Analyzing Boston 311 Responses with Gaussian Mixture Models

Jane Huang Harvard University
Isadora Nun Harvard University
Weiwei Pan Harvard University
Francisco Rivera Harvard University

Abstract

Fill in later. We approximate the distribution of the response times, longitude, and latitude of 311 requests as a three-component Gaussian mixture model.

1 Introduction

In Boston, thousands of requests are made each week to city non-emergency services to address issues such as graffiti, potholes, and broken traffic signals [2]. Requests made to non-emergency services fall under the umbrella of Boston 311.[1] Ensuring that requests are fulfilled in a timely manner and that services are accessible to all segments of the population is essential for maintaining the safety and well-being of city residents.

Traditionally, requests for city non-emergency services have been made through phone calls, but after smartphone technology became increasingly common, the Citizens Connect app was introduced in 2009 to allow Boston residents to report problems to 311 through a mobile phone interface [1]. The app developers state their motivation as: "Residents report public issues directly from their smart phones into the City's work order management system so that it gets immediately to the right person in City Hall to fix the problem...We were interested in seeing if we could engage more or different residents" [1].

To assess the extent to which the 311 app may facilitate more efficient responses to populations not as well-served through the traditional mode of contacting non-emergency services, we seek to model and compare the joint distribution of response times and locations for 311 requests made via constituent calls and those made via the mobile phone app. [WHICH FIGURES DO WE WANT TO SHOW HERE?] Because we are interested in identifying whether there are hidden subpopulations of requests that are distinguishable by the observed locations and response times, we use Gaussian mixture models to approximate the distributions.

[1]When city non-emergency services was rebranded recently as Boston 311, the Citizens Connect App was renamed BOS:311. We refer to the app as Citizens Connect throughout for continuity.

2 Data

Records of Boston's 311 service requests were obtained from `https://data.cityofboston.gov/City-Services/311-Service-Requests/awu8-dc52Opensinnewwindow`. We first selected all closed complaints that had been opened in 2015 through constituent calls or the Citizens Connect App. The descriptors extracted for the data included the times that the complaints were opened and closed by the city (reported to the nearest second) and the longitude and latitude of the source of the complaint. A new variable, response time, was defined as the difference between the reported closing and opening times of a 311 complaint. Because the numerical scale of response times is much larger than those of longitude and latitude, the values for each variable were rescaled to have zero mean and unit variance. (The mean response time is 13.2 days with standard deviation 29.1 days. The mean latitude is 42.326 with standard deviation 0.034. The mean longitude is -71.083 with standard deviation 0.035). The data were subsequently split apart into two sets based on whether they originated from constituent calls or Citizens Connect.

Because the dataset is very large, a randomly selected subsample was used for the analysis to reduce computational expense. For the expectation maximization estimates, the data were downsampled to provide $\sim 20,000$ points each for constituent call and Citizens Connect App analysis. For Gibbs sampling, the data were downsampled to provide $\sim 10,000$ points for each group.

3 Methods

3.1 Bayesian Gaussian Mixture Models

Each 311 request is described by a three-dimensional vector consisting of x_n = (response time, longitude,

latitude). We model the distribution of these datapoints as a mixture of K Gaussian components with means $\mu = \{\mu_k : 1 \leq k \leq K\}$ and covariance matrices $\Sigma = \{\Sigma_k : 1 \leq k \leq K\}$. The mixture coefficients for the model (i.e., the fraction of the population belonging to each component) is represented by π, a vector with K elements.

The forms of the equations describing Gaussian mixture models and the conditional distributions are adopted from [3, 4].

Since the component membership is not known, we specify the component membership indicators as $\mathbf{Z} = (z_1, \ldots, z_N)$. Each indicator for datapoint x_n is a K-element vector z_n, defined such that

$$z_{nk} = \begin{cases} 1, & x_n \text{ belongs to the } k\text{-th component} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Hence, the likelihood of the observations is

$$L(\mathbf{X}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{nk}}. \tag{2}$$

The prior for π is a Dirichlet distribution. For Σ, we use an inverse-Wishart prior, which is conjugate to the multivariate normal distribution and ensures the selection of a positive-definite matrix. The hyperparameters for the inverse-Wishart prior are the scale matrix S_0 and the degrees of freedom ν_0. These priors can be set to be weakly informative, which is useful because it is difficult to assess *a priori* what the components in the 311 data are. In addition, conjugate priors facilitate the use of Gibbs sampling to explore the posterior distribution.

To summarize, our model is described by the following:

$$\pi \sim Dir(\alpha_0) \tag{3}$$
$$\Sigma_k \sim \text{invWish}(S_0, \nu_0) \tag{4}$$
$$\mu_k|\Sigma_k \sim \mathcal{N}(m_0, V_0) \tag{5}$$
$$z_n|\pi \sim \prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{6}$$
$$x_n|\mathbf{Z}, \mu, \Sigma \sim \prod_{k=1}^{K} \mathcal{N}(\mu_k, \Sigma_k)^{z_{nk}} \tag{7}$$

3.2 Expectation maximization

3.2.1 Simulated Annealing

3.2.2 Gibbs sampling

For the Gibbs sampler, each iteration requires the following steps:

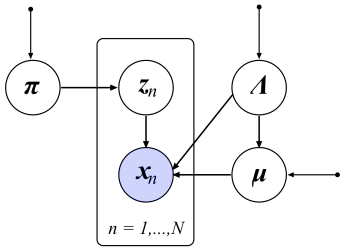

Figure 1: NEED TO UPDATE THIS FIGURE TO USE SIGMA INSTEAD OF GAMMA

1. Each datapoint's indicator variable Z_n is drawn from a multinomial distribution with the event probabilities given by

$$p(z_{nk} = 1|x_n, \mu, \Sigma, \pi) \propto \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \tag{8}$$

2. The mixture coefficients are drawn from the conditional distribution

$$p(\pi|\mathbf{Z}) = \text{Dir}(\{\alpha_{0,k} + N_k\}) \tag{9}$$

where N_k is the number of observations assigned to each cluster.

3. The component means are drawn from the conditional distribution

$$p(\mu_k|\Sigma_k, \mathbf{Z}, \mathbf{X}) = \mathcal{N}(\mu_k|m_k, V_k) \tag{10}$$

where

$$V_k^{-1} = V_0^{-1} + N_k \Sigma_k^{-1} \tag{11}$$

and

$$m_k = V_k(\Sigma_k^{-1} N_k \overline{x}_k + V_0^{-1} m_0) \tag{12}$$

with $\overline{x}_k$ defined as the mean value of the observations assigned to component k.

4. Finally, the component covariance matrices are drawn from the conditional distribution

$$p(\Sigma_k|\mu_k, \mathbf{Z}, \mathbf{X}) = IW(\Sigma_k|S_k, \nu_0 + N_k) \tag{13}$$

where

$$S_k = S_0 + \sum_{n=1}^{N} z_{nk}(x_n - \mu_k)(x_n - \mu_k)^\top \tag{14}$$

One common issue with using Gaussian mixture models is known as the "label-switching problem," which occurs when component identities are ambiguous because the likelihood is the same if component labels are exchanged. We follow the suggestion of Gelman et al. [3] to resolve the ambiguity by defining $\pi_1 > \pi_2 > \pi_3$. However, this type of identifiability constraint may not

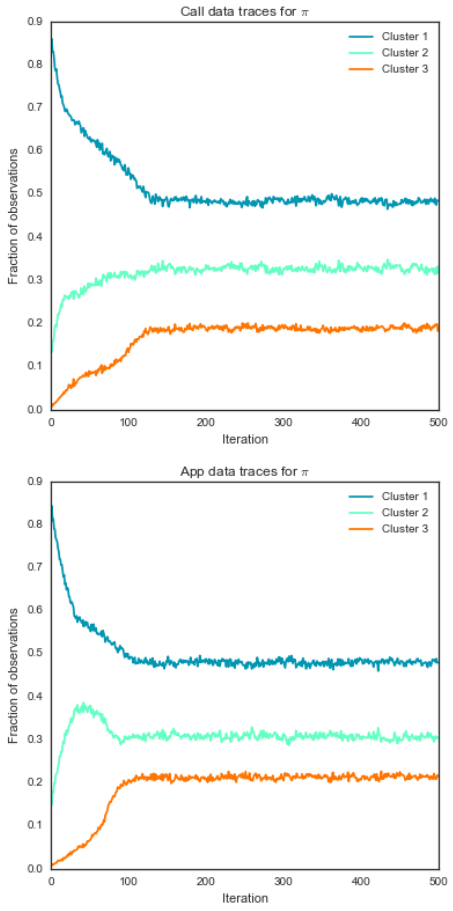

Figure 2: Gibbs sampler traces for the π parameter

always break symmetry as desired, particularly if the fractions are similar in size [5].

We use the MAP estimate obtained from expectation maximization to initialize the values of π, μ, and Σ for the Gibbs sampler. We then set $\alpha_0 = (1, 1, 1)$, $m_0 = (1, 1, 1)$, $\nu_0 = 3$ (the number of components in the mixture model), and S_0 and V_0 to the identity matrices to maintain weakly informative priors.

4 Results

4.1 Expectation maximization

Using the Bayesian Information Criterion [insert equation here], $K = 3$ was determined to be the optimal number of components to use to model the observations in the simulated annealing and Gibbs sampling analysis.

4.2 Simulated annealing

4.3 Gibbs sampling

The Gibbs sampler was run for a total of 500 iterations for both the constituent call and app data. Inspection of the trace plots suggest that convergence is achieved after 150 iterations (see Fig. 2 for sample trace plots), so a burn-in of 150 iterations is applied for calculation of the posterior means and uncertainties, listed in Table 1. DISCUSSION/COMPARISON OF RESULTS

4.4 Posterior predictive

To assess how well the model describes the data, we simulate data with the posterior predictive distribution. The posterior predictive distribution is generated by sampling from the posterior obtained through Gibbs sampling, then using those parameter values to sample from the likelihood.

4.5 Comparing results from different methods

5 Conclusion

References

[1] Boston's Mayor's Office of New Urban Mechanics. Citizens Connect. `http://www.cityofboston.gov/news/Default.aspx?id=20283`.

[2] City of Boston Mayor's Office. Mayor Walsh Launches Boston 311. Press Release, August 2015. `http://www.cityofboston.gov/news/Default.aspx?id=20283`.

[3] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2nd edition, 2004.

[4] J. H. Jones and S. Jackman. Bayesian hierarchical mixture models for high-risk births. Population Association of American 2008 Annual Meeting, 2008.

[5] M. Stephens. Dealing with label switching in mixture models. *J. R. Statist. Soc. B*, 62:795–809, 2000.

Table 1: Gibbs results

Parameter	Constituent Call	App data
π	$\begin{pmatrix} .484 \pm .006 & .328 \pm .006 & .189 \pm .005 \end{pmatrix}$	
μ_1	$\begin{pmatrix} -.419 \pm .001 & -.13 \pm .02 & -.11 \pm .02 \end{pmatrix}$	
Σ_1	$\begin{pmatrix} .0018 & -.0063 & -.0065 \\ -.0063 & 1.1 & .60 \\ -.0065 & .60 & 1.1 \end{pmatrix}$	
μ_2		
Σ_2		
μ_3		
Σ_3		

Variables listed in order of (response time, longitude, latitude) in standardized units. The mean response time is 13.2 days with standard deviation 29.1 days. The mean latitude is 42.326 with standard deviation 0.034. The mean longitude is -71.083 with standard deviation 0.035.