Notes on the Hierarchical Beta Binomial Model

Kevin Quinn

October 31, 2017

1 The Model

Suppose we have the following model:

$$
\begin{aligned}
Y_i &\overset{ind.}{\sim} \mathcal{B}inomial(m_i, \theta_i), \qquad i = 1, \ldots, n \\
\theta_i &\overset{ind.}{\sim} \mathcal{B}eta(\alpha, \beta), \qquad i = 1, \ldots, n \\
\alpha &\sim \mathcal{E}xponential(a) \\
\beta &\sim \mathcal{E}xponential(b)
\end{aligned}
$$

where it is assumed that α and β are a priori independent.

Some particular examples of this model:

- clustered survey responses
  - i indexes geographic units (census blocks)
  - m_i is the number of survey respondents in geographic unit i
  - Y_i is the number of "yes" responses to a particular yes/no question
  - θ_i is the fraction of "yes" responders among all residents of geographic unit i who are in the sampling frame
  - The distribution of geographic-unit-specific "yes" response fractions in the population of geographic units is $\mathcal{B}eta(\alpha, \beta)$
  - Our prior beliefs about α and β are consistent with independent exponential distributions
- law school placements
  - i indexes law schools
  - m_i is the number of 3Ls in law school i
  - Y_i is the number of 3Ls who have full times jobs lined up after they graduate
  - θ_i is school i's underlying placement rate.
  - The distribution of underlying placement rates in the population of law schools is $\mathcal{B}eta(\alpha, \beta)$
  - Our prior beliefs about α and β are consistent with independent exponential distributions
- baseball statistics

- i indexes baseball players
- m_i is the number of at bats for player i
- Y_i is the number of hits player i got in m_i at bats
- θ_i is player i's underlying propensity to get hits (his/her underlying batting average).
- The distribution of underlying batting averages in the population of players is $\mathcal{B}eta(\alpha, \beta)$
- Our prior beliefs about α and β are consistent with independent exponential distributions

Note the different interpretations of what θ_i represents as well as what constitutes the population.

2 The Posterior Distribution

In what follows, I'll use the following shorthand:

$$f_{bin}(y|m, \theta) = \binom{m}{y}\theta^y(1-\theta)^{m-y}$$

$$f_{beta}(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$f_{exp}(\alpha|a) = \begin{cases} a\exp(-a\alpha) & \text{if } \alpha \geq 0 \\ 0 & \text{if } \alpha < 0 \end{cases}$$

$$f_{bb}(y|m, \alpha, \beta) = \binom{m}{y}\frac{B(y+\alpha,\ m-y+\beta)}{B(\alpha,\ \beta)}$$

where Γ is the gamma function and B is the beta function. $\mathbf{y}$ denotes the n-vector of the realized values of Y_is. Also, we'll let $\boldsymbol{\theta}$ denote the n-vector of θ_i values.

We can write the posterior for α, β, and $\boldsymbol{\theta}$ as:

$$p(\alpha, \beta, \boldsymbol{\theta}|\mathbf{y}) \propto \left[f_{exp}(\alpha|a)f_{exp}(\beta|b)\prod_{i=1}^{n} f_{beta}(\theta_i|\alpha, \beta)\right]\left\{\prod_{i=1}^{n} f_{bin}(y_i|m_i, \theta_i)\right\}$$

The terms in the square brackets are our prior for $(\alpha, \beta, \boldsymbol{\theta})$. The piece in the curved brackets is the likelihood given $\boldsymbol{\theta}$. Note that y_i is assumed to be conditionally independent of (α, β) given θ_i. Thus, after conditioning on $\boldsymbol{\theta}$, α and β don't show up in the likelihood.

We can write the posterior for α and β marginalized over $\boldsymbol{\theta}$ as:

$$\begin{aligned} p(\alpha, \beta|\mathbf{y}) &\propto f_{exp}(\alpha|a)f_{exp}(\beta|b)\prod_{i=1}^{n}\int_0^1 f_{beta}(\theta_i|\alpha, \beta)f_{bin}(y_i|m_i, \theta_i)d\theta_i \\ &\propto f_{exp}(\alpha|a)f_{exp}(\beta|b)\prod_{i=1}^{n} f_{bb}(y_i|m_i, \alpha, \beta) \end{aligned} \tag{1}$$

3 Fitting the Model

The posterior distribution for this model is not a member of a known parametric family of distributions (it's not normal, beta, etc.). Thus we can't rely on existing analytic results or canned sampling procedures to summarize it. In what follows I'll outline two approaches to sampling from $p(\alpha, \beta, \boldsymbol{\theta}|\mathbf{y})$ that rely on Markov chain Monte Carlo (MCMC) methods. Chapter 5 of Jackman (2009) provides more detail on MCMC methods.

3.1 Strategy 1: Metropolis Sampling After Marginalizing over θ

The idea here is to sample (α, β) from $p(\alpha, \beta|\mathbf{y})$ and to then sample $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\mathbf{y}, \alpha, \beta)$ where the values of α and β we condition on are those drawn from the first step. In essence, this is just a variation on the method of composition. $p(\boldsymbol{\theta}|\mathbf{y}, \alpha, \beta)$ is easy to sample from—each $[\theta_i|\mathbf{y}, \alpha, \beta] \overset{ind.}{\sim} \mathcal{B}eta(\alpha + y_i,\ \beta + m_i - y_i)$.

Sampling (α, β) from $p(\alpha, \beta|\mathbf{y})$ is more complicated. To do that we'll use a *random walk Metropolis algorithm.*

The full algorithm for sampling from $p(\alpha, \beta, \boldsymbol{\theta}|\mathbf{y})$ is the following:

1. set α and β to some initial values
2. Repeat the following M times:
   (a) generate a *candidate value* $(\alpha_{can}, \beta_{can})' \sim \mathcal{N}((\alpha, \beta)', \mathbf{V})$ where $\mathbf{V}$ is a fixed, user-defined variance-covariance matrix.
   (b) calculate $r = \frac{p(\alpha_{can}, \beta_{can}|\mathbf{y})}{p(\alpha, \beta|\mathbf{y})}$ (using Equation 1).
   (c) generate $U \sim \mathcal{U}niform(0, 1)$
   (d) if $U < r$ set $(\alpha, \beta) \leftarrow (\alpha_{can}, \beta_{can})$, otherwise leave α and β at their current values.
   (e) sample $[\theta_i|\mathbf{y}, \alpha, \beta] \overset{ind.}{\sim} \mathcal{B}eta(\alpha + y_i,\ \beta + m_i - y_i)$ for $i = 1, \ldots, n$
   (f) store the values of α, β, and $\boldsymbol{\theta}$

This produces M draws approximately from the joint posterior distribution of $(\alpha, \beta, \boldsymbol{\theta})$.

3.2 Strategy 2: Metropolis within Gibbs

A second approach does not analytically marginalize over $\boldsymbol{\theta}$. Here we iteratively sample from $[\alpha, \beta|\mathbf{y}, \boldsymbol{\theta}]$ and $[\boldsymbol{\theta}|\mathbf{y}, \alpha, \beta]$. If we could directly sample from these conditional distributions this would be an example of *Gibbs sampling.* Directly sampling from $[\alpha, \beta|\mathbf{y}, \boldsymbol{\theta}]$ is not possible so we need to use a Metropolis step for this piece. This gives rise to what is sometimes called a *Metropolis within Gibbs* setup.

Before proceeding, note that

$$p(\alpha, \beta|\mathbf{y}, \boldsymbol{\theta}) \propto f_{exp}(\alpha|a) f_{exp}(\beta|b) \prod_{i=1}^{n} f_{beta}(\theta_i|\alpha, \beta) \tag{2}$$

(the likelihood for $\mathbf{y}$ drops out because $\mathbf{y}$ and (α, β) are conditionally independent given $\boldsymbol{\theta}$.)

The full algorithm for sampling from $p(\alpha, \beta, \boldsymbol{\theta}|\mathbf{y})$ is the following:

1. set α and β to some initial values
2. Repeat the following M times:
   (a) generate a *candidate value* $(\alpha_{can}, \beta_{can})' \sim \mathcal{N}((\alpha, \beta)', \mathbf{V})$ where $\mathbf{V}$ is a fixed, user-defined variance-covariance matrix.
   (b) calculate $r = \frac{p(\alpha_{can}, \beta_{can}|\mathbf{y}, \boldsymbol{\theta})}{p(\alpha, \beta|\mathbf{y}, \boldsymbol{\theta})}$ (using Equation 2).
   (c) generate $U \sim \mathcal{U}niform(0, 1)$
   (d) if $U < r$ set $(\alpha, \beta) \leftarrow (\alpha_{can}, \beta_{can})$, otherwise leave α and β at their current values.

(e) sample $[\theta_i|\mathbf{y}, \alpha, \beta] \overset{ind.}{\sim} \mathcal{B}eta(\alpha + y_i,\ \beta + m_i - y_i)$ for $i = 1, \ldots, n$

(f) store the values of α, β, and $\boldsymbol{\theta}$

This produces M draws approximately from the joint posterior distribution of $(\alpha, \beta, \boldsymbol{\theta})$.

References

Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*. Wiley.