An Improved Analysis of Alternating Minimization for Structured Multi-Response Regression

Sheng Chen
The Voleon Group
chen2832@umn.edu

Arindam Banerjee
Dept. of Computer Science & Engineering
University of Minnesota, Twin Cities
banerjee@cs.umn.edu

Abstract

Multi-response linear models aggregate a set of vanilla linear models by assuming correlated noise across them, which has an unknown covariance structure. To find the coefficient vector, estimators with a joint approximation of the noise covariance are often preferred than the simple linear regression in view of their superior empirical performance, which can be generally solved by alternating-minimization-type procedures. Due to the non-convex nature of such joint estimators, the theoretical justification of their efficiency is typically challenging. The existing analyses fail to fully explain the empirical observations due to the assumption of resampling on the alternating procedures, which requires access to fresh samples in each iteration. In this work, we present a resampling-free analysis for the alternating minimization algorithm applied to the multi-response regression. In particular, we focus on the high-dimensional setting of multi-response linear models with structured coefficient parameter, and the statistical error of the parameter can be expressed by the complexity measure, Gaussian width, which is related to the assumed structure. More importantly, to the best of our knowledge, our result reveals for the first time that the alternating minimization with random initialization can achieve the same performance as the well-initialized one when solving this multi-response regression problem. Experimental results support our theoretical developments.

1 Introduction

We consider the following multi-response linear model \cite{1, 5, 18} with $m$ real-valued outputs,

$$ y = X\theta^* + \eta, \quad \text{where} \quad \eta = \Sigma^{1/2} \tilde{\eta} $$

(1)

where $y \in \mathbb{R}^m$ is the response vector and $X \in \mathbb{R}^{m \times p}$ consists of $m$ $p$-dimensional feature vectors, and $\tilde{\eta} \in \mathbb{R}^m$ is a zero-mean isotropic noise vector. The $m$ responses share the same underlying parameter $\theta^* \in \mathbb{R}^p$, which corresponds to the so-called pooled model \cite{17}. Without loss of generality, the counterpart of (1) with response-specific parameters can be equivalently written in the above form, by block-diagonalizing rows of $X$ and concatenating different parameters into a single vector. What makes this model different from vanilla linear models is the correlated noise $\eta$ across responses, which is assumed to be a linear transformation of $\tilde{\eta}$. The noise covariance of $\eta$ is given by $\text{Cov}(\eta) = \Sigma_s$. This model has found numbers of real-world applications, such as econometrics \cite{17}, computational biology \cite{22} and climate informatics \cite{14, 15}, just to name a few.

In practice, we are given $n$ observations of $(X, y)$, denoted by $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^n$, while the noise covariance structure $\Sigma_s$ between responses is typically unknown. Our goal is to estimate the

$^*$This work was done when the author studied at University of Minnesota, Twin Cities
parameter $\theta$, ideally together with $\Sigma$. In this work, we additionally focus on the high-dimensional regime, where the true parameter $\theta$ is assumed to possess certain low-complexity structure measured by some function $f : \mathbb{R}^p \rightarrow \mathbb{R}_+$, which can be either convex (e.g., norms) or non-convex (e.g., $L_0$ cardinality). For the low-dimensional setting, it has been shown, both empirically and theoretically, that simultaneously estimating $\Sigma$ and $\theta$ leads to better performance than ordinary least squares [20].

Inspired by this fact, we consider the joint estimator of $(\Sigma, \theta)$ in high dimension as follows,

$$
\hat{(\theta_n, \Sigma_n)} = \arg\min_{\theta \in \mathbb{R}^p, \Sigma \succeq 0} \frac{1}{2n} \sum_{i=1}^n \left\| \Sigma^{-\frac{1}{2}} (y_i - X_i \theta) \right\|^2_2 + \frac{1}{2} \log |\Sigma| \quad \text{s.t.} \quad f(\theta) \leq \lambda ,
$$

which corresponds to the constrained maximum likelihood estimator (MLE) of $(\Sigma, \theta)$ when the noise is multivariate Gaussian. The structural assumption on $\theta$, is encoded by the inequality constraint. Though the noise structure is accounted in this joint estimator, one challenge faced by the associated optimization problem is the non-convexity of the objective function. In the light of the simplicity of the marginal optimization over $\Sigma$ and $\theta$ when the other is fixed, a popular approach to dealing with such problem is alternating minimization (AltMin), i.e., alternately solving for $\Sigma$ (and $\theta$) while keeping $\theta$ (and $\Sigma$) fixed. For problem (2), the update of AltMin can be written as

$$
\Sigma_{(t+1)} = \frac{1}{n} \sum_{i=1}^n \left( y_i - X_i \hat{\theta}_{(t)} \right) \left( y_i - X_i \hat{\theta}_{(t)} \right)^T ,
$$

$$
\hat{\theta}_{(t+1)} = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^n \left\| \Sigma_{(t+1)}^{-\frac{1}{2}} (y_i - X_i \theta) \right\|^2_2 \quad \text{s.t.} \quad f(\theta) \leq \lambda ,
$$

which will be executed for a number of iterations, say $T$. Here the new $\Sigma_{(t+1)}$ is obtained by computing the empirical covariance of the residues estimated at $\hat{\theta}_{(t)}$. Though $f$ can potentially be non-convex, the update of $\hat{\theta}_{(t+1)}$ is merely solving a constrained least squares problem, for which various algorithms are guaranteed to find the global minimum under mild conditions on data [21, 33]. Generally speaking, both steps are easy to implement, which makes AltMin more attractive compared with other optimization algorithms that jointly update $\theta$ and $\Sigma$. In the low-dimensional setting, the AltMin algorithm for multi-response regression was initially proposed by [28]. For the high-dimensional counterpart with sparse parameters, previous works [32, 23, 31] considered the regularized MLE approaches, which are also solved by AltMin-type algorithms. Unfortunately, none of those works provide finite-sample statistical guarantees for their algorithms. The first attempt to establish the non-asymptotic error bound of this AltMin approach is made by [20] for low-dimensional regime, with a brief extension to sparse parameter setting using iterative hard thresholding method [21]. But they did not allow more general structure of the parameter. One of the closely related work is [10] with a focus on general parameter structures captured by norms. They proposed an alternating estimation framework, in which the generalized Dantzig selector [8] is used for $\theta$-step as an alternative to the regularized and the constrained estimators.

The AltMin technique has also been applied to many other estimation problems, such as matrix completion [19], phase retrieval [27], and mixed linear regression [44]. However, the current theoretical understanding of AltMin is still incomplete. Including the aforementioned works, the statistical guarantees for non-convex AltMin procedures are often shown under the resampling assumption, which assumes that each iteration receives a fresh sample. Albeit this can be achieved by partitioning the data into disjoint subsets and using different batches in each update, people seldom do so in practice, as it usually results in worse performance than using all data in every iteration. From the theoretical perspective, the resampling assumption oversimplifies the analysis of the algorithm used in practice, which may otherwise require sophisticated proof techniques [36].

In this paper, we aim at a better way to bound the statistical error of the above AltMin procedure for general structure-inducing $f$. In principal, non-asymptotic statistical analyses for the high dimension typically involve bounding suprema of stochastic processes [26, 34, 42, 30]. The difficulty of analyzing AltMin lies in the dependency between the data and the obtained iterates, and the lack of independence prevents applications of various concentration inequalities to the suprema of the target processes. The resampling assumption facilitates the analysis of AltMin by assuming access to new data that are independent of previous iterates. In contrast to resampling, we here resort to uniform bounds to tackle the dependency issue. That is, instead of dealing with the processes involving the specific iterates generated by AltMin, we try to bound their worst-case counterparts that consider all possible iterates.
before running AltMin. This solution to dependency ends up with more complicated stochastic processes, which need careful treatment. By applying generic chaining \cite{43}, an advanced tool from probability theory, we are able to obtain the desired bounds for the processes under consideration, and eventually express the error bound in terms of a complexity measure called Gaussian width \cite{16,7} (see Section 3.1). In particular, we analyze the AltMin procedure under two different choices of initialization, one with an arbitrarily initialized iterate and the other starting at a point close to $\theta_*$. The $L_2$-error for both types of AltMin is shown to converge geometrically to certain minimum achievable error $e_{\text{min}}$ with overwhelming probability, i.e.,

$$
\|\hat{\theta}(T) - \theta_*\|_2 \leq e_{\text{min}} + \rho_* T \cdot (\|\hat{\theta}_0 - \theta_*\|_2 - e_{\text{min}})
$$

(5)

where $\rho_* < 1$ is the contraction factor and $e_{\text{min}}$ is given by

$$
e_{\text{min}} = O \left( \frac{w(C) + m}{\sqrt{n}} \right) \quad \text{(arbitrary initialization)} ,
$$

(6)

$$
e_{\text{min}} = O \left( \frac{w(C)}{\sqrt{n}} \right) \quad \text{(good initialization)} .
$$

(7)

Here $w(C)$ is the Gaussian width of a set $C$ related the structure of $\theta_*$ (see Definition 2). Surprisingly the error for good initializations matches the resampling-based result up to some constant, which requires more fresh data to achieve such a bound. In general, our work improves the results in \cite{20,10} in several aspects. First, our analysis does not rely on the resampling assumption. Second, our statistical guarantees work for general sub-Gaussian noise while \cite{20} and \cite{10} only considered Gaussian noise. Third, we allow the complexity function $f$ to be non-convex, whereas \cite{10} required $f$ being a norm. Last but not least, our result suggests that when the amount of data is adequate for the error bound \cite{6} to meet the requirement of good initialization, the AltMin with arbitrary initialization can achieve the same level of error as the well-initialized one. Although this type of guarantee for arbitrary initializations was discovered for other problems \cite{43}, it has not been revealed for the multi-response regression, and our proof technique is also different from the existing ones.

The rest of the paper is organized as follows. In Section 2 we outline our strategy for combating non-convexity and present the algorithmic details of the AltMin procedure for structured multi-response regression. In Section 3 we present the statistical guarantees for the AltMin algorithm under suitable probabilistic assumptions. We provide some experimental results in Section 4 and conclude in Section 5. All proofs are deferred to the supplementary material.

2 Strategy to Conquer Non-Convexity

For many statistical estimation problems, we can construct the estimator of the underlying model parameter $w_*$, by minimizing certain loss function on the given sample $D$,

$$
\hat{w} = \arg\min_{w \in \mathcal{W}} L(w; D) .
$$

(8)

In order to show the recovery guarantee for non-convex estimation, there are mainly two commonly-used strategies. One strategy is to show certain local convergence in a neighborhood $\mathcal{N}$ of the global minimizer $\hat{w}$ of \cite{6,20,39,45,25}. With a proper initialization inside $\mathcal{N}$, subsequent iterates produced by some local search might be able to converge to $\hat{w}$, whose statistical error is expected to be small. This strategy is particularly suitable for the noiseless setting, as $\hat{w}$ is equal to $w_*$, and most of the existing works use gradient descent type or its variants as workhorse algorithms. The other strategy is to show that there is no spurious local minima of $L$ under the assumed statistical models, so that any optimization algorithms that provably converge to local minima will suffice for a good estimation \cite{34,35,8,13,24,12}.

For our multi-response regression problem, however, it is difficult to apply the aforementioned strategies. First, bounding the statistical error of the global minimizer is nontrivial in the noisy setting, especially when the objective $L(w)$ involves more than one set of variables like the multi-response regression, let alone characterizing the equivalence of all local minima. Second, the gradient-based local search is inefficient for the problem \cite{2}, since the update of $\Sigma$ involves matrix inversion and projection onto positive semidefinite (PSD) cone. In contrast, AltMin procedure has a closed-form solution to $\Sigma$-step, which is preferred in this setting.
In this work, we consider another strategy for the non-convex estimation in which \( w ( w_0 ) \) is composed of two parameters, \( a \) and \( b ( a_\ast , b_\ast ) \). The loss \( L \) is assumed to jointly non-convex over \( a \) and \( b \), but might be marginally convex w.r.t. \( a ( b ) \) when \( b ( a ) \) is fixed. When the marginal subproblems are easy to solve, alternating minimization procedure is appealing for the purpose of estimation, which is true for the multi-response regression. The AltMin algorithm executes the following updates,

\[
\hat{a}_{t+1} = \text{argmin}_{a \in A} L(a, \hat{b}_t; D), \quad \hat{b}_{t+1} = \text{argmin}_{b \in B} L(\hat{a}_{t+1}, b; D). \tag{9}
\]

The basic idea for showing the statistical guarantees of AltMin is to derive the statistical error bounds for both the \( a \)- and \( b \)-steps when the other parameter is fixed to the latest estimate. Since both subproblems in (9) are usually simpler, the separate errors might be easier to characterize than considered jointly, which are ideally of the form,

\[
d_1 ( \hat{a}_{t+1} , a_\ast ) \leq e_1 \left( d_2 ( \hat{b}_{t+1} , b_\ast ) \right), \quad d_2 ( \hat{b}_{t+1} , b_\ast ) \leq e_2 \left( d_1 ( \hat{a}_{t+1} , a_\ast ) \right). \tag{10}
\]

The function \( d_1 \) (respectively \( d_2 \)) characterizes the closeness between \( \hat{a}_{t+1} \) and \( a_\ast \), \( \hat{b}_{t+1} \) and \( b_\ast \), which is nonnegative with \( d_1 ( a_\ast , a_\ast ) = 0 \) (\( d_2 ( b_\ast , b_\ast ) = 0 \)) but not necessarily a metric. The choice of \( d_1 \) and \( d_2 \) depends on the goal of analysis for the problem under consideration, and a suitable combination of \( d_1 \) and \( d_2 \) may facilitate the proof. The upper bound \( e_1 \) (respectively \( e_2 \)) may depend on other quantities such as \( n \), but our emphasis is the dependence on the estimation accuracy of \( b \) (\( a \)). It is natural to expect that \( e_1 ( e_2 ) \) will shrink as \( \hat{b}_t \) (\( \hat{a}_t \)) moves closer to \( b_\ast \) (\( a_\ast \)). Under this condition, we can apply the bounds in (10) alternatingly and recursively

\[
d_1(\hat{a}(T), a_\ast) \leq e_1 \left( d_2 \left( \hat{b}(T-1), b_\ast \right) \right) \leq \ldots \leq e_1 \left( \underbrace{e_2 \left( \ldots e_1 \left( d_2 \left( \hat{b}(0), b_\ast \right) \right) \ldots \right) }_{\text{composition of } T e_1(\cdot) \text{ and } T-1 e_2(\cdot)} \right) \tag{11}
\]

\[
d_2(\hat{b}(T), b_\ast) \leq e_2 \left( d_1 \left( \hat{a}(T), a_\ast \right) \right) \leq \ldots \leq e_2 \left( \underbrace{e_1 \left( \ldots e_1 \left( d_2 \left( \hat{b}(0), b_\ast \right) \right) \ldots \right) }_{\text{composition of } T e_2(\cdot) \text{ and } T e_1(\cdot)} \right) \tag{12}
\]

which may imply the error of \( \hat{a}(T) \) and \( \hat{b}(T) \) under other metrics of interest as well. Compared with the previous strategies, one notable difference of our treatment is that we do not care about the optimization convergence of AltMin, as we neither characterize the error of any local minimizers of \( L(\cdot) \) nor show any iterate convergence to those minimizers. Instead the ingredients we need are simply the statistical error bounds in (10). Given this fact, our analysis can be extended to the alternating estimation (AltEst) procedure (10) that need not optimize a joint objective over \( a \) and \( b \) and certainly cannot be handled by the earlier strategies.

In order to get (10), the analysis for each AltMin step is often confronted with a technical challenge due to the dependency between data and the iterates obtained so far, which is bypassed by many existing analyses via the resampling assumption. Essentially the resampling-based result states that with any fixed \( \hat{b}(t) (\hat{a}(t+1)) \), given a fresh sample \( D(t) \) independent of \( \hat{b}(t) (\hat{a}(t+1)) \), the next iterate \( \hat{a}(t+1) (\hat{b}(t+1)) \) satisfies the corresponding bound in (10) with high probability. To avoid the resampling, we leverage the idea of uniform bounds [40], which aims to show that given a sample \( D \), the bounds in (10) hold uniformly with high probability for all possible value of the input \( \hat{b}(t) \) and \( \hat{a}(t+1) \). This argument asks for no fresh data in each iteration, and the probability of the error bounds being true does not deteriorate with growing number of iterations. For structured multi-response regression, we will focus on the AltMin procedure shown in Algorithm [1]. For the rest of the paper, \( C_0 , C_1 , c_0 , c_1 \) and so on are reserved for absolute constants.

### 3 Statistical Guarantees of Alternating Minimization

In this section, we apply the resampling-free analysis strategy introduced in Section [2] to the multi-response regression problem, for which \( a = \Sigma \) and \( b = \theta \). First we introduce a few notations. Given a set \( A \subseteq \mathbb{R}^p \), define \( \text{cone} A = \{ c \cdot a \mid c \geq 0, a \in A \} \). We denote the smallest and the largest eigenvalue of \( \Sigma \), as \( \sigma_- \) and \( \sigma_+ \), and assume \( \text{Diag}(\Sigma) = I_{n \times n} \) throughout the paper for simplicity. In addition, we drop the subscripts indexing the iteration, and analyze both \( \Sigma \)-update and \( \theta \)-update in


\textbf{Algorithm 1} Alternating minimization for multi-response regression

\textbf{Input:} Number of iterations $T$, Data $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^n$ and Tuning parameter $\lambda$

\textbf{Output:} Estimated $\hat{\theta}\left(\tau\right)$

1: Initialize $\hat{\theta}\left(0\right)$ (e.g., solving [4] with $\hat{\Sigma}\left(0\right) = I)$
2: for $t := 0$ to $T - 1$ do
3: \hspace{1em} Compute $\hat{\Sigma}\left(t+1\right)$ according to (3)
4: \hspace{1em} Compute $\hat{\theta}\left(t+1\right)$ by solving (4)
5: end for
6: return $\hat{\theta}\left(T\right)$

a broader setting, where the other parameter is fixed as a generic input in certain regions, i.e.,

\begin{equation}
\hat{\hat{\Sigma}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{X}_i \theta)(y_i - \mathbf{X}_i \theta)^T ,
\end{equation}

\begin{equation}
\hat{\hat{\theta}}(\Sigma) = \arg\min_{f(\theta) \leq f(\theta_\ast)} \frac{1}{2n} \sum_{i=1}^{n} \left\| \Sigma^{-\frac{1}{2}} (y_i - \mathbf{X}_i \theta) \right\|_2^2.
\end{equation}

Note that here the tuning parameter $\lambda$ in (3) for the $\theta$-step is set as $\lambda = f(\theta_\ast)$, which will be kept for the rest of the analysis. Given the recent progress in non-convex optimization [3], we also assume that $\hat{\theta}(\Sigma)$ can be solved globally despite the potential non-convexity of $f$. The input regions we consider for $\theta$ and $\Sigma$ are respectively given by

\begin{equation}
\mathcal{R} = \left\{ \theta \in \mathbb{R}^p \mid f(\theta) \leq f(\theta_\ast) \right\} ,
\end{equation}

\begin{equation}
\mathcal{M}(c_0) = \left\{ \hat{\Sigma}(\theta) \in \mathbb{R}^{n \times m} \mid \theta \in \mathcal{R}, \|\theta - \theta_\ast\|_2 \leq c_0 \right\} ,
\end{equation}

in which $c_0$ is the error tolerance to be specified for the initialization. Note that the input region $\mathcal{M}(c_0)$ implicitly depends on $\mathcal{R}$ as well as the sample $\mathcal{D} = \{(x, y)\}_{i=1}^n$ used for computing $\hat{\Sigma}(\theta)$.

\subsection{3.1 Preliminaries}

To apply the proof strategy for AltMin, we first define the distance function $d_1$ and $d_2$.

\textbf{Definition 1 (distance functions)} The distance functions for $\Sigma$-step and $\theta$-step are defined as

\begin{equation}
d_1(\Sigma, \Sigma_\ast) = \frac{\xi(\Sigma)}{\xi(\Sigma_\ast)} - 1, \text{ where } \xi(\Sigma) = \frac{\sqrt{\text{Tr}(\Sigma^{-1} \Sigma_\ast \Sigma^{-1})}}{\text{Tr}(\Sigma^{-1})},
\end{equation}

\begin{equation}
d_2(\theta, \theta_\ast) = \|\theta - \theta_\ast\|_2.
\end{equation}

Although $d_1$ may look odd at first glance, it actually arises as a natural choice after we fix $d_2$, as the $L_2$-error of $\theta$ is our primary goal in the statistical analysis. It is worth noting that $\xi(\Sigma)$ is minimized at $\Sigma = \Sigma_\ast$. The following definition is critical to the analysis for general structures of $\theta_\ast$. [7].

\textbf{Definition 2 (error spherical cap)} For a structure-inducing $f$, its error spherical cap is defined as

\begin{equation}
C = \text{cone } \left\{ \mathbf{u} \in \mathbb{R}^p \mid f(\theta_\ast + \mathbf{u}) \leq f(\theta_\ast) \right\} \cap S^{p-1},
\end{equation}

where $S^{p-1} = \{ \mathbf{u} \mid \|\mathbf{u}\|_2 = 1 \}$ is the unit sphere of $\mathbb{R}^p$.

The probabilistic analysis of $d_1$ and $d_2$ is built upon the concept of sub-Gaussian vectors and matrices, which are defined below.

\textbf{Definition 3 (sub-Gaussian vector and matrix)} A vector $\mathbf{x} \in \mathbb{R}^p$ is said to be sub-Gaussian if its $\psi_2$-norm satisfies,

\begin{equation}
\|\mathbf{x}\|_{\psi_2} = \sup_{\mathbf{u} \in S^{p-1}} \|\langle \mathbf{x}, \mathbf{u} \rangle\|_{\psi_2} \leq \kappa < +\infty ,
\end{equation}

\[\[\]\]
Another key ingredient in the analysis is the complexity measure of the parameter structure captured where $\|\cdot\|_{\psi_2}$ with any input $\Sigma \in \mathcal{M}(+\infty)$. A matrix $X \in \mathbb{R}^{n \times p}$ is sub-Gaussian if the following $\psi_2$-norm for $X$ is finite,

$$\|X\|_{\psi_2} = \sup_{u \in \mathbb{S}^{p-1}} \left\|u^T \Gamma_v^{-\frac{1}{2}} X^T v\right\|_{\psi_2} \leq \kappa < +\infty,$$

(21)

where $\Gamma_v = \mathbb{E}[X^T v v^T X]$. Further, $\Gamma_v$ for any $v \in \mathbb{S}^{m-1}$ is assumed to satisfy the condition $0 < \mu^- \leq \lambda_{\min}(\Gamma_v) \leq \lambda_{\max}(\Gamma_v) \leq \mu^+ < +\infty$, for some constants $\mu^-$ and $\mu^+$.

This definition is adopted from [41][20]. If rows of $X$ are i.i.d. copies of an isotropic sub-Gaussian random vector $x$ with $\|x\|_{\psi_2} \leq \kappa$, it is not difficult to verify that $\|X\|_{\psi_2} \leq C \kappa$ for a universal constant $C$, and $\mu^- = \mu^+ = 1$. Our assumptions on $\{X_i\}$ and $\{\hat{\eta}_i\}$ are given below.

(A1) The designs $X_1, \ldots, X_n$ are i.i.d. copies of a sub-Gaussian $X$ with parameter $\kappa$, $\mu^-$ and $\mu^+$.

(A2) The isotropic noises $\hat{\eta}_1, \ldots, \hat{\eta}_n$ are i.i.d. copies of a sub-Gaussian $\eta$ with parameter $\tau$.

Another key ingredient in the analysis is the complexity measure of the parameter structure captured by $C$, which turns out to be the notion of Gaussian width [16].

**Definition 4 (Gaussian width)** The Gaussian width $\omega(A)$ of a set $A \subseteq \mathbb{R}^p$ is defined as

$$\omega(A) = \mathbb{E}_{x \sim N(0, I)} \left[ \sup_{u \in A} \langle g, u \rangle \right].$$

(22)

Gaussian width is easy to calculate or bound for the error spherical caps induced by many $f$ of interest [17][9]. Based on Gaussian width, the proofs of the error bounds utilize a powerful tool from probability theory, called generic chaining [37]. We refer the interested readers to the recent monograph [38] and references therein.

### 3.2 Error Bound for Arbitrary Initializations

Given the definitions of distance function $d_1$ and $d_2$, we first focus on the separate error bounds for the $\Sigma$-step and the $\theta$-step in [13] and [14]. To allow arbitrary initializations, we consider the tolerance of initialization error $\varepsilon_0 = +\infty$, which appears in the definition of $\mathcal{M}(\varepsilon_0)$.

**Lemma 1 (error bound for $\Sigma$-estimation)** Under the assumptions (A1) and (A2), if the sample size $n \geq C_0 \max \left\{1, \tau^4, \kappa^4 \left(\frac{\sigma^2 \mu^+}{\sigma^2 \mu^-}\right)^2\right\} \cdot \max \left\{m, \frac{w_1(C)}{m}\right\}$, with probability at least $1 - C_2 \exp(-C_1 m)$, $\hat{\Sigma}(\theta)$ given in [13] is invertible for any $\theta \in \mathbb{R}$ and its error satisfies

$$d_1 \left(\hat{\Sigma}(\theta), \Sigma_\star\right) \leq C_3 \tau^2 \sqrt{\frac{m}{n}} + C_4 \frac{\mu^+}{\sigma^\star} \cdot d_2 \left(\theta, \theta_\star\right).$$

(23)

**Remark:** If $\theta = \theta_\star$, the $\Sigma$-step computes the sample covariance of the noise, for which $d_2(\theta, \theta_\star) = 0$, and the remaining $O\left(\sqrt{\frac{m}{n}}\right)$ term in (23) is the typical statistical rate for covariance estimation.

**Lemma 2 (error bound for $\theta$-estimation)** Under the assumptions (A1) and (A2), if the sample size $n \geq C_0 \max \left\{1, \tau^4, \kappa^4 \left(\frac{\sigma^2 \mu^+}{\sigma^2 \mu^-}\right)^2\right\} \cdot \max \left\{m, \frac{w_1(C)}{m}\right\}$, then with probability at least $1 - C_2 \exp(-C_1 m)$, the following bound holds for $\hat{\theta}(\Sigma)$ given in [14] with any input $\Sigma \in \mathcal{M}(+\infty)$,

$$d_2 \left(\hat{\theta}(\Sigma), \theta_\star\right) \leq (1 + d_1 (\Sigma, \Sigma_\star)) \cdot \frac{C_4 \kappa \sqrt{\mu^+}}{\mu^- \sqrt{\text{Tr}(\Sigma_\star^{-1})}} \cdot \frac{m + w(C)}{\sqrt{n}},$$

(24)

where $\xi(\Sigma)$ is given in Definition [7].
Remark: For $\Sigma = \Sigma_*$ and $\Sigma = I$, $\theta$-step corresponds to the oracle estimator $\hat{\theta}_{\text{or}}$ and the ordinary least squares (OLS) estimator $\hat{\theta}_{\text{odn}}$, respectively, i.e.,

$$\hat{\theta}_{\text{or}} = \arg\min_{f(\theta) \leq f(\theta_0)} \frac{1}{2m} \sum_{i=1}^{n} \left\| \Sigma_0^{-\frac{1}{2}} (y_i - X_i \theta) \right\|_2^2,$$

(25)

$$\hat{\theta}_{\text{odn}} = \arg\min_{f(\theta) \leq f(\theta_0)} \frac{1}{2m} \sum_{i=1}^{n} \left\| y_i - X_i \theta \right\|_2^2.$$

(26)

An analysis similar to [10] shows that with high probability the $L_2$-errors of $\hat{\theta}_{\text{or}}$ and $\hat{\theta}_{\text{odn}}$ satisfy

$$\left\| \hat{\theta}_{\text{or}} - \theta_{\text{oc}} \right\|_2 \leq \frac{C_1 \kappa \sqrt{\mu^+}}{\mu^- \sqrt{\text{Tr}(\Sigma_*)^2}} \cdot \frac{w(C)}{\sqrt{n}} \equiv e_{\text{or}},$$

(27)

$$\left\| \hat{\theta}_{\text{odn}} - \theta_{\text{oc}} \right\|_2 \leq \frac{C_1 \kappa \sqrt{\mu^+}}{\mu^- \sqrt{m}} \cdot \frac{w(C)}{\sqrt{n}} \equiv e_{\text{odn}},$$

(28)

which indicates that the oracle estimator improves the OLS by a factor of

$$\frac{e_{\text{or}}}{e_{\text{odn}}} = \sqrt{\frac{m}{\text{Tr}(\Sigma_*)^2}}.$$  

(29)

In practice, this improvement can be significant, especially when there is strong cross-correlation among the responses, such that $\Sigma_*$ is close to singular.

By assembling Lemma 1 and 2, we obtain the following theorem for the error of AltMin, which exhibits a geometrical convergence to certain minimum achievable error.

**Theorem 1 (error bound for arbitrarily-initialized AltMin)** Under the assumptions (A1) and (A2), if the sample size $n \geq C_0 \cdot \max \left\{ 1, \tau^2, \kappa^4 \left( \frac{\mu^+ \sigma_0}{\mu^+ - \sigma_0} \right)^2, \kappa^2 \left( \frac{\mu^+ - \sigma_0}{\mu^+} \right)^2 \right\} \cdot \max \left\{ \frac{w(C)}{m}, m \right\}$, and $\theta_{(0)}$ is a feasible initialization (i.e., $f(\theta_{(0)}) \leq f(\theta_*)$), then with probability at least $1 - C_2 \exp(-C_1 m)$, the following error bound holds for $\hat{\theta}_{(T)}$ returned by Algorithm 2

$$\left\| \hat{\theta}_{(T)} - \theta_* \right\|_2 \leq e_{\text{min}} + \rho_n^T \cdot \left( \left\| \hat{\theta}_{(0)} - \theta_{*} \right\|_2 - e_{\text{min}} \right),$$

(30)

in which $\rho_n$ and $e_{\text{min}}$ satisfy the inequalities below with $\delta_n = C_5 \tau^2 \sqrt{\frac{\mu}{n}} \leq \frac{1}{4}$,

$$\rho_n \leq \frac{C_3 \kappa \mu^+}{\mu^- \sqrt{\sigma_0} \sqrt{\text{Tr}(\Sigma_*)^2}} \cdot \frac{m + w(C)}{\sqrt{n}} \leq \frac{1}{2},$$

(31)

$$e_{\text{min}} \leq \frac{C_4 \kappa \sqrt{\mu^+}}{\mu^- \sqrt{\text{Tr}(\Sigma_*)^2}} \cdot \frac{m + w(C)}{\sqrt{n}} \cdot \frac{1 + \delta_n}{1 - \rho_n}.$$  

(32)

Remark: The inequality (30) indicates that the upper bound of the error for AltMin procedure will decrease geometrically to the minimum achievable error $e_{\text{min}}$ with rate $\rho_n$. Though the initialization condition $f(\theta_{(0)}) \leq f(\theta_*)$ may not be true for arbitrary $\theta_{(0)}$, it should be satisfied by the first iterate $\theta_{(1)}$, from which Theorem 1 starts to apply.

Note that the $\rho_n$ in (30) not only controls the convergence rate of error, but also affects the value of $e_{\text{min}}$. The $e_{\text{min}}$ is of the same order as the right-hand side of (24) with $\Sigma = \Sigma_*$, which has an extra additive $O \left( \frac{m}{\sqrt{n}} \right)$ term compared with $e_{\text{or}}$. This is due to the uniformity considered for the $\theta$-step over all $\Sigma \in \mathcal{M}(+\infty)$. To improve the bound for AltMin, we can consider a small $e_0$ for $\mathcal{M}(e_0)$.

### 3.3 Improved Bound with Good Initializations

As discussed above, we consider a smaller input region $\mathcal{M}(e_0)$ for the $\theta$-step with $e_0 = \sqrt{\frac{\sigma_0}{\mu^+}}$. Before presenting the results, we introduce the set called error spherical sector.
Definition 5 (error spherical sector) For a structure-inducing \( f \), its error spherical sector is defined as
\[
S = \text{cone}\{u \in \mathbb{R}^p \mid f(\theta_* + u) \leq f(\theta_*)\} \cap B^{p-1},
\]
where \( B^p = \{u \mid \|u\|_2 \leq 1\} \) is the unit ball of \( \mathbb{R}^p \).

Geometrically \( S \) is closely related to the previously defined \( C \) in \([19]\), and their Gaussian widths satisfy \( w(S) \leq w(C) + c \) for some universal constant \( c \). Based on this definition, the following theorem characterizes the sharpened error of AltMin under good initializations.

**Theorem 2 (error bound for well-initialized AltMin)** Under the assumptions (A1) and (A2), if the sample size \( n \geq C_0 \cdot \max\left\{1, \tau^4, \kappa^4 \left(\frac{\mu^+}{\mu^+ - \sigma^+}\right)^2, \kappa^2 \left(\frac{\mu^+}{\mu^+ - \sigma^+}\right)^2\right\}\cdot \max\left\{\frac{w^4(C)}{m^3} + m^2 \mid \mathbb{E}^2\right\}, \]

and a feasible initialization \( \theta_0 \) satisfies \( \|\theta_0 - \theta_*\|_2 \leq \sqrt{\frac{\sigma^2_{\theta^*}}{\mu}} \), then with probability at least \( 1 - C_2 \exp\left(-C_1 \cdot \min\{w^2(C), m\}\right) \), the error bound \( \|\hat{\theta}_m - \theta_*\| \leq \sqrt{\frac{\sigma^2_{\theta^*}}{\mu}} \) holds for \( \hat{\theta}_m \) returned by Algorithm 1 with \( \rho_n \) and \( \epsilon_{\min} \) satisfying
\[
\rho_n \leq \frac{C_3 \kappa \mu^{\star}}{\mu^{\star} \sqrt{\sigma^{\star} \text{Tr}(\Sigma^{\star})}}, \quad \frac{w(S)}{\sqrt{n}} \leq \frac{1}{2},
\]
\[
\epsilon_{\min} \leq \frac{C_3 \kappa \mu^{\star}}{\mu^{\star} \sqrt{\text{Tr}(\Sigma^{\star})}}, \quad \frac{w(S)}{\sqrt{n}} \cdot \frac{1 + \delta_n}{1 - \rho_n},
\]
where \( \delta_n \) is the same as the one given in Theorem 1.

**Remark:** Since \( w(S) \) only differs from \( w(C) \) by a constant, the above error bound matches the order of the oracle error \( \epsilon^{\text{orc}} \). For instance, if \( \theta_* \) is \( s \)-sparse and \( f = \|\cdot\|_0 \), then \( w(S) \) and \( \epsilon_{\min} \) satisfy,
\[
\frac{w(S)}{\sqrt{n}} \sim O\left(\sqrt{s \log p}\right) \quad \Rightarrow \quad \epsilon_{\min} = O\left(\frac{\sqrt{s \log p}}{n}\right)
\]

The initialization condition is a result of setting a small value of \( \epsilon_0 \), which yields an improved version of Lemma 2, so that we can obtain a better bound in Theorem 2. A reasonably good initialization of \( \theta_0 \) can be obtained by solving OLS \( \hat{\theta}_{\text{ode}} \), whose error bound is given \([27]\). On the other hand, the iterates obtained by running arbitrarily-initialized AltMin may also satisfy the initialization requirements as Theorem 1 guarantees a moderate error. Once the requirements are met during the iteration, the arbitrarily-initialized AltMin can attain this sharper bound as well as the well-initialized.

## 4 Experiments

In this section, we present some experimental results to support our theoretical analysis. Specifically, we focus on the sparsity structure of \( \theta_* \), and consider \( L_0 \)-cardinality as complexity function \( f \). Throughout the experiment, we fix problem dimension \( p = 1000 \), sparsity level of \( \theta_* \) \( s = 20 \), and number of iterations \( T = 10 \). Entries of \( X \) of \( \eta \) are generated by i.i.d. standard Gaussian, and \( \theta_* = [1, \ldots, 1, -1, \ldots, -1, 0, \ldots, 0]^T \). \( \Sigma_* \) is given as a block diagonal matrix with \( \Sigma' = \begin{bmatrix} 1_a & a \\ a & 1 \end{bmatrix} \)

replicated along the diagonal. All the plots are obtained based on the average over 100 random trials.

First we set \( a = 0.9, m = 10 \), and vary sample size \( n \) from 30 to 80. We run the AltMin initialized by both OLS and Gaussian random vector, where \( \theta \)-step is solved by the hard-thresholding pursuit (HTP) algorithm \([11]\). The error plots are shown in Figure 1. Second, we fix \( m = 10 \), and vary the parameter \( a \) in \( \Sigma_* \) from 0.5 to 0.9 for \( n = 30, 40, 50 \) and 60. The plots in Figure 2(a) shows the error of AltMin against \( a \). As indicated by \([29]\), the improvement of the oracle least squares over the ordinary one is amplified with increasingly large \( a \). Figure 2(b) compares the actual ratio of \( \epsilon^{\text{orc}} \) to \( \epsilon^{\text{ode}} \) and the suggested one. Finally we fix \( a = 0.8 \), and the number of responses \( m \) ranges from 10 to 18 for \( n = 30, 40, 50 \) and 60. The results are presented in Figure 2(c) and 2(d).
Figure 1: (a) A phase transition is observed for the randomly-initialized AltMin around $n = 40$, whose error is on a par with the well-initialized for $n \geq 40$. This coincides with the remark for Theorem 2. Also, the error of AltMin is close to the oracle estimator, which is significantly better than OLS. (b) Our theoretical results suggest that a larger sample size leads to smaller $\rho_n$, thus AltMin converge faster as shown in the plots.

Figure 2: (a) With $\alpha$ varying from 0.5 to 0.9, the responses become increasingly correlated and the error of AltMin reduces more quickly. (b) The actual ratio of $e_{\text{orc}}$ to $e_{\text{odn}}$ is very close the predicted one given by (29). (c) As $m$ increases from 10 to 18, the error of AltMin does not decrease drastically. The main reason is the increasingly large error in the estimation of $\Sigma_*$. (d) Compared with the error of OLS, the advantage of AltMin becomes marginal with growing $m$, while its gap with the oracle estimator is widened.

5 Conclusions

In this paper, we investigate the alternating minimization (AltMin) algorithm for high-dimensional multi-response linear models, which allow general structures of the underlying parameter. In particular, we present a resampling-free analysis for the statistical error of the non-convex AltMin procedure. Our error bound matches the resampling-based result up to some constant, which is of the same order as the oracle estimator. Above all, the error bounds suggest that the arbitrarily-initialized AltMin is able to attain the same level of estimation error as the one with good initializations.

Acknowledgements

The research was supported by NSF grants IIS-1563950, IIS-1447566, IIS-1447574, IIS-1422557, CCF-1451986, CNS-1314560, IIS-0953274, IIS-1029711, NASA grant NNX12AQ39A, and gifts from Adobe, IBM, and Yahoo.
References


