STAT 431 — Applied Bayesian Analysis — Course Notes

Introduction to Computational Methods Part 1

Spring 2019

Notation:

$$\begin{aligned}
\boldsymbol{\theta} &= (\theta_1, \ldots \theta_p) = \text{parameters} \\
\boldsymbol{y} &= \text{data} \\
p(\boldsymbol{\theta}) &= \text{prior density} \\
p(\boldsymbol{\theta} \mid \boldsymbol{y}) &= \text{posterior density} \\
p(\theta_i \mid \boldsymbol{y}) &= \text{posterior marginal density for } \theta_i
\end{aligned}$$

Notice: When $\boldsymbol{\theta}$ has a continuous posterior distribution, most Bayesian inference tasks involve integration —

- Computing a normalizing constant $p(\boldsymbol{y})$:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \;=\; \frac{p(\boldsymbol{\theta})\, p(\boldsymbol{y} \mid \boldsymbol{\theta})}{p(\boldsymbol{y})}$$

  so

$$p(\boldsymbol{y}) \;=\; \int p(\boldsymbol{\theta})\, p(\boldsymbol{y} \mid \boldsymbol{\theta})\, d\boldsymbol{\theta}$$

- Computing a posterior expectation:

  For some function g, might want

$$\mathrm{E}\big(g(\boldsymbol{\theta}) \mid \boldsymbol{y}\big) \;=\; \int g(\boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta}$$

  (perhaps a posterior mean or variance of θ_i)

- Computing a posterior marginal density:

$$p(\theta_i \mid \boldsymbol{y}) \;=\; \int_{\substack{\text{all } \theta_j \\ j \neq i}} p(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta}_{(-i)}$$

  where $\boldsymbol{\theta}_{(-i)}$ is $\boldsymbol{\theta}$ with θ_i removed.

- Computing a posterior probability:

  For $H_0 : \boldsymbol{\theta} \in \Theta_0$,

$$\mathrm{P}(H_0 \mid \boldsymbol{y}) \;=\; \int_{\Theta_0} p(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta}$$

Other things that might involve integration (directly or indirectly) include finding a posterior quantile or obtaining (and working with) a posterior predictive distribution.

Numeric Integration

Goal: Approximate

$$\int_D f(\boldsymbol{x})\, d\boldsymbol{x}$$

Idea: Partition D into N regions $D_1, \ldots D_N$, with representative points

$$\boldsymbol{x}_1 \in D_1, \quad \ldots \quad \boldsymbol{x}_N \in D_N$$

and use

$$\sum_{j=1}^{N} f(\boldsymbol{x}_j) \cdot \mathrm{area}(D_j) \qquad \text{where} \quad \mathrm{area}(D_j) \ = \int_{D_j} d\boldsymbol{x}$$

[Draw 3-D example ...]

For example, the midpoint rule (in one dimension):

$$\int_a^b f(x)\,dx \quad \approx \quad \sum_{j=1}^{N} f(x_j) \cdot \frac{b-a}{N}$$

$$\text{where} \quad x_j \;=\; a \,+\, (b-a)\,\frac{j-\frac{1}{2}}{N}$$

[Draw example ...]

Note: Accuracy requires N large and f (somewhat) continuous and smooth.

R function `integrate()` uses an adaptive algorithm for one-dimensional integration.

Example: Proportion of people like us with pets

Recall: $y = 12$ out of $n = 70$

$$Y \mid \pi \sim \text{binomial}(n, \pi)$$

$$L(\pi; y) \propto \pi^{12}(1-\pi)^{58}$$

Let's use the Jeffreys prior

$$\pi \sim \text{beta}(1/2,\ 1/2)$$

Using conjugacy, the posterior is

$$\pi \mid y \sim \text{beta}(12.5,\ 58.5)$$

so

$$\mathrm{E}(\pi \mid y) = \frac{12.5}{12.5 + 58.5} = 0.17606$$

Let's pretend we don't know the posterior, and we want to approximate

$$
\begin{aligned}
\mathrm{E}(\pi \mid y) &= \int \pi \cdot p(\pi \mid y)\, d\pi \\
&= \int \pi \, \frac{p(\pi)\, L(\pi; y)}{p(y)}\, d\pi \\
&= \frac{\displaystyle\int \pi\, p(\pi)\, L(\pi; y)\, d\pi}{\displaystyle\int p(\pi)\, L(\pi; y)\, d\pi}
\end{aligned}
$$

So we need to approximate two integrals ...

R Example 8.1:

Population Proportion: Numeric Integration

Example: Are Bike Owners Less Likely to Ride the Bus?

Data (from class survey):

- Among $n_1 = 19$ bike owners, $y_1 = 8$ ride the bus
- Among $n_2 = 51$ non-bike owners, $y_2 = 25$ ride the bus

$$\begin{aligned}
\pi_1 &= \text{population proportion of owners who ride} \\
\pi_2 &= \text{population proportion of non-owners who ride} \\
\boldsymbol{y} &= (y_1, y_2)
\end{aligned}$$

Want

$$\mathrm{P}(\pi_1 < \pi_2 \mid \boldsymbol{y})$$

Likelihood:

$$\begin{aligned} L(\pi_1, \pi_2; \boldsymbol{y}) &= p(\boldsymbol{y} \mid \pi_1, \pi_2) \\ &= p(y_1 \mid \pi_1)\, p(y_2 \mid \pi_2) \\ &\propto \pi_1^{y_1}(1-\pi_1)^{n_1-y_1}\, \pi_2^{y_2}(1-\pi_2)^{n_2-y_2} \end{aligned}$$

We'll use a product-Jeffreys prior:

$$\pi_1, \pi_2 \sim \text{indep. } \mathrm{beta}(1/2, 1/2)$$

Need to compute

$$
\begin{aligned}
\mathrm{P}(\pi_1 < \pi_2 \mid \boldsymbol{y}) &= \int_0^1 \int_0^{\pi_2} p(\pi_1, \pi_2 \mid \boldsymbol{y})\, d\pi_1\, d\pi_2 \\
&= \frac{\displaystyle\int_0^1 \int_0^{\pi_2} p(\pi_1, \pi_2) L(\pi_1, \pi_2; \boldsymbol{y})\, d\pi_1\, d\pi_2}{\displaystyle\int_0^1 \int_0^1 p(\pi_1, \pi_2) L(\pi_1, \pi_2; \boldsymbol{y})\, d\pi_1\, d\pi_2}
\end{aligned}
$$

[Draw integration region ...]

R Example 8.2:

Comparing Population Proportions: Numeric Integration

Numeric integration works well for low-dimensional problems (few parameters).

For high-dimensional problems, simulation is often better ...

Any method using randomized simulation (sampling) for approximation is called a **Monte Carlo** method.

Independent Sampling

Idea: Randomly generate samples from the posterior.

Then use the **empirical distribution** of the samples to estimate aspects of the actual posterior.

Let data be $\boldsymbol{y}$, and let the sample from the posterior for parameter θ be

$$\theta^{(1)}, \; \cdots \; \theta^{(N)}$$

A posterior sample may be used to approximate many things —

- a mean:

$$\mathrm{E}(\theta \mid \boldsymbol{y}) \approx \frac{1}{N}\sum_{k=1}^{N}\theta^{(k)} = \text{sample mean of } \theta^{(k)}\text{s}$$

- a variance:

$$\mathrm{Var}(\theta \mid \boldsymbol{y}) \approx \text{sample variance of } \theta^{(k)}\text{s}$$

- a (lower) quantile q_p for probability p:

  [Draw quantile ...]

  Round pN to the nearest integer: $[pN]$

  Then use the $[pN]$th order statistic from the sample.

  (The order statistics are the $\theta^{(k)}$s re-ordered from least to greatest.)

- a 95% equal-tailed credible interval: form an estimated version of

$$(q_{0.025},\ q_{0.975})$$

- probability of $H_0 : \theta \in \Theta_0$

$$\text{fraction of } \theta^{(k)}\text{s in } \Theta_0 \quad = \quad \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}\big(\theta^{(k)} \in \Theta_0\big)$$

where $\mathbb{1}$ represents the indicator function:

$$\mathbb{1}(\theta \in \Theta_0) \quad = \quad \begin{cases} 1 & \theta \in \Theta_0 \\ 0 & \theta \notin \Theta_0 \end{cases}$$

- a mean of a function of θ:

$$\mathrm{E}(g(\theta) \mid \boldsymbol{y}) \quad \approx \quad \frac{1}{N} \sum_{k=1}^{N} g\big(\theta^{(k)}\big)$$

This is based on the fact that

$$g\big(\theta^{(1)}\big), \quad \ldots \quad g\big(\theta^{(N)}\big)$$

is a random sample from the posterior of $g(\theta)$.

Since these approximations are subject to random sampling variability, we need an assessment of their accuracy.

For example, the approximation

$$\frac{1}{N}\sum_{k=1}^{N} g\big(\theta^{(k)}\big) \qquad \text{of} \qquad \mathrm{E}(g(\theta) \mid \boldsymbol{y})$$

has an approximate *standard error* of

$$\frac{s_g}{\sqrt{N}} \qquad \text{where} \qquad s_g \;=\; \sqrt{\text{sample var. of } g\big(\theta^{(k)}\big)\text{s}}$$

(why?)

This is the **Monte Carlo error** for the mean approximation.

Example: Jevons's Coins Comparison

For $n_1 = 24$ coins minted before 1830:

$$\bar{y}_1 \ = \ 7.8730 \qquad s_1 \ = \ 0.05353$$

For $n_2 = 123$ newer coins (1860's):

$$\bar{y}_2 \ = \ 7.9725 \qquad s_2 \ = \ 0.01409$$

Assume independent samples from $\mathrm{N}(\mu_1, \sigma_1^2)$ and $\mathrm{N}(\mu_2, \sigma_2^2)$.

We want inference about $\mu_1 - \mu_2$.

Take "independent" "standard" (product-Jeffreys) priors:

$$\mu_1,\, \sigma_1^2,\, \mu_2,\, \sigma_2^2 \quad \sim \quad \frac{1}{\sigma_1^2} \cdot \frac{1}{\sigma_2^2}\, d\mu_1\, d\sigma_1^2\, d\mu_2\, d\sigma_2^2$$

(Let $\boldsymbol{y}$ represent both samples together.)

Then you can show that

$$(\mu_1, \sigma_1^2) \quad \text{and} \quad (\mu_2, \sigma_2^2) \quad \text{are posterior-independent}$$

with posterior distributions $(i = 1, 2)$

$$\mu_i \mid \sigma_i^2, \boldsymbol{y} \;\; \sim \;\; \mathrm{N}(\bar{y}_i, \sigma_i^2/n_i)$$

$$\sigma_i^2 \mid \boldsymbol{y} \;\; \sim \;\; \mathrm{IG}\left(\frac{n_i - 1}{2},\, \frac{n_i - 1}{2} s_i^2\right)$$

So we know that, under the posterior, μ_1 and μ_2 have independent t distributions. (Why?)

But this means that $\mu_1 - \mu_2$ has no simple-form posterior density.

We can easily randomly sample from the posterior distribution of $\mu_1 - \mu_2$ using R ...

R Example 8.3:

Comparing Normal Means: Independent Sampling

The 95% *Welch interval* is an approximate frequentist confidence interval for $\mu_1 - \mu_2$, used here for comparison:

$$\bar{y}_1 - \bar{y}_2 \quad \pm \quad t_{0.025,\,\mathrm{df}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where

$$\mathrm{df} \;=\; \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) \,+\, (s_2^2/n_2)^2/(n_2 - 1)}$$

Sampling: Concepts, Notation, Facts

Suppose we want a (joint) sample of

$$\boldsymbol{\theta} \;=\; (\theta_1, \ldots, \theta_p) \quad \sim \quad \text{some joint distribution } \mathcal{D}$$

where the joint distribution has a density

$$p(\theta_1, \ldots, \theta_p)$$

We will say

$$\boldsymbol{\theta}^{(1)},\ \boldsymbol{\theta}^{(2)},\ \ldots\ \boldsymbol{\theta}^{(N)}$$

is a **sample** from $\mathcal{D}$ if

$$\boldsymbol{\theta}^{(k)} \;\sim\; \mathcal{D} \qquad \text{for each } k$$

Notice: No need for independence — it could be a **dependent sample**.

Notation:

$$\boldsymbol{\theta}^{(k)} \;=\; \left(\theta_1^{(k)},\ \ldots\ \theta_p^{(k)}\right)$$

Fact:

$$\theta_j^{(1)},\ \ldots\ \theta_j^{(N)}$$

is a sample from the marginal distribution of θ_j (under $\mathcal{D}$)

Notation:

$$\boldsymbol{\theta}_{(-j)} \;=\; \boldsymbol{\theta} \text{ without } \theta_j$$

Then

$$p(\theta_j \mid \boldsymbol{\theta}_{(-j)})$$

is called the **full conditional density** for θ_j (corresponding to its **full conditional distribution**).

Fact: If $\boldsymbol{\theta} \sim \mathcal{D}$ and we sample

$$\tilde{\theta}_1 \quad \text{from} \quad p(\,\cdot \mid \boldsymbol{\theta}_{(-1)})$$

(i.e. $\tilde{\theta}_1$ is sampled from the full conditional of θ_1), then

$$\left(\tilde{\theta}_1,\, \boldsymbol{\theta}_{(-1)}\right) \ \sim \ \mathcal{D}$$

Fact: If $\boldsymbol{\theta} \sim \mathcal{D}$ and we sample

$$\tilde{\theta}_1 \quad \text{from} \quad p(\,\cdot \mid \boldsymbol{\theta}_{(-1)})$$

(i.e. $\tilde{\theta}_1$ is sampled from the full conditional of θ_1), then

$$\left(\tilde{\theta}_1,\, \boldsymbol{\theta}_{(-1)}\right) \;\sim\; \mathcal{D}$$

Similarly for any element of $\boldsymbol{\theta}$: If we sample

$$\tilde{\theta}_j \quad \text{from} \quad p(\,\cdot \mid \boldsymbol{\theta}_{(-j)})$$

then

$$\left(\theta_1,\, \ldots\, \theta_{j-1},\, \tilde{\theta}_j,\, \theta_{j+1},\, \ldots\, \theta_p\right) \;\sim\; \mathcal{D}$$

Basic Gibbs Sampling

Based on the full conditionals ...

For simplicity, suppose the model has two parameters:

$$\boldsymbol{\theta} \;=\; (\theta_1, \theta_2)$$

Given data $\boldsymbol{y}$, the (joint) posterior is

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \;=\; p(\theta_1, \theta_2 \mid \boldsymbol{y})$$

for which the full conditionals are

$$p(\theta_1 \mid \theta_2, \boldsymbol{y}) \quad \text{and} \quad p(\theta_2 \mid \theta_1, \boldsymbol{y})$$

Idea: Alternate between sampling from the full conditional for θ_1 and the full conditional for θ_2 (once each time), updating each value after sampling.

[Diagram ...]

Result: A sequence of **iterates**

$$\underbrace{\theta_1^{(1)},\ \ \theta_2^{(1)}}_{\boldsymbol{\theta}^{(1)}},\ \ \underbrace{\theta_1^{(2)},\ \ \theta_2^{(2)}}_{\boldsymbol{\theta}^{(2)}},\ \ \underbrace{\theta_1^{(3)},\ \ \theta_2^{(3)}}_{\boldsymbol{\theta}^{(3)}},\ \ \ldots$$

(The initial value $\theta_1^{(1)}$ may be chosen deterministically or at random — ideally it shouldn't matter.)

Note: These samples will generally be *dependent* because each is sampled based on the previous one.

Concern: Why would sampling from the full conditionals necessarily be any easier than direct sampling from the posterior?

Answer: You can often make the full conditionals easy to sample by using semi-conjugacy to choose the prior.

Notice:

$$p(\theta_1 \mid \theta_2, \boldsymbol{y}) \quad = \quad \frac{p(\theta_1, \theta_2 \mid \boldsymbol{y})}{p(\theta_2 \mid \boldsymbol{y})} \quad \underset{\text{in } \theta_1}{\propto} \quad p(\theta_1, \theta_2 \mid \boldsymbol{y})$$

So the joint posterior is actually a kernel of the full conditional for θ_1. (Similarly for θ_2.)

Perhaps this kernel is from a known family (such as when the prior is semi-conjugate).

Algorithm:

1. Choose initial value $\theta_1^{(1)}$ and sample $\theta_2^{(1)}$ from $p\big(\theta_2 \mid \theta_1^{(1)}, \boldsymbol{y}\big)$

2. For $k = 2$ to N,

    2.1 Sample $\theta_1^{(k)}$ from $p\big(\theta_1 \mid \theta_2^{(k-1)}, \boldsymbol{y}\big)$

    2.2 Sample $\theta_2^{(k)}$ from $p\big(\theta_2 \mid \theta_1^{(k)}, \boldsymbol{y}\big)$

3. Use iterates $\boldsymbol{\theta}^{(1)}, \ldots \boldsymbol{\theta}^{(N)}$ for inference.

The iterates form a “path of samples”:

[Illustrate path ...]

Fact: If $\theta_1^{(1)}$ is drawn from its posterior marginal $p(\theta_1 \mid \boldsymbol{y})$, then the sequence of iterates

$$\boldsymbol{\theta}^{(1)},\ \boldsymbol{\theta}^{(2)},\ \ldots\ \boldsymbol{\theta}^{(N)}$$

is a (dependent) sample from the posterior.

Principle: Under certain conditions, regardless of the value of $\theta_1^{(1)}$, the iterates will converge in distribution to the posterior.

So, "eventually" (for k large enough)

$$\boldsymbol{\theta}^{(k)},\ \ldots\ \boldsymbol{\theta}^{(N)}$$

will be *approximately* a (dependent) sample from the posterior.

Issues (addressed later):

- Where to start? (choosing $\theta_1^{(1)}$)
- How long until "close enough" to posterior?
- How many samples needed?
- How to detect problems?

Example: Normal Sample, *Semi*-Conjugate Prior

$$\underbrace{Y_1, \ldots Y_n}_{\boldsymbol{Y}} \mid \mu, \sigma^2 \quad \sim \quad \text{i.i.d. } \mathrm{N}(\mu, \sigma^2)$$

$$\left.\begin{aligned} \mu &\sim \mathrm{N}(\mu_0, \sigma_0^2) \\ \sigma^2 &\sim \mathrm{IG}(\alpha, \beta) \end{aligned}\right\} \text{independent}$$

Recall: NOT conjugate — no easy direct way to sample from the posterior.

But this prior *is* semi-conjugate ...

Getting the full conditional for μ is just like treating σ^2 as known — recall, under a $\mathrm{N}(\mu_0, \sigma_0^2)$ prior,

$$\mu \mid \sigma^2, \boldsymbol{y} \;\; \sim \;\; \mathrm{N}(\mu_1, 1/\tau_1^2)$$

where

$$\mu_1 \;\; = \;\; \frac{\tau_0^2 \mu_0 \; + \; n\tau^2 \bar{y}}{\tau_0^2 \; + \; n\tau^2} \qquad\qquad \tau_1^2 \;\; = \;\; \tau_0^2 \; + \; n\tau^2$$

with

$$\tau_0^2 \;\; = \;\; 1/\sigma_0^2 \qquad\qquad \tau^2 \;\; = \;\; 1/\sigma^2$$

Getting the full conditional for σ^2 is just like treating μ as known — recall, under an $\mathrm{IG}(\alpha, \beta)$ prior,

$$\sigma^2 \mid \mu, \boldsymbol{y} \;\; \sim \;\; \mathrm{IG}(\alpha + n/2,\; \beta + n\nu/2)$$

where

$$\begin{aligned} \nu &= \frac{1}{n} \sum_i (y_i - \mu)^2 \\ &= \frac{n-1}{n} s^2 \; + \; (\bar{y} - \mu)^2 \end{aligned}$$

The Gibbs sampler just alternates between sampling from these full conditionals ...

We illustrate with Jevons's coin data ...

Recall the (fully) conjugate prior:

$$\begin{aligned} \mu \mid \sigma^2 &\sim \mathrm{N}(\mu_0, \sigma^2/\kappa) \\ \sigma^2 &\sim \mathrm{IG}(\alpha, \beta) \end{aligned}$$

You can show

$$\mathrm{E}(\mu) \;=\; \mu_0 \qquad\qquad \mathrm{Var}(\mu) \;=\; \frac{\beta/(\alpha-1)}{\kappa}$$

We will choose a *semi*-conjugate prior that matches the means and variances of the conjugate prior used previously (Example 7.1).

R Example 8.4:

Gibbs Sampler for Semi-Conjugate Prior (Normal Sample)

Generalize: Gibbs Sampler for p Parameters

[Diagram of sampling ...]

Difficult situations for Gibbs sampling:

- Parameters have high posterior correlation

- Posterior has multiple modes (offset from each other)

Markov Chain Monte Carlo (MCMC)

A sequence of random variables

$$X_0,\ X_1,\ X_2,\ \ldots$$

is a **Markov chain (MC)** if, for each $t \geq 2$, X_t is conditionally independent of

$$X_0,\ \ldots\ X_{t-2}$$

given X_{t-1}.

That is,

$$p(x_t \mid x_{t-1}, \ldots x_0) \;\; = \;\; p(x_t \mid x_{t-1}).$$

X_t is the **state** of the MC at time t.

The **transition kernel** is the conditional density

$$p(x_t \mid x_{t-1})$$

which determines how X_t can be generated based on X_{t-1}.

The kernel is **time-invariant** if it does not depend on t.
(Similarly for the MC.)

More generally, MCs may be sequences of random *vectors* ...

A Gibbs sampler is a time-invariant Markov chain:

- $\boldsymbol{\theta}^{(k)}$ is generated using only $\boldsymbol{\theta}^{(k-1)}$
- the distributions used in the generation of $\boldsymbol{\theta}^{(k)}$ do not depend on k (except through the value of $\boldsymbol{\theta}^{(k-1)}$)

Under certain conditions, states of a time-invariant Markov chain converge in distribution to a unique distribution $\mathcal{D}$ as $t \to \infty$:

$$X_t \underset{t \to \infty}{\Longrightarrow} \mathcal{D}$$

for (almost) any X_0.

(The "certain conditions" are technical and often difficult to check.)

Under certain conditions, states of a time-invariant Markov chain converge in distribution to a unique distribution $\mathcal{D}$ as $t \to \infty$:

$$X_t \underset{t\to\infty}{\Longrightarrow} \mathcal{D}$$

for (almost) any X_0.

(The "certain conditions" are technical and often difficult to check.)

For a Gibbs sampler, the distribution $\mathcal{D}$ is the posterior:

$$\boldsymbol{\theta}^{(k)} \underset{k\to\infty}{\Longrightarrow} \text{the posterior}$$

for (almost) any choice of $\boldsymbol{\theta}^{(1)}$.

Practical approach to running a Gibbs sampler:

(1) Choose several different **initial values** ($\boldsymbol{\theta}^{(1)}$s).

(Better if they are far apart.)

(2) For each initial value, run a separate **chain** for N **iterations**.

(3) **Monitor** the chains for:

- whether they seem to be converging to the same distribution
- how many iterations until convergence

Increase N if necessary.

(4) If converged, declare the first B iterates

$$\boldsymbol{\theta}^{(1)}, \ \ldots \ \boldsymbol{\theta}^{(B)}$$

of each chain to be a **burn-in** period.

Ignore the burn-in iterates, and use the rest for inference.

(5) Estimate the Monte Carlo error in your inferences.

Run more iterations until it is sufficiently small.

Note: Some samplers also need an initial period of **adaptation** to find a good sampling scheme when semi-conjugacy does not hold. Only the iterates after both burn-in *and* adaptation should be used for inference.