Empirical Analysis of Initialization Procedures for Fitting Mixed Linear Regression

Andrei Kasputin (ak7671) Andrew Hopen (ah182)

1. Introduction

In linear regression, some response variable is assumed to be a (possibly noisy) affine combination of measured covariates. In mixed linear regression, the same is true, with a twist. Data are assumed to belong to one of multiple latent classes, with each class having its own affine relationship between covariates and response. For example, in Figure 1 we see a subset of New York City taxi rides from January 2019 [1], with trip distance on the x axis and trip cost on the y axis. We can see that the data are split into two classes: one where cost increases linearly with distance, and another with constant cost. In fact, the increasing class is standard rides, and the constant class is rides from the airport JFK. This information happens to be present in this publicly available dataset, so we can easily partition the data into two sets “Standard” and “JFK”, and fit a linear regression to each. But what if we had no such labels? Could we recover the groupings and correctly estimate the regression parameters?

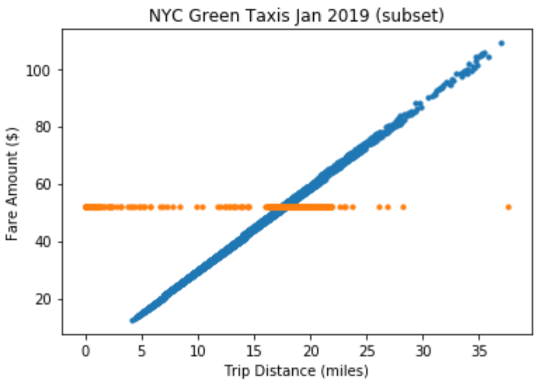


Figure 1. Curated taxi data

This problem is easily solved by a simple variant of the Expectation-Maximization algorithm, assuming we know the number of latent classes. We start by randomly initializing parameter vectors for each class. Then we alternate between assigning each sample to the class whose parameters best predict its response, and refitting the regressions with the new assignments. (Note that these are hard assignments, so this version of EM is more an analog to k-means than to mixture models.)

Though the EM algorithm provably decreases the error at each iteration, it can only guarantee convergence to some local minimum, and it does not necessarily converge quickly. Noisy data, feature covariance, class imbalance, and misspecification of the latent structure can compound these problems. For example, in Figure 2 we see a similar plot to Figure 1, but with extra data included. Here we can see (at least) three issues that were previously absent. First, for the Standard rides, there is a lot of noise in the response. Possibly we need to include more covariates. Second, there is a new, uncommon class of rides that have constant, low cost. These might be the result of faulty meters, for example. Third, the JFK rides are much less common than Standard rides. Due to all these complications, if we try to fit a mixed linear regression to this data, we essentially never recover the Standard/JFK classes; they no longer constitute the global minimum (Figure 3).

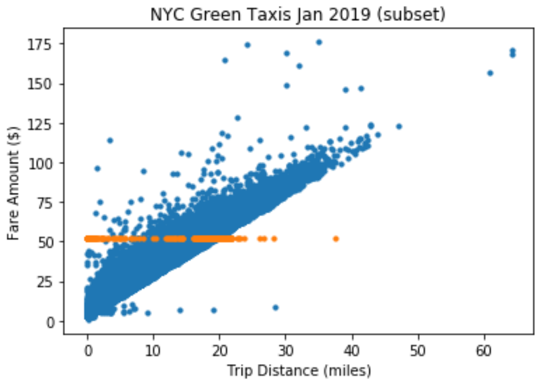


Figure 2. Full taxi data

Obviously we cannot expect EM for mixed regression converge to our desired solution under arbitrarily disadvantageous conditions. But it does not come with strong guarantees in any situation. In this paper, we explore a proposed improvement to the standard EM algorithm by Yi et al [2], which comes with guarantees under extremely regular conditions. In the next section, we explain the improvement. In the remainder of the paper, we compare the empirical performance of the two methods on simulated data sets that conform to the assumptions of Yi et al to varying degrees.

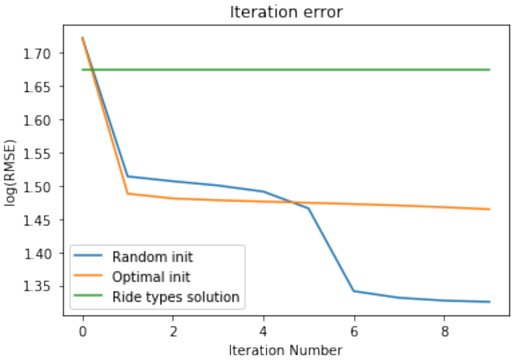


Figure 3. Can't recover ride types on complex data

2. State of the Art

The main contribution of Yi et al is an improvement to the initialization step. We mentioned in the previous section that parameter vectors are initialized randomly before the EM algorithm begins. For the case with two latent classes z_1 and z_2, Yi et al construct a matrix M from the data (X, y) such that the eigenvectors of $\mathbb{E}(M)$ are the true parameter vectors β_1 and β_2. As sample size N increases, M becomes an arbitrarily good approximation of $\mathbb{E}(M)$, so in practice we can just calculate the eigenvectors of M, and have an initialization that approximates the true parameter vectors.

The authors take this a step further. Since the eigenvectors of $\mathbb{E}(M)$ will be orthogonal, but parameter vectors need not be orthogonal, they implement a grid search over the unit circle on the plane spanned by the eigenvectors. They choose as the initializations the vectors on this grid that result in the lowest squared error on the data.

3. Methodology

Our original plan for this project was to apply the initialization procedure to the NYC taxi data mentioned in the Introduction. Since different ride types (e.g. Standard, JFK) have different charging schemes, the data lends itself nicely to mixed regression. However, as we saw, other (truly) latent constructs and class imbalance meant that the ride types don't correspond to global minima of the mixed regression procedure (see Figure 3).

When these problems became clear, we decided to instead assess the initialization procedure on synthetic data. The empirical results from Yi et al use a stringent simulation procedure: Covariates are drawn from a standard normal with zero correlation. Parameters are generated randomly orthonormally on the unit sphere. Latent class memberships are drawn uniformly at random with $p = 0.5$. Responses are noiseless.

We attempt to recreate the findings from Yi et al. Then we violate the simulation assumptions in various ways to see how the new initialization procedure performs under more realistic conditions.

4. Results

First we tried recreating the plot from Yi et al (Figure 4). This plot compares the speed of convergence (averaged over 200 trials) of random initialization versus the initialization proposed in the paper, for a synthetic dataset containing 300 samples of 10-dimensional data. (Note that the error metric here is the maximum of the two distances between the estimated parameter vector and its corresponding true parameter vector. We use this same error metric throughout for consistency.) We additionally tested the random initialization with a grid search on the unit circle that lies on the span of the initialized parameter vectors, akin to what is done for the improved initialization. We did this to disentangle the effect of the eigenvector formulation from the effect of simply being able to test multiple starting points. While we were basically able to replicate the findings for the optimized initialization, we had much more success with random initialization than Yi et all (Figure 5). The optimized initialization converged to the correct parameters only one percentage point more often than random initialization. Furtheremore, we see that the "optimal" initialization provides no benefit once grid search is controlled for. We do not know why the classic initialization scheme performs so much better for us than for Yi et al. Though the paper is not explicit about the random initialization procedure, we cannot replicate those poor results even with particularly antagonistic initialization schemes.

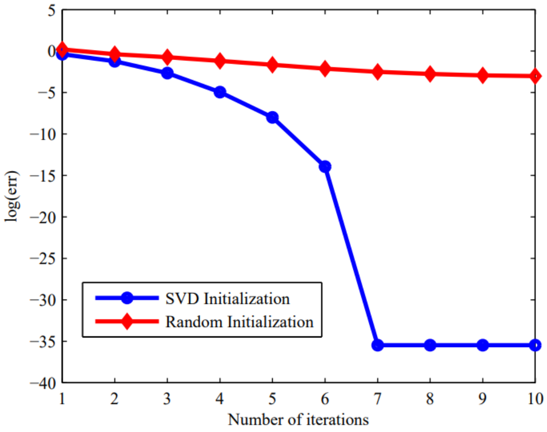


Figure 4. Convergence plot from Yi et al

We chose to test the three initializations (Random, Random + Grid Search, Optimized) for all combinations of three settings:

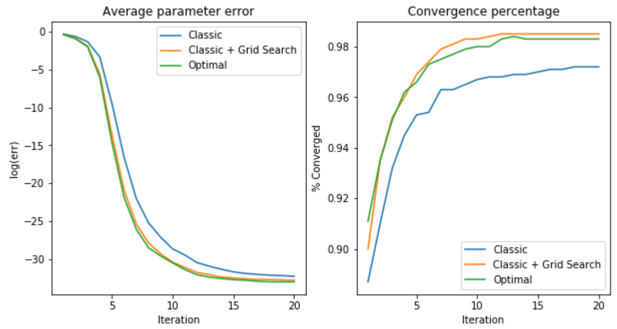


Figure 5. Attempt to recreate convergence speedup

- Data Dimensionality: 3 (Fig 6), 9 (Fig 7), and 27 (Fig 8)
- Covariate Correlations: None, Low $\left[Avg(abs(corr)) < 0.2\right]$, and Moderate $\left[0.35 < Avg(abs(corr)) < 0.45\right]$
- Signal-to-noise ratio: 999 and 3

We chose to assess data correlation and noise because these are some of the most fundamental ways that real-world data do not conform to the assumptions of Yi et all. We chose to assess different dimensionalities to see if the benefits of the initialization procedure change with dimensionality, holding constant the number of samples.

Increasing data dimensionality, decreasing SNR, or increasing covariance negatively impacts the convergence of the EM algorithm. (In all cases, for lower SNR convergence happens with less precision; see y-axes). For moderate covariance the algorithm did not converge for any data dimensionality and SNR. The proposed initialization results in the fastest convergence for low-dimensional data. However, for 9 dimensions it is outperformed by random initialization with grid search.

5. Discussion

It's difficult to determine the meaning of our results, given that we could not replicate the poor baseline to which Yi et al were comparing their method. Nevertheless, we were surprised that the benefits of the initialization procedure seemed to come in the lower dimensions. Intuitively, the benefit of the proposed procedure is that it picks its initialization from a 2D plane that approximates the 2D plane spanned by the true parameter vectors. In higher dimensions, a bad approximation to the true plane is more costly, so the performance of random initialization with grid search should degrade faster than the eigenvector-based initialization with grid search.

One possible explanation for this is that in high dimensions, the parameter space is especially dense with local minima. In Figure 9, we see that for 27 dimensions and no noise, there is actually a decrease in the convergence percentage after the first iteration. We did not think this would be possible, but apparently even when the parameter estimates are within a small tolerance of the actual parameters, the EM algorithm sometimes moves toward other minima. We note that this phenomenon occurs less for the optimized initialization procedure.

One other notable result is that for in all conditions where true convergence was common (zero and low covariance in 3D and 9D, zero covariance in 27D), the optimized initialization has a bigger advantage at later iterations when there is noise than in the noiseless case, though all curves flatten out similarly. This means that in these regimes, the optimized initialization achieves bad local minima less often than the random initialization, though they both converge at similar rates. This bodes well for the practical use of this initialization procedure. However, in most cases this advantage was mostly offset by running grid search on the randomized initialization.

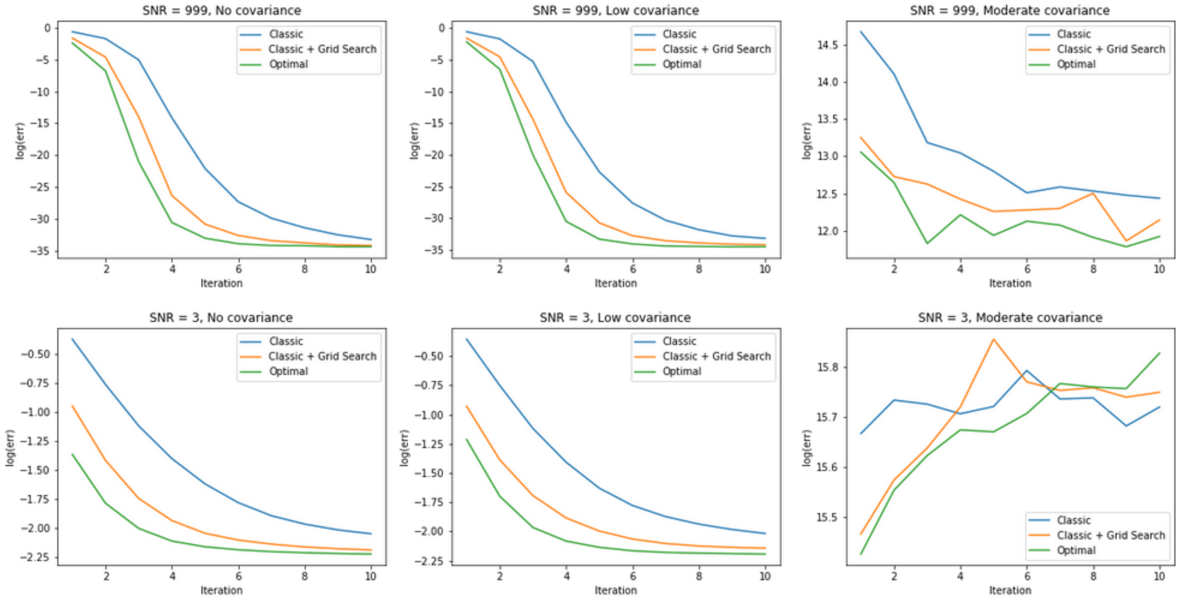
The most obvious way to expand on this analysis is to compare algorithm convergence over a denser grid of conditions, and for addional types of conditions. Other violations of Yi et al's simulation assumptions that we could test include

- Uncentered/unstandardized data
- Non-normal data distributions
- Latent class imbalance
- Non-orthonormal parameter vectors
- More or less course grid search
- Presence of an intercept term

References

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

[2] Xinyang Yi, Constantine Caramanis, Sujay Sanghavi. Alternating minimization for mixed linear regression. 2013.



SNR = 999, No covariance

SNR = 999, Low covariance

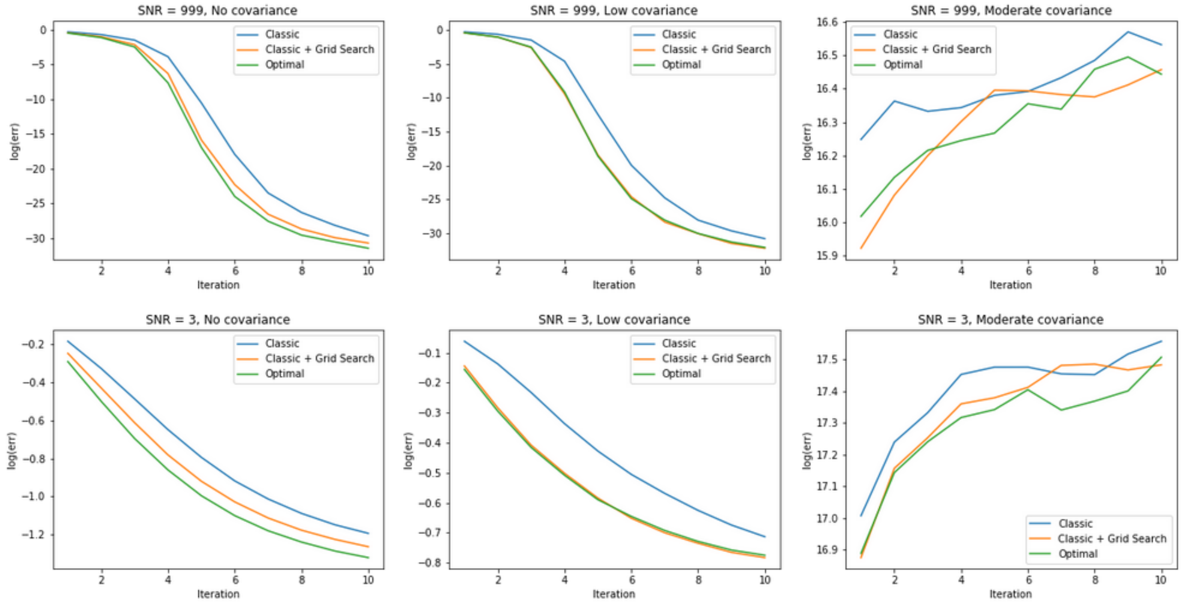
SNR = 999, Moderate covariance

SNR = 3, No covariance

SNR = 3, Low covariance

SNR = 3, Moderate covariance

Figure 6. Convergence for 3-dimensional data



SNR = 999, No covariance

SNR = 999, Low covariance

SNR = 999, Moderate covariance

SNR = 3, No covariance

SNR = 3, Low covariance

SNR = 3, Moderate covariance

Figure 7. Convergence for 9-dimensional data

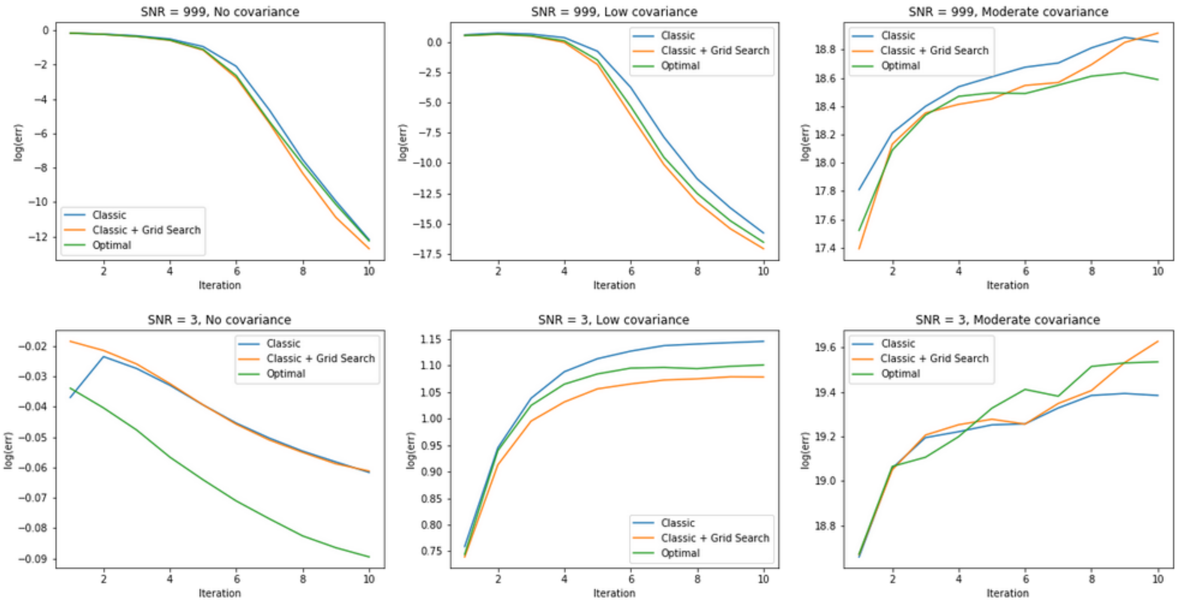


Figure 8. Convergence for 27-dimensional data

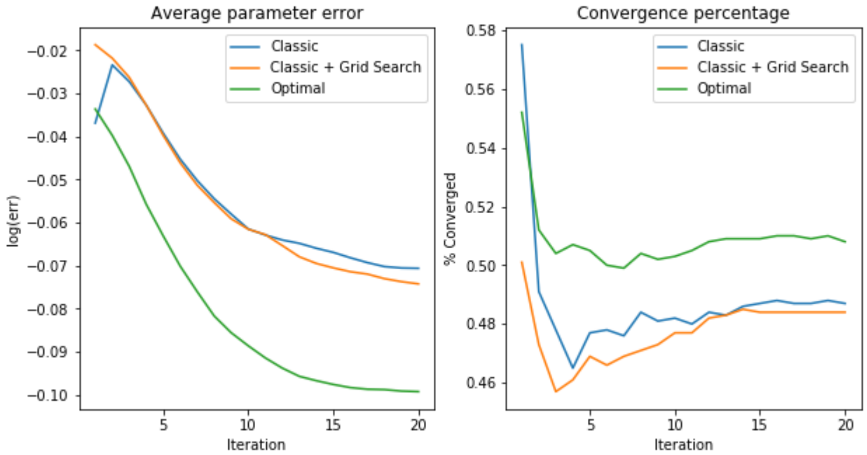


Figure 9. EM can "unconverge"