Geometric ergodicity of Metropolis-Hastings algorithms for conditional simulation in generalised linear mixed models

O. F. CHRISTENSEN olefc@math.auc.dk
Department of Mathematical Sciences, Aalborg University, 9220 Aalborg Øst, Denmark

J. MØLLER jm@math.auc.dk
Department of Mathematical Sciences, Aalborg University, 9220 Aalborg Øst, Denmark

R. P. WAAGEPETERSEN rw@math.auc.dk
Department of Mathematical Sciences, Aalborg University, 9220 Aalborg Øst, Denmark

Abstract. Conditional simulation is useful in connection with inference and prediction for a generalised linear mixed model. We consider random walk Metropolis and Langevin-Hastings algorithms for simulating the random effects given the observed data, when the joint distribution of the unobserved random effects is multivariate Gaussian. In particular we study the desirable property of geometric ergodicity, which ensures the validity of central limit theorems for Monte Carlo estimates.

Keywords: conditional simulation; generalised linear mixed model; geometric ergodicity; Langevin-Hastings algorithm; Markov chain Monte Carlo; random walk Metropolis algorithm.

MSC2000: 65C05, 65C40, 62J12

1 Introduction

Generalised linear mixed models (GLMMs) (see e.g. Breslow and Clayton, 1993) with correlated random effects are important for modelling of many types of correlated data. In particular Diggle *et al.* (1998) use GLMMs to model spatially correlated binary and count data, and this is the main inspiration for our work. Conditional simulation of the random effects given observed data from a GLMM is useful in connection with prediction and Monte Carlo maximum likelihood estimation, see Diggle *et al.* (1998) and McCulloch (1997), respectively.

In this paper we consider random walk Metropolis and Langevin-Hastings algorithms for conditional simulation in a GLMM with correlated Gaussian random effects. In particular, we study the desirable property of geometric ergodicity. Geometric ergodicity ensures the validity of central limit theorems for Monte Carlo estimates, and justifies assessment of the precision of a Monte Carlo estimate by estimation of the asymptotic variance in the limiting normal distribution (see e.g. Roberts and Rosenthal, 1998a). In an empirical study we compare asymptotic variances for random walk Metropolis and Langevin-Hastings algorithms and demonstrate the advantage of using the Langevin-Hastings algorithm.

The paper builds upon results on geometric ergodicity of Langevin-Hastings and random walk Metropolis algorithms in Roberts and Tweedie (1996a) and Jarner and Hansen (2000), respectively. The examples of target densities considered in these papers are useful to illustrate the results, but do not relate much to applications in statistics where Markov chain Monte Carlo (MCMC) methods are used. In contrast, we consider a specific class of target densities, which is useful in practice for many types of data.

Section 2 outlines generalised linear mixed models and describes the random walk Metropolis and Langevin-Hastings algorithms. Our results on geometric ergodicity are presented in Section 3. Section 4 contains an empirical comparison between the two algorithms. A discussion of other Langevin-type algorithms is given in Section 5.

2 Background

2.1 Generalised linear mixed models with correlated random effects

Generalised linear mixed models (GLMMs) (Breslow and Clayton, 1993; Lee and Nelder, 1996) are extensions of generalised linear models (GLMs) (McCullagh and Nelder, 1989) that allow additional sources of variability due to unobservable random effects. In this article we consider GLMMs where the joint distribution of the random effects is multivariate Gaussian. Such models and the notation used throughout this paper are briefly described below.

Suppose $y = (y_1, \dots, y_n)^{\mathrm{T}}$ is a realisation of an n-dimensional random vector $Y = (Y_1, \dots, Y_n)^{\mathrm{T}}$, S_i is an unobserved random variable associated with Y_i, $i = 1, \dots, n$, and $S_{n+1}, \dots, S_q$ are additional unobserved variables — in a spatial setting the additional variables may correspond to locations chosen for prediction, or they may be auxiliary variables introduced for computational convenience. We assume the following: $S = (S_1, \dots, S_n, S_{n+1}, \dots, S_q)^{\mathrm{T}}$ follows a q-dimensional normal distribution with mean zero and covariance matrix Σ; the conditional distribution of Y_i given S has a density $f(y_i; M_i)$ with respect to counting or Lebesgue measure, which only depends on the conditional mean $M_i = \mathrm{E}[Y_i|S_i]$, $i = 1, \dots, n$; and $Y_1, \dots, Y_n$ are conditionally independent given S. So the conditional density of $Y|S$ is

$$f(y|S) = \prod_{i=1}^{n} f(y_i; M_i). \tag{1}$$

We restrict attention to the case where the density $f(\cdot;\mu)$ is of an exponential family form

$$f(z;\mu) = \exp\left(z g_c(\mu) + b(z) - a(g_c(\mu))\right), \quad z \in \Omega, \tag{2}$$

where $\Omega \subseteq \mathbb{R}$ is the support of the density, μ is the mean parameter, and a, b, g_c are real functions; g_c is called the canonical link function. We assume that M_i is related to S_i by a link function g so that

$$g(M_i) = S_i + d_i^{\mathrm{T}}\beta, \tag{3}$$

where $d_i \in \mathbb{R}^p$ is a vector of covariates, $\beta \in \mathbb{R}^p$ is a vector of regression parameters, and the superscript T denotes transposition of vectors and matrices. Note that g is part of the model specification. We assume that g is strictly increasing, continuous and two times differentiable; these conditions are satisfied in the special case where $g = g_c$. By (3), the range of $g(M_i)$ must be the entire real line. So the mean parameter space $g^{-1}(\mathbb{R})$ is an open interval denoted $\mathcal{M} =]m_1; m_2[$.

GLMMs form a flexible class of models for many types of non-Gaussian data including count, binary and positive data. Many types of non-linear dependencies between the conditional means M_i and the linear predictors $S_i + d_i^{\mathrm{T}}\beta$, $i = 1, \ldots, n$, can further be specified through the link function g. We consider in particular the three following cases:

(*i*) The Poisson-log normal model where

$$f(z;\mu) = \exp\left(z \log\mu - \log(z!) - \mu\right), \quad z = 0, 1, \ldots, \tag{4}$$

is a Poisson density, $\mathcal{M} =]0; \infty[$, and $g(\mu) = g_c(\mu) = \log\mu$ is the canonical log-link.

(*ii*) The binomial-logit model where

$$f(z;\mu) = \exp\left(z\log(\mu/(N-\mu)) + \log\left(N^{-N}\binom{N}{z}\right) + N\log(N-\mu)\right), \quad z = 0, \ldots, N,$$

is a binomial density, $N > 0$ is a given integer, $\mathcal{M} =]0; N[$, and $g(\mu) = g_c(\mu) = \log(\mu/(N-\mu))$ is the canonical logit-link.

(*iii*) The exponential-log model where

$$f(z;\mu) = \exp(-z/\mu - \log\mu), \quad z > 0,$$

is an exponential density, $\mathcal{M} =]0; \infty[$ and $g(\mu) = \log\mu$ is the log-link. Here the link function $g = g_c$ would not have been valid, since the range of $g_c(\mu) = -1/\mu$ is strictly contained in $\mathbb{R}$.

In a Bayesian analysis one would introduce priors for β and Σ, but throughout this paper we consider β and Σ as fixed.

2.2 Description of algorithms

Diggle *et al.* (1998) use a so-called single-site updating algorithm for their posterior simulations. This is computationally demanding since each update of a random effect involves the computation of the conditional variance given the other random effects. Here we consider two algorithms where all the random effects are updated simultaneously.

Suppose that $\Sigma = KK^{\mathrm{T}}$ where K is a $q \times d$ matrix, and let $S = K\Gamma$ where Γ follows a d-dimensional standard normal distribution. For example, in a spatial setting we typically have $d = q$ and K equal to a square root matrix of Σ (see Section 4 for further details), while in a variance component model, $K = E\mathrm{diag}\{\sigma_1, \ldots, \sigma_d\}$ where E is a design matrix and $\sigma_i > 0$, $i = 1, \ldots, d$. Simulations of $S|Y = y$ can be obtained by transforming simulations from the distribution of $\Gamma \mid Y = y$, and this may be advantageous when Σ is not strictly positive definite.

Note that by (1), the log density of $\Gamma|Y = y$ is

$$\log f(\gamma|y) = \mathrm{const}(y) - \frac{1}{2}\|\gamma\|^2 + \sum_{i=1}^{n} \log f(y_i; \mu_i), \tag{5}$$

with

$$\mu_i = \mu_i(s) = g^{-1}(s_i + d_i^{\mathrm{T}}\beta) \tag{6}$$

where $s = (s_1, \ldots, s_n)^{\mathrm{T}} = Q\gamma$ and Q denotes the upper $n \times d$ submatrix of K.

2.2.1 Gaussian random walk Metropolis algorithm

The updates in this algorithm are given by two steps. Suppose that γ is the current state. First, a proposal γ' is generated from a multivariate normal distribution with mean γ and covariance matrix hI, where $h > 0$ is a user-specified parameter. Secondly, we return γ' with probability

$$\alpha(\gamma, \gamma') = 1 \wedge \frac{f(\gamma'|y)}{f(\gamma|y)};$$

otherwise the state γ is retained.

Using Lemmas 1.1. and 1.2. in Mengersen and Tweedie (1996) and Corollary 2 in Tierney (1994) one can verify that the algorithm produces an ergodic Markov chain with stationary distribution given by $f(\cdot|y)$.

2.2.2 Langevin-Hastings algorithm

More efficient algorithms can be obtained by adapting the proposal density to the target distribution. Consider the Langevin diffusion given by the stochastic differential equation

$$d\Gamma_t = \frac{1}{2}\nabla(\Gamma_t)dt + dW_t,\ t > 0, \tag{7}$$

where $(W_t)_{t\geq 0}$ is a multidimensional Wiener process and

$$\nabla(\gamma) = \frac{\partial}{\partial\gamma}\log f(\gamma|y) = -\gamma + Q^{\mathrm{T}}\left\{(y_i-\mu_i)\frac{g_c'(\mu_i)}{g'(\mu_i)}\right\}_{i=1}^{n} \tag{8}$$

is the gradient of the log target density. Under weak conditions, the Langevin diffusion has $f(\cdot|y)$ as equilibrium density. This is the initial motivation for constructing Metropolis-Hastings algorithms based on introducing Metropolis-Hastings accept/reject steps in discretisations of the Langevin diffusion.

The Langevin-Hastings algorithm considered in this paper is based on a first-order discretisation of the Langevin diffusion, the so-called Euler-discretisation (see Roberts and Tweedie, 1996a). The proposal distribution is thus a multivariate normal distribution with mean $\xi(\gamma) = \gamma + (h/2)\nabla(\gamma)$ and covariance matrix hI, $h > 0$, and the acceptance probability is

$$\alpha(\gamma,\gamma') = 1 \wedge \frac{f(\gamma'|y)\exp(-\frac{1}{2h}\left\|\gamma-\xi(\gamma')\right\|^2)}{f(\gamma|y)\exp(-\frac{1}{2h}\left\|\gamma'-\xi(\gamma)\right\|^2)}. \tag{9}$$

By the Lemmas 1.1. and 1.2. in Mengersen and Tweedie (1996) and Corollary 2 in Tierney (1994), the resulting Markov chain is ergodic with stationary distribution $f(\cdot|y)$.

Using the gradient to adapt the proposal kernel to the target density may lead to much better convergence and mixing properties than for an ordinary random walk Metropolis chain, see Roberts and Rosenthal (1998b) and Section 4. By Roberts *et al.* (1997) and Roberts and Rosenthal (1998b), the number of iterations required to obtain convergence is $O(d^{-1})$ for the random walk algorithm and $O(d^{-1/3})$ for the Langevin-Hastings algorithm, so the benefit of using Langevin-Hastings increases as the dimension increases.

3 Geometric ergodicity

An MCMC algorithm is geometrically ergodic if there exist a nonnegative function V and constants $0 < r < \infty$, $0 < \rho < 1$, such that for any state γ,

$$\sup_{|\chi|\leq V}\left|\int \chi(\gamma')P^{(m)}(\gamma,d\gamma') - E[\chi(\Gamma)|y]\right| \leq V(\gamma)r\rho^m, \quad m = 1,2,\ldots \tag{10}$$

where $P^{(m)}$ is the m-step transition kernel for the Markov chain. An important implication of geometric ergodicity is that a central limit theorem (CLT) holds for the ergodic averages (Roberts and Tweedie, 1996b). Let $\Gamma(j)$, $j \geq 0$ be the Markov chain started at an arbitrary initial state $\Gamma(0)$, and suppose that ψ is a function with $\psi^2 \leq V$. Then there exist a number $\sigma_\psi^2 > 0$ such that

$$\sqrt{m}\left(\frac{1}{m}\sum_{j=1}^{m}\psi(\Gamma(j)) - \mathrm{E}[\psi(\Gamma)|y]\right) \xrightarrow{\sim} N(0,\sigma_\psi^2). \tag{11}$$

In Section 3.1 below we verify that the random walk Metropolis algorithm is geometrically ergodic for a large class of GLMMs including the models (*i*)-(*iii*). The situation is different for the Langevin-Hastings algorithm which is not geometrically ergodic for the Poisson-log normal model (*i*) and the exponential-log normal model (*iii*). In Section 3.2 we show that a truncated version of the Langevin-Hastings algorithm is geometrically ergodic for any GLMM.

The log-density (5) differs from the types of densities considered in Roberts and Tweedie (1996a) and Jarner and Hansen (2000) by the likelihood term $\sum_{i=1}^{n} \log f(y_i; \mu_i)$. The geometric ergodicity results rely much on the term $-\frac{1}{2}||\gamma||^2$ in (5), i.e. on Γ being a priori Gaussian, while the main effort in the proofs of geometric ergodicity is to control the likelihood term.

3.1 Geometric ergodicity of random walk Metropolis

In the following we let $\nabla(\cdot)$ be as in (8) and define $n(\gamma) = \gamma/||\gamma||$ and $m(\gamma) = \nabla(\gamma)/||\nabla(\gamma)||$.

Geometric ergodicity of random walk Metropolis algorithms is studied in Jarner and Hansen (2000) for target densities which are 'super-exponential', i.e. densities for which

$$n(\gamma) \cdot \nabla(\gamma) \to -\infty \quad \text{as} \quad ||\gamma|| \to \infty. \tag{12}$$

Theorem 4.3 in Jarner and Hansen (2000) states that the random walk Metropolis algorithm is geometrically ergodic if, in addition to (12),

$$\limsup_{||\gamma|| \to \infty} n(\gamma) \cdot m(\gamma) < 0. \tag{13}$$

Using (8) we see that

$$n(\gamma) \cdot \nabla(\gamma) = -||\gamma|| + \frac{1}{||\gamma||} \sum_{i=1}^{n} s_i (y_i - g^{-1}(s_i + d_i^{\mathrm{T}} \beta)) \frac{g_c'(g^{-1}(s_i + d_i^{\mathrm{T}} \beta))}{g'(g^{-1}(s_i + d_i^{\mathrm{T}} \beta))}.$$

Combining this with the fact that g_c is increasing, it can be seen that $f(\cdot|y)$ is super-exponential when $g = g_c$ as in the models (*i*) and (*ii*). Also for the model (*iii*) where g is not the canonical link function, $f(\cdot|y)$ is super-exponential. Condition (13) holds for model (*ii*), but not for the models (*i*) and (*iii*). The following theorem, however, ensures geometric ergodicity in any of the cases (*i*)–(*iii*), see Remark 1 below.

Theorem 1. *Assume that*

$$\limsup_{\mu \to m_1, m_2} \frac{1}{g'(\mu)} \left| \frac{1}{\mu - y_i} + \frac{g_c''(\mu)}{g_c'(\mu)} - \frac{g''(\mu)}{g'(\mu)} \right| < \infty, \tag{14}$$

and

$$\limsup_{\mu \to m_1, m_2} \frac{y_i - \mu}{g(\mu)} \frac{g_c'(\mu)}{g'(\mu)} \leq 0, \tag{15}$$

hold for $i = 1, \dots, n$ (where $\limsup_{\mu \to m_1, m_2} \cdots = \max\{\limsup_{\mu \to m_1} \cdots, \limsup_{\mu \to m_2} \cdots\}$). Assume also that the covariance matrix QQ^T of $(S_1, \dots, S_n)$ is invertible. Then the Gaussian random walk Metropolis algorithm for conditional simulation of $\Gamma | Y = y$ is geometrically ergodic, with V in (10) equal to $V(\gamma) = f(\gamma|y)^{-1/2}$.

Proof. Geometric ergodicity follows from Theorem 4.1 in Jarner and Hansen (2000) provided (I) $f(\cdot \mid y)$ is super-exponential and (II)

$$\liminf_{\|\gamma\| \to \infty} \int_{A(\gamma)} q(\gamma'; \gamma, h) d\gamma' > 0, \tag{16}$$

where $q(\gamma'; \gamma, h) \propto \exp(-\|\gamma' - \gamma\|^2/(2h))$ denotes the Gaussian proposal density and

$$A(\gamma) = \{\gamma' \in \mathbb{R}^d \mid f(\gamma'|y) \geq f(\gamma|y)\}$$

is the acceptance region for $\gamma \in \mathbb{R}^d$.

Re (I) Define $R(\gamma)$ to be equal to $\{\cdots\}$ in (8). Super-exponentiality is then implied by the inequality

$$\limsup_{\|\gamma\| \to \infty} (Q\gamma)^{\mathrm{T}} R(\gamma)/\|\gamma\|^2 \leq 0. \tag{17}$$

To verify (17) we fix $i \in \{1, \dots, n\}$ and let $r_i(s_i) = R(\gamma)_i$, where $s = Q\gamma$. From (8) it follows that

$$s_i r_i(s_i) = \frac{y_i - \mu_i}{g(\mu_i)} \frac{g_c'(\mu_i)}{g'(\mu_i)} \frac{g(\mu_i)}{s_i} s_i^2,$$

where $\mu_i = g^{-1}(s_i + d_i^{\mathrm{T}} \beta)$. By (15) and the fact that $g(\mu_i)/s_i \to 1$ when $|s_i| \to \infty$, we see that for a given $\epsilon > 0$, $s_i r_i(s_i) < \epsilon s_i^2$, when $|s_i|$ is sufficiently large. Hence, by continuity of $s_i \mapsto s_i r_i(s_i)$, there exists a $k_1^i > 0$ such that $s_i r_i(s_i) < \epsilon s_i^2 + k_1^i$ for all s_i. Therefore,

$$(Q\gamma)^{\mathrm{T}} R(\gamma) \leq \epsilon \|s\|^2 + \sum_{i=1}^{n} k_1^i \leq \epsilon \tilde{\lambda}_d \|\gamma\|^2 + \sum_{i=1}^{n} k_1^i,$$

where $\tilde{\lambda}_d$ is the maximal eigenvalue of $Q^{\mathrm{T}} Q$. Since $\epsilon > 0$ is arbitrary, (17) holds.

Re (II) The main part of the proof is to show that there exists a $\delta > 0$ such that

$$\liminf_{\|\gamma\| \to \infty} \inf_{\gamma' \in B(\gamma, \delta)} m(\gamma') \cdot m(\gamma) > 1/5, \tag{18}$$

where $B(\gamma, \delta) = \{\gamma' \in \mathbb{R}^d | \|\gamma' - \gamma\| < \delta\}$. Using this result we can construct a fixed size cone

$$W(\gamma) = \{\gamma + a\xi \mid \xi \in \mathbb{R}^d, \ \ \|\xi\| = 1, \ \ \|\xi - m(\gamma)\| < 1/10, \ \ 0 \leq a < \delta\},$$

which by arguments similar to the proof of Theorem 4.3 in Jarner and Hansen (2000) is contained in $A(\gamma)$ for $\|\gamma\|$ sufficiently large; see also the discussion on page 354 in Jarner and Hansen (2000). Equation (16) then holds since

$$\liminf_{\|\gamma\| \to \infty} \int_{A(\gamma)} q(\gamma'; \gamma, h) d\gamma' \geq \liminf_{\|\gamma\| \to \infty} \int_{W(\gamma)} q(\gamma'; \gamma, h) d\gamma' = \int_{W(0)} q(\gamma'; 0, h) d\gamma' > 0.$$

Now we verify that there exists a $\delta > 0$ such that (18) holds. Let the eigenvalues of $\Sigma_1 = QQ^{\mathrm{T}}$ be $0 < \lambda_1 \leq \ldots \leq \lambda_n$, fix $i \in \{1, \ldots, n\}$, and observe that

$$\log\left(\frac{|r_i(s_i)|}{|r_i(s_i')|}\right) = \left(\log(|y_i - \mu_i|) + \log\left(\frac{g_c'(\mu_i)}{g'(\mu_i)}\right)\right) - \left(\log(|y_i - \mu_i'|) + \log\left(\frac{g_c'(\mu_i')}{g'(\mu_i')}\right)\right), \tag{19}$$

where $g(\mu_i') = s_i' + d_i^{\mathrm{T}}\beta$. Since g_c and g are increasing, $r_i(s_i)$ and $r_i(s_i')$ must have the same sign when $|s_i|$ is sufficiently large and $|s_i - s_i'| < k_2$ for some $k_2 > 0$. So by differentiating $\log(|r_i(s_i)|)$ with respect to s_i for $|s_i|$ sufficiently large, and using the mean-value theorem on (19), it follows that

$$\limsup_{|s_i|\to\infty} \sup_{|s_i - s_i'| < k_2} \left|\log\left(\frac{r_i(s_i)}{r_i(s_i')}\right)\right| \leq k_2 \limsup_{\tilde{\mu}_i \to m_1, m_2} \frac{1}{g'(\tilde{\mu}_i)} \left|\frac{1}{\tilde{\mu}_i - y_i} + \frac{g_c''(\tilde{\mu}_i)}{g_c'(\tilde{\mu}_i)} - \frac{g''(\tilde{\mu}_i)}{g'(\tilde{\mu}_i)}\right|.$$

By (14) we can choose k_2 so small that the right hand side of the above inequality is less than $\log(1 + \sqrt{\lambda_1/\lambda_n})$, and obtain

$$\begin{aligned}(r_i(s_i) - r_i(s_i'))^2 &= \left(\sqrt{r_i(s_i)/r_i(s_i')} - \sqrt{r_i(s_i')/r_i(s_i)}\right)^2 r_i(s_i) r_i(s_i') \\ &< \left(\sqrt{1 + \sqrt{\lambda_1/\lambda_n}} - \sqrt{1/(1 + \sqrt{\lambda_1/\lambda_n})}\right)^2 r_i(s_i) r_i(s_i') < (\lambda_1/\lambda_n)\left(r_i(s_i)^2 + r_i(s_i')^2\right)/2\end{aligned}$$

when $|s_i|$ is sufficiently large and $|s_i - s_i'| < k_2$.

Therefore, by continuity of $r_i(s_i)$, there exists a $k_3^i > 0$ such that $(r_i(s_i) - r_i(s_i'))^2 < (\lambda_1/\lambda_n)(r_i(s_i)^2 + r_i(s_i')^2)/2 + k_3^i$ when $|s_i - s_i'| < k_2$ for all s_i. Using this and the following two inequalities: $(R(\gamma) - R(\gamma'))^{\mathrm{T}}\Sigma_1(R(\gamma) - R(\gamma')) \leq \lambda_n \|R(\gamma) - R(\gamma')\|^2$ and $R(\gamma)^{\mathrm{T}}\Sigma_1 R(\gamma) \geq \lambda_1 \|R(\gamma)\|^2$, we see that

$$(R(\gamma) - R(\gamma'))^{\mathrm{T}}\Sigma_1(R(\gamma) - R(\gamma')) \leq (R(\gamma)^{\mathrm{T}}\Sigma_1 R(\gamma) + R(\gamma')^{\mathrm{T}}\Sigma_1 R(\gamma'))/2 + 2k_0, \tag{20}$$

when $\gamma' \in B(\gamma, \delta)$, where $k_0 = \sum_{i=1}^n k_3^i/2$ and $\delta = k_2/\sqrt{\tilde{\lambda}_d}$.

Let $\gamma' \in B(\gamma, \delta)$. By (20) it follows that

$$\begin{aligned}R(\gamma)^{\mathrm{T}}\Sigma_1 R(\gamma') &= (R(\gamma)^{\mathrm{T}}\Sigma_1 R(\gamma) + R(\gamma')^{\mathrm{T}}\Sigma_1 R(\gamma') - (R(\gamma) - R(\gamma))^{\mathrm{T}}\Sigma_1(R(\gamma) - R(\gamma')))/2 \\ &\geq (R(\gamma)^{\mathrm{T}}\Sigma_1 R(\gamma) + R(\gamma')^{\mathrm{T}}\Sigma_1 R(\gamma'))/4 - k_0.\end{aligned}$$

Using this and the inequality $\sqrt{v} \leq v/(8\delta) + 2\delta$, we obtain

$$\begin{aligned}\nabla(\gamma) \cdot \nabla(\gamma') &= (\|\gamma\|^2 + \|\gamma'\|^2 - \|\gamma - \gamma'\|^2)/2 + R(\gamma)^{\mathrm{T}}\Sigma_1 R(\gamma') - (Q\gamma)^{\mathrm{T}} R(\gamma') - (Q\gamma')^{\mathrm{T}} R(\gamma) \\ &\geq \|\gamma\|^2/2 + \|\gamma'\|^2/2 - \delta^2/2 + (R(\gamma)^{\mathrm{T}}\Sigma_1 R(\gamma) + R(\gamma')^{\mathrm{T}}\Sigma_1 R(\gamma'))/4 - k_0 \\ &\quad - (Q\gamma')^{\mathrm{T}} R(\gamma') - \delta\sqrt{R(\gamma')^{\mathrm{T}}\Sigma_1 R(\gamma')} - (Q\gamma)^{\mathrm{T}} R(\gamma) - \delta\sqrt{R(\gamma)^{\mathrm{T}}\Sigma_1 R(\gamma)} \\ &\geq \|\nabla(\gamma)\|^2/8 + (3/8)(\|\gamma\|^2 - 2(Q\gamma)^{\mathrm{T}} R(\gamma)) \\ &\quad + \|\nabla(\gamma')\|^2/8 + (3/8)(\|\gamma'\|^2 - 2(Q\gamma')^{\mathrm{T}} R(\gamma')) - k_0 - 9\delta^2/2.\end{aligned}$$

By (17), we see that $\|\gamma\|^2 - 2(Q\gamma)^{\mathrm{T}}R(\gamma) > 0$, when $\|\gamma\|$ is large. Therefore,

$$m(\gamma)\cdot m(\gamma') > \left(\frac{\|\nabla(\gamma)\|}{\|\nabla(\gamma')\|} + \left(\frac{\|\nabla(\gamma)\|}{\|\nabla(\gamma')\|}\right)^{-1}\right)/8 - \frac{k_0 - 9\delta^2/2}{\|\nabla(\gamma)\|\|\nabla(\gamma')\|} \geq \frac{2}{8} - \frac{k_0 - 9\delta^2/2}{\|\nabla(\gamma)\|\|\nabla(\gamma')\|}.$$

Consequently (18) follows, since $\|\nabla(\gamma)\| \to \infty$ when $\|\gamma\| \to \infty$.

□

Remark 1: When the canonical link $g = g_c$ is used, condition (15) is always satisfied and condition (14) simplifies to $\limsup_{\mu\to m_1,m_2} g_c'(\mu)|\mu - y_i| > 0$ for $i = 1,\dots,n$. The latter condition is easily verified for models (*i*) and (*ii*). For model (*iii*), condition (14) follows by using that $g_c''(\mu)/g_c'(\mu) = g''(\mu)/g'(\mu)$, and condition (15) is easily verified.

Remark 2: If $f(\cdot|\gamma)$ is super-exponential, then for any $t > 0$ there exists a $c_t > 0$ such that $f(\gamma|y)^{-1/2} \geq c_t \exp(t\|\gamma\|)$, $\gamma \in \mathbb{R}^d$. Under the conditions of Theorem 1, the CLT (11) therefore holds for functions ψ satisfying $\psi(\gamma)^2 \leq \exp(t\|\gamma\|)$ for some $t > 0$.

3.2 Geometric ergodicity of Langevin-Hastings

General conditions assuring geometric ergodicity of the Langevin-Hastings algorithm are given in Roberts and Tweedie (1996a). However, as shown in Roberts and Tweedie (1996a) and in the following proposition, the Langevin-Hastings algorithm is not always geometrically ergodic.

Proposition 1. *Suppose that the covariance matrix QQ^{T} of $(S_1,\dots,S_n)$ is invertible. The Langevin-Hastings algorithm is not geometrically ergodic for any $h \in]0;\infty[$ in the case of the model (i) or the model (iii).*

Proof. See Appendix B. □

The reason why the Langevin-Hastings algorithm is not geometrically ergodic for the models (*i*) and (*iii*) is that components of the gradient $\nabla(\gamma)$ may increase at an exponential rate as a function of γ. For certain values of γ the algorithm therefore proposes extremely large jumps which are rejected. In the sequel we discuss a modification of the Langevin-Hastings algorithm where $\nabla(\gamma)$ is replaced by

$$\nabla(\gamma)^{\mathrm{trunc}} = -\gamma + Q^{\mathrm{T}}R(\gamma) \tag{21}$$

and $R(\gamma)$ is a bounded function. We refer to this modified algorithm as the truncated Langevin-Hastings algorithm. In general, we choose the function R so that $\nabla(\gamma)^{\mathrm{trunc}} = \nabla(\gamma)$ for most values of γ in the 'center' of the target distribution $f(\cdot \mid y)$. For instance, if the canonical link function is used, then (8) simplifies to

$$\nabla(\gamma) = -\gamma + Q^{\mathrm{T}}\left\{y_i - \mu_i\right\}_{i=1}^{n}, \tag{22}$$

so for model (i), for example, we can take $R(\gamma) = \{y_i - (\mu_i \wedge H)\}_{i=1}^n$, where $0 < H < \infty$ is a truncation constant (for the model (ii), $\mathcal{M} =]0; N[$ is bounded, so here truncation is actually not needed). For the model (iii) we can let $R(\gamma) = \{y_i/(\mu_i \vee H) - 1\}_{i=1}^n$. In Theorem 2 below we establish geometric ergodicity for the truncated Langevin-Hastings algorithm.

Theorem 2. *Assume that $0 < h < 2$ and $R(\gamma)$ is bounded. Then the truncated Langevin-Hastings algorithm is geometrically ergodic, with V in (10) equal to $V(\gamma) = \exp(t\|\gamma\|)$ for an arbitrary $t > 0$.*

Proof. Below we give only a sketch of the proof as it follows the same line as the proof of Theorem 4 in Møller *et al.* (1998).

Let

$$q(\gamma, \gamma') = (2\pi h)^{-d/2} \exp(-\|\gamma' - \xi(\gamma)\|^2/(2h))$$

denote the proposal density. For a given $\epsilon > 0$, set $B_\epsilon(\gamma) = \{\gamma' : \|\gamma' - \xi(\gamma)\| < S_\epsilon\}$, where $S_\epsilon > 0$ is chosen so that $\int_{\mathbb{R}^d \setminus B_\epsilon(\gamma)} q(\gamma, \gamma') d\gamma' < \epsilon$.

Consider a proposal $\gamma' \in B_\epsilon(\gamma)$. Since $\gamma' - \xi(\gamma)$ is bounded and $\xi(\gamma) = (1-h/2)\gamma + O(1)$, we have that

$$\|\gamma'\|/\|\gamma\| \to (1 - h/2) < 1 \text{ as } \|\gamma\| \to \infty. \tag{23}$$

Hence $B_\epsilon(\gamma) \subseteq \{\gamma' : \|\gamma'\| < \|\gamma\|\}$ as $\|\gamma\| \to \infty$.

As in Møller *et al.* (1998) we now obtain geometric ergodicity from Theorem 4.1 in Roberts and Tweedie (1996a) by showing that a proposal $\gamma' \in B_\epsilon(\gamma)$ is always accepted when $\|\gamma\| \to \infty$. Let

$$\begin{aligned}
J_1 &= \left(\|\gamma\|^2 - \|\gamma'\|^2\right) h/8, \\
J_2 &= \left(\|Q^{\mathrm{T}} R(\gamma)\|^2 - \|Q^{\mathrm{T}} R(\gamma')\|^2\right) h/8, \\
J_3 &= \sum_{i=1}^n \left(\log f(y_i; \mu_i') - \log f(y_i; \mu_i)\right), \\
J_4 &= \left((Q\gamma)^{\mathrm{T}} R(\gamma) - (Q\gamma')^{\mathrm{T}} R(\gamma')\right)(1 - h/2)/2 + \left((Q\gamma)^{\mathrm{T}} R(\gamma') - (Q\gamma')^{\mathrm{T}} R(\gamma)\right)/2,
\end{aligned}$$

where $\mu_i = \mu_i(Q\gamma)$ and $\mu_i' = \mu_i(Q\gamma')$ are given by (6). The proposal γ' is accepted if $J_1 + J_2 + J_3 + J_4 \geq 0$. In the sequel we verify this condition in the case where $\|\gamma\| \to \infty$ and $\gamma' \in B_\epsilon(\gamma)$.

Combining (23) with the boundedness of $R(\gamma')$, the terms J_2 and J_4 are seen to be $o(J_1)$, where $J_1 \to \infty$. Now let $i \in \{1, \ldots, n\}$ be fixed, and consider the i'th term in J_3. Note that if $\|\gamma\| \to \infty$ in such a way that $|(Q\gamma)_i| \to \infty$ then $(Q\gamma')_i = (1-h/2)(Q\gamma)_i + O(1)$. Therefore, there exist functions l_i and u_i such that either

(I) $\mu_i \to m_1$, $\mu_i' \in [l_i(\mu_i); u_i(\mu_i)]$, $u_i(\mu_i) \to m_1$, and $\mu_i < l_i(\mu_i) < u_i(\mu_i)$ when μ_i is sufficiently close to m_1,

(II) $\mu_i \to m_2$, $\mu_i' \in [l_i(\mu_i); u_i(\mu_i)]$, $l_i(\mu_i) \to m_2$, and $l_i(\mu_i) < u_i(\mu_i) < \mu_i$ when μ_i is sufficiently close to m_2, or

(III) both μ_i and μ_i' stays inside a compact subset of $\mathcal{M}$.

Using (24) in case (I), (25) in case (II) (see the Appendix), and the continuity of $f(y_i; \cdot)$ in case (III), we see that $\log f(y_i; \mu_i') - \log f(y_i; \mu_i)$ is bounded below. Hence J_3 is bounded below. Therefore $J_1 + J_2 + J_3 + J_4 \geq 0$ when $\|\gamma\|$ is sufficiently large and $\gamma' \in B_\epsilon(\gamma)$. □

4 Empirical study of algorithms

In this section we study the performance of the random walk Metropolis and the truncated Langevin-Hastings algorithms applied to conditional simulation in the Poisson-log normal model (i) given weed count data y observed at $n = 250$ locations placed in a 20×14 grid; see Figure 1. We assume that the covariance matrix Σ is specified by the exponential correlation function so that $\mathrm{E}[S_i S_j] = \sigma^2 \exp(-d_{ij}/\alpha)$, where d_{ij} is the distance between the locations i and j. In Christensen *et al.* (2000) a full Bayesian analysis of the data set is performed. Below we fix the regression parameters β and the covariance parameters α and σ^2 at values equal to the posterior means computed in Christensen *et al.* (2000).

Roberts *et al.* (1997), Roberts and Rosenthal (1998b), and Breyer and Roberts (2000) show for certain classes of target densities that the proposal variance h should be tuned to obtain acceptance rates close to 0.23 for the random walk Metropolis algorithm and 0.57 for the Langevin-Hastings algorithm. Strictly speaking these results do not cover truncated Langevin-Hastings and the type of target densities considered here, but we find the rates 0.23 and 0.57 useful as guidelines. The truncation constant in Section 3.2 is chosen to be $H = 50$, which is roughly two times the maximal observed count.

The matrix K in the decomposition $\Sigma = KK^{\mathrm{T}}$ is calculated using either Cholesky factorisation or the circulant embedding technique (Wood and Chan, 1994; Dietrich and Newsam, 1993) based on the two-dimensional fast Fourier transform (FFT). In the latter case $(S_1, \ldots, S_{250})$ is embedded in a circulant stationary field $(S_1, \ldots, S_{250}, S_{251}, \ldots, S_{2048})$ defined on a 64×32 extended grid containing the original 20×14 grid.

4.1 Comparison of asymptotic variances for truncated Langevin-Hastings and random walk Metropolis

By Theorem 1 and Theorem 2 the algorithms are geometrically ergodic. For a MCMC estimate $\bar{S}_i$ of the conditional expectation $\mathrm{E}[S_i|y]$, we thus have a central limit theorem: $\sqrt{m}(\bar{S}_i - \mathrm{E}[S_i|y)]) \overset{\sim}{\to} N(0, \sigma_i^2)$ (see (11)), where m is the size of the MCMC sample used to calculate $\bar{S}_i$, and σ_i^2 is the asymptotic variance. We use the initial monotone sequence estimate in Geyer (1992) to estimate σ_i^2 at four representative locations; see Table 1.

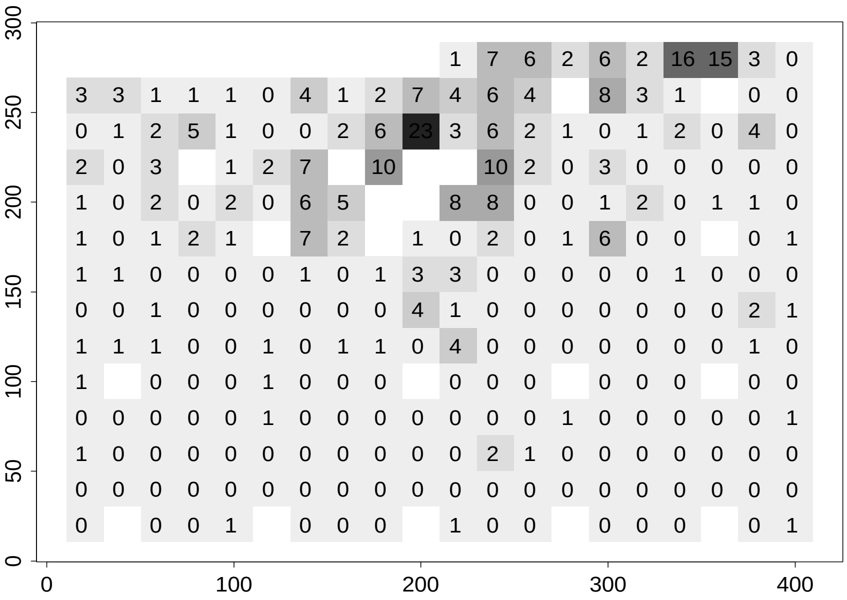

Figure 1: Counts of weed plants.

Algorithm	Accept.	Asymp. var.				CPU/
	rates	(200,240)	(20,280)	(200,140)	(320,60)	1000 itr.
Chol. LH	0.57	0.05	7.20	0.70	6.05	3.0
Chol. RW	0.23	0.46	188.32	15.70	131.55	1.8
FFT LH	0.56	1.41	6.92	5.37	7.23	9.1
FFT RW	0.23	126.52	400.88	309.68	510.89	5.1

Table 1: Results for Langevin-Hastings (LH) and random walk Metropolis (RW) algorithms using Cholesky (Chol.) or FFT implementations. For each case, acceptance rates and asymptotic variances for MCMC estimates are based on each 10th of 500000 iterations, while CPU times are for generation of 1000 iterations. The coordinates of the four locations refer to Figure 1.

If we consider e.g. the location $(320, 60)$ and the Cholesky implementation, then the estimated asymptotic variance is 22 times larger when using random walk Metropolis instead of Langevin-Hastings. For a given required precision of the Monte Carlo estimate one thus need a 22 times larger sample size with the random walk Metropolis algorithm than with the Langevin-Hastings algorithm. With the reported computing times on a 400 Mhz workstation, one would need to run the random walk Metropolis algorithm 13 times longer than the Langevin-Hastings algorithm. Considering the FFT implementation, the asymptotic variance for location $(320, 60)$ is 71 times larger for random walk Metropolis than for Langevin-Hastings.

With the FFT implementation we simulate a random field of dimension $q = 2048$ whereas the dimension for the Cholesky implementation is only $q = 250$. The improvement, when comparing Langevin-Hastings with random walk Metropolis, being largest for the FFT implementation is thus in accordance with the theoretical results in Roberts and Rosenthal (1998b), c.f. the end of Section 2.2.2.

4.2 Effect of truncation

The untruncated Langevin-Hastings algorithm is not geometrically ergodic (see Proposition 1). Therefore, the performance of the algorithm may depend much on the choice of initial value. Considering the Cholesky implementation, a sensible initial value $\gamma^{(1)}$ is the solution to $s^{(1)} = Q\gamma$, where

$$s_i^{(1)} = \log(y_i + 0.01) - d_i^{\mathrm{T}}\beta,\ i = 1, \ldots, 250,$$

in which case $f(y|s^{(1)})$ approximates the maximum of $f(y|\cdot)$. Another obvious initial value $\gamma^{(2)}$ is the unconditional mean $\gamma^{(2)} = 0$. With these initial values for γ both the untruncated and the truncated algorithm converges quickly to equilibrium (and the generated output is in fact identical for the two algorithms when the same seed is used for the random number generator). If we on the other hand choose a less sensible starting value and for example let $\gamma^{(3)}$ solve $s^{(3)} = Q\gamma$ where $s_i^{(3)} = 10$, $i = 1, \ldots, 250$, then the untruncated algorithm gets stuck in the initial value (no accept in the first 100000 iterations). The truncated algorithm however still converges quickly to equilibrium, see Figure 2.

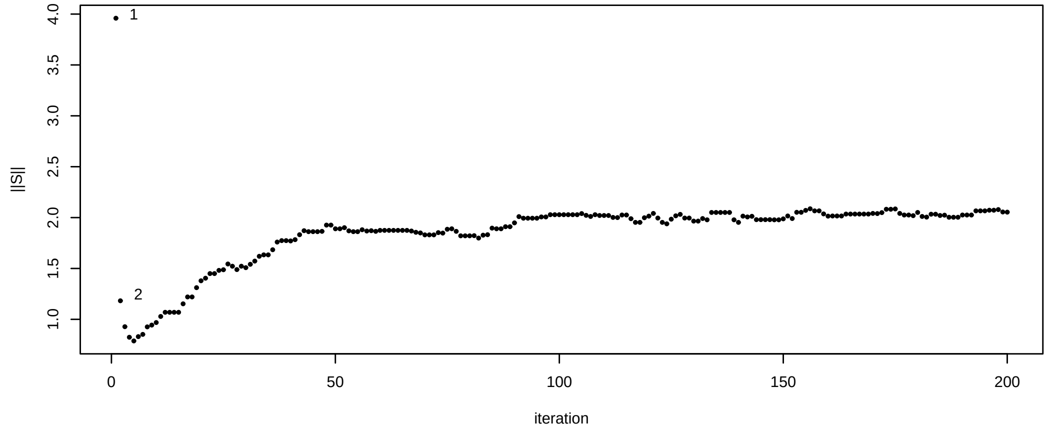

Figure 2: Timeseries of $\|S\|$ when the starting value $\gamma^{(3)}$ solves $s^{(3)} = Q\gamma$ where $s_i^{(3)} = 10$, $i = 1, \ldots, 250$. The initial state of the timeseries $\|s^{(3)}\| = 10\sqrt{250} \approx 158$ is omitted on the plot. The first two states after $s^{(3)}$ are marked with 1 and 2 in the plot.

5 Discussion

An alternative to the Langevin-Hastings algorithm, which is based on the Euler-discretisation, is to construct a proposal kernel based on a more refined discretisation of the Langevin diffusion as suggested in Stramer and Tweedie (1999) and further studied in the multidimensional case in Roberts and Stramer (2000). As in Section 2.2.2, let $\nabla(\gamma)$ denote the gradient of the log target density. Further, let $J(\gamma)$ be the second derivative of the log target density. The so-called local linearisation scheme (Ozaki, 1992; Shoji and Ozaki, 1998; Stramer and Tweedie, 1999) applied to the Langevin diffusion gives rise to a proposal kernel of the form $N(\mu_\gamma, K_\gamma)$ where

$$\mu_\gamma = \gamma + J(\gamma)^{-1}(\exp(hJ(\gamma)/2) - I)\nabla(\gamma)$$

and

$$K_\gamma = J(\gamma)^{-1}(\exp(hJ(\gamma)) - I).$$

The examples studied in Stramer and Tweedie (1999) and Roberts and Stramer (2000) show that much faster convergence to the equilibrium distribution may be obtained when using an algorithm based on local linearisation instead of the simple Euler-discretisation. However, we can verify that local linearisation applied to the Langevin diffusion for the Poisson-log normal model (*i*) does not yield a geometrically ergodic algorithm in the one-dimensional case. Moreover, in the multidimensional case, the conditions of Theorem 4.1 in Roberts and Stramer (2000) for geometric ergodicity are rather restrictive, and are not satisfied for either of the models (*i*) and (*iii*). The calculation of μ_γ and K_γ may finally be very time consuming when the dimension of γ is high.

A second alternative to the truncated Langevin-Hastings is to use a mixture of random walk Metropolis and Langevin-Hastings. A practical problem is that the proposal variance needs to be tuned for both the random walk Metropolis and the Langevin-Hastings updates. We prefer truncated Langevin-Hastings due to its fine convergence properties.

The results in this paper are restricted to situations where the regression parameter β and the covariance matrix Σ are fixed as e.g. when MCMC is used for likelihood computation, see Geyer and Thompson (1992) and McCulloch (1997). At present we do not know how to handle geometric ergodicity in the much more complicated situation where priors are imposed on β and Σ. However, in the case where a Gaussian prior is used for β, and Σ is fixed, our results on geometric ergodicity are still valid for random walk or truncated Langevin-Hastings algorithms with simultaneous updating of β and γ.

Acknowledgement: This research was supported by the European Union's network "Statistical and Computational Methods for the Analysis of Spatial Data. ERB-FMRX-CT96-0096", by MaPhySto – Centre for Mathematical Physics and Stochastics, funded by a grant from The Danish National Research Foundation – and by the Danish Informatics Network in the Agricultural Sciences, funded by a grant from the Danish Research Councils. We are also grateful to Gareth Roberts and Osnat Stramer for showing us their preprint and to Søren Jarner and Gareth Roberts for helpful discussions concerning geometric ergodicity of the algorithms. We finally thank the referees for valuable comments.

Appendix A

In the proof of Theorem 2 we need the following lemma, where we recall that $\mathcal{M} =]m_1; m_2[$ is an open interval.

Lemma 1. *Consider any $z_0 \in \Omega$, and any functions $l, u :]m_1; m_2[\mapsto]m_1; m_2[$ such that as $\mu \to m_1$, $\mu < l(\mu) < u(\mu)$ and $u(\mu) \to m_1$, while as $\mu \to m_2$, $l(\mu) < u(\mu) < \mu$ and $l(\mu) \to m_2$. Then*

$$\liminf_{\mu \to m_1} \inf_{\mu' \in [l(\mu); u(\mu)]} \{\log f(z_0; \mu') - \log f(z_0; \mu)\} > -\infty \tag{24}$$

and

$$\liminf_{\mu \to m_2} \inf_{\mu' \in [l(\mu); u(\mu)]} \{\log f(z_0; \mu') - \log f(z_0; \mu)\} > -\infty. \tag{25}$$

Proof. Since g_c is strictly monotone and continuous, $g_c(\mathcal{M}) =]\theta_1; \theta_2[$ is an open interval and we can reformulate (24) and (25) to conditions concerning the asymptotic behaviour of $e(\theta') - e(\theta)$, where $e(\theta) = z_0\theta - a(\theta)$. Observe that $e(\cdot)$ is continuous and $g_c([l(\mu); u(\mu)])$ is a compact interval. Thereby we obtain that (24) and (25) hold if

$$\liminf_{\theta \to \theta_1} \{e(t(\theta)) - e(\theta)\} > -\infty \tag{26}$$

and

$$\liminf_{\theta \to \theta_2} \{e(t(\theta)) - e(\theta)\} > -\infty \tag{27}$$

hold for any function $t :]\theta_1; \theta_2[\mapsto]\theta_1; \theta_2[$ such that as $\theta \to \theta_1$, $t(\theta) > \theta$ and $t(\theta) \to \theta_1$, while as $\theta \to \theta_2$, $t(\theta) < \theta$ and $t(\theta) \to \theta_2$.

Assume that (26) is false. Then there exists a strictly decreasing sequence $\{\theta_k\}$ so that $\theta_k \to \theta_1$ and $e(t(\theta_k)) - e(\theta_k) \to -\infty$ as $k \to \infty$. Note that by thinning of the sequence $\{\theta_k\}$, we can assume that $\{t(\theta_k)\}$ is strictly decreasing.

On one hand, for sufficiently large k,

$$\partial e(t(\theta_k))/\partial\theta \leq (e(t(\theta_k)) - e(\theta_k))/(t(\theta_k) - \theta_k) < 0,$$

where the first inequality follows by concavity of $e(\cdot)$.

On the other hand, let ν denote the counting or Lebesgue reference measure used in (2), let $\theta_0 \in]\theta_1;\theta_2[$ and observe that

$$\begin{aligned}\limsup_{\theta\to\theta_1} e(\theta) &= \limsup_{\theta\to\theta_1} \left\{ -\log\left(\int_\Omega \exp(\theta(z-z_0)+b(z))\nu(dz) \right) \right\} \\ &\le \limsup_{\theta\to\theta_1} \left\{ -\log\left(\int_{]-\infty;z_0]\cap\Omega} \exp(\theta(z-z_0)+b(z))\nu(dz) \right) \right\} \\ &\le -\log\left(\int_{]-\infty;z_0]\cap\Omega} \exp(\theta_0(z-z_0)+b(z))\nu(dz) \right) < \infty.\end{aligned}$$

Hence, $\{e(\theta_k)\}$ is bounded above. Consequently, $e(t(\theta_k)) = (e(t(\theta_k))-e(\theta_k))+e(\theta_k) \to -\infty$ as $k\to\infty$. So by thinning further the sequence $\{\theta_k\}$, we can assume that both sequences $\{e(t(\theta_k))\}$ and $\{t(\theta_k)\}$ are strictly decreasing. Using the concavity of $e(\cdot)$ we now obtain that

$$\partial e(t(\theta_k))/\partial\theta \ge (e(t(\theta_{k-1})) - e(t(\theta_k)))/(t(\theta_{k-1}) - t(\theta_k)) > 0.$$

This gives a contradiction. Hence (26) is true. By similar arguments we see that (27) holds. □

Appendix B

Theorem 4.2 in Roberts and Tweedie (1996a) gives a general condition which implies that the Langevin-Hastings algorithm is not geometrically ergodic. However, this condition does not cover the situation considered in Proposition 1.

Proof of Proposition 1. It follows from Proposition 5.1 in Roberts and Tweedie (1996b) that the algorithm is not geometrically ergodic, if there exists a sequence of Borel sets $\{M_k\}$ in $\mathbb{R}^d$ with Lebesgue measure $\nu(M_k) > 0$ so that

$$\limsup_{k\to\infty} \sup_{\gamma\in M_k} \mathrm{Acc}(\gamma) = 0, \tag{28}$$

where

$$\mathrm{Acc}(\gamma) = \int_{\mathbb{R}^d} \left(1 \wedge \frac{q(\gamma',\gamma)f(\gamma'|y)}{q(\gamma,\gamma')f(\gamma|y)} \right) q(\gamma,\gamma')d\gamma'.$$

For a given $\epsilon > 0$, letting $B_\epsilon(\gamma)$ be defined as in the proof of Theorem 2, then $q(\gamma',\gamma)/q(\gamma,\gamma') \le \exp(S_\epsilon^2/(2h))$ if $\gamma' \in B_\epsilon(\gamma)$, since $q(\gamma,\gamma') \ge (2\pi h)^{-d/2}\exp(-S_\epsilon^2/(2h))$

and $q(\gamma', \gamma) \leq (2\pi h)^{-d/2}$. Hence

$$\begin{aligned}
\mathrm{Acc}(\gamma) &< \epsilon + \int_{B_\epsilon(\gamma)} \left(1 \wedge \frac{q(\gamma', \gamma) f(\gamma'|y)}{q(\gamma, \gamma') f(\gamma|y)}\right) q(\gamma, \gamma') d\gamma' \\
&\leq \epsilon + \sup_{\gamma' \in B_\epsilon(\gamma)} \{q(\gamma', \gamma) f(\gamma'|y)/(q(\gamma, \gamma') f(\gamma|y))\} \qquad (29) \\
&\leq \epsilon + \exp\left(\sup_{\gamma' \in B_\epsilon(\gamma)} \{\log f(\gamma'|y) - \log f(\gamma|y)\}\right) \exp(S_\epsilon^2/(2h)).
\end{aligned}$$

Inserting (5) in (29), it follows easily that (28) holds if there exists $\{M_k\}$ with $\nu(M_k) > 0$ so that for all $\epsilon > 0$,

$$\limsup_{k\to\infty} \sup_{\gamma \in M_k} \sup_{\gamma' \in B_\epsilon(\gamma)} \left\{-(\|\gamma'\|^2 - \|\gamma\|^2)/2 + J_3(\gamma, \gamma')\right\} = -\infty, \qquad (30)$$

where $J_3(\gamma, \gamma')$ is defined as J_3 in the proof of Theorem 2.

We consider first the Poisson case where $R(\gamma) = \left\{y_i - \exp(s_i + d_i^{\mathrm{T}}\beta)\right\}_{i=1}^n$.

Define

$$M_k = \{\gamma | \|\gamma\| < k\delta_0, \quad s = Q\gamma, \quad (ki - 1/2) < s_i < ki, \quad i = 1, \ldots, n\},$$

where δ_0 is a constant which is determined below, so that M_k becomes a nonempty open subset of $\mathbb{R}^d$ — this ensures that $\nu(M_k) > 0$. The openness follows by noticing that M_k is of the form $M_k = \{\gamma | Q\gamma \in A_k\} \cap C_k$, where A_k is an open box, C_k is an open ball, and the function $\gamma \mapsto Q\gamma$ is obviously continuous. The nonemptyness is verified by constructing a $\tilde{\gamma} \in M_k$ as follows. Since Q has full row rank we can without loss of generality assume that $Q = [\tilde{Q} \quad \bar{Q}]$, where $\tilde{Q}$ is an invertible $n \times n$ matrix. For a given $\tilde{s} \in A_k$, define $\tilde{\gamma} = ((\tilde{Q}^{-1}\tilde{s})^{\mathrm{T}}, 0, \ldots, 0)^{\mathrm{T}}$ and observe that

$$\|\tilde{\gamma}\| = \|\tilde{Q}^{-1}\tilde{s}\| \leq \|\tilde{s}\|/\sqrt{\tilde{\lambda}_0} \leq n^{3/2}k/\sqrt{\tilde{\lambda}_0}\,,$$

where $\tilde{\lambda}_0 > 0$ is the smallest eigenvalue in $\tilde{Q}\tilde{Q}^{\mathrm{T}}$. Letting $\delta_0 > n^{3/2}/\sqrt{\tilde{\lambda}_0}$ in the definition of M_k, we see that $\tilde{\gamma} \in M_k$.

We now study the asymptotic behaviour of the term $J_3(\gamma, \gamma')$ in (30), which consist of terms $\log f(y_i; \mu_i') - \log f(y_i; \mu_i)$, where $\mu_i = \exp(s_i + d_i^{\mathrm{T}}\beta)$ and $\mu_i' = \exp(s_i' + d_i^{\mathrm{T}}\beta)$ with $s' = Q\gamma'$. Let $\gamma \in M_k$, $\gamma' \in B_\epsilon(\gamma)$ and $k \to \infty$. Then $\mu_i \to \infty$. Since $\xi(\gamma) = (1 - h/2)\gamma + Q^{\mathrm{T}}R(\gamma)h/2$,

$$\gamma' = (1 - h/2)\gamma + Q^{\mathrm{T}}R(\gamma)h/2 + \phi(\gamma', \gamma), \qquad (31)$$

where $\|\phi(\gamma', \gamma)\| \leq S_\epsilon$. Using this we see that

$$s' = (1 - h/2)s + QQ^{\mathrm{T}}R(\gamma)h/2 + Q\phi(\gamma', \gamma),$$

and consequently, for each $i = 1, \dots, n$,

$$\mu'_i = \tilde{\phi}_i(\gamma', \gamma)(\mu_i)^{1-h/2} \exp((QQ^{\mathrm{T}} R(\gamma))_i h/2),$$

where $\tilde{\phi}_i(\gamma', \gamma)$ is bounded. By definition of M_k, $(QQ^{\mathrm{T}} R(\gamma))_i$ behaves asymptotically as a real constant $C_i \neq 0$ times $\exp(s_{j_i})$, where $j_i = \max\{j | (QQ^{\mathrm{T}})_{ij} \neq 0\}$. Hence for each $i = 1, \dots, n$, either (I) or (II) hold (according to whether C_i is positive or negative), where

(I) $\mu_i \to \infty$ and $\mu'_i > \mu_i$, when k is sufficiently large,

(II) $\mu_i \to \infty$ and $\mu'_i \to 0$.

In case (I), $\log f(y_i; \mu'_i) - \log f(y_i; \mu_i)$ is bounded above, which follows from (25) with $\mu = \mu'_i$ and $\mu' = \mu_i$. In case (II) we observe from (4) that

$$\log f(y_i; \mu'_i) - \log f(y_i; \mu_i) = y_i \log \mu'_i - \mu'_i - y_i \log \mu_i + \mu_i \leq \mu_i \leq \exp(d_i^{\mathrm{T}} \beta) \exp(ki/n),$$

where the inequalities hold for k being sufficiently large. Thereby

$$\limsup_{k\to\infty} \sup_{\gamma \in M_k} \sup_{\gamma' \in B_\epsilon(\gamma)} J_3(\gamma, \gamma') / \exp(kn) < \infty. \tag{32}$$

We next pay attention to the other term $\|\gamma'\|^2 - \|\gamma\|^2$ in (30). Let $\gamma \in M_k$ and $\gamma' \in B_\epsilon(\gamma)$ be given. Combining (31) with the inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ and the positive definiteness of QQ^{T},

$$\|\gamma'\|^2 - \|\gamma\|^2 \geq \lambda_0 \|R(\gamma)\|^2 h^2/12 - (1 + (1-h/2)^2)\|\gamma\|^2 - S_\epsilon^2,$$

where $\lambda_0 > 0$ is the smallest eigenvalue of QQ^{T}. Then by definitions of $R(\gamma)$ and M_k,

$$\begin{aligned}\|\gamma'\|^2 - \|\gamma\|^2 \geq & \sum_{i=1}^n \lambda_0 \left(\exp(2(d_i^{\mathrm{T}}\beta + s_i)) - 2y_i \exp(d_i^{\mathrm{T}}\beta + s_i) + y_i^2 \right) \frac{h^2}{12} \\ & - (1 + (1-h/2)^2) k^2 \delta_0^2 - S_\epsilon^2 \\ \geq & c_0 + c_1 \sum_{i=1}^n \exp(2ki) - c_2 \sum_{i=1}^n \exp(ki) - c_3 k^2,\end{aligned}$$

where $c_0 \in \mathbb{R}$, $c_1 > 0$, $c_2 > 0$ and $c_3 > 0$ are constants (more precisely, they do not depend on (γ, γ', k)). Hence

$$\limsup_{k\to\infty} \sup_{\gamma \in M_k} \sup_{\gamma' \in B_\epsilon(\gamma)} \{-(\|\gamma'\|^2 - \|\gamma\|^2)/2\} / \exp(2kn) < 0. \tag{33}$$

Combining (33) and (32) we see that (30) holds for all $\epsilon > 0$ and hence the proposition is verified in the Poisson case.

For the exponential error distribution the proof follows the same line as above, except that we now define

$$M_k = \{\gamma | \|\gamma\| < k\delta_0, \quad s = Q\gamma, \quad (-ki - 1/2) < s_i < -ki, \quad i = 1, \dots, n\}.$$

We therefore omit the details. □

References

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Statist. Assoc.* **88**, 9–25.

Breyer, L. A. and Roberts, G. O. (2000). From Metropolis to diffusions: Gibbs states and optimal scaling. *Stoch. Proc. Appl.* **90**, 181–206.

Christensen, O. F., Møller, J. and Waagepetersen, R. (2000). Analysis of spatial data using generalized linear mixed models and Langevin-type Markov chain Monte Carlo. *Research Report R-00-2009*, Department of Mathematical Sciences, Aalborg University.

Dietrich, C. R. and Newsam, G. N. (1993). A fast and exact method for multidimensional Gaussian stochastic simulation. *Water Resources Research* **29**, 2861–2869.

Diggle, P., Tawn, J. A. and Moyeed, R. A. (1998). Model-based geostatistics (with discussion). *Appl. Statist.* **47**, 299–350.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statist. Sci.* **7**, 473–511.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. R. Statist. Soc.* B **54**, 657–699.

Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stoch. Proc. Appl.* **85**, 341–361.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *J. R. Statist. Soc.* B **58**, 619–678.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall, London, 2nd edition.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Am. Statist. Assoc.* **92**, 162–170.

Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24**, 101–121.

Møller, J., Syversveen, A. R. and Waagepetersen, R. (1998). Log Gaussian Cox processes. *Scand. J. Statist.* **25**, 451–482.

Ozaki, T. (1992). A bridge between nonlinear time series models and nonlinear stochastics dynamical systems: a local linearization approach. *Stat. Sin.* **2**, 113–135.

Roberts, G. O. and Rosenthal, J. S. (1998a). Markov chain Monte Carlo: some practical implications of theoretical results. *Canad. J. Statist.* **26**, 5–31.

Roberts, G. O. and Rosenthal, J. S. (1998b). Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Statist. Soc.* B **60**, 255–268.

Roberts, G. O. and Stramer, O. (2000). Tempered Langevin diffusions and algorithms. In preparation.

Roberts, G. O. and Tweedie, R. L. (1996a). Exponential convergence of Langevin diffusions and their discrete approximations. *Bernoulli* **2**, 341–363.

Roberts, G. O. and Tweedie, R. L. (1996b). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83**, 95–110.

Roberts, G. O., Gelman, A. and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7**, 110–120.

Shoji, T. and Ozaki, T. (1998). A statistical method of estimation and simulation for systems of stochastic differential equations. *Biometrika* **85**, 240–243.

Stramer, O. and Tweedie, R. L. (1999). Langevin-type models II: self-targeting candidates for MCMC algorithms. *Methodol. Comput. Appl. Probab.* **1**, 307–328.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22**, 1701–1762.

Wood, A. T. A. and Chan, G. (1994). Simulation of stationary Gaussian processes in $[0, 1]^d$. *J. Comput. Graph. Statist.* **3**, 409–432.