Point Estimates

1 Data

We work with a year of 311 service request data and, in the rest of the notebook, analyze the differences/similaritys between geographic/response time distribution between service requests generated by constituent calls and service requests generated by the Citizens Connect App.

We filter the full 311 service request data set for closed requests that are opened between `January 01, 2015 12:00AM` and `January 01, 2016 12:00AM`. This data set is then split into `call_data`, requests generated by constituent calls, and `app_data`, requests generated by the Citizens Connect App. We then choose a random subset from each `call_data` and `app_data` (by randomizing and then slicing), obtaining two datasets each of approximately 20,000 data points.

For better model fitting, we standardize each feature of the pooled set of sliced `call_data` and sliced `app_data`, so that each feature has mean zero and standard deviation 1.

2 Model Selection

The Bayesian Information Criterion (BIC) score is a weighted difference between the log-likelihood of a model and the complexity of the model (as well as the size of the data),

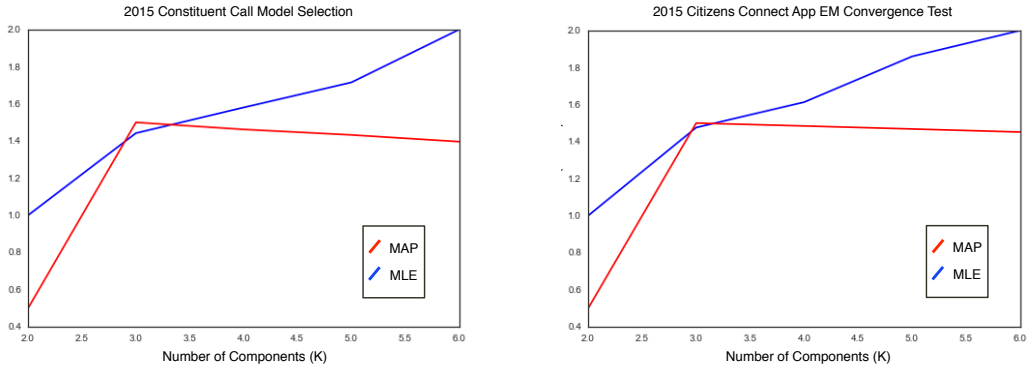
$$\log \mathcal{L}(X|\theta^{MLE}, \mathcal{M}) - \frac{1}{2}\kappa_{\mathcal{M}} * \log(N), \tag{1}$$

where $\mathcal{L}$ is the log-likelihood of the data, X, given the model, $\mathcal{M}$, and the maximum likelihood parameters of the model, θ^{MLE}. The BIC score approximates the evidence for the data $\log p(X|\mathcal{M})$ given the model $\mathcal{M}$, under the assumption that data distribution is an exponential family. In model selection, higher BIC scores indicate a preferable balance between fewer model parameters (preventing overfitting) and better model fit. In a Bayesian framework, we want to select the model with the largest posterior probability. This means choosing the model with the largest integrated complete likelihood (ICL). A BIC-like approximation of the ICL is proposed by Biernacki et al ([1])to be

$$\log \mathcal{L}(X, Z|\theta^{MAP}, \mathcal{M}) - \frac{1}{2}\kappa_{\mathcal{M}} * \log(N), \tag{2}$$

where X is the data and Z is the latent cluster labels. From Figure 2, we see that the most appropriate number of clusters for MAP estimates is $K = 3$.

Figure 1: Model Selection Using BIC Scores



(a) BIC score plot for Constituent Call Data

(b) BIC score plot for Citizens Connect Data

3 Computing Point Estimates

We compute point estimates, maximum likelihood and maximum a posteriori estimates, for our model via expectation maximization. We note that EM algorithms are sensitive to initialization. For example, if the means of the components are initialized in such a way where some component is initially assigned few or no points, this will result in the estimates for Σ being singular for the next iteration of the algorithm. To address the problem of choosing appropriate initializations, we will use the following initialization regime:

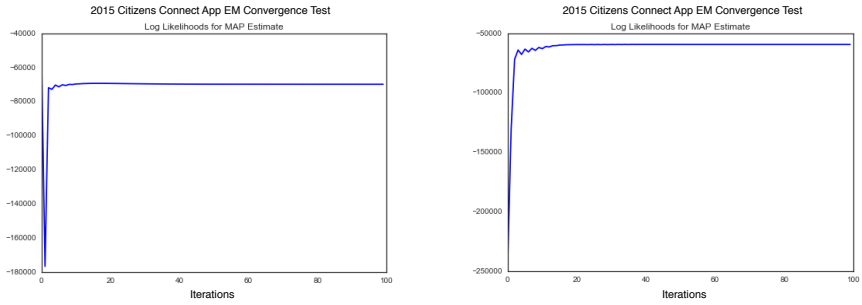
1. For MLE, we initialize the parameters μ, π, Σ using the means, mixing coefficients and scatter matrices for the components obtained from K-means or simulated annealing.

2. For MAP, we initialize the hyper-parameters of our model randomly and initialize the parameters μ, π, Σ using MLE estimates.

Observing that MLE tends to overfit for Gaussian mixture models, in the end, we will use the MAP estimate for our cluster analysis (the EM update equations for MLE and MAP are found in the ipython notebook `EM for MLE and MAP`).

Convergence of the algorithms is checked by generating log-likelihood plots (against the number of iterations of the algorithm). The log-likelihood plots for MAP estimates on constituent call and app data is shown in Figure 4. The log-likelihood plots for MLE on the data is included in the ipython notebook `EM for MLE and MAP`.

Figure 2: Convergence of EM Algorithms



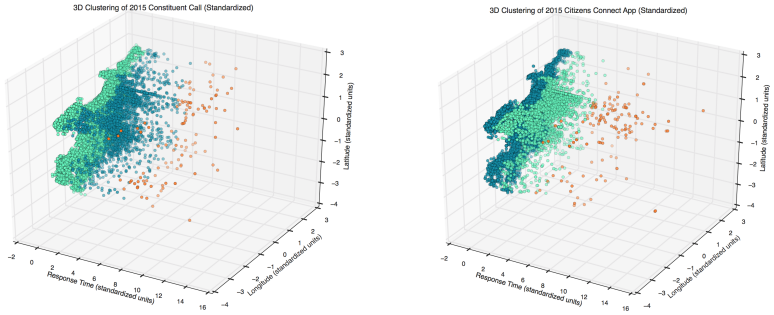
(a) Log Likelihood for MAP (Call Data) (b) Log Likelihood for MAP (App Data)

4 Cluster Analysis

Based on our MAP estimate of the model parameters, we produce a hard clustering of our data - each data point, $\mathbf{x}_n$, is attributed to the cluster, k for which $\pi_k p(\mathbf{x}_n|\mu, \Sigma)$ is the largest, over $1 \leq k \leq K$. The clustering is visualized on axes:

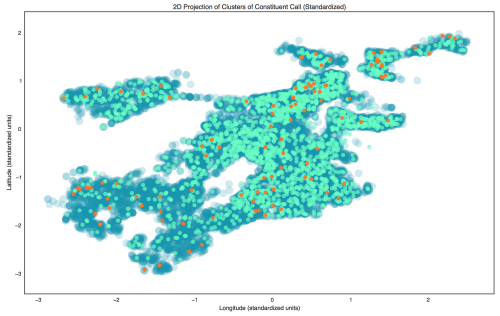
$$x = \text{response time (standardized)}, \quad y = \text{longitude (standardized)}, \quad z = \text{latitude (standardized)}. \tag{3}$$

Figure 3: Clustering from Point Estimates

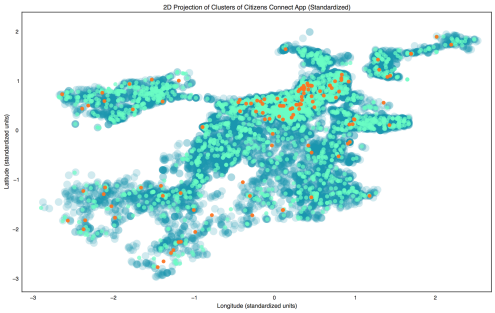


(a) MAP Clustering (Call Data) (b) MAP Clustering (App Data)

Figure 4: Clustering from Point Estimates

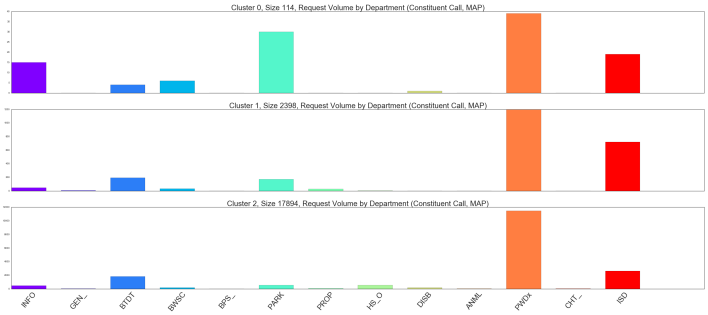


(a) MAP Clustering (Call Data)

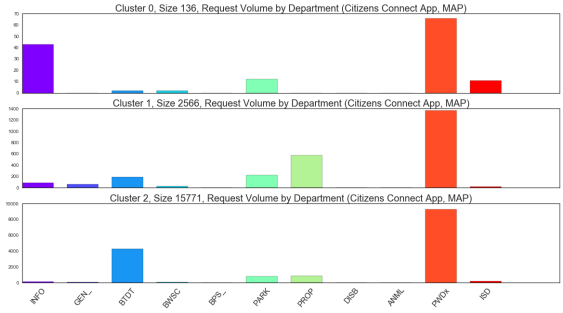


(b) MAP Clustering (App Data)

Figure 5: Clustering from Point Estimates



(a) MAP Clustering (Call Data)



(b) MAP Clustering (App Data)

References

[1] C. Biernacki and G. Govaert. Choosing models in model-based clustering and discriminant analysis. *J. Stat. Comput. Sim.*, 1(64):4971, 1999.