



# Bayesian Model Determination for Geostatistical Regression Models

Devin S. Johnson

*Department of Mathematical Sciences*

*and*

*Institute of Arctic Biology*

*University of Alaska Fairbanks*

# Sponsor

This work is funded by the U.S. EPA STAR funded program STARMAP at Colorado State University, Department of Statistics

*The work reported here was developed under the STAR Research Assistance Agreement CR-829095 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University. This presentation has not been formally reviewed by EPA. The views expressed here are solely those of presenter and the STARMAP, the Program he represents. EPA does not endorse any products or commercial services mentioned in this presentation.*

# Whiptail lizard abundance

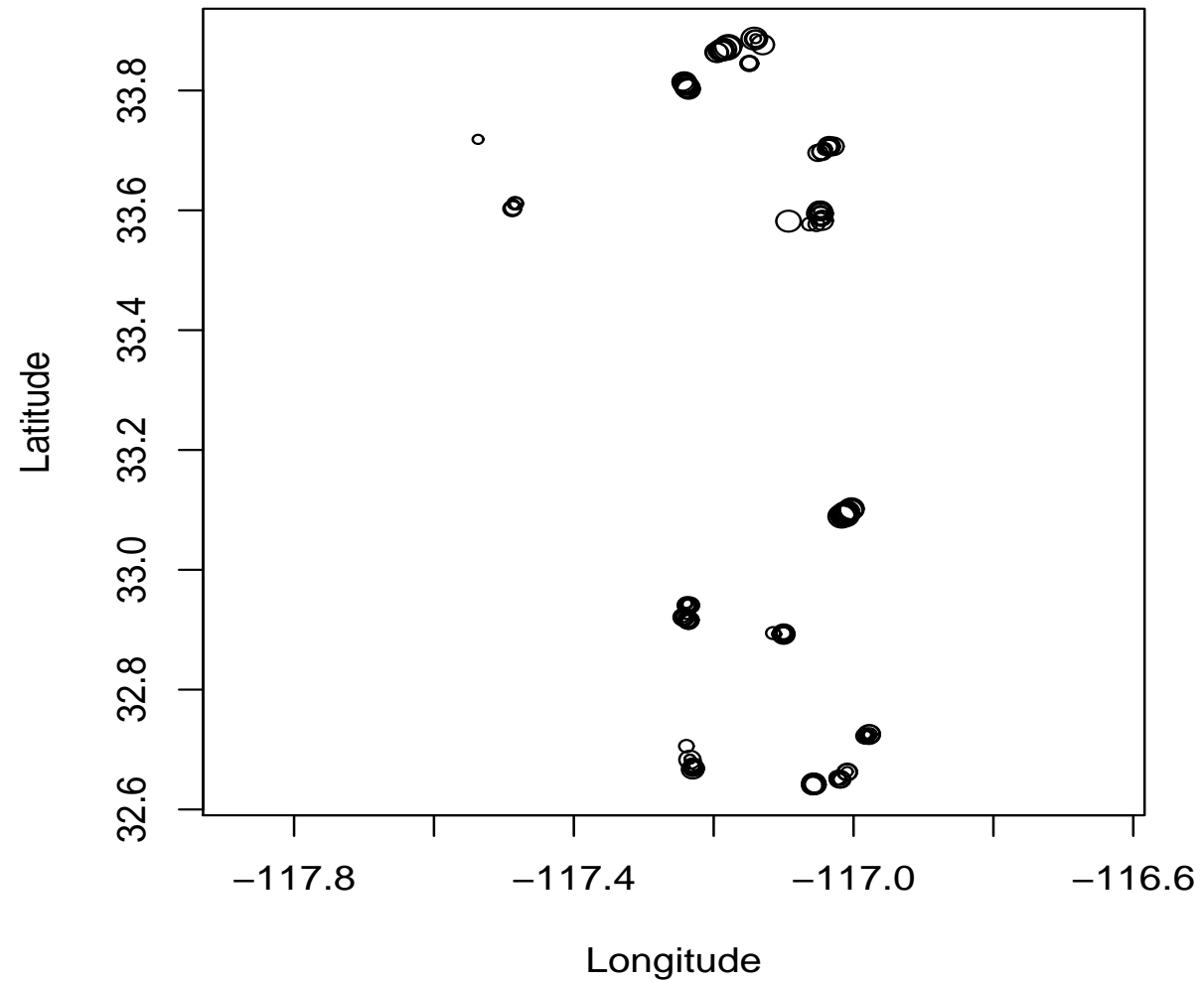
- Ver Hoef, Cressie, Fisher, Case (2001)
- Data measures abundance of the Orange-throated whiptail lizard in southern California
- $n = 148$  locations in 21 regions
- Response variable: Average number of lizards caught per day (log transformed)

$$Z(s) = \ln(\text{average \# caught at location } s)$$

# Whiptail lizard abundance

- Site covariates collected:
  - ◆ ant abundance (3 levels)
  - ◆  $\ln(\% \text{ sandy soil})$
  - ◆ elevation
  - ◆ bare rock indicator
  - ◆ % cover
  - ◆  $\ln(\% \text{ chaparral plants})$
- **Goal of analysis:** Determine which (if any) environmental covariates are useful for explaining lizard abundance

# Whiptail lizard locations



# Spatial correlation

- Close proximity of many observations calls independence into question
- Traditional regression subset procedures do not take spatial correlation into account.
- Large model space may prohibit some procedures that do take correlation into account
- Previous researchers have noted that failing to take spatial correlation into account can affect model selection results.

# Spatial regression models

$$Z(s) = \beta_0 + X_1(s)\beta_1 + \cdots + X_p(s)\beta_p + \delta(s)$$

where

- $s \in \mathcal{D} \subset \mathbb{R}^k$  is a spatial location
- $Z(s)$  is the response variable of interest
- $X_i(s)$  is an explanatory variable,  $i = 1, \dots, p$
- $\{\delta(s); s \in \mathcal{D}\}$  is a Gaussian random field with covariance function that decreases with distance

# Covariance function

$$\text{Cov}\{\delta(s), \delta(s')\} = \sigma^2(1 - \eta)\rho(h; \phi)$$

$$\text{Var}\{\delta(s)\} = \sigma^2$$

where,

- $h = s - s'$
- $\eta$  is the nugget parameter ( $0 < \eta < 1$ )
- $\sigma^2$  is the sill ( $0 < \sigma^2 < \infty$ )
- $\phi$  are the spatial correlation parameters

# Spatial correlation function

Exponential:

$$\rho(h; \phi) = \exp \left\{ -(h' \phi h)^{1/2} \right\}$$

Spherical:

$$\rho(h; \phi) = 1 - \frac{3}{2}(h' \phi h)^{1/2} + \frac{1}{2}(h' \phi h)^{3/2}$$

where  $\phi$  is a positive definite matrix

# Bayesian inference (Likelihood)

**Data:**  $Z = (Z(s_1), \dots, Z(s_n))'$

**Likelihood:**

$$L(Z|\beta, \sigma^2, \eta, \theta) = \mathcal{N}_n(Z; X\beta, \Sigma)$$

where

$$\Sigma = \sigma^2 \{ \eta I + (1 - \eta) \Omega \}$$

and

$$\Omega = [\rho(h_{ij}; \phi)]$$

# Bayesian inference (Posterior)

Bayesian inference is based on the conditional distribution of the parameters given  $Z$

$$\pi(\beta, \sigma^2, \eta, \phi | Z) \propto \mathcal{N}_n(Z; X\beta, \Sigma) \pi(\beta, \sigma^2, \eta, \phi)$$

where  $\pi(\beta, \sigma^2, \eta, \phi)$  is the *prior* distribution of the parameters.

Typical choices are:

$$\begin{aligned} \beta &\sim \mathcal{N}_p(\mu, V) & \sigma^2 &\sim \text{IG}(\nu/2, \lambda/2) \\ \eta &\sim \mathcal{U}(0, 1) & \phi &\sim \mathcal{W}(\Phi, \gamma) \end{aligned}$$

# Model selection issue

- In typical geostatistical analysis prediction is primary goal
  - ◆ covariance function is the object of interest
- In spatial regression,  $\beta_0, \dots, \beta_p$  is the primary focus
  - ◆ Determine relationship between covariates and response
  - ◆ Spatial correlation is often a nuisance.

# Incorporating model uncertainty

- Model incorporated as another parameter,  $M$  with sample space  $\mathcal{M} = \{m_0, \dots, m_K\}$
- For each  $m_k$  we have  $\theta_k = (\beta_k, \sigma^2, \eta, \phi)$
- Inference for the model can be made through the posterior distribution

$$P(\theta_k, m_k | Z) \propto P(Z | \theta_k, m_k) P(\theta_k | m_k) P(m_k)$$

# Posterior model probabilities

- Conditional distribution of the model given observed data

$$\begin{aligned} P(m_k|Z) &\propto \int P(Z|\theta_k, m_k)P(\theta_k|m_k)P(m_k)d\theta_k \\ &= P(Z|m_k)P(m_k) \end{aligned}$$

- The PMP is almost surely intractable
- There are several approximations for small model spaces

# Markov Chain Monte Carlo

- Appropriate for large model spaces
- Objective: Draw a sample  $(\theta_M^{(1)}, M^{(1)}), \dots, (\theta_M^{(N)}, M^{(N)})$  from  $P(\theta_k, m_k | Z)$ 
  - ◆ Construct a Markov chain with stationary distribution  $P(\theta_k, m_k | Z)$
  - ◆ Approximate  $P(m_k | Z)$  by the proportion of  $M^{(i)}$  that equal  $m_k$
  - ◆  $P(\beta_j \neq 0 | Z) = \sum_{k: \beta_j \neq 0} P(m_k | Z)$

# Reverse-Jump MCMC

For current state  $x = (\theta_k, m_k)$

1. Propose move of type  $i$  to  $m_{k'}$  from distribution  $J_i(x)$
2. Draw  $\theta_{k'}$  from  $G_i(x, m_{k'})$
3. Accept new state  $x'$  with probability

$$\min \left\{ 1, \frac{P(x'|Z)J_i(x')G_i(x')}{P(x|Z)J_i(x)G_i(x)} \right\}$$

# Difficulty with RJMCMC

- **Low acceptance rate:** Even if the appropriate model is chosen, bad parameter proposals will hinder mixing
- **Conjecture:** Proposals distributions  $G(x)$  close to  $P(\theta_{k'} | m_{k'}, Z)$  will produce the best results

Acceptance probability for  $P(\theta_{k'} | m_{k'}, Z)$

$$\min \left\{ 1, \frac{P(m_{k'} | Z) J(m_{k'})}{P(m_k | Z) J(m_k)} \right\}$$

# Partial analytic RJMCMC

- Godsill (2000) for AR order selection
- Use parameter proposal
  - ◆ Propose  $\beta_{k'} \sim P(\beta_{k'} | m_{k'}, \sigma^2, \eta, \phi, Z)$
  - ◆ Set  $(\sigma_{k'}^2, \eta_{k'}, \phi_{k'}) = (\sigma_k^2, \eta_k, \phi_k)$
- Acceptance probability

$$\min \left\{ 1, \frac{P(m_{k'} | \sigma^2, \eta, \phi, Z) J(m_{k'})}{P(m_k | \sigma^2, \eta, \phi, Z) J(m_k)} \right\}$$

No need to actually simulate  $\beta_{k'}$  values

# Model acceptance probability

- Suppose  $P(\beta_k|m_k) = \mathcal{N}_p(\mu_k, V_k)$
- Since  $Z = X\beta_k + \delta$ ,

$$Z|m_k, \sigma^2, \eta, \phi \sim \mathcal{N}_n(X_k\mu_k, X_kV_kX_k' + \Sigma)$$

$$\Rightarrow P(m_k|Z, \sigma^2, \eta, \phi)$$

$$\propto \exp \left\{ -\frac{1}{2}(Z - X_k\mu_k)'(X_kV_kX_k' + \Sigma)^{-1}(Z - X_k\mu_k) \right\} \\ \times P(m_k)$$

# Sampler for spatial regression

1. Update  $\beta_k$  from  $P(\beta_k | \dots)$  (Gibbs)
2. Update  $\sigma^2$  from  $P(\sigma^2 | \dots)$  (Gibbs)
3. Update  $\eta$  from  $P(\eta | \dots)$   
(Metropolis-within-Gibbs)
4. Update  $\phi$  from  $P(\phi | \dots)$   
(Metropolis-within-Gibbs)
5. Update  $m_k$  using partial analytic RJMCMC
6. goto step 1

# Whiptail lizard analysis

- Site covariates collected:
  - ◆ ant abundance (3 levels)
  - ◆  $\ln(\% \text{ sandy soil})$
  - ◆ elevation
  - ◆ bare rock indicator
  - ◆ % cover
  - ◆  $\ln(\% \text{ chaparral plants})$
- **Goal of analysis:** Determine which (if any) environmental covariates are useful for explaining lizard abundance

# Priors for analysis

We propose using the standard priors for the following parameters

- $M \sim 1/2^7$
- $\beta_k \sim \mathcal{N}_p(0, 100I)$  ( $\mathcal{N}(\cdot, \cdot)$  update)
- $\sigma^2 \sim \mathcal{IG}(0.01, 0.01)$  ( $\mathcal{IG}(\cdot, \cdot)$  update)
- $\eta \sim \mathcal{U}(0, 1)$

# Prior for $\phi$

We found the following to be a more flexible distribution than the standard Wishart

- Set  $\phi = LL'$  where  $L$  is a lower triangular matrix (Cholesky decomposition)
- Elements of  $L$  are  $(\alpha_{11}, \alpha_{21}, \alpha_{22})$
- Use priors
  - ◆  $\alpha_{11} \sim \text{Log}\mathcal{N}(0, 10)$
  - ◆  $\alpha_{21} \sim \mathcal{N}(0, 10)$
  - ◆  $\alpha_{22} \sim \text{Log}\mathcal{N}(0, 10)$

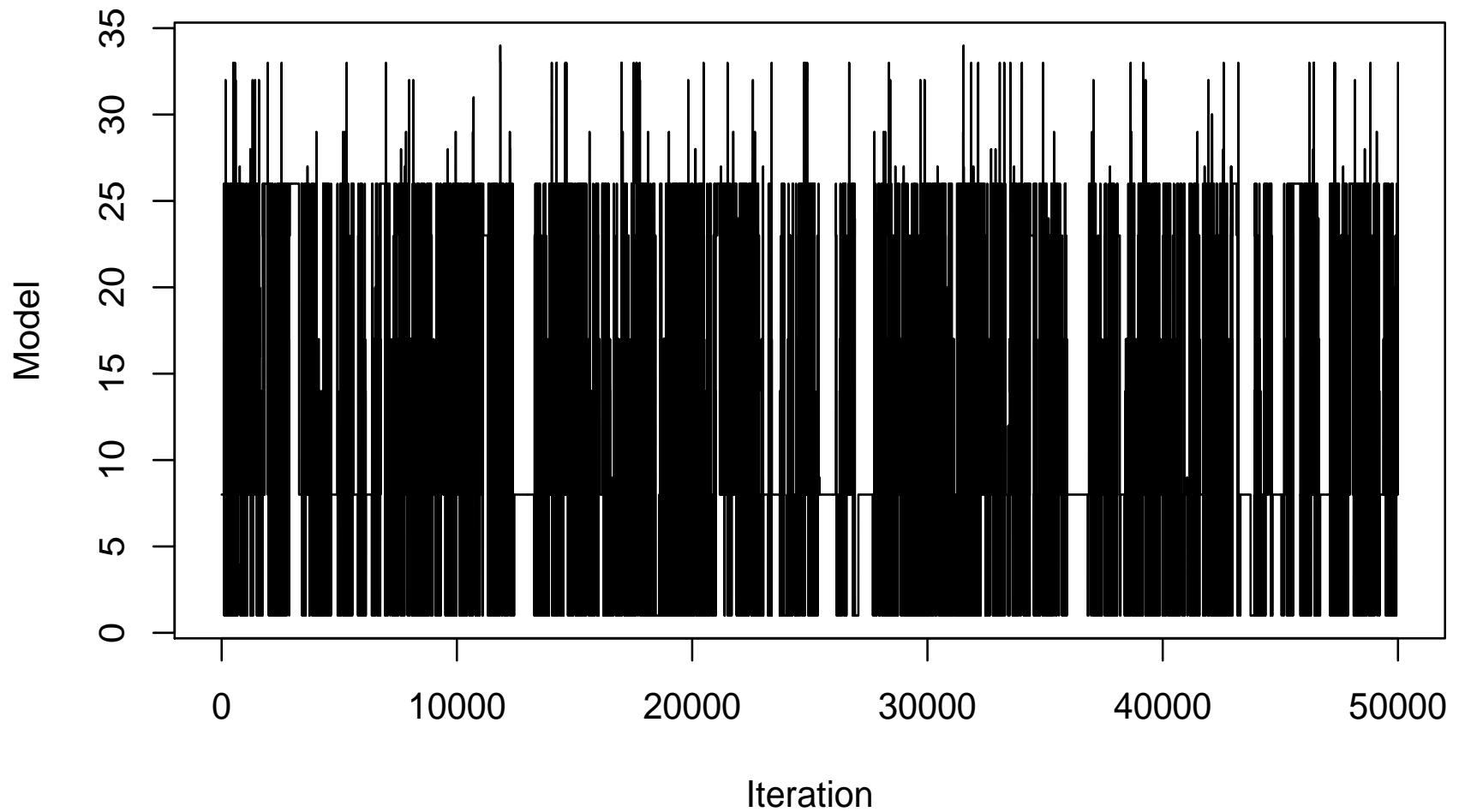
# Model jump proposal

## Random walk model proposals

- Select a covariate uniformly with probability  $1/7$
- If the covariate *is not* in the current model state, propose it for addition
- If the covariate *is* in the current model, propose it for deletion

Proposal mechanism is symmetric  
(i.e.  $J(m_{k'}) / J(m_k) = 1$ )

# Model chain



# Model chain summary

**Table 1.** Top five models measured by PMP for whiptail lizard spatial regression

Variables in Model	PMP
$\ln(\% \text{ Sandy soil})$	0.49
Intercept only	0.19
$\text{Ant}_1, \ln(\% \text{ Sandy soil})$	0.14
$\text{Ant}_1$	0.07
$\text{Ant}_2, \ln(\% \text{ Sandy soil})$	0.04

Chain visited 32 out of 128 possible models

# Posterior parameter probabilities

Table 2. Posterior probabilities of non-zero regression coefficients

Covariate	$P(\beta_j \neq 0 Z)$
Ant <sub>1</sub>	0.22
Ant <sub>2</sub>	0.07
ln(% Sandy soil)	0.71
Elevation	0.00
Bare Rock	0.01
% Cover	0.01
ln(% chaparral)	0.01

# Independence model ( $\eta = 1$ )

**Table 3.** Top five models measured by PMP for whiptail lizard regression (Independence model)

Variables in Model	PMP
$Ant_1, \ln(\% \text{ Sandy soil}), \ln(\% \text{ chaparral})$	0.46
$Ant_1, \ln(\% \text{ Sandy soil})$	0.22
$Ant_1, \ln(\% \text{ chaparral})$	0.18
$Ant_1, Ant_2, \ln(\% \text{ Sandy soil}), \ln(\% \text{ chaparral})$	0.04
$Ant_1, Ant_2, \ln(\% \text{ Sandy soil})$	0.02

Chain visited 34 out 128 possible models

# Posterior parameter probabilities

**Table 4.** Posterior probabilities of non-zero regression coefficients (Independence model)

Covariate	$P(\beta_j \neq 0 Z)$
Ant <sub>1</sub>	0.99
Ant <sub>2</sub>	0.09
ln(% Sandy soil)	0.77
Elevation	0.00
Bare Rock	0.03
% Cover	0.00
ln(% chaparral)	0.72

# Generalized linear models

## Data Model

$$Z(s_i) | \lambda(s_i) \sim f(\cdot; g^{-1}\{\lambda(s_i)\})$$

where  $E[Z(s_i) | \lambda(s)] = g^{-1}\{\lambda(s_i)\}$

## Parameter model

$$\lambda = (\lambda(s_1), \dots, \lambda(s_n))' \sim \mathcal{N}_n(X\beta; \Sigma)$$

where  $\Sigma$  is defined by a geostatistical covariance

# Note

Given  $\lambda$ , the model  $M$  for the covariate coefficients is independent of  $Z$  (hierarchical centering)

Therefore we can proceed as before, this time using

$$\begin{aligned} & P(m_k | Z, \lambda, \sigma^2, \eta, \phi) \\ &= P(m_k | \lambda, \sigma^2, \eta, \phi) \\ &\propto \exp \left\{ -\frac{1}{2} (\lambda - X_k \mu_k)' (X_k V_k X_k' + \Sigma)^{-1} (\lambda - X_k \mu_k) \right\} \\ &\quad \times P(m_k) \end{aligned}$$

in the acceptance ratio for model jumps

# Updates for Poisson data

To have an ergodic chain  $\lambda$  must be updated

## Langevin-Hastings Update

Use proposal

$$\lambda' \sim \mathcal{N}_n \left( \lambda + \frac{h}{2} \frac{\partial}{\partial \lambda} \log P(\lambda | \dots), hI \right).$$

e.g. for Poisson data

$$\frac{\partial}{\partial \lambda} \log P(\lambda | \dots) = Z - \exp(\lambda) + \Sigma^{-1}(\lambda - X\beta_k)$$

# Simulated data set

## Data

$$Z(s_i) \sim \text{Poisson}\{\lambda(s_i)\}; \quad i = 1, \dots, 100$$

## Random effect

$$\lambda(s) = 2 + 0.25X_1(s) + \delta(s)$$

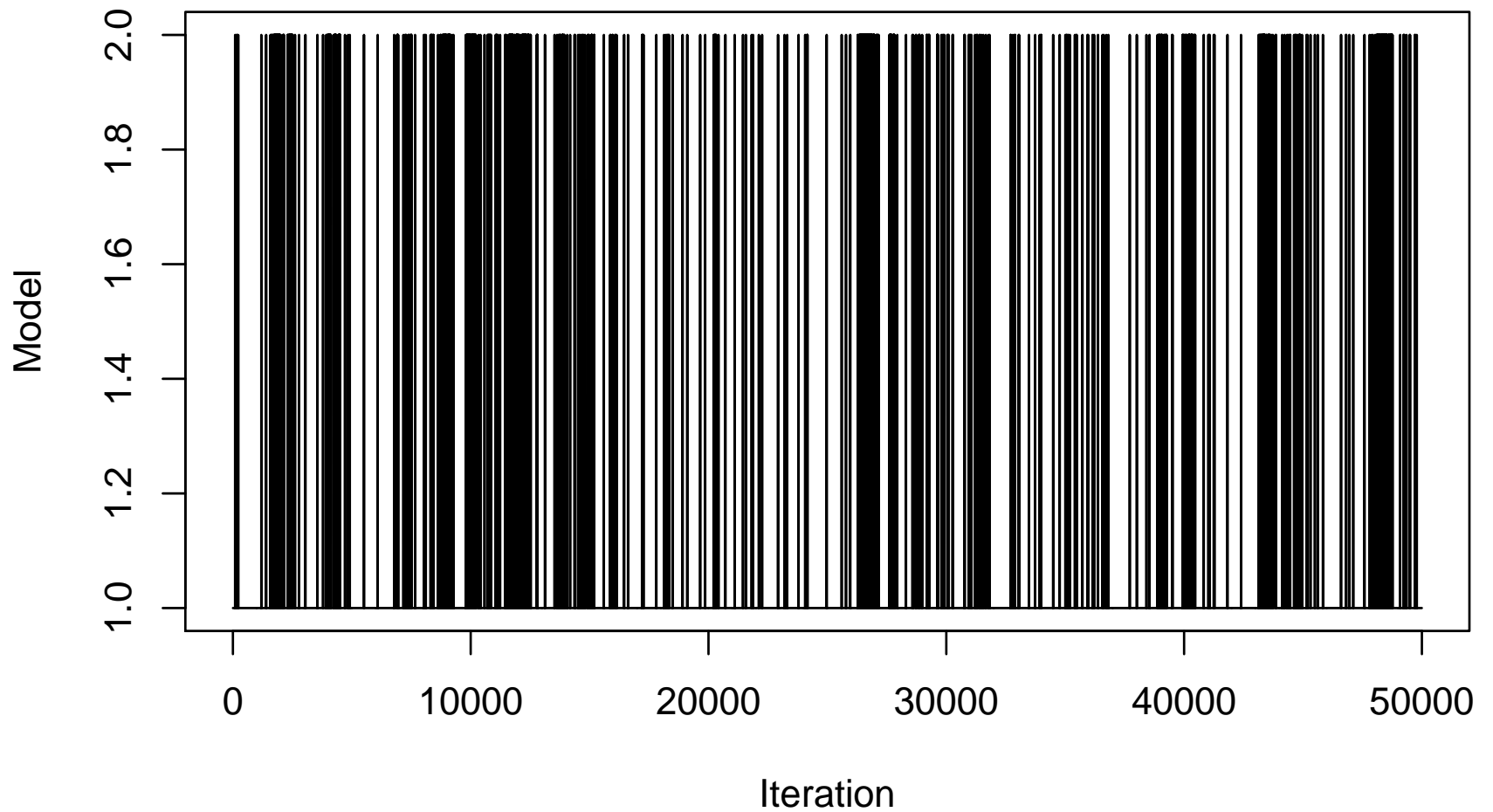
## Spatial model

$$\sigma^2 = 1, \eta = 0, \phi = I, 10 \times 10 \text{ Domain}$$

## Covariates

$$X_1(s) \sim \sqrt{(12/10)}t_{12}, \quad X_2(s) \sim t_3$$

# Model trace



# Posterior model analysis

**Table 5.** PMP for simulated GLM regression

Variables in Model	PMP
$X_1$	0.95
$X_1, X_2$	0.05
$X_2$	0.00
Intercept only	0.00

Chain visited 2 out of 4 possible models

# Discussion

- Failure to account for spatial correlation can lead to incorrect model inference
- Independence model selection procedures tend to add significant covariates to account for ignored correlation structure
- Partial analytic RJMCMC provides a straightforward method of model update in an MCMC sampler
- Straightforward extension to generalized linear spatial models