Bayesian linear model

Abhi Datta

04.21.2017

Review of last lecture

- Data analysis using geoR package
    - Residual plots
    - Empirical variograms
    - MLE or variogram model fitting
    - Uncertainty quantified kriging
- Map projections

Why do we need Bayesian models for spatial data

- The classical MLE based approach is limited in scope.
- For example, uncertainty quantification for the covariance parameters is tricky
  - Need to leverage asymptotic results
  - Increasing and fixed domain asymptotics for irregular sptial data
  - Parameters often not identifiable (Zhang 2006)
- The Bayesian approach expands the class of models and easily handles:
  - repeated measures or multiple data sources
  - unbalanced or missing data
  - spatial misalignment and change of support
  - varying coefficient models

  – and many other settings that are precluded (or much more complicated) in classical settings.

Basics of Bayesian inference

- We start with a model (likelihood) $f(y \mid \theta)$ for the observed data $y = (y_1, \ldots, y_n)'$ given unknown parameters θ (perhaps a collection of several parameters).

- Add a prior distribution $p(\theta \mid \lambda)$, where λ is a vector of hyper-parameters.

Basics of Bayesian inference

- We start with a model (likelihood) $f(y \mid \theta)$ for the observed data $y = (y_1, \ldots, y_n)'$ given unknown parameters θ (perhaps a collection of several parameters).

- Add a prior distribution $p(\theta \mid \lambda)$, where λ is a vector of hyper-parameters.

- If λ are known/fixed, then the posterior distribution of θ is given by:

$$p(\theta \mid y, \lambda) = \frac{p(\theta \mid \lambda) \times f(y \mid \theta)}{p(y \mid \lambda)} = \frac{p(\theta \mid \lambda) \times f(y \mid \theta)}{\int f(y \mid \theta) p(\theta \mid \lambda) d\theta}.$$

  We refer to this formula as *Bayes Theorem*.

Basics of Bayesian inference

- If λ are unknown, we assign a prior, $p(\lambda)$, and seek:

$$p(\theta, \lambda \,|\, y) \propto p(\lambda)p(\theta \,|\, \lambda)f(y \,|\, \theta)/p(y).$$

  The proportionality constant does not depend upon θ or λ:

$$p(y) = \int p(\lambda)p(\theta \,|\, \lambda)f(y \,|\, \theta)d\lambda d\theta$$

- The above represents a *joint* posterior from a *hierarchical model*. The *marginal* posterior distribution for θ is:

$$p(\theta \,|\, y) \propto \int p(\lambda)p(\theta \,|\, \lambda)f(y \,|\, \theta)d\lambda.$$

Basic of Bayesian inference

- Point estimation: simply choose an appropriate distribution summary: posterior mean, median or mode.
- Bayesian *credible sets:*. A $100(1-\alpha)\%$ credible set C for θ satisfies

$$P(\theta \in C \mid y) = \int_C p(\theta \mid y) d\theta \geq 1 - \alpha.$$

- The interval between the $\frac{\alpha}{2}^{th}$ and $(1-\frac{\alpha}{2})^{th}$ quantiles of $p(\theta \mid y)$ is a $100(1-\alpha)\%$ Bayesian *credible interval*.
- Often direct calculation of quantiles, modes and means are not straightforward.

Sampling-based inference:

- Approximate the posterior distribution $p(\theta \mid y)$ by drawing samples $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)}\}$ from it.
- $p(\theta \mid y) \approx \frac{1}{M} \sum_{i=1}^{M} I(\theta = \theta^{(i)})$
- Numerical integration can be replaced by "Monte Carlo integration".

$$E_{\theta \mid y}(g(\theta)) \approx \frac{1}{M} \sum_{i=1}^{M} g(\theta^{(i)})$$

- Sample quantiles approximate posterior quantiles

A simple example: Normal data and normal priors

- Example: Say $y = (y_1, \ldots, y_n)'$, where $y_i \overset{iid}{\sim} N(0, \sigma^2)$; assume σ is *known*.
- $\theta \sim N(\mu, \tau^2)$, i.e. $p(\theta) = N(\theta \,|\, \mu, \tau^2)$; μ, τ^2 are known.
- Posterior distribution of θ

A simple example: Normal data and normal priors

- Example: Say $y = (y_1, \ldots, y_n)'$, where $y_i \stackrel{iid}{\sim} N(0, \sigma^2)$; assume σ is *known*.
- $\theta \sim N(\mu, \tau^2)$, i.e. $p(\theta) = N(\theta \,|\, \mu, \tau^2)$; μ, τ^2 are known.
- Posterior distribution of θ

$$
\begin{aligned}
p(\theta|y) &\propto N(\theta \,|\, \mu, \tau^2) \times \prod_{i=1}^{n} N(y_i \,|\, \theta, \sigma^2) \\
&= N\left(\theta \,|\, \frac{\sigma^2}{\sigma^2 + n\tau^2}\mu + \frac{n\tau^2}{\sigma^2 + n\tau^2}\bar{y}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)
\end{aligned}
$$

Bayesian Linear Model

- $y_i \overset{\text{iid}}{\sim} N(x_i'\beta, \sigma^2)$,
- Assume prior $\beta \sim N(\mu, V)$
- $p(\beta \,|\, \sigma^2, y) \propto N(y \,|\, X\beta, \sigma^2 I) \times N(\beta \,|\, \mu, V)$

Bayesian Linear Model

- $y_i \stackrel{\text{iid}}{\sim} N(x_i'\beta, \sigma^2)$,
- Assume prior $\beta \sim N(\mu, V)$
- $p(\beta \,|\, \sigma^2, y) \propto N(y \,|\, X\beta, \sigma^2 I) \times N(\beta \,|\, \mu, V)$
- $\beta \sim N((X'X/\sigma^2 + V^{-1})^{-1}X'y/\sigma^2, (X'X/\sigma^2 + V^{-1})^{-1})$

Super useful result:
$p(\beta) \propto \prod_{i=1}^n \exp\left(-\frac{1}{2}(y_i - X_i\beta)'Q_i(y_i - X_i\beta)\right) \Rightarrow$
$\beta \sim N(B^{-1}b, B^{-1})$ where $B = \sum_{i=1}^n X_i'Q_iX_i$ and
$b = \sum_{i=1}^n X_i'Q_iy_i$

Bayesian Linear Model

- $\beta \sim N((X'X/\sigma^2 + V^{-1})^{-1}X'y/\sigma^2, (X'X/\sigma^2 + V^{-1})^{-1})$
- If $V^{-1} = 0$, then
  $p(\beta \,|\, \sigma^2, y) = N(\beta \,|\, (X^TX)^{-1}X^Ty,\, \sigma^2(X^TX)^{-1})$.
- $V^{-1} = 0$ corresponds to $p(\beta) \propto 1$ which is not a valid density as $\int 1 = \infty$. So why is it that we are even discussing them?
- If the priors are *improper* (that's what we call them), as long as the resulting posterior distributions are valid we can still conduct legitimate statistical inference on them.

Marginal and conditional distributions

- $\beta \,|\, \sigma^2, y \sim N((X'X/\sigma^2 + V^{-1})^{-1}X'y/\sigma^2, (X'X/\sigma^2 + V^{-1})^{-1})$
- $p(\beta \,|\, \sigma^2, y)$ would have been the desired posterior distribution had σ^2 been known.
- If σ^2 is unknown, $p(\beta \,|\, \sigma^2, y)$ is called the the conditional posterior distribution of β.
- The *marginal posterior* distribution by integrating out σ^2 is:

$$p(\beta \mid y) = \int p(\beta \mid \sigma^2, y) p(\sigma^2 \mid y) d\sigma^2$$

- Can we bypass the integration and still do inference on $\theta \,|\, y$?

Composition Sampling

- Suppose $\theta = (\theta_1, \theta_2)$ and we know how to sample from the *marginal posterior distribution* $p(\theta_2|y)$ and the *conditional distribution* $P(\theta_1 \,|\, \theta_2, y)$.

- Goals: Draw samples from the marginal posterior $p(\theta_1 \,|\, y)$ and from the joint distribution: $p(\theta_1, \theta_2 \,|\, y)$

Composition Sampling

- Suppose $\theta = (\theta_1, \theta_2)$ and we know how to sample from the *marginal posterior distribution* $p(\theta_2|y)$ and the *conditional distribution* $P(\theta_1 \mid \theta_2, y)$.
- Goals: Draw samples from the marginal posterior $p(\theta_1 \mid y)$ and from the joint distribution: $p(\theta_1, \theta_2 \mid y)$
- We do this in two stages using *composition sampling*:
  - First draw $\theta_2^{(j)} \sim p(\theta_2 \mid y),\ j = 1, \ldots M$.
  - Next draw $\theta_1^{(j)} \sim p\left(\theta_1 \mid \theta_2^{(j)}, y\right)$.

Composition Sampling

- *Composition sampling*:
    - First draw $\theta_2^{(j)} \sim p(\theta_2 \mid y),\ j = 1, \ldots M$.
    - Next draw $\theta_1^{(j)} \sim p\left(\theta_1 \mid \theta_2^{(j)}, y\right)$.
- This sampling scheme produces *exact* samples, $\{\theta_1^{(j)}, \theta_2^{(j)}\}_{j=1}^M$ from the posterior distribution $p(\theta_1, \theta_2 \mid y)$.
- Gelfand and Smith (*JASA*, 1990) demonstrated *automatic marginalization*: $\{\theta_1^{(j)}\}_{j=1}^M$ are samples from $p(\theta_1 \mid y)$ and (of course!) $\{\theta_2^{(j)}\}_{j=1}^M$ are samples from $p(\theta_2 \mid y)$.
- In effect, composition sampling has performed the following "integration":

$$p(\theta_1 \mid y) = \int p(\theta_1 \mid \theta_2, y) p(\theta_2 \mid y) d\theta.$$

Composition Sampling for Bayesian Linear Model

- $y_i \overset{\text{iid}}{\sim} N(x_i'\beta, \sigma^2)$, $p(\beta) \propto 1$
- Assume an Inverse Gamma ($IG(a, b)$) prior for σ^2, i.e.,

$$p(\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{a+1} exp(-b/\sigma^2)$$

- Marginal posterior distribution of σ^2 is:

$$p(\sigma^2 \,|\, y) = IG\left(\sigma^2 \,|\, a + \frac{n-p}{2}, b + \frac{(n-p)s^2}{2}\right),$$

where $s^2 = \hat{\sigma}^2 = \frac{1}{n-p} y^T (I - P_X) y$.

- If $a = b = 0$, i.e., $p(\sigma^2) \propto 1/\sigma^2$, then $\sigma^2 \,|\, y \sim IG(\sigma^2 \,|\, (n-p)/2, (n-p)s^2/2)$ and $E(\sigma^2 \,|\, y) = \hat{\sigma^2}$.
Striking similarity with the classical result!

Composition sampling for Bayesian Linear Model

- Now we are ready to carry out composition sampling from $p(\beta, \sigma^2 \mid y)$ as follows:
  - Draw M samples from $p(\sigma^2 \mid y)$:
  $$\sigma^{2(j)} \sim IG\left(\frac{n-p}{2}, \frac{(n-p)s^2}{2}(n-p)\right), \; j = 1, \ldots M$$
  - For $j = 1, \ldots, M$, draw from $p(\beta \mid \sigma^{2(j)}, y)$:
  $$\beta^{(j)} \sim N\left((X^T X)^{-1} X^T y, \, \sigma^{2(j)} (X^T X)^{-1}\right)$$
- The resulting samples $\{\beta^{(j)}, \sigma^{2(j)}\}_{j=1}^{M}$ represent M samples from $p(\beta, \sigma^2 \mid y)$.
- $\{\beta^{(j)}\}_{j=1}^{M}$ are samples from the marginal posterior distribution $p(\beta \mid y)$. This is a *multivariate t* density:

$$p(\beta \mid y) = \frac{\Gamma(n/2)}{(\pi(n-p))^{p/2}\Gamma((n-p)/2)|s^2(X^T X)^{-1}|}\left[1 + \frac{(\beta - \hat{\beta})^T (X^T X)(\beta - \hat{\beta})}{(n-p)s^2}\right]^{-n/2}.$$

Bayesian predictions

- Suppose we want to predict new observations, say $\tilde{y}$, based upon the observed data y. We will specify a *joint* probability model $p(\tilde{y}, y \,|\, , \theta)$, which defines the *conditional predictive distribution*:

$$p(\tilde{y} \,|\, y, \theta) = \frac{p(\tilde{y}, y \,|\, , \theta)}{p(y \,|\, \theta)}.$$

- Bayesian predictions follow from the *posterior predictive* distribution that averages out the θ from the conditional predictive distribution with respect to the posterior:

$$p(\tilde{y} \,|\, y) = \int p(\tilde{y} \,|\, y, \theta) p(\theta \,|\, y) d\theta.$$

- This can be evaluated using composition sampling:
  - First obtain: $\theta^{(j)} \sim p(\theta \,|\, y)$, $j = 1, \ldots M$
  - For $j = 1, \ldots, M$ sample $\tilde{y}^{(j)} \sim p(\tilde{y} \,|\, y, \theta^{(j)})$
- The $\{\tilde{y}^{(j)}\}_{j=1}^{M}$ are samples from the posterior predictive distribution $p(\tilde{y} \,|\, y)$.

Bayesian predictions from the linear model

- Suppose we have observed the new predictors $\tilde{X}$, and we wish to predict the outcome $\tilde{y}$. We specify $p(\tilde{y}, y \mid \theta)$ to be a normal distribution:

$$\begin{pmatrix} y \\ \tilde{y} \end{pmatrix} \sim N\left(\begin{bmatrix} X \\ \tilde{X} \end{bmatrix} \beta, \sigma^2 I \right)$$

- Note $p(\tilde{y} \mid y, \beta, \sigma^2) = p(\tilde{y} \mid \beta, \sigma^2) = N(\tilde{y} \mid \tilde{X}\beta, \sigma^2 I)$.

- The *posterior predictive* distribution:

$$\begin{aligned} p(\tilde{y} \mid y) &= \int p(\tilde{y} \mid y, \beta, \sigma^2) p(\beta, \sigma^2 \mid y) d\beta d\sigma^2 \\ &= \int p(\tilde{y} \mid \beta, \sigma^2) p(\beta, \sigma^2 \mid y) d\beta d\sigma^2. \end{aligned}$$

- By now we are comfortable evaluating such integrals:
  - First obtain: $(\beta^{(j)}, \sigma^{2(j)}) \sim p(\beta, \sigma^2 \mid y)$, $j = 1, \ldots, M$
  - Next draw: $\tilde{y}^{(j)} \sim N(\tilde{X}\beta^{(j)}, \sigma^{2(j)} I)$.

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, we have $y = N(X\beta + w, \tau^2 I)$, $w \sim N(0, C(\phi))$
- Assuming stationarity, $C(\phi) = \sigma^2 R(\phi)$ where $R(\phi)$ is the correlation matrix
- Marginalised model: $y \sim N(X\beta, \sigma^2 R + \tau^2 R(\phi))$
- Even if we assume ϕ is known and σ^2 and *taus* are given Inverse Gamma priors, composition sampling does not help here
- How to do inference on the Bayesian parameters?

References

- BCG book chapters 5.1, 5.2.1 and 5.2.2
- Berger, J., Bernardo, J., and Sun, D (2009) *The Formal Definition of Reference Priors.* The Annals of Statistics, 37(2), 905–938.
- 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.