



Bayesian Model Selection 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.

Program

- Introduction
- Geostatistical regression
- Bayesian model determination
- Computational issues
- Whipstall lizard abundance
- Extension to GLMM
- Analysis of simulated Poisson spatial data

Introduction

- Environmental data is often collected within a spatial domain
- Failing to take spatial correlation into account can affect regression model selection results.
- Previous methods to account for spatial correlation in model selection
 - ◆ Ver Hoef et al. (2001), Spatial stepwise subset selection
 - ◆ Thompson (2001), Bayes factor approx.
 - ◆ Hoeting et al. (2005), Spatial AICC

Benefits of current approach

- Allows inclusion of expert knowledge in selection of regression covariates
 - ◆ Spatial stepwise and AICC methods treat all covariates equally
- Selection over large model spaces
 - ◆ In both the Bayes factor and AICC methods parameters in each model must be separately estimated
- Extension to GLMMs

Geostatistical regression models

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

where

- $s \in \mathcal{D} \subset \mathbb{R}^2$ 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,

- $h = s - s'$
- η is the nugget parameter ($0 < \eta < 1$)
- σ^2 is the sill ($0 < \sigma^2 < \infty$)
- ϕ 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 ϕ is a 2×2 positive definite matrix

Model selection issue

- In typical geostatistical analysis prediction is primary goal
 - ◆ covariance function is the object of interest
 - ◆ covariates can be a nuisance
- In spatial regression, β_0, \dots, β_p is the primary focus
 - ◆ Determine relationship between covariates and response
 - ◆ Spatial correlation is often a nuisance.

Prior distributions

Bayesian inference requires a choice of prior distributions for model parameters Typical choices for geostatistical regression models are:

- $\beta \sim \mathcal{N}_p(\mu, V)$
- $\sigma^2 \sim \mathcal{IG}(\alpha/2, \nu/2)$
- $\eta \sim \mathcal{U}(0, 1)$
- $\phi \sim \mathcal{W}(\Phi, \gamma)$

Reparameterization of ϕ

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

- Set $\phi = A\Psi A$ where A is a diagonal matrix with positive elements (A_k) and Ψ is a correlation matrix with offdiagonal element ψ
- Let $(\alpha_1, \alpha_2) = (\log A_1, \log A_2)$
- Use priors
 - ◆ $(\alpha_1, \alpha_2) \sim \mathcal{N}_2(\alpha_0, \Sigma_\alpha)$
 - ◆ $\psi \sim \mathcal{U}(-1, 1)$

Bayesian model determination

- 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 model probability (PMP)

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

Model prior distribution

A classic model prior is derived by treating inclusion of the p coefficients as a series of independent Bernoulli trials with probability π_j . The result is the following prior

$$P(m_k) = \prod_{j=1}^p \pi_j^{I_j} (1 - \pi_j)^{1-I_j},$$

where I_j is the indicator that $\beta_j \neq 0$

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 (2001) 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 β_k from $P(\beta_k | \dots)$ (Gibbs)
2. Update σ^2 from $P(\sigma^2 | \dots)$ (Gibbs)
3. Update η from $P(\eta | \dots)$
(Metropolis-within-Gibbs)
4. Update ϕ from $P(\phi | \dots)$
(Metropolis-within-Gibbs)
5. Update m_k using partial analytic RJMCMC
6. goto step 1

Whiptail lizard abundance

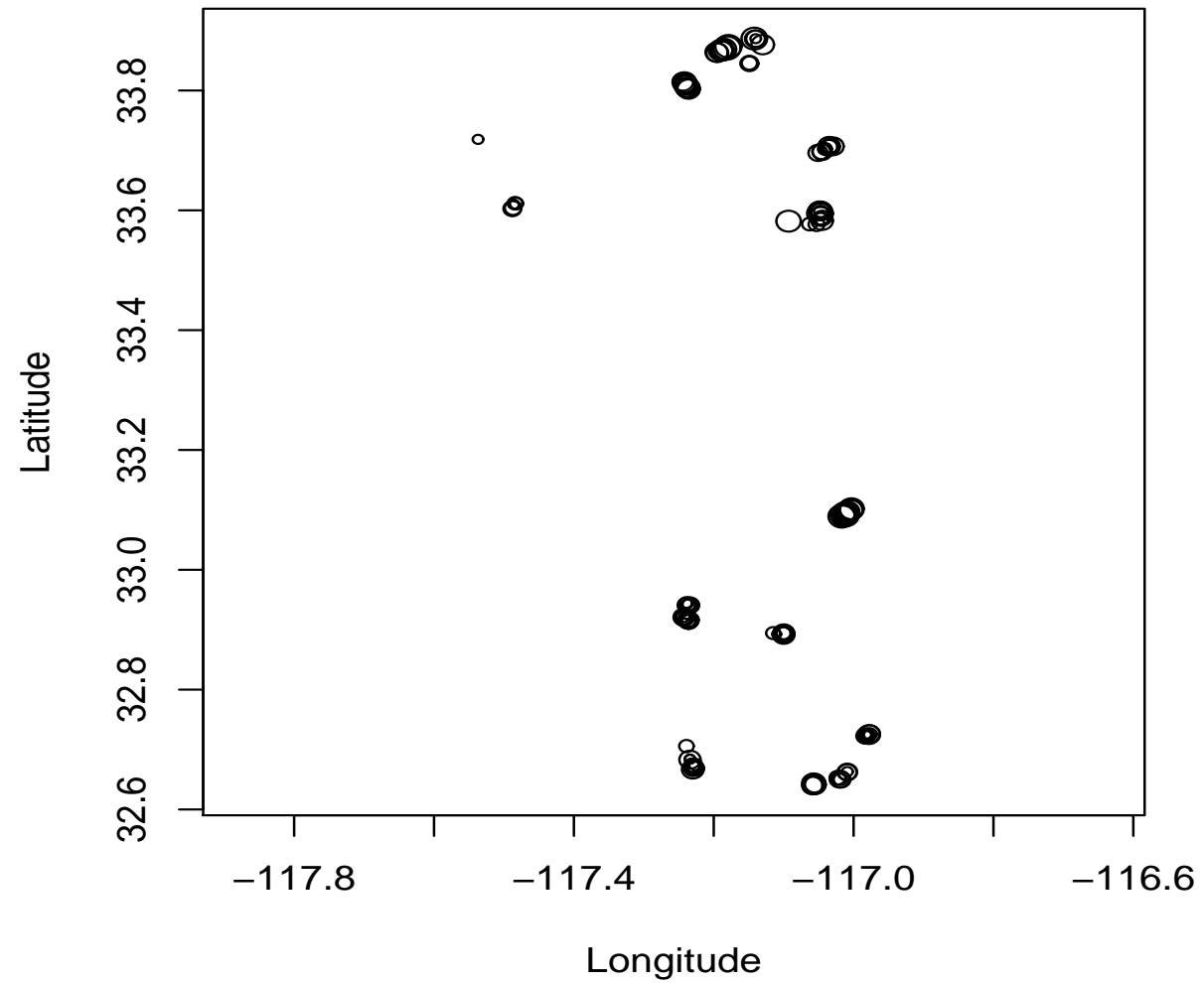
- Ver Hoef, Cressie, Fisher, Case (2001)
- Data measures abundance of the Orange-throated whiptail lizard in southern California
- $n = 148$ locations in 21 regions
- Response variable: Average number of lizards caught per day (log transformed)

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

Whiptail lizard analysis

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

Whiptail lizard locations



Priors for analysis

We propose using the standard priors for the following parameters

- $M \sim 1/2^7$
- $\beta_k \sim \mathcal{N}_p(0, 10\sigma^2(X_k'X_k)^{-1})$ ($\mathcal{N}(\cdot, \cdot)$ update)
- $\sigma^2 \sim \mathcal{IG}(0.005, 0.005)$ ($\mathcal{IG}(\cdot, \cdot)$ update)
- $\eta \sim \mathcal{U}(0, 1)$
- $(\alpha_1, \alpha_2) \sim \mathcal{N}(0, 100I)$
- $\psi \sim \mathcal{U}(-1, 1)$

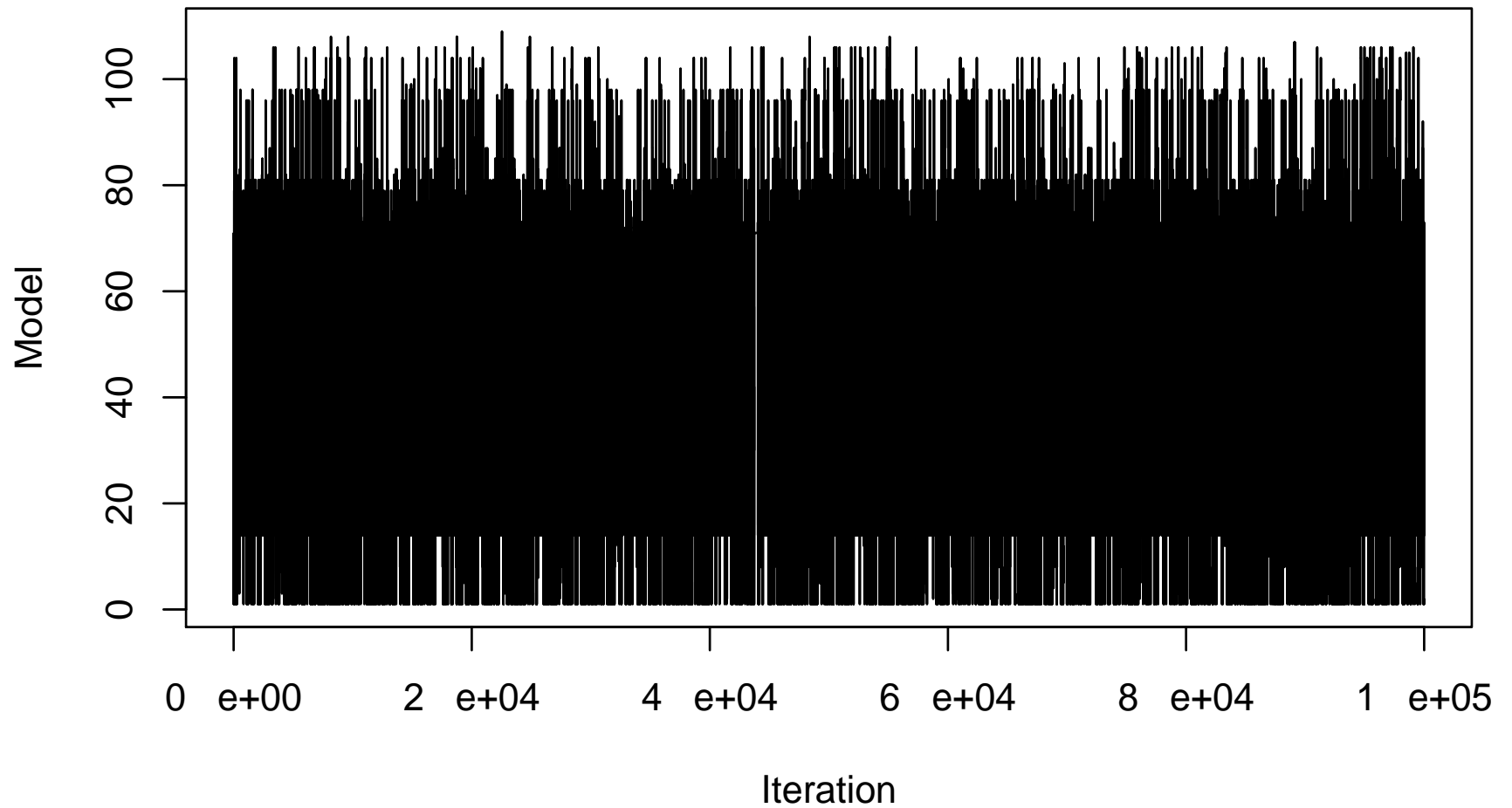
Model jump proposal

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

Top five models measured by
PMP for whiptail lizard spatial regression

Variables in Model	PMP
Sandy soil	0.17
Ant ₁ , Sandy soil	0.16
Sandy soil, % Cover	0.12
Ant ₁ , Sandy soil, % Cover	0.09
Ant ₁	0.05

Chain visited 109 out of 128 possible models

Posterior parameter probabilities

Environmental Covariate	Spatial	Indep.
Ant ₁	0.43	0.99
Ant ₂	0.13	0.51
ln % Sandy Soil	0.83	0.88
Elevation	0.18	0.65
Bare Rock	0.05	0.49
% Cover	0.34	0.39
ln % Chaparral	0.05	0.79

Generalized linear models

Data Model

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

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

Parameter model

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

where Σ is defined by a geostatistical covariance

Model acceptance

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

Therefore we can proceed as before, this time using

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

in the acceptance ratio for model jumps

Updates for Poisson data

To have an ergodic chain Z must be updated

Langevin-Hastings Update

Use proposal

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

e.g. for Poisson data

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

Simulated data set

Data

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

Random effect

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

Spatial model

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

Covariates

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

Posterior model analysis

Table 5. PMP for simulated GLM regression

Variables in Model	PMP
X_1	0.69
X_1, X_2	0.31
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

Current research

- Robustness of GLMM method to hierarchical centering (Christensen (2004) notes hierarchical centering can produce chains with high autocorrelation)
- Accounting for transformation/link uncertainty
- Simulation and theoretical investigation of "signal to noise effect" (as covariates become very influential does inclusion of a spatial covariance change selection results)