UNIVERSITY OF TEXAS AT SAN ANTONIO

Hastings-within-Gibbs Algorithm: Introduction and Application on Hierarchical Model

Liang Jiang

April 2010

1 ABSTRACT

In this paper, common MCMC algorithms are introduced including Hastings-within-Gibbs algorithm. Then it is applied to a hierarchical model with simulated data set. "Fix-scan" technique is used to update the latent variables in the model. And the results are studied to explore the problems of the algorithm.

2 A SHORT INTRODUCTION OF MCMC

Markov Chain Monte Carlo (MCMC) algorithms are now widely used in all areas of statistics due to its flexibility and generality. The problems that MCMC addresses are from a broad range of diverse disciplines, including physics, engineering, computer science, mathematics and statistics. Here we focus on the problem that involves generating samples from a probability distribution, say $\pi(\boldsymbol{x})$. Although the exact form of the function $\pi(\cdot)$ is known, direct generating may be difficult or impossible perhaps because of the complexity of the function form of $\pi(\boldsymbol{x})$ or the high dimensionality of $\boldsymbol{x}$. MCMC resolve these difficulties by instead generating from a Markov chain whose invariant distribution is $\pi(\cdot)$.

Among MCMC algorithms, there are two fundamental mechanisms: one simplifies the high dimensional problems by successively generating from different subsets of $\boldsymbol{x}$; the other involves an accept/reject rule to "correct" an arbitrary Markov chain so that invariant distribution of $\pi(\cdot)$ is guaranteed. The first one is the spirit of *Gibbs sampler*, and the latter one is essentially *Metropolis-Hastings algorithm*. There are a lot of ways to apply these two methods and their improved versions, and they also can be combined for application in many different ways depending on the problems.

2.1 GIBBS SAMPLER

Given the target multivariate distribution $\pi(\boldsymbol{x}) = \pi(x_1, ..., x_p)$, the Gibbs samplers successively and repeatedly generates samples for each of the random variables, X_i, from the *full conditional distribution* $(X_i|X_1, ..., X_{i-1}, X_{i+1}, ..., X_p)$. The samples obtained in this way are guaranteed to converge to the stationary distribution $\pi(x_1, ..., x_p)$ under mild regularity conditions, Roberts

and Smith (1994) [8]. So for sufficiently large number of iterations, say N, the samples, $(\boldsymbol{X}^{(0)},...,\boldsymbol{X}^{(N)})$, can be seen as realizations from $\pi(\boldsymbol{x})$. The full algorithm is described as below.

1. Set up initial value, $\boldsymbol{X}^{(0)} = (X_1^{(0)},...,X_p^{(0)})$.

2. For iteration k from 1 to N, do the following steps,
   for random variable i from 1 to p,
   generate sample $X_i^{(k)}$ from $(X_i|X_1,...,X_{i-1},X_{i+1},...,X_p)$.

3. Return the values $(\boldsymbol{X}^{(0)},...,\boldsymbol{X}^{(N)})$.

This is the most widely used version of the Gibbs sampler, referred as "DUGS" in Roberts and Sahu (1997) [10], in which the components are updated in a natural ordering. The second way to update components in reverse direction $p, p-1,...,1$, referred as "REGS". The third way is "RSGS", in which we first generate a uniform value i from $\{1,...,p\}$ then update X_i and repeat this p times. The last one is "RPGS", in which a random permutation $Z = (z_1,...,z_p)$ of $\{1,...,p\}$ is generated to determine the order of updating. Roberts and Sahu (1997) [10] studied the rates of convergence for these updating strategies for Gaussian target distributions.

The price that Gibbs samples pays for reducing the dimensionality of random variables $\boldsymbol{X}$ is slow convergence and high correlation when the components of $\boldsymbol{X}$ exhibit heavy dependence. Detailed description of the relationship between correlation and convergence can be found in Roberts and Sahu (1997) [10]. More details about Gibbs sampler can be found in Gelfand (2000) [6] and Gentle (2004) [7].

2.2 Metropolis-Hastings algorithm

Unlike Gibbs sampler, the Metropolis-Hastings algorithm doesn't require the ability of generating samples from all the full conditional distributions. Instead, a *proposal or candidate distribution* is chosen given the current value of random variables, $\boldsymbol{X}^{(k)}$. Then the M-H algorithm is defined by two steps: first, generate a proposal value, $\boldsymbol{X}^*$, from the proposal distribution, $q(\cdot,\boldsymbol{X}^{(k)})$; second, the proposal value is accepted as the next value with the

probability

$$\alpha(\boldsymbol{X}^{(k)}, \boldsymbol{X}^*) = \begin{cases} min\{\frac{\pi(\boldsymbol{X}^*)q(\boldsymbol{X}^{(k)}, \boldsymbol{X}^*)}{\pi(\boldsymbol{X}^{(k)})q(\boldsymbol{X}^*, \boldsymbol{X}^{(k)})}, 1\} & if\ \pi(\boldsymbol{X}^{(k)})q(\boldsymbol{X}^*, \boldsymbol{X}^{(k)}) > 0 \\ 1 & otherwise \end{cases}; \tag{2.1}$$

if it is rejected, then the current value is taken as the next value in the Markov chain. The full algorithm is described as below.

1. Set up initial value, $\boldsymbol{X}^{(0)} = (X_1^{(0)}, ..., X_p^{(0)})$.

2. For iteration k from 1 to N, do the following steps,
   a) generate a proposal value, $\boldsymbol{X}^* \sim q(\cdot, \boldsymbol{X}^{(k)})$
   b) let

$$\boldsymbol{X}^{(k+1)} = \begin{cases} \boldsymbol{X}^* & if \quad Unif(0,1) \leq \alpha(\boldsymbol{X}^{(k)}, \boldsymbol{X}^*) \\ \boldsymbol{X}^{(k)} & otherwise \end{cases}.$$

3. Return the values $(\boldsymbol{X}^{(0)}, ..., \boldsymbol{X}^{(N)})$.

The M-H algorithm also pays a price for its flexibility. If the proposal distribution is poorly chosen, either the acceptance rate is low, or the Markov chain moves throughout the support of the invariant distribution too slow (even could be stuck around one place). In both cases, it leads to low efficiency of Monte Carlo sampling. Also, the choice of proposal distribution is application-dependent. One proposal that works well on one target distribution may be extremely poor on another.

RANDOM WALK PROPOSAL

One family of proposal distributions is given by $q(\boldsymbol{X}^*, \boldsymbol{X}) = q(\boldsymbol{X}^* - \boldsymbol{X})$. This is equivalent to drawn samples from $\boldsymbol{X}^* = \boldsymbol{X} + z$, where z follows the distribution $q(\cdot)$. Since the proposal is equal to the current value plus a "noise", this algorithm is called *random walk M-H*. The common choices of $q(\cdot)$ include multivariate normal distribution and multivariate t distribution.

INDEPENDENCE PROPOSAL

Hastings (1970) [5] introduced a second family of proposal distributions, $q(\boldsymbol{X}^*, \boldsymbol{X}) = q(\boldsymbol{y})$. It is usually referred as *independence M-H*, because the

proposal is drawn independently of the current value $\boldsymbol{X}^{(k)}$. For this proposal to work and not get stuck in the tails of $\pi(\cdot)$, it is necessary that $q(\boldsymbol{y})$ has thicker tails than $\pi(\cdot)$.

TAILORED PROPOSAL

Chib and Greenberg (1994, 1995) [1] [2] suggested to match the proposal distribution to the target by using a multivariate normal or multivariate t distribution which has same location for mode as target does and the dispersion given by inverse of the Hessian evaluated at the mode. This proposal distribution can be written as $q(\cdot) = f(\cdot|\boldsymbol{m}, V)$ where

$$\boldsymbol{m} = arg\,max\,log\pi(\boldsymbol{x}) \tag{2.2}$$

$$V = \tau\{-\frac{\partial^2 log\pi(\boldsymbol{x})}{\partial \boldsymbol{x}\partial \boldsymbol{x}'}\}^{-1}_{\boldsymbol{x}=\boldsymbol{m}} \quad . \tag{2.3}$$

3 HASTINGS-WITHIN-GIBBS ALGORITHM

3.1 GENERALIZED LINEAR SPATIAL MODELS

When it is important to model non-Gaussian sampling mechanism or a non-Gaussian distribution of response random variable is of interest, a more flexible and useful model frame should be employed – *generalized linear spatial model* (GLSM), first proposed by Diggle et al. (1998) [3]. 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 μ_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 [4] section 3.6);
- $\boldsymbol{\Sigma}$ is a variance-covariance matrix with entries $\sigma_{ij} = \sigma^2\rho(u_{ij})$: σ^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 [4] 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) [3].

3.2 MCMC FOR GLSM

For the model of interest – GLSM, full conditional distributions need to be obtained before MCMC algorithms can be conducted. Now let's consider the details of each of these conditional distributions.

First, the structure of GLSM implies that

$$p(Y|\theta,S) = \prod_i p(Y_i|\theta,S_i) \tag{3.4}$$

for which it follows that

$$p(\theta|S,Y) \propto \prod_i p(Y_i|\theta,S_i)\pi(\theta) \tag{3.5}$$

where $\pi(\theta)$ is the prior distribution of θ.

Second, if let $\pi(\eta)$ be the prior distribution of η, Bayes' Theorem immediately implies that

$$p(\eta|S) \propto p(S|\eta)\pi(\eta). \tag{3.6}$$

Finally, the conditional distribution for S is

$$p(S|\theta,\eta,Y) \propto p(Y|\theta,S)p(S|\eta). \tag{3.7}$$

By taking a deeper look at equation (3.7) and considering that the conditional distribution $p(S|\eta)$ follows multivariate Gaussian, the full conditional distribution of each latent variable S_i is

$$p(S_i|S_{-i},\theta,\eta,Y) \propto p(Y|\theta,S)p(S_i|S_{-i},\eta) \tag{3.8}$$

in which $p(S_i|S_{-i},\eta)$ reduces to normal distribution

$$[S_i|S_{-i},\eta] \sim N(-\sum_{j\neq i} Q_{ij}S_jQ_{ij}^{-1}, Q_{ii}^{-1}) \tag{3.9}$$

where Q_{ij} is the (i,j) element of the inverse matrix of Σ.

Following equations (3.4), (3.5), (3.6) and (3.8), the "fixed-scan" Hastings-within-Gibbs algorithm, used in Diggle et al. (1998) [3], is described as below.

1. Step 0: choose initial value for θ,η and S (for the Poisson log-spatial model $S(x_i)^{(0)} = log(Y_i + 0.01) - d(x_i)'\eta^{(0)}$)
2. Step 1: update all the components of η
   a) choose a proposed value η' from prior $p(\eta)$ (uniform prior was used by Diggle et al.)

b) accept η' with probability $min\{\frac{p(S|\eta')}{p(S|\eta)}, 1\}$, otherwise keep η

3. Step 2: update S one by one for all locations
   a) choose a proposed value S_i' for the ith location from $p(S_i'|S_{-i}, \eta)$
   b) accept S_i' with probability $min\{\frac{p(Y_i|S_i', \theta)}{p(Y_i|S_i, \theta)}, 1\}$, otherwise keep S_i
   c) repeat for all S_i, $i = 1, ..., n$
4. Step 3: update all the components of θ
   a) choose a proposed value θ' from proposal distribution $p(\theta'|\theta)$
   b) accept θ' with probability $min\{\frac{p(Y|\theta', S)p(\theta|\theta')}{p(Y|\theta, S)p(\theta'|\theta)}, 1\}$, otherwise keep θ

3.3 Results for Simulated data

One data sets, named as "data45" and shown in figure 3.1, is simulated from the Poisson log-spatial model with matern correlation function. The parameters are $D\boldsymbol{\beta} = \beta = 0.5, \sigma^2 = 2, \phi = 0.2, \kappa = 1.5$.

By using the MCMC algorithm described in previous section on true model, we generated a Markov chain for β, σ^2, ϕ while $\kappa = 1.5$ is fixed. The first 1000 iterations were discarded as "burn-in" period, and every 100^{th} iteration of the following 11000 iterations were stored which provided a sample of 1100 values from the posterior distribution. The illustration of the chains and the approximated densities are show in figure 3.2, and the autocorrelation for each parameter and the cross-correlation among them are shown in figure 3.3 and 3.4. Table 3.1 shows the mean, median and 95% interval for the posterior samples.

Table 3.1: Summary of the posterior samples of β, σ^2, ϕ for "data45".

parameter	true value	posterior mean	posterior median	95% interval
β	0.5	0.65	0.66	[-0.08, 1.26]
σ^2	2.0	1.52	1.39	[0.59, 2.86]
ϕ	0.2	0.13	0.13	[0.08, 0.19]

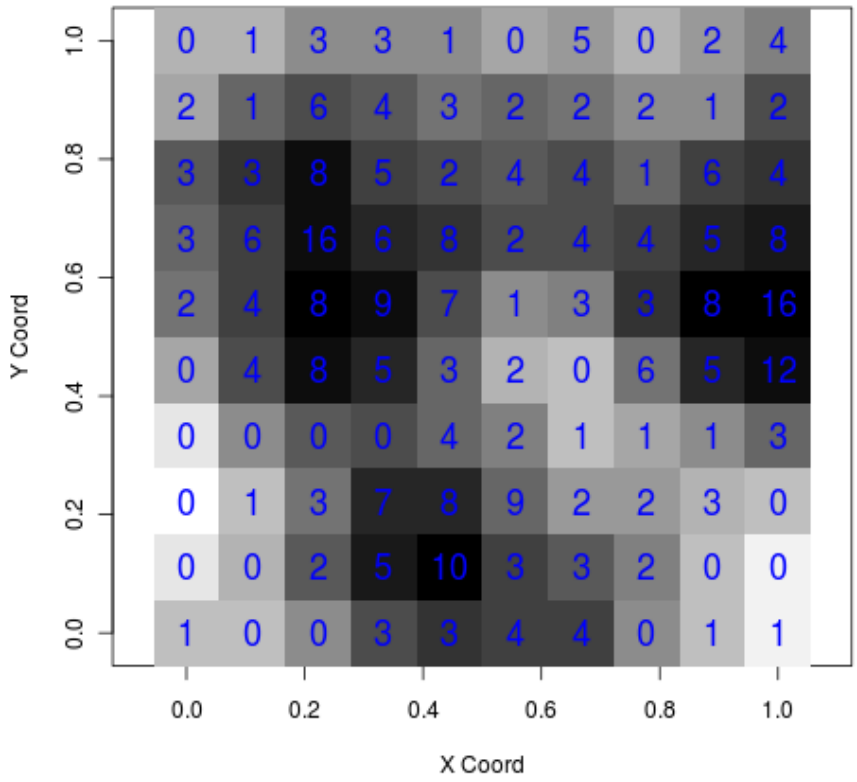

Figure 3.1: "data45" data: a simulated data set from the Poisson log-spatial model in a 10×10 unit grid.

4 CONCLUSION

From the results of "data45", we can clearly see several problems suffered by the Hastings-within-Gibbs algorithm,

1. slow mixing and strong autocorrelation: the chains for β have strong autocorrelation, figure 3.3, (autocorrelations for latent variables S_i are also strong);

2. significant dependence: cross-correlations among σ^2, ϕ, κ, figure 3.4;

3. 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 for GLSM, especially when the number of latent variables is large. To resolve these problems and improve the algorithm, addition techniques are required in future study.

REFERENCES

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

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

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

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

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

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

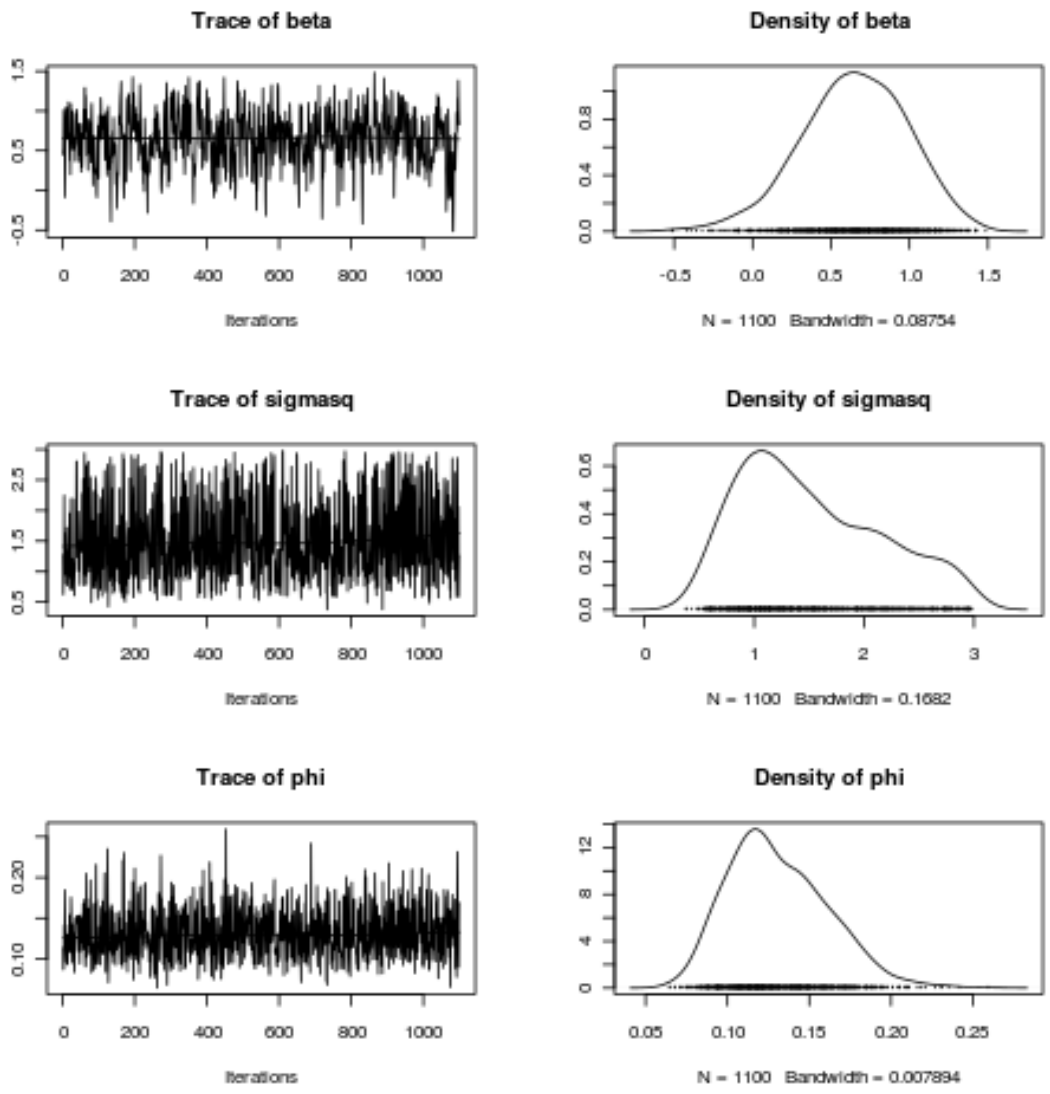

Figure 3.2: The Markov chains and approximated densities for posterior samples of β, σ^2, ϕ for "data45".

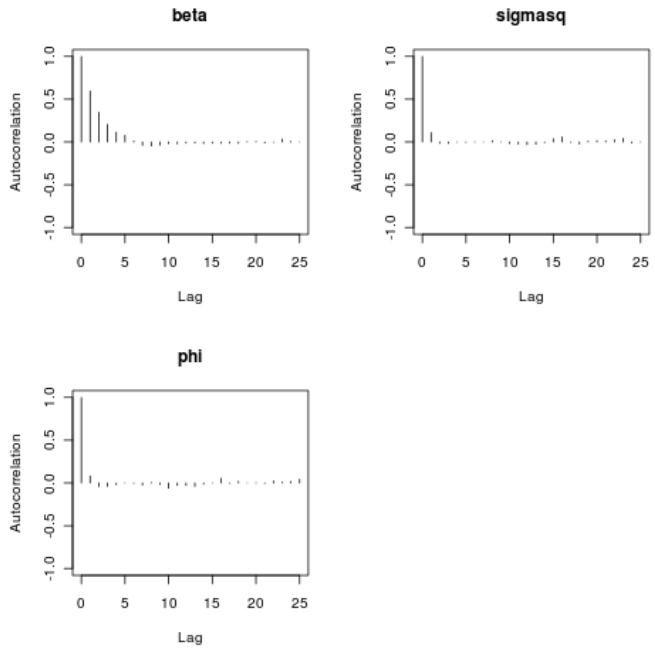

Figure 3.3: The autocorrelation for posterior samples of β, σ^2, ϕ for "data45".

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

[8] 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.

[9] 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.

[10] 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.

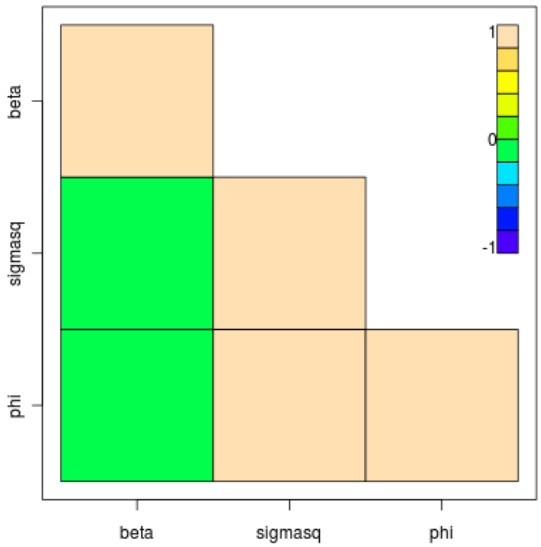

Figure 3.4: The cross-correlation among posterior samples of β, σ^2, ϕ for "data45".