Introduction to Bayesian Linear Model

Abhi Datta

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

abhidatta.com @dattascience

Why do we need Bayesian models for spatial data

- Uncertainty quantification for the covariance parameters using MLE is tricky
  - Need to leverage asymptotic results
  - Increasing and fixed domain asymptotics for irregular spatial data
  - Parameters often not identifiable (Zhang 2006)
- The Bayesian approach expands the class of models and easily handles:
  - binary or count outcomes
  - unbalanced or missing data
  - spatial misalignment and change of support
  - second-stage models
  - varying coefficient models
  - repeated measures or multiple data sources

– and many other settings where estimation and prediction can be complicated in the classical framework.

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 (known) 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 (known) hyper-parameters.

- The posterior distribution of θ is given by:

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

  as the proportionality constant $p(y)$ does not depend upon θ. We refer to this formula as Bayes Theorem.

A simple example: Normal data and normal priors

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

A simple example: Normal data and normal priors

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

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

A simple example: Normal data and normal priors

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

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

- When $\tau^2 \to \infty$ or $n \to \infty$, $\theta \mid y \sim N(\bar{y}, \sigma^2/n)$, i.e., same as the classical result

Improper priors

- In the previous example, $\theta \mid y \sim N(\bar{y}, \sigma^2/n)$ when $\tau^2 = \infty$
- However, $\tau^2 = \infty \Rightarrow p(\theta) \propto 1$ 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.

Basic of Bayesian inference

- Point estimation: simply choose an appropriate distribution summary: posterior mean, median or mode.
- Interval estimation: A $100(1-\alpha)\%$ Bayesian 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" to get posterior means of any functional $g(\theta)$ of θ:

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

- Sample quantiles approximate posterior quantiles. Easy calculation of credible intervals.

Bayesian Linear Model

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

Bayesian Linear Model

- $y_i \overset{\text{iid}}{\sim} N(x_i'\beta, \sigma^2)$,
- Assume prior $\beta \sim N(\mu, V)$ and σ^2 to be known
- $p(\beta \mid \sigma^2, y) \propto N(y \mid X\beta, \sigma^2 I) \times N(\beta \mid \mu, V)$
- $\beta \mid y \sim N((X'X/\sigma^2 + V^{-1})^{-1} X'y/\sigma^2, (X'X/\sigma^2 + V^{-1})^{-1})$

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 $B = \sum_{i=1}^n X_i' Q_i X_i$ and
$b = \sum_{i=1}^n X_i' Q_i y_i$

Bayesian Linear Model

- $\beta \mid y \sim N((X'X/\sigma^2 + V^{-1})^{-1}X'y/\sigma^2, (X'X/\sigma^2 + V^{-1})^{-1})$
- If $V^{-1} = 0$, then
  $p(\beta \mid y) = N(\beta \mid (X^T X)^{-1} X^T y,\, \sigma^2 (X^T X)^{-1})$.
- Note the striking similarity to the MLE and its variance !!
- $V^{-1} = 0$ corresponds to $p(\beta) \propto 1$ (another example of an improper prior)

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 $R := R(\phi)$ is the correlation matrix
- Marginalized model: $y \sim N(X\beta, \sigma^2 R + \tau^2 I)$
- Letting $\theta = (\beta, \sigma^2, \tau^2, \phi)$ and $p(\theta)$ the prior, we have $p(\theta \,|\, y) \propto$

$$\frac{1}{\sqrt{|\sigma^2 R + \tau^2 I|}} \exp\left(-\frac{1}{2}(y - X\beta)'(\sigma^2 R + \tau^2 I)^{-1}(y - X\beta)\right) \times p(\theta).$$

- We will use `rstan` to sample from this non-standard posterior

Sampling using Stan

- Subset of Dataset 3 from previous lectures

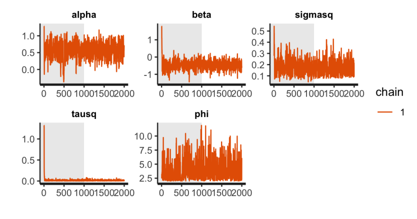


Figure: Trace plots

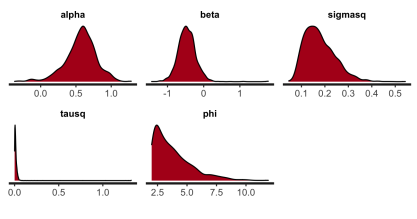
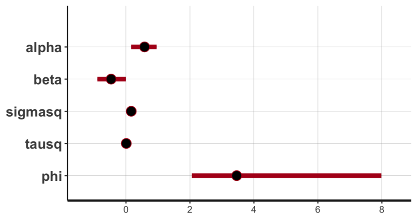


Figure: Posterior densities

Sampling using Stan

- Posterior estimates:



- Comparison with MLE:

	mean	2.5%	97.5%	mle
alpha	0.5690	0.1607	0.9589	0.6024
beta	-0.4559	-0.8949	-0.0029	-0.4787
sigmasq	0.1748	0.0867	0.3208	0.1429
tausq	0.0126	0.0002	0.0423	0.0000
phi	3.8635	2.0593	7.9799	4.5996

Review of lecture

- Basics of Bayesian inference – priors, posteriors, sampling, posterior (point and interval) estimates
- Example: Bayesian linear model
- Bayesian spatial GP model analysis using `rstan`