Taxi Trip Density Estimation in NYC*

Roger Fan†

May 26, 2016

1 Introduction

Demand estimation is a crucial problem for taxicab companies (and potentially other ride-share services), allowing them to more effectively plan and adjust deployments. Using a dataset of New York City taxi trip data, we show how clustering and density estimation techniques can be applied to this problem. Although taxi trip originations are not perfect indicators of demand, as a completed trip requires both the demand for a trip and a taxi available to supply it, we use taxi origination locations to hopefully proxy for taxi demand.

We use a Gaussian mixture model (GMM) to estimate the density of trip origination locations in New York City. This method allows us to effectively condense the information from millions of taxi trip originations to a manageable number of Gaussian distributions over the city.

We expect, however, that the demand for taxis changes over time, a factor that is vital to many applications of this analysis. To better handle this issue, we design a extension to the standard Gaussian mixture model that allows for time-varying mixing weights and use an Expectation-Maximization (EM) algorithm to estimate it.

2 Data

The NYC Taxi and Limousine Commision (TLC) provides extensive data on taxicab trips in New York City for the last several years.[1] This data includes both yellow cabs, which primarily pick up street hails in Manhattan and at the airports, and green cabs, which can only be hailed in northern Manhattan and the outer boroughs.

*2016 Applied Statistics Qualifying Exam report.

†rogerfan@umich.edu

[1]Available at http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml.

The primary variables of interest are the time and location of each trip's pickup and dropoff. Location data is encoded as latitude and longitude, which we will be able to use directly, and time data is recorded to the second, though we will primarily be using data at the hourly frequency in this analysis.

To simplify estimation and computation, we will focus on taxi rides conducted during a single week, December 7-10, 2015. We omit Friday-Sunday of the week to avoid complicating the analysis with weekend data, and so are left with four days of weekday data, consisting of just over 1.7 million individual taxi rides. Of these, around 25 thousand are missing location data, which is a small enough proportion that it is safe to simply omit them. And finally, we omit around 100 outlier trips that have nonsensical or extreme pickup or dropoff locations, leaving a final training dataset of 1,688,673 taxi trips.

3 Density Estimation

We can consider approximating taxi demand to be a density estimation problem, where each trip's pickup location is a draw from an underlying distribution. Estimating taxi demand then simply becomes a problem of estimating the (2-dimensional) density of taxi originations.

3.1 Gaussian Mixture Model

In order to estimate the density, we will assume that joint density for X comes from a J-component Gaussian mixture model and maximize the likelihood using the EM algorithm, which has been the standard estimation procedure for mixture models since Dempster et al. (1977) first introduced the EM algorithm. We introduce a latent variable Z that indicates group membership and have a generative model defined by

$$\begin{aligned} Z &\sim \text{Categorical}(\pi_1, \ldots, \pi_J) \\ X \mid Z = j &\sim \text{Gaussian}(\mu_j, \Sigma_j) \end{aligned} \tag{1}$$

The EM algorithm updates are therefore as follows.

1. E-step: Estimate the group probabilities $p_{ij}^{(t+1)}$ conditional on the current parameter estimates.
$$p_{ij}^{(t+1)} = \frac{\pi_j^{(t)} P(X_i \mid \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{j'=1}^{J} \pi_{j'}^{(t)} P(X_i \mid \mu_{j'}^{(t)}, \Sigma_{j'}^{(t)})} \tag{2}$$
2. M-step: Estimate the parameters using maximum likelihood estimators conditional on

the group probability estimates.

$$\begin{aligned}
\pi_j^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n} p_{ij}^{(t+1)} \\
\mu_j^{(t+1)} &= \frac{\sum_{i=1}^{n} p_{ij}^{(t+1)} X_i}{\sum_{i=1}^{n} p_{ij}^{(t+1)}} \\
\Sigma_j^{(t+1)} &= \frac{\sum_{i=1}^{n} p_{ij}^{(t+1)} (X_i - \mu_j^{(t+1)})(X_i - \mu_j^{(t+1)})^T}{\sum_{i=1}^{n} p_{ij}^{(t+1)}}
\end{aligned} \tag{3}$$

3.2 Time-Varying Mixing Components

However, the model described in Equation 1 has a major weakness: it does not use temporal information. We expect that the demand for taxis could vary significantly over different time periods. To illustrate this, Figure 1 plots taxi originations for two time periods, 7am-9am and 10am-12am.[2] Even using just the raw data, we can visually identify several areas that exhibit differences. In particular, the morning taxi demand seems to be higher in southwest Brooklyn (-74.00, 40.68), northern Queens (-73.92, 40.76), the Bronx (-73.90, 40.82), and northern Manhattan (-73.95, 40.81), while nighttime demand is higher in Williamsburg in northern Brooklyn (-73.94, 40.71). The morning periods in particular are as we might expect, all communities that commute into the city during morning rush hour.

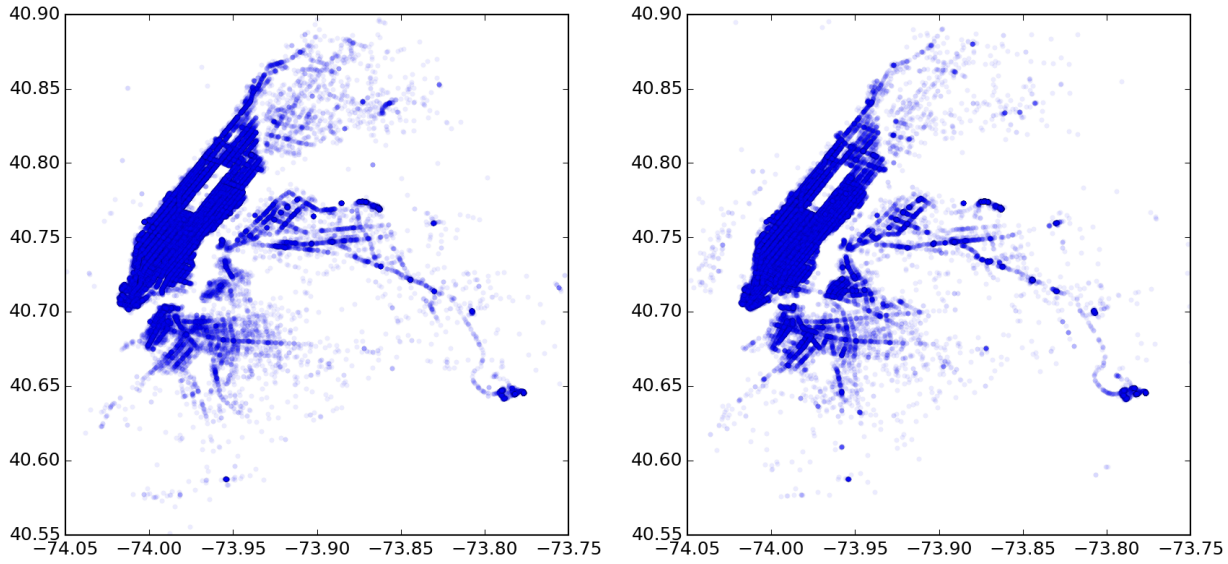
This motivates extending the Gaussian mixture model to incorporate temporal information. We will assume that the component distributions are the same over time, as if clusters are approximating neighborhoods then it seems reasonable to assume that the location and shape of each cluster is constant. But we will allow the mixing proportions to vary between time periods, allowing the demand from each neighborhood to shift over time. This, for instance, allows for phenomenon such as commuter neighborhoods that have high demand during rush hour but little during the evening. It will also hopefully allow us to identify clusters that are only visible during specific time periods.

To account for this, we use an additional observed variable $Y \in \{1, \ldots, K\}$ that tracks the time period of each obervation X. We will assume that Y is fixed and exogenous. Then our model becomes

$$\begin{aligned}
Z \mid Y = k &\sim \text{Categorical}(\pi_{k1}, \ldots, \pi_{kJ}) \\
X \mid Z = j &\sim \text{Gaussian}(\mu_j, \Sigma_j)
\end{aligned} \tag{4}$$

The EM algorithm is similar to the one described in Section 3.1, replacing the group proba-

[2]These subsamples have approximately the same number of observations.



(a) 7am-9am

(b) 10pm-12am

Figure 1: Taxi originations during morning rush hour (7am-9am) and at night (10pm-12am).

bility update in Equation 2 with

$$p_{ij}^{(t+1)} = \frac{\pi_{Y_i j}^{(t)} P(X_i \mid \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{j'=1}^{J} \pi_{Y_i j'}^{(t)} P(X_i \mid \mu_{j'}^{(t)}, \Sigma_{j'}^{(t)})} \tag{5}$$

And replacing the mixing component update in Equation 3 with

$$\pi_{kj}^{(t+1)} = \frac{\sum_{i=1}^{n} \mathbb{1}(Y_i = k) p_{ij}^{(t+1)}}{\sum_{i=1}^{n} \mathbb{1}(Y_i = k)} \tag{6}$$

Note that the advantage of dividing time into a (small) number of categories and estimating each category's mixing components separately is that it does not add much computational complexity over the standard GMM model. A richer model might be to assume the mixing proportion for each cluster varies smoothly over time and then use a local averaging or kernel regression method instead of Equation 6, but this would add a significant computational burden.

4 Results

Figure 2 shows the estimated density of the basic GMM model (Equation 1) with $J = 30$ components as well as each of the cluster means. We can see that the contours of the estimated density visually correspond well to the raw origination data. Several of the

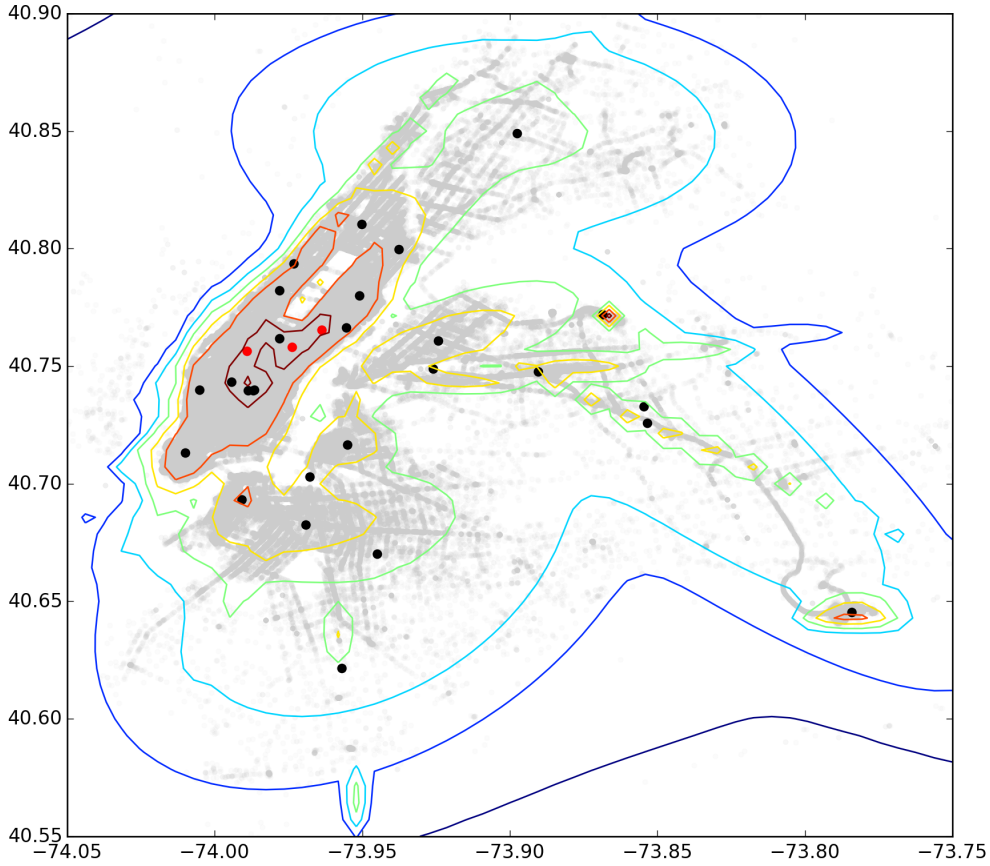


Figure 2: Estimated cluster centers and density of NYC taxi originations using a Gaussian mixture model with $J = 30$ components.

important features of NYC are identifiable, including the overall shape and size of the outer boroughs, the locations of JFK and LaGuardia airports (the two hotspots to the east), and the overall shape of Manhattan and Midtown, including a cool spot in Central Park. The red centers indicate the three clusters with the highest estimated mixing proportions, which together make up around 37% of originations.

Though this is effective for summarizing and visualizing the data and captures many of the relevant features of the data and city, it is not particularly effective for planning deployments or estimating demand since it does not allow for time-dependent predictions. Therefore, our next step is to estimate a model that incorporates temporal information.

To estimate the GMM model with time-varying mixing proportions described in Equation 4, we first need to divide time into categories. Attempting to adhere to common-sense work day divisions, we propose using six categories: early morning (2am-7am), morning rush hour (7am-9am), work day (9am-4pm), evening rush hour (4pm-6pm), evening (6pm-10pm), and night (10pm-2am). Figure 3 shows the frequency of taxi pickups over time as well as

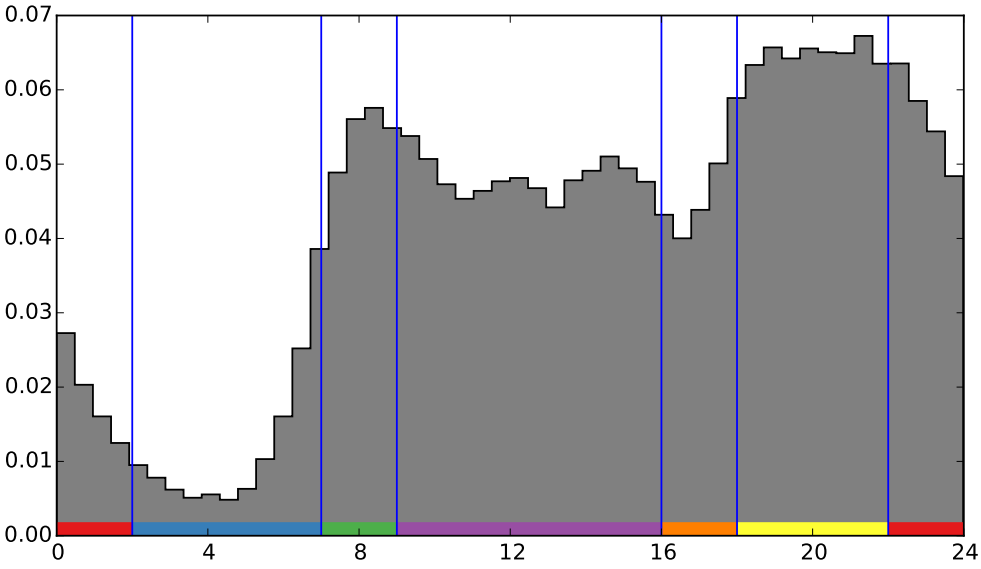


Figure 3: Taxi pickups over time with proposed time categories.

these proposed categories. We can see that the categories seem to correspond well to natural divisions in data, with boundaries that roughly track change points in the frequency over time.

Figure 4 shows the estimated centers for this model. We can see that the centers outside of Manhattan are nearly identical to those in Figure 2, but that the centers inside Manhattan are fairly different. It is notable that the estimated component distributions are noticeably different than the basic GMM model, hopefully adding time-varying mixing proportions is allowing the model to identify previously difficult-to-separate components.

Table 1 shows the evolution of mixing proportions for a subset of the clusters. We can see that there is significant variation across time for many of the clusters, and that the patterns can also be very different. JFK (cluster 5) and LaGuardia (cluster 6) have similar patterns over time except for in the early morning, where JFK has a relatively high percentage of taxi originations and LaGuardia has almost none. Or consider southwest Brooklyn (cluster 7), which is primarily busy during rush hour and the work day, compared to Williamsburg (cluster 8), which has more originations at night and in the early morning and very few during the day. Clusters in Mahattan can also be very different, as the Lower East Side (cluster 1) seems to be a night-life area, while the Upper East Side (cluster 0) has relatively few originations at night. Also note that cluster 1 is one of those that the basic GMM model was unable to find; it's unique temporal pattern allows this model to identify it.

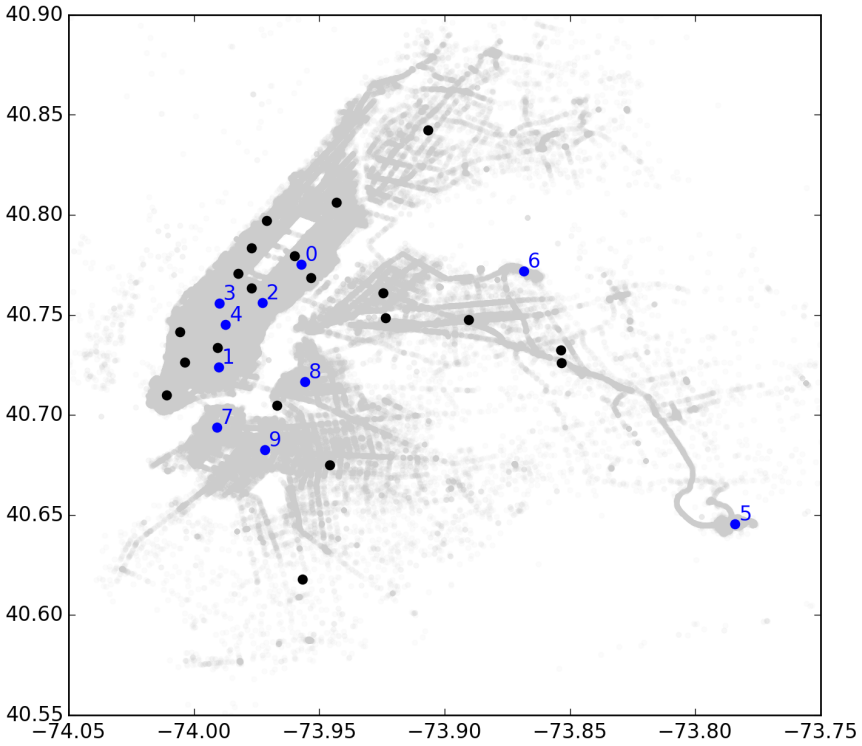


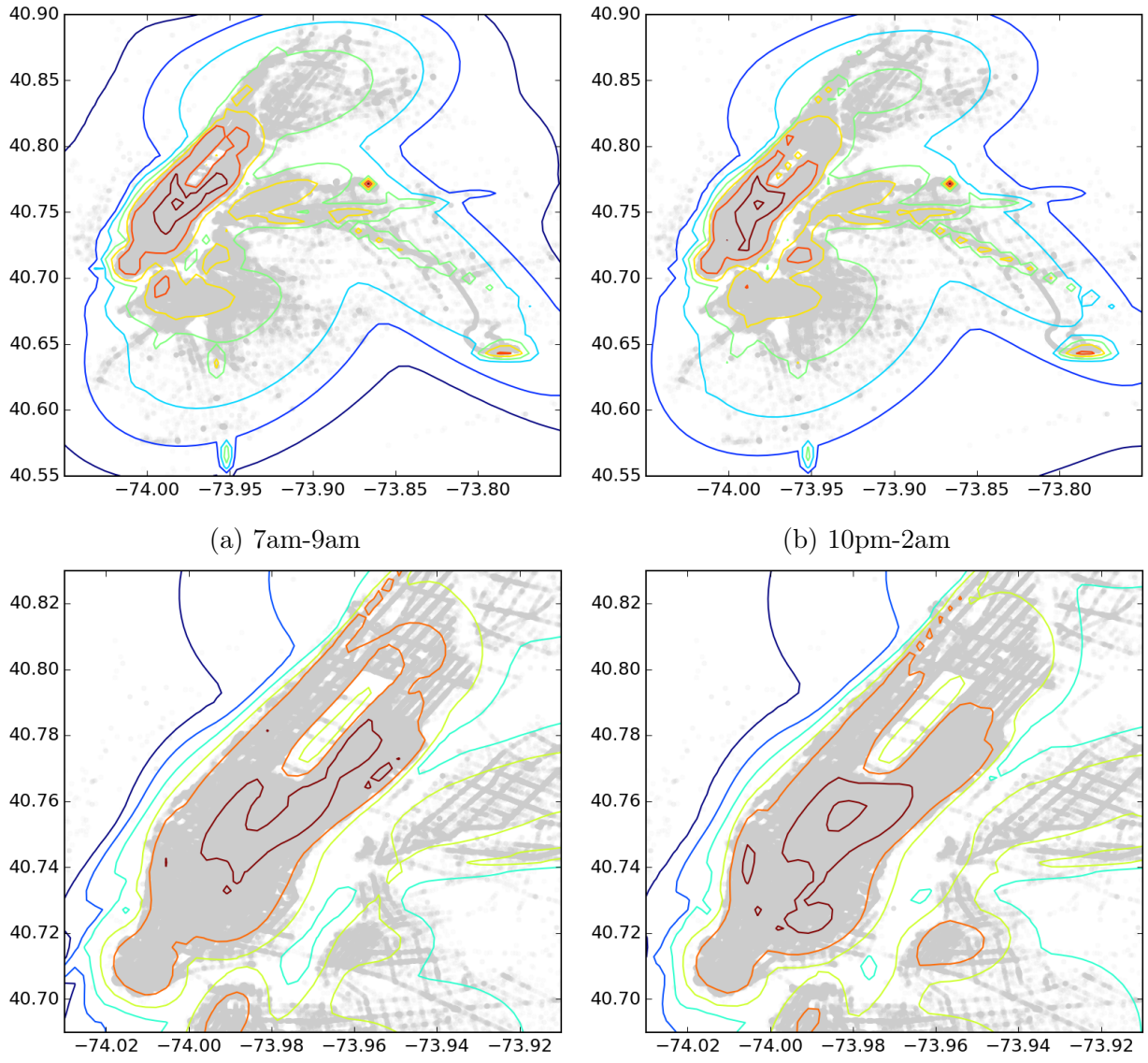
Figure 4: Estimated cluster centers of NYC taxi originations using a Gaussian mixture model with time-varying mixing proportions. $K = 30$ components and $K = 6$ time categories are used.

	Manhattan					Airports		Boroughs		
	0	1	2	3	4	5	6	7	8	9
02:00-07:00	8.10	5.72	8.97	14.38	17.72	2.09	0.16	1.18	1.45	1.62
07:00-09:00	8.53	1.82	11.91	8.56	18.40	1.27	1.87	1.86	0.23	1.58
09:00-16:00	7.19	1.83	11.95	6.77	18.91	1.72	3.58	1.20	0.25	1.23
16:00-18:00	8.06	1.68	9.86	5.64	15.20	2.64	3.73	1.54	0.36	1.59
18:00-22:00	6.78	3.76	12.51	7.14	17.64	1.75	2.54	1.30	0.73	1.71
22:00-02:00	4.19	8.61	12.23	8.55	18.02	1.69	1.96	0.97	2.04	1.92

Table 1: Estimated mixing proportions for a subset of clusters.

Figure 5 compares the estimated distributions for morning rush hour and late-night. We can see that many of the same patterns visible in Figure 1 are also clear here, where the northern and southern boroughs have more originations in the morning while Williamsburg has more originations at night. But using the estimated distributions, we can also identify patterns in Manhattan that were impossible to see with the raw data. For instance, we can clearly see the nighttime hotspot in southern Manhattan that is the Lower East Side, and

we can see how northern Manhattan and the Upper East Side have fewer night originations.



(a) 7am-9am

(b) 10pm-2am

(c) 7am-9am

(d) 10pm-2am

Figure 5: Estimated trip origination densities for two time periods using GMM with time-varying mixing proportions.

4.1 Evaluating Fit

In order to evaluate the fit of these models, we use a test dataset consisting of 426,194 observations from December 15, 2015. We consider three models. The first is the standard GMM model described in Equation 1. The second, which we call subset-GMM (ssGMM), is the same GMM model estimated separately for each of the six time category subsets. And

the third is the GMM model with time-varying mixing proportions (tvGMM) as described in Equation 4. Note that GMM is nested inside tvGMM which is in turn nested in ssGMM.

Table 2 presents diagnostics for these three models. For each model, we calculate the log-likelihood contribution for each time period of the test data, as well as the total test log-likelihood. We also calculate the Bayesian information criterion (BIC) for each model on the training data to evaluate in-sample performance.[3]

	GMM	ssGMM	tvGMM
02:00-07:00	125126	**128015**	127873
07:00-09:00	235173	236006	**237738**
09:00-16:00	773825	769287	**780562**
16:00-18:00	211914	210956	**214326**
18:00-22:00	601819	600960	**606997**
22:00-02:00	292844	296174	**298605**
Total log-lik	2240702	2241399	**2266101**
In-Sample BIC	-17732572	-17732375	**-17934791**

Table 2: Out-of-sample log-likelihoods and in-sample BIC.

We can see that, overall, tvGMM has significantly better out-of-sample fit than either GMM or ssGMM. In fact, tvGMM outperforms ssGMM and GMM on all but one of the subsets considered. ssGMM and GMM have very similar out-of-sample performance in terms of log-likelihood. In terms of BIC, tvGMM again clearly has the best performance, with GMM and ssGMM performing similarly. It seems that the penalty for the many more free parameters in ssGMM outweighs any performance gains it might have.

tvGMM seems to provide a much more effective and parsimonious way allow for variation across time periods while sharing information when appropriate. It performs significantly better than both GMM and ssGMM in terms of in-sample and out-of-sample diagnostics.

5 Conclusion

In an attempt to proxy for taxi demand in NYC, we have used data on taxi trip origination locations and times to conduct density estimation using Gaussian mixture models. Standard GMMs can effectively estimate the underlying density of taxi originations and recover features of the city and surrouding boroughs, but do not incorporate time, which is an essential factor for many possible applications of this analysis. In order to remedy this, we

[3]BIC is calculated as $p \log n - 2 \log L$, where p is the number of estimated parameters, n is the number of observations, and L is the maximized likelihood of the model. Lower BICs indicate more parsimonious models.

design a computationally feasible extension to the GMM that allows for time-varying mixing proportions and present a modified EM algorithm to estimate this model. This model effectively recovers temporal patterns in the density and both its in-sample and out-of-sample performance are superior to both the standard GMM and separate GMMs estimated on each subset.

One improvement to the analysis would be to use data-driven methods to determine the hyperparameters. Due to computational concerns, hyperparameters such as the number of clusters and the time categories were simply chosen, instead of being optimized. Incorporating model selection based on cross-validation, AIC/BIC, or similar techniques could improve performance.

As briefly mentioned in Section 3.2, a possible extension to this model is to allow for the mixing proportions to be an arbitrary smooth function of time instead of effectively a step function. For some smooth functions $\pi_1, \ldots, \pi_J$, the corresponding model is then described by:

$$\begin{aligned} Z \mid Y = y &\sim \text{Categorical}(\pi_1(y), \ldots, \pi_J(y)) \\ X \mid Z = j &\sim \text{Gaussian}(\mu_j, \Sigma_j) \end{aligned} \tag{7}$$

With a modified EM algorithm where the probability estimates in the M-step are replaced with Nadaraya-Watson estimators (as described in Chapter 6 of Hastie et al., 2001).

$$\pi_j^{(t+1)}(y) = \frac{\sum_{i=1}^n K(y, Y_i) p_{ij}^{(t+1)}}{\sum_{i=1}^n K(y, Y_i)} \tag{8}$$

For some kernel function K (other non- or semi-parametric regression estimates could also likely be used). This is a much richer model that can estimate complicated temporal patterns, and applications like planning taxi deployment would likely benefit from the finer temporal resolution. However, the regression step within each EM iteration makes this algorithm computationally difficult when applied to large datasets such as this one.

Another weakness of GMM-based models is that New York is a city with strong structure, where possible pickup locations are restricted by various parks, waterways, highways, legal boundaries, etc. Gaussian distributions are not well-suited to estimating these kinds of hard boundaries. Kernel density estimators with kernel warping such as those proposed by Zhou and Matteson (2015) could be used to simultaneously respect these structures and share information over time, but doing so in a way that scales computationally could be challenging.

Finally, when it comes to applications like taxi deployment, the number of trips is not the only concern. Factors such as trip destination, length, and overall fares are all potentially

important as well, and data on all these factors is also provided by the TLC. Incorporating this additional information would be important to move from simply modeling demand to modeling potential revenue or profit.

References

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Hastie, T., Tibshirani, R., , and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York, NY.

Zhou, Z. and Matteson, D. S. (2015). Predicting Melbourne ambulance demand using kernel warping. arXiv:1507.00363.