Alternating Minimization for Mixed Linear Regression

Xinyang Yi YIXY@UTEXAS.EDU
Constantine Caramanis CONSTANTINE@MAIL.UTEXAS.EDU
Sujay Sanghavi SANGHAVI@MAIL.UTEXAS.EDU
Department of Electrical and Computer Engineering,
The University of Texas at Austin,
Austin, TX, 78712

Abstract

Mixed linear regression involves the recovery of two (or more) unknown vectors from unlabeled linear measurements; that is, where each sample comes from exactly one of the vectors, but we do not know which one. It is a classic problem, and the natural and empirically most popular approach to its solution has been the EM algorithm. As in other settings, this is prone to bad local minima; however, each iteration is very fast (alternating between guessing labels, and solving with those labels).

In this paper we provide a new initialization procedure for EM, based on finding the leading two eigenvectors of an appropriate matrix. We then show that with this, a re-sampled version of the EM algorithm provably converges to the correct vectors, under natural assumptions on the sampling distribution, and with nearly optimal (unimprovable) sample complexity. This provides not only the first characterization of EM's performance, but also much lower sample complexity as compared to both standard (randomly initialized) EM, and other methods for this problem.

1. Introduction

In this paper we consider the *mixed linear regression* problem: we would like to recover vectors from linear observations of each, except that these are *unlabeled*. In particular, consider for $i = 1, \dots, N$

$$y_i = \langle \mathbf{x}_i, \beta_1^* \rangle \, z_i \, + \, \langle \mathbf{x}_i, \beta_2^* \rangle \, (1 - z_i) \, + \, w_i,$$

where each z_i is either 1 or 0, and w_i is noise independent of everything else. A value $z_i = 1$ means the i^{th} measurement comes from β_1^*, and $z_i = 0$ means it comes from β_2^*. Our objective is to infer $\beta_1^*, \beta_2^* \in \mathbb{R}^k$ given $(y_i, \mathbf{x}_i), i = 1, \dots, N$; in particular, we do not have access to the labels z_i. For now[1], we do not make *a priori* assumptions on the β's; thus we are necessarily in the regime where the number of samples, N, exceeds the dimensionality, k ($N > k$).

We show in Section 4 that this problem is NP-hard in the absence of any further assumptions. We therefore focus on the case where the measurement vectors $\mathbf{x}_i$ are independent, uniform Gaussian vectors in $\mathbb{R}^p$. While our algorithm works in the noisy case, our performance guarantees currently apply only to the setting of no noise, i.e., $w_i = 0$.

Mixed linear regression naturally arises in any application where measurements are from multiple latent classes and we are interested in parameter estimation. See (Deb & Holmes, 2000) for application of mixed linear regression in health care and work in (Grün et al., 2007) for some related dataset.

The natural, and empirically most popular, approach to solving this problem (as with other problems with missing information) is the Expectation-Maximization, or EM, algorithm; see e.g.Viele & Tong 2002. In our context, EM involves iteratively alternating between updating estimates for β_1, β_2, and estimates for the labels; typically, unless there is specific side-information, the initialization is random. Each step can be solved in closed form, and hence is very computationally efficient. However, as widely acknowledged, there has been to date no way to analytically pre-determine the performance of EM; as in other contexts, it is prone to getting trapped in local minima (Wu, 1983).

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

[1] As we discuss in more detail below, some work has been done in the sparse version of the problem, though the work we are aware of does not give an efficient algorithm with performance guarantees on $\|\hat{\beta}_i - \beta^*\|$, $i = 1, 2$.

Contribution of our paper: We provide the first analytical guarantees on the performance of the EM algorithm for mixed linear regression. A key contribution of our work, both algorithmically and for analysis, is the initialization step. In particular, we develop an initialization scheme, and show that with this EM will converge at least exponentially fast to the correct β's and finally recover ground truth exactly, with $O(k \log^2 k)$ samples for a problem of dimension k. This sample complexity is optimal, up to logarithmic factors, in the dimension and in the error parameter. We are investigating the proposed algorithm in the noisy case, while in this paper we only present noiseless result.

1.1. Related Work

There is of course a huge amount of work in both latent variable modeling, and finite mixture models; here we do not attempt to cover this broad spectrum, but instead focus on the most relevant work, pertaining directly to mixed linear regression.

The work in (Viele & Tong, 2002) describes the application of the EM algorithm to the mixed linear regression problem, both with bayesian priors on the frequencies for each mixture, and in the non-parametric setting (i.e. where one does not *a priori* know the relative fractions from each β). More recently, in the high dimension case when $N < k$ but the βs to be recovered are sparse, the work in (Stadler et al., 2010) proposes changing the vanilla EM for this problem, by adding a Lasso penalty to the β update step. For this method, and sufficient samples, they show that there exists a local minimizer which selects the correct support. This can be viewed as an interesting extension of the known fact about EM, that it has efficient local minima, to the sparse case; however there are no guarantees that any (or even several) runs of this modified EM will actually *find* this good local minimum.

In recent years, an interesting line of work (e.g., (Hsu & Kakade, 2012), (Anandkumar et al., 2012)) has shown the possibility of resolving latent variable models via considering spectral properties of appropriate third-order tensors. Very recent work (Chaganty & Liang, 2013) applies this approach to mixed linear regression. Their method suffers from high sample complexity; in the setting of our problem, their theoretical analysis indicates $N > O(k^6)$. Additionally, this method has much higher computational complexity than the methods in our paper (both EM, and the initialization), due to the fact that they need to work with third-order tensors.

A quite similar problem that attracts extensive attention is subspace clustering, where the goal is to learn an unknown number of linear subspaces of varying dimensions from sample points. Putting our problem in this setting, each sample $(y, \mathbf{x})$ is a vector in $\mathcal{R}^{k+1}$; the points from β_1 correspond to one k-dimensional subspace, and from β_2 to another k-dimensional subspace. Note that this makes for a very hard instance of subspace clustering, as not only are the dimensions of each subspace very high (only one less than ambient), but the projections of the points in the first k coordinates are exactly the same. Even without the latter restriction, one typical method (Vidal et al., 2003), (Elhamifar & Vidal, 2012) – as an example – requires $N \geq O\left(k^2\right)$ to have unique solution.

1.2. Notation

For matrix X, we use $\sigma_i(X)$ to denote the ith singular value of X. We denote the spectral, or operator, norm by $\|X\| := \max_i \sigma_i(X)$. For any vector $\mathbf{x}$ and scalar p, $\|\mathbf{x}\|_p$ is defined as the usual ℓ_p norm. For two vectors $\mathbf{x}, \mathbf{y}$ we use $\langle \mathbf{x}, \mathbf{y} \rangle$ to denote their inner product and $\mathbf{x} \otimes \mathbf{y}$ to denote their outer product. $\mathbf{x}^T$ is transpose of $\mathbf{x}$. We define $T(\mathbf{x}, \mathbf{y})$ to be the subspace spanned by $\mathbf{x}$ and $\mathbf{y}$. The operator $\mathcal{P}_{T(\mathbf{x},\mathbf{y})}$ is the orthogonal projection on $T(\mathbf{x}, \mathbf{y})$. We use N denote number of sample. k is dimension of unknown parameters.

2. Algorithm

In this section we describe the classical EM algorithm as is applied to our problem of mixed linear regression, and our new initialization procedure. Since our analytical results are currently only for the noiseless case, we focus here on EM for this setting, even though EM and also our initialization procedure easily apply to the general setting. The iterations of EM involve alternating between *(a)* given current β_1, β_2, partitioning the samples into J_1 (which are more likely to have come from β_1) and J_2 (respectively, from β_2), and then *(b)* updating each of β_1, β_2 given the new sample sets J_1, J_2 corresponding to each, respectively. Both parts of the iteration are extremely efficient, and can be scaled easily to large problem sizes. In the typical application, in the absence of any extraneous side information, the initial $\beta^{(0)}$'s are chosen at random.

It is not hard to see that each iteration of the above procedure results in a decrease in the loss function

$$
\begin{aligned}
&\mathcal{L}(\beta_1, \beta_2) \\
&= \sum_i \min_{z_i \in \{0,1\}} \left(y_i - \langle \mathbf{x}_i, z_i \beta_1 + (1 - z_i)\beta_2 \rangle\right)^2 . \qquad (1)
\end{aligned}
$$

Note that $\mathcal{L}$, being the minimum of several convex functions, is neither convex nor concave; hence, while EM is guaranteed to converge, all that can be said *a priori* is that it will reach a local minimum. Indeed, our hardness result in Section 4 confirms that for general $\mathbf{x}_i$, this must be the case. Yet even for the Gaussian case we consider, this has essentially been the state of analytical understanding of EM for this problem to date; in particular there are no global

Algorithm 1 EM (noiseless case)

```
input Initial $\beta_1^{(0)}, \beta_2^{(0)}$, number of iterations $t_0$, samples $\{(y_i, \mathbf{x}_i), i = 1, 2, ..., N\}$
for $t = 0, \cdots, t_0 - 1$ do
    {EM Part I: Guess the labels}
    $J_1, J_2 \leftarrow \emptyset$
    for $i = 1, 2, \cdots, N$ do
        if $\left| y_i - \langle \mathbf{x}_i, \beta_1^{(t)} \rangle \right| < \left| y_i - \langle \mathbf{x}_i, \beta_2^{(t)} \rangle \right|$ then
            $J_1 \leftarrow J_1 \cup \{i\}$
        else
            $J_2 \leftarrow J_2 \cup \{i\}$
        end if
    end for
    {EM Part II: Solve least squares}
    $\beta_1^{(t+1)} \leftarrow \mathrm{argmin}_{\beta \in \mathbb{R}^k} \|\mathbf{y}_{J_1} - \mathbf{X}_{J_1}\beta\|_2$
    $\beta_2^{(t+1)} \leftarrow \mathrm{argmin}_{\beta \in \mathbb{R}^k} \|\mathbf{y}_{J_2} - \mathbf{X}_{J_2}\beta\|_2$
end for
output $\beta_1^{(t_0)}, \beta_2^{(t_0)}$
```

guarantees on convergence to the true solutions, under any assumptions, as far as we are aware.

The main algorithmic innovation of our paper is to develop a more principled initialization procedure. In practice, this allows for faster convergence, and with fewer samples, to the true β_1^*, β_2^*. Additionally, it allows us to establish global guarantees for EM, when EM is started from here. We now describe this initialization.

2.1. Initialization

Our initialization procedure is based on estimating the top rank 2 subspace of the following the response weighted covariance matrix:

$$M := \frac{1}{N}\sum_{i=1}^{N} y_i^2 \mathbf{x}_i \otimes \mathbf{x}_i,$$

where $\otimes$ represents the outer product of two vectors. As we will show later, the second moment construction M is an unbiased estimator of a matrix whose top two eigenvectors span the same space spanned by the true β_1^*, β_2^*. We now present the idea, and then formally describe the procedure.

Idea: The expected value of M is given by

$$\mathbb{E}[M] \;=\; p_1 A_1 + p_2 A_2,$$

where p_1, p_2 are the fractions of observations of β_1^*, β_2^* respectively, and the matrices A_i, $i = 1, 2$, are given by

$$A_i \;:=\; \mathbb{E}\left[\, \langle \mathbf{x}, \beta_i^* \rangle^2 \, \mathbf{x} \otimes \mathbf{x} \right],$$

where the expectation is over the random vector $\mathbf{x}$, which in our setting is uniform normal. It is not hard to see that this matrix evaluates to

$$A_i \;=\; I \,+\, 2(\beta_i^* \otimes \beta_i^*)$$

where I is the identity matrix. One can expect that with sufficient samples N, the top-2 eigenspace of M will be a decent approximation of the space spanned by β_1^*, β_2^*. However, generally β_1^*, β_2^* are not identifiable from top-2 eigenvectors of M or even $\mathbb{E}[M]$. Note that even for the expected matrix $\mathbb{E}(M)$, when $p_1 = p_2$ and $\|\beta_1^*\| = \|\beta_2^*\|$, the top two eigenvectors will not be β_1^*,β_2^*. We thus need to run a simple 1-dimensional grid search on the unit circle in this space to find good approximations to the individual vectors β_1^*, β_2^*, as opposed to just the space spanned by them. Our algorithm uses the empirical loss of every candidate pair, produced by the grid search, in order to select a good initial starting point.

The details of the above idea are given below, along with the formal description of our procedure, in Algorithm 2.

Algorithm 2 Initialization

```
input Grid resolution $\delta$, samples $\{(y_i, \mathbf{x}_i), i = 1, 2, ..., N\}$
$M \leftarrow \frac{1}{N}\sum_{i=1}^{N} y_i^2 \mathbf{x}_i \otimes \mathbf{x}_i$
Compute top 2 eigenvectors $\mathbf{v}_1, \mathbf{v}_2$ of $M$
{Make the grid points}
$G \leftarrow \{\mathbf{u} : \mathbf{u} = \mathbf{v}_1 \cos(\delta t) + \mathbf{v}_2 \sin(\delta t), t = 0, 1, ..., \lceil \frac{2\pi}{\delta} \rceil\}$
{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 $\beta_1^{(0)}, \beta_2^{(0)}$
```

Choice of grid resolution δ. In section 4, we show that it's sufficient to choose $\delta < c\|\beta_1^* - \beta_2^*\|_2 \sqrt{\min\{p_1, p_2\}}^3$ for some universal constant c. Even we have no knowledge of gound truth, successful choice of δ relies on a conservative estimation of $\|\beta_1^* - \beta_2^*\|_2$ and $\min\{p_1, p_2\}$. Note that *this upper bound does not scale with problem size*. The number of candidate pairs is actually independent of (k, N).

Search avoidance method using prior knowledge of proportions. When p_1, p_2 are known, approximation of β_1^*, β_2^* can be computed from the top two eigenvectors of M in closed form. Suppose $(\mathbf{v}_b^*, \lambda_b^*), b = 1, 2$ are eigenvectors and eigenvalues of $\mathbb{E}[(M - I)/2]$. We define

$$sign(b) = \begin{cases} 1, \; b = 1 \\ -1, \; b = 2 \end{cases}$$

It is easy to check that when $\lambda_1^* \neq \lambda_2^*$ (we use $-b$ to denote

$\{1,2\} \setminus b$),

$$\beta_b^* = \sqrt{\frac{1-\Delta_b^*}{2}}\mathbf{v}_b^* + sign(b)\sqrt{\frac{1+\Delta_b^*}{2}}\mathbf{v}_{-b}^*, \ b = 1,2, \tag{2}$$

where

$$\Delta_b^* = \frac{(\lambda_b^* - \lambda_{-b}^*)^2 + p_b^2 - p_{-b}^2}{2(\lambda_{-b}^* - \lambda_b^*)p_b}, \ b = 1,2.$$

Duplicate eigenvalues. $\lambda_1^* = \lambda_2^*$ if and only if $p_1 = p_2$ and $\langle \beta_1^*, \beta_2^* \rangle = 0$. In this case $\{\beta_1^*, \beta_2^*\}$ are not identifiable from spectral structure of $\mathbb{E}(M)$ because any linear combination of $\{\beta_1^*, \beta_2^*\}$ is an eigenvector of $\mathbb{E}(M)$. We go back to Algorithm 2 in this case.

Based on the above analysis, we propose an alternative initialization method using proportion information when eigenvalues are nonidentical, in Algorithm 3.

Algorithm 3 Init with proportion information

```
input $p_1, p_2$, samples $\{(y_i, \mathbf{x}_i), i = 1, 2, ..., N\}$
  $M \leftarrow \frac{1}{N}\sum_{i=1}^{N} y_i^2 \mathbf{x}_i \otimes \mathbf{x}_i$
  Compute top 2 eigenvectors and eigenvalues $(\mathbf{v}_b, \lambda_b), b = 1, 2$ of $(M - I)/2$
  Compute $\beta_1^{(0)}, \beta_2^{(0)}$ via equation (2) ( use empirical version, i.e., remove superscript $*$)
output $\beta_1^{(0)}, \beta_2^{(0)}$
```

In Section 3, we demonstrate empirically the importance of this initialization technique; we show that EM initialized randomly has remarkably slower performance compared to EM initialized by Algorithm 2. Our theoretical results presented in Section 4, confirm this observation analytically.

3. Numerical Results

In this section, we present the empirical performance of our algorithm on synthetic data set. The results highlight in particular two important features of our results. First, the simulations corroborate our theoretical results given in Section 4, which show that our algorithm is nearly optimal (unimprovable) in terms of sample complexity. Indeed, we show here that EM+SVD succeeds when given about as many samples as dimensions (in the absence of additional structure, e.g., sparsity, it is not possible to do better). Second, our results show that the SVD initialization seems to be critical: without it, EM's performance is significantly degraded.

Experiment Settings. Each input vector $\mathbf{x}_i$ are generated independently from standard Guassian distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}$). We then choose the mixture labels for each sample with equal probability, i.e., we set $p_1 = p_2 = 0.5$. Also, in each trial, we generate β_1^* and β_2^* randomly but keep $\langle \beta_1^*, \beta_2^* \rangle = 1.73$. This constant 1.73 is arbitrarily chosen here. In this case, β_1^* and β_2^* are non-orthogonal and it's impossible to recover them from the SVD step due to ambiguity. We run algorithm 2 with a fairly coarse grid: $\delta = 0.3$. We also test algorithm 3 using $p_1 = p_2$. The following metric which stands for global optimality is used

$$\text{err}^{(t)} := \max\{\|\beta_1^{(t)} - \beta_1^*\|_2, \|\beta_2^{(t)} - \beta_2^*\|_2\}. \tag{3}$$

Here t is the sequence of number of iterations.

Sample Complexity. In figure 1 we empirically investigate how the number of samples N needed for exact recovery scales with the dimension k. Each point in Figure 1 represents 1000 trials, and the corresponding value of N is the number of samples at which the success rate was greater than 0.99. We use algorithm 2 for initialization. In figure 2, we show the phase transition curves with a few (N, k) pairs.

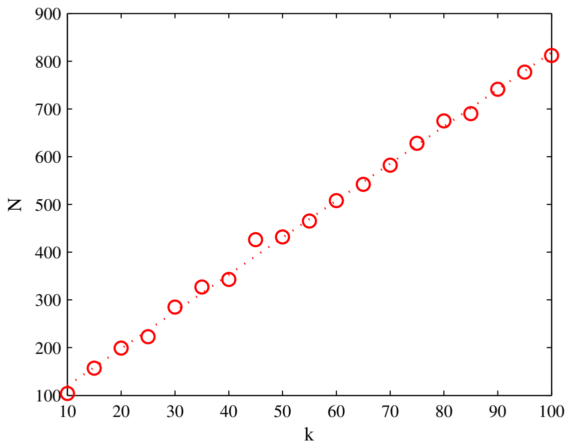

Figure 1. Number of samples needed for success rate greater than 0.99 using SVD+EM. The dotted line is the least square fit of the experimental data.

Effect of Initialization. We compare our eigenvector-based Initialization + EM with the usual randomly initialized EM. For $N = 300$ samples and $k = 10$ dimensions, figure 3 shows how the error err converges as a function of the iterations. Each curve is averaged over 200 trials. We observe that the final error of SVD+EM is about 10^{-35}. The level of noise results from float computation. For each trial, the blue and green curves show that exact recovery occurred after 7 iterations. This is possible since we are in the noiseless case.

As can be clearly seen, initialization has a profound effect on the performance of EM in our setting; it allows for exact recovery with high probability in a small number of iterations, while random initialization does not.

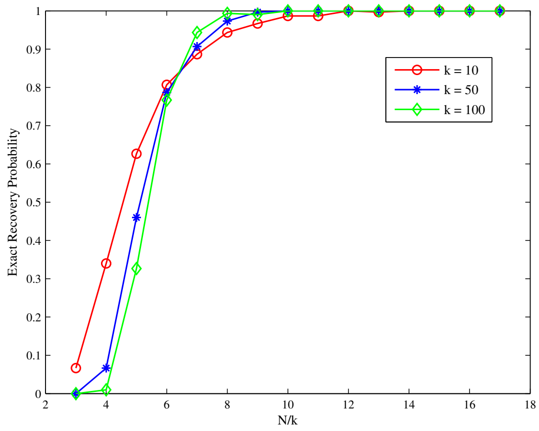

Figure 2. Success probability vs. normalized number of samples, i.e., N/k.

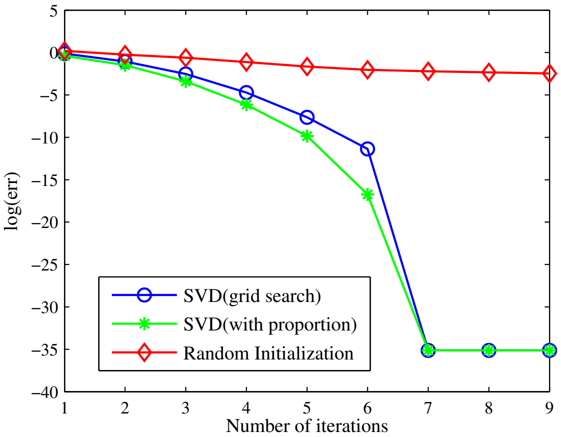

Figure 3. This figure compares the decay in error, as a function of iteration count of EM, with and without our initialization. As can be seen, initialization allows for exact recovery (the 10^{-35} error is precision of Matlab) in a small number of iterations, while the standard random initialization is still not close.

4. Main Results

In this section, we present the main results of our paper: provable statistical guarantees for EM, initialized with our Algorithm 2, in solving the mixed linear regression problem. We first show that for general $\{\mathbf{x}_i\}$, the problem is NP-hard, even without noise. Then, we focus on the setting where each measurement vector $\mathbf{x}_i$ is iid and sampled from the uniform normal distribution $\mathcal{N}(0, I)$. We also assume that the true vectors β_1^*, β_2^* are equal in magnitude, which without loss of generality, we assume is 1. Intuitively, equal magnitudes represents a hard case, as in this setting the y_i's from the two β's are statistically identical[2].

Our proof can be broken into two key results. We first show that using $O(k \log^2 k)$ samples, with high probability our initialization procedure returns $\beta_1^{(0)}, \beta_2^{(0)}$ which are within a constant distance of the true β_1^*, β_2^*. We note that for our scaling guarantees to hold, this constant need only be independent of the dimension, and in particular, it need not depend on the final desired precision. Results with a $1/\text{error}$ or even $1/\text{error}^2$ dependence – as would be required in order for the SVD step alone to obtain an approximation of β_i^*, $i = 1, 2$, to within some error tolerance, are exponentially worse than what our two-step algorithm guarantees.

We then show that, given this good initialization, at any subsequent step t with current estimate $(\beta_1^{(t)}, \beta_2^{(t)})$, doing one step of the EM iteration with samples that are independent of these $\beta_i^{(t)}$ results in the error decreasing by a factor of half, hence implying geometric convergence. As we explain below, our analysis providing this guarantee depends on using a new set of samples, i.e., the analysis does not allow re-use samples across iterations, as typically done in EM. We believe this is an artifact of the analysis; and of course, in practice, reusing the samples in each iteration seems to be advantageous.

Thus, our analytical results are for *resampled versions* of EM and the initialization scheme, which we state as Algorithms 4 and 5 below. Essentially, resampling involves splitting the set of samples into disjoint sets, and using one set for each iteration of EM; otherwise the algorithm is identical to before. Since we have geometric decrease in the error, achieving an ϵ accuracy comes at an additional cost of a factor $\log(1/\epsilon)$ in the sample complexity, as compared to what may have been possible with the non-resampled case. We then show that when $\epsilon \leq O\left(1/k^2\right)$, the error decays to be zero with high probability. In other words, we need in total $O\left(\log k\right)$ iterations in order to do exact recovery. Additionally, and the main contribution of this paper, the resampled version given here, represents the only known algorithm, EM or otherwise, with provable global statistical guarantees for the mixed linear regression problem, with sample complexity close to $O(k)$.

Similarly, in the initialization procedure, for analytical guarantees we require two separate sets of samples: one set $\mathcal{S}_*$ for finding the top-2 eignespace, and another set $\mathcal{S}_+$ for evaluating the loss function for grid points.

First, we provide the hardness result for the case of general

[2] In particular, each y_i has mean 0, and variance $\|\beta_1^*\|^2$ if it comes from the first vector, and $\|\beta_2^*\|^2$ if it comes from the second. Having them be equal, i.e. $\|\beta_1^*\|^2 = \|\beta_2^*\|^2$, makes the y_is statistically identical.

Algorithm 4 EM with resampling

```
input Initial $\beta_1^{(0)}, \beta_2^{(0)}$, number of iterations $t_0$, samples $\{(y_i, \mathbf{x}_i), i = 1, 2, ..., N\}$
  Partition the samples $\{(y_i, \mathbf{x}_i)\}$ into $t_0$ disjoint sets: $\mathcal{S}_1, ..., \mathcal{S}_{t_0}$.
  for $t = 1, \cdots, t_0$ do
      Use $\mathcal{S}_t$ to run lines 2 to 13 in algorithm 1.
  end for
output $\beta_1^{(t_0)}, \beta_2^{(t_0)}$
```

Algorithm 5 Initialization with resampling

```
input Grid resolution $\delta$, samples $\{(y_i, \mathbf{x}_i), i = 1, 2, ..., N\}$
  Partition the samples $\{(y_i, \mathbf{x}_i)\}$ into two disjoint sets: $\mathcal{S}_*, \mathcal{S}_+$
  $M \leftarrow \frac{1}{|\mathcal{S}_*|} \sum_{i \in \mathcal{S}_*} y_i^2 \mathbf{x}_i \otimes \mathbf{x}_i$
  Compute top 2 eigenvectors $\mathbf{v}_1, \mathbf{v}_2$ of $M$
  {Make the grid points}
  $G \leftarrow \{\mathbf{u} : \mathbf{u} = \mathbf{v}_1 \cos(\delta t) + \mathbf{v}_2 \sin(\delta t), t = 0, 1, ..., \lceil \frac{2\pi}{\delta} \rceil\}$
  {Pick the pair that has the lowest loss}
      $$\beta_1^{(0)}, \beta_2^{(0)} \leftarrow \arg \min_{\mathbf{u}_1, \mathbf{u}_2 \in G} \mathcal{L}_+(\mathbf{u}_1, \mathbf{u}_2)$$
  where this loss $\mathcal{L}_+$ is evaluated as in (1) using samples in $\mathcal{S}_+$
output $\beta_1^{(0)}, \beta_2^{(0)}$
```

$\{\mathbf{x}_i\}$.

Proposition 4.1. *Deciding if a general instance of the mixed linear equations problem specified by $(\mathbf{y}, X)$ has a solution, β_1, β_2, is NP-hard.*

The proof follows via a reduction from the so-called SUBSETSUM problem, which is known to be NP-hard(Garey & Johnson, 1979). We postpone the details to the supplemental material.

We now state two theoretical guarantees of the initialization algorithms. Recall that the error $err^{(t)}$ is as given in (3), and p_1, p_2 are the fractions of observations that come from β_1^*, β_2^* respectively.

The following result guarantees a good initialization (algorithm 5) *without requiring sample complexity that depends on the final target error of the ultimate solution*. Essentially, it says that we obtain an initialization that is *good enough* using $O(k \log^2 k)$ samples.

Proposition 4.2. *Given any constant $\widehat{c} < 1/2$, with probability at least $1 - c_3 k^{-2}$ Algorithm 5 produces an initialization $(\beta_1^{(0)}, \beta_2^{(0)})$, satisfying*

$$err^{(0)} \leq \widehat{c} \min\{p_1, p_2\} \|\beta_1^* - \beta_2^*\|_2,$$

as long as we choose grid resolution $\delta \leq \frac{2}{11}\widehat{c}\|\beta_1^* - \beta_2^*\|_2 \sqrt{\min\{p_1, p_2\}}^3$, and the number of samples $|\mathcal{S}_*|$ and $|\mathcal{S}_+|$ satisfy:

$$\begin{aligned} |\mathcal{S}_*| &\geq c_1 \left(\frac{1}{\widetilde{\delta}}\right)^2 k \log^2 k \\ |\mathcal{S}_+| &\geq \left(\frac{c_2}{\min\{p_1, p_2\}}\right) k, \end{aligned}$$

where c_1, c_2 and c_3 depend on $\hat{c}$ and $\min\{p_1, p_2\}$ but not on the dimension, k, and where

$$\widetilde{\delta} = \frac{\delta^2}{384}(1 - \sqrt{1 - 4(1 - \langle \beta_1^*, \beta_2^* \rangle^2) p_1 p_2}).$$

Algorithm 3 can be analyzed without resampling argument. The input sample set is $\mathcal{S}_*$, we have the following conclusion.

Proposition 4.3. *Consider initialization method in algorithm 3. Given any constant $\widehat{c} < 1/2$, with probablity at least $1 - \frac{1}{k^2}$, the approach produces an initialization $(\beta_1^{(0)}, \beta_2^{(0)})$ satisfying*

$$err^{(0)} \leq \widehat{c} \min\{p_1, p_2\} \|\beta_1^* - \beta_2^*\|_2,$$

if

$$|\mathcal{S}_*| \geq c_1 \left(\frac{1}{\widetilde{\delta}}\right)^2 k \log^2 k.$$

Here c_1 is a constant that depends on $\widehat{c}$. And

$$\sqrt{\widetilde{\delta}} = \widehat{c} \sqrt{\min\{p_1, p_2\}}^3 \|\beta_1^* - \beta_2^*\|_2 (\sqrt{1 - \kappa}) \kappa,$$

where $\kappa = \sqrt{1 - 4(1 - \langle \beta_1^*, \beta_2^* \rangle^2) p_1 p_2}$.

Comparing the obtained upper bound of $\widetilde{\delta}$ with that in proposition 4.2, we note there is an additional κ factor. Actually, κ represents the gap between top two eigenvectors of $\mathbb{E}(M)$. This factor characterizes the hardness of identifying two vectors from search avoiding method.

The proofs of proposition 4.2 and 4.3 relies on standard concentration results and eigenspace perturbation analysis. We postpone the details to supplemental materials.

The main theorem of the paper guarantees geometric decay of error, assuming a good initialization. Essentially, this says that to achieve error less than ϵ, we need $\log(1/\epsilon)$ iterations, each using $O(k)$ samples. Again, we note the absence of higher order dependence on the dimension, k, or anything other than the mild dependence on the final error tolerance, ϵ.

Theorem 4.4. *Consider one iteration in algorithm 4. For fixed $(\beta_1^{(t-1)}, \beta_2^{(t-1)})$, there exist absolute constants $\widetilde{c}, c_1, c_2$ such that if*

$$err^{(t-1)} \;\leq\; \widetilde{c}\,\min\{p_1, p_2\}\,\|\beta_1^* - \beta_2^*\|_2,$$

and if the number of samples in that iteration satisfies

$$|\mathcal{S}_t| \;\geq\; \left(\frac{c_1}{\min\{p_1, p_2\}}\right) k,$$

then with probability greater than $1 - \exp(-c_2 k)$ we have a geometric decrease in the error at the next stage, i.e.

$$err^{(t)} \leq \frac{1}{2} err^{(t-1)}$$

Note that the decrease factor $1/2$ is arbitrarily chosen here. To put the above results together, we choose the constant $\widehat{c}$ in proposition 4.2 and 4.3 to be less than the constant $\widetilde{c}$ in Theorem 4.4. Then, in each iteration of alternating minimization, with $O(k)$ fresh samples, the error decays geometrically by a constant factor with probability greater than $1 - \exp -ck$. Suppose we are satisfied with error level ϵ, resampling regime requires $O\left(k \log^2 k + k \log(1/\epsilon)\right)$ number of samples.

Let J_b^* denote the set of samples generated from $\beta_b^*, b = 1, 2$. It's not hard to observe that in noiseless case, exact recovery occurs when $J_b = J_b^*$. The next result shows that when $\epsilon < \frac{c}{k^2}\|\beta_1^* - \beta_2^*\|_2$, fresh $\Theta(k)$ samples will be clustered correctly which results in exact recovery.

Proposition 4.5. *(Exact Recovery) There exist absolute constants c_1, c_2 such that if*

$$err^{(t-1)} \leq \frac{c_1}{k^2}\|\beta_1^* - \beta_2^*\|_2$$

and

$$\frac{1}{\min\{p_1, p_2\}} k < |\mathcal{S}_t| < c_2 k,$$

then with probability greater than $1 - \frac{1}{k}$,

$$err^{(t)} = 0.$$

By setting $\epsilon = O\left(1/k^2\right)$, it turns out that exact recovery needs totally $O\left(k \log^2 k\right)$ samples. On using alternating minimization, approximation error will decay geometrically in the first place. Then when error hits some level, exact recovery occurs and the ground truth is found. Simulation results in figure 3 supports our conclusion.

5. Proof of Theorem 4.4

In this section, we provide the proofs of our main theorem: we show that with a good starting point, EM exhibits geometric convergence, reducing the error by a factor of 2 at each iteration. The following lemma is crucial.

Lemma 5.1. *Assume $\mathbf{x} \in \mathbb{R}^k$ is a standard normal random vector. Let u, v be two fixed vectors in $\mathbb{R}^k$. Define $\alpha_{(u,v)} := \cos^{-1}\frac{(v-u)^\top(v+u)}{\|u+v\|_2\|u-v\|_2}$, $\alpha_{(u,v)} \in [0, \pi]$. Let $\Sigma = \mathbb{E}(\mathbf{x}\mathbf{x}^\top | (\mathbf{x}^\top u)^2 > (\mathbf{x}^\top v)^2)$. Then,*

(1)

$$\sigma_{\max}(\Sigma) = 1 + \frac{\sin\alpha_{(u,v)}}{\alpha_{(u,v)}}, \tag{4}$$

$$\sigma_{\min}(\Sigma) = 1 - \frac{\sin\alpha_{(u,v)}}{\alpha_{(u,v)}}, \tag{5}$$

(2)

$$\mathbb{P}\left[(\mathbf{x}^\top u)^2 > (\mathbf{x}^\top v)^2\right] \begin{cases} > \dfrac{1}{2} & \|u\|_2 > \|v\|_2 \\ \leq \dfrac{\|u\|_2}{\|v\|_2} & \|u\|_2 < \|v\|_2 \end{cases} \tag{6}$$

To simplify notation, we drop the iteration index t, and let (β_1, β_2) denote the input to the EM algorithm, and (β_1^+, β_2^+) denote its output. Similarly, we write $\mathrm{err} := \max_i \|\beta_i - \beta_i^*\|$ and $\mathrm{err}^+ := \max_i \|\beta_i^+ - \beta_i^*\|$. We denote by J_1^* and J_2^* the sets of samples that come from β_1^* and β_2^* respectively, and similarly we denote the sets produced by the "E" step using the current iteration (β_1, β_2) by J_1 and J_2. Thus we have:

$$J_1^* := \{i \in \mathcal{S}_t : y_i = \mathbf{x}_i^\top \beta_1^*\},$$

and

$$J_1 := \{i \in \mathcal{S}_t : (y_i - \mathbf{x}_i^\top \beta_1)^2 < (y_i - \mathbf{x}_i^\top \beta_2)^2\},$$

and similarly for J_2^* and J_2.

We define a diagonal matrix $W \in \mathbb{R}^{\mathcal{S}_t \times \mathcal{S}_t}$ to pick out the rows in J_1 when used for left multiplication: to this end, let $W_{ii} = 1$ if $i \in J_1$, and zero otherwise. Let W^* be defined similarly, using J_1^*. Thus, β_1^+ is the least squares solution to $W\mathbf{y} = WX\beta$, and β_2^+ is the least squares solution to $(I - W)\mathbf{y} = (I - W)X\beta$, and

$$\mathbf{y} = W^* X \beta_1^* + (I - W^*) X \beta_2^*.$$

Observing that $W^2 = W$, we have that β_1^+ has closed form

$$\beta_1^+ = (X^\top W X)^{-1} X^\top W \mathbf{y}.$$

By simple algebraic calculation, we find

$$\beta_1^+ - \beta_1^* = (X^\top W X)^{-1} X^\top (W W^* - W) X (\beta_1^* - \beta_2^*).$$

In order to bound the magnitude of the error and hence of the right hand side, we write

$$\|\beta_1^+ - \beta_1^*\|_2 \leq AB, \tag{7}$$

where

$$\begin{aligned} A &= \|(X^\top W X)^{-1}\| \\ B &= \left\|X^\top (W - WW^*) X(\beta_1^* - \beta_2^*)\right\|_2. \end{aligned}$$

Bounding A. Observe that $X^\top W X = \sum_{i \in J_1} \mathbf{x}_i \mathbf{x}_i^\top$. Decomposing $J_1 = (J_1 \cap J_1^*) \cup (J_1 \cap J_2^*)$, we have

$$\sigma_{\min}(X^\top W X) \geq \sigma_{\min}(\sum_{i \in J_1 \cap J_1^*} \mathbf{x}_i \mathbf{x}_i^\top).$$

We need to control this quantity. We do so by lower bounding the number of terms in $J_1 \cap J_1^*$, and also the smallest singular value of the matrix $\Sigma = \mathbb{E}\left[\{\mathbf{x}_i \mathbf{x}_i^\top | i \in J_1 \cap J_1^*\}\right]$.

If the current error satisfies

$$\text{err} \leq \frac{\|\beta_1^* - \beta_2^*\|_2}{2}, \tag{8}$$

we have $\|\beta_1^* - \beta_2\|_2 > \|\beta_1^* - \beta_1\|_2$. Now, from Lemma 5.1, we have

$$\mathbb{P}\left[(\mathbf{x}_i^\top(\beta_1^* - \beta_1))^2 < (\mathbf{x}_i^\top(\beta_1^* - \beta_2))^2\right] > \frac{1}{2}$$

and

$$\sigma_{\min}(\Sigma) \geq (1 - \frac{2}{\pi}).$$

Using Hoeffding's inequality, with probability greater than $1 - e^{-\frac{1}{8}p_1|\mathcal{S}_t|}$, we have the bound $|J_1 \cap J_1^*| \geq \frac{1}{4}p_1|\mathcal{S}_t|$. By a standard concentration argument (see, e.g., (Vershynin, 2010) Corollary 50), we conclude that for any $\eta \in (0, 1 - \frac{2}{\pi})$, there exists a constant c_3, such that if

$$|\mathcal{S}_t| \geq c_3 \frac{k}{\eta p_1}, \tag{9}$$

then

$$A \leq \frac{4}{(1 - \frac{2}{\pi} - \eta)p_1|\mathcal{S}_t|}, \tag{10}$$

with probability at least $1 - e^{-k}$.

Bounding B. Let $Q := X^\top(W - WW^*)X$. We have

$$B^2 \leq \sigma_{\max}(Q)(\beta_1^* - \beta_2^*)^\top Q(\beta_1^* - \beta_2^*).$$

Moreover,

$$\begin{aligned} & (\beta_1^* - \beta_2^*)^\top Q(\beta_1^* - \beta_2^*) \\ = & \sum_{i \in J_1 \cap J_2^*} (\mathbf{x}_i^\top(\beta_1^* - \beta_2^*))^2 \\ \leq & \sum_{i \in J_1 \bigcap J_2^*} 2(\mathbf{x}_i^\top(\beta_1^* - \beta_1))^2 + 2(\mathbf{x}_i^\top(\beta_2^* - \beta_1))^2 \\ \leq & \sum_{i \in J_1 \bigcap J_2^*} 2(\mathbf{x}_i^\top(\beta_1^* - \beta_1))^2 + 2(\mathbf{x}_i^\top(\beta_2^* - \beta_2))^2. \end{aligned}$$

The last inequality results from the decision rule labeling β_1 and β_2. This immediately implies that

$$B \leq 2\sigma_{\max}(Q)\text{err}. \tag{11}$$

Using Lemma 5.1, $\sigma_{\max}(\mathbb{E}\left[\mathbf{x}_i \mathbf{x}_i^\top | i \in J_1 \cap J_2^*\right]) \leq 2$. Following Theorem 39 in (Vershynin, 2010), we claim that there exist constants c_4, c_5 such that with probability greater than $1 - 2e^{-c_4 k}$,

$$\sigma_{\max}(Q) \leq |J_1 \cap J_2^*|(2 + \max(\hat{\eta}, \hat{\eta}^2))$$

where $\hat{\eta} = c_5\sqrt{\frac{k}{|J_1 \cap J_2^*|}}$. Letting $c_6 = 2 + c_5^2$, we have

$$\sigma_{\max}(Q) \leq c_6 \max(k, |J_1 \cap J_2^*|).$$

Now using again Lemma 5.1, we find

$$\mathbb{E}\left[|J_1 \cap J_2^*|\right] \leq \frac{2\text{err}^{(t-1)}}{\|\beta_1^* - \beta_2^*\|_2} p_2|\mathcal{S}_t|.$$

By Hoeffding's inequality, with high probability

$$|J_1 \cap J_2^*| \leq 2\mathbb{E}\left[|J_1 \cap J_2^*|\right].$$

Now we can combine the bounds on A (10) and on B (11). Setting $\eta = (1 - \frac{2}{\pi})/2$, when

$$\text{err} \leq \frac{0.18}{64c_6} p_1\|\beta_1^* - \beta_2^*\|_2, \tag{12}$$

and

$$|\mathcal{S}_t| \geq \frac{16c_6}{0.18}\frac{k}{p_1}, \tag{13}$$

we conclude that

$$\|\beta_1^+ - \beta_1^*\|_2 \leq \frac{1}{2}\text{err}.$$

Repeating the steps for β_2^+, we obtain a similar result, and hence we conclude: $\text{err}^+ \leq \frac{1}{2}\text{err}$, as claimed.

6. Conclusion and Future Work

In this paper, we provide a sufficient condition when alternating minimization, as a popular method for solving non-convex optimization problem, achieves global optimum when it is used for solving mixed linear regression. Based on that, under mild conditions, we show a novel spectral initialization algorithm is guaranteed to satisfy the sufficient condition of alternating minimization with near optimal sample complexity. As a future direction, it would be interesting and challenging to analyze alternating minimization in the face of noise.

Acknowledgement

The authors would like to acknowledge NSF grants 1302435, 0954059, 1017525 and DTRA grant HDTRA1-13-1-0024 for supporting this research.

References

Anandkumar, Anima, Ge, Rong, Hsu, Daniel, Kakade, Sham M., and Telgarsky, Matus. Tensor decompositions for learning latent variable models. *CoRR*, abs/1210.7559, 2012.

Chaganty, A. and Liang, P. Spectral experts for estimating mixtures of linear regressions. In *International Conference on Machine Learning (ICML)*, 2013.

Deb, Partha and Holmes, Ann M. Estimates of use and costs of behavioural health care: a comparison of standard and finite mixture models. *Health Economics*, 9(6): 475–489, 2000.

Elhamifar, Ehsan and Vidal, René. Sparse subspace clustering: Algorithm, theory, and applications. *CoRR*, abs/1203.1005, 2012.

Garey, M.R. and Johnson, D.S. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Series of Books in the Mathematical Sciences. W. H. Freeman, 1979.

Grün, Bettina, Leisch, Friedrich, et al. Applications of finite mixtures of regression models. *URL: http://cran. r-project. org/web/packages/flexmix/vignettes/regression-examples. pdf*, 2007.

Hsu, Daniel and Kakade, Sham M. Learning gaussian mixture models: Moment methods and spectral decompositions. *CoRR*, abs/1206.5766, 2012.

Stadler, Nicolas, Buhlmann, Peter, and Geer, Sara. 1-penalization for mixture regression models. *TEST*, 19 (2):209–256, 2010. ISSN 1133-0686.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *ArXiv e-prints*, November 2010.

Vidal, René, Ma, Yi, and Sastry, Shankar. Generalized principal component analysis (gpca). In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pp. I–621. IEEE, 2003.

Viele, Kert and Tong, Barbara. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12(4), 2002. ISSN 0960-3174. URL http://dx.doi.org/10.1023/A%3A1020779827503.

Wu, CF. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.