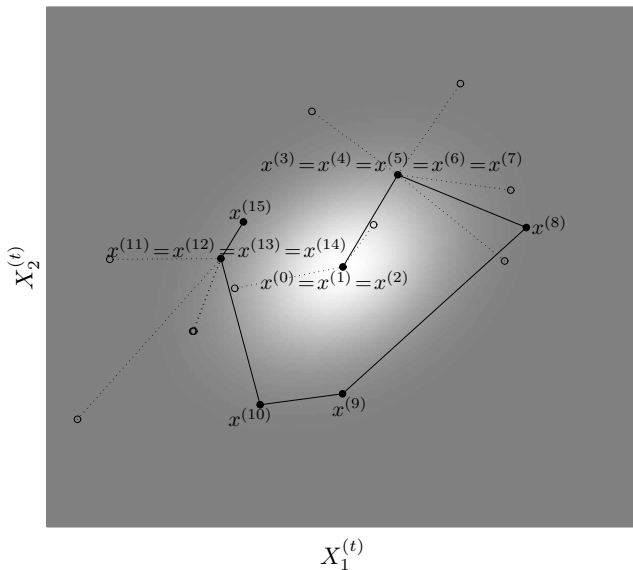
Starting with  $\mathbf{X}^{(0)} := (X_1^{(0)}, \dots, X_p^{(0)})$  iterate for  $t = 1, 2, \dots$

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

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

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

# Illustration of the Metropolis-Hastings method



# Basic properties of the Metropolis-Hastings algorithm

- The probability that a newly proposed value is accepted given  $\mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}$  is

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

- The probability of remaining in state  $\mathbf{X}^{(t-1)}$  is

$$\mathbb{P}(\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)} | \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}) = 1 - a(\mathbf{x}^{(t-1)}).$$

- The probability of acceptance does not depend on the normalisation constant:

If  $f(\mathbf{x}) = C \cdot \pi(\mathbf{x})$ , then

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

# The Metropolis-Hastings Transition Kernel

## Lemma 5.1

The transition kernel of the Metropolis-Hastings algorithm is

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

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

## 5.2 Convergence properties



# Theoretical properties

## Proposition 5.1

The Metropolis-Hastings kernel satisfies the *detailed balance condition*

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

Thus  $f(\mathbf{x})$  is the invariant distribution of the Markov chain  $(\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \dots)$ . Furthermore the Markov chain is reversible.

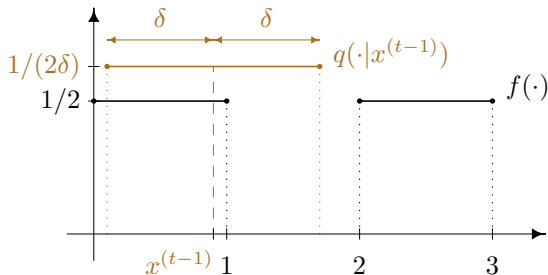
## Example 5.1: Reducible Metropolis-Hastings

Consider the target distribution

$$f(x) = (\mathbb{I}_{[0,1]}(x) + \mathbb{I}_{[2,3]}(x))/2.$$

and the proposal distribution  $q(\cdot | \mathbf{x}^{(t-1)})$ :

$$X | X^{(t-1)} = x^{(t-1)} \sim \text{U}[x^{(t-1)} - \delta, x^{(t-1)} + \delta]$$



Reducible if  $\delta \leq 1$ : the chain stays either in  $[0, 1]$  or  $[2, 3]$ .



## Further theoretical properties

- The Markov chain  $(\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \dots)$  is irreducible if  $q(\mathbf{x}|\mathbf{x}^{(t-1)}) > 0$  for all  $\mathbf{x}, \mathbf{x}^{(t-1)} \in \text{supp}(f)$ : every state can be reached in a single step.  
(less strict conditions can be obtained, see e.g. (see Roberts & Tweedie, 1996))
- The chain is aperiodic, if there is positive probability that the chain remains in the current state, i.e.  $\mathbb{P}(\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)}) > 0$ ,

# An ergodic theorem

## Theorem 5.1

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  $\mathbf{X}^{(0)}$ .

Interpretation: We can approximate expectations by their empirical counterparts using a single Markov chain.

## 5.3 Random-walk Metropolis

### 5.4 Choosing the proposal distribution

# Random-walk Metropolis: Idea

- In the Metropolis-Hastings algorithm the proposal is from  $\mathbf{X} \sim q(\cdot | \mathbf{X}^{(t-1)})$ .
- A popular choice for the proposal is  $q(\mathbf{x} | \mathbf{x}^{(t-1)}) = g(\mathbf{x} - \mathbf{x}^{(t-1)})$  with  $g$  being a *symmetric* distribution, thus

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

- Probability of acceptance becomes

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

- We accept ...
  - every move to a more probable state with probability 1.
  - moves to less probable states with a probability  $f(\mathbf{X})/f(\mathbf{x}^{(t-1)}) < 1$ .

# Random-walk Metropolis: Algorithm

## Random-Walk Metropolis

Starting with  $\mathbf{X}^{(0)} := (X_1^{(0)}, \dots, X_p^{(0)})$  and using a symmetric random walk proposal  $g$ , iterate for  $t = 1, 2, \dots$

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

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

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

Popular choices for  $g$  are (multivariate) Gaussians or t-distributions (the latter having heavier tails)

## Example 5.2: Bayesian probit model (1)

- Medical study on infections resulting from birth by Cesarean section
- 3 influence factors:
  - indicator whether the Cesarean was planned or not ( $z_{i1}$ ),
  - indicator of whether additional risk factors were present at the time of birth ( $z_{i2}$ ), and
  - indicator of whether antibiotics were given as a prophylaxis ( $z_{i3}$ ).
- Response variable: number of infections  $Y_i$  that were observed amongst  $n_i$  patients having the same covariates.

# births		planned	risk factors	antibiotics
infection	total			
$y_i$	$n_i$	$z_{i1}$	$z_{i2}$	$z_{i3}$
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



## Example 5.2: Bayesian probit model (2)

- Model for  $Y_i$ :

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

where  $\mathbf{z}_i = [1, z_{i1}, z_{i2}, z_{i3}]^T$  and  $\Phi(\cdot)$  being the CDF of a  $N(0, 1)$ .

- Prior on the parameter of interest  $\boldsymbol{\beta}$ :  $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbb{I}/\lambda)$ .
- The posterior density of  $\boldsymbol{\beta}$  is

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

## Example 5.2: Bayesian probit model (3)

Use the following random walk Metropolis algorithm (50,000 samples).

Starting with any  $\beta^{(0)}$  iterate for  $t = 1, 2, \dots$ :

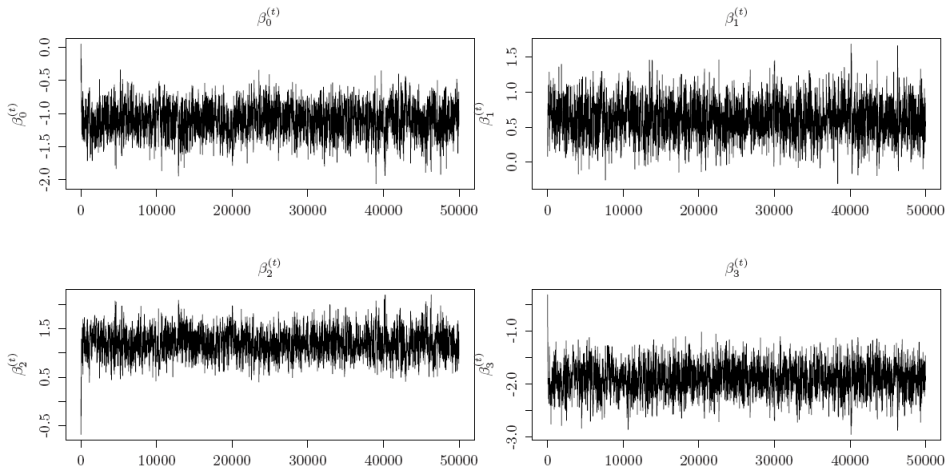
1. Draw  $\epsilon \sim N(\mathbf{0}, \Sigma)$  and set  $\beta = \beta^{(t-1)} + \epsilon$ .
2. Compute

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

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

(for the moment we use  $\Sigma = 0.08 \cdot \mathbb{I}$ , and  $\lambda = 10$ ).

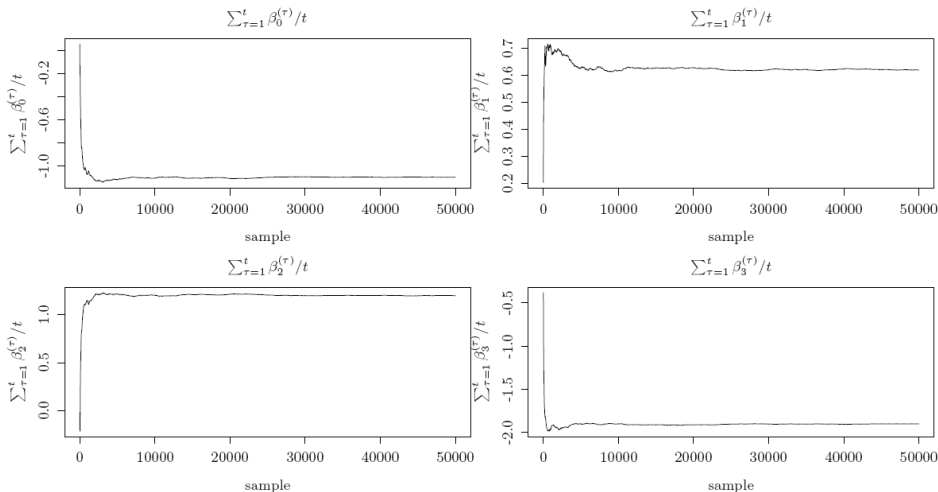
## Example 5.2: Bayesian probit model (4)



Convergence of the  $\beta_j^{(t)}$  is to a distribution, not a value!

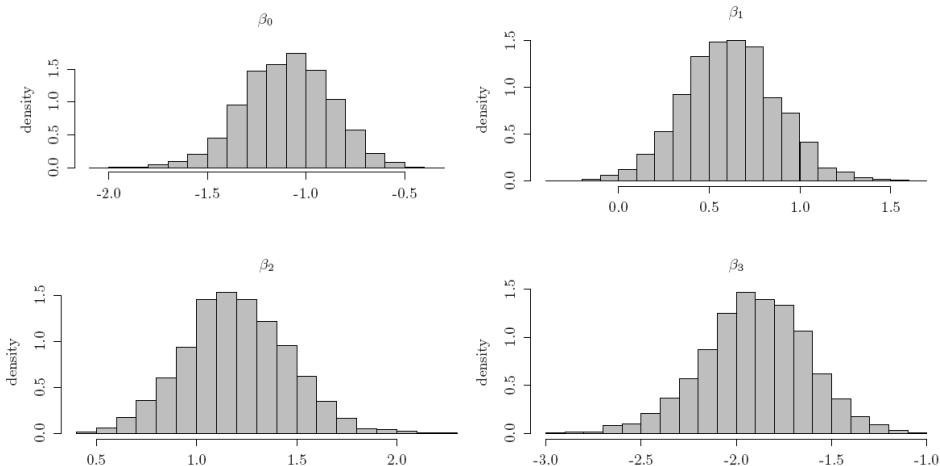


## Example 5.2: Bayesian probit model (5)



Convergence of cumulative averages  $\sum_{\tau=1}^t \beta_j^{(\tau)} / t$  is to a value.

## Example 5.2: Bayesian probit model (6)



## Example 5.2: Bayesian probit model (7)

		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

# Choosing a good proposal distribution

- Ideally: Markov chain with small correlation  $\rho(\mathbf{X}^{(t-1)}, \mathbf{X}^{(t)})$  between subsequent values.  
 $\leadsto$  fast exploration of the support of the target  $f$ .
- Two sources for this correlation:
  - the correlation between the current state  $\mathbf{X}^{(t-1)}$  and the newly proposed value  $\mathbf{X} \sim q(\cdot | \mathbf{X}^{(t-1)})$   
(can be reduced using a proposal with high variance)
  - the correlation introduced by retaining a value  $\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)}$  because the newly generated value  $\mathbf{X}$  has been rejected  
(can be reduced using a proposal with small variance)
- Trade-off for finding the ideal compromise between:
  - fast exploration of the space (good mixing behaviour)
  - obtaining a large probability of acceptance
- For multivariate distributions: covariance of proposal should reflect the covariance structure of the target.

## Example 5.3: Choice of proposal (1)

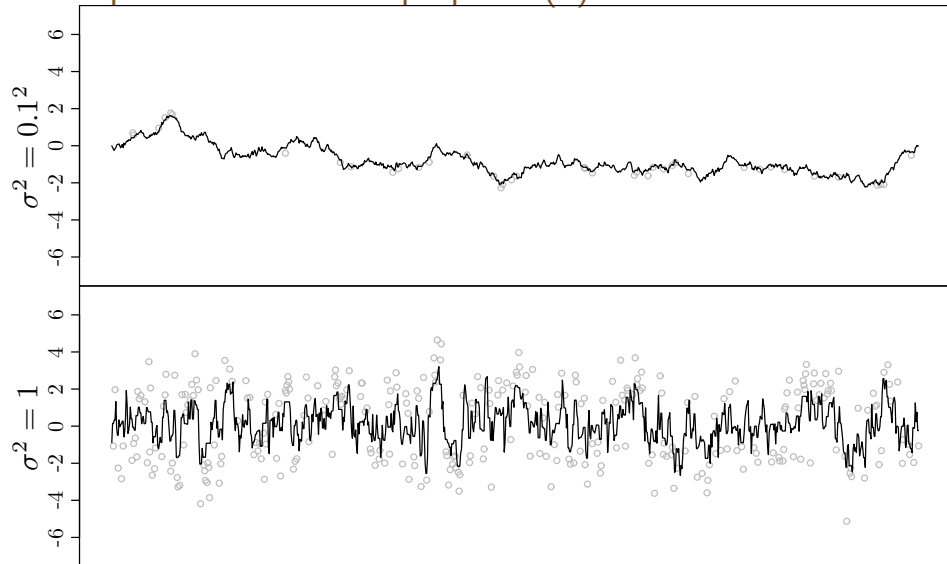
- Target distribution, we want to sample from:  $N(0, 1)$  (i.e.  $f(\cdot) = \phi_{(0,1)}(\cdot)$ )
- We want to use a random walk Metropolis algorithm with

$$\varepsilon \sim N(0, \sigma^2)$$

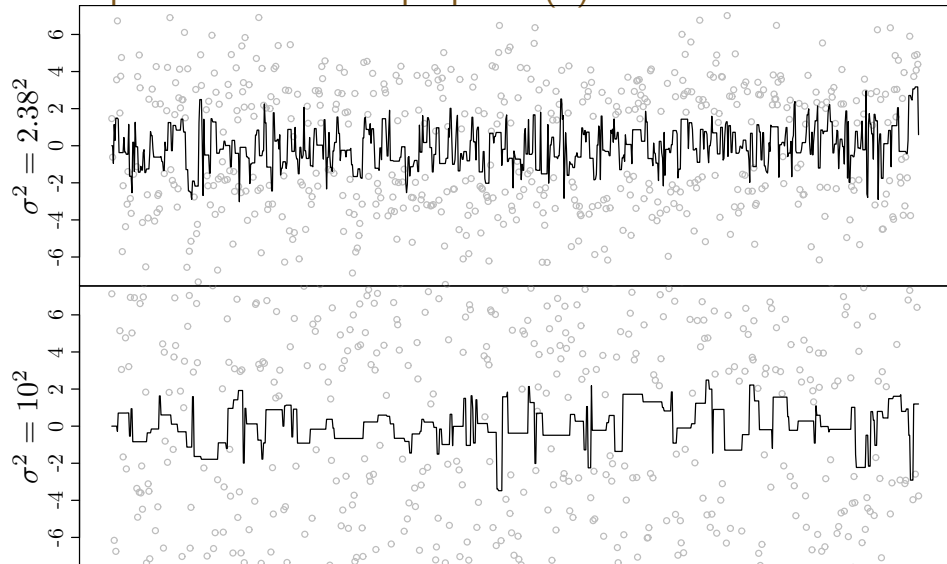
- What is the optimal choice of  $\sigma^2$ ?
- We consider four choices  $\sigma^2 = 0.1^2, 1, 2.38^2, 10^2$ .



## Example 5.3: Choice of proposal (2)



## Example 5.3: Choice of proposal (3)



## Example 5.3: Choice of proposal (4)

	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)

Suggests: Optimal choice is  $2.38^2 > 1$ .

## Example 5.4: Bayesian probit model (revisited)

- So far we used:  $\text{Var}(\epsilon) = 0.08 \cdot \mathbb{I}$ .
- Better choice: Let  $\text{Var}(\epsilon)$  reflect the covariance structure
- Frequentist asymptotic theory:  $\text{Var}(\hat{\beta}^{\text{m.l.e}}) = (\mathbf{Z}'\mathbf{D}\mathbf{Z})^{-1}$   
 $\mathbf{D}$  is a suitable diagonal matrix
- Better choice:  $\text{Var}(\epsilon) = 2 \cdot (\mathbf{Z}'\mathbf{D}\mathbf{Z})^{-1}$
- Increases rate of acceptance from 13.9% to 20.0% and reduces autocorrelation:

$\Sigma = 0.08 \cdot \mathbf{I}$	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$
Autocorrelation $\rho(\beta_j^{(t-1)}, \beta_j^{(t)})$	0.9496	0.9503	0.9562	0.9532
$\Sigma = 2 \cdot (\mathbf{Z}'\mathbf{D}\mathbf{Z})^{-1}$	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$
Autocorrelation $\rho(\beta_j^{(t-1)}, \beta_j^{(t)})$	0.8726	0.8765	0.8741	0.8792

(in this example  $\det(0.08 \cdot \mathbb{I}) = \det(2 \cdot (\mathbf{Z}'\mathbf{D}\mathbf{Z})^{-1})$ )