Regularized EM Algorithms: A Unified Framework and Statistical Guarantees

Authored by:

Constantine Caramanis
Xinyang Yi

Abstract

Latent models are a fundamental modeling tool in machine learning applications, but they present significant computational and analytical challenges. The popular EM algorithm and its variants, is a much used algorithmic tool; yet our rigorous understanding of its performance is highly incomplete. Recently, work in [1] has demonstrated that for an important class of problems, EM exhibits linear local convergence. In the high-dimensional setting, however, the M-step may not be well defined. We address precisely this setting through a unified treatment using regularization. While regularization for high-dimensional problems is by now well understood, the iterative EM algorithm requires a careful balancing of making progress towards the solution while identifying the right structure (e.g., sparsity or low-rank). In particular, regularizing the M-step using the state-of-the-art high-dimensional prescriptions (e.g., 'a la [19]) is not guaranteed to provide this balance. Our algorithm and analysis are linked in a way that reveals the balance between optimization and statistical errors. We specialize our general framework to sparse gaussian mixture models, high-dimensional mixed regression, and regression with missing variables, obtaining statistical guarantees for each of these examples.

1 Paper Body

We give general conditions for the convergence of the EM method for high-dimensional estimation. We specialize these conditions to several problems of interest, including high-dimensional sparse and low-rank mixed regression, sparse gaussian mixture models, and regression with missing covariates. As we explain below, the key problem in the high-dimensional setting is the M -step. A natural idea is to modify this step via appropriate regularization, yet choosing the appropriate sequence of regularizers is a critical problem. As we know from the theory of regularized M-estimators (e.g., [19]) the regularizer should be chosen proportional to the target estimation error. For EM, however, the target

estimation error changes at each step. The main contribution of our work is technical: we show how to perform this iterative regularization. We show that the regularization sequence must be chosen so that it converges to a quantity controlled by the ultimate estimation error. In existing work, the estimation error is given by the relationship between the population and empirical M -step operators, but this too is not well defined in the highdimensional setting. Thus a key step, related both to our algorithm and its convergence analysis, is obtaining a different characterization of statistical error for the high-dimensional setting. Background and Related Work EM (e.g., [8, 12]) is a general algorithmic approach for handling latent variable models (including mixtures), popular largely because it is typically computationally highly scalable, and easy to implement. On the flip side, despite a fairly long history of studying EM in theory (e.g., [12, 17, 21]), 1

very little has been understood about general statistical guarantees until recently. Very recent work in [1] establishes a general local convergence theorem (i.e., assuming initialization lies in a local region around true parameter) and statistical guarantees for EM, which is then specialized to obtain near-optimal rates for several specific low-dimensional problems ? low-dimensional in the sense of the classical statistical setting where the samples outnumber the dimension. A central challenge in extending EM (and as a corollary, the analysis in [1]) to the high-dimensional regime is the M -step. On the algorithm side, the M -step will not be stable (or even well-defined in some cases) in the high-dimensional setting. To make matters worse, any analysis that relies on showing that the finite-sample M -step is somehow ?close? to the M -step performed with infinite data (the population-level M -step) simply cannot apply in the high-dimensional regime. Recent work in [20] treats high-dimensional EM using a truncated M -step. This works in some settings, but also requires specialized treatment for every different setting, precisely because of the difficulty with the M -step. In contrast to work in [20], we pursue a high-dimensional extension via regularization. The central challenge, as mentioned above, is in picking the sequence of regularization coefficients, as this must control the optimization error (related to the special structure of ? ?), as well as the statistical error. Finally, we note that for finite mixture regression, St?adler et al.[16] consider an ‘1 regularized EM algorithm for which they develop some asymptotic analysis and oracle inequality. However, this work doesn?t establish the theoretical properties of local optima arising from regularized EM. Our work addresses this issue from a local convergence perspective by using a novel choice of regularization.

2

Classical EM and Challenges in High Dimensions

The EM algorithm is an iterative algorithm designed to combat the nonconvexity of max likelihood due to latent variables. For space concerns we omit the standard derivation, and only give the definitions we need in the sequel. Let Y , Z be random variables taking values in Y,Z, with joint distribution f? (y, z) depending on model parameter ? ? ? ? Rp . We observe samples of Y but not of the latent variable Z. EM seeks to maximize a lower bound on the maximum likelihood function for ?. Letting ?? (z—y) denote the conditional distribution

of Z given Y = y, letting y?? (y) denote the marginal distribution of Y , and defining the function n Z 1X 0 Qn (? —?) := ?? (z—yi) log f?0 (yi , z)dz, (2.1) n i=1 Z one iteration of the EM algorithm, mapping ? (t) to ? (t+1) , consists of the following two steps: ? E-step: Compute function Qn (?—? (t)) given ? (t) . ? M-step: ? (t+1) ? Mn (?) := arg max?0 ?? Qn (? 0 —? (t)). We can define the population (infinite sample) versions of Qn and Mn in a natural manner: Z Z 0 ? Q(? —?) := y? (y) ?? (z—y) log f?0 (y, z)dzdy Y

M(?)

=

Z 0

arg max Q(? —?). 0 ? ??

(2.2) (2.3)

This paper is about the high-dimensional setting where the number of samples n may be far less than the dimensionality p of the parameter ?, but where ? exhibits some special structure, e.g., it may be a sparse vector or a low-rank matrix. In such a setting, the M -step of the EM algorithm may be highly problematic. In many settings, for example sparse mixed regression, the M -step may not even be well defined. More generally, when n p, Mn (?) may be far from the population version, M(?), and in particular, the minimum estimation error kMn (? ?) ? M(? ?)k can be much larger than the signal strength k? ? k. This quantity is used in [1] as well as in follow-up work in [20], as a measure of statistical error. In the high dimensional setting, something else is needed.

3

Algorithm

The basis of our algorithm is the by-now well understood concept of regularized high dimensional estimators, where the regularization is tuned to the underlying structure of ? ? , thus defining a regu2

larized M -step via

Mrn (?) := arg max Qn (? 0 —?) ? ?n R(? 0), 0

(3.1)

? ??

where R(?) denotes an appropriate regularizer chosen to match the structure of ? ? . The key chal(t) lenge is how to choose the sequence of regularizers {?n } in the iterative process, so as to control optimization and statistical error. As detailed in Algorithm 1, our sequence of regularizers attempts to match the target estimation error at each step of the EM iteration. For an intuition of what this might look like, consider the estimation error at step t: kMrn (? (t)) ? ? ? k2 . By the triangle inequality, we can bound this by a sum of two terms: the optimization error and the final estimation error: kMrn (? (t)) ? ? ? k2 ? kMrn (? (t)) ? Mrn (? ?)k2 + kMrn (? ?) ? ? ? k2 . (3.2) (t)

Since we expect (and show) linear convergence of the optimization, it is natural to update ?n via a (t) (t?1) recursion of the form ?n = ??n +? as in (3.3), where the first term represents the optimization error, and ? represents the final statistical error, i.e., the last term above in (3.2). A key part of our analysis shows that this error (and hence ?) is controlled by k?Qn (? ? —?) ? ?Q(? ? —?)kR? , which in turn can be bounded uniformly for a variety

of important applications of EM, including the three discussed in this paper (see Section 5). While a technical point, it is this key insight that enables the right choice of algorithm and its analysis. In the cases we consider, we obtain min-max optimal rates of convergence, demonstrating that no algorithm, let alone another variant of EM, can perform better. Algorithm 1 Regularized EM Algorithm Input Samples {yi }ni=1 , regularizer R, number of iterations T , initial parameter ? (0) , initial regu(0) larization parameter ?n , estimated statistical error ?, contractive factor ? ¡ 1. 1: For t = 1, 2, . . . , T do 2: Regularization parameter update: (t?1) ?(t) + ?. n ? ??n

3: 4:

E-step: Compute function Qn (?—? Regularized M-step:

(t?1)

(3.3)

) according to (2.1).

? (t) ? Mrn (? (t?1)) := arg max Qn (?—? (t?1)) ? ?(t) n ? R(?). ???

5: End For Output ? (T) .

4

Statistical Guarantees

We now turn to the theoretical analysis of regularized EM algorithm. We first set up a general analytical framework for regularized EM where the key ingredients are decomposable regularizer and several technical conditions on the population based Q(?—?) and the sample based Qn (?—?). In Section 4.3, we provide our main result (Theorem 1) that characterizes both computational and statistical performance of the proposed variant of regularized EM algorithm. 4.1

Decomposable Regularizers

Decomposable regularizers (e.g., [3, 6, 14, 19]), have been shown to be useful both empirically and theoretically for high dimensional structural estimation, and they also play an important role in our analytical framework. Recall that for R : Rp ? R+ a norm, and a pair of subspaces (S, S) in Rp such that S ? S, we have the following definition: Definition 1 (Decomposability). Regularizer R : Rp ? R+ is decomposable with respect to (S, S) if ? R(u + v) = R(u) + R(v), for any u ? S, v ? S . Typically, the structure of model parameter ? ? can be characterized by specifying a subspace S such that ? ? ? S. The common use of a regularizer is thus to penalize the compositions of solution that 3

live outside S. We are interested in bounding the estimation error in some norm k ? k. The following quantity is critical in connecting R to k ? k. Definition 2 (Subspace Compatibility Constant). For any subspace S ? Rp , a given regularizer R and some norm k ? k, the subspace compatibility constant of S with respect to R, k ? k is given by R(u) ?(S) := sup . u?S{0} kuk

As is standard, the dual norm of R is defined as R? (v) := supR(u)?1 u, v . To simplify notation, we let kukR := R(u) and kukR? := R? (u). 4.2

Conditions on Q(?—?) and Qn (?—?)

Next, we review three technical conditions, originally proposed by [1], on the population level Q(?—?) function, and then we give two important conditions that the empirial function Qn (?—?) must satisfy, including one that characterizes the statistical error. It is well known that performance of EM algorithm

is sensitive to initialization. Following the lowdimensional development in [1], our results

are local, and apply to an r-neighborhood region around ? ? : B(r; ? ?) := u ? ?, ku ? ? ? k ? r . We first require that Q(?—? ?) is self consistent as stated below. This is satisfied, in particular, when ? ? maximizes the population log likelihood function, as happens in most settings of interest [12]. Condition 1 (Self Consistency). Function Q(?—? ?) is self consistent, namely ? ? = arg max Q(?—? ?). ???

We also require that the function Q(?—?) satisfies a certain strong concavity condition and is smooth over ?. Condition 2 (Strong Concavity and Smoothness (?, ?, r)). Q(?—? ?) is ?-strongly concave over ?, i.e.,

? (4.1) Q(?2 —? ?) ? Q(?1 —? ?) ? ?Q(?1 —? ?), ?2 ? ?1 ? ? k?2 ? ?1 k2 , ? ?1 , ?2 ? ?. 2 ? For any ? ? B(r; ?), Q(?—?) is ?-smooth over ?, i.e.,

? Q(?2 —?) ? Q(?1 —?) ? ?Q(?1 —?), ?2 ? ?1 ? ? k?2 ? ?1 k2 , ? ?1 , ?2 ? ?. (4.2) 2 The next condition is key in guaranteeing the curvature of Q(?—?) is similar to that of Q(?—? ?) when ? is close to ? ? . It has also been called First Order Stability in [1]. Condition 3 (Gradient Stability (?, r)). For any ? ? B(r; ? ?), we have

?Q(M(?)—?) ? ?Q(M(?)—? ?) ? ? k? ? ? ? k. The above condition only requires that the gradient be stable at one point M(?). This is sufficient for our analysis. In fact, for many concrete examples, one can verify a stronger version of Condition

3 that is ?Q(? 0 —?) ? ?Q(? 0 —? ?) ? ? k? ? ? ? k, ? ? 0 ? B(r; ? ?). Next we require two conditions on the empirical function Qn (?—?), which is computed from finite number of samples according to (2.1). Our first condition, parallel to Condition 2, imposes a curvature constraint on Qn (?—?). In order to guarantee that the estimation error k? (t) ? ? ? k in step t of the EM algorithm is well controlled, we would like Qn (?—? (t?1)) to be strongly concave at ? ? . However, in the setting where n p, there might exist directions along which Qn (?—? (t?1)) is flat, e.g., as in mixed linear regression and missing covariate regression. In contrast with Condition 2, we only require Qn (?—?) to be strongly concave over a particular set C(S, S; R) that is defined in terms of the subspace pair (S, S) and regularizer R. This set is defined as follows:

p

C(S, S; R) := u ? R : ?S ? (u) R ? 2 ? ?S (u) R + 2 ? ?(S) ? u , (4.3) where the projection operator ?S : Rp ? Rp is defined as ?S (u) := arg minv?S kv ? uk. The restricted strong concavity (RSC) condition is as follows. 4

Condition 4 (RSC (?n , S, S, r,T?)). For any fixed ? ? B(r; ? ?), with probability at least 1 ? ?, we have that for all ? 0 ? ? ? ? ? C(S, S; R),

?n Qn (? 0 —?) ? Qn (? ? —?) ? ?Qn (? ? —?), ? 0 ? ? ? ? ? k? 0 ? ? ? k2 . 2 The above condition states that Qn (?—?) is strongly concave in directions ? 0 ? ? ? that belong to C(S, S; R). It is instructive to compare Condition 4 with a related condition proposed by [14] for analyzing high dimensional M-estimators. They require the loss function to be strongly convex over the cone {u ? Rp : k?S ? (u)kR . k?S (u)kR }. Therefore our

restrictive set (4.3) is similar to the cone but has the additional term 2?(S)kuk. The main purpose of the term 2?(S)kuk is to allow the regularization parameter ?n to jointly control optimization and statistical error. We note that while Condition 4 is stronger than the usual

RSC condition in M-estimator, in typical settings

the difference is immaterial. This is because ?S (u) R is within a constant factor of ?(S) ? u , and hence checking RSC over C amounts to checking it over k?S ? (u)kR . ?(S)kuk, which is indeed what is typically also done in the M-estimator setting. Finally, we establish the condition that characterizes the achievable statistical error. Condition 5 (Statistical Error (?n , r, ?)). For any fixed ? ? B(r; ? ?), with probability at least 1 ? ?, we have

?Qn (? ? —?) ? ?Q(? ? —?) ? ? ?n . (4.4) R This quantity replaces the term kMn (?)?M(?)k which appears in [1] and [20], and which presents problems in the high dimensional regime. 4.3

Main Results

In this section, we provide the theoretical guarantees for a resampled version of our regularized EM algorithm: we split the whole dataset into T pieces and use a fresh piece of data in each iteration of regularized EM. As in [1], resampling makes it possible to check that Conditions 4-5 are satisfied without requiring them to hold uniformly for all ? ? B(r; ? ?) with high probability. Our empirical results indicate that it is not in fact required and is an artifact of the analysis. We refer to this resampled version as Algorithm 2. In the sequel, we let m := n/T to denote the sample complexity in each iteration. We let ? := supu?Rp {0} kuk? /kuk, where k ? k? is the dual norm of k ? k. For Algorithm 2, our main result is as follows. The proof is deferred to the Supplemental Material. Theorem 1. Assume the model parameter ? ? ? S and regularizer R is decomposable with respect to (S, S) where S ? S ? Rp . Assume r ¿ 0 is such that B(r; ? ?) ? ?. Further, assume function Q(?—?), defined in (2.2), is self consistent and satisfies Conditions 2-3 with parameters (?, ?, r) and (?, r). Given n samples and T iterations, let m := n/T . Assume Qm (?—?), computed from any m i.i.d. samples according to (2.1), satisfies Conditions 4-5 with parameters (?m , S, S, r, 0.5?/T) ??? and (?m , r, 0.5?/T). Let ?? := 5 ?? , and assume 0 ¡ ? ¡ ? and 0 ¡ ?? ? 3/4. Define m ? := r?m /[60?(S)] and assume ?m is such that ?m ? ?. Consider Algorithm 2 with initialization ? (0) ? B(r; ? ?) and with regularization parameters given by 1 ? ?t t ?m ?(t) k? (0) ? ? ? k + ?, t = 1, 2, . . . , T (4.5) m =? 1?? 5?(S) for any ? ? [3?m , 3?], ? ? [?? , 3/4]. Then with probability at least 1 ? ?, we have that for any t ? [T], 5 1 ? ?t ?(S)?. (4.6) k? (t) ? ? ? k ? ?t k? (0) ? ? ? k + ?m 1 ? ? The estimation error is bounded by a term decaying linearly with number of iterations t, which we can think of as the optimization error and a second term that characterizes the ultimate estimation error of our algorithm. With T = O(log n) and suitable choice of ? such that ? = O(?n/T), we bound the ultimate estimation error as 1 k? (T) ? ? ? k . ?(S)?n/T . (4.7) (1 ? ?)?n/T 5

We note that overestimating the initial error, k? (0) ?? ? k is not important, as it may slightly increase the overall number of iterations, but will not impact the ultimate estimation error. The constraint ?m . r?m /?(S) ensures that ?

(t) is contained in B(r; ? ?) for all t ? [T]. This constraint is quite mild in the sense that if ?m = ?(r?m /?(S)), ? (0) is a decent estimator with estimation error O(?(S)?m /?m) that already matches our expectation.

5

Examples: Applying the Theory

Now we introduce three well known latent variable models. For each model, we first review the standard EM algorithm formulations, and discuss the extensions to the high dimensional setting. Then we apply Theorem 1 to obtain the statistical guarantee of the regularized EM with data splitting (Algorithm 2). The key ingredient underlying these results is to check the technical conditions in Section 4 hold for each model. We postpone these tedious details to the Supplemental Material. 5.1

Gaussian Mixture Model

We consider the balanced isotropic Gaussian mixture model (GMM) with two components where the distribution of random variables (Y, Z) ? Rp ? {?1, 1} is characterized as Pr (Y = y—Z = z) = ?(y; z ? ? ? , ? 2 Ip), Pr(Z = 1) = Pr(Z = ?1) = 1/2. Here we use ?(?—?, ?) to denote the probability density function of N (?, ?). In this example, Z is the latent variable that indicates the cluster id of each sample. Given n i.i.d. samples {yi }ni=1 , function Qn (?—?) defined in (2.1) corresponds to n

M QGM (? 0 —?) = ? n

1 X w(yi ; ?)kyi ? ? 0 k22 + (1 ? w(yi ; ?))kyi + ? 0 k22 , 2n i=1 ky??k2

ky??k2

(5.1)

ky+?k2

where w(y; ?) := exp (? 2?2 2)[exp (? 2?2 2) + exp (? 2?2 2)]?1 . We assume ? ? ? B0 (s; p) := {u ? Rp : — supp(u)— ? s}. Naturally, we choose the regularizer R(?) to be the ‘1 norm. We define the signal-to-noise ratio SNR := k? ? k2 /?. Corollary 1 (Sparse Recovery in GMM). There exist constants ?, C such that if SNR ? ?, n/T ? 2 [80C(k? ? k? + ?)/k? ? k2] s log p, ? (0) ? B(k? ? k2 /4; ? ?); then with probability at least 1 ? T /p p ? (0) Algorithm 2 with parameters ? = C(k? ? k? + ?) T log p/n, ?n/T = 0.2k? (0) ? ? ? k2 / s, any ? ? [1/2, 3/4] and ‘1 regularization generates ? (t) that has estimation error r 5C(k? ? k? + ?) s log p (t) ? t (0) ? k? ? ? k2 ? ? k? ? ? k2 + T , for all t ? [T]. 1?? n

(5.2)

Note that by p setting T log(n/ log p), the order of final estimation error turns out to be (k? ? k? + ?) (s log p)/n) p log (n/ log p). The minimax rate for estimating s-sparse vector in a single Gaussian cluster is s log p/n, thereby the rate is optimal on (n, p, s) up to a log factor. 5.2

Mixed Linear Regression

Mixed linear regression (MLR), as considered in some recent work [5, 7, 22], is the problem of recovering two or more linear vectors from mixed linear measurements. In the case of mixed linear regression with two symmetric and balanced components, the response-covariate pair (Y, X) ? R ? Rp is linked through Y = hX, Z ? ? ? i + W, where W is the noise term and Z is the latent

variable that has Rademacher distribution over {?1, 1}. We assume X ? N (0, Ip), W ? N (0, ? 2). In this setting, with n i.i.d. samples {yi , xi }ni=1 of pair (Y, X), function Qn (?—?) then corresponds to n

LR QM (? 0 —?) = ? n

1 X w(yi , xi ; ?)(yi ? hxi , ? 0 i)2 + (1 ? w(yi , xi ; ?))(yi + hxi , ? 0 i)2 , 2n i=1 (5.3) 6

2

2

2

where w(y, x; ?) := exp (? (y?hx,?i))[exp (? (y?hx,?i)) + exp (? (y+hx,?i))]?1 . 2? 2 2? 2 2? 2 We consider two kinds of structure on ? ? : Sparse Recovery. Assume ? ? ? B0 (s; p). Then let R be the ‘1 norm, as in the previous section. We define SNR := k? ? k2 /?. Corollary 2 (Sparse recovery in MLR). There exist constant ?, C, C 0 such that if SNR ? ?, n/T ? 2 C 0 [(k? ? k2 + ?)/k? ? k2] s log p, ? (0) ? B(k? ? k2 /240, ? ?); then with probability at least 1 ? T /p p ? (0) Algorithm 2 with parameters ? = C(k? ? k2 + ?) T log p/n, ?n/T = k? (0) ? ? ? k2 /(15 s), any ? ? [1/2, 3/4] and ‘1 regularization generates ? (t) that has estimation error r 15C(k? ? k2 + ?) s log p (t) ? t (0) ? k? ? ? k2 ? ? k? ? ? k2 + T , for all t ? [T]. 1?? n Performing T

log(n/(s log p)) iterations gives us estimation rate (k? ? k2 + p ?) (s log p/n) log (n/(s log p)) which is near-optimal on (s, p, n). The dependence on k? ? k2 , which also appears in the analysis of EM in the classical (low dimensional) setting [1], arises from fundamental limits of EM. Removing such dependence for MLR is possible by convex relaxation [7]. It is interesting to study how to remove it in the high dimensional setting. Low Rank Recovery. Second we consider the setting where the model parameter is a matrix ?? ? Rp1 ?p2 with rank(??) = ? min(p1 , p2). We further assume X ? Rp1 ?p2 is an i.i.d. Gaussian matrix, i.e., entries of X are independent random variables with distribution 1). We apply PpN1 ,p(0, 2 nuclear norm regularization to serve the low rank structure, i.e, R(?) = i=1 —si (?)—, where si (?) is the ith singular value of ?. Similarly, we let SNR := k?? kF /?. Corollary 3 (Low rank recovery in MLR). There exist constant ?, C, C 0 such that if SNR ? ?, 2 n/T ? C 0 [(k?? kF + ?)/k?? kF] ?(p1 + p2), ?(0) ? B(k?? kF /1600, ??); thenpwith probability at least 1 ? T exp(?p1 ? p2) Algorithm 2 with parameters ? = C(k?? kF + ?) T (p1 + p2)/n, ? (0) ?n/T = 0.01k?(0) ? ?? kF / 2?, any ? ? [1/2, 3/4] and nuclear norm regularization generates ?(t) that has estimation error (t)

k?

?

t

(0)

? ? kF ? ? k?

100C 0 (k?? kF + ?) ? ? kF + 1?? ?

r

2?(p1 + p2) T , for all t ? [T]. n

The standard low rank matrix recovery with a single component, including other sensing matrix designs beyond the Gaussianity, has been studied extensively (e.g., [2, 4, 13, 15]). To the best of our knowledge, the theoretical study of the mixed low rank matrix recovery has not been considered. 5.3

Missing Covariate Regression

As our last example, we consider the missing covariate regression (MCR) problem. To parallel standard linear regression, {yi , xi }ni=1 are samples of (Y, X) linked through Y = hX, ? ? i + W . However, we assume each entry of xi is missing independently with probability ? (0, 1). Thereei takes the form fore, the observed covariate vector x

xi,j with probability 1 ? x ei,j = . ? otherwise We assume the model is under Gaussian design X ? N (0, Ip), W ? N (0, ? 2). We refer the reader to our Supplementary Material for the specific Qn (?—?) function. In high dimensional case, we assume ? ? ? B0 (s; p). We define ? := k? ? k2 /? to be the SNR and ? := r/k? ? k2 to be the relative contractivity radius. In particular, let ? := (1 + ?)?. Corollary 4 (Sparse Recovery in MCR). There exist constants C, C 0 , C0 , C1 such that if (1+?)? ? C0 ¡ 1, ¡ C1 , n/T ? C 0 max{? 2 (??)?1 , 1}s log p, ? (0) ? B(?k? ? k2 , ? ?); then with probp (0) ability at least 1 ? T /p Algorithm 2 with parameters ? = C? T log p/n, ?n/T = k? (0) ? ? ? ? k2 /(45 s), any ? ? [1/2, 3/4] and ‘1 regularization generates ? (t) that has estimation error r 45C? s log p (t) ? t (0) ? k? ? ? k2 ? ? k? ? ? k2 + T , for all t ? [T], 1?? n 7

Unlike the previous two models, we require an upper bound on the signal to noise ratio. This unusual constraint p is in fact unavoidable [10]. By optimizing T , the order of final estimation error turns out to be ? s log p/n log(n/(s log p)).

6

Simulations

We now provide some simulation results to back up our theory. Note that while Theorem 1 requires resampling, we believe in practice this is unnecessary. This is validated by our results, where we apply Algorithm 1 to the four latent variable models discussed in Section 5. Convergence Rate. We first evaluate the convergence of Algorithm 1 assuming only that the initialization is a bounded distance from ? ? . For a given error ?k? ? k2 , the initial parameter ? (0) is picked randomly from the sphere centered around ? ? with radius ?k? ? k2 . We use Algorithm 1 with T = 7, (0) ? = 0.7, ?n in Theorem 1. The choice of the critical parameter ? is given in the Supplementary Material. For every single trial, we report estimation error k? (t) ? ? ? k2 and optimization error k? (t) ? ? (T) k2 in every iteration. We plot the log of errors over iteration t in Figure 1. 2 Est error Opt error

Log error

-1

Log error

1 Est error Opt error

0

-2 -3

3 Est error Opt error
0
-2
Log error
0
-4
-1 -2
-6
-4 -5
-8
-6
-10
0
1
2
3
4
5
6
7
Est error Opt error
2
Log error
1
1 0 -1
-3
0
1
2
Number of iterations
3
4
5
6
-4
7
-2 0
1
2
Number of iterations
(a) GMM
3
4
5
6
-3

7
0
1
2
Number of iterations
(b) MLR(sparse)
3
4
5
6
7
Number of iterations
(c) MLR(low rank)
(d) MCR

Figure 1: Convergence of regularized EM algorithm. In each panel, one curve is plotted from single independent trial. Settings: (a,b,d) (n, p, s) = (500, 800, 5); (d) (n, p, ?) = (600, 30, 3); (a-c) SNR = 5; (d) (SNR,) = (0.5, 0.2); (a-d) ? = 0.5. Statistical Rate. We now evaluate the statistical rate. We set T = 7 and compute estimation error on ?b := ? (T) . In Figure 2, we plot k?b ? ? ? k2 over normalized sample complexity, i.e., n/(s log p) for s-sparse parameter and n/(?p) for rank ? p-by-p parameter. We refer the reader to Figure 1 for other settings. We observe that the same normalized sample complexity leads to almost identical estimation error in practice, which thus supports the corresponding statistical rate established in Section 5. 0.4 p = 200 p = 400 p = 800
1.4 p = 200 p = 400 p = 800
0.18 0.16 0.14
0.3 0.25
1
0.12
0.6
0.1
0.15
0.4
10
15
20
25
30
n/(s log p)
(a) GMM
5
10
15
20
25

30
1.4
1 4
5
6
7
n/(?p)
(b) MLR(sparse)
1.6
1.2
3
n/(s log p)
p = 200 p = 400 p = 800
1.8
0.8
0.2
5
2 p = 25 p = 30 p = 35
1.2
? ? ?? kF k?
0.35
k?? ? ? ? k2
k?? ? ? ? k2
0.2
k?? ? ? ? k2
0.22
(c) MLR(low rank)
8
5
10
15
20
25
30
n/(s log p)
(d) MCR

Figure 2: Statistical rates. Each point is an average of 20 independent trials. Settings: (a,b,d) s = 5; (c) ? = 3.

Acknowledgments The authors would like to acknowledge NSF grants 1056028, 1302435 and 1116955. This research was also partially supported by the U.S. Department of Transportation through the Data-Supported Transportation Operations and Planning (D-STOP) Tier 1 University Transportation Center.

8

2 References

[1] Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. arXiv preprint arXiv:1408.2156, 2014. [2] T Tony Cai and Anru Zhang. Rop: Matrix recovery via rank-one projections. The Annals of Statistics, 43(1):102?138, 2015. [3] Emmanuel Candes and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. The Annals of Statistics, pages 2313?2351, 2007. [4] Emmanuel J Cand'es and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. Information Theory, IEEE Transactions on, 57(4):2342? 2359, 2011. [5] Arun Tejasvi Chaganty and Percy Liang. Spectral experts for estimating mixtures of linear regressions. arXiv preprint arXiv:1306.3729, 2013. [6] Yudong Chen, Sujay Sanghavi, and Huan Xu. Improved graph clustering. Information Theory, IEEE Transactions on, 60(10):6440?6455, Oct 2014. [7] Yudong Chen, Xinyang Yi, and Constantine Caramanis. A convex formulation for mixed regression with two components: Minimax optimal rates. In Conf. on Learning Theory, 2014. [8] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological), pages 1?38, 1977. [9] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In Advances in Neural Information Processing Systems, pages 2726? 2734, 2011. [10] Po-Ling Loh and Martin J Wainwright. Corrupted and missing predictors: Minimax bounds for highdimensional linear regression. In Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on, pages 2601?2605. IEEE, 2012. [11] Jinwen Ma and Lei Xu. Asymptotic convergence properties of the em algorithm with respect to the overlap in the mixture. Neurocomputing, 68:105?129, 2005. [12] Geoffrey McLachlan and Thriyambakam Krishnan. The EM algorithm and extensions, volume 382. John Wiley & Sons, 2007. [13] Sahand Negahban, Martin J Wainwright, et al. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. The Annals of Statistics, 39(2):1069?1097, 2011. [14] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In Advances in Neural Information Processing Systems, pages 1348?1356, 2009. [15] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM review, 52(3):471?501, 2010. [16] Nicolas St?adler, Peter B?uhlmann, and Sara Van De Geer. L1-penalization for mixture regression models. Test, 19(2):209?256, 2010. [17] Paul Tseng. An analysis of the em algorithm and entropy-like proximal point methods. Mathematics of Operations Research, 29(1):27?44, 2004. [18] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. arXiv preprint arXiv:1011.3027, 2010. [19] Martin J Wainwright. Structured regularizers for high-dimensional problems: Statistical and computational issues. Annual Review of Statistics and Its Application, 1:233?253, 2014. [20] Zhaoran Wang, Quanquan Gu, Yang Ning,

and Han Liu. High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. arXiv preprint arXiv:1412.8729, 2014. [21] C.F.Jeff Wu. On the convergence properties of the em algorithm. The Annals of statistics, pages 95?103, 1983. [22] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. arXiv preprint arXiv:1310.3745, 2013.

9