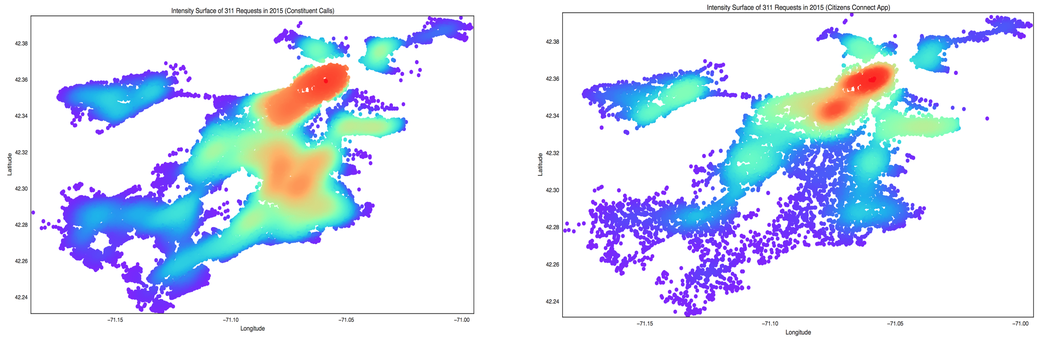
Analyzing Boston 311 Requests with Gaussian Mixture Models

Jane Huang, Isadora Nun, Weiwei Pan and Francisco Rivera
AM 207 Spring 2016

Introduction:

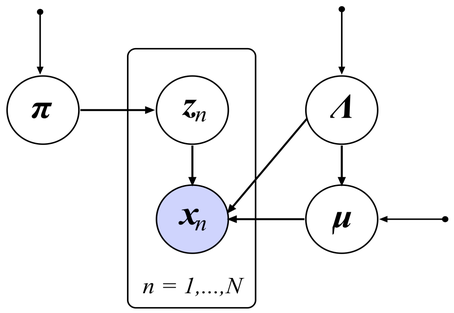
Traditionally, Boston residents have been able to report problems such as potholes or graffiti through phone calls to city non-emergency services. Ubiquitous smartphone technology now also allows Boston residents to contact 311 services through the Citizens Connect App.

Goal: Compare the effectiveness of responses to requests made through constituent calls and Citizens Connect App by modeling the geographic and response time distribution with Gaussian mixture models.



311 request volume intensity surface for constituent calls and Citizens Connect App (2015)

The Model



Our K-component Gaussian mixture model is defined as:

$$\pi \sim Dir(\alpha_0) \quad \text{(Mixture Coefficient)}$$

$$\Lambda_k \sim Wish(W_0, \nu_0) \quad \text{(Component Precision Matrix)}$$

$$\mu_k|\Lambda_k \sim \mathcal{N}(\eta_0, (\beta_0\Lambda_k)^{-1}) \quad \text{(Component Mean)}$$

$$z_n|\pi \sim \prod_{k=1}^{K} \pi_k^{z_{nk}} \quad \text{(Label)}$$

$$x_n|Z, \mu, \Lambda \sim \prod_{k=1}^{K} \mathcal{N}(\mu_k, \Lambda_k)^{z_{nk}} \quad \text{(Likelihood)}$$

Note that the set of hyper-parameters of our model is

$$\theta = (\alpha_0, W_0, \nu_0, \eta_0, \beta_0).$$

Methods

We analyze a randomly drawn subset of 2015 reports made to Boston 311 through either constituent calls or the Citizens Connect App.

This dataset is publicly available at:

$$\texttt{https://data.cityofboston.gov/}.$$

Each datapoint, $\mathbf{x}_n$, is a three-dimensional vector with features:

$$(\textit{response time}, \textit{longitude}, \textit{latitude}).$$

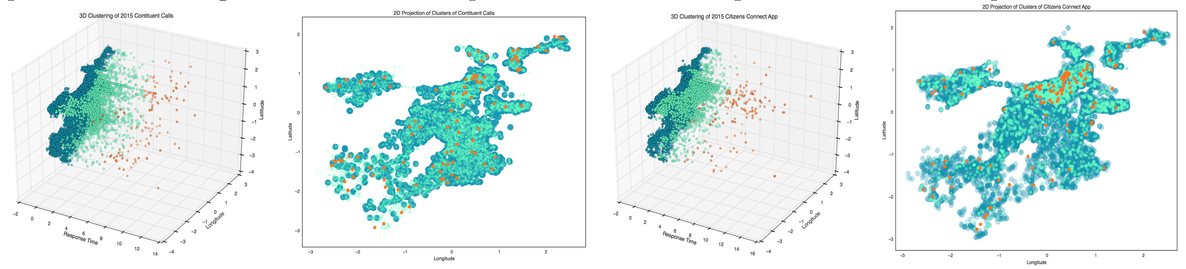
Using the Bayes Information Criterion, we chose a model with $K = 3$ components.

We compute both maximum likelihood estimates and maximum a posteriori estimates for the component parameters through expectation maximization. We compared these point estimates to the results for simulated annealing.

The point estimates are used to initialize a Gibbs sampler to obtain the posterior distribution. These samples are then used to generate the posterior predictive distribution to compare to the original data.

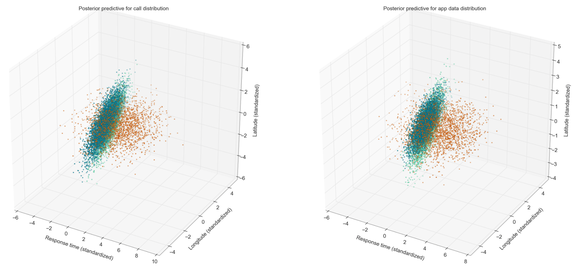
Results for Latent Clusters in 311 Data

Clustering based on MAP Estimates Our different estimation techniques consistently indicate that for both the Citizens Connect and constituent call data, the largest cluster is characterized by a short response time, the next largest cluster is characterized by a somewhat longer response time, and the smallest cluster is characterized by a very slow mean response time. For example, for MAP, the results for constituent calls and the Citizens Connect App are similar, with the largest cluster centered at 4.5 days, the next largest at 50 days, and the smallest at 266 days. The covariances between geography and response time are larger for the slowest response time cluster compared to the faster response time clusters.



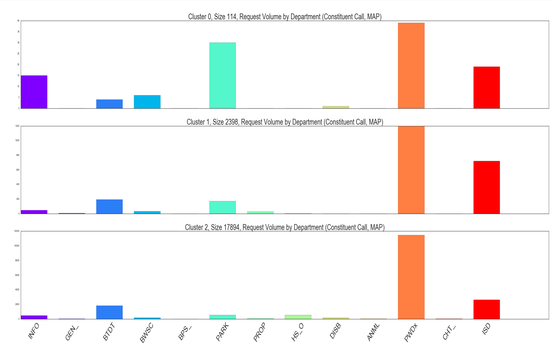
(a) 3D & 2D projections of clusters (constituent calls) **(b)** 3D & 2D projections of clusters (Citizens Connect App)

Posterior predictive distribution We show the posterior predictive distributions generated from our Gibbs sampler. While the full complexity of the Boston 311 data is not completely described by the three component Gaussian mixture model, a mixture model is still useful for understanding how the 311 requests spread apart into components differentiated by response times.

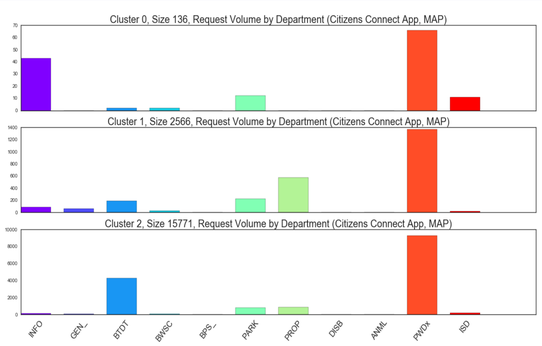


(b) Posterior predictive distributions for call and app data

Component Analysis



(a) Cluster profiles (constituent call)



(a) Cluster profiles (Citizens Connect App)

After identifying the clusters, we examined the qualitative data associated with each request to understand the composition of each cluster by type of service requested. While for both Citizens Connect App and constituent call requests, the clustering is comparable in terms of slow, medium, and fast response times, the differing compositions between the two sets of clusters hints that the reasons for the lagging response times may be associated with different latent characteristics. These results provide a foundation for building more detailed models examining predictors of lagging response times.