

# Bayesian inference for spatial GP models

Abhi Datta

04.28.2017

# Review of last lecture

- Bayesian Principles – Bayes theorem, posterior inference, credible intervals
- Bayesian Linear model
- Conjugate Normal-Inverse Gamma priors for  $(\beta, \sigma^2)$
- Sampling based inference – Monte Carlo Integration
- Composition sampling – Scope and limitations

# Bayesian inference for spatial linear model

- $y(s) = x(s)' \beta + w(s) + \epsilon(s)$ ,  $w(s) \sim GP(0, C(\cdot, \cdot | \phi))$ ,  
 $\epsilon \stackrel{\text{iid}}{\sim} N(0, \tau^2)$
- For  $n$  locations, **unmarginalized model**:  $y \sim N(X\beta + w, \tau^2 I)$ ,  
 $w \sim N(0, \sigma^2 R(\phi))$
- **Marginalized model**:  $y \sim N(X\beta, \sigma^2 R(\phi) + \tau^2 I)$
- Assume  $\phi$  is known,  $\sigma^2 \sim IG(a_\sigma, b_\sigma)$ ,  $\tau^2 \sim IG(a_\tau, b_\tau)$  and  
 $\beta \sim N(\mu, V)$
- Composition sampling does not help with either of the models
- How to do Bayesian inference ?

# Unmarginalized model

- Likelihood:  $N(y | X\beta + w, \tau^2 I) \times N(w | 0, \sigma^2 R(\phi)) \times N(\beta | \mu, V) \times IG(\sigma^2 | a_\sigma, b_\sigma) \times IG(\tau^2 | a_\tau, b_\tau)$

# Unmarginalized model

- Likelihood:  $N(y | X\beta + w, \tau^2 I) \times N(w | 0, \sigma^2 R(\phi)) \times N(\beta | \mu, V) \times IG(\sigma^2 | a_\sigma, b_\sigma) \times IG(\tau^2 | a_\tau, b_\tau)$
- Observe that
  - $\beta | \sigma^2, \tau^2, w, y \sim N(\mu^*, V^*)$
  - $w | \sigma^2, \tau^2, \beta, y \sim N(m, C^*)$
  - $\sigma^2 | \beta, \tau^2, w, y \sim IG(a_\sigma^*, b_\sigma^*)$
  - $\tau^2 | \beta, \sigma^2, w, y \sim IG(a_\tau^*, b_\tau^*)$
- Can we use these nice *full conditionals* to obtain posterior inference?

# Unmarginalized model

- Likelihood:  $N(y | X\beta + w, \tau^2 I) \times N(w | 0, \sigma^2 R(\phi)) \times N(\beta | \mu, V) \times IG(\sigma^2 | a_\sigma, b_\sigma) \times IG(\tau^2 | a_\tau, b_\tau)$
- Observe that
  - $\beta | \sigma^2, \tau^2, w, y \sim N(\mu^*, V^*)$
  - $w | \sigma^2, \tau^2, \beta, y \sim N(m, C^*)$
  - $\sigma^2 | \beta, \tau^2, w, y \sim IG(a_\sigma^*, b_\sigma^*)$
  - $\tau^2 | \beta, \sigma^2, w, y \sim IG(a_\tau^*, b_\tau^*)$
- Can we use these nice *full conditionals* to obtain posterior inference?
- **Yes!** Via Gibbs sampling

# Gibbs sampling

- Suppose that  $\theta = (\theta_1, \theta_2)$  and we seek the posterior distribution  $p(\theta_1, \theta_2 | \mathbf{y})$ .
- For many interesting hierarchical models, we have access to *full conditional distributions*  $p(\theta_1 | \theta_2, \mathbf{y})$  and  $p(\theta_2 | \theta_1, \mathbf{y})$ .
- The *Gibbs sampler* proposes the following sampling scheme. Set starting values  $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})$  For  $j = 1, \dots, M$ 
  - Draw  $\theta_1^{(j)} \sim p(\theta_1 | \theta_2^{(j-1)}, \mathbf{y})$
  - Draw  $\theta_2^{(j)} \sim p(\theta_2 | \theta_1^{(j)}, \mathbf{y})$
- This constructs a *Markov Chain* and, after an initial “burn-in” period when the chains are trying to find their way,  $\{\theta_1^{(j)}, \theta_2^{(j)}\}_{j=M_0+1}^M$  will be *Markov Chain Monte Carlo (MCMC)* samples from  $p(\theta_1, \theta_2 | \mathbf{y})$ , where  $M_0$  is the burn-in period..

# Gibbs sampling

- More generally, if  $\theta = (\theta_1, \dots, \theta_p)$  are the parameters in our model, we provide a set of initial values  $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_p^{(0)})$  and then performs the  $j$ -th iteration, say for  $j = 1, \dots, M$ , by updating successively from the *full conditional* distributions:

$$\theta_1^{(j)} \sim p(\theta_1^{(j)} | \theta_2^{(j-1)}, \dots, \theta_p^{(j-1)}, \mathbf{y})$$

$$\theta_2^{(j)} \sim p(\theta_2^{(j)} | \theta_1^{(j)}, \theta_3^{(j-1)}, \dots, \theta_p^{(j-1)}, \mathbf{y})$$

...

(the generic  $k^{th}$  element)

$$\theta_k^{(j)} \sim p(\theta_k^{(j)} | \theta_1^{(j)}, \dots, \theta_{k-1}^{(j)}, \theta_{k+1}^{(j-1)}, \dots, \theta_p^{(j-1)}, \mathbf{y})$$

...

$$\theta_p^{(j)} \sim p(\theta_p^{(j)} | \theta_1^{(j)}, \dots, \theta_{p-1}^{(j)}, \mathbf{y})$$

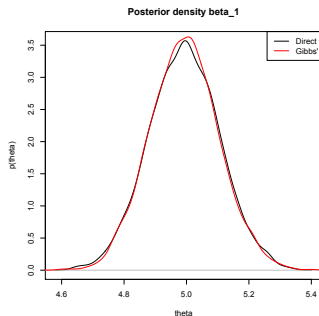
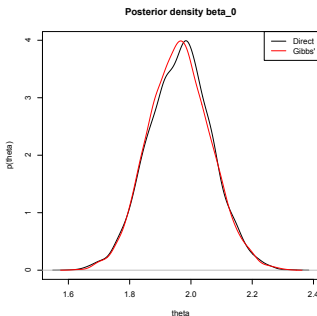


## Example

- $Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_i, 1)$  for  $i = 1, \dots, n$  where  $\beta_0 = 2$ ,  $\beta_1 = 5$  (both unknown) and  $n = 100$
- We assume independent priors  $\beta_0 \sim N(0, \gamma_0)$  and  $\beta_1 \sim N(0, \gamma_1)$  where  $\gamma_0 = 100$  and  $\gamma_1 = 10$
- Gibbs sampling (Gelfand and Smith, 1990):
  - $\beta_0 | \beta_1, Y \sim N(1'(Y - \beta_1 X)/(n + 1/\gamma_0), 1/(n + 1/\gamma_0))$
  - $\beta_1 | \beta_0, Y \sim N(X'(Y - \beta_0 1)/(\sum_{i=1}^n X_i^2 + 1/\gamma_1), 1/(\sum_{i=1}^n X_i^2 + 1/\gamma_1))$
- Direct approach:  $(\beta_0, \beta_1 | Y) \sim N(V_\beta(1, X)'Y, V_\beta)$  where  $V_\beta = ((1, X)'(1, X) + \text{diag}(1/\gamma_0, 1/\gamma_1))^{-1}$

## Example

- $Y_i \stackrel{\text{ind}}{\sim} N(\beta_0 + \beta_1 X_i, 1)$  for  $i = 1, \dots, n$  where  $\beta_0 = 2$ ,  $\beta_1 = 5$  (both unknown) and  $n = 100$
- We assume independent priors  $\beta_0 \sim N(0, \gamma_0)$  and  $\beta_1 \sim N(0, \gamma_1)$  where  $\gamma_0 = 100$  and  $\gamma_1 = 10$

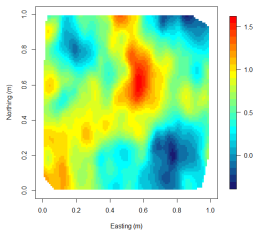


# Block Gibbs update

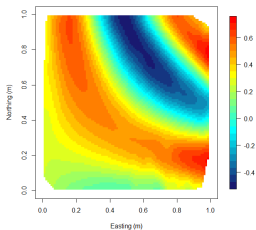
- Recall for the spatial model we had full conditionals for vectors  $\beta$  and  $w$
- Let  $\theta = (\theta_1, \theta_2, \dots, \theta_k) = (\eta'_1, \eta'_2, \dots, \eta'_m)$  where  $\eta_j$  are blocks of  $\theta_i$ 's
- One can use the Gibbs updates for the blocks  $\eta_j$ 's instead of using the individual updates for  $\theta_i$
- In many models, the block full conditionals are easier to obtain, substantially reduces computation and improves rate of convergence

# Data analysis

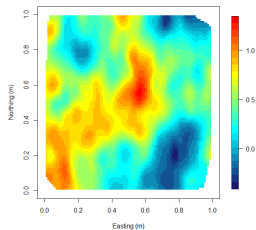
- Dataset 3 from Lecture 1
- True model:  $y(s) \sim N(0.2 - 0.3x(s) + w(s), 0.01)$ ,  
 $w(s) \sim GP, Cov(w(s_i), w(s_j)) = 0.25 * exp(-2||s_i - s_j||)$



(a)  $y(s)$



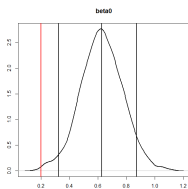
(b)  $x(s)$



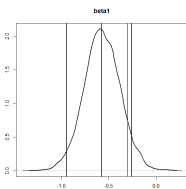
(c)  $w(s)$

## Parameter posteriors

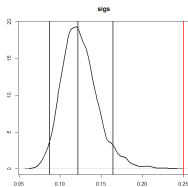
- $\phi$  is kept fixed at 4.23 (estimated value from variogram fitting)
- Gibbs sampler for  $w$ ,  $\beta$ ,  $\sigma^2$  and  $\tau^2$



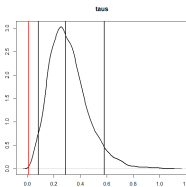
(a)  $\beta_0$



(b)  $\beta_1$



(c)  $\sigma^2$

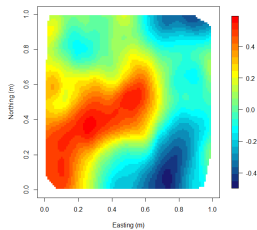


(d)  $\tau^2$

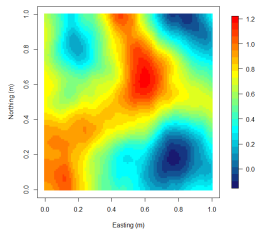
# Posterior predictive distributions

- For the unmarginalized model, posterior samples for  $w(s)$  are already generated for all  $s$  in the training data locations  $S$
- Posterior predictive distributions  $\tilde{y}(s)$  can be obtained using composition sampling:
  - If  $s_0 \notin S$ , generate samples from  $w(s_0) | y$  using  $w(s_0)^{(j)} | \cdot \sim N(c(s_0)C^{-1}w^{(j)}, \sigma^{2(j)}(1 - c(s_0)C^{-1}c(s_0))/\sigma^2))$
  - $c(s_0) = \text{cov}(w(s_0), w)$  and  $C = \text{var}(w)$
  - If  $\phi$  was also sampled, replace  $c$  and  $C$  by  $c^{(j)}$  and  $C^{(j)}$
  - For any  $s$ , generate  $\tilde{y}(s)^{(j)} = N(x(s)'\beta^{(j)}, \tau^{2(j)})$

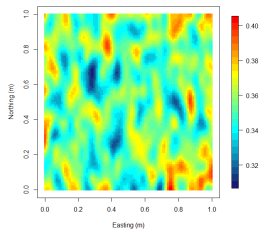
# Posterior surfaces



(a)  $w(s) | y$



(b)  $\tilde{y}(s) | y$



(c)  $\text{var}(\tilde{y}(s) | y)$

# Marginalized model

- Unmarginalized model has  $n$  additional parameters ( $w$ )
- May lead to slow MCMC convergence
- Marginalized model:  $y \sim N(X\beta, \sigma^2 R(\phi) + \tau^2 I)$
- **Pros:** Only  $p + 3$  parameters
- **Cons:** Even the full conditionals are not useful (except for  $\beta$ )
- How to do MCMC?



# Metropolis algorithm

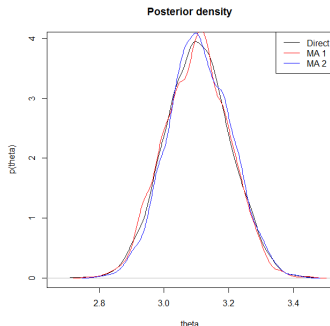
- We want to draw sample from a density  $p(\theta) = f(\theta)/K$
- Begin with an initial  $\theta^0$
- Choose a function  $q(x | y)$  such that
  - $q(x | y)$  is a valid density function in  $x$  for every value of  $y$
  - $q(x | y) = q(y | x)$
  - e.g.  $q(x | y) \sim N(x | y, \lambda) = \frac{1}{\sqrt{2\pi\lambda}} \exp(-\frac{1}{2\lambda}(x - y)^2)$
  - If  $\theta$  is multivariate one can choose  $q(x | y) \sim N(x | y, \Sigma)$
- $q$  is called the proposal density
- If  $\theta$  is multivariate, choose  $q$  to be a multivariate proposal density

# Metropolis algorithm

- At the  $i^{th}$  iteration, generate  $\theta^*$  from  $q(\cdot | \theta_{i-1})$
- Calculate the ratio  $r = f(\theta^*)/f(\theta_{i-1})$
- If  $r \geq 1$ , accept the new value i.e  $\theta_i = \theta^*$
- If  $r < 1$ :
  - Accept the new value i.e  $\theta_i = \theta^*$  with probability  $r$
  - Keep the old value i.e  $\theta_i = \theta_{i-1}$  with probability  $1 - r$
- The sample  $(\theta_i)_{i=N_b}^N$  is a sample from  $p(\theta)$  where  $N_b$  is a burn-in period used
- An overall rate of acceptance around 30% – 50% is desirable

## Example

- $Y_i \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2)$  for  $i = 1, \dots, n$  where  $\theta = 3$  (unknown),  $\sigma^2 = 1$  (known) and  $n = 100$
- Prior:  $\theta \sim N(\mu, \tau^2)$  where  $\mu = 0$  and  $\tau^2 = 10$
- Metropolis algorithm:  
$$p(\theta | Y) \propto \exp\left(-\frac{n}{2\sigma^2}(\bar{y} - \theta)^2 - \frac{1}{2\tau^2}(\theta - \mu)^2\right)$$
- Direct approach:  $\theta | Y \sim N\left(\frac{\frac{n\bar{y}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\right)$

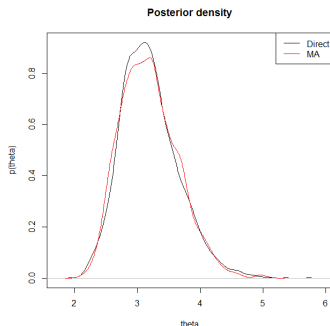


# Jacobian adjustment

- Often the parameter of interest  $\theta$  is not supported on the entire real line but on a part of it e.g.  $[0, 1]$ ,  $(0, \infty)$  etc.
- The normal proposal density is easy to use but has the entire real line as support
- One can choose a transformation  $g$  such that  $\eta = g(\theta)$  is supported on the real line
- Generate new  $\eta^*$  using the normal proposal density
- Use the inverse transformation to obtain  $\theta^* = g^{-1}(\eta^*)$
- The likelihood for  $\eta$  will be given by  $p(\eta) = p(\theta)/|g'(\theta)|$
- Calculate
$$r = p(\eta^*)/p(\eta_{i-1}) = p(\theta^*)/p(\theta_{i-1}) \times |g'(\theta_{i-1})|/|g'(\theta^*)|$$

## Example

- $Y_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$  for  $i = 1, \dots, n$  where  $\sigma^2 = 4$  (unknown)
- $\sigma^2$  is supported on  $(0, \infty)$ . So, we use log transformation
- Prior:  $\sigma^2 \sim \text{IG}(\alpha, \beta)$  where  $\alpha = 2$  and  $\beta = 1$
- Metropolis algorithm:  
 $p(\sigma^2 | Y) \propto (\sigma^2)^{-1-\alpha-n/2} \exp(-(\beta + \sum_{i=1}^n y_i^2/2)/\sigma^2)$
- Direct approach:  
 $\sigma^2 | Y \sim \text{Inverse Gamma}(\alpha + n/2, \beta + \sum_{i=1}^n y_i^2/2)$



# Metropolis-Hastings Algorithm

- Allows for asymmetric proposal densities
- We want to draw sample from a density  $p(\theta) = f(\theta)/K$
- Let  $q(x | y)$  denote the proposal density
- Calculate the ratio  $r = \frac{f(\theta^*)q(\theta_{i-1} | \theta^*)}{f(\theta_{i-1})q(\theta^* | \theta_{i-1})}$
- Useful if  $f$  is asymmetric
- Reduces to Metropolis algorithm if  $q$  is symmetric

## Metropolis within Gibbs

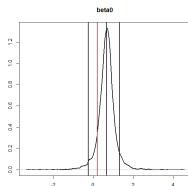
- For the marginalized model, doing MH for the entire vector  $(\beta', \sigma^2, \tau^2, \phi)'$  may be slow if  $p$  is large
- Also,  $\beta$  has nice normal full conditionals
- One can use a **Metropolis Random Walk (RW) step** for the univariate full conditional target densities inside the Gibbs sampler
- Example: MCMC steps for the marginalized model:
  - (a) Gibbs for  $\beta$ :  $\beta^{(j)} \sim N((X'X)^{-1}X'y, \tau^{2(j-1)}(X'X)^{-1})$
  - (b) RW for  $\phi$  from target density  $N(y | X\beta^{(j)}, \sigma^{2(j-1)}R(\phi) + \tau^{2(j-1)}I) \times p(\phi)$
  - (c) RW for  $\sigma^2$  from  $N(y | X\beta^{(j)}, \sigma^2R(\phi^{(j)}) + \tau^{2(j-1)}I) \times p(\sigma^2)$
  - (d) RW for  $\tau^2$  from target density  $N(y | X\beta^{(j)}, \sigma^{2(j)}R(\phi^{(j)}) + \tau^2I) \times p(\phi)$

## *Nimble* package

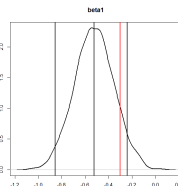
- <https://r-nimble.org/>
- Implements the MCMC for you
- You only need to specify the model and initialize the MCMC !
- We run the MCMC for the marginalized model for dataset 3 in Nimble



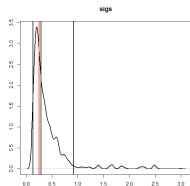
# Parameter posteriors



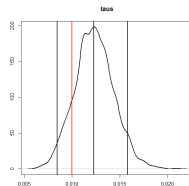
(a)  $\beta_0$



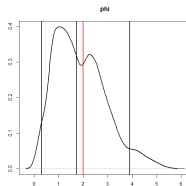
(b)  $\beta_1$



(c)  $\sigma^2$



(d)  $\tau^2$



(e)  $\phi$

# Recovering $w$

- The marginalized model integrates out the  $w$ 's
- We can recover them after the MCMC
- $w | y, \beta, \sigma^2, \tau^2, \phi \sim N(V_w(y - X\beta)/\tau^2, V_w)$  where  $V_w = (I/\tau^2 + R(\phi)^{-1}/\sigma^2)^{-1}$
- Use composition sampling

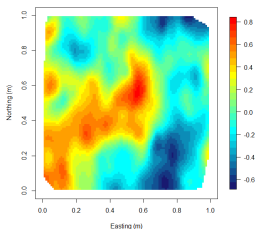


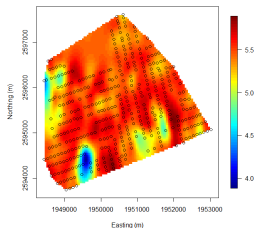
Figure:  $w(\mathbf{s}) | y$

# Predictions for the marginalized model

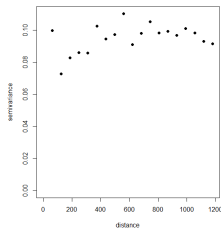
- Two ways to do predict  $\tilde{y}(s) | y$  using composition sampling
- If you have already recovered  $w$ 
  - Similar to the unmarginalized model
  - Generate  $w(s_0) | w, params$  and then  $\tilde{y}(s_0) | w(s_0), params$
- Direct approach (not requiring samples of  $w$ ):
  - $c(s_0) = cov(w(s_0), w)$  and  $\Sigma = \sigma^2 R(\phi) + \tau^2 I$
  - Generate samples of  $\tilde{y}(s_0) | y, params \sim N(x(s_0)' \beta + c(s_0)' \Sigma^{-1} (y - X \beta), \sigma^2 + \tau^2 - c(s_0)' \Sigma^{-1} c(s_0))$

# BEF data analysis using *spBayes* package

- <https://cran.r-project.org/web/packages/spBayes/spBayes.pdf>
- Runs MCMC for a variety of spatial linear models



(a) log biomass



(b) Variogram

# MCMC free Bayesian inference

- The marginalized model can be reparametrized as:  
 $N(y, X\beta, \sigma^2(R(\phi) + \alpha I))$  where  $\alpha = \tau^2/\sigma^2$
- If  $\phi$  and  $\alpha$  is fixed, we can do exact conjugate sampling
- *bayesGeostatExact* does that
- Fixed values of  $\phi$  and  $\alpha$  can be chosen from the variogram

	2.5%	25%	50%	75%	97.5%
Intercept	-0.351	0.647	1.205	1.773	2.711
ELEV	0.000	0.000	0.000	0.001	0.001
SLOPE	-0.015	-0.011	-0.009	-0.006	-0.002
Brightness	-0.002	0.006	0.010	0.014	0.023
Greenness	-0.002	0.003	0.006	0.009	0.014
Wetness	0.009	0.017	0.021	0.025	0.032
$\sigma^2$	0.072	0.078	0.082	0.086	0.094
$\tau^2$	0.014	0.016	0.016	0.017	0.019

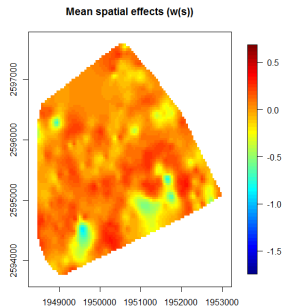
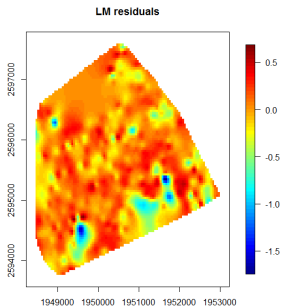
# Full Bayesian inference

- *spLM* function

	2.5%	25%	50%	75%	97.5%
(Intercept)	-0.037	0.762	1.301	1.786	3.042
ELEV	0.000	0.000	0.000	0.001	0.001
SLOPE	-0.016	-0.011	-0.008	-0.005	-0.001
Brightness	-0.001	0.007	0.011	0.016	0.025
Greenness	-0.004	0.002	0.005	0.008	0.012
Wetness	0.009	0.017	0.021	0.025	0.032
$\sigma^2$	0.041	0.052	0.068	0.085	0.104
$\tau^2$	0.005	0.018	0.037	0.050	0.065
$\phi$	0.003	0.007	0.009	0.011	0.015

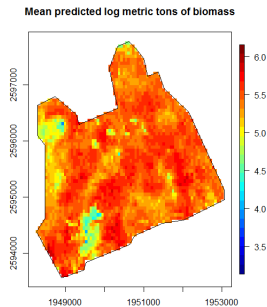
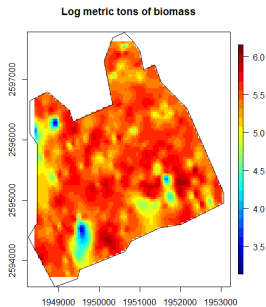
# Recovery of $w(s)$

- *spRecover* function



# Kriging

- *spPredict* function





# What we covered today

- Gibbs sampler
- MH algorithm
- Writing your own MCMC
- Using Nimble package to run the MCMC
- Using spBayes package to run the MCMC

# References

- BCG book chapters 5.3.1, 5.3.2 and 6.3.1
- **Expository article on Gibbs sampler:** Casella, G. and George, E.I. (1992), Explaining the Gibbs Sampler, *The American Statistician*, 46, 167-174.
- **Expository article on MH algorithm:** Chib, S. and Greenberg, E. (1995), Understanding the Metropolis-Hastings Algorithm, *The American Statistician*, 49, 327-335.
- Great slides on convergence diagnostics of Markov Chains [http://www.stat.missouri.edu/~dsun/8640/convergence\\_print.pdf](http://www.stat.missouri.edu/~dsun/8640/convergence_print.pdf)
- Gelfand, A., and Adrian F. M. Smith. (1990). *Sampling-Based Approaches to Calculating Marginal Densities*. *Journal of the American Statistical Association*, 85(410), 398-409.
- 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(1), 27-40.