Capital Bikeshare Station and Ride Analysis

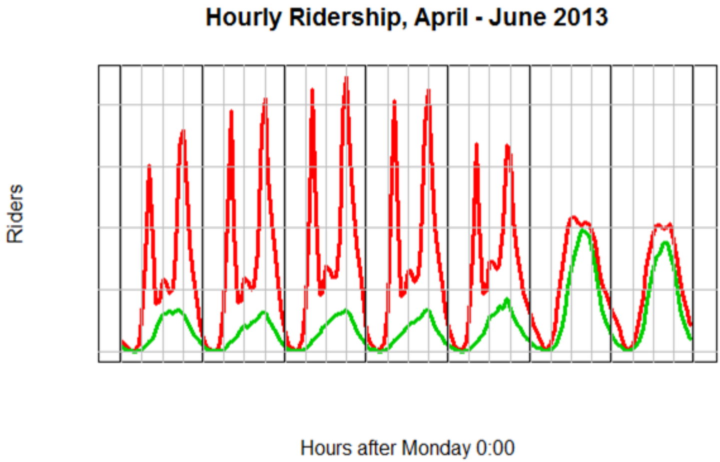
March 26, 2014

Bikeshare Station

The System

- Started in 2010. As of 4th quarter 2013 there are over 300 stations and 750,000 rides/quarter
- Data made freely available with start and end date, time, and station, and rider type
- `0h 5m 41s, 6/30/2013 23:51, Florida Ave & R St NW, 31503, 6/30/2013 23:56, 5th & K St NW, 31600, W01380, Subscriber`
- Large Scale system concerns, other systems

Registered vs. Casual Riders



Our Work

- Applying Expectation-Maximization algorithm to the ride data for different models in R Statistical Software
- Want to cluster ride data by a latent variable to analyze different variables (rider type, start/end station, time of day, start and end station pairs)
- Identify traffic flow, similar stations, ridership patterns
- Following similar work done on Velib‘ system in Paris

Expectation Maximization - 1

- Given data $(x_i, z_i) \sim f(x, z; \lambda)$, where λ is an unknown parameter vector
- Can estimate λ, using e.g. maximum likelihood
- **What if the z_i are unobserved?**
- Try to estimate λ from just the x_i
- Try to find expected values of the z_i as well

Expectation Maximization - 2

- Iterative procedure (Expectation step followed by maximization step) that is proven to converge
- Expectation (E) step: Compute (update) the expected value of the unobserved z_i, given the data x_i and the current value of λ
- Maximization (M) step: Maximize the log likelihood function to compute λ, given the data and the current expected z_i values

Model I: Clusters of Stations

- Station number $1 \leq i \leq N \approx 230$, time $t \in \{0, \ldots, 23\}$, day $d \in \{1, \ldots, 91\}$, cluster ℓ
- X_{itd} = start (or end) count at station i at time t on day d
- $Z_{i\ell} = 1$ iff station i is in cluster ℓ $Z_{i\ell} = 0$ otherwise

Model I Assumptions

- Conditioned on $Z_{i\ell} = 1$, assume $X_{itd} \sim \mathfrak{P}(\alpha_i \cdot \lambda_{\ell t})$ (Poisson)
- Suitable independence assumptions
- α_i = mean hourly count at station i
- $\lambda_{\ell t}$ = relative hourly intensities for cluster ℓ

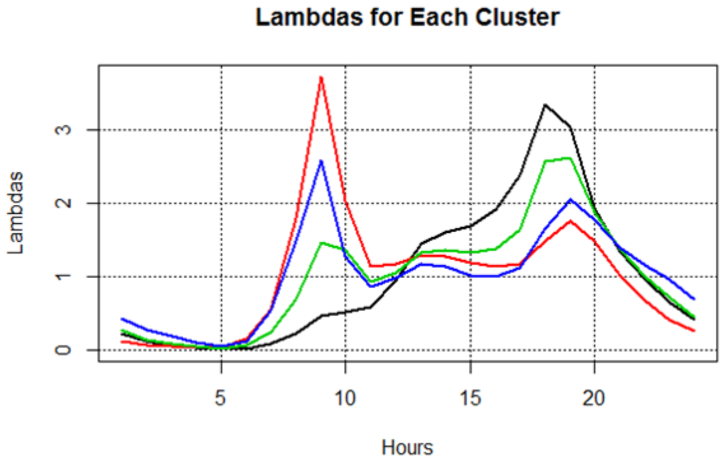
EM Algorithm for This Case

- Only need to update the $\lambda_{\ell t}$ and the expected values for $Z_{i\ell}$. The α_i are computed only once.
- E-Step: Calculate expected values of $Z_{i\ell}$ with Bayes Rule, using the data X_{itd} and the current estimates of the $\lambda_{\ell t}$.
- M-Step: Maximum likelihood estimate of the $\lambda_{\ell t}$, using the data X_{itd} and expected values of $Z_{\ell t}$

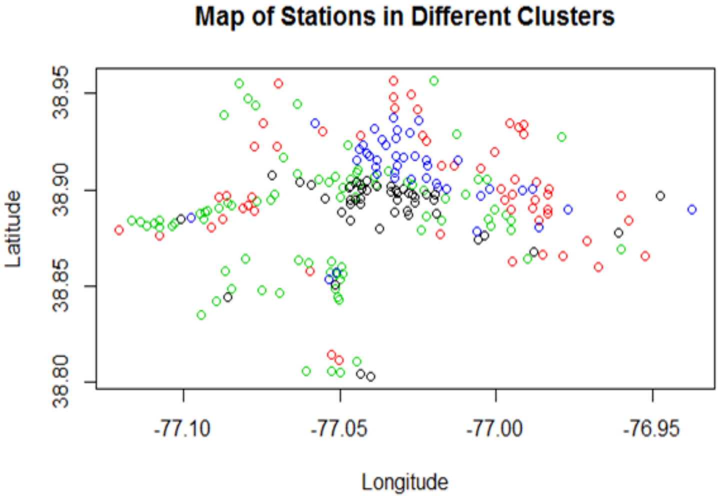
Algorithm Implementation

- Simplify update equations to run in R quickly
- Use matrix algebra and array manipulation
- Implementation very efficient even on personal laptop

Results for Model I: Intensities



Results for Model I: Clusters



Model II: Clusters of Stations

- Start and end station i, j, time t, day d, cluster ℓ
- X_{ijtd} = ride count from station i to j at time t on day d
- $Z_{ij\ell} = 1$ iff station pair (i, j) is in cluster ℓ, $Z_{ij\ell} = 0$ otherwise

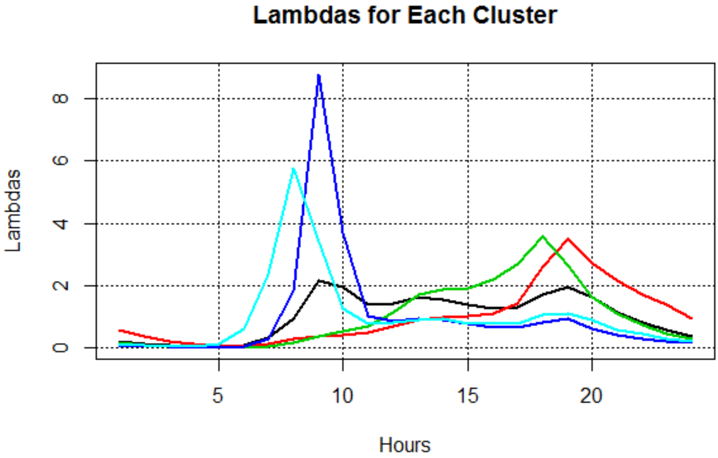
Model II Assumptions

- Conditioned on $Z_{ij\ell} = 1$, assume $X_{ijtd} \sim \mathfrak{P}(\alpha_{ij} \cdot \lambda_{\ell t})$ (Poisson)
- Suitable independence assumptions
- α_{ij} = mean hourly ride count from station i to station j
- $\lambda_{\ell t}$ = relative hourly intensities for cluster ℓ

Model II: Clusters of Station Pairs

- Clusters according to Station-Station pair ride count
- EM Algorithm applied to ride data with station pair scaling factor α_{ij}
- There are many station pairs with few or no rides between them
- These station pairs cannot be assigned to a cluster (there is no estimated $Z_{ij\ell}$ that is close to 1)

Results for Model II: Intensities



Trips from 5th and F Street (Gallery Place)