Classifying Riders

Given the results from last chapter, there is a clear need to understand what kinds of riders are in this data set. Ride Report, because they need to respect the privacy of their users, cannot identify individual riders in the data they provide to clients, yet our model results show that differentiating riders is crucial to getting good estimates in our models. However, there is one potential solution to this problem: If we can identify clusters of riders in the data set that give us nearly the same information as grouping by individuals did, these could be provided by Ride Report without nearly the same level of risks to user privacy as identifying individual riders.

In this chapter, we identify predictors that differentiate types of riders and then use these variables to identify clusters of riders. To tie these predictors into our model, we test to see how these predictors do as rider-level predictors for the intercepts and even some coefficients. We also compare random intercepts with riders to random intercept models done by cluster.

Characterizing riders

By aggregating observations in the data set, we can compute for each rider:

- Frequency of rides, in rides per week
- Proportion of rides on weekends
- Patterns in time of day on weekdays
- Patterns in ride length

The last two are a little vague. How can we compute variables that describe patterns of rides? Summary statistics like mean and variance can help, but really don't give a great description of rider patterns. What use would mean time of day be? Instead, we can transform time of day and length patterns into high dimensional variables and then do principle component analysis (PCA) to create summaries that best describe those distributions. To maintain some interprebility, we keep our PCA analysis for time of day and length patterns separate.

Characterizing Rider Length and Time Patterns

We want to extract features from riders' time and length patterns that will define a feature space in which we can easily identify useful clusters of riders. There are many ways to identify such features. The approach we will take here follows the heuristic of capturing the information in a histogram of each rider's time of day pattern.

Define the number of rides in hour $h \in \{0, \dots, 23\}$ for rider j as

$$n(h, j) = \left|\left\{i \mid X_i^{\text{rider}} = j, \lfloor t_i \rfloor = h\right\}\right|$$

(Recall that $t \in [0, 24)^n$, i.e it is measured in hours since midnight.)

We can then define the rider time of day matrix as P, where

$$P_{ij} = \frac{n(i, j)}{\sum_{h=0}^{23} n(h, j)}.$$

Now we can ask, what are the principal components of these column vectors?

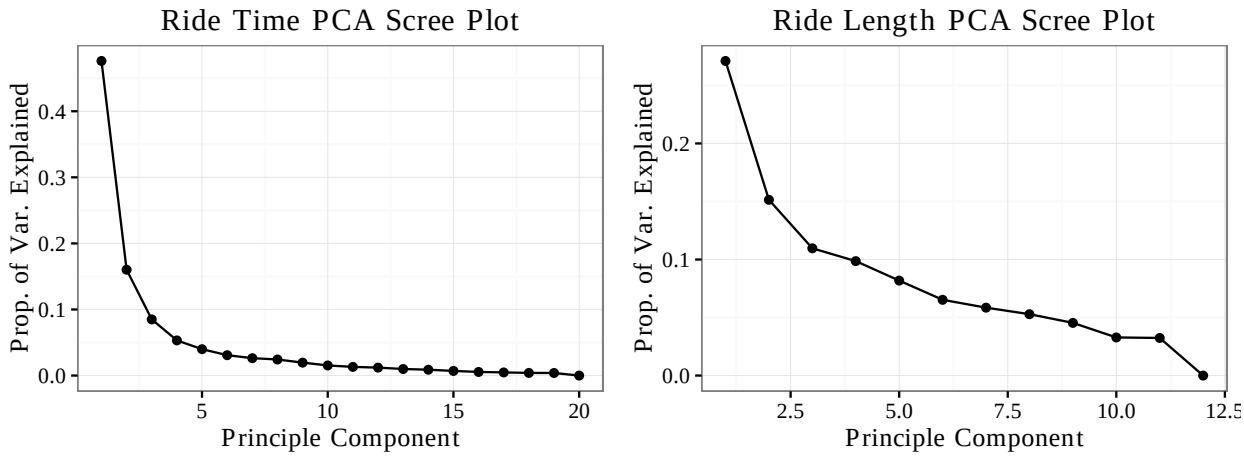

Figure 1: Scree plots for time of day and ride length principal component analyses.

Models with Rider-level predictors

We now have several variables that differentiate riders. Howe well do they predict our rider intercepts?

Now let U_j be the vector of rider-level variables. Then our model will be

$$Y_i \sim \text{Bernoulli}\left(\text{logit}^{-1}(\alpha_{j[i]} + X_i\beta)\right), \tag{1}$$

where,

$$\alpha_j \sim N(\gamma_0 + U_j\gamma) \tag{2}$$

This model should be comparable to Model 2.[1]

Cluster Intercepts Versus Rider Intercepts

Would our clusters be

[1]Though we would prefer to use a model similar to Model 4 from the previous chapter (the one with smoothing splines for time of day) the current additive mixed models package `gamm4` (which uses `lme4` to fit the mixed models part) does not support estimating the variability in group-level estimates. Instead, we fit these models in Stan, a probabilitic modeling language that does full Bayesian statistical inference with Markov-chain Monte Carlo sampling. Unfortunately, in this package, smoothing splines would have to be coded by hand and we lacked the expertise to write the functions to fit smoothing splines ourselves.

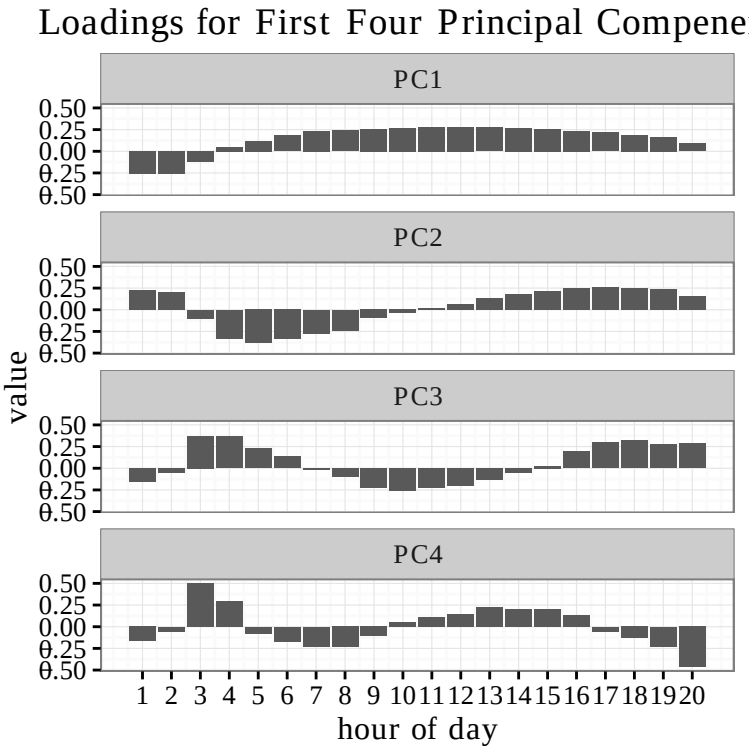

Figure 2: Loadings of first four principal components for time of day.

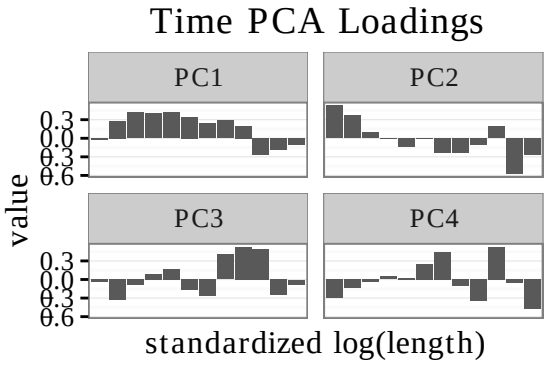

Figure 3: Loadings of first four principal components for ride length.

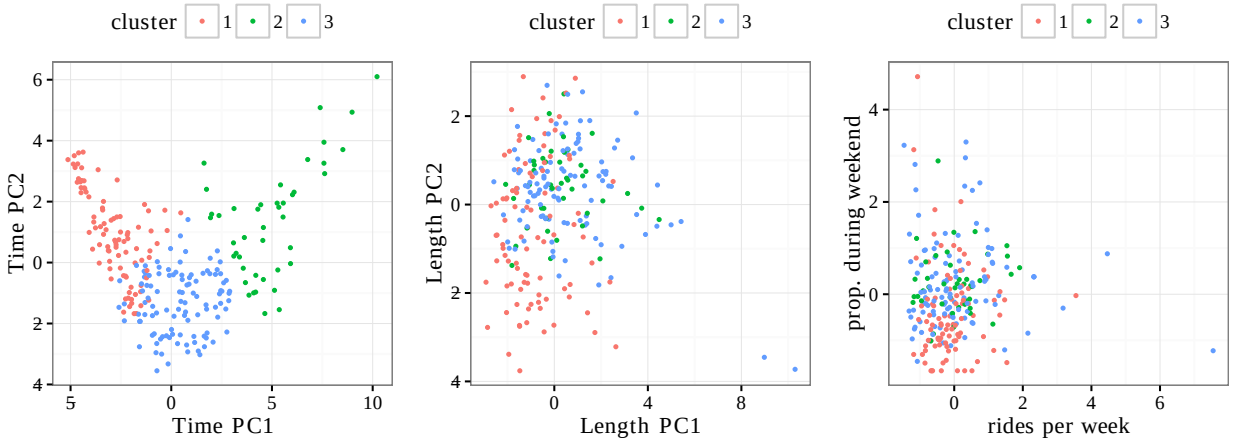

Figure 4: Scatter plots of clustering variables with 3 clusters identified by k-means clustering.