Bayesian Statistics in One Hour

Patrick Lam

Outline

Introduction

Bayesian Models

Applications

- Missing Data
- Hierarchical Models

Outline

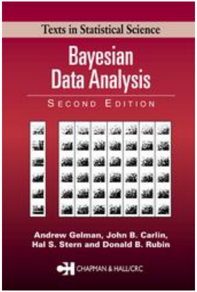
Introduction

Bayesian Models

Applications

- Missing Data
- Hierarchical Models

References



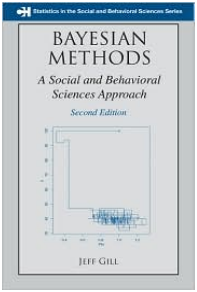
Texts in Statistical Science

Bayesian Data Analysis

SECOND EDITION

Andrew Gelman, John B. Carlin, Hal S. Stern and Donald B. Rubin

CHAPMAN & HALL/CRC



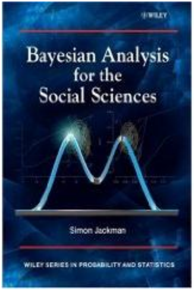
Statistics in the Social and Behavioral Sciences Series

BAYESIAN METHODS

A Social and Behavioral Sciences Approach

Second Edition

JEFF GILL



WILEY

Bayesian Analysis for the Social Sciences

Simon Jackman

WILEY SERIES IN PROBABILITY AND STATISTICS

Western, Bruce and Simon Jackman. 1994. "Bayesian Inference for Comparative Research." *American Political Science Review* 88(2): 412-423.

Jackman, Simon. 2000. "Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo." *American Journal of Political Science* 44(2): 375-404.

Jackman, Simon. 2000. "Estimation and Inference Are Missing Data Problems: Unifying Social Science Statistics via Bayesian Simulation." *Political Analysis* 8(4): 307-332.

Introduction

Three general approaches to statistics:

- frequentist (Neyman-Pearson, hypothesis testing)
- likelihood (what we've been learning all semester)
- Bayesian

Today's goal: Contrast {frequentist, likelihood} with Bayesian, with emphasis on Bayesian versus likelihood.

We'll go over some of the Bayesian critiques of non-Bayesian analysis and *non-Bayesian critiques of Bayesian analysis*.

Probability

Objective view of probability (non-Bayesian):

- The relative frequency of an outcome of an experiment over repeated runs of the experiment.
- The observed proportion in a population.

Subjective view of probability (Bayesian):

- Individual's degree of belief in a statement
- Defined personally (how much money would you wager on an outcome?)
- Can be influenced in many ways (personal beliefs, prior evidence)

Bayesian statistics is convenient because it does not require repeated sampling or large *n* assumptions.

Maximum Likelihood

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
&= p(y|\theta)k(y) \\
&\propto p(y|\theta)
\end{aligned}
$$

$$
L(\theta|y) = p(y|\theta)
$$

There is a fixed, true value of θ, and we maximize the likelihood to estimate θ and make assumptions to generate uncertainty about our estimate of θ.

Bayesian

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &\propto p(y|\theta)p(\theta) \end{aligned}$$

- θ is a random variable.
    - θ is stochastic and changes from time to time.
    - θ is truly fixed, but we want to reflect our uncertainty about it.
- We have a prior subjective belief about θ, which we update with the data to form posterior beliefs about θ.
- The posterior is a probability distribution that must integrate to 1.
- The prior is usually a probability distribution that integrates to 1 (proper prior).

θ as Fixed versus as a Random Variable

Non-Bayesian approach (θ fixed):

- Estimate θ with measures of uncertainty (SE, CIs)
- 95% Confidence Interval: 95% of the time, θ is in the 95% interval that is estimated each time.
  - $\text{P}(\theta \in 95\% \text{ CI}) =$ 0 or 1
- $\text{P}(\theta > 2) =$ 0 or 1

Bayesian approach (θ random):

- Find the posterior distribution of θ.
- Take quantities of interest from the distribution (posterior mean, posterior SD, posterior credible intervals)
- We can make probability statements regarding θ.
  - 95% Credible Interval: $\text{P}(\theta \in 95\% \text{ CI}) = 0.95$
  - $\text{P}(\theta > 2) = (0, 1)$

Critiques

$$\text{Posterior} \quad = \quad \text{Evidence} \times \text{Prior}$$

NB: *Bayesians introduce priors that are not justifiable.*
B: Non-Bayesians are just doing Bayesian statistics with uninformative priors, which may be equally unjustifiable.

NB: *Unjustified Bayesian priors are driving the results.*
B: Bayesian results $\approx$ non-Bayesian results as n gets larger (the data overwhelm the prior).

NB: *Bayesian is too hard. Why use it?*
B: Bayesian methods allow us to easily estimate models that are too hard to estimate (cannot computationally find the MLE) or unidentified (no unique MLE exists) with non-Bayesian methods. Bayesian methods also allow us to incorporate prior/qualitative information into the model.

Outline

Introduction

Bayesian Models

Applications

- Missing Data
- Hierarchical Models

Running a Model

Non-Bayesian:

1. Specify a probability model (distribution for Y).
2. Find MLE $\hat{\theta}$ and measures of uncertainty (SE, CI). Assume $\hat{\theta}$ follows a (multivariate) normal distribution.
3. Estimate quantities of interest analytically or via simulation.

Bayesian:

1. Specify a probability model (distribution for Y and priors on θ).
2. Solve for posterior and summarize it (mean, SD, credible interval, etc.). We can do both analytically or via simulation.
3. Estimate quantities of interest analytically or via simulation.

There is a Bayesian way to do any non-Bayesian parametric model.

A Simple (Beta-Binomial) Model

The Los Angeles Lakers play 82 games during a regular NBA season. In the 2008-2009 season, they won 65 games. Suppose the Lakers win each game with probability π. Estimate π.

We have 82 Bernoulli observations or one observation Y, where

$$Y \sim \mathrm{Binomial}(n, \pi)$$

with $n = 82$.

Assumptions:

- Each game is a Bernoulli trial.
- The Lakers have the same probability of winning each game.
- The outcomes of the games are independent.

We can use the beta distribution as a prior for π since it has support over [0,1].

$$
\begin{aligned}
p(\pi|y) &\propto p(y|\pi)p(\pi) \\
&= \mathrm{Binomial}(n,\pi) \times \mathrm{Beta}(\alpha,\beta) \\
&= \binom{n}{y}\pi^{y}(1-\pi)^{(n-y)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{(\alpha-1)}(1-\pi)^{(\beta-1)} \\
&\propto \pi^{y}(1-\pi)^{(n-y)}\pi^{(\alpha-1)}(1-\pi)^{(\beta-1)}
\end{aligned}
$$

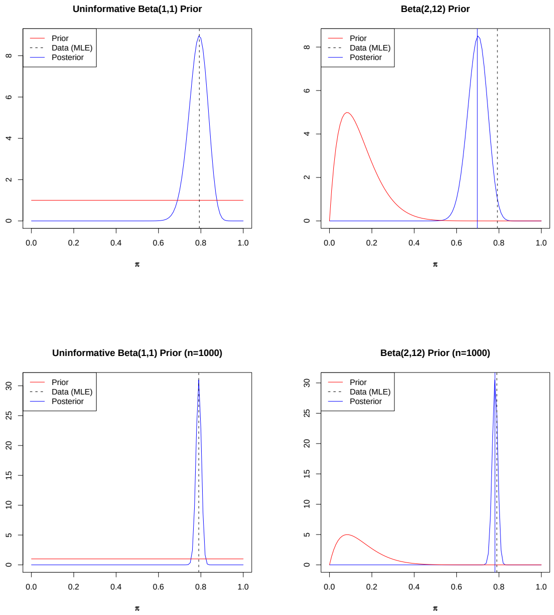
$$
p(\pi|y) \propto \pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1}
$$

The posterior distribution is simply a $\mathrm{Beta}(y+\alpha, n-y+\beta)$ distribution. Effectively, our prior is just adding $\alpha - 1$ successes and $\beta - 1$ failures to the dataset.

Bayesian priors are just adding pseudo observations to the data.

Since we know the posterior is a $\text{Beta}(y + \alpha, n - y + \beta)$ distribution, we can summarize it analytically or via simulation with the following quantities:

- posterior mean
- posterior standard deviation
- posterior credible intervals (credible sets)
- highest posterior density region


Uninformative Beta(1,1) Prior
Prior
Data (MLE)
Posterior
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
π
Beta(2,12) Prior
Prior
Data (MLE)
Posterior
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
π
Uninformative Beta(1,1) Prior (n=1000)
Prior
Data (MLE)
Posterior
0
5
10
15
20
25
30
0.0
0.2
0.4
0.6
0.8
1.0
π
Beta(2,12) Prior (n=1000)
Prior
Data (MLE)
Posterior
0
5
10
15
20
25
30
0.0
0.2
0.4
0.6
0.8
1.0
π

In the previous model, we had

$$\mathrm{Beta}(y+\alpha, n-y+\beta) \quad = \quad \frac{\mathrm{Binomial}(n,\pi) \times \mathrm{Beta}(\alpha,\beta)}{p(y)}$$

We knew that the likelihood × prior produced something that looked like a Beta distribution up to a constant of proportionality.

Since the posterior must be a probability distribution, we know that it is a Beta distribution and we can easily solve for the normalizing constant (although we don't need to since we already have the posterior).

When the posterior is the same distribution family as the prior, we have **conjugacy**.

Conjugate models are great because we can find the exact posterior, but we almost never have conjugacy except in very simple models.

What Happens When We Don't Have Conjugacy?

Consider a Poisson regression model with Normal priors on $\boldsymbol{\beta}$.

$$
\begin{aligned}
p(\boldsymbol{\beta}|\mathbf{y}) &\propto \prod_{i=1}^{n} \text{Poisson}(\lambda_i) \times \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\lambda_i &= \exp(\mathbf{x_i}\boldsymbol{\beta})
\end{aligned}
$$

$$
\begin{aligned}
p(\boldsymbol{\beta}|\mathbf{y}) \propto & \prod_{i=1}^{n} \frac{\exp(-e^{\mathbf{x_i}\boldsymbol{\beta}})\exp(\mathbf{x_i}\boldsymbol{\beta})^{y_i}}{y_i!} \times \\
& \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})\right)
\end{aligned}
$$

Likelihood × prior doesn't look like any distribution we know (non-conjugacy) and normalizing constant is too hard to find, so how do we find our posterior?

MCMC to the Rescue

Ideal Goal: Produce **independent** draws from our posterior distribution via simulation and summarize the posterior by using those draws.

Markov Chain Monte Carlo (MCMC): a class of algorithms that produce a chain of simulated draws from a distribution where each draw is **dependent** on the previous draw.

Theory: If our chain satisfies some basic conditions, then the chain will **eventually converge** to a stationary distribution (in our case, the posterior) and we have approximate draws from the posterior.

But there is no way to know for sure whether our chain has converged.

Algorithm 1: Gibbs Sampler

Let $\boldsymbol{\theta}^t = (\theta_1^t, \dots, \theta_k^t)$ be the tth draw of our parameter vector $\boldsymbol{\theta}$.

Draw a new vector $\boldsymbol{\theta}^{t+1}$ from the following distributions:

$$\begin{array}{rcl} \theta_1^{t+1} & \sim & p(\theta_1 | \theta_2^t, \dots, \theta_k^t, \mathbf{y}) \\ \theta_2^{t+1} & \sim & p(\theta_2 | \theta_1^{t+1}, \dots, \theta_k^t, \mathbf{y}) \\ & \vdots & \\ \theta_k^{t+1} & \sim & p(\theta_k | \theta_1^{t+1}, \dots, \theta_{k-1}^{t+1}, \mathbf{y}) \end{array}$$

Repeat m times to get m draws of our parameters from the approximate posterior (assuming convergence).

Requires that we know the conditional distributions for each θ. What if we don't?

Algorithm 2: Metropolis-Hastings

Draw a new vector $\boldsymbol{\theta}^{t+1}$ in the following way:

1. Specify a jumping distribution $J_{t+1}(\boldsymbol{\theta}^*|\boldsymbol{\theta}^t)$ (usually a symmetric distribution such as the multivariate normal).
2. Draw a proposed parameter vector $\boldsymbol{\theta}^*$ from the jumping distribution.
3. Accept $\boldsymbol{\theta}^*$ as $\boldsymbol{\theta}^{t+1}$ with probability min(r,1), where

$$r = \frac{p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{p(\mathbf{y}|\boldsymbol{\theta}^t)p(\boldsymbol{\theta}^t)}$$

If $\boldsymbol{\theta}^*$ is rejected, then $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t$.

Repeat m times to get m draws of our parameters from the approximate posterior (assuming convergence).

M-H always works, but can be very slow.

Outline

Introduction

Bayesian Models

Applications
- Missing Data
- Hierarchical Models

Missing Data

Suppose we have missing data. Define D as our data matrix and M as our missingness matrix.

$$D = \left(\begin{array}{cccc} y & x_1 & x_2 & x_3 \\ 1 & 2.5 & 432 & 0 \\ 5 & 3.2 & 543 & 1 \\ 2 & ? & 219 & ? \\ ? & 1.9 & ? & 1 \\ ? & 1.2 & 108 & 0 \\ ? & 7.7 & 95 & 1 \end{array}\right) \qquad M = \left(\begin{array}{cccc} y & x_1 & x_2 & x_3 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{array}\right)$$

One non-Bayesian approach to dealing with missing data is **multiple imputation**.

Multiple Imputation

We need a method for filling in the missing cells of D as a first step **before** we go to the analysis stage.

1. Assume a joint distribution for D_{obs} and M:

$$p(D_{obs}, M|\phi, \gamma) = \int p(D|\phi)p(M|D_{obs}, D_{mis}, \gamma)dD_{mis}$$

If we assume MAR, then

$$p(D_{obs}, M|\phi, \gamma) \quad \propto \quad \int p(D|\phi)dD_{mis}$$

2. Find ϕ, which characterizes the full D.

   $$L(\phi|D_{obs}) \quad = \quad \prod_{i=1}^{n} \int p(D_{i,obs}|\phi) dD_{mis}$$

   How do we do this integral?

   Assume Normality since the marginals of a multivariate Normal are Normal.

   $$L(\mu, \Sigma|D_{obs}) \quad = \quad \prod_{i=1}^{n} N(D_{i,obs}|\mu_{obs}, \Sigma_{obs})$$

3. Find $\hat{\mu}, \hat{\Sigma}$ and its distribution via EM algorithm and bootstrapping.
4. Draw m μ, Σ values, then use them to predict values for D_{mis}.

What we end up with is m datasets with missing values imputed.

We then run our regular analyses on the m datasets and combine the results using Rubin's rule.

Bayesian Approach to Missing Data

Bayesian paradigm: Everything unobserved is a random variable.

So we can set up the missing data as a "parameter" that we need to find.

$$\begin{aligned} p(D_{mis}, \phi, \gamma | D_{obs}, M) \quad &\propto \quad p(D_{obs}|D_{mis}, \phi)p(D_{mis}|\phi)p(M|D_{obs}, D_{mis}, \gamma) \\ & \quad\quad p(\gamma)p(\phi) \end{aligned}$$

If we assume MAR:

$$p(D_{mis}, \phi, \gamma | D_{obs}, M) \propto p(D_{obs}|D_{mis}, \phi)p(D_{mis}|\phi)p(\phi)$$

Use Gibbs Sampling or M-H to sample both D_{mis} and ϕ.

We don't have to assume normality of D to integrate over D_{mis}. We can just drop the draws of D_{mis}.

We can also incorporate both imputation and analyses in the same model.

$$p(\boldsymbol{\theta}, D_{mis}, \phi, \gamma | D_{obs}, M) \propto p(\mathbf{y}_{obs}, \mathbf{y}_{mis} | \boldsymbol{\theta}) p(D_{obs} | D_{mis}, \phi) p(D_{mis} | \phi) \\ p(\phi) p(\boldsymbol{\theta})$$

Again, find the posterior via Gibbs Sampling or M-H.

Moral: We can easily set up an application specific Bayesian model to incorporate missing data.

Multilevel Data

Suppose we have data in which we have $i = 1, \ldots, n$ observations that belong to one of $j = 1, \ldots, J$ groups.

Examples:

- students within schools
- multiple observations per country
- districts within states

We can have covariates on multiple levels. How do we deal with this type of data?

The Fixed Effects Model

We can let the intercept α vary by group:

$$y_i = \alpha_{j[i]} + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$$

This is known as the **fixed effects** model.

This is equivalent to estimating dummy variables for $J - 1$ groups.

We can use this model if we think that there is something inherent about the groups that affects our dependent variable.

However, the fixed effects model involves estimating many parameters, and also cannot take into account group-level covariates.

Hierarchical Model

A more flexible alternative is to use a **hierarchical model**, also known as a multilevel model, mixed effects model, or random effects model.

$$\begin{aligned} y_i &= \alpha_{j[i]} + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i \\ \alpha_j &\sim N(\alpha, \sigma_\alpha^2) \end{aligned}$$

Instead of assuming a completely different intercept for each group, we can assume that the intercepts are drawn from a common (Normal) distribution.

We can incorporate group-level covariates in the following way:

$$\begin{aligned} y_i &= \alpha_{j[i]} + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i \\ \alpha_j &\sim N(\gamma_0 + u_{j1}\gamma_1, \sigma_\alpha^2) \end{aligned}$$

or equivalently

$$\begin{aligned} y_i &= \alpha_{j[i]} + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i \\ \alpha_j &= \gamma_0 + u_{j1}\gamma_1 + \eta_j \\ \eta_j &\sim N(0, \sigma_\alpha^2) \end{aligned}$$

This is a relatively difficult model to estimate using non-Bayesian methods. The `lme4()` package in R can do it.

Bayesian Hierarchical Model

$$\begin{aligned} y_i &= \alpha_{j[i]} + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i \\ \alpha_j &\sim N(\gamma_0 + u_{j1}\gamma_1, \sigma_\alpha^2) \end{aligned}$$

We can do hierarchical models easily using Bayesian methods.

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{y}) \propto p(\mathbf{y} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) p(\boldsymbol{\alpha} | \boldsymbol{\gamma}) p(\boldsymbol{\gamma}) p(\boldsymbol{\beta})$$

Solve for the joint posterior using Gibbs Sampling or M-H.

We incorporate data with more than two levels easily as well.

Conclusion

There are pros and cons to using Bayesian statistics.

Pros:

- Incorporate outside/prior knowledge
- Estimate much more difficult models
- CI have more intuitive meaning
- Helps with unidentified models

Cons:

- It's hard(er)
- Computationally intensive
- Need defense of priors
- No guarantee of MCMC convergence

Statistical packages for Bayesian are also less developed (`MCMCpack()` in R, WinBUGS, JAGS).