Bayesian Model for New York City Taxi Rides

Harold M. Li

March 27, 2016

1 Description of Model

1.1 Introduction

In this Bayesian model, we treat each set of observations as the number of taxi rides in each Manhattan census tract on a hourly basis over an entire week. In mathematical notation,

$$(X_{h,c}), h = 1 \dots 168, c = 1 \dots 288$$

describes the number of taxi rates at hour h and census tract c. (There are 168 hours in a week, and 288 census tracts in Manhattan).

If we collect a year's worth of taxi ride data (i.e. 52 weeks), the total set of observations can be written as

$$(X_{h,c,i}), h = 1 \dots 168, c = 1 \dots 288, i = 1 \dots 52$$

Our goal here is to fit multiple sets of weekly taxi ride observations into a sensible Bayesian model.

1.2 The Bayesian Model

The actual taxi counts data varies for various census tracts. Some regions, like Times Square, have plenty of passengers that are picked up by cabs. As a result, for a given hour, the distribution of taxi pickups looks very much like a symmetric bell curve. However, for a less populated region, like Washington Heights, the distribution of taxi pickups would look like a skewed curve concentrated on small frequencies. As a result, we decide to model these counts as Poisson distributions, where the parameter varies across time and census tract.

$$X_{h,c} \sim \text{Poisson}(\lambda_{h,c}) \quad \forall h, c$$

However, this would resemble an unpooled method where the data across time and region would be totally independent. Given that there should be some kind of geospatial and time-wise connection between all these variables, we decide to derive the $\lambda_{h,c}$'s that share the same time period from a similar distribution.

$$\lambda_{h,c} \sim \text{Gamma}(\alpha_{h,c}, \beta_h) \quad \forall h, c$$

Notice that the λ's in the same hour of the week have the same β, which is the scale parameter of the Gamma distribution. This makes sense because the number of taxi rides across the city change significantly based on time of day. However, we still leave the α parameter different for all of them to allow some flexibility of the model. In addition, the α's will be deterministic to ease calculation of posterior conditionals.

The β's, on the other hand, will all be grouped to derive from a smaller set of Gamma distributions. The rationale for this is that taxi ride distributions among weekdays are very similar, as well as among weekdays. Thus, the 168 β's will now follow a set of 48 Gamma distributions - 24 of them devoted to weekdays, and the other 24 devoted to weekends.

$$\beta_h \sim \text{Gamma}(a_z, b_z) \quad \forall z = 1 \dots 48$$

Again, we let a_z's to be different among the 48 time situations, and let them be deterministic to ease calculations.

Finally, to tie all of these observations among different time periods and census tracts together, we model the b_z's to derive from a single Gamma distribution.

$$b_z \sim \text{Gamma}(p, q) \quad \forall z$$

With that, we have our Bayesian model. Next, we discuss how to obtain the posterior distribution of this model.

1.3 Simulating The Posterior Distribution

Deriving the full posterior distribution of a Bayesian model usually requires numerous calculations. Furthermore, simulating the distribution by computing integrals have shown to be frequently intractable. The Gibbs sampler is an effective algorithm that can simulate the posterior distribution given that we know the distribution of the posterior conditionals. The following equations shows how the posterior conditionals can be computed.

The full posterior distribution can be written as such:

$$\text{p}((\lambda_{h,c})_{\forall h,c}, (\beta_h)_{\forall h}, (b_z)_{\forall z} | (x_{h,c,i})_{\forall h,c,i}))$$

$$= \prod_{h,c} \mathrm{p}(x_{h,c,1\ldots N}|\lambda_{h,c}) \cdot \prod_{h,c} \mathrm{p}(\lambda_{h,c}|\beta_h) \cdot \prod_{h} \mathrm{p}(\beta_h|b_{f(h)}) \cdot \prod_{z} \mathrm{p}(b_z)$$

Note that the function f maps the 168 hours to the 48 time periods that define weekdays and weekends.

The log of the posterior distribution can be written as such:

$$\sum_{i=1}^{N} \sum_{h,c} \Big(x_{h,c,i} \log(\lambda_{h,c}) - \lambda_{h,c} - \log(x_{h,c,i}!) \Big)$$
$$+ \sum_{h,c} \Big((\alpha_{h,c} - 1) \ \log(\lambda_{h,c}) - \beta_h \lambda_{h,c} - \log\Gamma(\alpha_{h,c}) + \alpha_{h,c} \log(\beta_h) \Big)$$
$$+ \sum_{h} \Big((a_{f(h)} - 1) \log(\beta_h) - b_{f(h)} \beta_h - \log\Gamma(a_{f(h)}) + a_{f(h)} \log(b_{f(h)}) \Big)$$
$$+ \sum_{z} \Big((p-1) \log(a_z) - q a_z - \log\Gamma(p) + p \log(q) \Big)$$

1.3.1 Posterior Conditional on $\lambda_{h,c}$

To get the posterior conditional, we only the include in the full posterior distribution that contain λ:

$$\log \mathrm{p}(\lambda_{h,c}|(x_{h,c,i})_{\forall h,c,i}, (\beta_h)_{\forall h}, (b_z)_{\forall z})$$
$$= \sum_{i=1}^{N} \Big(x_{h,c,i} \log(\lambda_{h,c}) - \lambda_{h,c} \Big) + (\alpha_{h,c} - 1) \log(\lambda_{h,c}) - \beta_h \lambda_{h,c}$$
$$= \log(\lambda_{h,c}) \Big(\alpha_{h,c} + \sum_{i=1}^{N} x_{h,c,i} - 1 \Big) - \lambda_{h,c} \Big(n + \beta_h \Big)$$
$$\boxed{\sim \mathrm{Gamma}\Big(\alpha_{h,c} + \sum_{i=1}^{N} x_{h,c,i}, n + \beta_h \Big)}$$

1.3.2 Posterior Conditional on β_h

To get the posterior conditional, we only the include in the full posterior distribution that contain β:

$$\log \mathrm{p}(\beta_h|(x_{h,c,i})_{\forall h,c,i}, (\lambda_{h,c})_{\forall h,c}, (b_z)_{\forall z})$$

$$= \sum_{c=1}^{C} \Big(-\beta_h \lambda_{h,c} + \alpha_{h,c} \log \beta_h \Big) + (a_{f(h)} - 1) \log \beta_h - b_{f(h)} \beta_h$$

$$= \log \beta_h \Big(a_{f(h)} + \sum_{c=1}^{C} \alpha_{h,c} - 1 \Big) - \beta_h \Big(b_{f(h)} + \sum_{c=1}^{C} \lambda_{h,c} \Big)$$

$$\boxed{\sim \text{Gamma}\Big(a_{f(h)} + \sum_{c=1}^{C} \alpha_{h,c}, b_{f(h)} + \sum_{c=1}^{C} \lambda_{h,c} \Big)}$$

1.3.3 Posterior Conditional on b_z

To get the posterior conditional, we only the include in the full posterior distribution that contain b:

$$\log \text{p}(b_z | (x_{h,c,i})_{\forall h,c,i}, (\lambda_{h,c})_{\forall h,c}, (beta_h)_{\forall h})$$

$$= \sum_{h:f(h)=z} \Big(-b_z \beta_h + a_z \log b_z \Big) + (p-1) \log b_z - q b_z$$

$$= \log b_z \Big(p + n_z a_z - 1 \Big) - b_z \Big(q + \sum_{h:f(h)=z} \beta_h \Big)$$

$$\boxed{\sim \text{Gamma}\Big(p + n_z a_z, q + \sum_{h:f(h)=z} \beta_h \Big)}$$

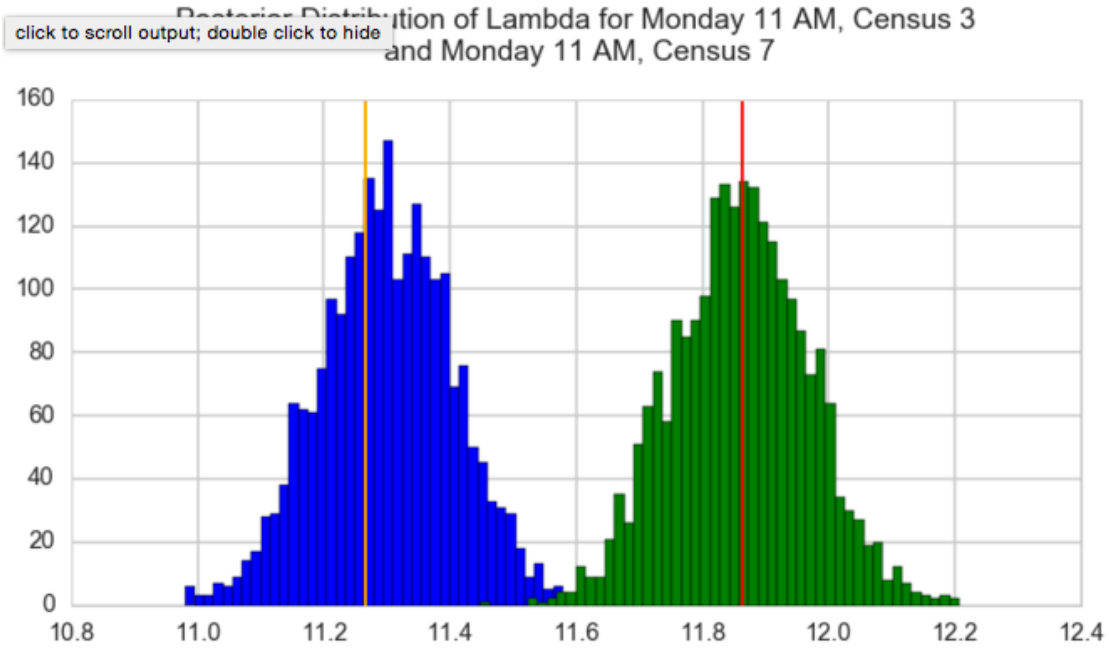
Note that n_z denotes the number of hourly periods that correspond to time segment z.

2 Testing the Model

We test the model by generating 1000 sets of weekly observations $(x_{h,c})$, which also calibrates $(\lambda_{h,c}), (\beta_h), (b_z)$ in the process. We set the parameters to be $p = 10, q = 2$, all a_z's to be 10, and all $\alpha_{h,c}$ to be 20. Then, we run the Gibbs Sampler to simulate the posterior distribution for the variables $(\lambda_{h,c}), (\beta_h), (b_z)$, and compare the distribution with the true $(\lambda_{h,c}), (\beta_h), (b_z)$'s that derived the weekly observations $(x_{h,c})$.

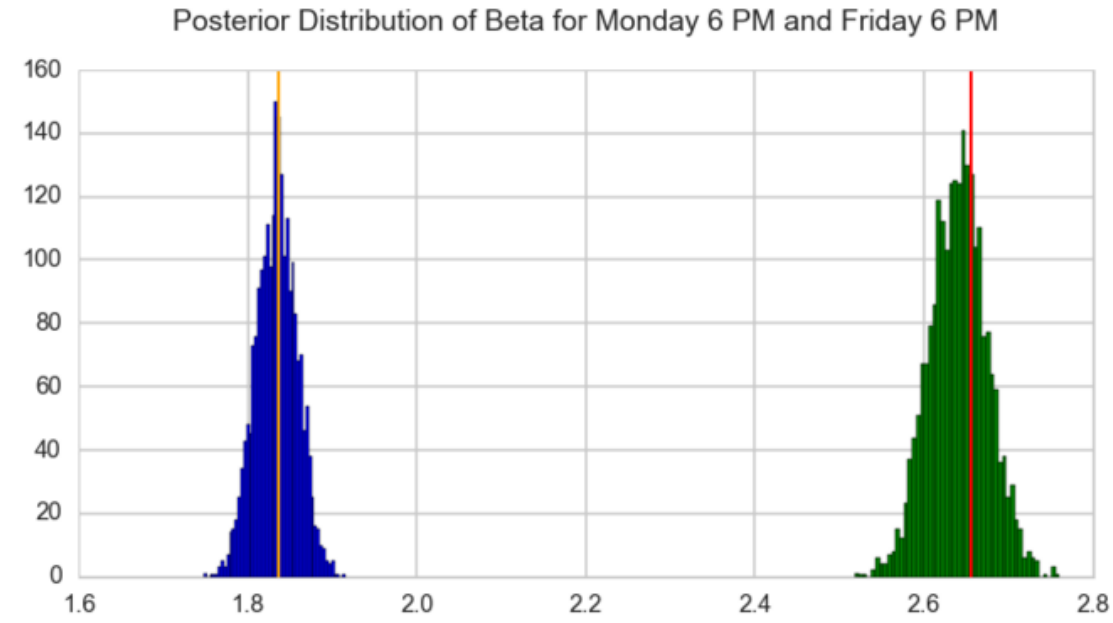
2.1 Posterior Distribution for $\lambda_{h,c}$

This is an example of λ posterior distributions for Monday 11 AM for two different census tracts. $(h = 11, c = 4, 7)$



2.2 Posterior Distribution for β_h

This is an example of β posterior distributions for Monday and Friday at 6 PM.
$(h = 18, 114)$



2.3 Posterior Distribution for b_z

This is an example of b posterior distributions for Weekday and Weekend periods at 6 PM. $(z = 18, 42)$