UNIVERSITY OF TEXAS AT SAN ANTONIO

# Improved Robust MCMC Algorithm for Hierarchical Models

Liang Jiang

July 2010

# 1 ABSTRACT

In this paper, three important techniques are discussed with details: 1) group updating scheme; 2) Langevin algorithm; 3) data-corrected parameterization. They largely improve the performance of Hastings-within-Gibbs algorithm. And these improvements are illustrated by applying them on a hierarchical model with Rongelap data.

# 2 INTRODUCTION

Generally, Hastings-within-Gibbs algorithm suffers several problems when applied to hierarchical models,

1. poor convergence: the chains for second layer parameters don't converge even for a large number of iterations;
2. slow mixing and strong autocorrelation: the chains for latent variables and coefficients of predictors have strong autocorrelation;
3. significant dependence: cross-correlations among latent variables and between hyper-parameters are large;
4. heavy computational work: each latent variable $S_i$ is updated individually during which expensive matrix operations are needed.

Thus, the Hastings-within-Gibbs algorithm requires a lot of computational work, slow mixing and inaccurate. It is not surprising that it usually doesn't work properly, especially when the number of latent variables is large.

In this paper, the following three techniques are discussed with details to resolve the problems in the algorithmic perspective.

- Group updating scheme: improve mixing, reduce autocorrelation and computation by updating related components in blocks.
- Langevin algorithm: improve convergence by including the information of the gradient of the target distribution into the proposal distribution.

- Data-corrected Parameterization: improve convergence and reduce autocorrelation by allowing flexible transformation of components based on the data.

# 3 The Hierarchical Model

To validate the efficiency of the proposed techniques, one hierarchical model is used as an example, which is *generalized linear spatial model* (GLSM), first proposed by Diggle et al. (1998) [9]. The complete model specification is

$$\begin{aligned} Y_i|S(\boldsymbol{x}_i) &\sim p(y_i|\mu_i),\ \ i = 1,...,n; \\ \mu_i &= g^{-1}(S(\boldsymbol{x}_i)) \\ S(\boldsymbol{x}) &\sim MVN(D\boldsymbol{\beta}, \boldsymbol{\Sigma}) \end{aligned} \tag{3.1}$$

where

- response variables $Y_i$ are conditional independent and follow a specific distribution $p(\cdot)$ with mean $\mu_i$;
- as before, $S(\boldsymbol{x}_i)$ belongs to a stationary Gaussian process with mean structure $D\boldsymbol{\beta}$ and covariance structure $\boldsymbol{\Sigma}$;
- $g^{-1}(\cdot)$ is a specific link function;
- $D$ is a known covariate matrix usually related to locations while $\boldsymbol{\beta}$ is its coefficient vector ($D\boldsymbol{\beta}$ together determines the "spatial trend" in response variables, Diggle and Ribeiro 2010 [10] section 3.6);
- $\boldsymbol{\Sigma}$ is a variance-covariance matrix with entries $\sigma_{ij} = \sigma^2\rho(u_{ij})$: $\sigma^2$ is a unknown constant variance and $\rho(u_{ij})$ belongs on one of the common families of correlation function.

Note that this model is also known as *spatial generalized linear model* (SGLM), and it is included in generalized linear mixed models category since the Gaussian process $S(\boldsymbol{x}_i)$ can serve as *random effects*.

## The Poisson log-spatial model

As the name implies, this model has logarithm link function and the conditional distribution of each response variable $Y_i$ is Poisson. The complete

model specification is

$$\begin{aligned} Y_i|S(\boldsymbol{x}_i) &\sim Poisson(\cdot|\mu_i) \\ \log\mu_i &= S(\boldsymbol{x}_i) \\ S(\boldsymbol{x}) &\sim MVN(D\boldsymbol{\beta}, \boldsymbol{\Sigma}). \end{aligned} \tag{3.2}$$

where $Y_i$ are conditional independent given the latent variables $S_i$; $D$ is a known covariate matrix usually related to locations; while $\boldsymbol{\beta}$ is its coefficient vector and $D\boldsymbol{\beta}$ determines the "trend" in response variables (Diggle and Ribeiro 2010 [10] section 3.6).

This model is naturally a good candidate for count data. For Rongelap Data in which the response variables are photon emission counts $Y_i$ over time-periods $t_i$ at locations $\boldsymbol{x}_i$. The Poisson log-linear model can be easily adopted,

$$\log\mu_i = \log t_i + S(\boldsymbol{x}_i) \tag{3.3}$$

with powered exponential correlation function as shown in Diggle et al. (1998) [9].

# 4 Group Updating

In Gibbs sampler, random variables can be partitioned into groups (blocks). For example,

$$\boldsymbol{x} = (x_1, ..., x_p) \longrightarrow (\boldsymbol{y}_1, ..., \boldsymbol{y}_s) \tag{4.1}$$

where the $i$th groups, $\boldsymbol{y}_i$, contains $r_i \leq 1$ components and $\sum_{i=1}^{s} r_i = p$. Then the groups are updated by following the procedure of Gibbs sampler.

It is generally believed that *grouping (blocking)* of the components leads to faster convergence rate, as indicated in Amit and Grenander (1991) [1] "the larger the blocks that are updated simultaneously - the faster the convergence", because grouping "moves any high correlation ... from the Gibbs sampler over to the random vector generator" (Seewald 1992 [29]). Liu (1994) [18] and Liu et al. (1994) [19] revealed the benefit of grouping strategy (as well as *collapsing*) in the use of three-component Gibbs samplers. Roberts and Sahu (1997) [27] provided theoretical results on the role of grouping in the context of Gibbs Markov chains for multivariate normal target distributions. They proved that the grouped Gibbs sampler, DUGS

more specifically, has a faster convergence rate if all partial correlations of a Gaussian target density are non-negative. However, it is necessary to point out that group updating may demand more computational effort and even reduce the convergence rate in certain case as shown in Whittaker (1990) [31].

As a general rule, highly correlated components are candidates to be grouped, Gentle et al. (2004) [16]. In GLSM, random effects $S_i$ are natural choice because they are highly correlated and drawing samples from the posterior distribution $p(S|Y,\theta,\eta)$ is achievable via Metropolis-Hastings algorithm without too much extra computational effort. Actually, by group updating $S$ instead of "fix-scan" one by one, overall computational work is significantly reduced due to large number of latent variables in GLSM. The components of $\boldsymbol{\beta}$ can be updated in one group as well if there are more than one coefficients.

# 5 Langevin-Hastings algorithm

## 5.1 Weakness of random walk algorithm

Though random walk algorithm is the most commonly used Metropolis-Hastings algorithm due to its easy implementation for many diverse problems, it suffers slow convergence frequently because of two reasons.

First, its efficiency depends crucially on the scaling of the proposal density. If the variance of proposal distribution is too small, the Markov chain will converge slow because of the small moves of increments. And if the proposal variance is too large, the acceptance rate of the moves will be too small. For this issue, a few practical rules of thumb was proposed to provide guidelines for scaling the proposal, Besag and Green (1993) [2] and Besag et al. (1995) [4]. And Roberts et al. (1997) [26] proved that optimal performance is achieved under quite general conditions when "tune the proposal variance so that the average acceptance rate is roughly 1/4".

The second reason is that it conducts moves around the current point by following proposal distributions and completely ignores all the information in target distributions.

## 5.2 LANGEVIN ALGORITHM

In contrast to random walk algorithm, Langevin algorithm utilizes local information of target density and can be significantly more efficient, especially in high dimensional problems.

Derived from *diffusion theory* (Grenander and Miller 1994 [17] and Phillips and Smith 1996 [23]), the basic idea of this approach consists of two steps: first, seeking a *diffusion equation* (or a *stochastic differential equation*) which produces a *diffusion* (or *continuous-time process*) with stationary distribution $\pi$; and then discretizing the process for implementation of the method. More specifically, the *Langevin diffusion process* is defined by the stochastic differential equation

$$\mathrm{d}X_t = \mathrm{d}B_t + \frac{1}{2} \bigtriangledown \log \pi(X_t)\, \mathrm{d}t \tag{5.1}$$

where $B_t$ is the standard Brownian motion. This process leaves $\pi$ as its stationary distribution. Roberts and Rosenthal (1998) [28] also stressed that the Langevin diffusion in (5.1) is the only non-explosive diffusion which is reversible with respect to $\pi$.

To implement the diffusion algorithm, a discretization step is required where (5.1) is replaced by a random walk like transition

$$x^{(t+1)} = x^{(t)} + \frac{\sigma^2}{2} \bigtriangledown \log \pi(x^{(t)}) + \sigma \epsilon_t \tag{5.2}$$

where $\epsilon_t \sim N_p(0, I_p)$ and $\sigma^2$ corresponds to the step size of discretization. However, the Markov chain (5.2) could be very different from that of original diffusion process (5.1) and Roberts and Tweedie (1995) [25] showed that the chain (5.2) may be transient which makes $\pi$ no longer the stationary distribution.

To correct this negative behavior, Besag (1994) [3] suggested to apply M-H acceptance/rejection rule on moderating the discretization step, which means (5.2) is treated as a proposal in M-H algorithm. Thus the full Langevin Algorithm is described as below.

1. Given $X^{(t)}$, a random variable $X^*$ is proposed by

$$X^* = X^{(t)} + \frac{\sigma^2}{2} \bigtriangledown \log\{\pi(X^{(t)})\} + \sigma \epsilon_t \tag{5.3}$$

   where $\sigma$ is user-specified parameter.

2. Set $X^{(t+1)} = X^*$ with probability

$$\alpha = \min\{1, \frac{\pi(X^*)q(X^{(t)}, X^*)}{\pi(X^{(t)})q(X^*, X^{(t)})}\} \tag{5.4}$$

where

$$q(x, x^*) \propto \exp[-\frac{1}{2\sigma^2}\|x - x^* - \frac{\sigma^2}{2} \bigtriangledown \log\pi(x)\|^2]. \tag{5.5}$$

Otherwise, set $X^{(t+1)} = X^{(t)}$.

As a result, this algorithm includes the gradient information of the target density into the proposal density. Roberts et al. (1998) [28] showed that the optimal asymptotic scaling is achieved when the acceptance rate of this algorithm is tuned to around 0.574. Furthermore, they suggested the proposal variance should scale respect to the dimension as $p^{-1/3}$ and thus $O(p^{1/3})$ steps are required to converge comparing to $O(p)$ steps require by random walk algorithms for the same class of target densities. So the benefit of using Langevin algorithm increases as the dimension increases, which is desired for implementation in GLSM considering large number of latent variables are usually group updated.

To apply Langevin algorithm on updating the group of latent variables $S$ in GLSM, the gradient of target density usually can be obtained. In the case of difficult settings, numerical derivatives of exact gradient can be employed. In practice, Christensen et al. (2006) [8] suggested that choosing variance of discretization $\sigma^2 = \hat{l}^2/p^{1/3}$ with $\hat{l} = 1.65$ leads to optimal performance of the algorithms. Note that the Langevin algorithm is also desired for updating the group of coefficients $\boldsymbol{\beta}$.

# 6 Parameterization

## 6.1 CP v.s. NCP

It has been well recognized that convergence of MCMC methods, especially when using Gibbs sampler and related techniques, depends crucially on the choice of parameterization, Roberts and Sahu (1997) [27] and Papaspiliopoulos et al. (2007) [22].

Considering a hierarchical model in which $Y$ represents data, $X$ denotes the hidden layer and $\boldsymbol{\eta}$ denotes the unknown hyperparameters. The data $Y$

is independent of the parameters $\boldsymbol{\eta}$ conditional on $X$. This relationship can be revealed as follow

$$\eta \rightarrow X \rightarrow Y. \tag{6.1}$$

This known as *centered parameterization (CP)*, and the MCMC methods for generating samples from the posterior distribution $p(X, \boldsymbol{\eta}|Y)$ can be conducted in two steps,

1. sample $\boldsymbol{\eta}$ from $p(\boldsymbol{\eta}|X)$;

2. sample $X$ from $p(X|\boldsymbol{\eta}, Y)$.

From a modeling and interpretation perspective, CP is naturally used as a starting point. Plus, the independent property of the conditional posterior $p(\boldsymbol{\eta}|X, Y) = p(\boldsymbol{\eta}|X$ often leads to easy sampling of $\boldsymbol{\eta}$. And the analysis in Gelfand et al. (1995 and 1996) [12][13] showed that centered parameterization improved convergence for location parameters in a broad class of normal linear mixed models and generalized linear mixed models.

However, considering $X$ and $\boldsymbol{\eta}$ are generally strongly dependent a priori, the data $Y$ need to be strong informative about $X$ to diminish this dependence. Papaspiliopoulos et al. (2007) [22] also showed a situation that when the data are informative about $\boldsymbol{\eta}$ they still cannot diminish the prior dependence between $X$ and $\boldsymbol{\eta}$. Thus, there are many situations where the posterior dependence between $X$ and $\boldsymbol{\eta}$ is prohibitively strong that *non-centered parameterization (NCP)* is needed.

In NCP, a parameterization of an augmentation scheme $X$ is defined by any random pair $(\tilde{X}, \boldsymbol{\eta})$ together with a function $h$ such that

$$X = h(\tilde{X}, \boldsymbol{\eta}, Y), \tag{6.2}$$

and $\tilde{X}$ and $\boldsymbol{\eta}$ are *a priori independent*. The MCMC algorithm for generating from the posterior distribution $p(\tilde{X}, \boldsymbol{\eta}|Y)$ is then given by

1. sample $\boldsymbol{\eta}$ from $p(\boldsymbol{\eta}|\tilde{X}, Y)$;

2. sample $X$ from $p(\tilde{X}|\boldsymbol{\eta}, Y)$.

Another motivation behind the NCP is that the convergence properties of sampling from $p(\tilde{X}|\boldsymbol{\eta}, Y)$ could be better than from $p(X|\boldsymbol{\eta}, Y)$ in many cases, Papaspiliopoulos et al. (2003) [21].

As shown by the examples in Papaspiliopoulos et al. (2003 and 2007) [21][22], neither CP nor NCP are uniformly effective and they possess complementary strength that "when under the one parameterization, converges slowly; under the other it often converges much faster". Hence, the choice of parameterization is largely depending on how informative the particular realization of the data is for $X$. Also note that both CP and NCP are constructed based on the prior distributions of the model, and it would be more effective if parameterizing the posterior and take data into account. Two ways of data-based modifications were suggested in Papaspiliopoulos et al. [22].

## 6.2 Correcting the CP

Consider a linear parameterization

$$X = \sigma(\boldsymbol{\eta}, Y)\tilde{X} + \mu(\boldsymbol{\eta}, Y) \tag{6.3}$$

where $\mu(\boldsymbol{\eta}, Y) = E(X|\boldsymbol{\eta})$ and $\sigma^2(\boldsymbol{\eta}, Y) = Var(X|\boldsymbol{\eta})$. This can be seen as a first-order approximation of an NCP. When correcting this parameterization based on the data, it is natural to replace $\mu(\boldsymbol{\eta}, Y)$ with $\tilde{\mu}(\boldsymbol{\eta}, Y) = E(X|\boldsymbol{\eta}, Y)$ and $\sigma^2(\boldsymbol{\eta}, Y)$ with $\tilde{\sigma}^2(\boldsymbol{\eta}, Y) = Var(X|\boldsymbol{\eta}, Y)$. Then the new method allows the data to decide how much "centering" should be given in parameterization and as a result a NCP will be offered for "infinitely weak data" and a CP will be offered for "infinitely strong data". This parameterization method can be interpreted as a "data-corrected" *partially non-centered parameterization (PNCP)*, Papaspiliopoulos et al. (2003) [21]. PNCP is sometimes difficult to construct. When $\tilde{\mu}$ and $\tilde{\sigma}^2$ are not directly available, their approximation form can be used.

## 6.3 Correcting the NCP

To correct the NCP

$$X = h(\tilde{X}, \boldsymbol{\eta}) = h(\tilde{h}(X^*, \boldsymbol{\eta}, Y), \boldsymbol{\eta}), \tag{6.4}$$

it is natural to search for an approximate pivotal quantity $X^*$ and the function $\tilde{h}$ in (6.4) often relieves hard constraints on $X$ imposed by data. Several examples are given in Papaspiliopoulos et al. (2007) [22].

# 7 ROBUST MCMC ALGORITHM FOR GLSM

Let's consider the Poisson log-linear spatial model described in previous section. In the model specification (3.2), note that the relationship of the data $Y$, the latent variables $S$ and the parameters $\boldsymbol{\eta} = (\boldsymbol{\beta}, \phi, \sigma^2, \kappa)$ (in the case of matern correlation function) is naturally CP,

$$\eta \rightarrow S \rightarrow Y. \tag{7.1}$$

However, since Langevin algorithm is sensitive to inhomogeneity of the components with different variances, as mentioned by Roberts and Rosenthal (1998) [28], and considering the different characteristics of $(S, \boldsymbol{\beta}, \boldsymbol{\eta})$ (for sake of simplifying the notation, $\boldsymbol{\eta} = (\phi, \sigma^2, \kappa)$ will be used from now on), they should be updated in three blocks. Thus, *the goal is to find a parameterization of* $(S, \boldsymbol{\beta}, \boldsymbol{\eta})$ *so that after parameterization the posterior distributions of components in three blocks are approximately uncorrelated and in best case have equal variance and the dependence among three blocks are minimized.* The full parameterization should have the following form,

$$\begin{array}{rcl} S & \rightarrow & \tilde{S}(S; \boldsymbol{\beta}, \boldsymbol{\eta}, Y) \\ \boldsymbol{\beta} & \rightarrow & \tilde{\boldsymbol{\beta}}(\boldsymbol{\beta}; \boldsymbol{\eta}, Y) \\ \boldsymbol{\eta} & \rightarrow & \tilde{\boldsymbol{\eta}}(\boldsymbol{\eta}; Y) \end{array} \tag{7.2}$$

and the resulting algorithm will updata $\tilde{S}, \tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\eta}}$ respectively.

## 7.1 GAUSSIAN APPROXIMATION OF $p(S|y)$

The desired parameterization mentioned before usually are not easy to find except for multivariate Gaussian distribution. Thus, Christensen et al. (2006) [8] suggested to use a Gaussian approximation of the distribution $p(S|y)$ and then orthogonalize and standardize the approximated distribution. By differentiating $\log p(S|y)$ twice with respect to $S$, the covariance matrix of approximated Gaussian distribution is

$$\tilde{\Sigma} = (\Sigma^{-1} + \Lambda(\hat{S}))^{-1} \tag{7.3}$$

where $\Sigma$ is the covariance matrix of $S$ and $\Lambda(S)$ is a diagonal matrix with entries $-\frac{\partial^2}{\partial S_i^2} \log p(y_i|S_i)$, $i = 1, ..., n$, and $\hat{S}$ is a typical value of $S$ (the mode of

$p(S|y)$ or $p(y|S)$). For the Poisson log-linear spatial model, $\hat{S} = arg\,max\{p(y_i|S_i)\}$, $i = 1,...,n$, is suggested in Christensen's paper which leads to $\Lambda(\hat{S})_{ii} = y_i$. The reason for not using the mode of $p(S|y)$ is because of heavy computational work required by numerically finding the mode in GLSM. And using the current value of $S$ as $\hat{S}$ during updating would involve an intractable Jacobian matrix.

## 7.2 Parameterization of $S$ and $\boldsymbol{\beta}$

Let's initially assume a normal prior for $\boldsymbol{\beta}$, $p(\boldsymbol{\beta}) \sim N(\boldsymbol{\mu}, \Omega)$, and use a Tayler expansion around $\hat{S}$,

$$\log p(y|S) \approx -0.5(S-\hat{S})^T \Lambda(\hat{S})(S-\hat{S}) + c \tag{7.4}$$

where $c$ is a constant and note that the first order terms cancel with the choice of $\hat{S} = arg\,max\{p(y_i|S_i)\}$. Then the logarithm of conditional distribution of $(S, \boldsymbol{\beta})$ will be

$$\begin{aligned}
\log p(S, \boldsymbol{\beta}|\boldsymbol{\eta}, y) &\approx \log p(y|S) + \log p(S|\boldsymbol{\beta}, \boldsymbol{\eta}) + \log p(\boldsymbol{\beta}) \\
&\approx -0.5(S-\hat{S})^T \Lambda(\hat{S})(S-\hat{S}) - 0.5(S-D\boldsymbol{\beta})^T \Sigma^{-1}(S-D\boldsymbol{\beta}) \\
&\quad - 0.5(\boldsymbol{\beta}-\boldsymbol{\mu})^T \Omega^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}) \\
&= -0.5(S-\tilde{\Sigma}(\Lambda(\hat{S})\hat{S} + \Sigma^{-1}D\boldsymbol{\beta}))^T \tilde{\Sigma}^{-1}(S-\tilde{\Sigma}(\Lambda(\hat{S})\hat{S} + \Sigma^{-1}D\boldsymbol{\beta})) \\
&\quad - 0.5(\boldsymbol{\beta} - \tilde{\Omega}(D^T\Sigma^{-1}\tilde{\Sigma}\Lambda(\hat{S})\hat{S} + \Omega^{-1}\boldsymbol{\mu}))^T \tilde{\Omega}^{-1}(\boldsymbol{\beta} \\
&\quad - \tilde{\Omega}(D^T\Sigma^{-1}\tilde{\Sigma}\Lambda(\hat{S})\hat{S} + \Omega^{-1}\boldsymbol{\mu}))
\end{aligned} \tag{7.5}$$

where $\tilde{\Sigma} = (\Sigma^{-1} + \Lambda(\hat{S}))^{-1}$ from (7.3) and

$$\begin{aligned}
\tilde{\Omega} &= (\Omega^{-1} + D^T(\Sigma^{-1} - \Sigma^{-1}\tilde{\Sigma}\Sigma^{-1})D)^{-1} \\
&= (\Omega^{-1} + D^T(\Lambda(\hat{S})\Sigma + I_n)^{-1}\Lambda(\hat{S})D)^{-1}.
\end{aligned} \tag{7.6}$$

From (7.5), we can see that the parameterization

$$\tilde{S} = (\tilde{\Sigma}^{1/2})^{-1}(S - \tilde{\Sigma}(\Lambda(\hat{S})\hat{S} + \Sigma^{-1}D\boldsymbol{\beta})) \tag{7.7}$$

$$\tilde{\boldsymbol{\beta}} = (\tilde{\Omega}^{1/2})^{-1}(\boldsymbol{\beta} - \tilde{\Omega}(D^T\Sigma^{-1}\tilde{\Sigma}\Lambda(\hat{S})\hat{S} + \Omega^{-1}\boldsymbol{\mu})) \tag{7.8}$$

where $\tilde{\Sigma}^{1/2}$ and $\tilde{\Omega}^{1/2}$ are Cholesky decomposition, will provide approximately uncorrelated components of $(\tilde{S}_1,...,\tilde{S}_n)$ and $\tilde{\beta}_1,...,\tilde{\beta}_p$ with zero mean

and unit variance! This parameterization of $S$ in (7.7) can be interpreted as "data-base" PNCP introduced in section 2.8.2. In this case, when the data is "weak" $\Lambda(\cdot) \equiv 0$ resulting in $\tilde{S} = (\Sigma^{1/2})^{-1}(S - D\boldsymbol{\beta})$ which is NCP; and when the data is "strong", $\tilde{\Sigma} \approx \Lambda(\hat{S})^{-1}$ resulting in

$$\begin{aligned}\tilde{S} &= (\tilde{\Sigma}^{1/2})^{-1}(S - \hat{S}) \\ &= (Var[S|y]^{1/2})^{-1}(S - E[S|y])\end{aligned}$$

which is the standardized version of CP.

After parameterization, $\tilde{S}$ and $\tilde{\boldsymbol{\beta}}$ are updated in two separate blocks by using Langevin algorithm. As mentioned in section 2.7.2, $\hat{l}^2/n^{1/3}$ and $\hat{l}^2/p^{1/3}$ with $\hat{l} = 1.65$ are used as the variances of discretization respectively for $\tilde{S}$ and $\tilde{\boldsymbol{\beta}}$ to achieve optimal performance (the acceptance rates are tuned to around 0.574).

## 7.3 Parameterization of $\boldsymbol{\eta}$

The posterior correlation between $\phi$ and $\sigma$ is commonly strong and requires a parameterization to make algorithm efficient. Zhang (2004) [32] showed that the two parameter $\sigma^2$ and $\phi$ for exponential covariance function are not consistently estimable, but $\sigma^2/\phi$ is. Therefore $\sigma^2/\phi$ should be used for parameterization. And considering the posterior distribution of such parameters usually are heavily skewed, the final parameterizations are

$$\nu_1 = \log(\sigma^2/\phi) \tag{7.9}$$

$$\nu_2 = \log(\sigma). \tag{7.10}$$

In more general cases, for Matern correlation family Zhang showed $\sigma^2/\phi^{2\nu}$ is consistently estimable, which leads to the parameterization $\nu_1 = \log(\sigma^2/\phi^{2\nu})$, $\nu_2 = \log(\sigma)$.

## 7.4 Flat priors

For GLMMs with a known singular correlation matrix for the random effects, the conditions for proper posterior are given in Sun et al. (2000) [30]. Gelfand and Sahu (1999) [14] studied general conditions for posterior property with an improper prior for $\boldsymbol{\beta}$ in a GLM. The use of flat priors should

be taken care with caution, as demonstrated in Natarajan and McCulloch (1995) [20] that an improper prior on the variance for the random effects of GLMMs may lead to an improper posterior. However, Christencen (2002) [7] provided the following proposition that guarantees a posterior under certain conditions.

> Consider a realization $y = (y_1, ..., y_n)$ of the Poisson log-linear spatial model, and assume that $y_1, ..., y_m$ are positive and $y_{m+1}, ..., y_n$ are zero. Let $\kappa_+(\phi)$ denote the correlation function of $S_+ = (S_1, ..., S_m)$ and $D_+ = (d_1, ..., d_m)^T$ the corresponding $m \times p$ design matrix. Suppose that $\phi, \boldsymbol{\beta}, \sigma$ are a priori independent with densities $\pi_a, \pi_b, \pi_c$, where $\pi_b(\boldsymbol{\beta}) \propto 1$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$. Then the posterior is proper if
>
> 1. $D_+$ has rank $p$;
> 2. $\kappa_+(\phi)$ is invertible for all $\phi \in supp\, \pi_a$;
> 3. $(|D_+^T \kappa_+^{-1}(\phi) D_+ || \kappa_+(\phi)|)^{-1/2} \pi_a(\phi)$is integrable on $[0, \infty]$;
> 4. $\int_0^\infty \sigma^{p-m} \pi_c(\sigma)\, \mathrm{d}\sigma < \infty$.

The suggested priors by Christencen are

$$\begin{array}{ll} \pi_a(\phi) & \propto 1/\phi, \log\phi \in [a_1, a_2] \\ \pi_b(\boldsymbol{\beta}) & \propto 1, \boldsymbol{\beta} \in \mathbb{R}^p \\ \pi_c(\sigma) & \propto \sigma^{-1} \exp(-c/\sigma), \sigma > 0 \end{array} \tag{7.11}$$

which satisfy the proposition.

# 8 Results for Rongelap data

## 8.1 "Fix-scan" Hastings-within-Gibbs Algorithm

The Poisson log-spatial model with matern correlation function was assumed for Rongelap data and Hastings-within-Gibbs algorithm was applied. The first 1000 iterations were discarded as "burn-in" period, and every $100^{th}$ iteration of the following 10000 iterations were stored which provided a sample of 1000 values from the posterior distribution. The corresponding results are shown in table 8.1 and figure 8.1, 8.2, 8.3.

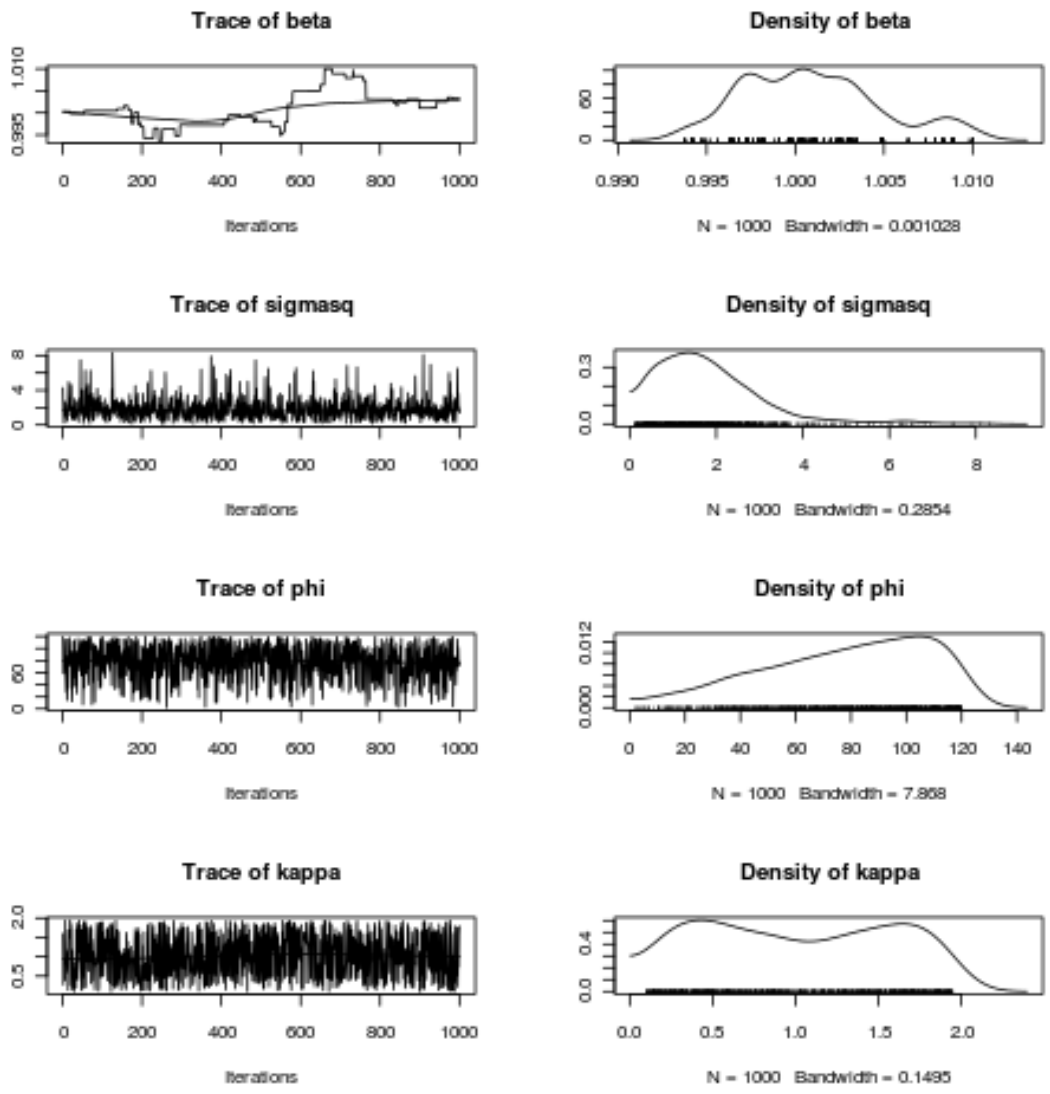


Figure 8.1: The Markov chains and approximated densities for posterior samples of $\beta, \sigma^2, \phi, \kappa$ for Rongelap data.

Table 8.1: Summary of the posterior samples of $\beta, \sigma^2, \phi, \kappa$ for Rongelap data.

| parameter | posterior mean | posterior median | 95% interval |
|---|---|---|---|
| $\beta$ | 1.00 | 1.00 | [0.99, 1.01] |
| $\sigma^2$ | 1.82 | 1.59 | [0.27, 5.27] |
| $\phi$ | 77.90 | 82.77 | [14.45, 118.57] |
| $\kappa$ | 1.02 | 0.99 | [0.14, 1.91] |

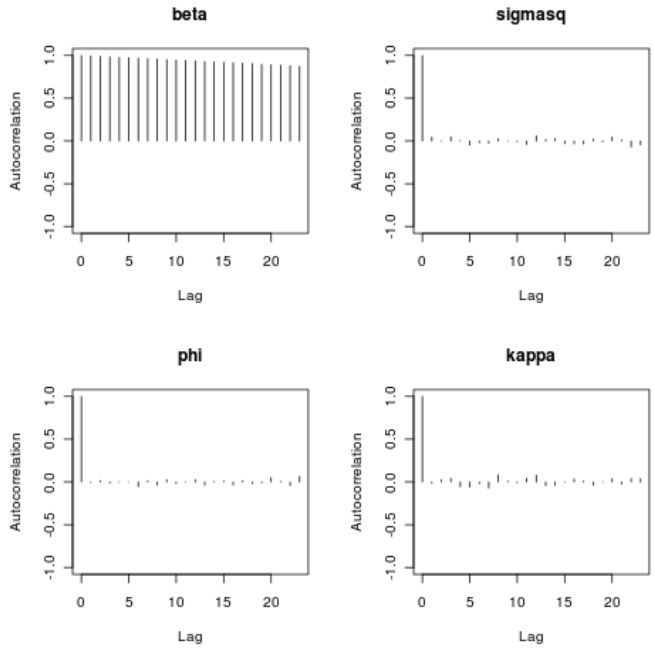


Figure 8.2: The autocorrelation for posterior samples of $\beta, \sigma^2, \phi, \kappa$ for Rongelap data.

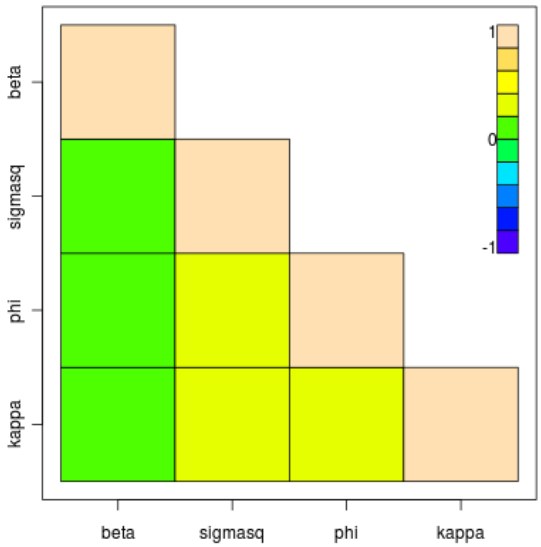


Figure 8.3: The cross-correlation among posterior samples of $\beta, \sigma^2, \phi, \kappa$ for Rongelap data.

## 8.2 ROBUST MCMC ALGORITHM

Robust MCMC algorithm describe in last section was applied on Rongelap data. The Poisson log-spatial model with powered exponential correlation function is assumed and set $\kappa = 1$ for simplicity. Total 10000 iterations with 2000 burning-in and 10 thinning is used, because the Markov chains converge much faster than when using "fix-scan" Hastings-within-Gibbs algorithm and it is enough to reveal significant improvement.

From figures 8.4, 8.5 and 8.6, it is clear to see that mixing of the chains is largely improved. The parameterization of $\beta$ improves its convergence as well as the mixing. To illustrate the effect of the parameterization of $(\phi, \sigma)$, scatter plots between $\phi$ and $\sigma$ and between their parameterized values $\nu_1 = \log\sigma$ and $\nu_2 = \log\frac{\sigma^2}{\phi}$ are plotted in figure 8.7. The heavy tails for $\phi$ and $\sigma$ are reduced and the strong correlation between them ,0.933, is reduced to -0.293.

Furthermore, Christensen et al. (2006) [8] compared the autocorrelation performance among CP, NCP and robust MCMC algorithm. As shown in figure 8.8, robust MCMC algorithm clearly outperform the other two, and CP is superior to NCP in this case because Rongelap data have very large counts with long recording periods ranging from 200 to 1800 seconds.

Table 8.2: Results for Rongelap data when using Robust MCMC algorithm.

| parameter | posterior mean | posterior median | 95% interval |
|---|---|---|---|
| $\beta$ | 1.82 | 1.82 | [1.55, 2.03] |
| $\sigma^2$ | 0.37 | 0.35 | [0.23, 0.78] |
| $\phi$ | 144.74 | 125.84 | [72.76, 337.01] |

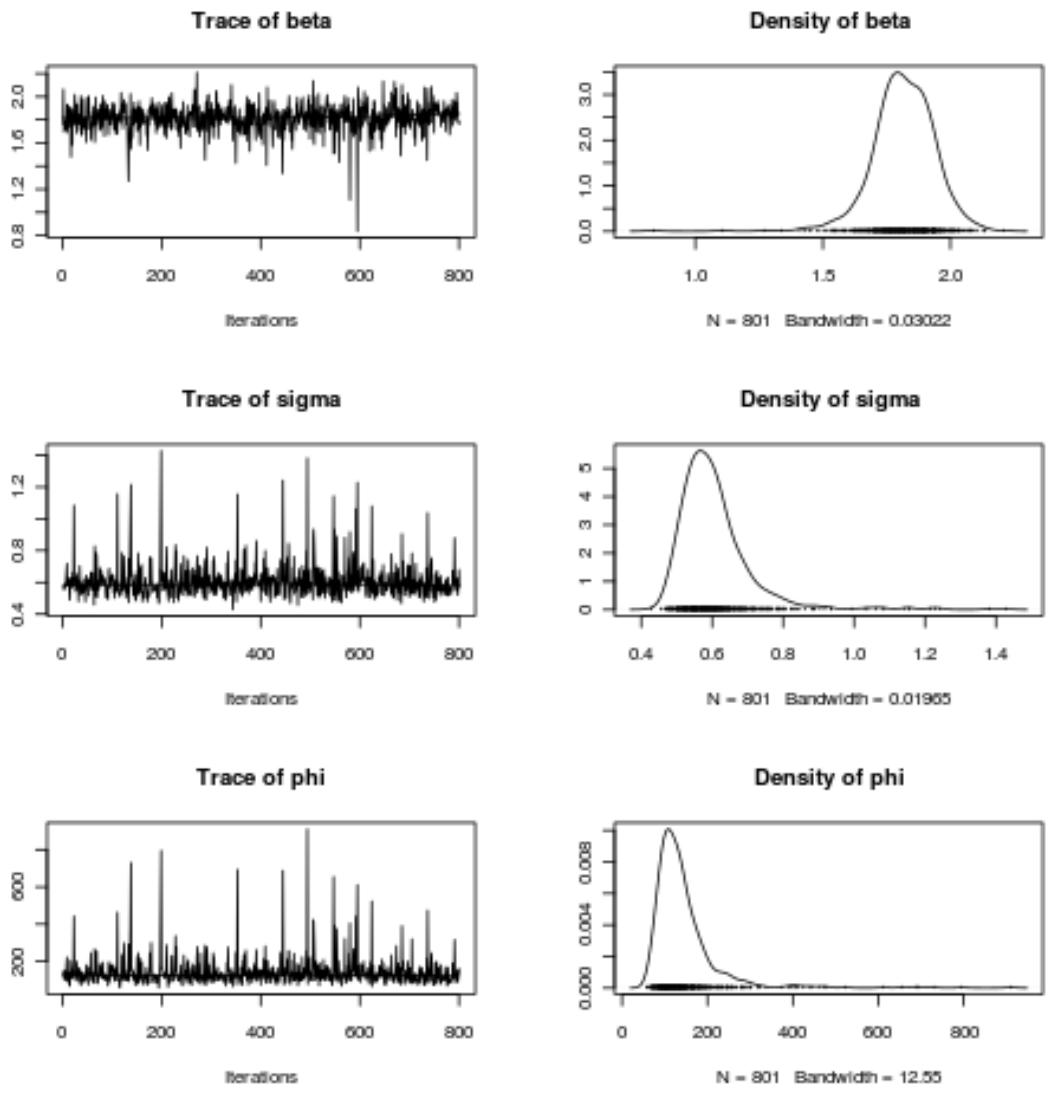


Figure 8.4: The Markov chains and approximated densities for Rongelap data when using Robust MCMC algorithm.

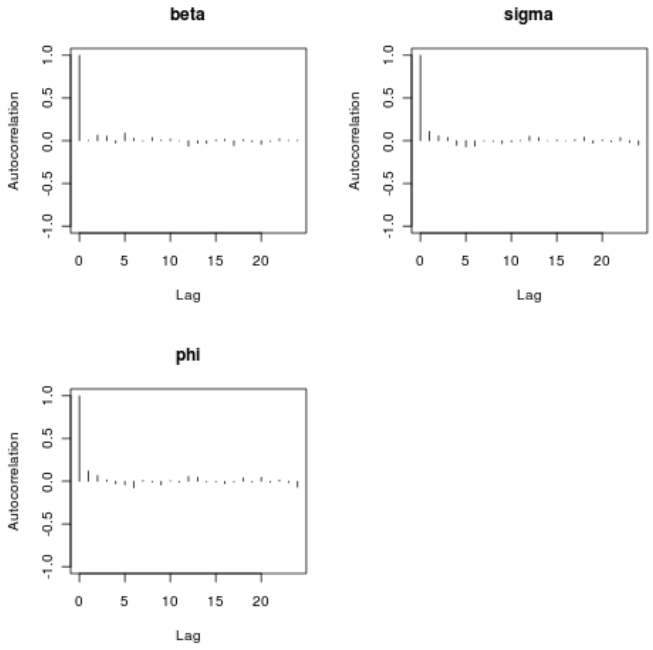


Figure 8.5: The autocorrelations for Rongelap data when using Robust MCMC algorithm.

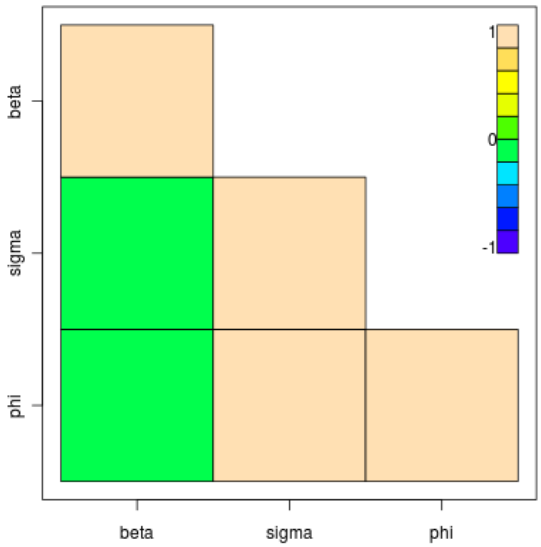


Figure 8.6: The cross-correlations for Rongelap data when using Robust MCMC algorithm.

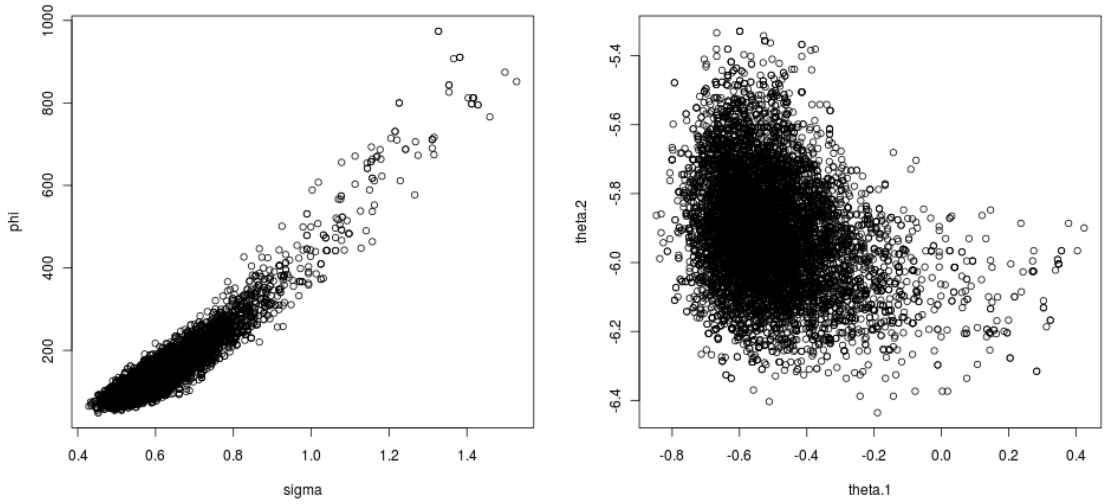


Figure 8.7: Scatter plots between $\phi$ and $\sigma$ (left), and between $\nu_1 = \log\sigma$ and $\nu_2 = \log\frac{\sigma^2}{\phi}$ (right) for Rongelap data when using Robust MCMC algorithm.

# REFERENCES

[1] Amit, Y. and Grenander, U. (1991). Comparing sweep strategies for stochastic relaxation.*J. Multivariate Analysis*, 37, No. 2, 197-222.

[2] Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian computation. J. Roy. *Statist. Soc. Ser. B*, 55, 25-38.

[3] Besag, J. (1994). Discussion of "Markov chains for exploring posterior distributions". *Ann. Statist.*, 22, 1734-1741.

[4] Besag, J., Green, P. J., Higdon, D. and Mmengeren, K. (1995). Bayesian computation and stochastic systems. *Statist. Sci.*, 10, 3-66.

[5] Chib, S. and Greenberg, E. (1994). Bayes inference for regression models with ARMA(p, q) errors. *Journal of Econometrics*, 64: 183-206.

[6] Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *American Statistician*, 49: 327-335.

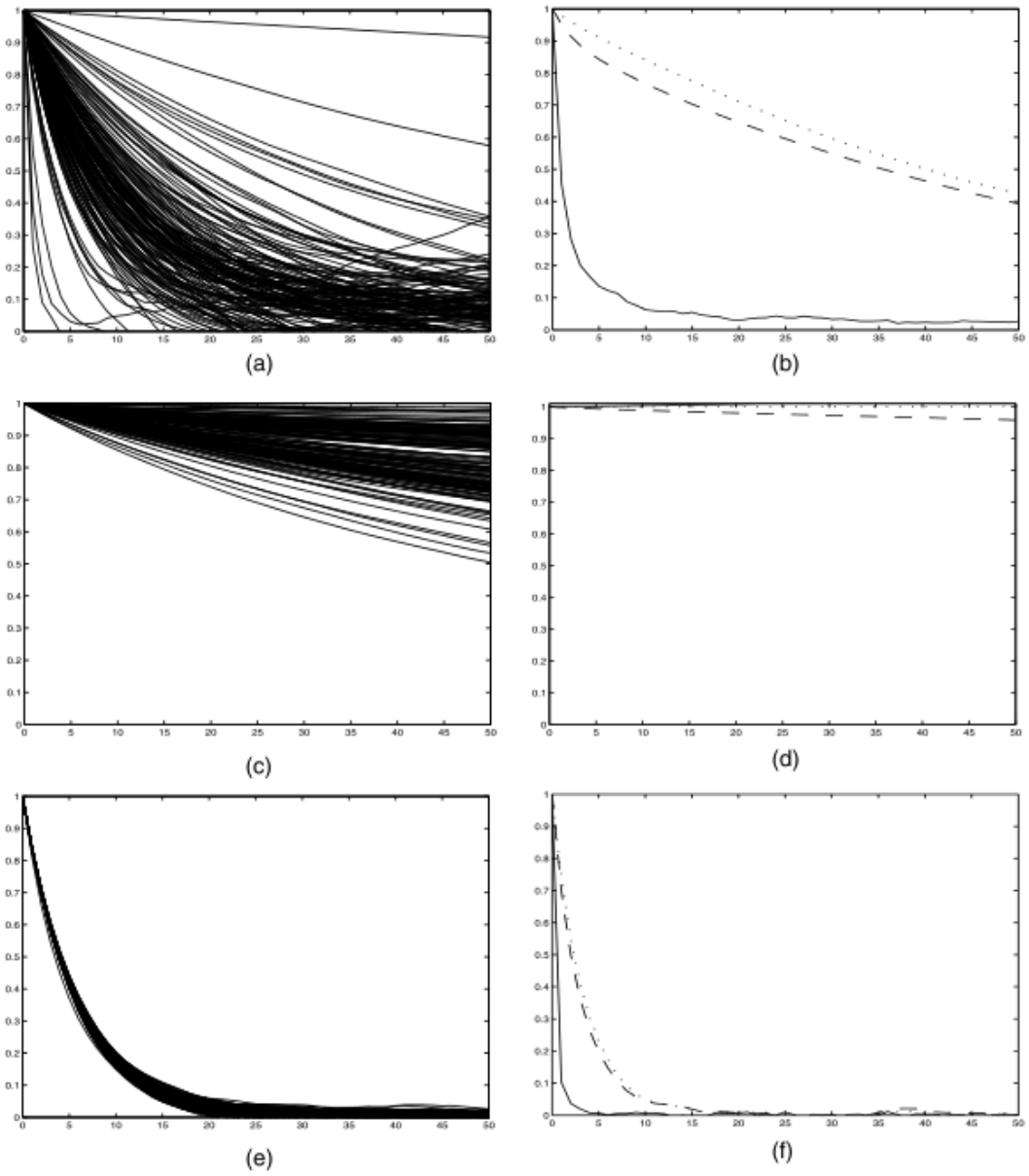


Figure 8.8: Autocorrelation performance among CP, NCP and robust MCMC algorithm. Left column: $S_1, ..., S_{157}$, right column: $\beta$ (solid), $\phi$ (dashed), $\sigma$ (dash-dotted); Top row: CP, middle row: NCP, bottom row: robust MCMC

[7] Christensen, O. F. (2002). *Methodology and Applications in Non-linear Model-based Geostatistics*, PhD dissertation, Aalborg University.

[8] Christensen, O. F., Roberts, G. O. and Sköld, M. (2006). Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *J. of Computational and Graphical Statistics*, Vol. 15, No. 1, 1-17.

[9] Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998). Model-based geostatistics. *J. Roy. Statist. Soc. Ser. C*, 47, 299-350.

[10] Diggle, P. J., Ribeiro, P. J. (2010). *Model-based geostatistics*, Springer Series in Statistics.

[11] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57: 97-109.

[12] Gelfand, A.E., Sahu, S. and Carlin, B. (1995). Efficient Parametrization for Normal Linear Mixed Effects Models. *Biometrika*, 82, 479-488.

[13] Gelfand, A.E., Sahu, S. and Carlin, B. (1996). Efficient Parametrization for Generalized Linear Mixed Effects Models. *Bayesian Statistics 5* (eds J.M. Bernardo, J. 0. Berger, A. P. Dawid and A. F. M. Smith), pp. 165-180, Oxford University Press, New York.

[14] Gelfand, A. E., Sahu, S. K. (1999). On the proprierity of posteriors and Bayesian identifiability in generalized linear models. *Journal of the American Statistical Association*, 94, 247-253.

[15] Gelfand, A. E. (2000). Gibbs Sampling. *Journal of the American Statistical Association*, 95, 1300-1304.

[16] Gentle, J.E. , HÃd'rdle, W. and Mori, Y. (eds) (2004). *Handbook of Computational Statistics: Concepts and Methods*. Berlin: Springer-Verlag.

[17] Grenander, U. and Miller, M. (1994). Representations of knowledge in complex systems (with discussion). *J. Royal Statist. Soc. Series B*, 56, 549-603

[18] Liu, J. S. (1994). The Collapsed Gibbs Sampler in Bayesian Computations With Applications to Gene Regulation Problem. *Journal of the American Statistical Association*, 89, 958-966.

[19] Liu, J. S., Wong, W. H. and Kong A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81, 27-40

[20] Natarajan, R., and McCulloch, C. E. (1995). A Note on the Existence of the Posterior Distribution for a Class of Mixed Models for Binomial Responses. *Biometrika*, 82, 639-643.

[21] Papaspiliopoulos, O., Roberts, G. O., and SkÃűld, M. (2003). Non-centered parameterizations for hierarchical models and data augmentation (with discussion). In *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 307-326.

[22] Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2007). A General Framework for the Parametrization of Hierarchical Models. *Statist. Sci.*, Volume 22, Number 1 (2007), 59-73.

[23] Phillips, D. B. and Smith, A. F. (1996), Bayesian model comparison via jump diffusions. *Markov chain Monte Carlo in Practice*, pp. 215-240, Chapman and Hall, New York.

[24] Roberts, G.O., and Smith, F.M. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications*, 49(2), 207-216.

[25] Roberts, G. and Tweedie, R. (1995). Exponential convergence for Langevin diffusions and their discrete approximations. Technical report, Statistics Laboratory, Univ. of Cambridge.

[26] Roberts, G. O., Gelman, A. and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms, *Annals of Applied Probability*, Vol. 7, No. 1, 110-120.

[27] Roberts, G. O. and Sahu, S. K. (1997). Updating Schemes, Correlation Structure, Blocking and Parameterisation for the Gibbs Sampler. *Journal of the Royal Statistical Society*, B, 59, 291-317.

[28] Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Vol. 60, No. 1, pp. 255-268.

[29] Seewald, (1992)Discussion on Parameterization issues in Bayesian inference (by S. E. Hills and A. F. M. Smith). *Bayesian Statistics 4* (eds J.M. Bernardo, J. 0. Berger, A. P. Dawid and A. F. M. Smith), pp. 241-243, Oxford: Oxford University Press.

[30] Sun, D., Speckman, P.L., and Tsutakawa, R.K. (2000). *Random eïňĂects in generalized linear mixed models (GLMMs)*. In Generalized Linear Models: A Bayesian Perspective, Dipak K. Dey, Bani K. Mallick and Sujit Ghosh, eds., Marcel Dekker Inc., 23-39.

[31] Whittaker, (1990). *Graphical Models in Applied Mathematical Multivariate Analysis*. New York: Wiley.

[32] Zhang, Hao (2004). Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics. *Journal of the American Statistical Association*, 99, 250-261.