Bayesian inference for spatial GP models

Abhi Datta

February 16, 2018

Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland

Review of last lecture

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

Bayesian inference for spatial linear model

- $y(s) = x(s)'\beta + w(s) + \epsilon(s)$, $w(s) \sim GP(0, C(\cdot, \cdot \,|\, \phi))$, $\epsilon \overset{\text{iid}}{\sim} N(0, \tau^2)$
- 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 ϕ 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: $N(y \mid X\beta + w, \tau^2 I) \times N(w \mid 0, \sigma^2 R(\phi) \times N(\beta \mid \mu, V) \times IG(\sigma^2 \mid a_\sigma, b_\sigma) \times IG(\tau^2 \mid a_\tau, b_\tau)$

Unmarginalized model

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

Unmarginalized model

- Likelihood: $N(y \mid X\beta + w, \tau^2 I) \times N(w \mid 0, \sigma^2 R(\phi) \times N(\beta \mid \mu, V) \times IG(\sigma^2 \mid a_\sigma, b_\sigma) \times IG(\tau^2 \mid a_\tau, b_\tau)$
- Observe that
  - $\beta \mid \sigma^2, \tau^2, w, y \sim N(\mu^*, V^*)$
  - $w \mid \sigma^2, \tau^2, \beta, y \sim N(m, C^*)$
  - $\sigma^2 \mid \beta, \tau^2, w, y \sim IG(a_\sigma^*, b_\sigma^*)$
  - $\tau^2 \mid \beta, \sigma^2, w, y \sim IG(a_\tau^*, b_\tau^*)$
- 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 $p(\theta_1, \theta_2 \mid y)$.
- For many interesting hierarchical models, we have access to *full conditional distributions* $p(\theta_1 \mid \theta_2, y)$ and $p(\theta_2 \mid \theta_1, y)$.
- The *Gibbs sampler* proposes the following sampling scheme. Set starting values $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})$ For $j = 1, \ldots, M$
    - Draw $\theta_1^{(j)} \sim p(\theta_1 \mid \theta_2^{(j-1)}, y)$
    - Draw $\theta_2^{(j)} \sim p(\theta_2 \mid \theta_1^{(j)}, y)$
- 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 $p(\theta_1, \theta_2 \mid y)$, where M_0 is the burn-in period..

Gibbs sampling

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

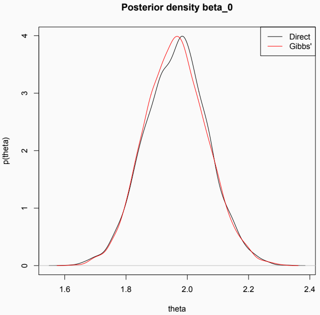
$$\theta_1^{(j)} \sim p(\theta_1^{(j)} \mid \theta_2^{(j-1)}, \ldots, \theta_p^{(j-1)}, y)$$
$$\theta_2^{(j)} \sim p(\theta_2 \mid \theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_p^{(j-1)}, y)$$
$$\ldots$$
$$\text{(the generic } k^{th} \text{ element)}$$
$$\theta_k^{(j)} \sim p(\theta_k | \theta_1^{(j)}, \ldots, \theta_{k-1}^{(j)}, \theta_{k+1}^{(j-1)}, \ldots, \theta_p^{(j-1)}, y)$$
$$\ldots$$
$$\theta_p^{(j)} \sim p(\theta_p \mid \theta_1^{(j)}, \ldots, \theta_{p-1}^{(j)}, y)$$

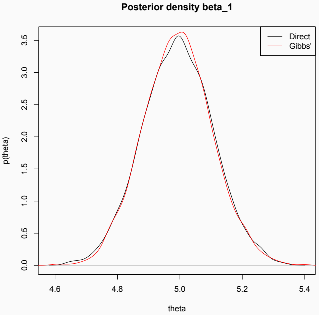
Example

- $Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, 1)$ for $i = 1, \ldots, n$ 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):
  - $\beta_0 \mid \beta_1, Y \sim N(1'(Y - \beta_1 X)/(n + 1/\gamma_0), 1/(n + 1/\gamma_0))$
  - $\beta_1 \mid \beta_0, Y \sim N(X'(Y - \beta_0 1)/(\sum_{i=1}^n X_i^2 + 1/\gamma_1), 1/(\sum_{i=1}^n X_i^2 + 1/\gamma_1))$
- Direct approach: $(\beta_0, \beta_1 \mid Y) \sim N\left(V_\beta (1, X)' Y, V_\beta\right)$ where $V_\beta = \left((1, X)'(1, X) + diag(1/\gamma_0, 1/\gamma_1)\right)^{-1}$

Example

- $Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, 1)$ for $i = 1, \ldots, n$ 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 $\gamma_1 = 10$



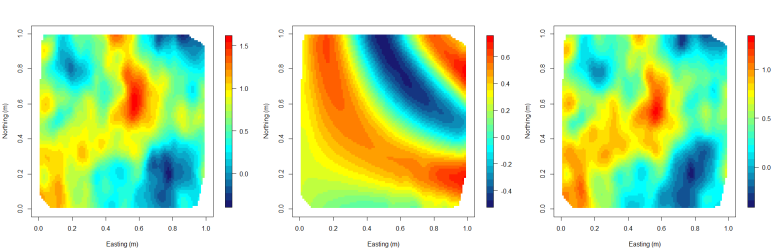


Block Gibbs update

- Recall for the spatial model we had full conditionals for vectors β and w
- Let $\theta = (\theta_1, \theta_2, \ldots, \theta_k) = (\eta_1', \eta_2', \ldots, \eta_m')$ where η_j are blocks of θ_i's
- One can use the Gibbs updates for the blocks η_j's instead of using the individual updates for θ_i
- In many models, the block full conditionals are easier to obtain, substantially reduces computation and improves rate of convergence

Data analysis

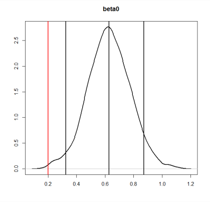
- Dataset 3 from Lecture 1
- True model: $y(s) \sim N(0.2 - 0.3x(s) + w(s), 0.01)$,
  $w(s) \sim GP,\ Cov(w(s_i), w(s_j)) = 0.25 * exp(-2||s_i - s_j||)$



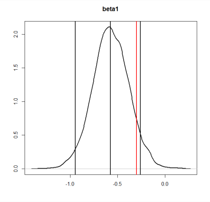
$y(s)$ $x(s)$ $w(s)$

Parameter posteriors

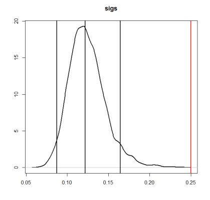
- ϕ is kept fixed at 4.23 (estimated value from variogram fitting)
- Gibbs sampler for w, β, σ^2 and τ^2



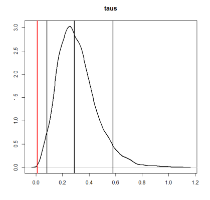
β_0



β_1



σ^2

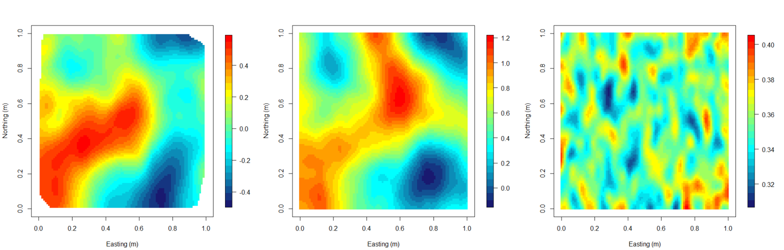


τ^2

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 $w(s_0)^{(j)} \,|\, \cdot \sim N(c(s_0)' C^{-1} w^{(j)}, \sigma^{2(j)}(1 - r(s_0)' R^{-1} r(s_0)))$
  - $c(s_0) = cor(w(s_0), w)$ and $R = cor(w)$
  - If ϕ 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



$w(s) \mid y$ $\quad\tilde{y}(s) \mid y$ $\quad var(\tilde{y}(s) \mid y)$

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 β)
- How to do MCMC?

Metropolis algorithm

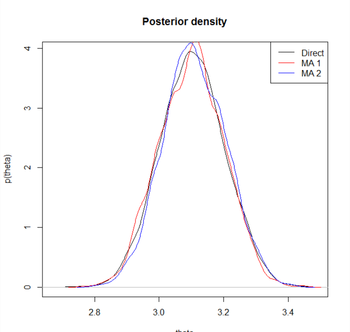
- We want to draw sample from a density $p(\theta) = f(\theta)/K$
- Begin with an initial θ^0
- Choose a function $q(x \mid y)$ such that
  - $q(x \mid y)$ is a valid density function in x for every value of y
  - $q(x \mid y) = q(y \mid x)$
  - e.g. $q(x \mid y) \sim N(x \mid y, \lambda) = \frac{1}{\sqrt{2\pi\lambda}}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

Metropolis algorithm

- At the i^{th} iteration, generate θ^* from $q(\cdot \mid \theta_{i-1})$
- 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 (controlled by the tuning parameter λ)

Example

- $Y_i \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$ for $i = 1, \ldots, n$ 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:
  $p(\theta \mid Y) \propto \exp(-\frac{n}{2\sigma^2}\left(\bar{y} - \theta)^2 - \frac{1}{2\tau^2}(\theta - \mu)^2\right)$
- Direct approach: $\theta \mid Y \sim N\left(\frac{\frac{n\bar{y}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\right)$

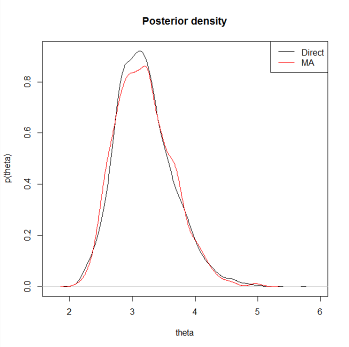


Jacobian adjustment

- Often the parameter of interest θ 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 η^* using the normal proposal density
- Use the inverse transformation to obtain $\theta^* = g^{-1}(\eta^*)$
- The likelihood for η 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

- $Y_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ for $i = 1, \dots, n$ where $\sigma^2 = 4$ (unknown)
- σ^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:
  $p(\sigma^2 \mid Y) \propto (\sigma^2)^{-1-\alpha-n/2} \exp(-(\beta + \sum_{i=1}^n y_i^2/2)/\sigma^2)$
- Direct approach:
  $\sigma^2 \mid Y \sim \text{Inverse Gamma}(\alpha + n/2, \beta + \sum_{i=1}^n y_i^2/2)$



Metropolis-Hastings Algorithm

- Allows for asymmetric proposal densities
- We want to draw sample from a density $p(\theta) = f(\theta)/K$
- Let $q(x \mid y)$ denote the proposal density
- Calculate the ratio $r = \frac{f(\theta^*)q(\theta_{i-1} \mid \theta^*)}{f(\theta_{i-1})q(\theta^* \mid \theta_{i-1})}$
- Useful if f is asymmetric
- Reduces to Metropolis algorithm is 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, β 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 β: $\beta^{(j)} \sim N((X'X)^{-1}X'y, \tau^{2(j-1)}(X'X)^{-1})$

  (b) RW for ϕ from target density $N(y \,|\, X\beta^{(j)}, \sigma^{2(j-1)}R(\phi) + \tau^{2(j-1)}I) \times p(\phi)$

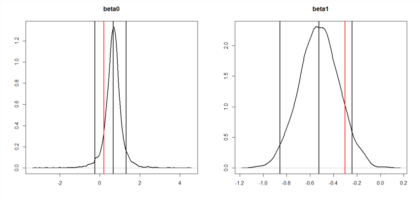
  (c) RW for σ^2 from $N(y \,|\, X\beta^{(j)}, \sigma^2 R(\phi^{(j)}) + \tau^{2(j-1)}I) \times p(\sigma^2)$

  (d) RW for τ^2 from target density $N(y \,|\, X\beta^{(j)}, \sigma^{2(j)}R(\phi^{(j)}) + \tau^2 I) \times p(\phi)$

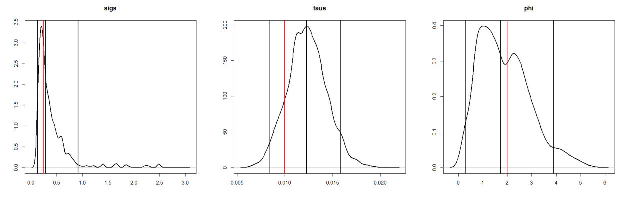
Nimble package

- `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



β_0 β_1



σ^2 τ^2 ϕ

Recovering w

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

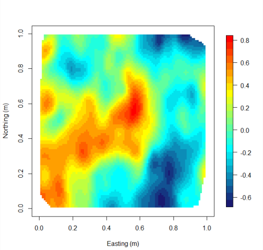


Figure: $w(\mathbf{s}) \mid y$

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 $w(s_0) \,|\, w, \mathit{params}$ and then $\tilde{y}(s_0) \,|\, w(s_0), \mathit{params}$
- 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 $\tilde{y}(s_0) \,|\, y, \mathit{params} \sim$
    $N(x(s_0)'\beta + c(s_0)'\Sigma^{-1}(y - X\beta), \sigma^2 + \tau^2 - c(s_0)'\Sigma^{-1}c(s_0))$

What we covered today

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

References

- 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`
- 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.