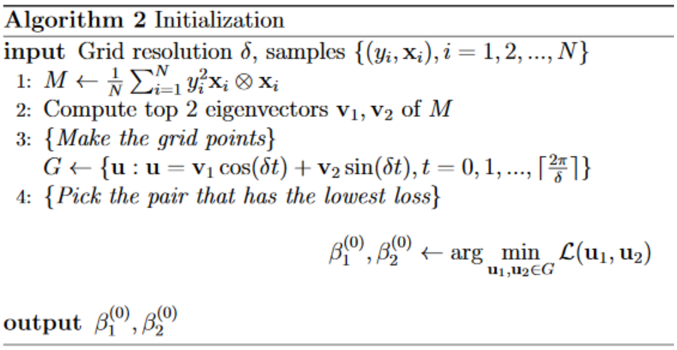
DS-GA 1013 Project Proposal

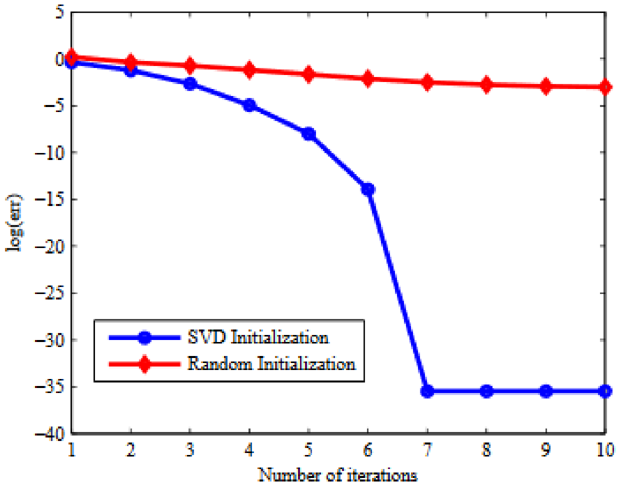
Andrew Hopen and Andrei Kapustin

For our project, we would like to explore a practical application of the method proposed in the paper "Alternating Minimization for Mixed Linear Regression" by Yi et al [1]. In mixed linear regression, we model some continuous outcome variable as a linear combination of features, with the additional complication that each sample belongs to one of multiple latent classes, and those classes have different regression parameters. Mixed linear models are fit with the EM algorithm: we start by randomly assigning latent membership and randomly initializing the parameter vectors, then we alternate between re-assigning the classes and re-estimating the regression parameters until convergence. Yi et al propose an improvement to the initialization step. Rather than starting with random parameter vectors, they construct a matrix M whose eigenvectors are unbiased estimates of the true parameter vectors (Fig 1a, steps 1-2). Then they search over the unit circle in the space spanned by those eigenvectors for better options, since the true parameter vectors need not be orthogonal (Fig 1a, steps 3-4). Yi et al show that for a constructed dataset, this initialization procedure dramatically speeds up convergence (Fig 1b).



```
Algorithm 2 Initialization
input Grid resolution $\delta$, samples $\{(y_i, \mathbf{x}_i), i = 1, 2, ..., N\}$
1: $M \leftarrow \frac{1}{N}\sum_{i=1}^{N} y_i^2 \mathbf{x}_i \otimes \mathbf{x}_i$
2: Compute top 2 eigenvectors $\mathbf{v}_1, \mathbf{v}_2$ of $M$
3: {Make the grid points}
   $G \leftarrow \{\mathbf{u} : \mathbf{u} = \mathbf{v}_1 \cos(\delta t) + \mathbf{v}_2 \sin(\delta t), t = 0, 1, ..., \lceil \frac{2\pi}{\delta} \rceil\}$
4: {Pick the pair that has the lowest loss}
   $\beta_1^{(0)}, \beta_2^{(0)} \leftarrow \arg \min_{\mathbf{u}_1, \mathbf{u}_2 \in G} \mathcal{L}(\mathbf{u}_1, \mathbf{u}_2)$
output $\beta_1^{(0)}, \beta_2^{(0)}$
```

(a) Initialization Algorithm



(b) Convergence Speedup

Figure 1: Yi et al Method

We plan to test this procedure on publicly available NYC Taxi Data [2]. This data includes information on all taxi rides in NYC. For example, Figure 2 shows a subset of "green taxi" (outer borough) rides in January 2019. We have color-coded the trips by "RateCode" and shown their corresponding best-fit lines. We see, for example, that trips from the airport JFK have a flat rate of $52, whereas standard rides show a slope of $2.29 per mile.

We would like to run the mixed linear regression procedure on this data to see how well we can recover RateCodes and the regression parameters, and to assess speedup when applying the proposed initialization method. We will run these tests for different numbers of covariates and different numbers of latent RateCodes, as permitted by the data.

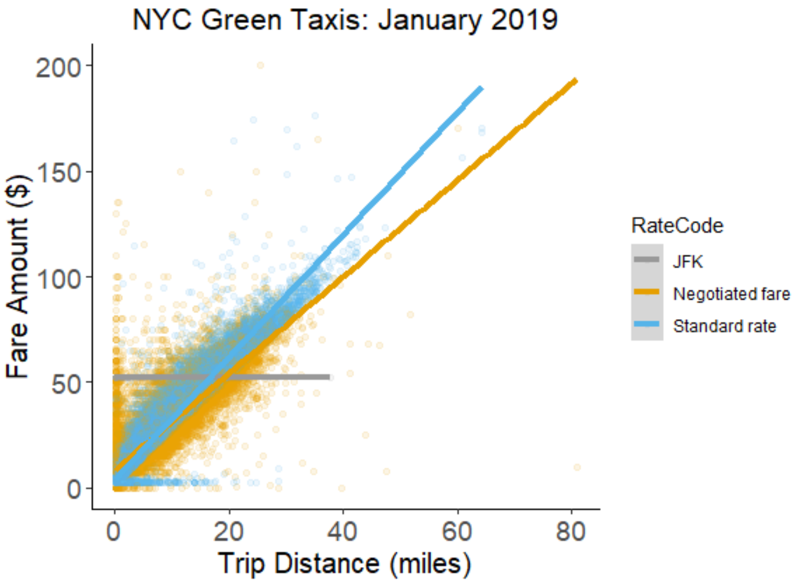


Figure 2: Real-World Mixed Regression

Bibliography

[1] Xinyang Yi, Constantine Caramanis, Sujay Sanghavi. *Alternating Minimization for Mixed Linear Regression*, 2013 https://arxiv.org/pdf/1310.3745v1.pdf

[2] NYC Taxi Data. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page