Journal of Machine Learning Research 25 (2024) 1-85 Submitted 3/23; Revised 10/23; Published 7/24

Statistical analysis for a penalized EM algorithm in high-dimensional mixture linear regression model

Ning Wang NINGWANGBNU@BNU.EDU.CN
Center for Statistics and Data Science
Beijing Normal University
Zhuhai, Guangdong, China
Xin Zhang HENRY@STAT.FSU.EDU
Qing Mai QMAI@FSU.EDU
Department of Statistics
Florida State University
Tallahassee, FL, 32306, USA

Editor: Xiaotong Shen

Abstract

The expectation-maximization (EM) algorithm and its variants are widely used in statistics. In high-dimensional mixture linear regression, the model is assumed to be a finite mixture of linear regression and the number of predictors is much larger than the sample size. The standard EM algorithm, which attempts to find the maximum likelihood estimator, becomes infeasible for such model. We devise a group lasso penalized EM algorithm and study its statistical properties. Existing theoretical results of regularized EM algorithms often rely on dividing the sample into many independent batches and employing a fresh batch of sample in each iteration of the algorithm. Our algorithm and theoretical analysis do not require sample-splitting, and can be extended to multivariate response cases. The proposed methods also have encouraging performances in numerical studies.

Keywords: EM algorithm, High-dimensional regression, Mixture model.

1. Introduction

Consider a univariate response $Y \in \mathbb{R}$ and a p-dimensional predictor $X \in \mathbb{R}^p$. The mixture linear regression model assumes that

$$Y = \beta_k^T X + \epsilon, \text{ for } k = 1, \cdots, K, \text{ with probability } \omega_k > 0, \tag{1}$$

where $K \geq 2$ is the number of mixtures, $\sum_{k=1}^K \omega_k = 1$, $\epsilon \sim N(0, \sigma^2)$, $\sigma^2 > 0$, is independent of X, and β_k is the p-dimensional regression coefficient vector that characterizes the linear relationship between Y and X in the k-th mixture. By introducing a latent variable $W \in \{1, \cdots, K\}$, independent of X, model (1) is equivalent to

$$\mathbb{P}(W = k) = \omega_k, \quad Y \mid (X, W = k) \sim N(\beta_k^T X, \sigma^2). \tag{2}$$

We consider the high-dimensional joint estimation of all the β_k's and provide a general estimation procedure with strong theoretical guarantees. The latent mixtures, indicated by

the latent variable W in (2), make the estimation problem much more challenging than the linear regression, especially in high dimensions. We assume that only a subset of predictors, indexed by $S \subset \{1, \cdots, p\}$, is relevant to the regression. Therefore, $(\beta_k)_{S^c} = 0$ for all k, where S^c is the complement of S. The group lasso penalty (Yuan and Lin, 2006) is naturally applied to those p-dimensional β_k vectors in the maximization steps of our regularized EM algorithm to select relevant variables across all the mixtures.

The mixture linear model and finite mixture models in general, are widely used to account for heterogeneity in data analysis (e.g., Turner, 2000; McLachlan and Peel, 2004; McLachlan et al., 2019). When the number of predictors is not large, the latent mixtures and the model parameters can be estimated using the expectation-maximization (EM, Dempster et al., 1977) algorithm. The EM algorithm is the dominant solution for finding the maximum likelihood estimators of mixture regression models like (1). When ϵ is not normally distributed, the EM algorithm has been modified and extended to robust fitting of mixture linear models under the t- and the Laplace distributions in Yao et al. (2014) and Song et al. (2014), respectively. Leisch (2004) provides computational and implementation details of mixture regression models.

Although the EM algorithm has been extensively used in mixture regression models, it is challenging to establish a rigorous theoretical characterization of the finite-sample estimates in the iterative algorithm. Some groundbreaking progress has been made in recent years. Balakrishnan et al. (2017) laid theoretical foundations for quantifying the EM updates' convergence within statistical precision of the ground truth. For mixture linear model, strong theoretical guarantees of the EM algorithm are often established by focusing on the model with two equal mixtures, $K = 2$ and $\omega_1 = \omega_2 = 1/2$, and symmetric regression coefficient vectors, $\beta_2 = -\beta_1$, (e.g., Kwon et al., 2021, and references therein). Then the EM algorithm is simplified substantially because $\omega_1 = \omega_2 = 1/2$ does not require estimation and, more importantly, the model parameters are reduced to a single vector $\beta \equiv \beta_1 = -\beta_2 \in \mathbb{R}^p$. Many theoretical results are established for the "sample-splitting" EM algorithm, which divides the full data into T equal batches and uses a new batch of samples in each iteration. Without sample splitting, theoretical analysis becomes much more challenging. This is because the function to be maximized in each EM iteration, namely the Q-function, involves both the random samples and the current parameter estimates, which are made independent by sample-splitting. See Balakrishnan et al. (2017) and Klusowski et al. (2019) for recent studies of the sample-splitting EM algorithm in mixture linear models. While the above-mentioned works all focus on low-dimensional models and unpenalized EM algorithm, regularized EM algorithm for high-dimensional mixture linear models is of growing interest in recent years. A selective review is as follows.

Khalili and Chen (2007) studied the variable selection for mixture linear regression with penalized likelihood. When p is fixed and n goes to infinity, they established variable selection consistency and root-n consistency for a possible local maximum. Städler et al. (2010) proposed a lasso-type penalty (Tibshirani, 1996) on the negative-log-likelihood function and obtained non-asymptotic convergence results for the global optimum. However, there is no guarantee for the EM algorithm to attain either the particular local maximum or the global maximum. To rigorously study the entire EM iterative solution sequence in high dimensions, existing results rely on the sample-splitting procedure. For example, convergence results based on sample-splitting algorithms are established for a truncated EM algorithm

(Wang et al., 2015), a penalized EM algorithm (Yi and Caramanis, 2015), and a stochastic EM algorithm (Zhu et al., 2017), all under the assumptions of $K = 2$, $\omega_1 = \omega_2 = 1/2$, $\beta_2 = -\beta_1$, and normally distributed predictors. More recently, with the help of sample splitting, Zhang et al. (2020) systematically studied estimation, confidence intervals, and large-scale hypotheses testing for the mixture linear model. Sample splitting is undoubtedly used to facilitate theoretical analyses, but is not desirable in practice. It remains unknown how to practically choose T, which is both the number of iterations and the number of sample batches, and how it affects the estimation. Moreover, it is believed (e.g., Zhang et al., 2020) that data splitting is unnecessary in numerical studies.

The contributions of this article are multi-fold. The first and most significant contribution is developing a practical penalized EM algorithm for a general high-dimensional mixture linear model and establishing its substantial theoretical guarantees. Our algorithm is equally applicable to random-X and fixed-X and multiple mixtures $K \geq 2$. In our theoretical analysis, we do not require sample splitting and allow a relatively general model. Specifically, we establish a non-asymptotic convergence rate for a two-mixture linear model with unknown proportions $\omega_1, \omega_2 \in (0, 1)$, unrelated two regression parameter vectors $\beta_1, \beta_2 \in \mathbb{R}^p$, and normally distributed predictors with unknown covariance structure. To our best knowledge, we established the first theoretical results for the high-dimensional mixture linear regression under such a general setting without sample splitting. Compared with the high-dimensional linear model literature, the theoretical analysis for the high-dimensional mixture model requires bounding the supremum of random processes and is much more challenging without sample splitting. Our general proof strategy is related to Cai et al. (2019), which studies the penalized EM algorithm for the Gaussian mixture model without data splitting. The complicated relationship between the random response Y and the random predictor X makes the theoretical studies of mixture linear regression even more challenging than the Gaussian mixture model, which only involves random X. Many new concentration results are needed for random processes involving both Y and X. For instance, unlike X_i in the Gaussian mixture model that is sub-Gaussian, the product term $X_i Y_i$ appears frequently in the EM iterates and estimates and is more difficult to bound (Adamczak, 2008). Even for the theoretical analysis of the population EM iterates, double expectations $\mathrm{E}_X\{\mathrm{E}_{Y|X}(\cdot)\}$ are needed than single expectation $\mathrm{E}_X(\cdot)$ in the Gaussian mixture model. With substantial efforts, we obtain a near optimal convergence rate of $\log n\sqrt{s\log p/n}$, with a small price $\log(n)$ to pay for not sample splitting.

The second contribution is our new theoretical insights on model misspecification. Specifically, we analyze how a fixed parameter value σ^2 in the penalized EM algorithm may affect the estimation of β. To the best of our knowledge, it has not been studied in the literature on the mixture linear regression model. In most theoretical studies considering the EM algorithm for the mixture linear regression, the variance σ^2 is often assumed to be the true parameter value σ_*^2 and is a fixed constant. Kwon et al. (2021) considered the convergence results when σ^2 is updated in each M step. However, they only considered a simplified case where the mixture proportions are known to be $1/2$ and are not updated in each M step. Besides, their theoretical studies are limited to low dimensions. In general, when the error variance σ^2 is not correctly specified in the penalized EM algorithm brings bias for estimation. However, for the mixture model with a relatively large signal-to-noise ratio, our theory indicates that the choice of σ^2 has a minor influence on the estimation, and thus an

accurate estimation for it is usually not necessary. This conclusion is further demonstrated by a simulation study.

The third contribution is that we extend the study for the mixture linear regression model to multiple response cases. For the mixture linear regression model with a multivariate response, the naive approach is fitting a mixture linear regression model separately for each element of the response. The major drawback of doing so is each observation may be identified into different clusters when we model each univariate response separately. We illustrated the advantages of considering multiple responses together than handling them separately from both theoretical and numerical aspects. The advantage of considering multiple responses together is also demonstrated in Hyun et al. (2023), which developed a sparse mixture linear regression model to estimate the time-varying data sets (i.e. at each time point, the data satisfy a mixture linear regression model). However, they are interested in developing a time-varying model to analyze a real-world dataset, but no theoretical study is conducted. In contrast, we rigorously characterize the advantage of considering multiple responses simultaneously with statistical theory.

The rest of the article is organized as follows. Section 2 contains the implementation details and discussions about the penalized EM algorithm. Section 3 presents the theory for the penalized EM algorithm and the influence of the choice of σ^2. Simulation studies are presented in Section 4. We then extend the mixture linear regression to multiple response cases and consider its theoretical studies in Section 5. In Section 6, we consider a real data example followed by a short discussion in Section 7. The appendix contains proofs for all the lemmas and theorems and additional implementation details.

2. Estimation

We assume that we collect n independent data points $\{(X_i, Y_i)\}_{i=1}^n$ from model (1). In this section, we do not make additional assumptions on X as our estimation procedure is equally applicable to random or fixed X and allows for both continuous and discrete predictors. Let $\theta = \{\omega_1, \cdots, \omega_k, \beta_1, \cdots, \beta_k\}$ be the unknown parameters to be estimated. In this section, we focus on the studies of the regression coefficients and treat σ^2 as known. The estimation for σ^2 is briefly discussed at the end of this section. We further show in Theorem 5 that misspecification of σ^2 has a relatively small impact on the final estimation.

To motivate our proposal, we first derive the standard EM algorithm and discuss its limitations. The EM algorithm aims to maximize the log-likelihood of $Y \mid X$ over θ, by iteratively updating the sequence of solutions $\{\widehat{\theta}^{(t)}, t = 0, 1, \ldots\}$ via the Expectation-step (E-step) and the Maximization-step (M-step). Recall that W is the latent variable representing the mixtures. Consider the $(t+1)$-th iteration with the current value $\widehat{\theta}^{(t)}$. In the E-step, we calculate the expectation of the log-likelihood of $W \mid (Y, X)$ at the parameter $\widehat{\theta}^{(t)}$. This is known as the Q-function,

$$Q(\theta \mid \widehat{\theta}^{(t)}) = -\frac{1}{2n} \sum_{i=1}^{n} \sum_{k=1}^{K} \widehat{\eta}_{i,k}(\widehat{\theta}^{(t)})(Y_i - X_i^T \beta_k)^2 + \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \widehat{\eta}_{i,k}(\widehat{\theta}^{(t)}) \log(\omega_k), \tag{3}$$

where $\widehat{\eta}_{i,k}(\widehat{\theta}^{(t)}) = \mathbb{P}(W_i = k \mid Y_i, X_i, \widehat{\theta}^{(t)})$. The estimated probability $\widehat{\eta}_{i,k}(\widehat{\theta}^{(t)})$ is given by

$$\widehat{\eta}_{i,k}(\widehat{\theta}^{(t)}) = \frac{\widehat{\omega}_k^{(t)} \phi_{\sigma^2}(Y_i - X_i^T \widehat{\beta}_k^{(t)})}{\sum_{k=1}^K \widehat{\omega}_k^{(t)} \phi_{\sigma^2}(Y_i - X_i^T \widehat{\beta}_k^{(t)})}, \tag{4}$$

where $\phi_{\sigma^2}(u)$ is the probability density function of $N(0, \sigma^2)$. Then, in the M-step, we update $\widehat{\theta}_k^{(t+1)} = \text{argmax}_\theta \, Q(\theta \mid \widehat{\theta}^{(t)})$ by maximizing (3).

Note that the standard EM algorithm is infeasible to high-dimensional problems. If $p \gg n$, even when the latent random variables W_i's are observed, the maximizer of the Q function is not well-defined. Moreover, as W_i is generally latent and unobserved, we need to calculate $\widehat{\eta}_{i,k}(\widehat{\theta}^{(t)})$ in (4), which involves the p-dimensional random vector X and the p-dimensional parameter vectors β_k's.

To estimate the linear mixture regression model in high dimensions, we modify the standard EM algorithm by encouraging sparsity. In high-dimensional statistics, it is often assumed that the coefficients have many elements as zero, i.e, most elements in β_k are zero. But we further assume that β_k has a joint sparsity structure, in that, for most j, we have $\beta_{1j} = \ldots = \beta_{Kj} = 0$, where β_{kj} represents the j-th element of β_k. The group sparsity facilitates the interpretation, as for each j, if $\beta_{1j} = \ldots = \beta_{Kj} = 0$, then the j-th element of X is unimportant for the prediction of Y regardless of which mixture the observation comes from. Moreover, the group sparsity benefits the E-step, because, with straightforward calculation, we can rewrite (4) as

$$\widehat{\eta}_{i,k}(\widehat{\theta}^{(t)}) = \widehat{\omega}_k^{(t)} / \big(\widehat{\omega}_k^{(t)} + \sum_{k' \neq k} \widehat{\omega}_{k'}^{(t)} \exp\big\{(\widehat{\beta}_{k'}^{(t)} - \widehat{\beta}_k^{(t)})^T X_i (Y_i - (\widehat{\beta}_k^{(t)} + \widehat{\beta}_{k'}^{(t)})^T X_i / 2)/\sigma^2\big\}\big). \tag{5}$$

Equation (5) shows an advantage of the group sparsity over the individual sparsity. It implies the j-th element in X is unimportant for the evaluation of $\widehat{\eta}_{i,k}(\widehat{\theta}^{(t)})$ if $\widehat{\beta}_{k'}^{(t)} - \widehat{\beta}_k^{(t)} = 0$ for all k, k'. The group sparsity guarantees that such a situation happens for most elements in X and $\widehat{\eta}_{i,k}(\widehat{\theta}^{(t)})$ is determined by a few elements in X.

With the sparsity assumption, we modify the EM algorithm by imposing the group lasso penalty (Yuan and Lin, 2006) on β_k, for $k = 1, \cdots, K$. Our the penalized EM algorithm replaces M-step by

$$\widehat{\theta}_k^{(t+1)} = \underset{\theta}{\text{argmax}} \left\{ 2Q(\theta \mid \widehat{\theta}^{(t)}) - \lambda_n^{(t+1)} \sum_{j=1}^p \sqrt{\sum_{k=1}^K \beta_{kj}^2} \right\}, \tag{6}$$

where $\lambda_n^{(t+1)} > 0$ is the tuning parameter at the $(t+1)$-th iteration and β_{kj} be the j-th element of β_k. In the E-step, we evaluate $\widehat{\theta}_k^{(t+1)}$ by (5), which is the same as in the standard EM algorithm. Clearly, our penalized EM algorithm reduces to the standard EM algorithm when $\lambda_n^{(t+1)} = 0$ for all $t = 0, 1, \ldots$.

The optimization in (6) is separable in ω_k and β_k. For ω, the updating equation is the same as in the standard EM algorithm, i.e, $\widehat{\omega}_k^{(t+1)} = \sum_{i=1}^n \widehat{\eta}_{i,k}(\widehat{\theta}^{(t)})/n$. For β_k, it amounts to minimizing the following objective function,

$$\ell(\beta_1, \cdots, \beta_K) = \sum_{k=1}^K \beta_k^T \widehat{\Sigma}_k^{(t+1)} \beta_k - 2 \sum_{k=1}^K (\widehat{\rho}_k^{(t+1)})^T \beta_k + \lambda_n^{(t+1)} \sum_{j=1}^p \sqrt{\sum_{k=1}^K \beta_{kj}^2}, \tag{7}$$

where $\widehat{\rho}_k^{(t+1)} = \sum_{i=1}^n \widehat{\eta}_{i,k}(\widehat{\theta}^{(t)}) X_i Y_i / n$ and $\widehat{\Sigma}_k^{(t+1)} = \sum_{i=1}^n \widehat{\eta}_{i,k}(\widehat{\theta}^{(t)}) X_i X_i^T / n$. The convex optimization in (7) can be done efficiently by the groupwise majorization descent algorithm (Yang and Zou, 2015). We provide implementation details in appendix Section B.

The tuning parameter $\lambda_n^{(t)}$ could either be fixed or varying across iterations. For theoretical consideration, in Algorithm 1, we set $\lambda_n^{(t+1)} = \kappa \lambda_n^{(t)} + C_\lambda \sqrt{\log(p)\log(n)^2/n}$, where $0 < \kappa < 1/2$ and C_λ are generic constants, for ease of showing the statistical convergence results. Note that $\lambda_n^{(t)}$ is at the order of $\sqrt{\log(p)\log(n)^2/n}$ when t is large. Thus, in practice, we fix $\lambda_n^{(t)} = \lambda$ for all t and tune λ by the Bayesian information criterion (see our numerical studies). For fixed $\lambda_n^{(t)} = \lambda$ over all iterations, our penalized EM algorithm is maximizing

$$L(\theta) - \lambda/2 \sum_{j=1}^{p} \sqrt{\sum_{k=1}^{K} \beta_{kj}^2}, \tag{8}$$

where $L(\theta)$ is the conditional log-likelihood of $Y \mid X$. The following lemma shows the convergence result of Algorithm 1.

Lemma 1 *If we set $\lambda_n^{(t)} = \lambda$ for all t in Algorithm 1, the objective function from* (8) *evaluated at $\widehat{\theta}^{(t+1)}$ is guaranteed to be no less than the objective function from* (8) *evaluated at $\widehat{\theta}^{(t)}$. That is, the sequence of iterates $\{\widehat{\theta}^{(t)}\}_{t=1}^{\infty}$ generated by Algorithm 1 monotonically increase the value of the objective function from* (8)*.*

In Algorithm 1, σ^2 is treated as a known parameter to facilitate theoretical studies. Treating σ^2 as known is also common in theoretical studies for mixture linear regression (Yi and Caramanis, 2015; Balakrishnan et al., 2017; Zhang et al., 2020, e.g.). We leave the detailed discussion for it in Section 3. In practice, the estimates of σ^2 can be updated straightforwardly in the EM algorithm as $n^{-1} \sum_{i=1}^n \sum_{k=1}^K \widehat{\eta}_{i,k}(\widehat{\theta}^{(t+1)})(Y_i - X_i^T \widehat{\beta}_k^{(t+1)})^2$. Similar estimates are also adopted in numerical studies of Zhang et al. (2020).

Regularization strategies are also used by Yi and Caramanis (2015), Cai et al. (2019), and Zhang et al. (2020) in high-dimensional EM algorithms. However, Cai et al. (2019) considers the clustering problem instead of regression problem. Yi and Caramanis (2015); Zhang et al. (2020) studies the regression problem with the addition of the lasso penalty (Tibshirani, 1996) instead of the group lasso penalty. However, a key difference between our algorithm and theirs is that they require data to be split into T batches, and the (penalized) EM algorithm iterates T times, using one independent batch at each iteration. The sample splitting is rarely performed in standard EM algorithms on low-dimensional data, as it may decrease the computation efficiency. Instead, the sample splitting is an attempt to circumvent technical difficulty in proving the convergence rate. Our proposed EM algorithm does not split sample, but we will show that it achieves a high level of accuracy regardless. Moreover, in addition to investigating the property of the penalized EM algorithm in estimating model (1), we also study the effect of misspecification of σ^2 and the estimation of mixture linear regression when there are multiple responses; see Theorem 5 and Section 5, respectively.

Algorithm 1 Group lasso penalized EM algorithm for model (1) in high dimensions

Input: Initial values $\widehat{\omega}_k^{(0)}$, $\widehat{\beta}_k^{(0)}$, for $k = 1, \cdots, K$, maximum iteration number T, data $\{X_i, Y_i; i = 1, \ldots, n\}$, and initial tuning parameter

$$\lambda_n^{(0)} = C_1(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \cdots \vee |\widehat{\omega}_K^{(0)} - \omega_K^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_2 \vee \cdots \vee \|\widehat{\beta}_K^{(0)} - \beta_K^*\|_2)/\sqrt{s} + C_\lambda \sqrt{\log(n)^2 \log(p)/s}$$

for some positive constants C_1 and C_λ.

Iterate: For $t = 0, \cdots, T-1$, do the following steps until convergence.

- For $i = 1, \cdots, n$, let

$$\widehat{\eta}_{i,k}(\widehat{\theta}^{(t)}) = \widehat{\omega}_k^{(t)} / \Big(\widehat{\omega}_k^{(t)} + \sum_{k' \neq k} \widehat{\omega}_{k'}^{(t)} \exp\big\{ (\widehat{\beta}_{k'}^{(t)} - \widehat{\beta}_k^{(t)})^T X_i \big(Y_i - (\widehat{\beta}_k^{(t)} + \widehat{\beta}_{k'}^{(t)})^T X_i / 2 \big) / \sigma^2 \big\} \Big).$$

- For $k = 1, \cdots, K$, update

$$\widehat{\omega}_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \widehat{\eta}_{i,k}(\widehat{\theta}^{(t)}),$$
$$\widehat{\rho}_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \big(\widehat{\eta}_{i,k}(\widehat{\theta}^{(t)}) X_i Y_i \big),$$
$$\widehat{\Sigma}_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \big(\widehat{\eta}_{i,k}(\widehat{\theta}^{(t)}) X_i X_i^T \big),$$

and update $\widehat{\beta}_k^{(t+1)}$ by minimizing

$$\sum_{k=1}^K \beta_k^T \widehat{\Sigma}_k^{(t+1)} \beta_k - 2 \sum_{k=1}^K (\widehat{\rho}_k^{(t+1)})^T \beta_k + \lambda_n^{(t+1)} \sum_{j=1}^p \sqrt{\sum_{k=1}^K \beta_{kj}^2}.$$

with $\lambda_n^{(t+1)} = \kappa \lambda_n^{(t)} + C_\lambda \sqrt{\log(p) \log(n)^2 / n}$, where $\kappa \in (0, 1/2)$.

Output: $\widehat{\theta}^{(t+1)} = \{\widehat{\omega}_1^{(t+1)}, \ldots, \widehat{\omega}_K^{(t+1)}, \widehat{\beta}_1^{(t+1)}, \ldots, \widehat{\beta}_K^{(t+1)}\}$

3. Theory

3.1 Preliminary

We begin this section with some notations. For numbers a and b, $a \vee b$ means $\max\{a, b\}$. For an integer n, we let $[n]$ denote the set $\{1, \cdots, n\}$. For a vector $x = (x_1, \cdots, x_p)^T$, $\|x\|_0$ is the number of nonzero elements in x, $\|x\|_1 = \sum_{i=1}^p |x_i|$, and $\|x\|_2 = \sqrt{\sum_{i=1}^p x_i^2}$. For a symmetric

matrix A, we denote $\lambda_{min}(A)$ and $\lambda_{max}(A)$ as the smallest and largest eigenvalues of A, respectively. The Frobenius norm of a matrix $A = (a_{ij})$ is defined as $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$. The ℓ_2 norm of a matrix A is $\|A\|_2 = \sqrt{\lambda_{max}(A^T A)}$. For a subset $\mathcal{A} \subseteq \{1, \cdots, p\}$, $\mathcal{A}^c$ denotes its complement. For two sequences of positive numbers a_n and b_n, $a_n = O(b_n)$ means $a_n \leq c b_n$ for a constant $c > 0$ for all n, $a_n = o(b_n)$ means that $a_n / b_n \to 0$ as $n \to \infty$, and $b_n \gg a_n$ means that $a_n = o(b_n)$. Let $\mathcal{S}^{p-1}$ be the unit sphere. For a positive integer $s \leq p/2$, let set $\Gamma_{2p}(2s) = \{\mu \in \mathbb{R}^{2p} : \|\mu_{S^c}\|_1 \leq 5\sqrt{2s}\|\mu_S\|_2 + 2\sqrt{2s}\|\mu\|_2$ for some $S \subset [2p]$ with $|S| = 4s\}$ and $\Gamma(s) = \Gamma_{2p}(2s)_{1:p}$, where $\Gamma_{2p}(2s)_{1:p} = \{\mu_{1:p} : \mu \in \Gamma_{2p}(2s)\}$. For a vector x and a symmetric matrix A, we define $\|x\|_{2,s} = \sup_{\|\mu\|_2 = 1, \mu \in \Gamma(s)} \langle x, \mu \rangle$, and $\|A\|_{2,s} = \sup_{\|\mu\|_2 = 1, \mu \in \Gamma(s)} |\mu^T A \mu|$.

We assume independent random predictors $X_i \sim N(0, \Sigma)$ in our theoretical analysis. This is less restrictive than the assumption of $X_i \sim N(0, I_p)$ in Yi and Caramanis (2015) and Balakrishnan et al. (2017) in that we allow the predictor to be correlated. In this section, we consider $K = 2$, which is a common assumption in theoretical analysis for high-dimensional EM algorithm (Yi and Caramanis, 2015; Cai et al., 2019; Zhang et al., 2020, e.g.). We re-define $\theta = \{\omega_1, \beta_1, \beta_2\}$ since $\omega_2 = 1 - \omega_1$. Let θ^* be the true value of θ, and $\widehat{\theta}^{(t)}$ be the estimate of θ at the t-th step in Algorithm 1. The true parameter space we consider is

$$\Theta^* = \{\theta^* : \omega_1^* \in (c_w, 1 - c_w), \|\beta_k^*\|_0 \leq s, \|\beta_k^*\|_2 \leq M_b, \text{ for } k = 1, 2\}.$$

This is a natural parameter space to consider. The condition $\omega_1^* \in (c_w, 1 - c_w)$ guarantees the sample size from each latent class is large enough. Condition on $\|\beta_k^*\|_2 \leq M_b$ is also similarly used in Yi and Caramanis (2015) under the data splitting framework and is milder than a similar condition $\|\beta\|_1 \leq M_b$ used in Cai et al. (2019), where the EM algorithm for Gaussian mixture model without data splitting is studied.

In theoretical studies, we first assume that the true value of σ^2 denoted as σ_*^2 as a known parameter, namely, the input σ^2 is σ_*^2 in Algorithm 1. Without loss of generality, we assume $\sigma_*^2 = 1$. Treating σ^2 as known is also common in state-of-the-art theoretical studies for mixture linear regression (Yi and Caramanis, 2015; Balakrishnan et al., 2017; Zhang et al., 2020, e.g.). Although we do not analyze β_k's and σ^2 simultaneously, we investigate how the choice of σ^2 influence the estimation of Algorithm 1 in Theorem 5.

Since $\sigma_*^2 = 1$ and $K = 2$, we simplify $\mathbb{P}(W_i = 1 \mid Y_i, X_i, \theta)$ as

$$\eta_{i,1}(\theta) = 1 / \big[1 + (\omega_2/\omega_1) \exp\{(\beta_2 - \beta_1)^T X_i \cdot (Y_i - (\beta_1 + \beta_2)^T X_i / 2)\}\big],$$

and let $\eta_{i,2}(\theta) = \mathbb{P}(W_i = 2 \mid Y_i, X_i, \theta) = 1 - \eta_{i,1}(\theta)$. The following quantities are used repeatedly in our theoretical analysis:

$$\widehat{\omega}_k(\theta) = \frac{1}{n} \sum_{i=1}^n \eta_{i,k}(\theta), \ \widehat{\rho}_k(\theta) = \frac{1}{n} \sum_{i=1}^n \eta_{i,k}(\theta) X_i Y_i,$$

$$\widehat{\Sigma}_k(\theta) = \frac{1}{n} \sum_{i=1}^n \eta_{i,k}(\theta) X_i X_i^T, \ \omega_k(\theta) = \mathrm{E}\big\{\frac{1}{n} \sum_{i=1}^n \eta_{i,k}(\theta)\big\},$$

$$\rho_k(\theta) = \mathrm{E}\big\{\frac{1}{n} \sum_{i=1}^n \eta_{i,k}(\theta) X_i Y_i\big\}, \ \Sigma_k(\theta) = \mathrm{E}\big\{\frac{1}{n} \sum_{i=1}^n \eta_{i,k}(\theta) X_i X_i^T\big\},$$

where the expectation is with respect to X_i and Y_i, $i = 1, \ldots, n$. Then we let $M(\theta) = \{\omega_k(\theta), \rho_k(\theta), \Sigma_k(\theta), k = 1, 2\}$, $M_n(\theta) = \{\widehat{\omega}_k(\theta), \widehat{\rho}_k(\theta), \widehat{\Sigma}_k(\theta), k = 1, 2\}$, and define $d_{2,s}\big(M(\theta_1), M(\theta_2)\big)$ and $d_2\big(M(\theta_1), M(\theta_2)\big)$ as

$$\max_{k=1,2}\{|\omega_k(\theta_1) - \omega_k(\theta_2)| \vee \|\rho_k(\theta_1) - \rho_k(\theta_2)\|_{2,s} \vee \|(\Sigma_k(\theta_1) - \Sigma_k(\theta_2))\beta_k^*\|_{2,s}\},$$
$$\max_{k=1,2}\{|\omega_k(\theta_1) - \omega_k(\theta_2)| \vee \|\rho_k(\theta_1) - \rho_k(\theta_2)\|_{2} \vee \|(\Sigma_k(\theta_1) - \Sigma_k(\theta_2))\beta_k^*\|_{2}\},$$

respectively, which are distances between $M(\theta_1)$ and $M(\theta_2)$.

Let $\Delta = \sqrt{(\beta_2^* - \beta_1^*)^T \Sigma (\beta_2^* - \beta_1^*)}$, which is a measure of the signal-to-noise ratio of the mixture linear regression model. We define the contraction basin $\mathcal{B}_{con}(\theta^*)$ as follows.

$$\mathcal{B}_{con}(\theta^*) = \{\theta : \omega_k \in (c_0, 1 - c_0), \|\beta_k - \beta_k^*\|_2 \le C_b\Delta,\ \beta_k - \beta_k^* \in \Gamma(s),\ \text{for } k = 1, 2\}.$$

Intuitively, the contraction basin requires that β_k is not far away from the true parameter β_k^*. Under the technical conditions shown later, an initialization $\widehat{\theta}^{(0)}$ falls in the contraction basin can guarantee the subsequent estimators $\widehat{\theta}^{(t)}$ in Algorithm 1 are all contained in the contraction basin.

3.2 Main results

We first introduce some technical conditions before stating the theoretical results.

(C1) The eigenvalues of Σ satisfy that $M_1 \le \lambda_{min}(\Sigma) \le \lambda_{max}(\Sigma) \le M_2$.

(C2) $n \gg s\log(p)$.

(C3) The signal-to-noise ratio $\Delta > C_1(c_0)$ for a constant $C_1(c_0)$ only depends on c_0, and $C_b < C_2(c_0, M_2)$ for a constant $C_2(c_0, M_2)$ only depends on c_0, M_2.

(C4) The initialization $\widehat{\theta}_0^{(0)} = (\widehat{\omega}_1^{(0)}, \widehat{\beta}_1^{(0)}, \widehat{\beta}_2^{(0)}) \in \mathcal{B}_{con}(\theta^*)$.

Condition (C1) is a standard assumption on the covariance Σ in high-dimensional statistics (Bickel and Levina, 2008; Cai et al., 2011). Condition (C2) is a common assumption in high dimensions on the relationship among (n, p, s) to guarantee consistent estimation (Meinshausen and Yu, 2009, e.g). In particular, it implies that the restrictive eigenvalue condition $\inf_{\mu\in\Gamma(s)\cap\mathcal{S}^{p-1}}\{\mu^T(\sum_{i=1} X_iX_i^T/n)\mu\} > \tau_0$ holds for a positive generic constant τ_0 with high probability, and is used for proving the concentration of $\widehat{\beta}_k^{(t)}$ in the t-th iteration. Condition (C3) has two requirements. The first one is that the signal-to-noise ratio is larger than a universal constant that does not depend on n and p so that the two mixtures are distinguishable. This requirement was also previously used in mixture linear model (e.g., Yi and Caramanis, 2015; Balakrishnan et al., 2017; Zhang et al., 2020). The second one is that, for the parameter β_k in the contraction basin, the distance $\|\beta_k - \beta_k^*\|_2$ is bounded by the signal-to-noise ratio multiplied by a generic constant independent of n and p. This requirement makes all the β_k in the contraction basin not too far away from the truth β_k^*. Condition (C4) ensures that the initialization is in the contraction basin. The contraction and concentration properties shown later guarantee that the estimates in each step of Algorithm 1 stay in the contraction basin.

Next, we present two lemmas about the linear convergence of the population EM updates and the concentration of the sample estimation to the population one in each EM iteration. The following two lemmas together with a key technical Lemma A.9 in the appendix are highly non-trivial and serve as the building blocks of the main theory for Algorithm 1.

Lemma 2 *Under conditions (C1) and (C3), if $\theta \in \mathcal{B}_{con}(\theta^*)$, then*

$$d_2\big(M(\theta), M(\theta^*)\big) \le \kappa_0\big(|\omega_1(\theta) - \omega_1^*| \vee \|\beta_1 - \beta_1^*\|_2 \vee \|\beta_2 - \beta_2^*\|_2\big).$$

for some $0 < \kappa_0 < \frac{1}{2\vee(64/\tau_0)}$.

Lemma 3 *Suppose that $\theta^* \in \Theta^*$. Under condition (C1), there exists a constant $C_{con} > 0$, such that with probability at least $1 - 4p^{-1}$,*

$$\sup_{\theta \in \mathcal{B}_{con}(\theta^*)} d_{2,s}(M(\theta), M_n(\theta)) \le C_{con}\sqrt{\frac{s\log(n)^2\log(p)}{n}}$$

Intuitively, Lemma 2 shows the computational contraction of Algorithm 1. It implies that, in each EM iterations, the expectations of the updated estimators converge to the true parameters at a linear rate. On the other hand, Lemma 3 establishes the statistical convergence rate of estimators to their expectations in each EM iteration. When the iteration steps are large enough, the computational error will be dominated by the statistical error, which means that further iterations can not improve the statistical convergence rate of the algorithm. Cai et al. (2019) also proved similar lemmas under the Gaussian mixture model, but our proof is more challenging, as we are interested in the mixture linear regression model. The unboundedness in both X and Y makes $M(\theta)$ more complicated and $M_n(\theta)$ have heavier tails.

Thanks to Lemmas 2 and 3, we can show the following result for Algorithm 1.

Theorem 4 *Under conditions (C1)–(C4), there exists a constant $0 < \kappa < 1/2$, such that $\widehat{\beta}_k^{(t+1)}$ obtained by Algorithm 1 satisfies, with probability $1 - 4p^{-1}$,*

$$\|\widehat{\beta}_k^{(t+1)} - \beta_k^*\|_2 = O\Big(\kappa^t(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2^*\|_2) + \sqrt{\frac{s\log(n)^2\log(p)}{n}}\Big).$$

Consequently, for $t \ge \{-\log(\kappa)\}^{-1}\log\{n(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2^*\|_2)\}$,

$$\|\widehat{\beta}_k^{(t+1)} - \beta_k^*\|_2 = O\Big(\sqrt{\frac{s\log(n)^2\log(p)}{n}}\Big).$$

We make several remarks on Theorem 4. Firstly, compared with existing results (Yi et al., 2014; Yi and Caramanis, 2015; Balakrishnan et al., 2017) requiring $\beta_2 = -\beta_1$ and $X \sim N(0, I_p)$, our model settings are more general. On one hand, note that $\beta_2 = -\beta_1$ is not just a location shift of the response variable. As an illustration, when $\beta_1 = (1, 1, 0)^T$ and $\beta_2 = (0.5, 2, 0)^T$, a simple location shift cannot reduce the model to the case where $\beta_2 = -\beta_1$. By removing the assumption that $\beta_2 = -\beta_1$, our theory is applicable to a larger model space. On the other hand, the predictors are not likely to be uncorrelated in practice. Thus, it is meaningful to extend the condition $X \sim N(0, I_p)$ to $X \sim N(0, \Sigma)$.

Although we adopt the group lasso penalty in the algorithm, the theoretical results can be naturally extended to the penalties that are decomposable (Negahban et al., 2012), such as the popular lasso penalty in the literature. From a more technical perspective, the lasso penalty is easier to handle in theoretical analysis than the group lasso penalty. The results in Theorem 4 also hold for the lasso penalty with minor modifications to the proof of Lemma A.9 (appendix Section F), where we would modify the penalized Q-function in Lemma A.9 as

$$\ell(\beta_1,\beta_2)=\sum_{k=1}^{2}\beta_k^T\widehat{\Sigma}_k^{(t+1)}\beta_k-2\sum_{k=1}^{2}(\widehat{\rho}_k^{(t+1)})^T\beta_k+\lambda_n^{(t+1)}\sum_{k=1}^{2}\|\beta_k\|_1.$$

The optimization is thus separable for each β_k. For each k, using the same technique as Lemma A.9, we can show that $\widehat{\beta}_k^{(t+1)}-\beta_k^*\in\Gamma(s)$ (under a new $\Gamma(s)=\{\mu\in\mathbb{R}^p:\|\mu_{S^c}\|_1\le 5\sqrt{s}\|\mu_S\|_2+2\sqrt{s}\|\mu\|_2\}$ that takes a simpler form compared with that in the group lasso), and $\|\widehat{\beta}_k^{(t+1)}-\beta_k^*\|_2\le\frac{4}{\tau_0}d_{2,s}\big(M_n(\widehat{\theta}^{(t)}),M(\theta^*)\big)+\frac{2}{\tau_0}\sqrt{s}\lambda_n^{(t+1)}$. Those two results are exactly the same as in Lemma A.9, which are applied to the proof in Section F.2 to obtain the concentration results for $\widehat{\beta}_k^{(t+1)}$. Therefore, our proof technique is more general and implies the same convergence rate for lasso penalty as stated in Theorem 4.

Moreover, unlike the extensive literature about the sample-splitting EM algorithms for mixture linear regression (Yi et al., 2014; Yi and Caramanis, 2015; Zhang et al., 2020), the convergence result in Theorem 4 does not require sample splitting. Sample splitting is not desirable, especially when we have a small sample size. Splitting a limited number of observations into T batches decreases the estimation efficiency, makes the estimation less stable and is rarely used in practice. Hence, it is meaningful to develop theoretical results without data splitting for mixture linear regression, as those in Theorem 4. To our best knowledge, Theorem 4 is the first theoretical result for the high-dimensional EM algorithm of mixture linear regression without data splitting. Also note that the convergence rate we obtain is nearly optimal. When the latent random variables W_i's are known, the optimal rate is $\sqrt{s\log(p)/n}$ (Ye and Zhang, 2010, e.g.), while when W_i's are unknown, Zhang et al. (2020) gives an estimation rate of $\sqrt{s\log(p)\log n/n}$ with sample splitting. Our result is slightly slower than these rates by the factors of $\log n$ and $\log^{1/2}n$, respectively. The additional $\log(n)$ terms are the price of no data splitting. Technically, to prove the convergence results without data splitting, we have to bound the tails of the supremum of unbounded random processes, which is much more challenging than bounding the random variables.

In Section 2, Lemma 1 only states that the algorithm can converge but does not answer if and when it can converge to the global optima or the ground truth. Theorem 4 answers it both computationally and statistically and provides more information. Starting with an initialization in the contraction basin, Theorem 4 says that the proposed algorithm can converge to the true parameters with a convergence rate containing both computational error and statistical error. It is a direct analysis to the output obtained by the algorithm. In the convergence, $\kappa^t(|\widehat{\omega}_1^{(0)}-\omega_1^*|\vee\|\widehat{\beta}_1^{(0)}-\beta_1^*\|_2\vee\|\widehat{\beta}_2^{(0)}-\beta_2^*\|_2)$ can be viewed as the computational error. It is an exponential function for t. Since $\kappa<1$, when $t\to\infty$, this term disappear. It can also be viewed as the geometric convergence, except that the convergence is to the ground truth instead of the global solution. The second term $\sqrt{s\log(n)^2\log(p)/n}$ is the statistical error, which cannot disappear no matter how many EM updates we run.

Combining those two errors, if $t \geq \{-\log(\kappa)\}^{-1}\log\{n(|\widehat{\omega}_1^{(0)}-\omega_1^*|\vee\|\widehat{\beta}_1^{(0)}-\beta_1^*\|_2\vee\|\widehat{\beta}_2^{(0)}-\beta_2^*\|_2)\}$, the computational error will be dominated by the statistical error. So, after this step, further EM updates will not improve the convergence rate of the algorithm's output. Note that $\{-\log(\kappa)\}^{-1}\log\{n(|\widehat{\omega}_1^{(0)}-\omega_1^*|\vee\|\widehat{\beta}_1^{(0)}-\beta_1^*\|_2\vee\|\widehat{\beta}_2^{(0)}-\beta_2^*\|_2)\}$ only involves a $\log n$ term. After almost finite EM updates, the proposed algorithm can provide a good enough estimation in practice.

Now we turn to the effect of the misspecification of σ^2. In most existing theoretical analysis including this one, σ^2 is usually treated as a known parameter in theoretical studies for the mixture linear regression model. There are two reasons for this treatment. On one hand, the regression coefficients are of primary interest in regression models. On the other, statistical analysis for mixture regression model with known σ^2 is already challenging. However, in practice σ^2 is almost never known. Statisticians often plug in an estimated value of σ^2, which differs from σ_*^2. Yet it is unclear how the misspecification affects the estimation of the mixture linear regression. In the following theorem, we obtain a non-asymptotic convergence result for Algorithm 1 with misspecified σ^2, which indicates that although a misspecified σ^2 brings bias to the estimation, it usually has a minor influence on the mixture linear regression model with a large signal-to-noise ratio.

Theorem 5 *Let $\widetilde{\beta}_k^{(t+1)}$ be the estimation of Algorithm 1 in the $(t+1)$-th EM step where σ^2 may not equal to σ_*^2. Let $\xi = \sigma_*(2\sigma_*^2/\sigma^2-1)^{-1}\cdot(\Delta/\sigma_*)^{-2}\cdot\left|1-\sigma_*^2/\sigma^2\right|$. Under Conditions (C1)-(C4) and $\sigma^2 < 2\sigma_*^2$, there exists a constant $0<\kappa<1/2$ such that, with probability $1-4p^{-1}$,*

$$\|\widetilde{\beta}_k^{(t+1)}-\beta_k^*\|_2 = O\Big(\kappa^t(|\widehat{\omega}_1^{(0)}-\omega_1^*|\vee\|\widehat{\beta}_1^{(0)}-\beta_1^*\|_2\vee\|\widehat{\beta}_2^{(0)}-\beta_2^*\|_2)+\sigma_*\sqrt{\frac{s\log(n)^2\log(p)}{n}}+\xi\Big).$$

Consequently, for $t \geq \{-\log(\kappa)\}^{-1}\log\{n(|\widehat{\omega}_1^{(0)}-\omega_1^*|\vee\|\widehat{\beta}_1^{(0)}-\beta_1^*\|_2\vee\|\widehat{\beta}_2^{(0)}-\beta_2^*\|_2)\}$,

$$\|\widetilde{\beta}_k^{(t+1)}-\beta_k^*\|_2 \leq c_1\xi + c_2\sigma_*\sqrt{\frac{s\log(n)^2\log p}{n}},$$

where c_1 and c_2 are generic constants that are not related to σ^2 and σ_*^2.

We make several remarks for Theorem 5. Firstly, The only difference between the convergence rates in Theorems 4 and 5 is the bias term $c_1\xi$ and is exactly 0 when $\sigma^2=\sigma_*^2$. Although not using σ_*^2 in Algorithm 1 may return a biased estimation, the bias is small when σ^2 is close to σ_*^2.

Secondly, the bias term is also a function of $(\Delta/\sigma_*)^{-2}$. When the signal-to-noise ratio Δ is large, the bias term is small regardless of the value of σ^2. Specifically, if $\Delta \gg O(n^{1/4}/[s\log(n)^2\log p]^{1/4})$, then the bias term is ignorable compared with $\sqrt{s\log(n)^2\log p/n}$. Intuitively, when $\Delta\to\infty$, the mixtures can be easily identified, which makes the mixture linear regression model reduce to linear regression models. It is well known that the choice of σ^2 in the linear regression model does not affect the maximum likelihood estimation for the coefficients. Thus, when $\Delta\to\infty$, the bias term disappears. However, the current parameter space Θ^* we consider restrict Δ to be bounded above by a constant. Removing this assumption can cause some statistics in the proof no longer to be sub-Gaussian or

sub-exponential, which makes the theoretical studies more challenging. The study for the case, where Δ can also diverge, is more challenging and beyond the scope of this paper.

Thirdly, for high-dimensional mixture linear regression model, the signal-to-noise ratio needs to be sufficiently large to make the mixtures identifiable. Hence, the bias term is usually small in practice. The numerical studies in Section 4.3 demonstrate that σ^2 has a minor influence on the performance of Algorithm 1 for relatively large Δ. For example, the empirical averaged biases based on 100 replicates are smaller than 0.05 when σ^2 is between 0.75 and 1.5 for the first two settings. Please see Section 4.3 for the detailed model settings and discussions.

Finally, Theorem 5 also connects the proposed algorithm with Lloyd's Algorithm, which is another popular algorithm for the mixture models. When $\sigma^2 \to 0$, the proposed algorithm can be viewed as a variant of Lloyd's Algorithm. For this scenario, the bias term is smaller than $2c_1(\Delta/\sigma_*)^{-2}$, which means that, even when σ^2 is seriously misspecified as 0, the bias can still be small when Δ is large. We also remark that the condition $\sigma^2 < 2\sigma_*^2$ is not necessary for the algorithm in practice, but only a requirement in theory due to technical reasons.

4. Simulation studies

4.1 Simulation set-up

In this section, we investigate the empirical performance of Algorithm 1. For practical initialization, we start with lasso regression (Tibshirani, 1996) of Y on X to roughly select important predictors and then apply the tensor power method (Anandkumar et al., 2014) on the selected variables to get initializations for β_k and mixing proportion ω_k, $k = 1, \cdots, K$. We use $T = 20$ as the maximum number of iterations in Algorithm 1, and stop the iterations when $\|(\widehat{\beta}_1^{(t+1)}, \ldots, \widehat{\beta}_K^{(t+1)}) - (\widehat{\beta}_1^{(t)}, \ldots, \widehat{\beta}_K^{(t)})\|_F < 10^{-3}$. We use a single tuning parameter $\lambda = \lambda_n^{(t)}$ for all $t = 0, 1, \ldots, T$ and choose λ based on the Bayesian information criterion.

We include the following methods: 1) *Oracle*, we fit the standard EM algorithm on the true subset of s relevant predictors. 2) *Initial*, after using the tensor power method to get the initialization for β_k, $k = 1, \cdots, K$, we plug them into (5) to get an estimation for the weight $\widehat{\eta}_{ik}$ to the i-th sample, $i = 1, \cdots, n$. Then, we label the i-th sample with $\text{argmax}_{k=1,\cdots,K} \widehat{\eta}_{ik}$. After that, we fit lasso regressions separately for each estimated mixtures to estimate β_k. Compared with the tensor power method, this method returns sparse estimations. 3) *GLLiM*, the Gaussian Local Linear Mapping EM algorithm (Deleforge et al., 2015) that is implemented in the R package `xLLiM`. 4) *HDEM*, our implementation of the high-dimensional EM algorithm proposed by Zhang et al. (2020). 5) *PSEM*, i.e., the post-selection EM algorithm, we fit the standard EM algorithm on the selected variables from Algorithm 1. and finally 6) *PEM*, the penalized EM algorithm (Algorithm 1).

We consider the following simulation models, where we first generate independent predictor $X_i \sim N(0, \Sigma)$ and error $\epsilon_i \sim N(0, 1)$ and then Y_i follows from the mixture linear regression (1). For all the four simulation examples, we fix the first $s = 10$ coefficients in β_k to be nonzero and consider both $p = 400$ and $p = 1000$ settings. The total sample size n is set to be 400 for models (M1)–(M3), where we have two mixtures, and 600 for model (M4), where we have three mixtures. The symmetric mixture assumption does not hold in any

of our models, $\beta_1 \neq -\beta_2$. In model (M5), we investigate the performance of the proposed algorithm when the group-wise sparse structure is violated.

- (M1) Two mixtures with $\omega_1 = \omega_2 = 0.5$, and auto-regressive covariance structure $[\Sigma]_{ij} = 0.3^{|i-j|}$ for $i, j = 1, \ldots, p$. The nonzero coefficients of β_1 is generated independently from $N(0, 1)$; and the nonzero coefficients of β_2 is $\beta_{2j} = \beta_{1j} + 2 \cdot \text{sgn}(\beta_{1j})$, $j = 1, \ldots, s$.

- (M2) Same as (M1) but with weaker signals:$\beta_{2j} = \beta_{1j} + 1 \cdot \text{sgn}(\beta_{1j})$, $j = 1, \ldots, s$.

- (M3) Same as (M1) but with a different covariance Σ based on Erdós-Rényi random graph. Let $\widetilde{\Sigma} = (\widetilde{\sigma}_{ij})$, where $\widetilde{\sigma}_{ij} = u_{ij}\delta_{ij}$, δ_{ij} follows Bernoulli(0.1) distribution, and $u_{ij} \sim \text{Uniform}[0.5, 1] \cup \text{Uniform}[-1, -0.5]$. Then let $\widetilde{\Sigma}_1 = (\widetilde{\Sigma} + \widetilde{\Sigma}^T)/2$ and $\Sigma^* = \widetilde{\Sigma}_1 + \{\max(-\lambda_{min}(\widetilde{\Sigma}_1), 0) + 0.05\}I_p$. Finally, Σ^* is standardized to have 1's on the diagonal.

- (M4) Three mixtures with $\omega_1 = \omega_2 = \omega_3 = 1/3$. The nonzero elements of β_1 and β_3 are -1 and 5, respectively, and the nonzero elements of β_2 are evenly spaced between 1 and 3.

- (M5) The same as (M1) but with the two mixtures consist of distinct sets of important variables. With $n = p = 400$, we set the first ten elements of β_1 as 1 and set the ten consecutive elements of β_2 as 2, starting with the j-th elements, $j \in \{1, 3, 5, 7, 9, 11\}$. As such, the number of shared non-zero elements between β_1 and β_2 is varying from 0 (no overlap of important variables) to 10 (joint sparsity structure).

4.2 Simulation results

The performances of those methods are evaluated using the following criteria. The parameter estimation errors are defined as $\sqrt{\sum_{k=1}^{K} \|\widehat{\beta}_k - \beta_k\|_2^2}$ for β and $\sum_{k=1}^{K} |\widehat{\omega}_k - \omega_k| \times 100$ for ω. The mixture estimation error is defined as $\sum_{i=1}^{n} I(\widehat{W}_i \neq W_i)/n \times 100$, where $\widehat{W}_i = \text{argmax}_{k=1,\ldots,K} \widehat{\eta}_{i,k}(\widehat{\theta})$ for $i = 1, \ldots, n$. We also consider the mean squared perdition error for the methods. We generate an independent testing data sets from the simulation model with the same sample size, use the estimated regression coefficients and mixture labels to predict the response, and then calculate the mean squared perdition error. For methods perform variable selection, we also recorded the true positive and false positive rates of variable selection.

The estimation results are summarized in Table 1, particularly for β's estimation we further plot the results from 100 replicates in Figure 1. As expected, the *oracle* method, i.e., the standard EM algorithm applied to the truly relevant $s = 10$ predictors, has the best performance. On the other hand, the method that does not perform variable selection, GLLiM, failed for all the simulation settings. The proposed method, *PEM*, has encouraging performances for all the simulation models; overall, it is comparable to the *oracle* method and has a slight edge over the very recent method *HDEM* (Zhang et al., 2020). The advantage of our method over *HDEM*, which updates each β_k separately via lasso penalized

M1		$p = 400$			$p = 1000$	
	β_k	ω_k	W_i	β_k	ω_k	W_i
Oracle	0.44 (0.04)	6.08 (0.55)	8.43 (0.14)	0.39 (0.01)	6.18 (0.45)	8.40 (0.13)
PSEM	0.57 (0.06)	6.31 (0.55)	9.17 (0.42)	0.76 (0.12)	6.28 (0.47)	10.50 (0.77)
GLLiM	9.63 (0.01)	68.18 (2.29)	33.23 (1.21)	9.68 (0.01)	74.20 (2.28)	34.24 (1.33)
Initial	5.14 (0.05)	32.14 (2.02)	41.75 (0.55)	5.34 (0.04)	35.84 (2.21)	42.46 (0.54)
HDEM	1.22 (0.09)	8.60 (0.61)	10.71 (0.61)	1.93 (0.18)	10.17 (1.18)	15.34 (1.32)
PEM	1.04 (0.05)	6.67 (0.58)	9.79 (0.43)	1.26 (0.09)	7.08 (0.52)	10.91 (0.76)
M2		$p = 400$			$p = 1000$	
Oracle	0.44 (0.03)	12.29 (0.84)	16.21 (0.21)	0.45 (0.03)	11.56 (0.69)	15.97 (0.17)
PSEM	0.64 (0.03)	13.67 (0.90)	17.78 (0.40)	0.70 (0.02)	14.06 (0.93)	18.15 (0.21)
GLLiM	6.74 (0.01)	67.01 (1.98)	33.95 (1.10)	6.77 (0.01)	68.60 (2.12)	33.04 (1.04)
Initial	2.79 (0.04)	46.08 (2.44)	44.03 (0.42)	2.99 (0.04)	55.76 (2.43)	44.75 (0.41)
HDEM	1.26 (0.06)	30.87 (1.21)	22.58 (0.63)	1.96 (0.10)	41.59 (2.15)	29.65 (1.11)
PEM	1.03 (0.04)	23.63 (1.32)	19.95 (0.44)	1.39 (0.08)	33.36 (2.31)	23.28(0.86)
M3		$p = 400$			$p = 1000$	
Oracle	0.56 (0.09)	8.16 (0.62)	10.09 (0.14)	0.48 (0.07)	7.60 (0.55)	9.70 (0.16)
PSEM	0.67 (0.09)	7.97 (0.59)	11.78 (0.64)	0.65 (0.07)	7.81 (0.56)	11.10 (0.52)
GLLiM	9.65 (0.01)	68.41 (1.95)	33.72 (1.11)	9.54 (0.01)	67.52 (2.15)	33.97(1.02)
Initial	5.02 (0.05)	36.47 (1.96)	43.41 (0.52)	5.30 (0.04)	43.71 (2.18)	43.79 (0.47)
HDEM	1.54 (0.12)	11.82 (0.73)	14.12 (0.94)	2.07 (0.18)	13.39 (1.02)	17.68 (1.31)
PEM	1.21 (0.08)	10.67 (0.72)	12.27 (0.63)	1.39 (0.07)	10.47 (0.73)	12.10 (0.54)
M4		$p = 400$			$p = 1000$	
Oracle	1.03 (0.28)	5.14 (0.32)	7.57 (0.10)	1.35 (0.38)	5.28 (0.53)	7.53 (0.11)
PSEM	1.80 (0.52)	5.21 (0.34)	10.82 (1.21)	3.56 (0.67)	6.58 (0.59)	16.53 (1.94)
GLLiM	17.38 (0.01)	52.15 (2.72)	33.79 (0.60)	17.40 (0.01)	54.10 (2.25)	34.99 (0.63)
Initial	13.16 (0.17)	18.56 (0.98)	49.62 (0.64)	13.54 (0.14)	18.48 (0.91)	51.21 (0.60)
HDEM	8.87 (0.86)	8.93 (0.65)	30.02 (2.51)	12.08 (0.73)	19.28 (1.28)	43.62 (2.42)
PEM	2.43 (0.42)	5.08 (0.30)	11.49 (1.23)	3.99 (0.57)	6.32 (0.46)	17.10 (1.94)

Table 1: Average estimation errors based on 100 replicates (standard errors in parentheses).

	p	Oracle	PSEM	GLLiM	Initial	HDEM	PEM
M1	400	1.02 (0.08)	1.21 (0.12)	65.54 (0.70)	12.63 (0.19)	1.73 (0.16)	1.39 (0.03)
	1000	0.94 (0.01)	1.68 (0.36)	64.71 (0.53)	13.44 (0.23)	3.06 (0.36)	1.80 (0.15)
M2	400	0.86 (0.01)	1.04 (0.04)	32.68 (0.32)	3.78 (0.06)	1.66 (0.08)	1.30 (0.05)
	1000	0.88 (0.04)	1.23 (0.07)	32.25 (0.25)	4.10 (0.06)	2.51 (0.12)	1.64 (0.08)
M3	400	1.27 (0.14)	1.41 (0.16)	46.50 (0.51)	9.22 (0.11)	2.27 (0.22)	1.59 (0.08)
	1000	1.04 (0.07)	1.47 (0.19)	44.47 (0.47)	10.42 (0.15)	3.75 (0.34)	2.05 (0.14)
M4	400	1.96 (0.41)	2.51 (0.60)	175.16 (1.74)	56.63 (1.23)	8.43 (1.44)	2.78 (0.52)
	1000	1.83 (0.34)	12.61 (2.93)	173.77 (1.76)	60.81 (1.28)	39.13 (2.89)	9.09 (1.86)

Table 2: Mean squared perdition error based on 100 replicates (standard errors in parentheses).

estimation in the M-step, can be explained by the group-wise penalization and joint estimation of all β_k's in the M-step. In model M4 with three mixtures, our approach of joint estimation has even more gains in estimation accuracy.

We report the mean squared prediction errors of the methods in Table 2. The results are similar as the estimation error, namely *oracle* method has the best performance, followed by *PSEM* and *PEM*. We note that *PSEM* performs slightly better than *PEM* under models

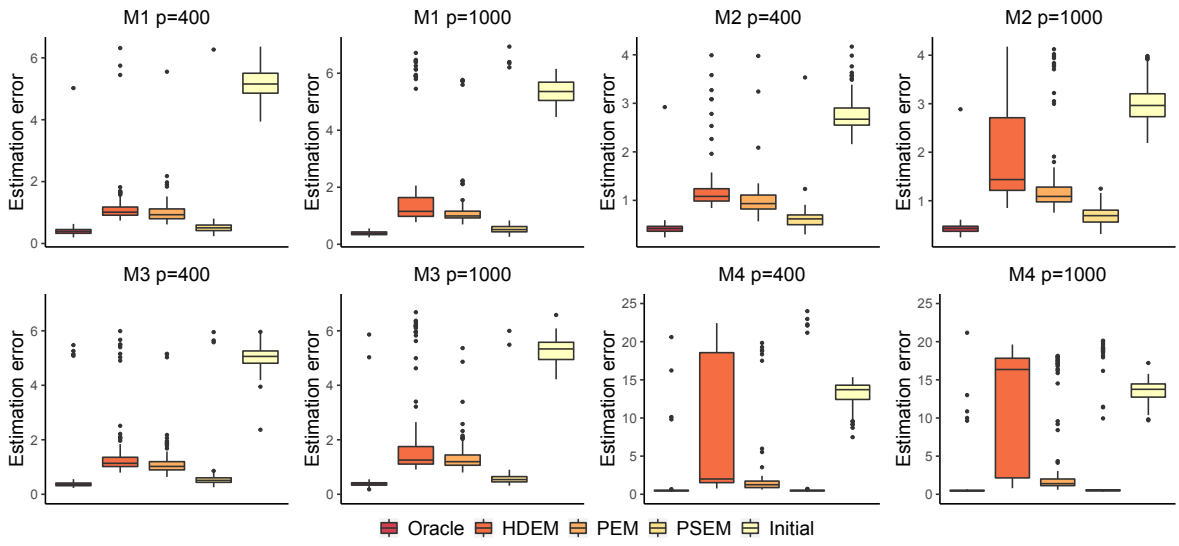

Figure 1: The estimation error of β_k's based on 100 replicates.

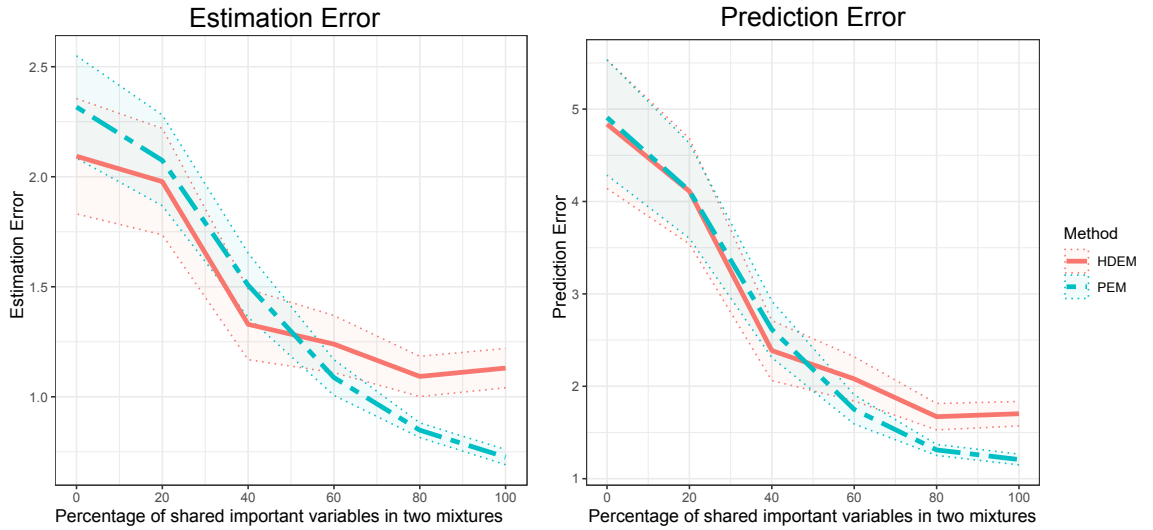

Figure 2: Model M5. Mean estimation error (left plot) and prediction error (right plot) and their standard errors (visualized by the length of interval on the plots).

M1–M3, where the variable selection results are excellent, but performs worse than *PEM* under model M4 with $p = 1000$, where the variable selection results are not as good as models M1–M3.

The sparsity pattern targeted by the group lasso encompasses scenarios where the sparsity patterns within two mixtures are different. In model (M5), we highlight the benefits of employing the group lasso when the two mixtures consist of distinct sets of important variables. Figure 2 shows the mean estimation error for the regression coefficients and prediction error when the number of shared non-sparse elements between β_1 and β_2 varies

	M1 $p = 400$		M1 $p = 1000$		M2 $p = 400$		M2 $p = 1000$	
	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Initial	83.2 (0.21)	8.3 (0.05)	76.4 (0.23)	3.2 (0.02)	88.3 (0.19)	8.9 (0.05)	79.6 (0.21)	4.3 (0.03)
HDEM	92.7 (0.09)	1.4 (0.03)	86.6 (0.15)	1.5 (0.03)	89.8 (0.01)	1.9 (0.02)	81.4 (0.25)	1.2 (0.01)
PEM	100 (0)	0.9 (0.02)	100 (0)	0.7 (0.02)	100 (0)	1.2 (0.01)	100 (0)	0.8 (0.01)
	M3 $p = 400$		M3 $p = 1000$		M4 $p = 400$		M4 $p = 1000$	
Initial	89.1 (0.16)	0.5 (0.05)	73.4 (0.25)	3.9 (0.03)	47.5 (0.31)	4.0 (0.05)	38.3 (0.28)	1.7 (0.02)
HDEM	95.0 (0.06)	2.6 (0.05)	80.2 (0.12)	1.7 (0.03)	83.4 (0.24)	12.6 (0.15)	64.7 (0.33)	6.7 (0.06)
PEM	100 (0)	1.3 (0.02)	100 (0)	0.5 (0.01)	100 (0)	2.7 (0.11)	100 (0)	3.8 (0.09)

Table 3: Average true positive rate (TPR) and false positive rate (FPR) for variable selection based on 100 replicates (standard errors in parentheses).

from 0 (no overlap of important variables in two mixtures) to 10 (exactly the same set of important variables in two mixtures). When this number is small, the group lass approach in PEM is not much worse than the lasso approach in HDEM. However, when the number increases, say more than 6, PEM starts to outperform HDEM even when the two mixtures have different sparsity patterns.

We also note that Algorithm 1 substantially improves over the *Initial* values. Even when the initialization is quite far from the truth, the proposed algorithm improves over iterations in terms of all the evaluation criteria. This is partly explained by the guaranteed monotonicity of our iterative algorithm in the penalized log-likelihood (see Lemma 1).

4.3 Simulations for misspecified σ^2 in Algorithm 1

In this section, we show some simulation results for using different σ^2 in Algorithm 1. The data is generated from (1). We set $K = 2$, $\omega_1 = \omega_2 = 1/2$, $p = 400$, $n = 400$, and the true value of σ^2 to be 1. The nonzero coefficients of β_1 is generated independently from $N(0, 1)$; and the nonzero coefficients of β_2 is $\beta_{2j} = \beta_{1j} + \delta \cdot \text{sgn}(\beta_{1j})$, $j = 1, \ldots, p$. The signal strength δ takes value from $\{0.75, 1, 2\}$ and the input variance σ^2 in Algorithm 1 takes value from $\{0.5, 0.75, 1, 1.5, 2\}$.

	$\sigma^2 = 0.5$	$\sigma^2 = 0.75$	$\sigma^2 = 1$	$\sigma^2 = 1.5$	$\sigma^2 = 2$
$\delta = 2$	1.19 (0.08)	1.05 (0.03)	1.05 (0.03)	1.06 (0.03)	1.13 (0.07)
$\delta = 1$	1.05 (0.05)	1.01 (0.04)	1.03 (0.04)	1.06 (0.03)	1.23 (0.05)
$\delta = 0.75$	1.50 (0.09)	1.51 (0.10)	1.60 (0.09)	2.43 (0.12)	3.28 (0.07)

Table 4: Coefficient estimation errors for Algorithm 1 with different input σ^2 based on 100 replicates. (standard errors in parentheses).

From Table 4 and Figure 3, when δ equals 1 and 2, using all the five σ^2 in Algorithm 1 returns good estimations. Although when σ^2 is 0.5 or 2, the performances are slightly worse, but can still capture the mixture information and give accurate enough estimates. The results are consistent with Theorem 5. For mixture linear regression models with a large signal-to-noise ratio, the choice of σ^2 has a minor influence on the performance of Algorithm 1. When δ is 0.75, the signal-to-noise ratio is not large enough, the performance of Algorithm 1 becomes worse, no matter the value of σ^2. This is mainly caused by the poor quality of the initialization. It deserves to notice that when σ^2 equals 0.5 or 0.75, the

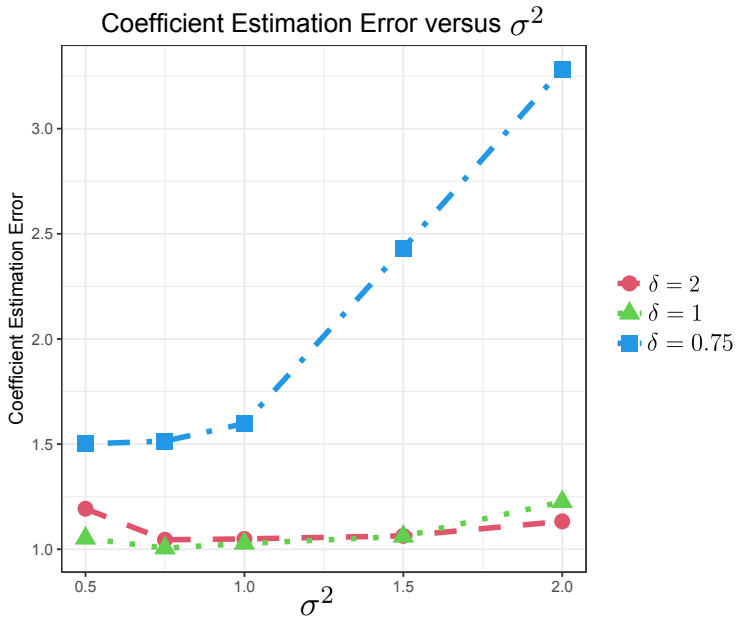

Figure 3: Coefficient estimation errors for Algorithm 1 with different input σ^2 based on 100 replicates.

performance of the algorithm is slightly better than taking $\sigma^2 = 1$. This result indicates that when the signal-to-noise ratio is relatively small, Algorithm 1 can be less sensitive to the initialization if σ^2 takes a smaller value than the true one.

5. Extension to multivariate mixture linear regression

5.1 Model and algorithm

In this section, we consider a generalization of the mixture linear regression model, where Y is a q-dimensional response. Specifically, we consider the model

$$\mathbb{P}(W = k) = \omega_k, \quad Y \mid (X, W = k) \sim N(\beta_k^T X, \Sigma_y), \tag{9}$$

where $\beta \in \mathbb{R}^{p\times q}$ and $\Sigma_y \in \mathbb{R}^{q\times q}$ is a symmetric and positive definite matrix. We will answer two questions: *Does the penalized EM algorithm still work and have similar convergence results for multiple response cases? If so, what is the advantage of considering q responses simultaneously than q mixture linear regression problems separately?*

As the mixture linear regression model, to handle the high-dimensionality of X, we assume the matrix coefficients β_k, $k = 1, \cdots, K$, to be sparse, and the sparsity patterns are the same for all the mixtures. We redefine $S = \{(i, j) : (\beta_k)_{ij} \neq 0,\ i = 1 \cdots, p,\ j = 1, \cdots, q\}$ and $s = |S|_0$ be the cardinality of S. We allow $p \gg n$ and q to grow linearly with s.

In model (9), our parameter of interest is $\theta = \{\omega_1, \cdots, \omega_k, \beta_1, \cdots, \beta_K\}$. The error covariance Σ_y is treated as a known parameter. The penalized EM algorithm for solving (9) is analogous to Algorithm 1 and is summarized in Algorithm 2. Similar to Algorithm 1,

Algorithm 2 Group lasso penalized EM algorithm for model (9)

Input: Initial values $\widehat{\omega}_k^{(0)}$, $\widehat{\beta}_k^{(0)}$, for $k = 1, \cdots, K$, maximum iteration number T, data $\{X_i, Y_i; i = 1, \ldots, n\}$, and initial tuning parameter

$$\lambda_n^{(0)} = C_1(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \cdots \vee |\widehat{\omega}_K^{(0)} - \omega_K^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_F \vee \cdots \vee \|\widehat{\beta}_K^{(0)} - \beta_K^*\|_F)/\sqrt{s} + C_\lambda \sqrt{\log(n)^2 \log(pq)/s}$$

for some positive constants C_1 and C_λ.

Iterate: For $t = 0, \cdots, T-1$, do the following steps until convergence.

- For $i = 1, \cdots, n$, let

$$\widehat{\eta}_{i,k}(\widehat{\theta}^{(t)}) = \widehat{\omega}_k^{(t)} / \Big(\widehat{\omega}_k^{(t)} + \sum_{k' \neq k} \widehat{\omega}_{k'}^{(t)} \exp\big\{ X_i^T (\widehat{\beta}_{k'}^{(t)} - \widehat{\beta}_k^{(t)}) \Sigma_y^{-1} \big(Y_i - (\widehat{\beta}_k^{(t)} + \widehat{\beta}_{k'}^{(t)})^T X_i / 2 \big) \big\} \Big).$$

- For $k = 1, \cdots, K$, update

$$\widehat{\omega}_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \widehat{\eta}_{i,k}(\widehat{\theta}^{(t)}),$$
$$\widehat{\rho}_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \big(\widehat{\eta}_{i,k}(\widehat{\theta}^{(t)}) X_i Y_i^T \Sigma_y^{-1} \big),$$
$$\widehat{\Sigma}_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \big(\widehat{\eta}_{i,k}(\widehat{\theta}^{(t)}) X_i X_i^T \big),$$

and update $\widehat{\beta}_k^{(t+1)}$ by minimizing

$$\sum_{k=1}^K \mathrm{tr}\big(\Sigma_y^{-1} \beta_k^T \widehat{\Sigma}_k^{(t+1)} \beta_k \big) - 2 \sum_{k=1}^K \mathrm{tr}\big((\widehat{\rho}_k^{(t+1)})^T \beta_k \big) + \lambda_n^{(t+1)} \sum_{j=1}^p \sum_{l=1}^q \sqrt{\sum_{k=1}^K (\beta_k)_{jl}^2}.$$

with $\lambda_n^{(t+1)} = \kappa \lambda_n^{(t)} + C_\lambda \sqrt{\log(pq)\log(n)^2/n}$, where $\kappa \in (0, 1/2)$ and $(\beta_k)_{jl}$ is the (j,l)-th element of β_k.

Output: $\widehat{\theta}^{(t+1)} = \{\widehat{\omega}_1^{(t+1)}, \ldots, \widehat{\omega}_K^{(t+1)}, \widehat{\beta}_1^{(t+1)}, \ldots, \widehat{\beta}_K^{(t+1)}\}$

the formula of $\lambda_n^{(t)}$ is mainly for theoretical consideration. In practice, we fix $\lambda_n^{(t)} = \lambda$ over all iterations.

5.2 Theoretical analysis

In this section, we consider the theoretical studies for Algorithm 2. We assume that Σ_y is a known parameter and is symmetric and positive definite. Recall that, even for the univariate-response mixture linear regression, theoretical analysis for unknown variance σ^2 is extremely challenging and has not been considered in high dimensions. As a result, for multiple-response mixture linear regression, it makes sense to assume that Σ_y is known as a necessary simplification. As such, we focus on the coefficients β_k as parameters of interest. To the best of our knowledge, there is little work for multiple-response mixture regression even in this simplified context.

We first define the following additional notations. For a matrix $A \in \mathbb{R}^{p\times q}$, $\|A\|_F$ is its Frobenius norm and $\|A\|_{F,s} = \sup_{\mu\in\mathbb{R}^{p\times q},\|\mu\|_F=1,\mathrm{vec}(\mu)\in\Gamma(s)}\langle A,\mu\rangle_F$. The true parameter space we consider now is

$$\Theta^* = \{\theta^* : \omega_1^* \in (c_w, 1-c_w), \|\mathrm{vec}(\beta_k^*)\|_0 \le s, \|\beta_k^*\|_F \le M_b, \text{for } k = 1,2\},$$

and the signal-to-noise ratio is re-defined as $\Delta = \sqrt{\mathrm{tr}((\beta_2^*-\beta_1^*)^T\Sigma(\beta_2^*-\beta_1^*)\Sigma_y)}/q$. Then we define $d_{F,s}\big(M(\theta_1), M(\theta_2)\big)$ and $d_F\big(M(\theta_1), M(\theta_2)\big)$ as

$$\max_{k=1,2}\{|\omega_k(\theta_1)-\omega_k(\theta_2)| \vee \|\rho_k(\theta_1)-\rho_k(\theta_2)\|_{F,s} \vee \|(\Sigma_k(\theta_1)-\Sigma_k(\theta_2))\beta_k^*\|_{F,s}\},$$
$$\max_{k=1,2}\{|\omega_k(\theta_1)-\omega_k(\theta_2)| \vee \|\rho_k(\theta_1)-\rho_k(\theta_2)\|_{F} \vee \|(\Sigma_k(\theta_1)-\Sigma_k(\theta_2))\beta_k^*\|_{F}\},$$

respectively, and the new constriction basin $\mathcal{B}_{con}(\theta^*)$ as

$$\mathcal{B}_{con}(\theta^*) = \{\theta : \omega_k \in (c_0, 1-c_0), \|\beta_k - \beta_k^*\|_F \le C_b\Delta,\ \mathrm{vec}(\beta_k-\beta_k^*) \in \Gamma(s),\ \text{for } k = 1,2\}.$$

We require the same technical conditions as the mixture linear regression with a modification to Condition (C3):

(C3') The new signal-to-noise ratio $\Delta = \sqrt{\mathrm{tr}((\beta_2^*-\beta_1^*)^T\Sigma(\beta_2^*-\beta_1^*)\Sigma_y)}/q > C_1(c_0)$ for a constant $C_1(c_0)$ only depends on c_0, and $C_b < C_2(c_0, M_2)$ for a constant $C_2(c_0, M_2)$ only depends on c_0, M_2.

The new signal-to-noise ratio reveals a fundamental difference of considering q responses together than q mixture linear regressions separately. To see this, consider the case $\Sigma_y = I_q$ and define $\beta_{k,j}$ as the j-th column of β_k, i.e, the regression coefficient of the j-th element of Y and X. Further define the individual signal-to-noise-ratio as $\Delta_j = \sqrt{(\beta_{1,j}-\beta_{2,j})^T\Sigma(\beta_{1,j}-\beta_{2,j})}$. If we apply Algorithm 1 to the j-th element of Y and X, for $j = 1,\ldots,q$, then we need all of Δ_j to be bounded below, as discussed in Section 3. In other words, when some Δ_j are small and the two mixtures are not well separated on this coordinate, Algorithm 1 may not be able to estimate the corresponding coefficients. However, if we consider all the responses simultaneously in multivariate mixture linear regression, we only require $\Delta = \sqrt{\sum_{j=1}^q \Delta_j^2}/q$, the average of q signal-to-noise ratios, to be greater than a constant. We can allow some responses to have small signal-to-noise ratios without rendering Algorithm 2 inapplicable. This is because all the response shares the common latent structure. The responses with large signal-to-noise ratios can help enhance the identification of the mixtures for the responses with small signal-to-noise ratios.

As a brief numerical illustration, we generated simulated datasets from (9) with $K = 2$, $n = 400$, $p = 100$, $s = 10$, $q = 2$, $\Sigma_y = I_q$ and $[\Sigma]_{ij} = 0.3^{|i-j|}$. The first 5 rows of β_k, $k = 1, 2$, are set to be nonzero. More specifically, the nonzero coefficients of the first and second columns of β_1 are generated from $N(0, 1)$; and the nonzero coefficients of β_2 are $(\beta_2)_{j1} = (\beta_1)_{j1} + 2 \cdot \mathrm{sgn}((\beta_1)_{j1})$ and $(\beta_2)_{j2} = (\beta_1)_{j2} + \delta \cdot \mathrm{sgn}((\beta_1)_{j2})$. The signal strength is 2 for the first response and δ for the second one. We vary δ from 0.5 to 2 with increments of 0.5. From Figure 4, we can see that the coefficient estimation errors for Algorithm 2 remain at a

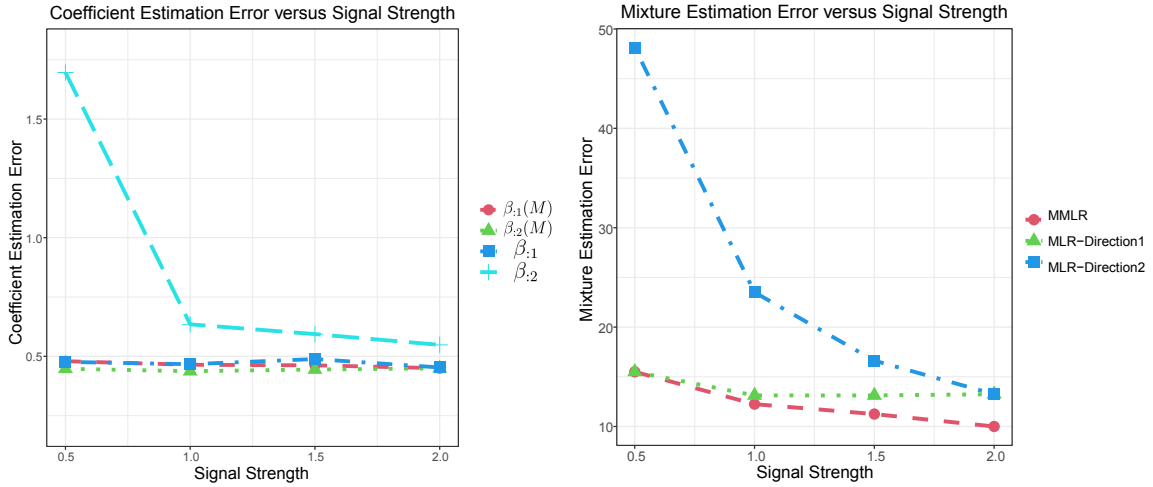

Figure 4: The reported results for each signal strength are medians of 100 replicates. In the left panel, $\beta_{:j}(M)$ represents the estimation error for the j-th column of β using Algorithm 2; $\beta_{:j}$ represents the estimation error for the j-th column of β using Algorithm 1 separately for each response. In the right panel, MMLR represents the estimation error for the mixtures using Algorithm 2; MLR-Direction1 and MLR-Direction2 represent the estimation error for the mixtures using Algorithm 1 separately for the first and second element of the response, respectively.

low level for all signal strengths and the mixture estimation error gains slightly improvement when the signal strength of the second response increases. When the signal strength of the second response is weak, the strong signal of the first response assists the recovery of the mixtures and helps increase the coefficient estimation accuracy of the second coefficient substantially. As a comparison, when using Algorithm 1 separately for each response, the estimation errors for the second coefficient and mixture increase quickly when the signal strength decreases. In terms of the mixture estimation error, it even performs like random guessing when the δ is 0.5.

The following lemmas and theorem rigorously show the statistical convergence for Algorithm 2.

Lemma 6 *Under conditions (C1),(C3'), if $q = O(s)$ and $\theta \in \mathcal{B}_{con}(\theta^*)$, then*

$$d_F\big(M(\theta), M(\theta^*)\big) \le \kappa_0\big(|\omega_1(\theta) - \omega_1^*| \vee \|\beta_1 - \beta_1^*\|_F \vee \|\beta_2 - \beta_2^*\|_F\big).$$

for some $0 < \kappa_0 < \frac{1}{2\vee(64/\tau_0)}$.

Lemma 7 *Suppose that $\theta^* \in \Theta^*$. Under condition (C1), there exists a constant $C_{con} > 0$, such that with probability at least $1 - 4(pq)^{-1}$,*

$$\sup_{\theta \in \mathcal{B}_{con}(\theta^*)} d_{F,s}(M(\theta), M_n(\theta)) \leq C_{con} \sqrt{\frac{s\log(n)^2\log(pq)}{n}}$$

Lemmas 6 & 7 can be interpreted similarly to Lemmas 1 and 2, respectively. However, since we are now interested in multivariate response, we further assume condition $q = O(s)$ in addition to the technical conditions (C1) to (C4). The assumption on q results from the signal-to-noise ratio. Note that $\Delta = O(s/q)$. In order for the condition $\Delta > C_1$ to hold, we need $q = O(s)$. However, when this assumption holds, q, i.e, the dimensionality of Y, has a relatively small effect on the estimation error, as we only have an additional factor of $\log q$ in Lemma 7 in comparison to Lemma 3.

Combine Lemmas 6 & 7 and we have the following theorem.

Theorem 8 *Under conditions (C1), (C2), (C3'), and(C4) and $q = O(s)$, there exists a constant $0 < \kappa < 1/2$, such that $\widehat{\beta}_k^{(t+1)}$ obtained by Algorithm 2 satisfies, with probability greater than $1 - 4(pq)^{-1}$,*

$$\|\widehat{\beta}_k^{(t+1)} - \beta_k^*\|_F = O\Big(\kappa^t(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_F \vee \|\widehat{\beta}_2^{(0)} - \beta_2^*\|_F) + \sqrt{\frac{s\log(n)^2\log(pq)}{n}}\Big).$$

Consequently, for $t \geq \{-\log(\kappa)\}^{-1}\log\{n(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_F \vee \|\widehat{\beta}_2^{(0)} - \beta_2^*\|_F)\}$,

$$\|\widehat{\beta}_k^{(t+1)} - \beta_k^*\|_F = O\Big(\sqrt{\frac{s\log(n)^2\log(pq)}{n}}\Big).$$

Again, the convergence rate for multivariate mixture linear regression model is similar to that for the mixture linear regression model with an additional $\log(q)$ term. The convergence rate grows slowly as a function of q. However, theoretical requirement $\Delta > C_1$ and practical consideration about the estimation for Σ_y restrict q to grow linearly with s and slower than n. It would be interesting to develop a method that allows q to grow faster, which we leave a challenging future topic.

6. Real data analysis

The cancer cell line encyclopedia (CCLE) dataset contains 8-point dose-response curves for 24 chemical compounds across over 400 cell lines, which is a publicly available dataset at `www.broadinstitute.org/ccle` and is also considered in Li et al. (2019). Because the cell lines are not consistent for different chemical compounds, we consider chemical compounds: Lapatinib, AZD6244, and PD-0325901, three popular chemical compounds for cancer treatment. Analogous to Li et al. (2019), we use the area under the dose-response curve, also known as the activity area, to measure the sensitivity of a drug for each cell line. Besides the drug information, the data also contains the expression data of 18,926 genes for each cell line. Aiming to identify the genes sensitive to the chemical compounds, we treat the active area as the response and the gene expressions as the predictor. After

identifying cell lines treated using all three chemical compounds, we get the sample size $n = 490$. We keep $p = 500$ genes that are highly correlated with three responses (for each gene, we calculated the sum of its absolute correlations with the three responses). Due to the complexity of cancer, we expect the data to be heterogeneous.

We consider $K = 2, 3, 4, 5, 6$, and 8 for the proposed algorithms and denote individual mixture linear regression using Algorithm 1 as PEM and multivariate mixture linear regression using Algorithm 2 as MPEM. In addition, we consider the methods used in the simulation studies including HDEM, Initial method, and PSEM and LASSO regression as competitors. We repeatedly split the whole data set into $1 : 4$ ratios, and use $1/5$ of the data as the testing samples and the rest as the training samples. The process is repeated 100 times. We report the mean squared prediction errors. Please see Table 5 for the overall prediction errors and Figures 5 and 6 for the boxplots. More detailed comparison results for each response are presented in Section A of the appendix.

When $K = 5$, the overall mean squared prediction errors for MPEM is the smallest. However, from Figure 5, we can see that MPEM is insensitive to K. When $K = 4, 5$ and 6, the mean squared prediction errors are very close. As a comparison, PEM is more sensitive to K and achieves the best prediction error when $K = 4$. Two possible reasons make MPEM perform better than PEM. The first reason is that considering the three responses together results in a larger signal-to-noise ratio, which facilitates the mixture identification and coefficients estimation. The second reason is that, unlike MPEM, the estimated mixtures are different for the three responses in PEM, which may reduce the estimation accuracy. Although PEM performs slightly worse than MPEM, it still outperforms the other competitors, especially the LASSO regression. It implies that the data set is quite heterogeneous, the proposed mixture linear regression approach improves the performance of a single linear regression significantly.

When $K = 4$ and 6, we also consider the estimated label for each observation based on the full data set. The number of observations in each estimated mixture is reported in Table 6. Based on this table, we recommend using $K = 2$, 3 and 3 for PEM to the three responses, respectively, and using $K = 5$ for MPEM. Except from the first response, the recommended K's make PEM and MPEM achieve the lowest or among the lowest prediction error. For the first response, PEM achieves the smallest prediction error when $K = 6$. However, it only separates the observations into 2 mixtures. Note that the estimated label for the ith observation is $\text{argmax}_{k=1,\cdots,K} \widehat{\eta}_{i,k}(\widehat{\theta})$, it can happen that no observation is classified into some mixtures. For the first response, the two mixtures are more balanced with the increase of K, which may be one reason makes the prediction error smaller. Another reason is that the relationship between the first response and the predictors may be complicated and non-linear, which can not be captured by a mixture linear regression model. Hence, PEM tends to use more mixtures to approximate the true relationship. Another phenomena we observe is that MPEM uses more mixtures than PEM. This happens since the true mixtures can be different for the three responses. By separating the samples into more mixtures, MPEM becomes more accurate and robust. In addition, we considered using BIC to determine the number of mixture regression components. We selected K using criterion $\text{argmin}_K - 2\ell(\widehat{\theta}_K) + \log(n)K\widehat{s}_K$, where $\ell(\cdot)$ is the log-likelihood function of the mixture linear regression model, $\widehat{\theta}_K$ is the estimations from Algorithm 1 using K mixtures, and s_K is the number of selected predictors using K. The selected K's for individual regression

are 2, 3, and 3 for the three responses, respectively, and 5 for the multivariate mixture regression. This result is consistent with the previously suggested values.

K	PEM	MPEM	HDEM	Initial	PSEM
2	0.413 (0.005)	0.388 (0.004)	0.420 (0.004)	0.612 (0.009)	0.422 (0.005)
3	0.275 (0.004)	0.276 (0.004)	0.347 (0.004)	0.493 (0.010)	0.322 (0.004)
4	0.254 (0.005)	0.238 (0.004)	0.324 (0.004)	0.443 (0.008)	0.293 (0.005)
5	0.276 (0.007)	**0.237** (0.004)	0.315 (0.005)	0.416 (0.008)	0.298 (0.005)
6	0.302 (0.007)	0.239 (0.004)	0.310 (0.004)	0.395 (0.007)	0.304 (0.005)
8	0.347 (0.006)	0.243 (0.004)	0.310 (0.005)	0.357 (0.006)	0.312 (0.005)

Table 5: Mean squared prediction errors based on 100 replicates (with standard errors provided in parentheses). For LASSO, the result is 0.948 (0.026).

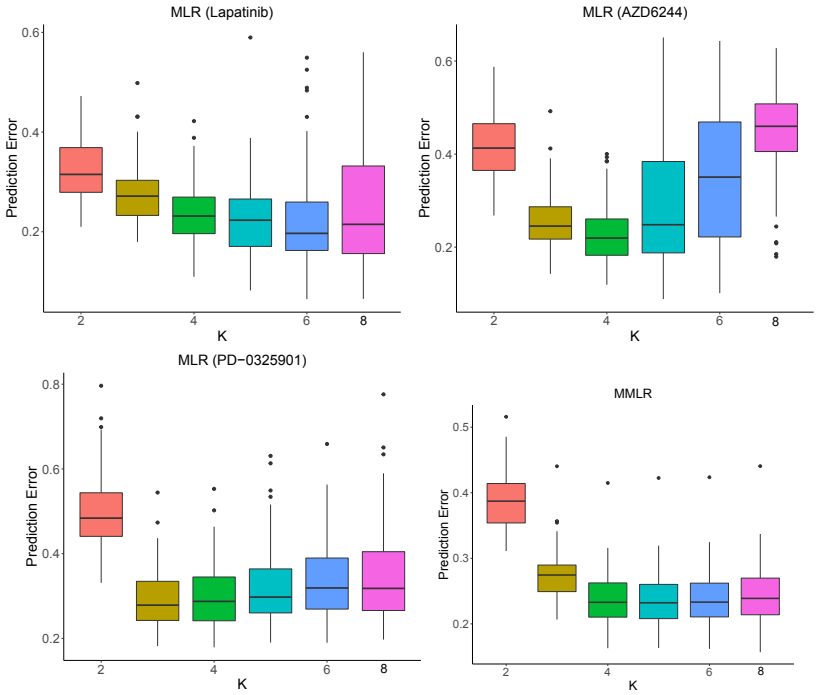

Figure 5: CCLE data: Boxplot for the mean squared prediction error

	K=4				K=6					
PEM (Lapatinib)	0	0	147	343	0	0	0	0	162	328
PEM (AZD6244)	0	87	162	241	0	0	0	85	187	218
PEM (PD-0325901)	0	105	170	215	0	0	0	101	180	209
MPEM	76	128	138	148	0	48	73	104	114	151

Table 6: Numbers of samples in each estimated mixtures.

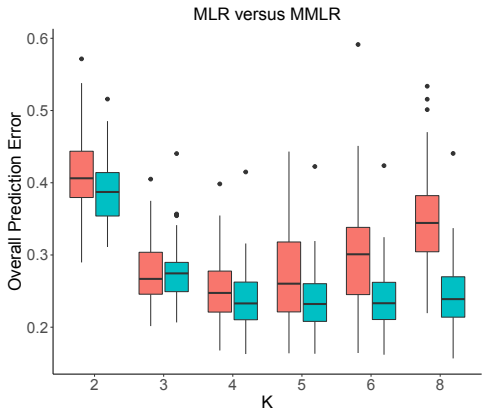

Figure 6: Overall prediction error for PEM (red) and MPEM (blue).

7. Discussion

The paper studies a group lasso penalized EM algorithm for high-dimensional mixture linear regression. We obtained an encouraging non-asymptotic convergence rate without data splitting and under a general model setting. Although our theory is for normally distributed predictors and two-mixtures regression, the penalized EM algorithm is applicable to essentially any predictors and to more than two mixtures. We then extended the mixture linear regression model and the penalized EM algorithm to the multivariate response case and established its non-asymptotic theory.

The theoretical study is currently for the model with two mixtures. It provides insights and foundations to the theoretical studies for cases with $K \geq 3$. We leave it as a future work. Besides, the error is assumed to follow a normal distribution, it is interesting to considering more general error assumptions, such as the student t-distribution, to achieve robustness. Finally, it is also meaningful future work to extend the studies to generalized mixture linear regression models.

Acknowledgments

We are grateful to the action editor and two reviewers for their insightful comments and valuable suggestions, which have helped us greatly in improving our manuscript. Research in this article is partly supported by grants CCF-1908969 (Mai), DMS-2053697 (Zhang), and DMS-2113590 (Zhang) from NSF, and 2023A1515110001 (Wang) from NSFC.

References

Radoslaw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of machine learning*

research, 15:2773–2832, 2014.

Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45 (1):77–120, 2017.

Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

T Tony Cai, Weidong Liu, and Xi Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, 2011.

T Tony Cai, Jing Ma, and Linjun Zhang. Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267, 2019.

Antoine Deleforge, Florence Forbes, and Radu Horaud. High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25 (5):893–911, 2015.

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)*, 39(1):1–22, 1977.

Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20: 1–29, 2015.

David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

Sangwon Hyun, Mattias Rolf Cape, Francois Ribalet, and Jacob Bien. Modeling cell populations measured by flow cytometry with covariates using sparse mixture of regressions. *The Annals of Applied Statistics*, 17(1):357 – 377, 2023.

Abbas Khalili and Jiahua Chen. Variable selection in finite mixture of regression models. *Journal of the american Statistical association*, 102(479):1025–1038, 2007.

Jason M Klusowski, Dana Yang, and WD Brinda. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory*, 65(6):3515–3524, 2019.

Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the minimax optimality of the em algorithm for learning two-component mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1405–1413. PMLR, 2021.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer Science & Business Media, 1991.

Friedrich Leisch. Flexmix: A general framework for finite mixture models and latent class regression in r. *Journal of Statistical Software*, 11(8):1–18, 2004.

Qianyun Li, Runmin Shi, and Faming Liang. Drug sensitivity prediction with high-dimensional mixture regression. *PloS one*, 14(2):e0212108, 2019.

Geoffrey J McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, 2004.

Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The annals of statistics*, 37(1):246–270, 2009.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557, 2012.

Gilles Pisier. Some applications of the metric entropy condition to harmonic analysis. In *Banach Spaces, Harmonic Analysis, and Probability Theory*, pages 123–154. Springer, 1983.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. In *Conference on Learning Theory*, pages 10–1. JMLR Workshop and Conference Proceedings, 2012.

Weixing Song, Weixin Yao, and Yanru Xing. Robust mixture regression model fitting by laplace distribution. *Computational Statistics & Data Analysis*, 71:128–137, 2014.

Nicolas Städler, Peter Bühlmann, and Sara Van De Geer. $\ell 1$-penalization for mixture regression models. *Test*, 19(2):209–256, 2010.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

T Rolf Turner. Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(3):371–384, 2000.

Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional em algorithm: Statistical optimization and asymptotic normality. *Advances in neural information processing systems*, 28:2512–2520, 2015.

Yi Yang and Hui Zou. A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141, 2015.

Weixin Yao, Yan Wei, and Chun Yu. Robust mixture regression using the t-distribution. *Computational Statistics & Data Analysis*, 71:116–127, 2014.

Fei Ye and Cun-Hui Zhang. Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *The Journal of Machine Learning Research*, 11:3519–3540, 2010.

Xinyang Yi and Constantine Caramanis. Regularized em algorithms: A unified framework and statistical guarantees. *Advances in Neural Information Processing Systems*, 28:1567–1575, 2015.

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pages 613–621, 2014.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Linjun Zhang, Rong Ma, T Tony Cai, and Hongzhe Li. Estimation, confidence intervals, and large-scale hypotheses testing for high-dimensional mixed linear regression. *arXiv preprint arXiv:2011.03598*, 2020.

Rongda Zhu, Lingxiao Wang, Chengxiang Zhai, and Quanquan Gu. High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm. In *International Conference on Machine Learning*, pages 4180–4188. PMLR, 2017.

Appendix

Appendix A contains additional results for the real data example. Appendix B contains implementation details and the convergence of Algorithm 1. Appendix C contains additional technical lemmas. Appendix D shows the proofs of Lemma 2 in the main paper. Appendix E shows the proofs of Lemma 3 in the main paper. Appendix F contains the proofs for Theorem 4. In Appendix G, we show the proof for Theorem 5. Appendix H contains the proofs for lemmas and the theorem in Section 5 of the main paper.

Appendix A. Additional results for real data analysis

In the paper, we only report the overall prediction error for all the three responses. The prediction error for each response is given in Table A.1.

	K=2					K=3				
	MLR	MMLR	HDEM	Initial	PSEM	MLR	MMLR	HDEM	Initial	PSEM
Lapatinib	0.325	0.337	0.265	0.318	0.302	0.276	0.353	0.234	0.242	0.266
	(0.006)	(0.007)	(0.004)	(0.006)	(0.005)	(0.005)	(0.007)	(0.004)	(0.006)	(0.005)
AZD6244	0.416	0.347	0.474	0.650	1.331	0.254	0.197	0.406	0.515	1.212
	(0.006)	(0.006)	(0.008)	(0.013)	(0.029)	(0.006)	(0.004)	(0.007)	(0.013)	(0.024)
PD-0325901	0.498	0.478	0.522	0.867	2.749	0.297	0.277	0.401	0.722	2.650
	(0.008)	(0.006)	(0.008)	(0.021)	(0.037)	(0.008)	(0.006)	(0.009)	(0.019)	(0.045)
Overall	0.413	0.388	0.420	0.612	1.461	0.275	0.276	0.347	0.493	1.376
	(0.005)	(0.004)	(0.004)	(0.009)	(0.020)	(0.004)	(0.004)	(0.005)	(0.009)	(0.021)
	K=4					K=5				
	MLR	MMLR	HDEM	Initial	PSEM	MLR	MMLR	HDEM	Initial	PSEM
Lapatinib	0.234	0.272	0.224	0.226	0.243	0.221	0.269	0.214	0.216	0.247
	(0.006)	(0.006)	(0.004)	(0.007)	(0.006)	(0.007)	(0.006)	(0.005)	(0.006)	(0.006)
AZD6244	0.228	0.206	0.377	0.446	0.608	0.292	0.207	0.371	0.428	0.455
	(0.006)	(0.005)	(0.010)	(0.012)	(0.020)	(0.014)	(0.005)	(0.010)	(0.011)	(0.010)
PD-0325901	0.299	0.234	0.372	0.658	2.385	0.316	0.233	0.360	0.605	1.006
	(0.008)	(0.005)	(0.009)	(0.016)	(0.051)	(0.009)	(0.005)	(0.012)	(0.018)	(0.033)
Overall	0.254	0.238	0.324	0.443	1.079	0.276	0.237	0.315	0.416	0.569
	(0.005)	(0.004)	(0.005)	(0.008)	(0.020)	(0.007)	(0.004)	(0.006)	(0.008)	(0.012)
	K=6					K=8				
	MLR	MMLR	HDEM	Initial	PSEM	MLR	MMLR	HDEM	Initial	PSEM
Lapatinib	0.219	0.273	0.205	0.209	0.246	0.248	0.277	0.200	0.201	0.263
	(0.009)	(0.006)	(0.005)	(0.006)	(0.005)	(0.012)	(0.006)	(0.005)	(0.006)	(0.005)
AZD6244	0.351	0.209	0.361	0.411	0.433	0.447	0.215	0.356	0.364	0.429
	(0.014)	(0.005)	(0.010)	(0.010)	(0.009)	(0.009)	(0.005)	(0.009)	(0.011)	(0.009)
PD-0325901	0.337	0.235	0.365	0.565	0.558	0.345	0.238	0.374	0.506	0.431
	(0.009)	(0.006)	(0.011)	(0.015)	(0.013)	(0.011)	(0.006)	(0.010)	(0.013)	(0.010)
Overall	0.302	0.239	0.310	0.395	0.412	0.347	0.243	0.310	0.357	0.375
	(0.007)	(0.004)	(0.006)	(0.007)	(0.08)	(0.006)	(0.004)	(0.005)	(0.006)	(0.006)

Table A.1: Mean squared prediction errors for each response and the overall mean squared prediction errors based on 100 replicates (Standard errors are given in parenthesis).

Appendix B. Implementation details and convergence of Algorithm 1

B.1 The groupwise majorization descent algorithm

Recall that in Algorithm 1 of the main paper, we update $\widehat{\beta}_k^{(t+1)}$, $k = 1, \cdots, K$, by minimizing

$$\sum_{k=1}^{K} \beta_k^T \widehat{\Sigma}_k^{(t+1)} \beta_k - 2 \sum_{k=1}^{K} (\widehat{\rho}_k^{(t+1)})^T \beta_k + \lambda_n^{(t+1)} \sum_{j=1}^{p} \sqrt{\sum_{k=1}^{K} \beta_{kj}^2}. \tag{A.1}$$

In this section, we elaborate on the application of the groupwise-majorization-descent algorithm (Yang and Zou, 2015) for solving this optimization.

Let $\beta = (\beta_1^T, \cdots, \beta_K^T)^T$, $L(\beta) = \sum_{k=1}^K \beta_k^T \widehat{\Sigma}_k^{(t+1)} \beta_k - 2 \sum_{k=1}^K (\widehat{\rho}_k^{(t+1)})^T \beta_k$, $\widetilde{\Sigma}^{(t+1)}$ be a block-wise diagonal matrix with the k-th diagonal block to be $\widehat{\Sigma}_k^{(t+1)}$, and $\widetilde{\rho}^{(t+1)} = \big((\widehat{\rho}_1^{(t+1)})^T, \cdots, (\widehat{\rho}_K^{(t+1)})^T\big)^T$. The function $L(\beta)$ can be equivalently written as $L(\beta) = \beta^T \widetilde{\Sigma}^{(t+1)} \beta - 2\beta^T \widetilde{\rho}^{(t+1)}$. Let groups $G_j = \{j, j+p, \cdots, j+(K-1)p\}$, $j = 1, \cdots, p$. Define β^j as the coefficient of β in group G_j, and H^j as the sub-matrix of H corresponding to group G_j. Let $\widetilde{\beta}$ be the current solution of β. By Taylor expansion, we have

$$L(\beta) \leq L(\widetilde{\beta}) + (\beta - \widetilde{\beta})^T U + \frac{1}{2} (\beta - \widetilde{\beta})^T H (\beta - \widetilde{\beta}),$$

where $U(\widetilde{\beta}) = -\nabla L(\widetilde{\beta})$ and $H = 2\widetilde{\Sigma}^{(t+1)}$. Following the Algorithm 1 of Yang and Zou (2015), the groupwise-majorization-descent algorithm for solving objective function (A.1) is summarized in Algorithm A.1.

Algorithm A.1 The groupwise-majorization-descent algorithm for solving $(\widehat{\beta}_1^{(t+1)}, \cdots, \widehat{\beta}_K^{(t+1)})$

- For $j = 1, \cdots, p$, compute r_j, which is the largest eigenvalue of H^j.
- Initialize $\widetilde{\beta}$.
- Repeat the following cyclic groupwise updates until convergence:
  For $j = 1, \cdots, p$, do the following steps:
  - Compute $U(\widetilde{\beta}) = -\nabla L(\widetilde{\beta})$.
  - Compute
  $$\widetilde{\beta}^j(\text{new}) = \frac{1}{r_j} (U^j + r_j \widetilde{\beta}^j) \big(1 - \frac{\lambda}{\|U^j + r_j \widetilde{\beta}^j\|_2}\big)_+.$$
  - Set $\widetilde{\beta}^j = \widetilde{\beta}^j(\text{new})$.
- Output $(\widehat{\beta}_1^{(t+1)}, \cdots, \widehat{\beta}_K^{(t+1)}) = \widetilde{\beta}$.

As is shown in Yang and Zou (2015), in the groupwise-majorization-descent algorithm, the objective function is strictly decreased after updating all groups in a cycle, unless

the solution does not change after each groupwise update. Thus, the objective function converges monotonically.

B.2 Proof of Lemma 1

The Algorithm 1 in the main paper includes two iterations: EM iteration and iteration for solving $\widehat{\beta}_k^{(t+1)}$, $k = 1, \cdots, K$. We have shown the convergence of the groupwise-majorization-descent algorithm used in second iteration. Thus, we only need to prove the convergence of the EM iteration.

Recall that the conditional log-likelihood function of $Y_i \mid X_i$ is given by

$$\ell(\theta; Y, X) = \frac{1}{n}\sum_{i=1}^{n} \log\big[\sum_{k=1}^{K} \frac{\omega_k}{\sqrt{2\pi}\sigma} \exp\big\{-\frac{(Y_i - X_i^T\beta_k)^2}{2\sigma^2}\big\}\big].$$

Consider the following regularized log-likelihood function of $Y_i \mid X_i$ with penalty $\lambda/2\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{K}\beta_{kj}^2}$,

$$\begin{aligned}\ell_1(\theta; Y, X) &= \ell(\theta; Y, X) - \lambda/2\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{K}\beta_{kj}^2} \\ &= \frac{1}{n}\sum_{i=1}^{n} \log\big[\sum_{k=1}^{K} \frac{\omega_k}{\sqrt{2\pi}\sigma} \exp\big\{-\frac{(Y_i - X_i^T\beta_k)^2}{2\sigma^2}\big\}\big] - \lambda/2\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{K}\beta_{kj}^2}.\end{aligned}$$

Because the Q-function

$$Q(\theta \mid \widehat{\theta}^{(t)}) = -\frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{K}\widehat{\eta}_{i,k}(\widehat{\theta}^{(t)})(Y_i - X_i^T\beta_k)^2 + \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\widehat{\eta}_{i,k}(\widehat{\theta}^{(t)})\log(\omega_k)$$

is a Minorization function of $\ell(\theta; Y, X)$ (See e.g. Hunter and Lange (2004) for more details about the definition of a Minorization function), we know that $Q(\theta \mid \widehat{\theta}^{(t)}) - \lambda/2\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{K}\beta_{kj}^2}$ is a Minorization function of $\ell_1(\theta; Y, X)$. The maximization of $Q(\theta \mid \widehat{\theta}^{(t)}) - \lambda/2\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{K}\beta_{kj}^2}$ can be found by the following two steps:

$$\begin{aligned}(\widehat{\beta}_1^{(t+1)}, \cdots, \widehat{\beta}_K^{(t+1)}) &= \underset{\beta_1, \cdots, \beta_k \in \mathbb{P}^p}{\operatorname{argmin}} \big\{\sum_{k=1}^{K}\beta_k^T\widehat{\Sigma}_k^{(t+1)}\beta_k - 2\sum_{k=1}^{K}(\widehat{\rho}_k^{(t+1)})^T\beta_k + \lambda\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{K}\beta_{kj}^2}\big\}, \\ \widehat{\omega}_k^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n}\widehat{\eta}_{i,k}(\widehat{\theta}^{(t)}),\ k = 1, \cdots, K.\end{aligned}$$

Those two updates are exactly those we use in the t-th EM iteration in Algorithm 1.

Thus, the update $(\widehat{\beta}_1^{(t+1)}, \cdots, \widehat{\beta}_K^{(t+1)}, \widehat{\omega}_1^{(t+1)}, \cdots, \widehat{\omega}_K^{(t+1)})$ makes the value of $\ell_1(\theta; Y, X)$ increase in each EM iteration of Algorithm 1. So the function value of $\ell_1(\theta; Y, X)$ converges monotonically. Combine this result with the convergence of the groupwise-majorization-descent algorithm, we know that, when $\lambda_n^{(t+1)} = \lambda$ for all t, Algorithm 1 converges.

Appendix C. Ancillary Lemmas

We first review the notations in the paper and introduce more notations that will be used in the Appendix. For numbers a and b, $a \vee b$ means $\max\{a, b\}$. For $n \in \mathbb{N}$, $[n]$ denotes the set $\{1, \cdots, n\}$. For numbers a and b, $a \vee b$ means $\max\{a, b\}$. For $n \in \mathbb{N}$, $[n]$ denotes the set $\{1, \cdots, n\}$. For a vector $x = (x_1, \cdots, x_p)$, $\|x\|_0$ is the number of non-zero elements in x, $\|x\|_1 = \sum_{i=1}^p |x_i|$, $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$, and $\|x\|_\infty = max_{i=1,\cdots,p}|x_i|$. The Frobenius norm of a matrix $A = (a_{ij})$ is defined as $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$. The ℓ_1 and ℓ_2 norms of a matrix A are defined as $\|A\|_1 = \sup_{\|x\|_1=1} \|Ax\|_1$ and $\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$, respectively. For a symmetric matrix A, we denote $\lambda_{min}(A)$ and $\lambda_{max}(A)$ as the smallest and largest eigenvalues of A, respectively. For a set $\mathcal{A}$, $\mathcal{A}^c$ denotes its complement. For two sequences of positive numbers a_n and b_n, $a_n = O(b_n)$ means $a_n \leq cb_n$ for some constant $c > 0$ for all n, $a_n = o(b_n)$ means that $a_n/b_n \to 0$ as $n \to \infty$, and $b_n \gg a_n$ means that $a_n = o(b_n)$. Let $\mathcal{B}_2^p$ and $\mathcal{S}^{p-1}$ be the unit Euclidean ball and the unit sphere, respectively. For a positive integer $s \leq p/2$, let set $\Gamma_{2p}(2s) = \{\mu \in \mathbb{R}^{2p} : \|\mu_{S^c}\|_1 \leq 5\sqrt{2s}\|\mu_S\|_2 + 2\sqrt{2s}\|\mu\|_2$ for some $S \subset [2p]$ with $|S| = 4s\}$ and $\Gamma(s) = \Gamma_{2p}(2s)_{1:p}$, where $\Gamma_{2p}(2s)_{1:p} = \{\mu_{1:p} : \mu \in \Gamma_{2p}(2s)\}$. For a vector X and a symmetric matrix A, we define $\|X\|_{2,s} = \sup_{\|\mu\|_2=1,\mu\in\Gamma(s)} \langle X, \mu\rangle$, and $\|A\|_{2,s} = \sup_{\|\mu\|_2=1,\mu\in\Gamma(s)} |\mu^T A\mu|$. To emphasize $\eta_{i,k}(\theta)$ is a random objects, we write it equivalently as $\eta_{k,\theta}(X_i, Y_i)$ in the proofs.

We first show some technical lemmas about the covering for $\Gamma(s)$. Because $\Gamma(s) = \Gamma_{2p}(2s)_{1:p}$. It suffices to find the covering for $\Gamma_{2p}(2s)$.

Lemma A.1 (Rudelson and Zhou (2012),Lemma 11) *Let $\mu_1, \cdots, \mu_M \in \mathbb{R}^p$ and $y \in \text{conv}(\mu_1, \cdots, \mu_M)$. There exists a set $L \subset [M]$, such that*

$$|L| \leq m = \frac{4 \max_{j\in[M]} \|\mu_j\|_2^2}{e\epsilon^2},$$

and a vector $y' \in \text{conv}(\mu_j : j \in L)$, such that

$$\|y' - y\|_2 \leq \epsilon.$$

Lemma A.2 (Rudelson and Zhou (2012),Lemma 21) *Let $\mu, \theta, x \in \mathbb{R}^q$ be vectors such that $\|\theta\|_2 = 1$, $\langle x, \theta\rangle \neq 0$, and μ is not parallel to x. Define $\phi : \mathbb{R} \to \mathbb{R}$ by:*

$$\phi(\lambda) = \frac{\langle x + \lambda\mu, \theta\rangle}{\|x + \lambda\mu\|_2}.$$

Assume $\phi(\lambda)$ has a local maximum at 0, then

$$\frac{\langle x + \mu, \theta\rangle}{\langle x, \theta\rangle} \geq 1 - \frac{\|\mu\|_2}{\|x\|_2}.$$

Lemma A.3 *Let $0 < s < p/2$ and $d = c_d s$ for some constant c_d, then*

$$\Gamma_p(s) \cap \mathcal{S}^{p-1} \subset 2\text{conv}(\cup_{|J|\leq d} E_J(p) \cap \mathcal{S}^{p-1}), \tag{A.2}$$

where conv *denotes the convex hull and $E_J(p) = \text{span}(e_j : j \in J)$.*

Proof The proof is analogous to that for Lemma 13 of Rudelson and Zhou (2012) with some modifications. For completeness, we present the proof here. We assume that $d < p$, otherwise the lemma is trivially true. For each $\mu \in \mathbb{R}^P$, let T_0 denote the locations of the s largest coefficients of μ in the absolute values. Decompose a vector $\mu \in \Gamma_p(s) \cap \mathcal{S}^{p-1}$ as

$$\mu = \mu_{T_0} + \mu_{T_0^c} \in \mu_{T_0} + (5\sqrt{s}\|\mu_{T_0}\|_2 + 2\sqrt{s}\|\mu\|_2) \cdot \text{absconv}(e_j : j \in T_0^c),$$

where absconv denotes the absolutely convex hull. Since

$$\|\mu_{T_0^c}\|_2^2 \leq \|\mu_{T_0^c}\|_1 \|\mu_{T_0^c}\|_\infty \leq (5\sqrt{s}\|\mu_{T_0}\|_2 + 2\sqrt{s}\|\mu\|_2)\frac{\|\mu_{T_0}\|_1}{s} \leq (5\|\mu_{T_0}\|_2 + 2\|\mu\|_2)\|\mu_{T_0}\|_2,$$

we have

$$1 = \|\mu\|_2^2 = \|\mu_{T_0}\|_2^2 + \|\mu_{T_0^c}\|_2^2 \leq 6\|\mu_{T_0}\|_2^2 + 2\|\mu\|_2\|\mu_{T_0}\|_2,$$

which implies $\|\mu_{T_0}\|_2 > \frac{1}{4}$.

Define $V = \{\mu_{T_0} + (5\sqrt{s}\|\mu_{T_0}\|_2 + 2\sqrt{s}\|\mu\|_2) \cdot \text{absconv}(e_j : j \in T_0^c) \mid \mu \in \Gamma_p(s) \cap \mathcal{S}^{p-1}\}$. We have $\Gamma_p(s) \cap \mathcal{S}^{p-1} \subset V \subset \Gamma_p(s)$ and V is compact. Therefore, V contains a base of $\Gamma_p(s)$, that is, for any $y \in \Gamma_p(s)/\{0\}$, there exists $\lambda > 0$ such that $\lambda y \in V$. For any nonzero $\nu \in \mathbb{R}^p$, we define

$$F(\nu) = \frac{\nu}{\|\nu\|_2}.$$

The function F is continuous on $\Gamma_p(s)/\{0\}$, and in particular, on V. Hence,

$$\Gamma_p(s) \cap \mathcal{S}^{p-1} = F(\Gamma_p(s)/\{0\}) = F(V).$$

By duality, inclusion (A.2) can be derived form the fact that the supremum of any linear functional over the left side of (A.2) does not exceed the supremum over the right side of it. By $\Gamma_p(s) \cap \mathcal{S}^{p-1} = F(\Gamma_p(s)/\{0\}) = F(V)$, it is enough to show that for any $\theta \in \mathcal{S}^{p-1}$, there exists $z' \in \mathbb{R}^p/\{0\}$ such that $\text{supp}(z') \leq d$ and $F(z')$ is well defined, which satisfies that

$$\max_{\nu \in V} \langle F(\nu), \theta) \rangle \leq 2\langle F(z'), \theta) \rangle. \tag{A.3}$$

For a given θ, we construct a d-sparse vector z', which satisfies (A.3). Let

$$z = \underset{\nu \in V}{\text{argmax}} \langle F(\nu), \theta) \rangle.$$

By definition of V, there exists $I \in [p]$ such that $|I| = s$, and for some $\eta_j \in \{1, -1\}$,

$$z = z_I + (5\sqrt{s}\|z_I\|_2 + 2\sqrt{s}\|z\|_2) \sum_{j \in I^c} \alpha_j \eta_j e_j,$$

where $\alpha_j \in [0, 1], \sum_{j \in I^v} \alpha_j \leq 1$, and $1 \geq \|z_I\|_2 \geq 1/4$.

Note that of $\alpha_i = 1$ for some $i \in I^c$, then z is a sparse vector, and we can set $z' = z$ in order for (A.3) to hold. We proceed by assuming $\alpha_i \in [0, 1)$ for all $i \in I^c$. We will

construct a required sparse vector z' via Lemma A.1. To satisfy the assumption of Lemma A.1, denote $e_{p+1} = 0$, $\eta_{p+1} = 1$, and set

$$\alpha_{p+1} = 1 - \sum_{j \in I^c} \alpha_j,$$

which makes $\alpha_{p+1} \in [0, 1]$. Let $y = z_{I^c}$, $\mathcal{M} = \{j \in I^c \cup p+1 : \alpha_j > 0\}$, and $\epsilon > 0$ that will be specified later. Applying Lemma A.3 with vector $\mu_j = (5\sqrt{s}\|z_I\|_2 + 2\sqrt{s}\|z\|_2)\eta_j e_j$ for $j \in \mathcal{M}$, construct a set $J' \subset \mathcal{M}$ satisfying

$$|J'| \le m := \frac{4 \max_{j \in I^c}(5\sqrt{s}\|z_I\|_2 + 2\sqrt{s}\|z\|_2)^2 \|e_j|_2^2}{\epsilon^2} \le \frac{196s}{\epsilon^2},$$

and a vector

$$y' = (5\sqrt{s}\|z_I\|_2 + 2\sqrt{s}\|z\|_2) \sum_{j \in J'} \beta_j \eta_j e_j,$$

where $\beta_j \in [0, 1]$ and $\sum_{j \in J'} \beta_j = 1$, such that $\|y - y'\|_2 \le \epsilon$.

Set $z' = z_{I^c} + y'$. By construction, $z' \in E_J$, where $J = (I \cup J') \cap [p]$ and $|J| \le s + m$. Furthermore, we have

$$\|z - z'\|_2 = \|y - y'\|_2 \le \epsilon.$$

For $\{\beta_j : j \in J'\}$ as above, we extend it to $\{\beta_j : j \in I^c \cup \{p+1\}\}$ by setting $\beta_j = 0$ if $j \in I^c \cup \{p+1\}/J'$ and write

$$z' = z_I + (5\sqrt{s}\|z_I\|_2 + 2\sqrt{s}\|z\|_2) \sum_{j \in I^c \cup \{p+1\}} \beta_j \eta_j e_j,$$

where $\beta_j \in [0, 1]$ and $\sum_{j \in I^c \cup \{p+1\}} \beta_j = 1$.

If $z' = z$, the conclusion holds. Otherwise, for some λ to br specified, consider the vector

$$z + \lambda(z' - z) = z_I + (5\sqrt{s}\|z_I\|_2 + 2\sqrt{s}\|z\|_2) \sum_{j \in I^c \cup \{p+1\}} [(1-\lambda)\alpha_j + \lambda\beta_j]\eta_j e_j.$$

We have $[(1-\lambda)\alpha_j + \lambda\beta_j] = 1$ and $(1-\lambda)\alpha_j + \lambda\beta_j \in [0, 1]$. Therefore, $\sum_{j \in I^c}[(1-\lambda)\alpha_j + \lambda\beta_j] \le 1$ and $z + \lambda(z' - z) \in V$.

Now we consider the following function ϕ,

$$\phi(\lambda) = \langle F(z + \lambda(z' - z)), \theta \rangle = \frac{\langle z + \lambda(z' - z), \theta \rangle}{\|z + \lambda(z' - z)\|_2}.$$

Since z is the maximizer of $\langle F(\nu), \theta \rangle$ for all ν, $\phi(\lambda)$ attains the local maximum at 0. Then by Lemma A.2, we have

$$\frac{\langle z', \theta \rangle}{\langle z, \theta \rangle} = \frac{\langle z + (z' - z), \theta \rangle}{\langle z', \theta \rangle} \ge 1 - \frac{\|z' - z\|_2}{\|z\|_2}.$$

It follows that

$$\frac{\langle F(z'),\theta\rangle}{\langle F(z),\theta\rangle} = \frac{\langle z'/\|z'\|_2,\theta\rangle}{\langle z/\|z\|_2,\theta\rangle} = \frac{\|z\|_2}{\|z'\|_2}\cdot\frac{\langle z',\theta\rangle}{\langle z,\theta\rangle} \geq \frac{\|z\|_2}{\|z\|_2+\|z'-z\|_2}\cdot\frac{\|z\|_2-\|z'-z\|_2}{\|z\|_2}$$
$$= \frac{\|z\|_2-\|z'-z\|_2}{\|z\|_2+\|z'-z\|_2} \geq 1-\frac{2\epsilon}{\|z\|_2+\epsilon}$$

Note that $\|z\|_2 \geq \|z_I\|_2 \geq 1/4$, setting $\epsilon = 1/12$. we have

$$\frac{\langle F(z'),\theta\rangle}{\langle F(z),\theta\rangle} \geq 1/2.$$

We have constructed a sparse z' such that (A.3) holds. Let $c_d = s + s\frac{196s}{\epsilon^2}$, we get the conclusion in Lemma A.3.

■

Lemma A.4 *The restrictive eigenvalue condition* $\inf_{\mu\in\Gamma(s)\cap\mathcal{S}^{p-1}} |\mu^T\sum_{i=1} X_iX_i^T\mu/n| > \tau_0$ *holds with high probability when* $n \gg s\log(p)$.

Proof Recall that $X_i \sim N(0,\Sigma)$. Let $\mathcal{C}_{2p}(2s) = \mathrm{conv}(\cup_{|J|\leq 2d}E_J(2p)\cap\mathcal{S}^{2p-1})$, and $\mathcal{C}(s) = \mathcal{C}_{2p}(2s)_{1:p}$. By Lemma A.3, $\Gamma(s)\subseteq\mathcal{C}(s)$. We will show a stronger conclusion that

$$\inf_{\nu\in\mathcal{C}(s)\cap\mathcal{S}^{p-1}} \nu^T\sum_{i=1}^n X_iX_i^T\nu/n \geq \tau_1.$$

Note that, for any $\nu\in\mathcal{C}(s)\cap\mathcal{S}^{p-1}$, $\nu^TX_i \sim N(0,\nu^T\Sigma\nu)$ and $\mathrm{E}(\nu^T\sum_{i=1}^n X_iX_i^T\nu/n) = \nu^T\Sigma\nu$. By Bernstein's inequality, we have

$$\mathbb{P}(|\nu^T\sum_{i=1}^n X_iX_i^T\nu/n - \nu^T\Sigma\nu| \geq t) \leq 2\exp(-cn\min\{t^2/L^2, t/L\}),$$

where $L = \|(\nu^TX_i)^2\|_{\psi_1} = 2\nu^T\Sigma\nu \leq M_2$. When $t \leq M_2$, we have

$$\mathbb{P}(|\nu^T\sum_{i=1}^n X_iX_i^T\nu/n - \nu^T\Sigma\nu| \geq t) \leq 2\exp(-c_1nt^2),$$

where $c_1 = c/L^2$.

Suppose that $\nu_1,\cdots,\nu_J$ is an ϵ-net of $\mathcal{C}(s)\cap\mathcal{S}^{p-1}$, by union bound

$$\mathbb{P}(|\nu_j^T\sum_{i=1}^n X_iX_i^T\nu_j/n - \nu_j^T\Sigma\nu_j| \geq t,\ j=1,\cdots,J) \leq 2|J|\exp(-c_1nt^2) \leq 2\exp(c_2s\log(p) - c_1nt^2).$$

Let $\Psi = \frac{1}{n}\sum_{i=1}^n X_iX_i^T$. We have

$$M_1 - t \leq \|\Psi^{1/2}\nu_j\|_2^2 \leq M_2 + t$$

with probability at least $1-2\exp(c_2 s\log(p)-c_1 nt^2)$, for all $j=1,\cdots,J$. It follows that

$$\sqrt{M_1-t}\le\|\Psi^{1/2}\nu_j\|_2\le\sqrt{M_2+t}.$$

Not that for any $\nu\in\mathcal{C}(s)\cap\mathcal{S}^{p-1}$, we can find a j such that $\|\nu-\nu_j\|_2\le\epsilon$ and

$$\|\Psi^{1/2}\nu_j\|_2-\|\Psi^{1/2}(\nu-\nu_j)\|_2\le\|\Psi^{1/2}\nu\|_2\le\|\Psi^{1/2}\nu_j\|_2+\|\Psi^{1/2}(\nu-\nu_j)\|_2.$$

The right hind side is upper bounded by $\sqrt{M_2+t}+\epsilon\sup_{\nu\in\mathcal{C}(s)\cap\mathcal{S}^{p-1}}\|\Psi^{1/2}\nu\|_2$. By taking supremum over all ν for $\|\Psi^{1/2}\nu\|_2$, we have

$$\sup_{\nu\in\mathcal{C}(s)\cap\mathcal{S}^{p-1}}\|\Psi^{1/2}\nu\|_2\le\frac{\sqrt{M_2+t}}{1-\epsilon}.$$

Meanwhile, the left hind side is lower bounded by $\sqrt{M_1-t}-\epsilon\sup_{\nu\in\mathcal{C}(s)\cap\mathcal{S}^{p-1}}\|\Psi^{1/2}\nu\|_2$. Thus

$$\inf_{\nu\in\mathcal{C}(s)\cap\mathcal{S}^{p-1}}\|\Psi^{1/2}\nu\|_2\ge\sqrt{M_1-t}-\frac{\epsilon\sqrt{M_2+t}}{1-\epsilon}.$$

Let $t=\frac{1}{2}M_1$ and $\epsilon=(1-\tau_1)\sqrt{M_1/2}/(\sqrt{M_2+M_1/2}+(1-\tau_1)\sqrt{M_1/2})$, we have

$$\inf_{\nu\in\mathcal{C}(s)\cap\mathcal{S}^{p-1}}\|\Psi^{1/2}\nu\|_2\ge\tau_1,$$

with probability at least $1-2\exp(c_2 s\log(p)-c_1 nM_1^2/4)$.

■

The following several lemmas are used for the proof of Lemma 3 in the main paper.

Lemma A.5 *Let $X_i(t)$, $i=1,\cdots,n$, be independent mean zero random processes indexed by points $t\in\mathcal{T}$, ϵ_i, $i=1,\cdots,n$, be independent Rademacher random variables and ϕ be a non-decreasing convex function. We have*

$$\mathrm{E}\{\phi(\sup_{t\in\mathcal{T}}\sum_{i=1}^n X_i(t))\}\le\mathrm{E}\{\phi(|2\sup_{t\in\mathcal{T}}\sum_{i=1}^n\epsilon_i X_i(t)|)\}$$

If we remove the mean zero assumption, then

$$\mathrm{E}\{\phi\big(\sup_{t\in\mathcal{T}}\sum_{i=1}^n(X_i(t)-\mathrm{E}X_i(t))\big)\}\le\mathrm{E}\{\phi(|2\sup_{t\in\mathcal{T}}\sum_{i=1}^n\epsilon_i X_i(t)|)\}.$$

Proof Let $Z_i(t)$ be an independent copy of $X_i(t)$, for $i=1,\cdots,n$. For the mean zero case, we have

$$\begin{aligned}\sup_{t\in\mathcal{T}}\sum_{i=1}^n X_i(t)&=\sup_{t\in\mathcal{T}}\sum_{i=1}^n\{X_i(t)-\mathrm{E}_Z(Z_i(t))\}\\&=\sup_{t\in\mathcal{T}}\mathrm{E}_Z\sum_{i=1}^n\{X_i(t)-Z_i(t)\}\\&\le\mathrm{E}_Z\sup_{t\in\mathcal{T}}\{\sum_{i=1}^n X_i(t)-\sum_{i=1}^n Z_i(t)\}\end{aligned}$$

Without mean zero assumption, we have

$$\begin{aligned}\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}\{X_i(t)-\mathrm{E}X_i(t)\} &= \sup_{t\in\mathcal{T}}\sum_{i=1}^{n}\{X_i(t)-\mathrm{E}_Z(Z_i(t))\}\\ &= \sup_{t\in\mathcal{T}}\mathrm{E}_Z\sum_{i=1}^{n}\{X_i(t)-Z_i(t)\}\\ &\le \mathrm{E}_Z\sup_{t\in\mathcal{T}}\{\sum_{i=1}^{n}X_i(t)-\sum_{i=1}^{n}Z_i(t)\}\end{aligned}$$

We can see that the term on the right hind side are the same for both cases. Thus, we only show the proof for the mean zero case. The proof for the case without mean zero assumption are exactly the same. We have

$$\begin{aligned}\phi\big(\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}X_i(t)\big) &\le \phi\big(\mathrm{E}_Z\sup_{t\in\mathcal{T}}\{\sum_{i=1}^{n}X_i(t)-\sum_{i=1}^{n}Z_i(t)\}\big)\\ &\le \mathrm{E}_Z\{\phi\big(\sup_{t\in\mathcal{T}}\{\sum_{i=1}^{n}X_i(t)-\sum_{i=1}^{n}Z_i(t)\}\big)\},\end{aligned}$$

where the second inequality follows by Jensen's inequality. Taking expectation with respect to X_i on both sides of last inequality, we have

$$\begin{aligned}\mathrm{E}_X\phi\big(\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}X_i(t)\big) &\le \mathrm{E}_{X,Z}\{\phi\big(\sup_{t\in\mathcal{T}}\{\sum_{i=1}^{n}X_i(t)-\sum_{i=1}^{n}Z_i(t)\}\big)\}\\ &= \mathrm{E}_{X,Z,\epsilon}\{\phi\big(\sup_{t\in\mathcal{T}}\{\sum_{i=1}^{n}\epsilon_i(X_i(t)-Z_i(t))\}\big)\}\\ &\le \mathrm{E}_{X,Z,\epsilon}\{\phi\big(|\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}\epsilon_iX_i(t)|+|\sup_{t\in\mathcal{T}}\epsilon_iZ_i(t)|\big)\}\\ &\le \frac{1}{2}\mathrm{E}_{X,\epsilon}\phi\big(2|\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}\epsilon_iX_i(t)|\big)+\frac{1}{2}\mathrm{E}_{Z,\epsilon}\phi\big(2|\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}\epsilon_iZ_i(t)|\big)\\ &= \mathrm{E}_{X,\epsilon}\phi\big(2|\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}\epsilon_iX_i(t)|\big)\end{aligned}$$

where we have the first equality because $\epsilon_i(X_i(t)-Z_i(t))$ has the same distribution as $X_i(t)-Z_i(t)$ and the last inequality because exp is convex. ■

Lemma A.6 (Cai et al. (2019), Lemma C.1) *Let $X_i(t)$, $i=1,\cdots,n$ be independent mean zero random processes indexed by points $t\in\mathcal{T}$, and ϵ_i, $i=1,\cdots,n$, be independent Rademacher random variables. Consider the Lipschitz functions $\psi_i(\cdot)$, for $i=1,\cdots,n$,*

with Lipschitz constant L that satisfies $\psi(0) = 0$. Then for any increasing convex function $\phi(\cdot)$ and a fixed $t_1 \in \mathcal{T}$, we have

$$\mathrm{E}\big\{\phi\big(|\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}\epsilon_i\psi_i(X_i(t))X_i(t_1)|\big)\big\} \le \mathrm{E}\big\{\phi\big(|2L\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}\epsilon_i X_i(t)X_i(t_1)\big)\big\}.$$

The next Lemma is about the tail inequality for supremum of unbounded random process.

Lemma A.7 (Adamczak (2008),Theorem 4) *Let $X_1, \cdots, X_n$ be independent random variables with values in a measurable space $(S, \mathcal{B})$ and let $\mathcal{F}$ be a countable class of measurable functions $f : S \to \mathbb{R}$. Assume that for every $f \in \mathcal{F}$ and every i, $\mathrm{E}f(X_i) = 0$, and for all i, $\|\sup_f |f(X_i)|\|_{\psi_\alpha} \le \infty$, where $0 \le \alpha \le 1$ and $\|\cdot\|_{\psi_\alpha}$ is the Orlicz ψ_α norm. Let $Z = \sup_{f\in\mathcal{F}}|\sum_{i=1}^{n} f(X_i)|$ and $\sigma^2 = \sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\mathrm{E}f(X_i)^2$. Then, for all $0 < \eta < 1$ and $\delta > 0$, there exists a constant $C = C(\eta, \delta)$, such that for all $t \ge 0$,*

$$\mathbb{P}(Z \ge (1+\eta)\mathrm{E}Z + t) \le \exp\Big(-\frac{t^2}{2(1+\delta)\sigma^2}\Big) + 3\exp\Big(-\big(\frac{t}{C\|\max_i \sup_{f\in\mathcal{F}}|f(X_i)|\|_{\psi_\alpha}}\big)^{1/\alpha}\Big),$$

$$\mathbb{P}(Z \ge (1-\eta)\mathrm{E}Z - t) \le \exp\Big(-\frac{t^2}{2(1+\delta)\sigma^2}\Big) + 3\exp\Big(-\big(\frac{t}{C\|\max_i \sup_{f\in\mathcal{F}}|f(X_i)|\|_{\psi_\alpha}}\big)^{1/\alpha}\Big).$$

The following Lemma shows some basic calculation results.

Lemma A.8

$$\int_{-\infty}^{\infty} e^{-a^2x^2+bx+c} = \sqrt{\frac{\pi}{a^2}}e^{\frac{b^2}{4a^2}+c}$$

$$\int_{0}^{\infty} xe^{-a^2x^2+bx}dx = \frac{1}{2a^2}\Big\{\frac{\sqrt{\pi}b}{2a}e^{b^2/4a^2}\widetilde{\Phi}(-\frac{b}{2a}) + 1\Big\}$$

$$\int_{-\infty}^{\infty} x^2e^{-a^2x^2+bx}dx = \frac{\sqrt{\pi}(2a^2+b^2)}{4a^5}e^{\frac{b^2}{4a}}$$

$$\int_{0}^{\infty} \Phi(ax)e^{-b^2x^2}x^2dx = \frac{\sqrt{2\pi}}{4b^3} - \frac{1}{2\sqrt{\pi}}[\frac{1}{b^3}\arctan(\frac{b}{a}) - \frac{a}{b^2(a^2+b^2)}]$$

$$\int_{0}^{\infty} \widetilde{\Phi}(ax)e^{-b^2x^2}x^2dx = \frac{1}{2\sqrt{\pi}}[\frac{1}{b^3}\arctan(\frac{b}{a}) - \frac{a}{b^2(a^2+b^2)}]$$

where $\Phi(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-x^2}dx$ and $\widetilde{\Phi}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-x^2}dx$.

Appendix D. Proof of Lemma 2

In this section, we aim to show that for $\theta \in \mathcal{B}_{con}(\theta^*)$,

$$d_2\big(M(\theta), M(\theta^*)\big) \le \kappa_0\big(|\omega_1(\theta) - \omega_1^*| \vee \|\beta_1 - \beta_1^*\|_2 \vee \|\beta_2 - \beta_2^*\|_2\big).$$

for some $0 < \kappa_0 < \frac{1}{2\vee(64/\tau_0)}$.

Before the proof, we first show what the parameters ρ_k^* and Σ^* are. The parameters ρ_k^* and Σ^* are not defined in the model, but we can get them from ω_k^* and β_k^*. We show the explicit form of them here. By definition

$$\begin{aligned}\rho_k^* = \rho_k(\theta^*) &= \mathrm{E}(\frac{1}{n}\sum_{i=1}^n \eta_{k,\theta^*}(X_i,Y_i)X_iY_i)\\ &= \mathrm{E}(\eta_{k,\theta^*}(X_i,Y_i)X_iY_i) = \mathrm{E}_{Y,X}(\mathrm{E}_{W|Y,X}(I(W_i=k))X_iY_i)\\ &= \mathrm{E}_{Y,X}(\mathrm{E}_{W|Y,X}(X_iY_iI(W_i=k))) = \mathrm{E}_{Y,X,W}(X_iY_iI(W_i=k))\\ &= \mathrm{E}_W\{I(W_i=k)\mathrm{E}_{X|W=k}(X_i\mathrm{E}_{Y|X,W=k})\} = \omega_k^*\Sigma\beta_k^*.\end{aligned}$$

Similarly, we have

$$\Sigma^* = \Sigma(\theta^*) = \omega_k^*\Sigma.$$

Note that we have $\Sigma^*\beta_k^* = \rho_k^*$.

We also have $\omega_k(\theta^*) = \mathrm{E}(\frac{1}{n}\sum_{i=1}^n \eta_{k,\theta^*}(X_i,Y_i)) = \mathrm{E}(\frac{1}{n}\sum_{i=1}^n \mathbb{P}(W_i = k \mid Y_i,X_i)) = \mathbb{P}(W_i=k) = \omega_1^*$.

Contraction for $\omega_1(\theta)$

Proof Recall that

$$\omega_1(\theta) - \omega_1(\theta^*) = \mathrm{E}\big(\frac{1}{n}\sum_{i=1}^n(\eta_{1,\theta}(X_i,Y_i) - \eta_{1,\theta^*}(X_i,Y_i))\big) = \mathrm{E}\big(\eta_{1,\theta}(X_i,Y_i) - \eta_{1,\theta^*}(X_i,Y_i)\big).$$

Let $\xi = (\omega_1, \beta_2-\beta_1, \beta_1+\beta_2)$. With a little abuse of notations, we can write $\eta_{1,\theta}(X_i,Y_i)$ as $\eta_{1,\xi}(X_i,Y_i)$. Let $\Delta_\xi = \xi - \xi^*$ and $\xi_u = \xi^* + u\Delta_\xi$. Then

$$\begin{aligned}&\mathrm{E}\{\eta_{1,\xi}(X_i,Y_i) - \eta_{1,\xi^*}(X_i,Y_i)\}\\ &= \mathrm{E}\big\{\int_0^1 \langle\frac{d\eta_{1,\xi}(X_i,Y_i)}{d\xi}\Big|_{\xi=\xi_u}, \Delta_\xi\rangle du\big\}\\ &= \mathrm{E}\big\{\int_0^1 \langle\frac{\partial\eta_{1,\xi}(X_i,Y_i)}{\partial\omega_1}\Big|_{\xi=\xi_u}, \Delta_{\omega_1}\rangle du\big\} + \mathrm{E}\big\{\int_0^1 \langle\frac{\partial\eta_{1,\xi}(X_i,Y_i)}{\partial(\beta_2-\beta_1)}\Big|_{\xi=\xi_u}, \Delta_{\beta_2-\beta_1}\rangle du\big\}\\ &\quad + \mathrm{E}\big\{\int_0^1 \langle\frac{\partial\eta_{1,\xi}(X_i,Y_i)}{\partial(\beta_2+\beta_1)}\Big|_{\xi=\xi_u}, \Delta_{\beta_2+\beta_1}\rangle du\big\}\\ &\le \Big|\int_0^1 \mathrm{E}(\frac{\partial\eta_{1,\xi}(X_i,Y_i)}{\partial\omega_1})\Big|_{\xi=\xi_u} du(\omega_1-\omega_1^*)\Big| + \Big|\langle\int_0^1 \mathrm{E}(\frac{\partial\eta_{1,\xi}(X_i,Y_i)}{\partial(\beta_2-\beta_1)})\Big|_{\xi=\xi_u} du, \Delta_{\beta_2-\beta_1}\rangle\Big|\\ &\quad + \Big|\langle\int_0^1 \mathrm{E}(\frac{\partial\eta_{1,\xi}(X_i,Y_i)}{\partial(\beta_2+\beta_1)})\Big|_{\xi=\xi_u} du, \Delta_{\beta_2+\beta_1}\rangle\Big|\\ &\le \underbrace{\sup_{\xi\in\mathcal{B}_{con}(\theta^*)}(|\mathrm{E}\frac{\partial\eta_{1,\xi}(X_i,Y_i)}{\partial\omega_1}|)|\omega_1-\omega_1^*|}_{(i)} + \underbrace{\sup_{\xi\in\mathcal{B}_{con}(\theta^*)}(\|\mathrm{E}\frac{\partial\eta_{1,\xi}(X_i,Y_i)}{\partial(\beta_2-\beta_1)}\|_2)\|_2\beta_1-\beta_1^*-\beta_2+\beta_2^*\|_2}_{(ii)}\\ &+ \underbrace{\sup_{\xi\in\mathcal{B}_{con}(\theta^*)}(\|\mathrm{E}\frac{\partial\eta_{1,\xi}(X_i,Y_i)}{\partial(\beta_1+\beta_2)}\|_2)\|\beta_1-\beta_1^*+\beta_2-\beta_2^*\|_2}_{(iii)}\end{aligned} \tag{A.4}$$

Thus, we only need to bound the three terms of last inequalities. By calculation, we have

$$\frac{\partial \eta_{1,\xi}(X_i,Y_i)}{\partial \omega_1} = \frac{\exp\{(\beta_2-\beta_1)^T X_i(Y_i-(\frac{\beta_1+\beta_2}{2})^T X_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^T X_i(Y_i-(\frac{\beta_1+\beta_2}{2})^T X_i)\}\big)^2}$$
$$\frac{\partial \eta_{1,\xi}(X_i,Y_i)}{\partial (\beta_2-\beta_1)} = -\omega_1(1-\omega_1)\frac{\exp\{(\beta_2-\beta_1)^T X_i(Y_i-(\frac{\beta_1+\beta_2}{2})^T X_i)\}X_i(Y_i-(\frac{\beta_1+\beta_2}{2})^T X_i)}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^T X_i(Y_i-(\frac{\beta_1+\beta_2}{2})^T X_i)\}\big)^2}$$
$$\frac{\partial \eta_{1,\xi}(X_i,Y_i)}{\partial (\beta_1+\beta_2)} = \omega_1(1-\omega_1)\frac{\exp\{(\beta_2-\beta_1)^T X_i(Y_i-(\frac{\beta_1+\beta_2}{2})^T X_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^T X_i(Y_i-(\frac{\beta_1+\beta_2}{2})^T X_i)\}\big)^2}\frac{X_i X_i^T(\beta_2-\beta_1)}{2}.$$

Note that $Y_i \sim_d Z_i + \psi_i^T X_i$, where $Z_i \sim N(0,1)$ independent of X_i and W_i, and ψ_i takes two values with $\mathbb{P}(\psi_i=\beta_1^*)=\mathbb{P}(W_i=1)=\omega_1^*$ and $\mathbb{P}(\psi_i=\beta_2^*)=\mathbb{P}(W_i=2)=1-\omega_1^*$. Define $\delta_i(\beta)=\psi_i-(\beta_1+\beta_2)/2$. We have $Y_i-(\frac{\beta_1+\beta_2}{2})X_i \sim_d Z_i+\delta_i(\beta)^T X_i$. Then

$$\frac{\partial \eta_{1,\xi}(X_i,Y_i)}{\partial \omega_1} = \frac{\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}\big)^2}$$

Define event $\mathcal{E}_i=\{|Z_i|\le 1/2|\delta_i(\beta)^T X_i| \mid X_i\}$. We have

$$\mathbb{P}(\mathcal{E}_i^c) \le \sum_{k=1}^2 \omega_k^* \mathbb{P}(|Z_i|\ge 1/2|\delta_i(\beta)^T X_i| \mid W_i=k, X_i) \le \sum_{k=1}^2 \exp\{-\frac{((\beta_k^*-\frac{\beta_1+\beta_2}{2})^T X_i)^2}{8}\}.$$

Note that on the event $\mathcal{E}_i$, we have $|(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)| \ge \frac{1}{2}|(\beta_2-\beta_1)^T X_i||\delta_i(\beta)^T X_i|$, and the function $f_1(t)=\frac{e^t}{(w_1+(1-w_1)e^t)^2} \le \frac{1}{\omega_1(1-\omega_1)}$ and $\sup_{|t|>|a|} f_1(t) \le \frac{1}{\min\{w_1^2,(1-\omega_1)^2\}}\exp(-|a|)$. Then

$$\begin{aligned}
&\mathrm{E}\Big(\frac{\partial \eta_{1,\xi}(X_i,Y_i)}{\partial \omega_1}\Big)\\
&= \mathrm{E}\Big(\frac{\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}\big)^2}\Big)\\
&= \mathrm{E}_{W,X}\Big\{\mathrm{E}_{Z|X,W}\Big(\frac{\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}\big)^2} \mid \mathcal{E}_i\Big)\mathbb{P}(\mathcal{E}_i)\\
&\quad + \mathrm{E}_{Z|X,W}\Big(\frac{\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}\big)^2} \mid \mathcal{E}_i^c\Big)\mathbb{P}(\mathcal{E}_i^c)\Big\}\\
&\le \frac{1}{\min\{w_1^2,(1-\omega_1)^2\}}\mathrm{E}_{W,X}\Big\{\exp\{-\frac{1}{2}|(\beta_2-\beta_1)^T X_i(\beta_W^*-\frac{\beta_1+\beta_2}{2})^T X_i|\}\\
&\quad + \frac{1}{\omega_1(1-\omega_1)}\big[\exp\{-\frac{((\beta_W^*-\frac{\beta_1+\beta_2}{2})^T X_i)^2}{8}\}\big]\Big\}\\
&\le \frac{1}{\min\{w_1^2,(1-\omega_1)^2\}}\mathrm{E}_W\mathrm{E}_{X|W}\Big\{\exp\{-\frac{1}{2}((\beta_2-\beta_1)^T X_i)^2\}+\exp\{-\frac{1}{2}((\beta_W^*-\frac{\beta_1+\beta_2}{2})^T X_i)^2\}\Big\}\\
&\quad + \frac{1}{\omega_1(1-\omega_1)}\mathrm{E}_W\mathrm{E}_{X|W}\Big\{\big[\exp\{-\frac{((\beta_W^*-\frac{\beta_1+\beta_2}{2})^T X_i)^2}{8}\}\big]\Big\}.
\end{aligned}$$

Note that $(\beta_2-\beta_1)^T X_i \mid W_i=k \sim N(0,(\beta_2-\beta_1)^T\Sigma(\beta_2-\beta_1))$, by Lemma A.8, we have

$$\begin{aligned}\mathrm{E}_{X|W=k}\big(\exp\{-\frac{1}{2}((\beta_2-\beta_1)^T X_i)^2\}\big) &= \frac{1}{\sqrt{1+(\beta_2-\beta_1)^T\Sigma(\beta_2-\beta_1)}}\\ &\le \frac{1}{\sqrt{(\beta_2-\beta_1)^T\Sigma(\beta_2-\beta_1)}}.\end{aligned}$$

Similarly, we have

$$\mathrm{E}_{X|W=k}\big(\exp\{-\frac{((\beta_k^*-\frac{\beta_1+\beta_2}{2})^T X_i)^2}{2}\}\big) \le \frac{1}{\sqrt{(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma(\beta_k^*-\frac{\beta_1+\beta_2}{2})}}.$$

Then

$$\begin{aligned}&\mathrm{E}(\frac{\partial\eta_{1,\xi}(X_i,Y_i)}{\partial\omega_1})\\ &= \mathrm{E}\Big(\frac{\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}\big)^2}\Big)\\ &\le \frac{1}{\min\{w_1^2,(1-\omega_1)^2\}}\sum_{k=1}^2\omega_k^*\Big\{\frac{1}{\sqrt{(\beta_2-\beta_1)^T\Sigma(\beta_2-\beta_1)}}+\frac{1}{\sqrt{(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma(\beta_k^*-\frac{\beta_1+\beta_2}{2})}}\Big\}\\ &\quad+\frac{1}{\omega_1(1-\omega_1)}\sum_{k=1}^2\omega_k^*\big(\frac{1}{\sqrt{4(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma(\beta_k^*-\frac{\beta_1+\beta_2}{2})}}\big).\end{aligned}$$

By the definition of the contraction basin $\mathcal{B}_{con}(\theta)$, we have

$$\begin{aligned}\sqrt{(\beta_2-\beta_1)^T\Sigma(\beta_2-\beta_1)} &\ge \sqrt{(\beta_2^*-\beta_1^*)^T\Sigma(\beta_2^*-\beta_1^*)}-\sqrt{(\beta_2-\beta_2^*-\beta_1+\beta_1^*)^T\Sigma(\beta_2-\beta_2^*-\beta_1+\beta_1^*)}\\ &\ge \Delta-4C_b\Delta M_2.\end{aligned}$$

When $C_b \le 1/(4M_2)$, $\sqrt{(\beta_2-\beta_1)^T\Sigma(\beta_2-\beta_1)} \ge c\Delta$. Similar conclusion also holds for $\sqrt{(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma(\beta_k^*-\frac{\beta_1+\beta_2}{2})}$. Hence,

$$\mathrm{E}(\frac{\partial\eta_{1,\xi}(X_i,Y_i)}{\partial\omega_1}) \le \frac{C_1}{\Delta},$$

for some positive constant C_1. Thus, when $\Delta \ge C_1/\kappa_0$, $\mathrm{E}(\frac{\partial\eta_{1,\xi}(X_i,Y_i)}{\partial\omega_1}) \le \kappa_0$.

To bound (ii), recall that

$$\frac{\partial\eta_{1,\xi}(X_i,Y_i)}{\partial(\beta_2-\beta_1)} = -\omega_1(1-\omega_1)\frac{\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}\big)^2}\cdot X_i(Z_i+\delta_i(\beta)^T X_i),$$

It follows that

$$
\begin{aligned}
&-\frac{1}{\omega_1(1-\omega_1)}\mathrm{E}\big(\frac{\partial \eta_{1,\xi}(X_i,Y_i)}{\partial(\beta_2-\beta_1)}\big)\\
&=\mathrm{E}_{W,X}\Big\{\mathrm{E}_{Z|X,W}\Big(\frac{\exp\{(\beta_2-\beta_1)^TX_i(Z_i+\delta_i(\beta)^TX_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^TX_i(Z_i+\delta_i(\beta)^TX_i)\}\big)^2}\cdot X_i(Z_i+\delta_i(\beta)^TX_i)\Big)\\
&=\mathrm{E}_{W,X}\Big\{\mathrm{E}_{Z|X,W}\Big(\frac{\exp\{(\beta_2-\beta_1)^TX_i(Z_i+\delta_i(\beta)^TX_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^TX_i(Z_i+\delta_i(\beta)^TX_i)\}\big)^2}\cdot X_i(Z_i+\delta_i(\beta)^TX_i)\mid\mathcal{E}_i\Big)\mathbb{P}(\mathcal{E}_i)\Big)\\
&\quad+\mathrm{E}_{W,X}\Big\{\mathrm{E}_{Z|X,W}\Big(\frac{\exp\{(\beta_2-\beta_1)^TX_i(Z_i+\delta_i(\beta)^TX_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^TX_i(Z_i+\delta_i(\beta)^TX_i)\}\big)^2}\cdot X_i(Z_i+\delta_i(\beta)^TX_i)\mid\mathcal{E}_i^c\Big)\mathbb{P}(\mathcal{E}_i^c)\Big)
\end{aligned}
$$

Let H_k be an orthogonal matrix whose first row is $(\beta_2-\beta_1)^T\Sigma^{1/2}/\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2$. Meanwhile, $(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}$ is in the span of the first two rows of H_k. Given $W_i=k$, we write X_i as $\Sigma^{1/2}H_k^TV_i$, where $V_i\sim N(0,I_p)$. Then $(\beta_2-\beta_1)^TX_i\mid(W_i=k)=\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2V_{i1}$, and $(\beta_k^*-\frac{\beta_1+\beta_2}{2})^TX_i\mid(W_i=k)=\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2(\lambda_1V_{i1}+\lambda_2V_{i2})$, where $\lambda_1^2+\lambda_2^2=1$. For notation simplicity, we write $\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2V_{i1}$ as $T_1(V_{i1})$, and $\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2(\lambda_1V_{i1}+\lambda_2V_{i2})$ as $T_2(V_{i1},V_{i2})$.

Then, given $W_i=k$,

$$
\begin{aligned}
\frac{\partial \eta_{1,\xi}(X_i,Y_i)}{\partial(\beta_2-\beta_1)}=&-\omega_1(1-\omega_1)\Sigma^{1/2}H_k^T\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}\\
&\cdot V_i(Z_i+T_2(V_{i1},V_{i2})).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
-\frac{1}{\omega_1(1-\omega_1)}\mathrm{E}\big(\frac{\partial \eta_{1,\xi}(X_i,Y_i)}{\partial(\beta_2-\beta_1)}\big)=\sum_{k=1}^{2}\omega_k^*\mathrm{E}_{V|W}\Big\{\mathrm{E}_{Z|V,W}\Big(&\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}\\
&\cdot V_i(Z_i+T_2(V_{i1},V_{i2}))\Big)\Big\}.
\end{aligned}
\tag{A.5}
$$

Because $V_{i1},\cdots,V_{ip}$ are independent, for $j=3,\cdots,p$, we have

$$
\begin{aligned}
&\sum_{k=1}^{2}\omega_k^*\mathrm{E}_{V|W}\Big\{\mathrm{E}_{Z|V,W}\Big(\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}\cdot V_{ij}(Z_i+T_2(V_{i1},V_{i2}))\Big)\Big\}\\
&=\sum_{k=1}^{2}\omega_k^*\mathrm{E}_{V|W}\Big\{V_{ij}\mathrm{E}_{Z|V,W}\Big(\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}(Z_i+T_2(V_{i1},V_{i2}))\Big)\Big\}\\
&=0,
\end{aligned}
$$

where the last equality follows since $\omega_1^*\mu_1+\omega_2^*\mu_2=0$. Hence, we only need to elements of (A.5). When $j=1$,

$$
\begin{aligned}
&\sum_{k=1}^{2}\omega_k^*\mathrm{E}_{V|W}\Big\{\mathrm{E}_{Z|V,W}\Big(\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}V_{i1}(Z_i+T_2(V_{i1},V_{i2}))\Big)\Big\}\\
&=\sum_{k=1}^{2}\omega_k^*\mathrm{E}_{V|W}\Big\{\mathrm{E}_{Z|V,W}\Big(\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}\\
&\quad\cdot T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\Big)\Big\}/\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2\\
&=\sum_{k=1}^{2}\omega_k^*\mathrm{E}_{V|W}\Big\{\mathrm{E}_{Z|V,W}\Big(\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}\\
&\quad\cdot T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\mid\mathcal{E}_i\Big)\mathbb{P}(\mathcal{E}_i)\Big\}/\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2\\
&\quad+\sum_{k=1}^{2}\omega_k^*\mathrm{E}_{V|W}\Big\{\mathrm{E}_{Z|V,W}\Big(\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}\\
&\quad\cdot T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\mid\mathcal{E}_i^c\Big)\mathbb{P}(\mathcal{E}_i^c)\Big\}/\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2\\
&\le\frac{1}{c_0^2\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2}\sum_{k=1}^{2}\omega_k^*\mathrm{E}_{X|W}\Big\{\exp\{-\frac{3}{2}((\beta_2-\beta_1)^TX_i)^2\}+\exp\{-\frac{3}{2}((\beta_k^*-\frac{\beta_1+\beta_2}{2})^TX_i)^2\}\Big\}\\
&\quad+\frac{1}{4c_0^2\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2}\sum_{k=1}^{2}\omega_k^*\mathrm{E}_{X|W}\Big\{\big[\exp\{-\frac{((\beta_k^*-\frac{\beta_1+\beta_2}{2})^TX_i)^2}{8}\}\big]\Big\}.
\end{aligned}
$$

In the last inequality, we transform V_i back to X_i, and then use the fact that $f_2(t)=|\frac{e^t t}{(w_1+(1-w_1)e^t)^2}|\le\frac{1}{4\min\{\omega_1^2,(1-\omega_1)^2\}}$ and $\sup_{|t|>|a|}f_2(t)\le\frac{1}{\min\{w_1^2,(1-\omega_1)^2\}}\exp(-|\frac{3a}{2}|)$. By Lemma A.8,

$$
\mathrm{E}_{W,V}\Big\{\mathrm{E}_{Z|V,W}\Big(\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}V_{i1}(Z_i+T_2(V_{i1},V_{i2}))\Big)\Big\}\le\frac{C_1}{\Delta}. \tag{A.6}
$$

Thus, when $\Delta\ge C_1/\kappa_0$, the last expectation is smaller that κ_0.

Note that on event $\mathcal{E}_i$, we also have $|(Z_i+\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2(\lambda_1V_{i1}+\lambda_2V_{i2}))|\le\frac{3}{2}\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2|(\lambda_1V_{i1}+\lambda_2V_{i2})|$, namely $|(Z_i+T_2(V_{i1},V_{i2}))|\le\frac{3}{2}\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2|(\lambda_1V_{i1}+\lambda_2V_{i2})|$.

$$
\begin{aligned}
&\mathrm{E}_{Z|V,W}\Big(\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}V_{i2}(Z_i+T_2(V_{i1},V_{i2}))\Big)\\
\le\ &\mathrm{E}_{Z|V,W}\Big(\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}V_{i2}(Z_i+T_2(V_{i1},V_{i2}))\mid\mathcal{E}_i\Big)\mathbb{P}(\mathcal{E}_i)\\
&+\mathrm{E}_{Z|V,W}\Big(\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}V_{i2}(Z_i+T_2(V_{i1},V_{i2}))\mid\mathcal{E}_i^c\Big)\mathbb{P}(\mathcal{E}_i^c)\\
\le\ &\frac{3}{2c_0^2}\Big(\exp\{-\frac{1}{2}\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2V_{i1}^2\}|V_{i2}|\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2|(\lambda_1V_{i1}+\lambda_2V_{i2})|\Big)\\
&+\frac{3}{2c_0^2}\Big(\exp\{-\frac{1}{2}\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1V_{i1}+\lambda_2V_{i2})^2\}|V_{i2}|\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2|(\lambda_1V_{i1}+\lambda_2V_{i2})|\Big)\\
&+\frac{1}{2c_0^2\sqrt{\pi}}\int_{\frac{1}{2}\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2|(\lambda_1V_{i1}+\lambda_2V_{i2})|}^{\infty}V_{i2}(Z_i+T_2(V_{i1},V_{i2}))\exp(-Z_i^2/2)dZ_i\\
\le\ &\frac{3}{2c_0^2}\Big(\underbrace{\exp\{-\frac{1}{2}\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2V_{i1}^2\}|V_{i2}|\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2|(\lambda_1V_{i1}+\lambda_2V_{i2})|\Big)}_{(I)}\\
&+\underbrace{\frac{3}{2c_0^2}\Big(\exp\{-\frac{1}{2}\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1V_{i1}+\lambda_2V_{i2})^2\}|V_{i2}|\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2|(\lambda_1V_{i1}+\lambda_2V_{i2})|\Big)}_{(II)}\\
&+\underbrace{\frac{1}{2c_0^2\sqrt{\pi}}\exp\{-\frac{1}{2}\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1V_{i1}+\lambda_2V_{i2})^2\}|V_{i2}|}_{(III)}\\
&\cdot\underbrace{(\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2|(\lambda_1V_{i1}+\lambda_2V_{i2})|+1)}_{(III)},
\end{aligned}
$$

where the last inequality is obtained by standard integration and the fact that $\int_t^\infty \exp(-x^2/2)dx \le \exp(-t^2/2)$.

Since $(III)=\frac{\sqrt{\pi}}{3}((I)+(II))$, we only need to bound (I) and (II). For the term (I), we have

$$
\begin{aligned}
\mathrm{E}\big\{(I)\big\} &= \|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2\mathrm{E}_{V_{i1}}\{\exp(-\frac{1}{2}\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2V_{i1}^2)\mathrm{E}_{V_{i2}}\big(|\lambda_1V_{i1}V_{i2}+\lambda_2V_{i2}^2|\big)\}\\
&\le \|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2\mathrm{E}_{V_{i1}}\{\exp(-\frac{1}{2}\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2V_{i1}^2)(\mathrm{E}_{V_{i2}}|\lambda_1V_{i1}V_{i2}|+\mathrm{E}_{V_{i2}}|\lambda_2V_{i2}^2|)\}\\
&\le \|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2\mathrm{E}_{V_{i1}}\big\{\exp(-\frac{1}{2}\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2V_{i1}^2)|\lambda_1V_{i1}|\}\\
&\quad+\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2\mathrm{E}_{V_{i1}}\big\{\exp(-\frac{1}{2}\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2V_{i1}^2)|\lambda_2|\},
\end{aligned}
$$

By Lemma A.8, we have

$$\begin{aligned}
&\mathrm{E}_{V_{i1}}\{\exp(-\frac{1}{2}\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2V_{i1}^2)\lambda_1|V_{i1}|\}\\
&=2|\lambda_1|/\sqrt{2\pi}\int_0^\infty v\exp(-\frac{1}{2}\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2v^2)\exp(-\frac{1}{2}v^2)dv\\
&=\frac{2|\lambda_1|/\sqrt{2\pi}}{\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2+1}\\
&\le\frac{C_2}{\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2},
\end{aligned}$$

Also, we have

$$\begin{aligned}
&\mathrm{E}_{V_{i1}}\{\exp(-\frac{1}{2}\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2V_{i1}^2)|\lambda_2|\}\\
&=|\lambda_2|/\sqrt{2\pi}\int_{-\infty}^\infty\exp(-\frac{1}{2}\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2v^2)\exp(-\frac{1}{2}v^2)dv\\
&=\frac{|\lambda_2|}{\sqrt{\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2+1}}\\
&\le\frac{C_3|\lambda_2|}{\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2}\le\frac{C_3|\lambda_2|}{(1-c_1)\Delta}
\end{aligned}$$

Thus

$$\begin{aligned}
\mathrm{E}\big((I)\big)&\le\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2(\frac{C_2}{\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2}+\frac{C_3|\lambda_2|}{\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2})\\
&\le\frac{C_2(1+c_1)}{2(1-c_1)\Delta}+C_3|\lambda_2|.
\end{aligned}$$

For the term (II), letting $v_1=\lambda_1V_{i1}+\lambda_2V_{i2}$ and $v_2=V_{i2}$, we have

$$\begin{aligned}
\mathrm{E}((II))&=\frac{\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2}{2\pi\lambda_1}\int_{-\infty}^\infty|v_2|\exp(-\frac{1}{2}v_2^2)dv_2\\
&\quad\cdot\int_{-\infty}^\infty|v_1|\exp(-\frac{1}{2}\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2v_1^2)\exp(-\frac{1}{2}v_1^2)dv_1\\
&\le\frac{C_4\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2}{\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2}\le\frac{C_4(1+c_1)}{(1-c_1)^2\Delta}
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathrm{E}_{W,V}\Big\{\mathrm{E}_{Z|V,W}\Big(\Big|\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}V_{i2}(Z_i+T_2(V_{i1},V_{i2}))\Big|\Big)\Big\}\\
&\le\sum_{k=1}^2\omega_k^*\Big(\frac{3C_2(1+c_1)}{4c_0^2(1-c_1)\Delta}+\frac{3C_3|\lambda_2|}{2c_0^2}+\frac{3C_4(1+c_1)}{2c_0^2(1-c_1)^2\Delta}\Big).
\end{aligned}$$

Note that

$$|\lambda_2| \leq \frac{\|(\beta_k^* - \beta_k)^T \Sigma^{1/2}\|_2}{\|(\beta_k^* - \frac{\beta_1+\beta_2}{2})^T \Sigma^{1/2}\|_2} \leq \frac{C_b\sqrt{M_2}}{1-c_1}.$$

When $\Delta \geq \frac{9C_2(1+c_1)}{4c_0^2(1-c_1)\kappa_0} \vee \frac{9C_4(1+c_1)}{2c_0^2(1-c_1)^2\kappa_0}$ and $C_b \leq \frac{2\kappa_0 c_0^2(1-c_1)}{9c_3\sqrt{M2}}$, we have

$$\mathrm{E}_{W,V}\Big\{\mathrm{E}_{Z|V,W}\Big(\frac{\exp\{T_1(V_{i1})(Z_i + T_2(V_{i1}, V_{i2}))\}}{\big(\omega_1 + (1-\omega_1)\exp\{T_1(V_{i1})(Z_i + T_2(V_{i1}, V_{i2}))\}\big)^2} V_{i2}(Z_i + T_2(V_{i1}, V_{i2}))\Big)\Big\} \leq \kappa_0.$$

Combined with the bounded shown in (A.6), we have $\|\mathrm{E}\frac{\partial \eta_{1,\xi}(X_i,Y_i)}{\partial(\beta_1-\beta_2)}\|_2 \leq \kappa_0$.

Finally, we bound the term (iii), recall that

$$\mathrm{E}\frac{\partial \eta_{1,\xi}(X_i,Y_i)}{\partial(\beta_2+\beta_1)} = \omega_1(1-\omega_1)\mathrm{E}\Big(\frac{\exp\{(\beta_2-\beta_1)^T X_i(Z_i + \delta_i(\beta)^T X_i)\}}{\big(\omega_1 + (1-\omega_1)\exp\{(\beta_2-\beta_1)^T X_i(Z_i + \delta_i(\beta)^T X_i)\}\big)^2} \cdot X_i X_i^T(\beta_2-\beta_1)\Big)$$

Recall that H_k is an orthogonal matrix whose first row is $(\beta_2-\beta_1)^T\Sigma^{1/2}/\|(\beta_k^* - \frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2$. Meanwhile, $(\beta_k^* - \frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}$ is in the span of the first two rows of H. Given $W_i = k$, we write X_i as $\Sigma^{1/2}H^T V_i$, where $V_i \sim N(H\Sigma^{-1/2}\mu_k, I_p)$. We have $(\beta_2-\beta_1)^T X_i \mid W_i = k = \|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2 V_{i1} = T_1(V_{i1})$, and $(\beta_k^* - \frac{\beta_1+\beta_2}{2})^T X_i \mid W_i = k = \|(\beta_k^* - \frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2(\lambda_1 V_{i1} + \lambda_2 V_{i2}) = T_2(V_{i1}, V_{i2})$, where $\lambda_1^2 + \lambda_2^2 = 1$. Then

$$\begin{aligned}
&\frac{1}{\omega_1(1-\omega_1)}\mathrm{E}\frac{\partial \eta_{1,\xi}(X_i,Y_i)}{\partial(\beta_2+\beta_1)} \\
&= \sum_{k=1}^{2} \omega_k^* \Sigma^{1/2} H_k^T \mathrm{E}_{Z,V|W}\Big(\frac{\exp\{T_1(V_{i1})(Z_i + T_2(V_{i1}, V_{i2}))\}}{\big(\omega_1 + (1-\omega_1)\exp\{\|(\beta_2-\beta_1)^T\Sigma^{1/2}V_{i1}(Z_i + T_2(V_{i1}, V_{i2}))\}\big)^2} \\
&\cdot V_i \|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2 V_{i1}\Big)
\end{aligned}$$

As we discussed before, because $\omega_1^*\mu_1 + \omega_2^*\mu_2 = 0$ and V_{ij} are independent with each other for all j, the 3 to p elements of the last vector are exactly zero. Thus, we only need to

bound the first two elements of the last vector. When $j = 1$,

$$
\begin{aligned}
&\mathrm{E}_{V|W}\Big\{\mathrm{E}_Z\Big(\frac{\exp\{T_1(V_{i1})(Z_i + T_2(V_{i1}, V_{i2}))\}}{\big(\omega_1 + (1-\omega_1)\exp\{\|(\beta_2-\beta_1)^T\Sigma^{1/2}V_{i1}(Z_i + T_2(V_{i1}, V_{i2}))\}\big)^2} \\
&\quad \cdot \|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2 V_{i1}^2\Big)\Big\} \\
&= \mathrm{E}_{V|W}\Big\{\mathrm{E}_Z\Big(\frac{\exp\{T_1(V_{i1})(Z_i + T_2(V_{i1}, V_{i2}))\}}{\big(\omega_1 + (1-\omega_1)\exp\{\|(\beta_2-\beta_1)^T\Sigma^{1/2}V_{i1}(Z_i + T_2(V_{i1}, V_{i2}))\}\big)^2} \\
&\quad \cdot \|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2 V_{i1}^2 \mid \mathcal{E}_i\Big)\mathbb{P}(\mathcal{E}_i)\Big\} \\
&\quad + \mathrm{E}_{V|W}\Big\{\mathrm{E}_Z\Big(\frac{\exp\{T_1(V_{i1})(Z_i + T_2(V_{i1}, V_{i2}))\}}{\big(\omega_1 + (1-\omega_1)\exp\{\|(\beta_2-\beta_1)^T\Sigma^{1/2}V_{i1}(Z_i + T_2(V_{i1}, V_{i2}))\}\big)^2} \\
&\quad \cdot \|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2 V_{i1}^2 \mid \mathcal{E}_i\Big)\mathbb{P}(\mathcal{E}_i)\Big\} \\
&\le \|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2\|_2 \mathrm{E}_{V|W}\Big\{\exp\{-\frac{1}{2}\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2 V_{i1}^2\}V_{i1}^2\Big\} \\
&\quad + \|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2 \mathrm{E}_{V|W}\Big\{\exp\{-\frac{1}{2}\|(\beta_k^* - \frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1 V_{i1} + \lambda_2 V_{i2})^2\}V_{i1}^2\Big\} \\
&\le \frac{C_5}{\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2} + \\
&\quad \|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2 \mathrm{E}_{V|W}\Big\{\exp\{-\frac{1}{2}\|(\beta_k^* - \frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1 V_{i1} + \lambda_2 V_{i2})^2\}V_{i1}^2\Big\}
\end{aligned}
$$

Let $v_1 = \lambda_1 V_{i1} + \lambda_2 V_{i2}$ and $v_2 = V_{i2}$, by Lemma A.8, we have

$$
\begin{aligned}
&\mathrm{E}_{V|W}\{\exp\{-\frac{1}{2}\|(\beta_k^* - \frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1 V_{i1} + \lambda_2 V_{i2})^2\}V_{i1}^2\} \\
&= \frac{1}{2\pi\lambda_1^3}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\{-\frac{1}{2}\|(\beta_k^* - \frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2 v_1^2\}(v_1 - \lambda_2 v_2)^2\exp(-\frac{1}{2}v_1^2) \\
&\quad \cdot \exp(-\frac{1}{2}v_2^2)dv_1 dv_2 \\
&\le \frac{C_6}{\|(\beta_k^* - \frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^3} + \frac{C_7}{\|(\beta_k^* - \frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2} + \frac{C_8\lambda_2^2}{\|(\beta_k^* - \frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2} \\
&\le \frac{C_6}{\|(\beta_k^* - \frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^3} + \frac{C_7}{\|(\beta_k^* - \frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2} + \frac{C_8 C_b^2 M_2}{(1-c_1)^2\|(\beta_k^* - \frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2}.
\end{aligned}
$$

When $j=2$,

$$
\begin{aligned}
&\mathrm{E}_{V|W}\Big\{\mathrm{E}_Z\Big(\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{\|(\beta_2-\beta_1)^T\Sigma^{1/2}V_{i1}(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}\\
&\quad\cdot\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2|V_{i1}V_{i2}|\Big)\Big\}\\
&=\mathrm{E}_{V|W}\Big\{\mathrm{E}_Z\Big(\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{\|(\beta_2-\beta_1)^T\Sigma^{1/2}V_{i1}(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}\\
&\quad\cdot\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2|V_{i1}V_{i2}|\mid\mathcal{E}_i\Big)\mathbb{P}(\mathcal{E}_i)\Big\}\\
&\quad+\mathrm{E}_{V|W}\Big\{\mathrm{E}_Z\Big(\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{\|(\beta_2-\beta_1)^T\Sigma^{1/2}V_{i1}(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}\\
&\quad\cdot\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2|V_{i1}V_{i2}|\mid\mathcal{E}_i\Big)\mathbb{P}(\mathcal{E}_i)\Big\}\\
&\le\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2\|_2\mathrm{E}_{V|W}\Big\{\exp\{-\frac{1}{2}\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2V_{i1}^2\}|V_{i1}V_{i2}|\Big\}\\
&\quad+\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2\mathrm{E}_{V|W}\Big\{\exp\{-\frac{1}{2}\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1V_{i1}+\lambda_2V_{i2})^2\}|V_{i1}V_{i2}|\Big\}\\
&\le\frac{C_5}{\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2}\\
&\quad+\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2\mathrm{E}_{V|W}\Big\{\exp\{-\frac{1}{2}\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1V_{i1}+\lambda_2V_{i2})^2\}|V_{i1}V_{i2}|\Big\}
\end{aligned}
$$

Again, letting $v_1=\lambda_1V_{i1}+\lambda_2V_{i2}$ and $v_2=V_{i2}$, when $W_i=k$, we have

$$
\begin{aligned}
&\mathrm{E}_V\{\exp\{-\frac{1}{2}\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1V_{i1}+\lambda_2V_{i2})^2\}|V_{i1}V_{i2}|\}\\
&\le\frac{1}{2\pi\lambda_1^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\{-\frac{1}{2}\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2v_1^2\}|v_2(v_1-\lambda_2v_2)|\exp(-\frac{1}{2}v_1^2)\cdot\exp(-\frac{1}{2}v_2^2)dv_1dv_2\\
&\le\frac{C_9}{\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2}+\frac{C_8|\lambda_2|}{\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2}\\
&\le\frac{C_9}{\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2}+\frac{C_{10}C_b\sqrt{M_2}}{(1-c_1)\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\mathrm{E}\Big(\Big|\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{\|(\beta_2-\beta_1)^T\Sigma^{1/2}V_{i1}(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}\\
&\quad\cdot V_i\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2V_{i1}\Big|\Big)\\
&\le\sum_{k=1}^{2}\omega_k^*\Big(\frac{C_6}{\|(\beta_W^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2}+\frac{C_7+C_9}{\|(\beta_W^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2}+\frac{(C_8+C_{10})C_b^2M_2}{(1-c_1)^2}\Big)\\
&\le\sum_{k=1}^{2}\omega_k^*\Big(\frac{C_6}{(1-c_1)^2\Delta^2}+\frac{C_7+C_9}{(1-c_1)\Delta}+\frac{C_8C_b^2M_2}{(1-c_1)^2}+\frac{C_{10}C_b\sqrt{M_2}}{(1-c_1)}\Big)
\end{aligned}
$$

When $\Delta \geq \frac{\sqrt{4C_6}}{(1-c_1)\sqrt{\kappa_0}} \vee \frac{4(C_7+C+9)}{(1-c_1)\kappa_0}$ and $C_b \leq \sqrt{\frac{\kappa_0(1-c_1)^2}{4C_8M_2}} \wedge \frac{\kappa_0(1-c_1)}{4C_{10}\sqrt{M2}}$,

$$\sum_{k=1}^{2}\omega_k^*\mathrm{E}_{Z,V|W}\Big(\Big|\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{\|(\beta_2-\beta_1)^T\Sigma^{1/2}V_{i1}(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2} \cdot V_{ij}\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2V_{i1}\Big|\Big) \leq \kappa_0$$

for $j=1,2$. It follows that $\|\mathrm{E}\frac{\partial\eta_{1,\xi}(X_i,Y_i)}{\partial(\beta_1+\beta_2)}\|_2 \leq \kappa_0$.

Plug the bounds for (i), (ii), and (iii) back into (A.4), we have

$$|\omega_k(\theta)-\omega_k^*| \leq \kappa_0(|\omega_1-\omega_1^*| \vee \|\beta_1^*-\beta_1\|_2 \vee \|\beta_2^*-\beta_2\|_2),$$

for sufficient large Δ and small C_b.

■

Contraction for $\rho_1(\theta)$

We aim to show that $\|\rho_1(\theta)-\rho_1(\theta^*)\|_2 \leq \kappa_0(|\omega_k-\omega_k^*| \vee \|\beta_1-\beta_1^*\|_2 \vee \|\beta_2^*-\beta_2\|_2)$.

Proof Recall that

$$\rho_1(\theta) = \mathrm{E}(\frac{1}{n}\sum_{i=1}^{n}\eta_{1,\theta}(X_i,Y_i)X_iY_i) = \mathrm{E}(\eta_{1,\theta}(X_i,Y_i)X_iY_i),$$

and $\xi=(\omega_1,\beta_2-\beta_1,\beta_1+\beta_2)$. Similar as what we do for $\omega_1(\theta)$, we have

$$\begin{aligned}
&\rho_1(\theta)-\rho_1(\theta^*)\\
&=\mathrm{E}\Big\{\int_0^1\Big(\frac{d\eta_{1,\xi}(X_i,Y_i)X_iY_i}{d\xi}\Big|_{\xi=\xi_u}\Big)^T\Delta_\xi du\Big\}\\
&=\mathrm{E}\Big\{\int_0^1\Big(\frac{\partial\eta_{1,\xi}(X_i,Y_i)X_iY_i}{\partial\omega_1}\Big|_{\xi=\xi_u}\Big)^T\Delta_{\omega_1}du\Big\}+\mathrm{E}\Big\{\int_0^1\Big(\frac{\partial\eta_{1,\xi}(X_i,Y_i)X_iY_i}{\partial(\beta_2-\beta_1)}\Big|_{\xi=\xi_u}\Big)^T\Delta_{\beta_2-\beta_1}du\Big\}\\
&\quad+\mathrm{E}\Big\{\int_0^1\Big(\frac{\partial\eta_{1,\xi}(X_i,Y_i)X_iY_i}{\partial(\beta_2+\beta_1)}\Big|_{\xi=\xi_u}\Big)^T\Delta_{\beta_2+\beta_1}du\Big\}\\
&=\sup_{\xi\in\mathcal{B}_{con}(\theta^*)}\mathrm{E}\Big(\frac{\partial\eta_{1,\xi}(X_i,Y_i)X_iY_i}{\partial\omega_1}\Big|_{\xi=\xi_u}\Big)^T\Delta_{\omega_1}+\sup_{\xi\in\mathcal{B}_{con}(\theta^*)}\mathrm{E}\Big(\frac{\partial\eta_{1,\xi}(X_i,Y_i)X_iY_i}{\partial(\beta_2-\beta_1)}\Big|_{\xi=\xi_u}\Big)^T\Delta_{\beta_2-\beta_1}\\
&\quad+\sup_{\xi\in\mathcal{B}_{con}(\theta^*)}\mathrm{E}\Big(\frac{\partial\eta_{1,\xi}(X_i,Y_i)X_iY_i}{\partial(\beta_2+\beta_1)}\Big|_{\xi=\xi_u}\Big)^T\Delta_{\beta_2+\beta_1}
\end{aligned}$$

It follows that

$$\begin{aligned}
\|\rho_1(\theta)-\rho_1^*\|_2 \leq &\sup_{\xi\in\mathcal{B}_{con}(\theta^*)}\|\mathrm{E}(\frac{\partial\eta_{1,\theta}(X_i,Y_i)}{\partial\omega_1}X_iY_i)\|_2|\omega_1-\omega_1^*|\\
&+\sup_{\xi\in\mathcal{B}_{con}(\theta^*)}\|\mathrm{E}(\frac{\partial\eta_{1,\theta}(X_i,Y_i)}{\partial(\beta_2-\beta_1)}X_i^T\cdot Y_i)\|_2\|\beta_1-\beta_1^*-\beta_2+\beta_2^*\|_2\\
&+\sup_{\xi\in\mathcal{B}_{con}(\theta^*)}\|\mathrm{E}(\frac{\partial\eta_{1,\theta}(X_i,Y_i)}{\partial(\beta_1+\beta_2)}X_i^T\cdot Y_i)\|_2\|\beta_1-\beta_1^*+\beta_2-\beta_2^*\|_2.
\end{aligned}$$

Note that

$$\begin{aligned}
&\|\mathrm{E}(\frac{\partial \eta_{1,\theta}(X_i,Y_i)}{\partial \omega_1}X_iY_i)\|_2 \\
&\le \underbrace{\|\mathrm{E}\frac{\exp\{(\beta_2-\beta_1)^TX_i(Z_i+\delta_i(\beta)^TX_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^TX_i(Z_i+\delta_i(\beta)^TX_i)\}\big)^2}\cdot X_i(Z_i+\delta_i(\beta)^TX_i)\|_2}_{(iv)} \\
&+ \underbrace{\|\mathrm{E}\frac{\exp\{(\beta_2-\beta_1)^TX_i(Z_i+\delta_i(\beta)^TX_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^TX_i(Z_i+\delta_i(\beta)^TX_i)\}\big)^2}\cdot X_iX_i^T\frac{\beta_1+\beta_2}{2}\|_2}_{(v)}
\end{aligned} \tag{A.7}$$

The bound for term (iv) is exactly with that for term (ii). For sufficient large Δ and small C_b, $(iv) \le \kappa_0/2$.

Then note that

$$\begin{aligned}
(v) &= \|\Sigma^{1/2}H_k^T\mathrm{E}\frac{\exp\{\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2V_{i1}(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^TX_i(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}V_iV_i^TH_k\Sigma^{1/2}\frac{\beta_1+\beta_2}{2}\|_2 \\
&\le \|\Sigma\|_2\|\frac{\beta_1+\beta_2}{2}\|_2\|\underbrace{\mathrm{E}\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^TX_i(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}V_iV_i^T}_{(I)}\|_2.
\end{aligned}$$

Because V_{ij} are independent with each other for all j and $\omega_1^*\mu_1+\omega_2^*\mu_2=0$, we know that except from the diagonal elements and the first 2×2 sub-matrix, matrix (I) is exactly zero. Thus, to bound the 2-norm of (I), we only need to considering the first 2×2 sub-matrix and the diagonal elements. Note that, for any j,q,

$$\begin{aligned}
&\mathrm{E}|\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}V_{ij}V_{iq}^T| \\
&\le \sum_{k=1}^{2}\big(\omega_k^*\mathrm{E}_{Z,V|W}\Big|\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\mathrm{expexp}\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}V_{ij}V_{iq}^T\mid\mathcal{E}_i\Big|\mathbb{P}(\mathcal{E}_i)\big) \\
&\quad+\sum_{k=1}^{2}\big(\omega_k^*\mathrm{E}_{Z,V|W}\Big|\frac{\exp\{T_1(V_{i1})(Z_i+\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})\|_2(\lambda_1V_{i1}+\lambda_2V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}V_{ij}V_{iq}^T\mid\mathcal{E}_i^c\Big|\mathbb{P}(\mathcal{E}_i^c)\big) \\
&\le \sum_{k=1}^{2}\big(\omega_k^*\mathrm{E}_{V|W}\big\{\exp(-1/2\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2V_{i1}^2) \\
&\quad+\exp(-1/2\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})\|_2^2(\lambda_1V_{i1}+\lambda_2V_{i2})^2)\big\}|V_{ij}V_{iq}|\big) \\
&\quad+\sum_{k=1}^{2}\big(\omega_k^*\mathrm{E}_{V|W}\big\{\exp(-1/8\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})\|_2^2(\lambda_1V_{i1}+\lambda_2V_{i2})^2)\big\}|V_{ij}V_{iq}|\big).
\end{aligned}$$

By Lemma A.8 and the same calculation procedure used before, we have

$$\mathrm{E}|\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^TX_i(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}V_{ij}V_{iq}^T|\le C_{11}\sum_{k=1}^{2}(\omega_k^*/\Delta+\omega_k^*/\Delta^2).$$

Thus, for large enough Δ, we have $(v) \le \kappa_0/2$. It follows that $\|\mathrm{E}(\frac{\partial \eta_{1,\theta}(X_i,Y_i)}{\partial \omega_1} X_i Y_i)\|_2 \le \kappa_0$ for sufficient large Δ and small C_b.

Next, recall that

$$\mathrm{E}(\frac{\partial \eta_{1,\theta}(X_i,Y_i)}{\partial \beta_2-\beta_1} X_i^T \cdot Y_i) = \omega_1(1-\omega_1)\mathrm{E}\Big(\frac{\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}\big)^2} \cdot X_i(Z_i+\delta(\beta)_i^T X_i)X_i^T(Z_i+\psi_i^T X_i)\Big)$$

Let H_k be an orthogonal matrix whose first row is $(\beta_2-\beta_1)^T\Sigma^{1/2}/\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2$ and satisfies that $(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2} \subseteq \mathrm{span}\{H_{k,1:}^T, H_{k,2:}^T\}$ and $(\beta_1+\beta_2)^T\Sigma^{1/2} \subseteq \mathrm{span}\{H_{k,1:}^T, H_{k,2:}^T, H_{k,3:}^T\}$, where $H_{k,j:}$ represents the j-th row of H_k. Let $X_i \mid W_i=k = \Sigma^{1/2}H_k^T V_i$, where $V_i \mid W_i = k \sim N(\Sigma^{1/2}H_k^T\mu_k, I_p)$. Then given $W_i=k$, we have $(\beta_2-\beta_1)^T X_i = \|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2 V_{i1} = T_1(V_{i1})$, $(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T X_i = \|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2(\lambda_1 V_{i1}+\lambda_2 V_{i2}) = T_2(V_{i1},V_{i2})$, and $(\beta_1+\beta_2)^T X_i = \|(\beta_1+\beta_2)^T\Sigma^{1/2}\|_2(\lambda_3 V_{i1}+\lambda_4 V_{i2}+\lambda_5 V_{i3}) = T_3(V_{i1},V_{i2},V_{i3})$, where $\lambda_1^2+\lambda_2^2=1$, $\lambda_3^2+\lambda_4^2+\lambda_5^2=1$ and

$$|\lambda_2| \le \frac{\|(\beta_k^*-\beta_k)^T\Sigma^{1/2}\|_2}{\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2} \le \frac{C_b\sqrt{M_2}}{1-c_1}.$$

Then we decompose $\mathrm{E}(\frac{\partial \eta_{1,\theta}(X_i,Y_i)}{\partial \beta_2-\beta_1} X_i^T \cdot Y_i)$ as

$$\begin{aligned}
&\mathrm{E}(\frac{\partial \eta_{1,\theta}(X_i,Y_i)}{\partial \beta_2-\beta_1} X_i^T \cdot Y_i)\\
&= \omega_1(1-\omega_1)\mathrm{E}(\frac{\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}\big)^2}\\
&\quad \cdot X_i(Z_i+\delta_i(\beta)^T X_i)X_i^T(Z_i+\delta_i(\beta)^T X_i))\\
&\quad + (\omega_1(1-\omega_1)/2)\mathrm{E}(\frac{\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^T X_i(Z_i+\delta_i(\beta)^T X_i)\}\big)^2}\\
&\quad \cdot X_i(Z_i+\delta_i(\beta)^T X_i)X_i^T(\beta_1+\beta_2)^T X_i)) \qquad \text{(A.8)}\\
&= \omega_1(1-\omega_1)\sum_{k=1}^2 \omega_k^* \mathrm{E}_{V|W_i=k}\Big\{\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}\\
&\quad \cdot \Sigma^{1/2}H_k^T V_i V_i^T(Z_i+T_2(V_{i1},V_{i2}))^2 H_k\Sigma^{1/2}\Big\}\\
&+ \omega_1(1-\omega_1)\sum_{k=1}^2 \omega_k^* \mathrm{E}_{V|W_i=k}\Big\{\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}\\
&\quad \cdot \Sigma^{1/2}H_k^T V_i V_i^T(Z_i+T_2(V_{i1},V_{i2}))(T_3(V_{i1},V_{i2},V_{i3}))H_k\Sigma^{1/2}\Big\}
\end{aligned}$$

Firstly, we bound the first term of (A.8). Because $\omega_1^*\mu_1+\omega_2^*\mu_2=0$, and V_{ij} are independent for all j, for j,l not in the first 2×2 block and the diagonal elements, we have

$$\sum_{k=1}^2 \omega_k^* \mathrm{E}_{V|W_i=k}\Big\{\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2} V_{ij}V_{il}^T(Z_i+T_2(V_{i1},V_{i2}))^2\Big\} = 0.$$

For j, l in the first 2×2 block and the diagonal elements, we have

$$
\begin{aligned}
&\mathrm{E}_{V|W_i=k}\Big\{\Big|\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}V_{ij}V_{il}(Z_i+T_2(V_{i1},V_{i2}))^2\Big|\Big\}\\
&\le \mathrm{E}_{V|W_i=k}\Big\{\Big|\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}V_{ij}V_{il}(Z_i+T_2(V_{i1},V_{i2}))^2\Big| \mid \mathcal{E}_i\Big\}\mathbb{P}(\mathcal{E}_i)\\
&\quad+\mathrm{E}_{V|W_i=k}\Big\{\Big|\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}V_{ij}V_{il}(Z_i+T_2(V_{i1},V_{i2}))^2\Big| \mid \mathcal{E}_i^c\Big\}\mathbb{P}(\mathcal{E}_i^c)\\
&\le \frac{3}{2c_0^2}\Big(\exp\{-\frac{1}{2}\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2V_{i1}^2\}|V_{ij}V_{il}|\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1V_{i1}+\lambda_2V_{i2})^2\Big)\\
&\quad+\frac{3}{2c_0^2}\Big(\exp\{-\frac{1}{2}\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1V_{i1}+\lambda_2V_{i2})^2\}|V_{ij}V_{il}|\\
&\qquad\cdot\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1V_{i1}+\lambda_2V_{i2})^2\Big)\\
&\quad+\frac{1}{2c_0^2\sqrt{\pi}}\int_{\frac{1}{2}\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1V_{i1}+\lambda_2V_{i2})^2}^{\infty}V_{ij}V_{il}(Z_i+T_2(V_{i1},V_{i2}))^2\exp(-Z_i^2/2)dZ_i\\
&\le \frac{3}{2c_0^2}\underbrace{\Big(\exp\{-\frac{1}{2}\|(\beta_2-\beta_1)^T\Sigma^{1/2}\|_2^2V_{i1}^2\}|V_{ij}V_{il}|\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1V_{i1}+\lambda_2V_{i2})^2\Big)}_{(I)}\\
&\quad+\underbrace{\frac{3}{2c_0^2}\Big(\exp\{-\frac{1}{2}\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1V_{i1}+\lambda_2V_{i2})^2\}|V_{ij}V_{il}|}_{(II)}\\
&\qquad\cdot\underbrace{\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1V_{i1}+\lambda_2V_{i2})^2\Big)}_{(II)}\\
&\quad+\underbrace{\frac{1}{2c_0^2\sqrt{\pi}}\exp\{-\frac{1}{2}\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2^2(\lambda_1V_{i1}+\lambda_2V_{i2})^2\}|V_{ij}V_{il}|}_{(III)}\\
&\qquad\cdot\underbrace{(2\|(\beta_k^*-\frac{\beta_1+\beta_2}{2})^T\Sigma^{1/2}\|_2|(\lambda_1V_{i1}+\lambda_2V_{i2})|+1)}_{(III)},
\end{aligned}
$$

where the last inequality is obtained by standard integration and the fact that $\int_t^\infty \exp(-x^2/2)dx \le \exp(-t^2/2)/t$.

Let $v_1 = \lambda_1 V_{i1} + \lambda_2 V_{i2}$ and $v_2 = V_{i2}$. By Lemma A.8, we have terms (I), (II) and (III) are all smaller than κ_0, for sufficient large Δ.

Then we bound the second term of (A.8). Because $\omega_1^*\mu_1 + \omega_2^*\mu_2 = 0$, and V_{ij} are independent for all j, for j, l not in the first 3×3 block and the diagonal elements, we have

$$
\begin{aligned}
&\sum_{k=1}^{2}\omega_k^*\mathrm{E}_{V|W_i=k}\Big\{\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}\\
&\quad V_{ij}V_{il}(T_2(V_{i1},V_{i2}))(Z_i+T_3(V_{i1},V_{i2},V_{i3}))\Big\}=0
\end{aligned}
$$

For j,l in the first 3×3 block and the diagonal elements, using the same techniques used before, for sufficient large Δ and small C_b, we have

$$\sum_{k=1}^{2}\omega_k^*\mathrm{E}_{V|W_i=k}\Big\{\Big|\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}\\ V_{ij}V_{il}(T_2(V_{i1},V_{i2}))(Z_i+T_3(V_{i1},V_{i2},V_{i3}))\Big|\Big\}\le\kappa_0$$

Thus

$$\|\mathrm{E}(\frac{\partial\eta_{1,\theta}(X_i,Y_i)}{\partial\beta_2-\beta_1}X_i^T\cdot Y_i)\|_2\le\kappa_0,$$

for sufficient large Δ and small C_b.

Finally, recall that

$$\begin{aligned}&\frac{1}{\omega_1(1-\omega_1)}\mathrm{E}(\frac{\partial\eta_{1,\theta}(X_i,Y_i)}{\partial\beta_2+\beta_1}X_i^T\cdot Y_i)\\ &=\mathrm{E}(\frac{\exp\{(\beta_2-\beta_1)^TX_i(Z_i+\delta_i(\beta)^TX_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^TX_i(Z_i+\delta_i(\beta)^TX_i)\}\big)^2}\cdot X_iX_i^T(\beta_2-\beta_1)X_i^T(Z_i+\psi_i^TX_i))\\ &=\underbrace{\mathrm{E}(\frac{\exp\{(\beta_2-\beta_1)^TX_i(Z_i+\delta_i(\beta)^TX_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^TX_i(Z_i+\delta_i(\beta)^TX_i)\}\big)^2}\cdot X_iX_i^T(\beta_2-\beta_1)X_i^T(Z_i+\delta_i(\beta)^TX_i))}_{(a1)}\\ &\quad+\underbrace{\mathrm{E}(\frac{\exp\{(\beta_2-\beta_1)^TX_i(Z_i+\delta_i(\beta)^TX_i)\}}{\big(\omega_1+(1-\omega_1)\exp\{(\beta_2-\beta_1)^TX_i(Z_i+\delta_i(\beta)^TX_i)\}\big)^2}\cdot X_iX_i^T(\beta_2-\beta_1)X_i^T(\frac{\beta_1^T+\beta_2^T}{2}X_i))}_{(a2)}\end{aligned}$$

For term $(a1)$, we have

$$(a1)=\sum_{k=1}^{n}\Big(\omega_k^*\mathrm{E}_{V|W_i=k}\big\{\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}\\ \cdot\Sigma^{1/2}H_k^TV_iV_i^TH_k\Sigma^{1/2}T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\Big)$$

Because $\omega_1^*\mu_1+\omega_2^*\mu_2=0$, and V_{ij} are independent for all j, for j,l not in the first 2×2 block and the diagonal elements, we have $(a1)=0$. For j,l not in the first 2×2 block and the diagonal elements, we have

$$\sum_{k=1}^{n}\Big(\omega_k^*\mathrm{E}_{V|W_i=k}\Big\{\Big|\frac{\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\}\big)^2}\\ \cdot V_{ij}V_{il}T_1(V_{i1})(Z_i+T_2(V_{i1},V_{i2}))\Big|\Big\}\Big)\le\kappa_0$$

for sufficient large Δ and small C_b. The bounds for terms $(a1)$ and $(a2)$ implies that

$$\|\mathrm{E}(\frac{\partial\eta_{1,\theta}(X_i,Y_i)}{\partial\beta_2+\beta_1}X_i^T\cdot Y_i)\|_2\le\kappa_0,$$

for sufficient large Δ and small C_b.

To sum up, we have

$$\|\rho_1(\theta) - \rho_1(\theta^*)\|_2 \leq \kappa_0(|\omega_1 - \omega_1^*| \vee \|\beta_1 - \beta_1^*\|_2 \vee \|\beta_2 - \beta_2^*\|_2).$$

■

Contraction for $\Sigma_1(\theta)$

We aim to show that $\|(\Sigma_1(\theta) - \Sigma_1(\theta^*))\beta_1^*\|_2 \leq \kappa_0(|\omega_1 - \omega_1^*| \vee \|\beta_1 - \beta_1^*\|_2 \vee \|\beta_2 - \beta_2^*\|_2)$.

Recall that

$$\Sigma_1(\theta)\beta_1^* = \mathrm{E}(\frac{1}{n}\sum_{i=1}^{n}\eta_{1,\theta}(X_i, Y_i)X_iX_i^T\beta_1^*) = \mathrm{E}(\eta_{1,\theta}(X_i, Y_i)X_iX_i^T\beta_1^*).$$

Similar as what we do for $\rho_1(\theta)$, we have

$$\begin{aligned}
&\Sigma_1(\theta)\beta_1^* - \Sigma_1(\theta^*)\beta_1^* \\
&= \mathrm{E}\big\{\int_0^1 \Big(\frac{d\eta_{1,\xi}(X_i, Y_i)X_iX_i^T\beta_1^*}{d\xi}\Big|_{\xi=\xi_u}\Big)^T\Delta_\xi du\big\} \\
&= \mathrm{E}\big\{\int_0^1 \Big(\frac{\partial\eta_{1,\xi}(X_i, Y_i)X_iX_i^T\beta_1^*}{\partial\omega_1}\Big|_{\xi=\xi_u}\Big)^T\Delta_{\omega_1} du\big\} + \mathrm{E}\big\{\int_0^1 \Big(\frac{\partial\eta_{1,\xi}(X_i, Y_i)X_iX_i^T\beta_1^*}{\partial(\beta_2-\beta_1)}\Big|_{\xi=\xi_u}\Big)^T\Delta_{\beta_2-\beta_1} du\big\} \\
&\quad + \mathrm{E}\big\{\int_0^1 \Big(\frac{\partial\eta_{1,\xi}(X_i, Y_i)X_iX_i^T\beta_1^*}{\partial(\beta_2+\beta_1)}\Big|_{\xi=\xi_u}\Big)^T\Delta_{\beta_2+\beta_1} du\big\} \\
&= \mathrm{E}\Big(\frac{\partial\eta_{1,\xi}(X_i, Y_i)X_iX_i^T\beta_1^*}{\partial\omega_1}\Big|_{\xi=\xi_u}\Big)^T du\Delta_{\omega_1} + \int_0^1 \mathrm{E}\Big(\frac{\partial\eta_{1,\xi}(X_i, Y_i)X_iX_i^T\beta_1^*}{\partial(\beta_2-\beta_1)}\Big|_{\xi=\xi_u}\Big)^T du\Delta_{\beta_2-\beta_1} \\
&\quad + \int_0^1 \mathrm{E}\Big(\frac{\partial\eta_{1,\xi}(X_i, Y_i)X_iX_i^T\beta_1^*}{\partial(\beta_2+\beta_1)}\Big|_{\xi=\xi_u}\Big)^T du\Delta_{\beta_2+\beta_1}
\end{aligned}$$

It follows that

$$\begin{aligned}
\|\Sigma_1(\theta)\beta_1^* - \Sigma_1(\theta^*)\beta_1^*\|_2 \leq &\sup_{\xi\in\mathcal{B}_{con}(\theta^*)} \|\mathrm{E}(\frac{\partial\eta_{1,\theta}(X_i, Y_i)}{\partial\omega_1}X_iX_i^T\beta_1^*)\|_2|\omega_1 - \omega_1^*| \\
&+ \sup_{\xi\in\mathcal{B}_{con}(\theta^*)} \|\mathrm{E}(\frac{\partial\eta_{1,\theta}(X_i, Y_i)}{\partial(\beta_2-\beta_1)}X_i^T(X_i^T\beta_1^*))\|_{2,s}\|\beta_1 - \beta_1^* - \beta_2 + \beta_2^*\|_2 \\
&+ \sup_{\xi\in\mathcal{B}_{con}(\theta^*)} \|\mathrm{E}(\frac{\partial\eta_{1,\theta}(X_i, Y_i)}{\partial(\beta_1+\beta_2)}X_i^T(X_i^T\beta_1^*))\|_{2,s}\|\beta_1 - \beta_1^* + \beta_2 - \beta_2^*\|_2.
\end{aligned}$$

Bounding these three terms is a simplified case of that for the contraction of $\rho_1(\theta)$. We omit the details here.

Appendix E. Proof of Lemma 3

We first prove that $\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\{\|\widehat{\rho}_1(\theta) - \rho_1(\theta)\|_{2,s}\} = O(\sqrt{\frac{s\log(n)^2\log(p)}{n}})$ with probability at least $1 - p^{-1}$.

For $\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\{\|\widehat{\rho}_1(\theta)-\rho_1(\theta)\|_{2,s}\}$, directly applying Lemma A.6 will convert the problem into bounding the product of 4 sub-Gaussian variables. To avoid this problem and get a sharper bound, we use the tail bound for unbounded random process given in Lemma A.7.

Proof Recall that

$$\widehat{\rho}_1(\theta)-\rho_1(\theta)=\frac{1}{n}\sum_{i=1}^{n}\eta_{1,\theta}(X_i,Y_i)X_iY_i-\mathrm{E}(\eta_{1,\theta}(X_i,Y_i)X_iY_i).$$

Let $\widetilde{W}^{\rho}=\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\|\widehat{\rho}_1(\theta)-\rho_1(\theta)\|_{2,s}$. By definition, we have

$$\widetilde{W}^{\rho}=\sup_{\mu\in\Gamma(s)\cap\mathcal{S}^{p-1}}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\langle\frac{1}{n}\sum_{i=1}^{n}\eta_{1,\theta}(X_i,Y_i)X_iY_i-\mathrm{E}(\eta_{1,\theta}(Y_i,X_i)X_iY_i),\mu\rangle$$

Let

$$W_{\mu}^{\rho}=\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\langle\frac{1}{n}\sum_{i=1}^{n}\eta_{1,\theta}(X_i,Y_i)X_iY_i-\mathrm{E}(\eta_{1,\theta}(Y_i,X_i)X_iY_i),\mu\rangle.$$

Because $\Gamma(s)\cap\mathcal{S}^{p-1}\subseteq\mathcal{C}(s)\cap\mathcal{S}^{p-1}$, we will bound

$$W^{\rho}=\sup_{\mu\in\mathcal{C}(s)\cap\mathcal{S}^{p-1}}W_{\mu}^{\rho}=\sup_{\mu\in\mathcal{C}(s)\cap\mathcal{S}^{p-1}}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\langle\frac{1}{n}\sum_{i=1}^{n}\eta_{1,\theta}(X_i,Y_i)X_iY_i-\mathrm{E}(\eta_{1,\theta}(Y_i,X_i)X_iY_i),\mu\rangle,$$

instead. Let $\nu_1,\cdots,\nu_{M_{net}}$ denote a 1/2-net of $\mathcal{C}(s)\cap\mathcal{S}^{p-1}$, we have

$$W_{\nu}^{\rho}\le W_{\mu_j}^{\rho}+|W_{\mu_j}^{\rho}-W_{\nu}^{\rho}|\le\max_{j\in[M_{net}]}W_{\mu_j}^{\rho}+W^{\rho}\|\nu-\mu_j\|_2\le\max_{j\in[M_{net}]}W_{\mu_j}^{\rho}+1/2W^{\rho}.$$

Thus $W^{\rho}\le 2\max_{j\in[M_{net}]}W_{\mu_j}^{\rho}$. So, instead of directly bounding the tail of W^{ρ}, we can first get the tail probability for $W_{\mu_j}^{\rho}$ for a fixed j, then using the union bound to get the tail probability for W^{ρ}.

Note that

$$\begin{aligned}\|\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\eta_{1,\theta}(X_i,Y_i)\mu_j^TX_iY_i-\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}}(\eta_{1,\theta}(X_i,Y_i)\mu_j^TX_iY_i)\|_{\psi_1}&\le c\|\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\eta_{1,\theta}(X_i,Y_i)\mu_j^TX_iY_i\|_{\psi_1}\\&\le c\|\mu_j^TX_iY_i\|_{\psi_1}<\infty,\end{aligned}$$

where we use the fact that $0<\eta_{1,\theta}(X_i,Y_i)<1$ and $\mu_j^TX_iY_i$ is sub-exponential.

We use Lemma A.7 to get the tail probability for $W_{\mu_j}^{\rho}$. We first bound $\mathrm{E}(W_{\mu_j}^{\rho})$ using similar procedure used in Adamczak (2008). Let $f(X_i,Y_i)=\eta_{1,\theta}(X_i,Y_i)\mu_j^TX_iY_i-\mathrm{E}(\eta_{1,\theta}(X_i,Y_i)\mu_j^TX_iY_i)$ for simplicity. We have $W_{\mu_j}^{\rho}\le\frac{1}{n}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}|\sum_{i=1}^{n}f(X_i,Y_i)|$.

Define truncated function and the remaining part of $f(X_i,Y_i)$ as

$$\begin{aligned}f_1&=f(X_i,Y_i)I(|\mu_j^TX_iY_i|\le\rho)-\mathrm{E}[f(X_i,Y_i)I(|\mu_j^TX_iY_i|\le\rho)],\\f_2&=f(X_i,Y_i)I(|\mu_j^TX_iY_i|>\rho)-\mathrm{E}[f(X_i,Y_i)I(|\mu_j^TX_iY_i|>\rho)],\end{aligned}$$

where $\rho = 8\mathrm{E}\max_i \sup_{\theta\in\mathcal{B}_{con}(\theta^*),\mu\in\Gamma(s)\cap\mathcal{S}^{p-1}} |f(X_i,Y_i)|$.

Let $Q = \max_i |\mu_j^T X_i Y_i|$. We have

$$\mathrm{E}Q = \int_0^\infty \mathbb{P}(Q > x)dx. \tag{A.9}$$

Note that $\|\mu_j^T X_i Y_i\|_{\psi_1} \le C$ we have $\mathbb{P}(|\mu^T X_i Y_i| > x\log n) \le \exp(-cx\log n)$. By union bound we have

$$\mathbb{P}(Q > x\log n) \le \sum_{i=1}^n \exp(-cx\log n) = \exp(-xc\log n + \log n).$$

When $\log n > 2$ and $x > 2/c$, we have $\mathbb{P}(Q > x\log n) \le \exp(-cx)$. By (A.9) and $\mathbb{P}(Q > x) < 1$, we have $\mathrm{E}Q \le c\log n$. It follows that $\rho \le C\log(n)$.

Note that

$$\begin{aligned}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\sum_{i=1}^n f(X_i,Y_i)| \le & \sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\sum_{i=1}^n f_1(X_i,Y_i) - \mathrm{E}f_1(X_i,Y_i)| \\ & + \sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\sum_{i=1}^n f_2(X_i,Y_i) - \mathrm{E}f_2(X_i,Y_i)|,\end{aligned}$$

where we use the fact that $\mathrm{E}(f_1(X_i,Y_i) + f_2(X_i,Y_i)) = 0$. It follows that

$$\begin{aligned}\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\sum_{i=1}^n f(X_i,Y_i)| \le & \mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\sum_{i=1}^n f_1(X_i,Y_i) - \mathrm{E}f_1(X_i,Y_i)| \\ & + 2\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\sum_{i=1}^n f_2(X_i,Y_i)|.\end{aligned} \tag{A.10}$$

By Markov inequality and the definition of $f_2(X_i,Y_i)$, we have

$$\mathbb{P}(\max_{k\le n} \sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\sum_{i=1}^k f_2(X_i,Y_i)| > 0) \le \mathbb{P}(\max_i |\mu_j^T X_i Y_i| > \rho) \le 1/8.$$

Then by Hoffmann-Jørgensen inequality (see e.g Ledoux and Talagrand (1991), Proposition 6.8)

$$\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\sum_{i=1}^n f_2(X_i,Y_i)| \le 8\mathrm{E}\max_i \sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |f(X_i,Y_i)| \le \rho. \tag{A.11}$$

Thus, we have

$$\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\frac{1}{n}\sum_{i=1}^n f_2(X_i,Y_i)| \le \frac{C\log n}{n}. \tag{A.12}$$

Next we go back to bound $\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\frac{1}{n}\big(\sum_{i=1}^n f_1(X_i,Y_i) - \mathrm{E}f_1(X_i,Y_i)\big)|$. By Lemme A.5 and Lemma A.6, we have

$$\begin{aligned}&\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\frac{1}{n}\big(\sum_{i=1}^n f_1(X_i,Y_i) - \mathrm{E}f_1(X_i,Y_i)\big)| \\ \le & C\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\frac{1}{n}\sum_{i=1}^n \epsilon_i(\beta_2-\beta_1)^T X_i(Y_i - \frac{(\beta_1+\beta_2)^T}{2}X_i)\mu_j^T X_i Y_i|,\end{aligned}$$

where $|\mu_j^T X_i Y_i| \leq \rho$ for all i. Under the condition $|\mu_j^T X_i Y_i| \leq \rho$, $\epsilon_i(\beta_2-\beta_1)^T X_i(Y_i - \frac{(\beta_1+\beta_2)^T}{2}X_i)\mu_j^T X_i Y_i$ is sub-exponential for any β_1 and β_2. Define set $\mathcal{T} = \{\nu_1 : \nu_1^T = (\beta_1^T, \beta_2^T) \in \mathcal{B}_{con}\}$. We have $D = \text{diam}(\mathcal{T}) \leq cC_b\Delta$.

By Corollary 5.2 of Dirksen (2015), we have

$$\begin{aligned}&\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}|\frac{1}{n}\sum_{i=1}^n \epsilon_i(\beta_2-\beta_1)^T X_i(Y_i - \frac{(\beta_1+\beta_2)^T}{2}X_i)\mu_j^T X_i Y_i| \\ &\leq C_1\big(\frac{1}{\sqrt{n}}\gamma_2(\mathcal{T},d)+\frac{1}{n}\gamma_1(\mathcal{T},d)\big)+C_2(\frac{1}{n}+\frac{1}{\sqrt{n}}),\end{aligned} \tag{A.13}$$

where $\gamma_1(\mathcal{T},d)$ and $\gamma_2(\mathcal{T},d)$ are Talagrand γ_1 and γ_2 functional (See Dirksen (2015) for details), and d is the ℓ_2-norm.

By Lemma A.3, $\mathcal{T} \in C\text{conv}(\cup_{|J|\leq 2c_d s}E_J(2p)\cap B_2^{2p})$. From the volumetric argument in Rudelson and Zhou (2012)[Section H.1], we know that the covering number $\mathcal{N}(\mathcal{T},d,\epsilon)$ satisfies that

$$\log(\mathcal{N}(\mathcal{T},d,\epsilon)) \leq C_4\big(s\log(ep/s)+s\log(1+2/\epsilon)\big). \tag{A.14}$$

Then note that

$$\gamma_\alpha(\mathcal{T},d) \leq C\rho\int_0^D \big(\log\mathcal{N}(T,d,\epsilon)\big)^{1/\alpha}d\epsilon,$$

we have

$$\gamma_2(\mathcal{T},d) = C\rho\int_0^D\sqrt{\log(\mathcal{N}(\mathcal{T},d,\epsilon))}d\epsilon = C\rho\int_0^D\sqrt{s}\Big(\log(\frac{ep}{s})+\log(1+2/\epsilon)\Big)^{1/2}d\epsilon \leq C_1\rho\sqrt{s\log p},$$

and

$$\gamma_1(\mathcal{T},d) = C\rho\int_0^D\log(\mathcal{N}(\mathcal{T},d,\epsilon))d\epsilon = C\rho\int_0^D s\Big(\log(\frac{ep}{s})+\log(1+2/\epsilon)\Big)d\epsilon \leq C_1\rho s\log p.$$

By (A.13), we know that

$$\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}|\frac{1}{n}\sum_{i=1}^n \epsilon_i(\beta_2-\beta_1)^T X_i(Y_i - \frac{(\beta_1+\beta_2)^T}{2}X_i)\mu_j^T X_i Y_i| \leq C\log n\sqrt{\frac{s\log p}{n}}.$$

Combine this result and (A.12), we have

$$\mathrm{E}(W_{\mu_j^\rho}) \leq \mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}|\frac{1}{n}\sum_{i=1}^n\frac{1}{n}f(X_i,Y_i)| \leq C\Big(\sqrt{\frac{s(\log n)^2\log p}{n}}+\frac{\log n}{n}\Big).$$

Note that $\|\max_i \sup_{\theta\in\mathcal{B}_{con}(\theta^*)} f(X_i,Y_i)\|_{\psi_1} \leq \|\max_i|\mu_j^T X_i Y_i|\|_{\psi_1}$ and $\|\max_i|\mu_j^T X_i Y_i|\|_{\psi_1} \leq C\log n\|\mu_j^T X_i Y_i\|_{\psi_1}$ (Pisier's inequality (Pisier, 1983)), we have $\|\max_i \sup_{\theta\in\mathcal{B}_{con}(\theta^*)} f(X_i,Y_i)\|_{\psi_1} \leq C\log n$. Then by Lemma A.7, we have

$$\mathbb{P}(W_{\mu_j^\rho} \geq C\big(\sqrt{\frac{s(\log n)^2\log p}{n}}+\frac{\log n}{n}\big)+t) \leq 4\max\big\{\exp(-C_5 nt^2), 3\exp(-\frac{C_6 nt}{\log n})\big\}$$

By union bound, we have

$$\mathbb{P}(W^\rho \geq C\big(\sqrt{\frac{s(\log n)^2\log p}{n}}+\frac{\log n}{n}\big)+t) \leq M_{net}\mathbb{P}(W_{\mu_j^\rho}) \\ \leq 4\max\big\{\exp(c_d s\log p - C_5 nt^2), 3\exp(c_d s\log p - \frac{C_6 nt}{\log n})\big\}.$$

Let $t = c\sqrt{\frac{s(\log n)^2\log p}{n}}$ for large enough generic constant c, when $n \gg s\log p$,

$$W^\rho = O(\sqrt{\frac{s(\log n)^2\log p}{n}})$$

with probability at least $1-p^{-1}$.

■

We then prove that $\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\{\|(\widehat{\Sigma}_1(\theta)-\Sigma_1(\theta))\beta_1^*\|_{2,s}\} = O(\sqrt{\frac{s\log(n)^2\log(p)}{n}})$ with probability at least $1-p^{-1}$.

Proof Recall that

$$(\widehat{\Sigma}_1(\theta)-\Sigma_1(\theta))\beta_1^* = \frac{1}{n}\sum_{i=1}^n(\eta_{1,\theta}(X_i,Y_i)X_iX_i^T\beta_1^* - E(\eta_{1,\theta}(X_i,Y_i)X_iX_i^T\beta_1^*)$$

Similar to the bound for $\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\{\|\widehat{\rho}_1(\theta)-\rho_1(\theta)\|_{2,s}\}$, we define $W^\Sigma = \sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\{\|(\widehat{\Sigma}_1(\theta)-\Sigma_1(\theta))\beta_1^*\|_{2,s}\}$, which can be written as

$$W^\Sigma = \sup_{\theta\in\mathcal{B}_{con}(\theta^*),\mu\in\Gamma(S)\cap\mathcal{S}^{p-1}}\langle\frac{1}{n}\sum_{i=1}^n(\eta_{1,\theta}(X_i,Y_i)-E(\eta_{1,\theta}(X_i,Y_i))X_iX_i^T\beta_1^*,\mu\rangle,$$

By the same proof procedure for $\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\|\widehat{\rho}_1(\theta)-\rho_1(\theta)\|_{2,s}$, we have

$$\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\{\|(\widehat{\Sigma}_1(\theta)-\Sigma_1(\theta))\beta_1^*\|_{2,s}\} = O(\sqrt{\frac{s\log(n)^2\log(p)}{n}})$$

with probability at least $1-p^{-1}$.

■

Finally, we prove that $\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\{|\widehat{\omega}_1(\theta)-\omega_1(\theta)|\} = O(\sqrt{\frac{s\log(p)}{n}})$ with probability at least $1-2p^{-1}$.

Proof Recall that

$$\widehat{\omega}_1(\theta) = \frac{1}{n}\sum_{i=1}^n\frac{\omega_1(\theta)}{\omega_1(\theta)+\omega_2(\theta)\exp\{(\beta_2-\beta_1)^TX_i(Y_i-\frac{(\beta_1+\beta_2)^TX_i}{2})\}}.$$

Let $C_\theta(X_i,Y_i) = (\beta_2-\beta_1)^TX_i(Y_i-\frac{(\beta_1+\beta_2)^TX_i}{2})$, and $\widehat{z}_\omega = \sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\{|\widehat{\omega}_1(\theta)-\omega_1(\theta)|\}$. We want to bound $\mathrm{E}(e^{\lambda\widehat{z}_\omega})$. Note that $\widehat{z}_\omega \leq \max\{\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}(\widehat{\omega}_1(\theta)-\omega_1(\theta)), \sup_{\theta\in\mathcal{B}_{con}(\theta^*)}((\omega_1(\theta)-$

$\widehat{\omega}_1(\theta_1))\}$, thus we first show the bound for $\widetilde{z}_\omega = \sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\{\widehat{\omega}_1(\theta) - \omega_1(\theta)\}$. The bound for $\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\{\omega_1(\theta) - \widehat{\omega}_1(\theta)\}$ can be obtained similarly.

Let ϵ_i, $i = 1, \cdots, n$ be i.i.d Rademacher random variables, by Lemma A.5, we have

$$\mathrm{E}(e^{\lambda\widetilde{z}_\omega}) \le \mathrm{E}\Big[\exp\Big(\Big|\frac{2\lambda}{n}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\sum_{i=1}^{n}\epsilon_i\big(\frac{\omega_1(\theta)}{\omega_1(\theta)+\omega_2(\theta)\exp\{C_\theta(X_i,Y_i)\}} - \omega_1(\theta)\big)\Big|\Big)\Big].$$

Note that $\psi(x) = \omega_1/(\omega_1 + (1-\omega_1)e^x) - \omega_1$ is Lipschitz with constant $(1-c_0)/c_0$ and $\psi(0) = 0$. By the Talagrand comparison inequality A.6, we have

$$\begin{aligned}
\mathrm{E}(e^{\lambda\widetilde{z}_\omega}) &\le \mathrm{E}\Big[\exp\Big(\Big|\frac{2\lambda}{n}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\sum_{i=1}^{n}\epsilon_i\big(\frac{\omega_1(\theta)}{\omega_1(\theta)+\omega_2(\theta)\exp\{C_\theta(X_i,Y_i)\}} - \omega_1(\theta)\big)\Big|\Big)\Big] \\
&\le \mathrm{E}\Big[\exp\Big(\Big|\frac{2\lambda}{n}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\sum_{i=1}^{n}\epsilon_i\frac{1-c_0}{c_0}C_\theta(Y_i)\Big|\Big)\Big] \\
&\le \underbrace{\mathrm{E}\Big[\exp\Big(\Big|\frac{2\lambda}{n}\frac{1-c_0}{c_0}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\sum_{i=1}^{n}(\epsilon_i(\beta_2-\beta_1)^TX_i(Y_i-\omega_1^*\beta_1^{*T}X_i-\omega_2^*\beta_2^{*T}X_i))\Big|\Big)\Big]}_{(i)} \\
&\times \underbrace{\mathrm{E}\Big[\exp\Big(\Big|\frac{2\lambda}{n}\frac{1-c_0}{c_0}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\sum_{i=1}^{n}(\epsilon_i(\beta_2-\beta_1)^TX_i(\omega_1^*\beta_1^{*T}X_i+\omega_2^*\beta_2^{*T}X_i-\frac{(\beta_1+\beta_2)^TX_i}{2}))\Big|\Big)\Big]}_{(ii)}
\end{aligned}$$

We first bound the term (i). Let $1/2_{net}(\Gamma(s)\cap S^{p-1})$ be a $1/2-net$ of $\Gamma(s)\cap S^{p-1}$ and M_{net} be its covering numbers. Note that for any $\nu_j \in S^{p-1}$, $(\nu_j^TX_i)^2$ and $\widetilde{Y}_i^2$ are sub-exponential. Let $\widetilde{Y}_i = Y_i - (\omega_1^*\beta_1^* + \omega_2^*\beta_2^*)^TX_i$. We have

$$\begin{aligned}
(i) &= \mathrm{E}\Big[\exp\Big(\Big|\frac{2\lambda}{n}\frac{1-c_0}{c_0}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\sum_{i=1}^{n}(\epsilon_i(\beta_2-\beta_1)^TX_i\widetilde{Y}_i)\Big|\Big)\Big] \\
&\le \mathrm{E}\Big[\exp\Big(\frac{2\lambda}{n}\frac{1-c_0}{c_0}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\|\beta_1-\beta_2\|_2\sup_{\nu\in\Gamma(s)\cap S^{p-1}}|\sum_{i=1}^{n}\epsilon_i\widetilde{Y}_i\nu^TX_i|\Big)\Big] \\
&\le \mathrm{E}\Big[\exp\Big(\frac{4\lambda}{n}\frac{1-c_0}{c_0}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\|\beta_1-\beta_2\|_2\sup_{\nu_j\in1/2_{net}(\Gamma(s)\cap S^{p-1})}|\sum_{i=1}^{n}\nu_j^TX_i\epsilon_i\widetilde{Y}_i|\Big)\Big].
\end{aligned}$$

Since $\|\beta_1-\beta_2\|_2 \le \|\beta_1-\beta_1^*\|_2 + \|\beta_2-\beta_2^*\|_2 + \|\beta_1^*-\beta_2^*\|_2$. We have $\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\|\beta_1-\beta_2\|_2 \le (2C_b\Delta + 2M_b)$. Since both X_{ij} and $\widetilde{Y}_i$ are i.i.d Gaussian, $X_{ij}\widetilde{Y}_i$ is sub-exponential. Hence, for sufficient small λ ($\lambda/n \to 0$),

$$\begin{aligned}
(i) &\le \mathrm{E}\Big[\exp\Big(\Big|\frac{4\lambda}{n}\frac{1-c_0}{c_0}(2C_b\Delta+2M_b)\sup_{\nu_j\in1/2_{net}(\Gamma(s)\cap S^{p-1})}|\sum_{i=1}^{n}\nu_j^TX_i\epsilon_i\widetilde{Y}_i|\Big)\Big] \\
&\le M_{net}\mathrm{E}\Big[\exp\Big(\Big|\frac{4\lambda}{n}\frac{1-c_0}{c_0}(2C_b\Delta+2M_b)|\sum_{i=1}^{n}\nu_j^TX_i\epsilon_i\widetilde{Y}_i|\Big)\Big] \\
&\le \exp\Big\{\frac{16\lambda^2(2C_b\Delta+2M_b)^2(1-c_0)^2}{nc_0^2}C_1 + c_ds\log(p)\Big\}.
\end{aligned}$$

Then we consider the term (ii). Note that for any $\nu_j \in S^{p-1}$, $(\nu_j^T X_i)^2$ is sub-exponential, for sufficient small λ ($\lambda/n \to 0$), we have

$$\begin{aligned}
(ii) \leq & \mathrm{E}\Big[\exp\Big(\Big|\frac{2\lambda(1-c_0)}{c_0 n}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\|\beta_2-\beta_1\|_2 \sup_{\theta\in\mathcal{B}_{con}(\theta^*)}(\max_k\{\|\omega_k^*(\beta_k^*-\frac{\beta_1+\beta_2}{2})\|_2\}) \\
& \cdot \sup_{\nu_1,\nu_2\in\Gamma(s)\cap S^{p-1}}\sum_{i=1}^n|\nu_1 X_i||\nu_2 X_i|\Big|\Big)\Big] \\
\leq & \mathrm{E}\Big[\exp\Big(\Big|\frac{2\lambda(1-c_0)}{c_0 n}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\|\beta_2-\beta_1\|_2 \sup_{\theta\in\mathcal{B}_{con}(\theta^*)}(\max_k\{\|\omega_k^*(\beta_k^*-\frac{\beta_1+\beta_2}{2})\|_2\}) \\
& \cdot \sup_{\nu\in\Gamma(s)\cap S^{p-1}}\sum_{i=1}^n(\nu X_i)^2\Big|\Big)\Big] \\
\leq & \mathrm{E}\Big[\exp\Big(\Big|\frac{8\lambda(1-c_0)}{c_0 n}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\|\beta_2-\beta_1\|_2 \sup_{\theta\in\mathcal{B}_{con}(\theta^*)}(\max_k\{\|\omega_k^*(\beta_k^*-\frac{\beta_1+\beta_2}{2})\|_2\}) \\
& \cdot \max_{\nu_j\in 1/2_{net}(\Gamma(s)\cap S^{p-1})}\sum_{i=1}^n(\nu_j X_i)^2\Big|\Big)\Big]
\end{aligned}$$

$$\begin{aligned}
& \leq M_{net}\exp\Big(\frac{64\lambda^2(1-c_0)^4c_2^2(2C_b+2M_b)^2}{c_0^2 n}C_1\Big) \\
& \leq \exp\Big(\frac{64\lambda^2(1-c_0)^4c_2^2(2C_b+2M_b)^2}{c_0^2 n}C_1+2c_d s\log(p)\Big).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\mathrm{E}(e^{\lambda\widetilde{z}_\omega}) &\leq \exp\Big\{\frac{16\lambda^2(2C_b+2M_b)^2(1-c_0)^2}{nc_0^2}C_1+\log(p)\Big\}\exp\{\frac{64\lambda^2(1-c_0)^4c_2^2(2C_b+2M_b)^2}{c_0^2 n}C_1+c_d s\log(p)\} \\
&= \exp\{c_1\frac{\lambda^2}{n}+c_2 s\log(p)\}.
\end{aligned}$$

By the Chernoff bound, letting $\lambda=\sqrt{ns\log(p)/c_1}$ and $t=(c_2+1)\sqrt{c_1 s\log(p)/n}$, we have

$$\mathbb{P}(\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\widehat{\omega}_1(\theta)-\omega_1(\theta)>t)\leq\mathbb{P}(\widetilde{z}_\omega>t)\leq e^{-\lambda t}\mathrm{E}(e^{\lambda\widetilde{z}_\omega})\leq e^{-s\log(p)}<\frac{1}{p}.$$

Similarly, we can show that $\mathbb{P}(\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\omega_1(\theta)-\widehat{\omega}_1(\theta)>t)\leq 1/p$.

To sum up, we have $\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\{|\widehat{\omega}_1(\theta)-\omega_1(\theta)|\}=O(\sqrt{\frac{s\log(p)}{n}})$ with probability at least $1-2/p$.

■

Appendix F. Proof of Theorem 4

F.1 Concentration of the estimator $\widehat{\beta}_k^{(t+1)}$

In this section, we will show the proof of the following Lemma.

Lemma A.9 *Suppose that* $\theta^* \in \Theta^*$ *and* $\widehat{\beta}_k^{(0)} \in \mathcal{B}_{con}$. *Let* $\lambda_n^{(t+1)} \geq 4C_{con}(\sqrt{\log(n)^2\log(p)/n}+ 8\kappa_0(|\widehat{\omega}_1^{(t)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(t)} - \beta_1^*\|_2 \vee \|\widehat{\beta}_2^{(t)} - \beta_2^*\|_2)/\sqrt{s})$ *for* κ_0 *defined before and some constant* C_{con} *and* $\widehat{\beta}_k^{(t+1)}$ *solved by*

$$(\widehat{\beta}_1^{(t+1)}, \widehat{\beta}_2^{(t+1)}) = \underset{\beta_1,\beta_2}{\operatorname{argmin}} \Big\{ \sum_{k=1}^{2} \beta_k^T \widehat{\Sigma}_k^{(t+1)} \beta_k - 2\sum_{k=1}^{2} (\widehat{\rho}_k^{(t+1)})^T \beta_k + \lambda^{(t+1)} \sum_{j=1}^{p} \sqrt{\sum_{k=1}^{2} \beta_{kj}^2} \Big\},$$

where $\widehat{\rho}_k^{(t+1)} = (\frac{1}{n}\sum_{i=1}^n \widehat{\eta}_{ik}(\widehat{\theta}^{(t+1)} Y_i X_i^T)^T$ *and* $\widehat{\Sigma}_k^{(t+1)} = (\frac{1}{n}\sum_{i=1}^n \widehat{\eta}_{ik}(\widehat{\theta}^{(t+1)} X_i X_i^T)$. *We have*

$$\widehat{\beta}_k^{(t+1)} - \beta_k^* \in \Gamma(s),$$

and

$$\|\widehat{\beta}_k^{(t+1)} - \beta_k^*\|_2 \leq \frac{4}{\tau_0} d_{2,s}\big(M_n(\widehat{\theta}^{(t)}), M(\theta^*)\big) + \frac{2}{\tau_0}\sqrt{s}\lambda_n^{(t+1)}.$$

Proof Because $(\widehat{\beta}_1^{(t+1)}, \cdots, \widehat{\beta}_k^{(t+1)})$ is the minimizer of

$$\sum_{k=1}^{2} \beta_k^T \widehat{\Sigma}_k^{(t+1)} \beta_k - 2\sum_{k=1}^{2} (\widehat{\rho}_k^{(t+1)})^T \beta_k + \lambda^{(t+1)} \sum_{j=1}^{p} \sqrt{\sum_{k=1}^{2} \beta_{kj}^2},$$

we have

$$\begin{aligned}
&\lambda^{(t+1)}\Big(\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(\widehat{\beta}_{kj}^{(t+1)})^2} - \sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(\beta_{kj}^*)^2}\Big) \\
&\leq 2\sum_{k=1}^{2}\Big\{(\beta_k^* - \widehat{\beta}_k^{(t+1)})^T \widehat{\Sigma}_k^{(t+1)} \beta_k^* - (\beta_k^* - \widehat{\beta}_k^{(t+1)})^T \widehat{\Sigma}_k^{(t+1)} (\beta_k^* - \widehat{\beta}_k^{(t+1)})/2 - (\widehat{\rho}_k^{(t+1)})^T (\beta_k^* - \widehat{\beta}_k^{(t+1)})\Big\} \\
&\leq 2\sum_{k=1}^{2}\Big\{(\beta_k^* - \widehat{\beta}_k^{(t+1)})^T \widehat{\Sigma}_k^{(t+1)} \beta_k^* - (\widehat{\rho}_k^{(t+1)})^T (\beta_k^* - \widehat{\beta}_k^{(t+1)})\Big\} \\
&= 2\sum_{k=1}^{2}\Big\{(\beta_k^* - \widehat{\beta}_k^{(t+1)})^T \big[\widehat{\Sigma}_k^{(t+1)} \beta_k^* - \widehat{\rho}_k^{(t+1)}\big]\Big\} \\
&= 2\sum_{k=1}^{2}\Big\{(\beta_k^* - \widehat{\beta}_k^{(t+1)})^T \big[(\widehat{\Sigma}_k^{(t+1)} - \Sigma_k^{(t+1)})\beta_k^* + \Sigma_k^{(t+1)}\beta_k^* - \rho_k^{(t+1)} - (\widehat{\rho}_k^{(t+1)} - \rho_k^{(t+1)})\big]\Big\} \\
&= 2\sum_{k=1}^{2}\Big\{(\beta_k^* - \widehat{\beta}_k^{(t+1)})^T \big[(\widehat{\Sigma}_k^{(t+1)} - \Sigma_k^{(t+1)})\beta_k^* - (\widehat{\rho}_k^{(t+1)} - \rho_k^{(t+1)})\big]\Big\} \\
&\quad + 2\sum_{k=1}^{2}\Big\{(\beta_k^* - \widehat{\beta}_k^{(t+1)})^T \big[\Sigma_k^{(t+1)}\beta_k^* - \rho_k^{(t+1)}\big]\Big\}
\end{aligned}$$

$$
\begin{aligned}
&= 2\sum_{k=1}^{2}\Big\{(\beta_k^* - \widehat{\beta}_k^{(t+1)})^T\big[(\widehat{\Sigma}_k^{(t+1)} - \Sigma_k^{(t+1)})\beta_k^* - (\widehat{\rho}_k^{(t+1)} - \rho_k^{(t+1)})\big]\Big\} \\
&\quad + 2\sum_{k=1}^{2}\Big\{(\beta_k^* - \widehat{\beta}_k^{(t+1)})^T\big[(\Sigma_k^{(t+1)} - \Sigma_k^*)\beta_k^* - \omega_k^*\Sigma\beta_k^* - \rho_k^* - (\rho_k^{(t+1)} - \rho_k^*)\big]\Big\} \\
&= 2\sum_{k=1}^{2}\Big\{(\beta_k^* - \widehat{\beta}_k^{(t+1)})^T\big[(\widehat{\Sigma}_k^{(t+1)} - \Sigma_k^{(t+1)})\beta_k^* - (\widehat{\rho}_k^{(t+1)} - \rho_k^{(t+1)})\big]\Big\} \\
&\quad + 2\sum_{k=1}^{2}\Big\{(\beta_k^* - \widehat{\beta}_k^{(t+1)})^T\big[(\Sigma_k^{(t+1)} - \Sigma_k^*)\beta_k^* - (\rho_k^{(t+1)} - \rho_k^*)\big]\Big\} \\
&\leq 2\sum_{k=1}^{2}\Big\{(\beta_k^* - \widehat{\beta}_k^{(t+1)})^T\big[(\widehat{\Sigma}_k^{(t+1)} - \Sigma_k^{(t+1)})\beta_k^* - (\widehat{\rho}_k^{(t+1)} - \rho_k^{(t+1)})\big]\Big\} \\
&\quad + 4\kappa_0\frac{|\widehat{\omega}_k^{(t)} - \omega_k| \vee \|\widehat{\beta}_k^{(t)} - \beta_k^*\|_2}{\sqrt{s}}\sqrt{s}\|\widehat{\beta}_k^{(t+1)} - \beta_k^*\|_2\}.
\end{aligned}
$$

Let $u_k^{(t+1)} = \widehat{\beta}_k^{(t+1)} - \beta_k^*$, we have

$$
\begin{aligned}
\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(\widehat{\beta}_{kj}^{(t+1)})^2} - \sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(\beta_{kj}^*)^2} &= \sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(\beta_{kj}^* + u_{kj}^{(t+1)})^2} - \sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(\beta_{kj}^*)^2} \\
&= \sum_{j\in S}\sqrt{\sum_{k=1}^{2}(\beta_{kj}^* + u_{kj}^{(t+1)})^2} + \sum_{j\in S^c}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2} - \sum_{j\in S}\sqrt{\sum_{k=1}^{2}(\beta_{kj}^*)^2} \\
&\geq \sum_{j\in S^c}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2} - \sum_{j\in S}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2}.
\end{aligned}
$$

Let $\widetilde{S}$ be a set of size $2s$, which contains S and the largest s coefficients of $\widehat{\rho}_k^{(t+1)} - \rho_k^{(t+1)}$. We have

$$
\begin{aligned}
&|(\beta_k^* - \widehat{\beta}_k^{(t+1)})^T(\widehat{\rho}_k^{(t+1)} - \rho_k^{(t+1)})| \\
&\leq \|\widehat{\rho}_k^{(t+1)} - \rho_k^{(t+1)}\|_{2,s}\|(\beta_k^* - \widehat{\beta}_k^{(t+1)})_{\widetilde{S}}\|_2 + \|(\widehat{\rho}_k^{(t+1)} - \rho_k^{(t+1)})_{\widetilde{S}^c}\|_\infty\|(\beta_k^* - \widehat{\beta}_k^{(t+1)})_{\widetilde{S}^c}\|_1 \\
&\leq C_{con}\sqrt{\log(n)^2\log(p)/n}\sqrt{s}\|(\beta_k^* - \widehat{\beta}_k^{(t+1)})_{\widetilde{S}}\|_2 + C_{con}\sqrt{\log(n)^2\log(p)/n}\|(\beta_k^* - \widehat{\beta}_k^{(t+1)})_{\widetilde{S}^c}\|_1,
\end{aligned}
$$

where in the last inequality, we use Lemma 2 and the fact that $\|(\widehat{\rho}_k^{(t+1)} - \rho_k^{(t+1)})_{\widetilde{S}^c}\|_\infty \leq \|(\widehat{\rho}_k^{(t+1)} - \rho_k^{(t+1)})_{\widetilde{S}}\|_2/\sqrt{s} \leq \|\widehat{\rho}_k^{(t+1)} - \rho_k^{(t+1)}\|_{2,s}/\sqrt{s}$. We have the same results for the term $|(\beta_k^* - \widehat{\beta}_k^{(t+1)})^T(\widehat{\Sigma}_k^{(t+1)} - \Sigma_k^{(t+1)})\beta_k^*|$.

Also, note that $\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2}\le\sum_{k=1}^{2}\|u_k^{(t+1)}\|_1\le 2\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2}$, we have

$$\begin{aligned}
&\lambda_n^{(t+1)}\Big\{\sum_{j\in S^c}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2}-\sum_{j\in S}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2}\Big\}\\
&\ge\lambda_n^{(t+1)}\Big\{\sum_{j\in \widetilde{S}^c}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2}-\sum_{j\in \widetilde{S}}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2}\Big\}\\
&\ge\lambda_n^{(t+1)}\Big\{\frac{1}{2}\sum_{k=1}^{2}\|(\mu_k^{(t+1)})_{\widetilde{S}^c}\|_1-\sum_{k=1}^{2}\sqrt{s}\|(\mu_k^{(t+1)})_{\widetilde{S}}\|_2\Big\}.
\end{aligned}$$

Then

$$\begin{aligned}
&\lambda_n^{(t+1)}\Big\{\frac{1}{2}\sum_{k=1}^{2}\|(\mu_k^{(t+1)})_{\widetilde{S}^c}\|_1-\sum_{k=1}^{2}\sqrt{s}\|(\mu_k^{(t+1)})_{\widetilde{S}}\|_2\Big\}\\
&\quad\le C_{con}\sqrt{\log(n)^2\log(p)/n}\sqrt{s}\sum_{k=1}^{2}\|(\mu_k)_{\widetilde{S}}\|_2+C_{con}\sqrt{\log(n)^2\log(p)/n}\sum_{k=1}^{2}\|(\mu_k^{(t+1)})_{\widetilde{S}^c}\|_1\\
&\qquad+4\kappa_0\frac{|\widehat{\omega}_k^{(t)}-\omega_k|\vee\|\widehat{\beta}_k^{(t)}-\beta_k^*\|_2}{\sqrt{s}}\sqrt{s}\|\mu_k^{(t+1)}\|_2\Big\}.
\end{aligned}$$

Let $\lambda_n^{(t+1)}\ge 4C_{con}\sqrt{\log(n)^2\log(p)/n}+8\kappa_0\frac{|\widehat{\omega}_k^{(t)}-\omega_k|\vee\|\widehat{\beta}_k^{(t)}-\beta_k^*\|_2}{\sqrt{s}}$, we have

$$\sum_{k=1}^{2}\|(u_k^{(t+1)})_{\widetilde{S}^c}\|_1\le 5\sum_{k=1}^{2}\sqrt{s}\|(u_k^{(t+1)})_{\widetilde{S}}\|_2+2\sqrt{s}\sum_{k=1}^{2}\|\mu_k^{(t+1)}\|_2.$$

Let $\mu^T=(\mu_1^T,\mu_2^T)^T$, and $S_1=\{\widetilde{S},\widetilde{S}+p\}$, where $\widetilde{S}+p$ means the collection of the index in $\widetilde{S}$ adds p. We have

$$\|(u^{(t+1)})_{S_1^c}\|_1\le 5\sqrt{2s}\|(u^{(t+1)})_{S_1}\|_2+2\sqrt{2s}\|\mu^{(t+1)}\|_2.$$

Next, we prove the second conclusion. Note that

$$\begin{aligned}
&\lambda^{(t+1)}\Big(\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(\widehat{\beta}_{kj}^{(t+1)})^2}-\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(\beta_{kj}^*)^2}\Big)\\
&\le 2\sum_{k=1}^{2}\Big\{(\beta_k^*-\widehat{\beta}_k^{(t+1)})^T\widehat{\Sigma}_k^{(t+1)}\beta_k^*-(\beta_k^*-\widehat{\beta}_k^{(t+1)})^T\widehat{\Sigma}_k^{(t+1)}(\beta_k^*-\widehat{\beta}_k^{(t+1)})/2-(\widehat{\rho}_k^{(t+1)})^T(\beta_k^*-\widehat{\beta}_k^{(t+1)})\Big\}
\end{aligned}$$

It follows that

$$\sum_{k=1}^{2}(\beta_k^*-\widehat{\beta}_k^{(t+1)})^T\widehat{\Sigma}_k^{(t+1)}(\beta_k^*-\widehat{\beta}_k^{(t+1)}) \le |2\sum_{k=1}^{2}\Big\{(\beta_k^*-\widehat{\beta}_k^{(t+1)})^T\widehat{\Sigma}_k^{(t+1)}\beta_k^*-(\widehat{\rho}_k^{(t+1)})^T(\beta_k^*-\widehat{\beta}_k^{(t+1)})\Big\}|$$
$$+\lambda_n^{(t+1)}\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(\widehat{\beta}_{kj}^{(t+1)}-\beta_{kj})^2}$$
$$\le |2\sum_{k=1}^{2}\Big\{(\beta_k^*-\widehat{\beta}_k^{(t+1)})^T\widehat{\Sigma}_k^{(t+1)}\beta_k^*-(\widehat{\rho}_k^{(t+1)})^T(\beta_k^*-\widehat{\beta}_k^{(t+1)})\Big\}|$$
$$+\lambda_n^{(t+1)}\sum_{k=1}\|\widehat{\beta}_k^{(t+1)}-\beta_k^*\|_1$$

Recall that

$$\widehat{\Sigma}_k^{(t+1)}=\frac{1}{n}\sum_{i=1}^{n}\eta_{k,\widehat{\theta}^{(t)}}(Y_i,X_i)X_iX_i^T.$$

By Lemma 1, we know that

$$|\frac{1}{n}\sum_{i=1}^{n}\eta_{k,\widehat{\theta}^{(t)}}(Y_i,X_i)-\mathrm{E}(\widehat{\omega}_k^{(t)})|=o(\sqrt{s\log(p)/n}).$$

It follows that $\frac{1}{n}\sum_{i=1}^{n}\eta_{k,\widehat{\theta}^{(t)}}(Y_i,X_i)>\tau_1$ for some positive constant τ_1. Define set $\mathcal{N}=\{i:\eta_{k,\widehat{\theta}^{(t)}}(Y_i,X_i)>\tau_1/2\}$. We have $|\mathcal{N}|/n>\tau_1/(2-\tau_1)$. By Lemma A.4, we have $\|\widehat{\Sigma}_k^{(t+1)}\|_{2,s}>\tau_0$ for some positive constant τ_0.

Then

$$\sum_{k=1}^{2}\|\beta_k^*-\widehat{\beta}_k^{(t+1)}\|_2^2\tau_0 \le 2\sum_{k=1}^{2}\Big\{(\beta_k^*-\widehat{\beta}_k^{(t+1)})^T\widehat{\Sigma}_k^{(t+1)}\beta_k^*-(\widehat{\rho}_k^{(t+1)})^T(\beta_k^*-\widehat{\beta}_k^{(t+1)})\Big\}|$$
$$+\lambda_n^{(t+1)}\sqrt{s}\sum_{k=1}\|\widehat{\beta}_k^{(t+1)}-\beta_k^*\|_2$$

Hence,

$$(\sum_{k=1}^{2}\|\beta_k^*-\widehat{\beta}_k^{(t+1)}\|_2)^2\cdot\tau_0/2 \le 2d_{2,s}(M_n(\widehat{\theta}^{t+1}),M(\theta^*))\sum_{k=1}^{2}\|\beta_k^*-\widehat{\beta}_k^{(t+1)}\|_2$$
$$+\lambda_n^{(t+1)}\sqrt{s}\sum_{k=1}^{2}\|\widehat{\beta}_k^{(t+1)}-\beta_k^*\|_2$$

It follows that

$$\sum_{k=1}^{2}\|\widehat{\beta}_k^{(t+1)}-\beta_k^*\|_2 \le 4/\tau_0 d_{2,s}(M_n(\widehat{\theta}^{t+1}),M(\theta^*))+2/\tau_0\sqrt{s}\lambda_n^{(t+1)}.$$

■

F.2 Proof of the main Theorem

The proof for the theorem is analogous to that in Section A.2 of Cai et al. (2019), we show it here for completeness. In the algorithm, we update $\lambda_n^{(t)}$ by $\lambda_n^{(t)} = \kappa\lambda_n^{(t-1)} + C_\lambda\sqrt{\log(n)^2\log(p)/n}$ and $\lambda_n^{(0)} = C_1\frac{(|\widehat{\omega_1}^{(0)}-\omega_1^*|\vee\|\widehat{\beta}_1^{(0)}-\beta_1^*\|_2\vee\|\widehat{\beta}_2^{(0)}-\beta_2^*\|_2)}{\sqrt{s}} + C_\lambda\sqrt{\log(n)^2\log(p)/n}$, with $C_1 = \tau_0/4$. Thus

$$\lambda_n^{(t)} = \kappa^t C_1\frac{(|\widehat{\omega_1}^{(0)}-\omega_1^*|\vee\|\widehat{\beta}_1^{(0)}-\beta_1^*\|_2\vee\|\widehat{\beta}_2^{(0)}-\beta_2^*\|_2)}{\sqrt{s}} + \frac{1-\kappa^{t+1}}{1-\kappa}C_\lambda\sqrt{\log(n)^2\log(p)/n}.$$

Let $\kappa = (1\vee 32/\tau_0)\cdot\kappa_0$. Because $\kappa_0 \le \frac{1}{2\vee 64/\tau_0}$, we have $\kappa\in(0,1/2)$. Let

$$\begin{aligned}
C^* &= \Big\{\big(\frac{2\kappa^2-4\kappa+2}{2\kappa^2-5\kappa+2}\big(\frac{4}{\tau_0}+\frac{8}{\tau_0(1-\kappa)}\big)\big)\vee\frac{1-\kappa}{1-2\kappa}\Big\}C_{con}\\
C_\lambda &= 4C_{con}+\frac{8\kappa_0}{1-\kappa}C^*.
\end{aligned}$$

We have (1):$\kappa\ge\kappa_0$, $C_1\kappa\ge 8\kappa_0$, and $\kappa_0/\tau_0 + 2C_1\kappa/\tau_0\le\kappa$; (2): $\kappa_0 C^*/(1-\kappa)+C_{con}\le C^*$, and $4C_{con}/\tau_0 + 2C_\lambda/((1-\kappa)\tau_0)\le C^*$.

Next, we use induction to show the following conclusions.

$$\begin{aligned}
&\lambda_n^{(t+1)} \ge 4C_{con}\sqrt{\frac{\log(n)^2\log(p)}{n}} + 8\kappa_0\frac{(|\widehat{\omega}_1^{(t)}-\omega_1^*|\vee\|\widehat{\beta}_1^{(t)}-\beta_1^*\|_2\vee\|\widehat{\beta}_2^{(t)}-\beta_2^*\|_2)}{\sqrt{s}}\\
&|\widehat{\omega}_1^{(t+1)}-\omega_1^*|\vee\|\widehat{\beta}_1^{(t+1)}-\beta_1\|_2\vee\|\widehat{\beta}_2^{(t+1)}-\beta_2\|_2\\
&\quad\le\kappa^{t+1}(|\widehat{\omega}_1^{(0)}-\omega_1^*|\vee\|\widehat{\beta}_1^{(0)}-\beta_1^*\|_2\vee\|\widehat{\beta}_2^{(0)}-\beta_2^*\|_2) + \frac{1-\kappa^{t+2}}{1-\kappa}C^*\sqrt{\frac{s\log(n)^2\log(p)}{n}}.
\end{aligned}$$

Firstly, note that

$$\begin{aligned}
d_{2,s}(M_n(\widehat{\theta}^{(1)}),M(\theta^*)) &\le d_{2,s}(M(\widehat{\theta}^{(0)}),M(\theta^*)) + C_{con}\sqrt{\frac{s\log(n)^2\log(p)}{n}}\\
&\le \kappa_0(|\widehat{\omega}_1^{(0)}-\omega_1^*|\vee\|\widehat{\beta}_1^{(0)}-\beta_1^*\|_2\vee\|\widehat{\beta}_2^{(0)}-\beta_2^*\|_2) + C_{con}\sqrt{\frac{s\log(n)^2\log(p)}{n}}\\
&\le \kappa(|\widehat{\omega}_1^{(0)}-\omega_1^*|\vee\|\widehat{\beta}_1^{(0)}-\beta_1^*\|_2\vee\|\widehat{\beta}_2^{(0)}-\beta_2^*\|_2) + \frac{1-\kappa^2}{1-\kappa}C^*\sqrt{\frac{s\log(n)^2\log(p)}{n}},
\end{aligned}$$

since $\kappa_0\le\kappa$, and $C_{con}\le\frac{\kappa_0}{1-\kappa}C^* + C_{con}\le C^*\le(1+\kappa)C^*$. It follows that

$$\begin{aligned}
\lambda_n^{(1)} &= \kappa\lambda_n^{(0)} + C_\lambda\sqrt{\frac{\log(n)^2\log(p)}{n}}\\
&= C_1\kappa\frac{(|\widehat{\omega}_1^{(0)}-\omega_1^*|\vee\|\widehat{\beta}_1^{(0)}-\beta_1^*\|_2\vee\|\widehat{\beta}_2^{(0)}-\beta_2^*\|_2)}{\sqrt{s}} + (\kappa C_\lambda + C_\lambda)\sqrt{\frac{\log(n)^2\log(p)}{n}}\\
&\ge 4C_{con}\sqrt{\frac{\log(n)^2\log(p)}{n}} + 8\kappa_0\frac{(|\widehat{\omega}_1^{(0)}-\omega_1^*|\vee\|\widehat{\beta}_1^{(0)}-\beta_1^*\|_2\vee\|\widehat{\beta}_2^{(0)}-\beta_2^*\|_2)}{\sqrt{s}},
\end{aligned}$$

since $(\kappa+1)C_\lambda \geq 4C_{con}$, and $C_1\kappa \geq 8\kappa_0$. Moreover, by Lemma A.9, we have

$$\begin{aligned}
\|\widehat{\beta}_k^{(1)} - \beta_k^*\|_2 &\leq 4/\tau_0 d_{2,s}(M_n(\widehat{\theta}^{(1)}), M(\theta^*)) + 2/\tau_0\sqrt{s}\lambda_n^{(1)} \\
&\leq 4/\tau_0(\kappa_0(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2^*\|_2) + C_{con}\sqrt{\frac{s\log(n)^2\log(p)}{n}}) \\
&+ 2/\tau_0\sqrt{s}(C_1\kappa\frac{(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2^*\|_2)}{\sqrt{s}} + (\kappa C_\lambda + C_\lambda)\sqrt{\frac{\log(n)^2\log(p)}{n}}) \\
&= (4/\tau_0\kappa_0 + 2/\tau_0 C_1\kappa)(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2^*\|_2) \\
&\quad + (4/\tau_0 C_{con} + 2/\tau_0(1+\kappa)C_\lambda)\sqrt{\frac{s\log(n)^2\log(p)}{n}} \\
&\leq \kappa(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2^*\|_2) + \frac{1-\kappa^2}{1-\kappa}C^*\sqrt{\frac{s\log(n)^2\log(p)}{n}},
\end{aligned}$$

since $4/\tau_0\kappa_0 + 2/\tau_0 C_1\kappa \leq \kappa$ and $4/\tau_0 C_{con} + 2/\tau_0(1+\kappa)C_\lambda \leq (1+\kappa)C^*$. Hence,

$$\begin{aligned}
|\widehat{\omega}_1^{(1)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(1)} - \beta_1^*\|_2 \vee \|\widehat{\beta}_2^{(1)} - \beta_2^*\|_2 \leq \kappa(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2^*\|_2) \\
+ \frac{1-\kappa^2}{1-\kappa}C^*\sqrt{\frac{s\log(n)^2\log(p)}{n}}.
\end{aligned}$$

Next we assume the conclusions hold at the t-th step, namely

$$\lambda_n^{(t)} \geq 4C_{con}\frac{\log(n)^2\log(p)}{n} + 8\kappa_0\frac{(|\widehat{\omega}_1^{(t-1)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(t-1)} - \beta_1^*\|_2 \vee \|\widehat{\beta}_2^{(t-1)} - \beta_2^*\|_2)}{\sqrt{s}}$$

$$|\widehat{\omega}_1^{(t)} - \omega_1^*| \vee \|\widehat{\beta}_k^{(t)} - \beta_k\|_2 \leq \kappa^t(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2^*\|_2) + \frac{1-\kappa^{t+1}}{1-\kappa}C^*\sqrt{\frac{s\log(n)^2\log(p)}{n}}.$$

Then

$$\begin{aligned}
&4C_{con}\sqrt{\frac{\log(n)^2\log(p)}{n}} + 8\kappa_0\frac{|\widehat{\omega}_1^{(t)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(t)} - \beta_1\|_2 \vee \|\widehat{\beta}_2^{(t)} - \beta_2\|_2}{\sqrt{s}} \\
&\leq 4C_{con}\sqrt{\frac{\log(n)^2\log(p)}{n}} + 8\kappa_0\frac{\kappa^t(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2^*\|_2) + \frac{1-\kappa^{t+1}}{1-\kappa}C^*\sqrt{\frac{s\log(n)^2\log(p)}{n}}}{\sqrt{s}} \\
&\leq 8\kappa_0\kappa^t\frac{|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2^*\|_2}{\sqrt{s}} + 4C_{con}\sqrt{\frac{\log(n)^2\log(p)}{n}} \\
&\quad + 8\kappa_0\frac{1-\kappa^{t+1}}{1-\kappa}C^*\sqrt{\frac{\log(n)^2\log(p)}{n}} \\
&\leq \kappa^{t+1}C_1\frac{|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2^*\|_2}{\sqrt{s}} + \frac{1-\kappa^{t+2}}{1-\kappa}C_\lambda\sqrt{\frac{\log(n)^2\log(p)}{n}} \\
&= \lambda_n^{(t+1)},
\end{aligned}$$

Since $\frac{1-\kappa^{t+2}}{1-\kappa}C_\lambda \geq 4C_{con} + 8\kappa_0\frac{1-\kappa^{t+1}}{1-\kappa}C^*$ and $C_1\kappa \geq 2\kappa_0$.

Then note that

$$
\begin{aligned}
&d_{2,s}(M_n(\widehat{\theta}^{(t+1)}, M(\theta^*)) \\
&\le d_{2,s}(M(\widehat{\theta}^{(t)}, M(\theta^*)) + C_{con}\sqrt{\frac{s\log(n)^2\log(p)}{n}} \\
&\le \kappa_0(|\widehat{\omega}_1^{(t)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(t)} - \beta_1\|_2 \vee \|\widehat{\beta}_2^{(t)} - \beta_2\|_2) + C_{con}\sqrt{\frac{s\log(n)^2\log(p)}{n}} \\
&\le \kappa_0(\kappa^t(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2\|_2) + \frac{1-\kappa^{t+1}}{1-\kappa}C^*\sqrt{\frac{s\log(n)^2\log(p)}{n}}) + C_{con}\sqrt{\frac{s\log(n)^2\log(p)}{n}} \\
&\le \kappa^{t+1}((|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2\|_2) + \frac{1-\kappa^{t+2}}{1-\kappa}C^*\sqrt{\frac{s\log(n)^2\log(p)}{n}},
\end{aligned}
$$

since $\kappa_0 \le \kappa$ and $\kappa_0\frac{1-\kappa^{t+1}}{1-\kappa}C^* + C_{con} \le \frac{1-\kappa^{t+2}}{1-\kappa}C^*$.

Then by Lemma A.9,

$$
\begin{aligned}
&\|\widehat{\beta}_k^{(t+1)} - \beta_k^*\|_2 \\
&\le 4/\tau_0 d_{2,s}(M_n(\widehat{\theta}^{(t+1)}, M(\theta^*)) + 2/\tau_0\sqrt{s}\lambda_n^{(t+1)} \\
&\le 4/\tau_0\big\{\kappa_0(\kappa^t(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2\|_2) + \frac{1-\kappa^{t+1}}{1-\kappa}C^*\sqrt{\frac{s\log(n)^2\log(p)}{n}}) + C_{con}\sqrt{\frac{s\log(p)}{n}}\big\} \\
&+ 2\tau_0\sqrt{s}\big\{\kappa^{t+1}C_1\frac{|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2^*\|_2}{\sqrt{s}} + \frac{1-\kappa^{t+2}}{1-\kappa}C_\lambda\sqrt{\frac{\log(p)}{n}}\big\} \\
&\le (4/\tau_0\kappa_0 + 2/\tau_0 C_1\kappa)\kappa^t(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2\|_2) \\
&\quad + (\kappa_0\frac{1-\kappa^{t+1}}{1-\kappa}C^* + 4/\tau_0 C_{con} + 2/\tau_0\frac{1-\kappa^{t+2}}{1-\kappa}C_\lambda)\sqrt{\frac{s\log(n)^2\log(p)}{n}} \\
&\le \kappa^{t+1}(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2\|_2) + \frac{1-\kappa^{t+2}}{1-\kappa}C^*\sqrt{\frac{s\log(n)^2\log(p)}{n}}.
\end{aligned}
$$

We complete the induction and have

$$
|\widehat{\omega}_1^{(t)} - \omega_1^*| \vee \|\widehat{\beta}_k^{(t)} - \beta_k\|_2 \le \kappa^t(|\widehat{\omega}_1^{(0)} - \omega_1^*| \vee \|\widehat{\beta}_1^{(0)} - \beta_1^*\|_2 \vee \|\widehat{\beta}_2^{(0)} - \beta_2^*\|_2) + \frac{1-\kappa^{t+1}}{1-\kappa}C^*\sqrt{\frac{s\log(n)^2\log(p)}{n}}.
$$

Appendix G. Proof of Theorem 5

In general, using $\sigma^2 \neq \sigma_*^2$ makes the algorithm return a biased estimation. We consider how large the bias is. When $\sigma^2 \neq \sigma_*^2$, the bias of the estimation is caused by $\rho_k(\theta^*) \neq \Sigma(\theta^*)\beta_k^*$. Thus, in the proof, we will give a upper bound for $\|\rho_k(\theta^*) - \Sigma_k(\theta^*)\beta^*\|_2$. Let $f(\cdot \mid X_i^T\beta_k^*, \sigma_*^2)$ be the probability density function for $N(X_i^T\beta_k^*, \sigma_*^2)$. We first treat X_i, $i = 1, \cdots, n$ as fixed, and will take expecation for the conditional upper bound with respect to X_i in the end of the proof. Recall that

$$
\rho_1(\theta^*) = \int_{-\infty}^{\infty} \frac{\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma^2)}{\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma^2) + \omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma^2)}\big(\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma_*^2) + \omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma_*^2)\big)X_i Y_i dY_i.
$$

$$
\Sigma_1(\theta^*) = \int_{-\infty}^{\infty} \frac{\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma^2)}{\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma^2) + \omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma^2)}\big(\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma_*^2) + \omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma_*^2)\big)X_i X_i^T dY_i.
$$

We consider the following term

$$
\begin{aligned}
I &= \int_{-\infty}^{\infty} \frac{\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma^2)}{\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma^2) + \omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma^2)} \big(\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma_*^2) + \omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma_*^2)\big) \\
&\quad \cdot (Y_i - X_i^T\beta_1^*) dY_i \\
&= \int_{-\infty}^{\infty} \frac{\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma^2)}{\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma^2) + \omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma^2)} \omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma_*^2)(Y_i - X_i^T\beta_1^*) dY_i \\
&\quad + \int_{-\infty}^{\infty} \frac{\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma^2)}{\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma^2) + \omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma^2)} \omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma_*^2)(Y_i - X_i^T\beta_1^*) dY_i \\
&= \int_{-\infty}^{\infty} \omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma_*^2)(Y_i - X_i^T\beta_1^*) dY_i \\
&\quad - \int_{-\infty}^{\infty} \frac{\omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma^2)}{\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma^2) + \omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma^2)} \omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma_*^2)(Y_i - X_i^T\beta_1^*) dY_i \\
&\quad + \int_{-\infty}^{\infty} \frac{\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma^2)}{\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma^2) + \omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma^2)} \omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma_*^2)(Y_i - X_i^T\beta_1^*) dY_i \\
&= \int_{-\infty}^{\infty} \frac{\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma^2)}{\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma^2) + \omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma^2)} \omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma_*^2)(Y_i - X_i^T\beta_1^*) dY_i \\
&\quad - \int_{-\infty}^{\infty} \frac{\omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma^2)}{\omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma^2) + \omega_2^* f(Y_i \mid X_i^T\beta_2^*, \sigma^2)} \omega_1^* f(Y_i \mid X_i^T\beta_1^*, \sigma_*^2)(Y_i - X_i^T\beta_1^*) dY_i \\
&\overset{\tilde{Z}_i = Y_i - X_i^T\beta_1^*}{=} \omega_1^*\omega_2^* \int_{-\infty}^{\infty} \frac{f(\tilde{Z} \mid 0, \sigma^2) f(\tilde{Z} \mid \delta, \sigma_*^2) - f(\tilde{Z} \mid 0, \sigma_*^2) f(Z_i \mid \delta, \sigma^2)}{\omega_1^* f(\tilde{Z} \mid 0, \sigma^2) + \omega_2^* f(\tilde{Z} \mid \delta, \sigma^2)} \tilde{Z} d\tilde{Z} \\
&\overset{Z_i = \tilde{Z}_i - \delta/2}{=} \omega_1^*\omega_2^* \int_{-\infty}^{\infty} \frac{f(Z_i \mid -\delta/2, \sigma^2) f(Z_i \mid \delta/2, \sigma_*^2) - f(Z_i \mid -\delta/2, \sigma_*^2) f(Z_i \mid \delta/2, \sigma^2)}{\omega_1^* f(Z_i \mid -\delta/2, \sigma^2) + \omega_2^* f(Z_i \mid \delta/2, \sigma^2)} (Z_i + \delta/2) dZ_i
\end{aligned}
$$

where $\delta = X_i^T(\beta_2^* - \beta_1^*)$. Without loss of generality, we assume that $\delta > 0$. We decompose I as follows.

$$
\begin{aligned}
I &= \omega_1^*\omega_2^* \int_0^{\infty} \frac{f(Z_i \mid -\delta/2, \sigma^2) f(Z_i \mid \delta/2, \sigma_*^2) - f(Z_i \mid -\delta/2, \sigma_*^2) f(Z_i \mid \delta/2, \sigma^2)}{\omega_1^* f(Z_i \mid -\delta/2, \sigma^2) + \omega_2^* f(Z_i \mid \delta/2, \sigma^2)} Z_i dZ_i \\
&\quad + \omega_1^*\omega_2^* \int_{-\infty}^{0} \frac{f(Z_i \mid -\delta/2, \sigma^2) f(Z_i \mid \delta/2, \sigma_*^2) - f(Z_i \mid -\delta/2, \sigma_*^2) f(Z_i \mid \delta/2, \sigma^2)}{\omega_1^* f(Z_i \mid -\delta/2, \sigma^2) + \omega_2^* f(Z_i \mid \delta/2, \sigma^2)} Z_i dZ_i \\
&\quad + \omega_1^*\omega_2^* \int_0^{\infty} \frac{f(Z_i \mid -\delta/2, \sigma^2) f(Z_i \mid \delta/2, \sigma_*^2) - f(Z_i \mid -\delta/2, \sigma_*^2) f(Z_i \mid \delta/2, \sigma^2)}{\omega_1^* f(Z_i \mid -\delta/2, \sigma^2) + \omega_2^* f(Z_i \mid \delta/2, \sigma^2)} \delta/2 dZ_i \\
&\quad + \omega_1^*\omega_2^* \int_{-\infty}^{0} \frac{f(Z_i \mid -\delta/2, \sigma^2) f(Z_i \mid \delta/2, \sigma_*^2) - f(Z_i \mid -\delta/2, \sigma_*^2) f(Z_i \mid \delta/2, \sigma^2)}{\omega_1^* f(Z_i \mid -\delta/2, \sigma^2) + \omega_2^* f(Z_i \mid \delta/2, \sigma^2)} \delta/2 dZ_i \\
&= I_1 + I_2 + I_3 + I_4.
\end{aligned}
$$

Note that the signs for I_1, I_2 and I_3 are the same, and the sign for I_4 are different from them. We have

$$|I| \leq |I_1 + I_2 + I_3| + |I_4|.$$

Next, we consider I_1 to I_4 separately. For I_1, we have

$$
\begin{aligned}
|I_1| &\le \omega_1^*\omega_2^*\Big|\int_0^\infty \frac{f(Z_i \mid -\delta/2,\sigma^2)f(Z_i \mid \delta/2,\sigma_*^2) - f(Z_i \mid -\delta/2,\sigma_*^2)f(Z_i \mid \delta/2,\sigma^2)}{\omega_2^* f(Z_i \mid \delta/2,\sigma^2)} Z_i dZ_i\Big| \\
&= \frac{\omega_1^*}{\sqrt{2\pi}\sigma_*}\int_0^\infty \Big\{\exp\Big(-\frac{(Z_i+\delta/2)^2}{2\sigma^2} - \frac{(Z_i-\delta/2)^2}{2\sigma_*^2} + \frac{(Z_i-\delta/2)^2}{2\sigma^2}\Big) - \exp\Big(-\frac{(Z_i+\delta/2)^2}{2\sigma_*^2}\Big)\Big\} Z_i dZ_i \\
&= \frac{\omega_1^*}{\sqrt{2\pi}\sigma_*}\Big|\int_0^\infty \Big\{\exp\Big(-\frac{(Z_i-\tilde{\Delta}\delta/2)^2}{2\sigma_*^2}\Big)\cdot\exp\Big(-\frac{\delta^2(1-\tilde{\Delta}^2)}{8\sigma_*^2}\Big)\Big\} Z_i dZ_i - \int_0^\infty \exp\Big(-\frac{(Z_i+\delta/2)^2}{2\sigma_*^2}\Big)\Big| \\
&= \frac{\omega_1^*}{\sqrt{2\pi}\sigma_*}\Big|\tilde{\Delta}\delta/2\int_{-\tilde{\Delta}\delta/2}^\infty \exp(-\frac{x^2}{2\sigma_*^2})dx\cdot\exp\Big(-\frac{\delta^2(1-\tilde{\Delta}^2)}{8\sigma_*^2}\Big) \\
&\quad + \tilde{\Delta}\delta/2\int_{-\tilde{\Delta}\delta/2}^\infty \exp(-\frac{x^2}{2\sigma_*^2})x dx\cdot\exp\Big(-\frac{\delta^2(1-\tilde{\Delta}^2)}{8\sigma_*^2}\Big) - \int_{\delta/2}^\infty \exp(-\frac{x^2}{2\sigma_*^2})x dx + \delta/2\int_{\delta/2}^\infty \exp(-\frac{x^2}{2\sigma_*^2})dx\Big| \\
&= \frac{\omega_1^*}{\sqrt{2\pi}\sigma_*}\Big|\tilde{\Delta}\delta/2\int_{-\tilde{\Delta}\delta/2}^\infty \exp(-\frac{x^2}{2\sigma_*^2})dx\cdot\exp\Big(-\frac{\delta^2(1-\tilde{\Delta}^2)}{8\sigma_*^2}\Big) + \delta/2\int_{\delta/2}^\infty \exp(-\frac{x^2}{2\sigma_*^2})dx\Big|
\end{aligned}
$$

where $\tilde{\Delta} = 1 - 2\sigma_*^2/\sigma^2$. Similarly, we have

$$
\begin{aligned}
|I_2| &\le \omega_1^*\omega_2^*\Big|\int_{-\infty}^0 \frac{f(Z_i \mid -\delta/2,\sigma^2)f(Z_i \mid \delta/2,\sigma_*^2) - f(Z_i \mid -\delta/2,\sigma_*^2)f(Z_i \mid \delta/2,\sigma^2)}{\omega_1^* f(Z_i \mid -\delta/2,\sigma^2)} Z_i dZ_i\Big| \\
&= \frac{\omega_2^*}{\sqrt{2\pi}\sigma_*}\Big|\tilde{\Delta}\delta/2\int_{-\tilde{\Delta}\delta/2}^\infty \exp(-\frac{x^2}{2\sigma_*^2})dx\cdot\exp\Big(-\frac{\delta^2(1-\tilde{\Delta}^2)}{8\sigma_*^2}\Big) + \delta/2\int_{\delta/2}^\infty \exp(-\frac{x^2}{2\sigma_*^2})dx\Big|
\end{aligned}
$$

and

$$
\begin{aligned}
|I_3| &\le \omega_1^*\omega_2^*\delta/2\Big|\int_0^\infty \frac{f(Z_i \mid -\delta/2,\sigma^2)f(Z_i \mid \delta/2,\sigma_*^2) - f(Z_i \mid -\delta/2,\sigma_*^2)f(Z_i \mid \delta/2,\sigma^2)}{\omega_2^* f(Z_i \mid \delta/2,\sigma^2)} dZ_i\Big| \\
&= \frac{\omega_1^*\delta}{2\sqrt{2\pi}\sigma_*}\Big|\int_{-\tilde{\Delta}\delta/2}^\infty \exp(-\frac{x^2}{2\sigma_*^2})dx\cdot\exp\Big(-\frac{\delta^2(1-\tilde{\Delta}^2)}{8\sigma_*^2}\Big) - \int_{\delta/2}^\infty \exp(-\frac{x^2}{2\sigma_*^2})dx\Big|
\end{aligned}
$$

and

$$
\begin{aligned}
|I_4| &\le \omega_1^*\omega_2^*\delta/2\Big|\int_0^\infty \frac{f(Z_i \mid -\delta/2,\sigma^2)f(Z_i \mid \delta/2,\sigma_*^2) - f(Z_i \mid -\delta/2,\sigma_*^2)f(Z_i \mid \delta/2,\sigma^2)}{\omega_1^* f(Z_i \mid -\delta/2,\sigma^2)} dZ_i\Big| \\
&= \frac{\omega_2^*\delta}{2\sqrt{2\pi}\sigma_*}\Big| - \int_{-\tilde{\Delta}\delta/2}^\infty \exp(-\frac{x^2}{2\sigma_*^2})dx\cdot\exp\Big(-\frac{\delta^2(1-\tilde{\Delta}^2)}{8\sigma_*^2}\Big) + \int_{\delta/2}^\infty \exp(-\frac{x^2}{2\sigma_*^2})dx\Big|
\end{aligned}
$$

Recall that the signs for I_1, I_2 and I_3 are the same, and the sign for I_4 are different from them, we have

$$\begin{aligned}
|I| \leq & \Big|\frac{1}{2\sqrt{2\pi}\sigma_*^2}\Big(\tilde{\Delta}\delta\int_{-\tilde{\Delta}\delta/2}^{\infty}\exp(-\frac{x^2}{2\sigma_*^2})dx\cdot\exp\big(-\frac{\delta^2(1-\tilde{\Delta}^2)}{8\sigma_*^2}\big)+\delta\int_{\delta/2}^{\infty}\exp(-\frac{x^2}{2\sigma_*^2})dx\Big)\\
&+\frac{\delta}{2\sqrt{2\pi}\sigma_*^2}\Big(\int_{-\tilde{\Delta}\delta/2}^{\infty}\exp(-\frac{x^2}{2\sigma_*^2})dx\cdot\exp\big(-\frac{\delta^2(1-\tilde{\Delta}^2)}{8\sigma_*^2}\big)-\int_{\delta/2}^{\infty}\exp(-\frac{x^2}{2\sigma_*^2})dx\Big)\Big|\\
=&\Big|\frac{1}{2\sqrt{2\pi}\sigma_*^2}(1+\tilde{\Delta})\delta\int_{-\tilde{\Delta}\delta/2}^{\infty}\exp(-\frac{x^2}{2\sigma_*^2})dx\cdot\exp\big(-\frac{\delta^2(1-\tilde{\Delta}^2)}{8\sigma_*^2}\big)\Big|\\
\overset{t=x/\sigma_*}{=}&\frac{\delta}{\sqrt{2\pi}}\big|1-\frac{\sigma_*^2}{\sigma^2}\big|\cdot\int_{-\frac{\tilde{\Delta}\delta}{2\sigma_*}}^{\infty}\exp(-t^2/2)dt\cdot\exp\big(-\frac{\delta^2(1-\tilde{\Delta}^2)}{8\sigma_*^2}\big)
\end{aligned}$$

Using the inequality that $\int_t^{\infty}\exp(-x^2/2)dx\leq\frac{1}{t}\exp(-t^2/2)$ when $t>0$, when $\sigma^2<2\sigma_*^2$, we have

$$|I|\leq\frac{2\sigma_*}{\sqrt{2\pi}(2\sigma_*^2/\sigma^2-1)}\cdot\big|1-\frac{\sigma_*^2}{\sigma^2}\big|\cdot\exp(-\frac{\delta^2}{8\sigma_*^2})$$

Let H be an orthogonal matrix whose first row is $(\beta_2^*-\beta_1^*)\Sigma^{1/2}/\|(\beta_2^*-\beta_1^*)\Sigma^{1/2}\|_2$. We write X_i as $\Sigma^{1/2}H^TV_i$, where $V_i\sim N(0,I_p)$. Then $\delta=(\beta_2^*-\beta_1^*)^TX_i=\|(\beta_2^*-\beta_1^*)\Sigma^{1/2}\|_2V_{i1}=\Delta V_{i1}$. Note that

$$\begin{aligned}
\|\rho_1(\theta^*)-\Sigma_1(\theta^*)\beta_1^*\|_2&=\|\mathrm{E}_{X_i}(IX_i)\|_2\\
&\leq\|\Sigma^{1/2}H^T\|_2\|\mathrm{E}_{V_i}(IV_i)\|_2\\
&=\|\Sigma^{1/2}H^T\|_2\|\mathrm{E}_{V_{i1}}(IV_{i1})\|_2
\end{aligned}$$

We have the last equality because when $j\geq2$, $\mathrm{E}_{V_{ij}}(IV_{ij})=0$.

Then we have

$$\begin{aligned}
\|\mathrm{E}_{V_{i1}}(IV_{i1})\|_2&\leq\mathrm{E}(|I||V_{i1}|)\\
&\leq\frac{2\sigma^*}{\pi(2\sigma_*^2/\sigma^2-1)}\cdot\big|1-\frac{\sigma_*^2}{\sigma^2}\big|\cdot\int_0^{\infty}\exp(-\frac{\Delta^2V_{i1}^2}{8\sigma_*^2})\exp(-V_{i1}^2/2)V_{i1}dV_{i1}\\
&\leq\frac{2\sigma^*}{\pi(\sigma_*^2/\sigma^2-1)}\cdot\big|1-\frac{\sigma_*^2}{\sigma^2}\big|\cdot\frac{4\sigma_*^2}{\Delta^2}\\
&=\frac{8\sigma_*(\Delta/\sigma_*)^{-2}}{\pi(2\sigma_*^2/\sigma^2-1)}\cdot\big|1-\frac{\sigma_*^2}{\sigma^2}\big|.
\end{aligned}$$

It follows that

$$M_{\text{bias}}=\|\rho_1(\theta^*)-\Sigma_1(\theta^*)\beta_1^*\|_2\leq\frac{8\|\Sigma^{1/2}\|_2\sigma_*(\Delta/\sigma_*)^{-2}}{\pi(2\sigma_*^2/\sigma^2-1)}\cdot\big|1-\frac{\sigma_*^2}{\sigma^2}\big|$$

To give a complete version of the proof, we first go back to the definitions of $M(\theta)$ and $M_n(\theta)$. When using σ^2 in the algorithm, the formula of $\widetilde{\eta}_{i,k}(\theta)$ is changed. Thus, we redefine terms

$$\widetilde{\eta}_{i,1}(\theta)=1-\widetilde{\eta}_{i,2}(\theta)=1/\big[1+(\omega_2/\omega_1)\exp\{(\beta_2-\beta_1)^TX_i\cdot(Y_i-(\beta_1+\beta_2)^TX_i/2)/\sigma^2\}\big],$$

and

$$\widetilde{\omega}_k(\theta) = \frac{1}{n}\sum_{i=1}^n \widetilde{\eta}_{i,k}(\theta),\ \widetilde{\rho}_k(\theta) = \frac{1}{n}\sum_{i=1}^n \widetilde{\eta}_{i,k}(\theta)X_iY_i,$$
$$\widetilde{\Sigma}_k(\theta) = \frac{1}{n}\sum_{i=1}^n \widetilde{\eta}_{i,k}(\theta)X_iX_i^T,\ \omega_k(\theta) = \mathrm{E}\big\{\frac{1}{n}\sum_{i=1}^n \widetilde{\eta}_{i,k}(\theta)\big\},$$
$$\rho_k(\theta) = \mathrm{E}\big\{\frac{1}{n}\sum_{i=1}^n \widetilde{\eta}_{i,k}(\theta)X_iY_i\big\},\ \Sigma_k(\theta) = \mathrm{E}\big\{\frac{1}{n}\sum_{i=1}^n \widetilde{\eta}_{i,k}(\theta)X_iX_i^T\big\},$$

where the expectation is with respect to X_i and Y_i, $i = 1, \ldots, n$. Then we let $\widetilde{M}(\theta) = \{\omega_k(\theta), \rho_k(\theta), \Sigma_k(\theta), k = 1, 2\}$, $\widetilde{M}_n(\theta) = \{\widetilde{\omega}_k(\theta), \widetilde{\rho}_k(\theta), \widetilde{\Sigma}_k(\theta), k = 1, 2\}$. For $\widetilde{M}$ and $\widetilde{M}_n$, we have two similar Lemmas as Lemmas 2 and 3:

Lemma A.10 *Under conditions (C1) and (C3), if $\theta \in \mathcal{B}_{con}(\theta^*)$, then*

$$d_2\big(\widetilde{M}(\theta), \widetilde{M}(\theta^*)\big) \le \kappa_0\big(|\omega_1(\theta) - \omega_1^*| \vee \|\beta_1 - \beta_1^*\|_2 \vee \|\beta_2 - \beta_2^*\|_2\big).$$

for some $0 < \kappa_0 < \frac{1}{2\vee(64/\tau_0)}$.

Lemma A.11 *Suppose that $\theta^* \in \Theta^*$. Under condition (C1), there exist a constant $C_{con} > 0$, such that with probability at least $1 - 4p^{-1}$,*

$$\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} d_{2,s}(\widetilde{M}(\theta), \widetilde{M}_n(\theta)) \le C_{con}\sigma_*\sqrt{\frac{s\log(n)^2\log(p)}{n}}$$

The proofs of those two Lemmas are almost identical to those for Lemmas 2 and 3. We omit the details here. We remark that $\widetilde{M}(\theta^*)$ is no longer $(\omega_1^*, \omega_1^*\Sigma\beta_1^*, \omega_1^*\Sigma)$ because of the bias caused by using σ^2. We have the following concentration results.

Lemma A.12 *Suppose that $\theta^* \in \Theta^*$ and $\widetilde{\beta}_k^{(0)} \in \mathcal{B}_{con}$. Let $\lambda_n^{(t+1)} \ge 4C_{con}(\sqrt{\log(n)^2\log(p)/n} + 8\kappa_0(|\widetilde{\omega}_k^{(t)} - \omega_k^*| \vee \|\widetilde{\beta}_k^{(t)} - \beta_k^*\|_2)/\sqrt{s}) + 4M_{bias}/\sqrt{s}$ and $\widetilde{\beta}_k^{(t+1)}$ be solved by*

$$(\widetilde{\beta}_1^{(t+1)}, \cdots, \widetilde{\beta}_k^{(t+1)}) = \underset{\beta_1,\cdots,\beta_k}{\operatorname{argmin}}\Big\{\sum_{k=1}^K \beta_k^T\widetilde{\Sigma}_k^{(t+1)}\beta_k - 2\sum_{k=1}^K(\widetilde{\rho}_k^{(t+1)})^T\beta_k + \lambda^{(t+1)}\sum_{j=1}^p\sqrt{\sum_{k=1}^K\beta_{kj}^2}\Big\},$$

where $\widetilde{\rho}_k^{(t+1)} = (\frac{1}{n}\sum_{i=1}^n \widetilde{\eta}_{ik}^{(t+1)}Y_iX_i^T)^T$, $\widetilde{\Sigma}_k^{(t+1)} = (\frac{1}{n}\sum_{i=1}^n \widetilde{\eta}_{ik}^{(t+1)}X_iX_i^T)$, and $\widetilde{\eta}_{i1}^{(t+1)} = \widetilde{\omega}_1^{(t)}/(\widetilde{\omega}_1^{(t)} + \widetilde{\omega}_2^{(t)}\exp((\widetilde{\beta}_2^{(t)} - \widetilde{\beta}_1^{(t)})^TX_i(Y_i - X_i^T)(\widetilde{\beta}_2^{(t)} + \widetilde{\beta}_1^{(t)})^T/2)/\sigma^2)$. We have

$$\widetilde{\beta}_k^{(t+1)} - \beta_k^* \in \Gamma(s),$$

and

$$\|\widetilde{\beta}^{(t+1)} - \beta^*\|_2 \le \frac{4}{\tau_0}\big(d_{2,s}(\widetilde{M}_n(\widetilde{\theta}^{t+1}), \widetilde{M}(\theta^*)) + M_{bias}\big) + \frac{2}{\tau_0}\sqrt{s}\lambda_n^{(t+1)}.$$

Proof Because $(\widetilde{\beta}_1^{(t+1)}, \cdots, \widetilde{\beta}_k^{(t+1)})$ is the minimizer of

$$\sum_{k=1}^{K} \beta_k^T \widetilde{\Sigma}_k^{(t+1)} \beta_k - 2 \sum_{k=1}^{K} (\widetilde{\rho}_k^{(t+1)})^T \beta_k + \lambda^{(t+1)} \sum_{j=1}^{p} \sqrt{\sum_{k=1}^{K} \beta_{kj}^2},$$

we have

$$\begin{aligned}
&\lambda^{(t+1)} \Big(\sum_{j=1}^{p} \sqrt{\sum_{k=1}^{2} (\widetilde{\beta}_{kj}^{(t+1)})^2} - \sum_{j=1}^{p} \sqrt{\sum_{k=1}^{2} (\beta_{kj}^*)^2} \Big) \\
&\leq 2 \sum_{k=1}^{2} \Big\{ (\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T \widetilde{\Sigma}_k^{(t+1)} \beta_k^* - (\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T \widetilde{\Sigma}_k^{(t+1)} (\beta_k^* - \widetilde{\beta}_k^{(t+1)})/2 - (\widetilde{\rho}_k^{(t+1)})^T (\beta_k^* - \widetilde{\beta}_k^{(t+1)}) \Big\} \\
&\leq 2 \sum_{k=1}^{2} \Big\{ (\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T \widetilde{\Sigma}_k^{(t+1)} \beta_k^* - (\widetilde{\rho}_k^{(t+1)})^T (\beta_k^* - \widetilde{\beta}_k^{(t+1)}) \Big\} \\
&= 2 \sum_{k=1}^{2} \Big\{ (\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T \big[\widetilde{\Sigma}_k^{(t+1)} \beta_k^* - \widetilde{\rho}_k^{(t+1)} \big] \Big\} \\
&= 2 \sum_{k=1}^{2} \Big\{ (\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T \big[(\widetilde{\Sigma}_k^{(t+1)} - \Sigma_k^{(t+1)}) \beta_k^* + \Sigma_k^{(t+1)} \beta_k^* - \rho_k^{(t+1)} - (\widetilde{\rho}_k^{(t+1)} - \rho_k^{(t+1)}) \big] \Big\} \\
&= 2 \sum_{k=1}^{2} \Big\{ (\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T \big[(\widetilde{\Sigma}_k^{(t+1)} - \Sigma_k^{(t+1)}) \beta_k^* - (\widetilde{\rho}_k^{(t+1)} - \rho_k^{(t+1)}) \big] \Big\} \\
&\quad + 2 \sum_{k=1}^{2} \Big\{ (\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T \big[\Sigma_k^{(t+1)} \beta_k^* - \rho_k^{(t+1)} \big] \Big\} \\
&= 2 \sum_{k=1}^{2} \Big\{ (\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T \big[(\widetilde{\Sigma}_k^{(t+1)} - \Sigma_k^{(t+1)}) \beta_k^* - (\widetilde{\rho}_k^{(t+1)} - \rho_k^{(t+1)}) \big] \Big\} \\
&\quad + 2 \sum_{k=1}^{2} \Big\{ (\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T \big[(\Sigma_k^{(t+1)} - \Sigma_k(\theta^*)) \beta_k^* + \Sigma_k(\theta^*) \beta_k^* - \rho_k(\theta^*) - (\rho_k^{(t+1)} - \rho_k(\theta^*)) \big] \Big\} \\
&\leq 2 \sum_{k=1}^{2} \Big\{ (\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T \big[(\widetilde{\Sigma}_k^{(t+1)} - \Sigma_k^{(t+1)}) \beta_k^* - (\widetilde{\rho}_k^{(t+1)} - \rho_k^{(t+1)}) \big] \Big\} \\
&\quad + 4\kappa_0 \frac{|\widehat{\omega}_1^{(t)} - \omega_1| \vee \|\widetilde{\beta}_1^{(t)} - \beta_1^*\|_2 \vee \|\widetilde{\beta}_2^{(t)} - \beta_2^*\|_2}{\sqrt{s}} \sqrt{s} \sum_{k=1}^{2} \|\widetilde{\beta}_k^{(t+1)} - \beta_k^*\|_2 + 2 \frac{M_{bias}}{\sqrt{s}} \sqrt{s} \sum_{k=1}^{2} \|\widetilde{\beta}_k^{(t+1)} - \beta_k^*\|_2.
\end{aligned}$$

Let $u_k^{(t+1)} = \widetilde{\beta}_k^{(t+1)} - \beta_k^*$, we have

$$\begin{aligned}\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(\widetilde{\beta}_{kj}^{(t+1)})^2} - \sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(\beta_{kj}^*)^2} &= \sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(\beta_{kj}^* + u_{kj}^{(t+1)})^2} - \sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(\beta_{kj}^*)^2} \\ &= \sum_{j\in S}\sqrt{\sum_{k=1}^{2}(\beta_{kj}^* + u_{kj}^{(t+1)})^2} + \sum_{j\in S^c}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2} - \sum_{j\in S}\sqrt{\sum_{k=1}^{2}(\beta_{kj}^*)^2} \\ &\geq \sum_{j\in S^c}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2} - \sum_{j\in S}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2}.\end{aligned}$$

Let $\widetilde{S}$ be a set of size $2s$, which contains S and the largest s coefficients of $\widehat{\rho}_k^{(t+1)} - \rho_k^{(t+1)}$. We have

$$\begin{aligned}&|(\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T(\widetilde{\rho}_k^{(t+1)} - \rho_k^{(t+1)})| \\ &\leq \|\widetilde{\rho}_k^{(t+1)} - \rho_k^{(t+1)}\|_{2,s}\|(\beta_k^* - \widetilde{\beta}_k^{(t+1)})_{\widetilde{S}}\|_2 + \|(\widetilde{\rho}_k^{(t+1)} - \rho_k^{(t+1)})_{\widetilde{S}^c}\|_\infty\|(\beta_k^* - \widetilde{\beta}_k^{(t+1)})_{\widetilde{S}^c}\|_1 \\ &\leq C_{con}\sqrt{\log(n)^2\log(p)/n}\sqrt{s}\|(\beta_k^* - \widetilde{\beta}_k^{(t+1)})_{\widetilde{S}}\|_2 + C_{con}\sqrt{\log(n)^2\log(p)/n}\|(\beta_k^* - \widetilde{\beta}_k^{(t+1)})_{\widetilde{S}^c}\|_1,\end{aligned}$$

where in the last inequality, we use Lemma 2 and the fact that $\|(\widetilde{\rho}_k^{(t+1)} - \rho_k^{(t+1)})_{\widetilde{S}^c}\|_\infty \leq \|(\widetilde{\rho}_k^{(t+1)} - \rho_k^{(t+1)})_{\widetilde{S}}\|_2/\sqrt{s} \leq \|\widetilde{\rho}_k^{(t+1)} - \rho_k^{(t+1)}\|_{2,s}/\sqrt{s}$. We have the same results for the term $|(\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T(\widetilde{\Sigma}_k^{(t+1)} - \Sigma_k^{(t+1)})\beta_k^*|$.

Also, note that $\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2} \leq \sum_{k=1}^{2}\|u_k^{(t+1)}\|_1 \leq 2\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2}$, we have

$$\begin{aligned}&\lambda_n^{(t+1)}\Big\{\sum_{j\in S^c}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2} - \sum_{j\in S}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2}\Big\} \\ &\geq \lambda_n^{(t+1)}\Big\{\sum_{j\in \widetilde{S}^c}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2} - \sum_{j\in \widetilde{S}}\sqrt{\sum_{k=1}^{2}(u_{kj}^{(t+1)})^2}\Big\} \\ &\geq \lambda_n^{(t+1)}\Big\{\frac{1}{2}\sum_{k=1}^{2}\|(\mu_k^{(t+1)})_{\widetilde{S}^c}\|_1 - \sum_{k=1}^{2}\sqrt{s}\|(\mu_k^{(t+1)})_{\widetilde{S}}\|_2\Big\}.\end{aligned}$$

Then

$$\begin{aligned}&\lambda_n^{(t+1)}\Big\{\frac{1}{2}\sum_{k=1}^{2}\|(\mu_k^{(t+1)})_{\widetilde{S}^c}\|_1 - \sum_{k=1}^{2}\sqrt{s}\|(\mu_k^{(t+1)})_{\widetilde{S}}\|_2\Big\} \\ &\leq C_{con}\sqrt{\log(n)^2\log(p)/n}\sqrt{s}\sum_{k=1}^{2}\|(\mu_k)_{\widetilde{S}}\|_2 + C_{con}\sqrt{\log(n)^2\log(p)/n}\sum_{k=1}^{2}\|(\mu_k)_{\widetilde{S}^c}\|_1 \\ &\quad + 4\kappa_0\frac{|\widetilde{\omega}_k^{(t)} - \omega_k| \vee \|\widetilde{\beta}_k^{(t)} - \beta_k^*\|_2}{\sqrt{s}}\sqrt{s}\sum_{k=1}^{2}\|\mu_k^{(t+1)}\|_2 + 2\frac{M_{bias}}{\sqrt{s}}\sqrt{s}\sum_{k=1}^{2}\|\mu_k^{(t+1)}\|_2.\end{aligned}$$

Let $\lambda_n^{(t+1)} \geq 4C_{con}\sqrt{\log(n)^2\log(p)/n} + 8\kappa_0 \frac{|\widehat{\omega}_1^{(t)}-\omega_1|\vee\|\widetilde{\beta}_1^{(t)}-\beta_1^*\|_2\vee\|\widetilde{\beta}_2^{(t)}-\beta_2^*\|_2}{\sqrt{s}} + \frac{4M_{bias}}{\sqrt{s}}$, we have

$$\sum_{k=1}^{2}\|(u_k^{(t+1)})_{\widetilde{S}^c}\|_1 \leq 5\sum_{k=1}^{2}\sqrt{s}\|(u_k^{(t+1)})_{\widetilde{S}}\|_2 + 4\sqrt{s}\sum_{k=1}^{2}\|\mu_k^{(t+1)}\|_2.$$

Let $\mu^T = (\mu_1^T, \mu_2^T)^T$, and $S_1 = \{\widetilde{S}, \widetilde{S}+p\}$, where $\widetilde{S}+p$ means the collection of the index in $\widetilde{S}$ adds p. We have

$$\|(u^{(t+1)})_{S_1^c}\|_1 \leq 5\sqrt{2s}\|(u^{(t+1)})_{S_1}\|_2 + 4\sqrt{2s}\|\mu^{(t+1)}\|_2.$$

Next, we prove the second conclusion. Note that

$$\begin{aligned}
&\lambda^{(t+1)}\Big(\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(\widetilde{\beta}_{kj}^{(t+1)})^2} - \sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(\beta_{kj}^*)^2}\Big)\\
&\leq 2\sum_{k=1}^{2}\Big\{(\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T\widetilde{\Sigma}_k^{(t+1)}\beta_k^* - (\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T\widetilde{\Sigma}_k^{(t+1)}(\beta_k^* - \widetilde{\beta}_k^{(t+1)})/2 - (\widetilde{\rho}_k^{(t+1)})^T(\beta_k^* - \widetilde{\beta}_k^{(t+1)})\Big\}
\end{aligned}$$

It follows that

$$\begin{aligned}
\sum_{k=1}^{2}(\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T\widetilde{\Sigma}_k^{(t+1)}(\beta_k^* - \widetilde{\beta}_k^{(t+1)}) &\leq |2\sum_{k=1}^{2}\Big\{(\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T\widetilde{\Sigma}_k^{(t+1)}\beta_k^* - (\widetilde{\rho}_k^{(t+1)})^T(\beta_k^* - \widetilde{\beta}_k^{(t+1)})\Big\}|\\
&\quad + \lambda_n^{(t+1)}\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{2}(\widetilde{\beta}_{kj}^{(t+1)} - \beta_{kj})^2}\\
&\leq |2\sum_{k=1}^{2}\Big\{(\beta_k^* - \widetilde{\beta}_k^{(t+1)})^T\widetilde{\Sigma}_k^{(t+1)}\beta_k^* - (\widetilde{\rho}_k^{(t+1)})^T(\beta_k^* - \widetilde{\beta}_k^{(t+1)})\Big\}|\\
&\quad + \lambda_n^{(t+1)}\sum_{k=1}^{2}\|\widetilde{\beta}_k^{(t+1)} - \beta_k^*\|_1
\end{aligned}$$

Recall that

$$\widetilde{\Sigma}_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\eta_{k,\widehat{\theta}^{(t)}}(Y_i, X_i)X_iX_i^T.$$

By Lemma 1, we know that

$$|\frac{1}{n}\sum_{i=1}^{n}\eta_{k,\widehat{\theta}^{(t)}}(Y_i, X_i) - \mathrm{E}(\widehat{\omega}_k^{(t)})| = o(\sqrt{s\log(p)/n}).$$

It follows that $\frac{1}{n}\sum_{i=1}^{n}\eta_{k,\widehat{\theta}^{(t)}}(Y_i, X_i) > \tau_1$ for some positive constant τ_1. Define set $\mathcal{N} = \{i : \eta_{k,\widehat{\theta}^{(t)}}(Y_i, X_i) > \tau_1/2\}$. We have $|\mathcal{N}|/n > \tau_1/(2-\tau_1)$. By Lemma A.4, we have $\|\widetilde{\Sigma}_k^{(t+1)}\|_{2,s} > \tau_0$ for some positive constant τ_0.

Then

$$\sum_{k=1}^{2}\|\beta_k^*-\widetilde{\beta}_k^{(t+1)}\|_2^2\tau_0\le 2\sum_{k=1}^{2}\Big\{(\beta_k^*-\widetilde{\beta}_k^{(t+1)})^T\widetilde{\Sigma}_k^{(t+1)}\beta_k^*-(\widetilde{\rho}_k^{(t+1)})^T(\beta_k^*-\widetilde{\beta}_k^{(t+1)})\Big\}|$$
$$+\lambda_n^{(t+1)}\sqrt{s}\sum_{k=1}\|\widetilde{\beta}_k^{(t+1)}-\beta_k^*\|_2$$

Hence,

$$(\sum_{k=1}^{2}\|\beta_k^*-\widetilde{\beta}_k^{(t+1)}\|_2)^2\cdot\tau_0/2\le 2\big(d_{2,s}(\widetilde{M}_n(\widetilde{\theta}^{t+1}),\widetilde{M}(\theta^*))+M_{bias}\big)\sum_{k=1}^{2}\|\beta_k^*-\widetilde{\beta}_k^{(t+1)}\|_2$$
$$+\lambda_n^{(t+1)}\sqrt{s}\sum_{k=1}^{2}\|\widetilde{\beta}_k^{(t+1)}-\beta_k^*\|_2$$

It follows that

$$\sum_{k=1}^{2}\|\widetilde{\beta}_k^{(t+1)}-\beta_k^*\|_2\le 4/\tau_0\big(d_{2,s}(\widetilde{M}_n(\widetilde{\theta}^{t+1}),\widetilde{M}(\theta^*))+M_{bias}\big)+2/\tau_0\sqrt{s}\lambda_n^{(t+1)}.$$

■

Theorem 5 then follows directly by Lemmas A.11, A.11, A.12, and analogous argument in Section F.2.

Appendix H. Proofs for Section 5

Note that Σ_y is known, the multivariate mixture linear regression model is equivalently to

$$\mathbb{P}(W_i=k)=\omega_k,\quad \check{Y}_i\mid(X_i,W_i=k)\sim N(\check{\beta}_k^TX,I_q),$$

where $\check{Y}_i=\Sigma_y^{1/2}Y_i$ and $\check{\beta}_k=\beta_k\Sigma_y^{1/2}$. Since Σ_y is positive definite, $\|(\beta_k-\beta_k^*)\|_F\eqsim\|(\beta_k-\beta_k^*)\Sigma_y^{1/2}\|_F=\|(\check{\beta}_k-\check{\beta}_k^*)\|_F$. In addition, since Σ_y is positive definite, the sparse pattern of $\check{\beta}_k$ is the same as β_k (The location of non-zero rows are the same for them). Thus, building the upper bound for $\check{\beta}_k$ is equivalent to building that for β_k. When we considering the equivalent model, the covariance matrix for Y_i is simplified to I_q. In the following part of this section, with a little abuse of notation, we write $\check{\beta}_k$ as β_k and $\check{Y}_i$ as Y_i for simplicity.

H.1 Proof of Lemma 6

The proof strategy is analogous to that for the mixture linear regression. However, due to the multi-dimensionality of the response, the proof is a little more complex. Since the proof for the contraction property of $\omega_1(\theta)$, $\rho_k(\theta)$ and $\Sigma_k(\theta)$ are similar to each other, we elaborate on the proof for $\rho_k(\theta)$, which is the most complex one, as an illustration. We aim to show that $\|\rho_1(\theta)-\rho_1(\theta^*)\|_F\le\kappa_0(|\omega_k-\omega_k^*|\vee\|\beta_1-\beta_1^*\|_F\vee\|\beta_2^*-\beta_2\|_F)$. Recall that

$$\rho_1(\theta)=\mathrm{E}(\frac{1}{n}\sum_{i=1}^{n}\eta_{1,\theta}(X_i,Y_i)X_iY_i^T)=\mathrm{E}(\eta_{1,\theta}(X_i,Y_i)X_iY_i^T),$$

$\xi = (\omega_1, \text{vec}(\beta_2 - \beta_1), \text{vec}(\beta_1 + \beta_2))$, $\Delta_\xi = \xi - \xi^*$, and $\xi_u = \xi^* + u\Delta_\xi$. Note that

$$
\begin{aligned}
&\text{vec}^T(\rho_1(\theta) - \rho_1(\theta^*)) \\
&= \text{E}\Big\{ \int_0^1 \Big(\frac{d\text{vec}(\eta_{1,\xi}(X_i, Y_i) X_i Y_i^T)}{d\text{vec}(\xi)}\Big|_{\xi=\xi_u} \Big)^T \Delta_\xi du \Big\} \\
&= \text{E}\Big\{ \int_0^1 \Big(\frac{\partial\text{vec}(\eta_{1,\xi}(X_i, Y_i) X_i Y_i^T)}{\partial\omega_1}\Big|_{\xi=\xi_u} \Big)^T \Delta_{\omega_1} du \Big\} + \text{E}\Big\{ \int_0^1 \Big(\frac{\partial\text{vec}(\eta_{1,\xi}(X_i, Y_i) X_i Y_i^T)}{\partial\text{vec}(\beta_2 - \beta_1)}\Big|_{\xi=\xi_u} \Big)^T \Delta_{\beta_2-\beta_1} du \Big\} \\
&\quad + \text{E}\Big\{ \int_0^1 \Big(\frac{\partial\text{vec}(\eta_{1,\xi}(X_i, Y_i) X_i Y_i^T)}{\partial\text{vec}(\beta_2 + \beta_1)}\Big|_{\xi=\xi_u} \Big)^T \Delta_{\beta_2+\beta_1} du \Big\} \\
&= \sup_{\xi\in\mathcal{B}_{con}(\theta^*)} \text{E}\Big(\frac{\partial\text{vec}(\eta_{1,\xi}(X_i, Y_i) X_i Y_i^T)}{\partial\omega_1}\Big|_{\xi=\xi_u} \Big)^T \Delta_{\omega_1} + \sup_{\xi\in\mathcal{B}_{con}(\theta^*)} \text{E}\Big(\frac{\partial\text{vec}(\eta_{1,\xi}(X_i, Y_i) X_i Y_i^T)}{\partial\text{vec}(\beta_2 - \beta_1)}\Big|_{\xi=\xi_u} \Big)^T \Delta_{\beta_2-\beta_1} \\
&\quad + \sup_{\xi\in\mathcal{B}_{con}(\theta^*)} \text{E}\Big(\frac{\partial\text{vec}(\eta_{1,\xi}(X_i, Y_i) X_i Y_i)}{\partial\text{vec}(\beta_2 + \beta_1)}\Big|_{\xi=\xi_u} \Big)^T \Delta_{\beta_2+\beta_1}
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\|\rho_1(\theta) - \rho_1^*\|_F \le & \sup_{\xi\in\mathcal{B}_{con}(\theta^*)} \|\text{E}(\frac{\partial\eta_{1,\theta}(X_i, Y_i)}{\partial\omega_1}\text{vec}(X_i Y_i^T))\|_2 |\omega_1 - \omega_1^*| \\
& + \sup_{\xi\in\mathcal{B}_{con}(\theta^*)} \|\text{E}(\frac{\partial\eta_{1,\theta}(X_i, Y_i)}{\partial\text{vec}^T(\beta_2 - \beta_1)}\text{vec}(X_i Y_i^T))\|_2 \|\beta_1 - \beta_1^* - \beta_2 + \beta_2^*\|_F \\
& + \sup_{\xi\in\mathcal{B}_{con}(\theta^*)} \|\text{E}(\frac{\partial\eta_{1,\theta}(X_i, Y_i)}{\partial\text{vec}^T(\beta_1 + \beta_2)}\text{vec}(X_i Y_i^T))\|_2 \|\beta_1 - \beta_1^* + \beta_2 - \beta_2^*\|_F.
\end{aligned}
$$

We show the bound for $\sup_{\xi\in\mathcal{B}_{con}(\theta^*)} \|\text{E}(\frac{\partial\eta_{1,\theta}(X_i,Y_i)}{\partial\text{vec}(\beta_2-\beta_1)}\text{vec}(X_i Y_i^T))\|_2 \|\beta_1 - \beta_1^* - \beta_2 + \beta_2^*\|_F$ as an illustration. The implementation of the bounds for the first and third terms on the right hand side of last inequality is similar. Recall that

$$
\begin{aligned}
\frac{\partial\eta_{1,\theta}(X_i, Y_i)}{\partial\text{vec}^T(\beta_2 - \beta_1)}\text{vec}(X_i Y_i^T) &= \omega_1(1-\omega_1)\Big(\frac{\exp\{X_i^T(\beta_2 - \beta_1)(Z_i + \delta_i(\beta)^T X_i)\}}{\big(\omega_1 + (1-\omega_1)\exp\{X_i^T(\beta_2 - \beta_1)(Z_i + \delta_i(\beta)^T X_i)\}\big)^2} \\
&\quad \cdot \text{vec}(X_i(Z_i + \delta(\beta)_i^T X_i))\text{vec}^T(X_i(Z_i + \psi_i^T X_i))\Big) \\
&= \omega_1(1-\omega_1)\Big(\frac{\exp\{X_i^T(\beta_2 - \beta_1)(Z_i + \delta_i(\beta)^T X_i)\}}{\big(\omega_1 + (1-\omega_1)\exp\{X_i^T(\beta_2 - \beta_1)(Z_i + \delta_i(\beta)^T X_i)\}\big)^2} \\
&\quad \cdot (I_q \otimes X_i)(Z_i + \delta(\beta)_i^T X_i)(Z_i + \psi_i^T X_i)^T (I_q \otimes X_i^T) \\
&= \omega_1(1-\omega_1)\Big(\frac{\exp\{X_i^T(\beta_2 - \beta_1)(Z_i + \delta_i(\beta)^T X_i)\}}{\big(\omega_1 + (1-\omega_1)\exp\{X_i^T(\beta_2 - \beta_1)(Z_i + \delta_i(\beta)^T X_i)\}\big)^2} \\
&\quad \cdot (I_q \otimes X_i)(Z_i + \delta(\beta)_i^T X_i)(Z_i + \delta(\beta)_i^T X_i - \frac{(\beta_2 + \beta_1)^T}{2} X_i)^T (I_q \otimes X_i^T),
\end{aligned}
$$

where $\mathbb{P}(\delta(\beta)_i = \beta_k^* - \frac{\beta_2+\beta_1}{2}) = \mathbb{P}(\psi = \beta_k^*) = \mathbb{P}(W_i = k) = \omega_k^*$. Let $H_k \in \mathbb{R}^{p\times p}$ be an othorgonal matrix satisfies that the first row is $\frac{(\Sigma^{1/2}(\beta_2-\beta_1))_{:1}}{\|(\Sigma^{1/2}(\beta_2-\beta_1))_{:1}\|_2}$, $(\Sigma^{1/2}(\beta_2 - \beta_1))_{:2} \in \text{span}(H_{k,1:}, H_{k,2:})$, $\cdots$, $(\Sigma^{1/2}(\beta_2 - \beta_1))_{:q} \in \text{span}(H_{k,1:}, \cdots, H_{k,q:})$, $(\Sigma^{1/2}(\beta_k^* - \frac{\beta_2+\beta_1}{2}))_{:1} \in$

$\text{span}(H_{k,1:},\cdots,H_{k,(q+1):}),\cdots,(\Sigma^{1/2}(\beta_k^*-\frac{\beta_2+\beta_1}{2}))_{:q}\in\text{span}(H_{k,1:},\cdots,H_{k,(2q):})$, and $(\Sigma^{1/2}(\beta_2+\beta_1))_{:1}\in\text{span}(H_{k,1:},\cdots,H_{k,(2q+1):}),\cdots,(\Sigma^{1/2}(\beta_2+\beta_1))_{:q}\in\text{span}(H_{k,1:},\cdots,H_{k,(3q):})$, where $A_{:j}$, $A_{l:}$ are the j-th column and l-th row of matrix A, respectively. We can write $X_i=\Sigma^{1/2}H_k^TV_i$, where $V_i\sim N(0,I_p)$. By definition, we have

$$\begin{aligned}
&\text{vec}^T(\beta_2-\beta_1)(I_q\otimes X_i)=V_i^TH_k^T\Sigma^{1/2}(\beta_2-\beta_1)=\|\Sigma(\beta_2-\beta_1)\|_FV_i^T\Lambda_1,\\
&=\underbrace{\|\Sigma^{1/2}(\beta_2-\beta_1)\|_F(\lambda_{11}V_{i1},\lambda_{12}V_{i1}+\lambda_{22}V_{i2},\cdots,\sum_{j=1}^{q}\lambda_{jq}V_{ij})}_{T_1^T},\\
&\text{vec}^T(\beta_k^*-\frac{\beta_2+\beta_1}{2})(I_q\otimes X_i)=V_i^TH_k^T\Sigma^{1/2}(\beta_k^*-\frac{\beta_2+\beta_1}{2})=\|\Sigma^{1/2}(\beta_k^*-\frac{\beta_2+\beta_1}{2})\|_FV_i^T\Lambda_2\\
&=\underbrace{\|\Sigma^{1/2}(\beta_k^*-\frac{\beta_2+\beta_1}{2})\|_F(\sum_{j=1}^{q}\lambda_{j1}^*V_{ij},\cdots,\sum_{j=1}^{2q}\lambda_{jq}^*V_{ij})}_{T_2^T},\\
&\text{vec}^T(\beta_2+\beta_1)(I_q\otimes X_i)=V_i^TH_k^T\Sigma^{1/2}(\beta_2+\beta_1)=\|\Sigma^{1/2}(\beta_2+\beta_1)\|_FV_i^T\Lambda_3\\
&=\underbrace{\|\Sigma^{1/2}(\beta_2+\beta_1)\|_F(\sum_{j=1}^{2q}\check{\lambda}_{j1}V_{ij},\cdots,\sum_{j=1}^{3q}\check{\lambda}_{jq}V_{ij})}_{T_2^T},
\end{aligned}$$

where Λ_1, Λ_2 and Λ_3 are $p\times q$ matrices and λ_{jl}, λ_{jl}^* and $\check{\lambda}_{jl}$ are the (j,l)-th element of Λ_1, Λ_2 and Λ_3, respectively. More specifically, when $j>l$, $\lambda_{jl}=0$, when $j>q+l$, $\lambda_{jl}^*=0$, and when $j>2q+l$, $\check{\lambda}_{jl}=0$. Also, we have $\sum_{j=1}^p\lambda_{jl}^2\le 1$, $\sum_{j=1}^p(\lambda_{jl}^*)^2\le 1$ and $\sum_{j=1}^p\check{\lambda}_{jl}^*\le 1$. Then we can write

$$\begin{aligned}
&\frac{1}{\omega_1(1-\omega_1)}\frac{\partial\eta_{1,\theta}(X_i,Y_i)}{\partial\text{vec}^T(\beta_2-\beta_1)}\text{vec}(X_iY_i^T)\\
&=(I_q\otimes\Sigma^{1/2})H_k^T\frac{\exp\{T_1^T(Z_i+T_2)\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1^T(Z_i+T_2)\}\big)^2}(I_q\otimes V_i)(Z_i+T_2^T)(Z_i+T_2^T-T_3^T/2)^T(I_q\otimes V_i^T)H_k(I_q\otimes\Sigma^{1/2}).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\text{E}\|\frac{\partial\eta_{1,\theta}(X_i,Y_i)}{\partial\text{vec}^T(\beta_2-\beta_1)}\text{vec}(X_iY_i^T)\|_2\\
&\le\frac{M_2}{4c_0^2}\text{E}\|\frac{\exp\{T_1^T(Z_i+T_2)\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1^T(Z_i+T_2)\}\big)^2}(I_q\otimes V_i)(Z_i+T_2^T)(Z_i+T_2^T-T_3^T/2)^T(I_q\otimes V_i)^T\|_2\\
&:=\frac{M_2}{4c_0^2}\text{E}\|L\|_2.
\end{aligned}$$

Note that the expectation of $L\in\mathbb{R}^{pq\times pq}$ is a block-wise diagonal matrix with block size $p\times p$. Also, only the first $3q\times 3q$ sub-matrix and the diagonal elements of each block matrix is non-zero. Thus, to bound the expectation of the 2-norm of the matrix, we only

need to consider those non-zero elements. For index (j,l) in the first $3q \times 3q$ sub-matrix and diagonal elements of the first $p \times p$ block matrix,

$$\mathrm{E}_{V|W_i=k}(|L_{jl}|) = \mathrm{E}_{V|W_i=k}\Big\{ \frac{\exp\{T_1^T(Z_i+T_2)\}}{\big(\omega_1+(1-\omega_1)\exp\{T_1^T(Z_i+T_2)\}\big)^2} |V_{ij}(Z_{ij}+T_{2,j})(Z_{il}+T_{2,l}-T_{3,l}/2)^T V_{il}| \Big\}.$$

Define events

$$\begin{aligned}
\mathcal{E}_1 &= \{Z_i : |T_1^T Z_i| < |T_1^T T_2| - |T_1^T T_1|/4\} \\
\mathcal{E}_2 &= \{Z_{ij} : |Z_{ij}| \le \|T_2\|_2\} \\
\mathcal{E}_3 &= \{Z_{il} : |Z_{il}| \le \|T_2\|_2\},
\end{aligned}$$

and $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$. We have

$$\mathbb{P}(\mathcal{E}_1^c) \le 2\exp\Big(-\frac{(|T_1^T T_2| - |T_1^T T_1|/4)^2}{2|T_1^T T_1|}\Big), \quad \mathbb{P}(\mathcal{E}_2^c) \le 2\exp(-|T_2^T T_2|/8), \quad \mathbb{P}(\mathcal{E}_3^c) \le 2\exp(-|T_2^T T_2|/8),$$

and on $\mathcal{E}_1$, $|T_1^T(Z_i+T_2)| \ge |T_1^T T_1|/4$. Note that when $\|T_2 - T_1/2\|_2$ (Controlled by C_b) is sufficient small,

$$\begin{aligned}
(|T_1^T T_2| - |T_1^T T_1|/4)^2 &\ge (|T_1^T T_1|/2 - |T_1^T(T_2 - T_1/2)| - |T_1^T T_1|/4)^2 \\
&= (|T_1^T T_1|/4 - \|T_1\|_2 \|P_{T_1}(T_2 - T_1/2)\|_2)^2 \\
&= \frac{1}{16}\|T_1\|_2^2 (\|T_1\|_2 - 4\|P_{T_1}(T_2 - T_1/2)\|_2)^2,
\end{aligned}$$

where P_{T_1} is the projection matrix on to the space spanned by T_1. Then we have

$$\begin{aligned}
\mathrm{E}_{V|W_i=k}(|L_{jl}|) &\le \mathrm{E}_{V|W_i=k}(|L_{jl}| \mid \mathcal{E})\mathbb{P}(\mathcal{E}) + \mathrm{E}_{V|W_i=k}(|L_{jl}| \mid \mathcal{E}^c)\mathbb{P}(\mathcal{E}^c) \\
&\le \mathrm{E}_{V|W_i=k}\Big\{ \frac{9}{8c_0^2}\exp(-T_1^T T_1/4)|V_{ij}V_{il}|T_2^T T_2 \Big\} \\
&\quad + \mathrm{E}_{V|W_i=k}\big\{ 2\exp(-(\|T_1\|_2 - 4\|P_{T_1}(T_2 - T_1/2)\|_2)^2/32) + 4\exp(-T_2^T T_2/2) \big\}
\end{aligned}$$

By the definition of the contraction basin $\mathcal{B}_{con}(\theta)$, we have

$$\|\Sigma^{1/2}(\beta_2 - \beta_1)\|_F \ge \|\Sigma^{1/2}(\beta_2^* - \beta_1^*)\|_F - \|\Sigma^{1/2}(\beta_2 - \beta_2^* - \beta_1 + \beta_1^*)\|_F \ge \Delta - 4C_b M_b$$

When $C_b \le 1/(4M_2)$, $\|\Sigma^{1/2}(\beta_2 - \beta_1)\|_F \ge c\Delta$. Similar conclusion also holds for $\|\Sigma^{1/2}(\beta_k^* - \frac{\beta_2+\beta_1}{2})\|_F$. Then recall that $T_1 = X_i(\beta_2 - \beta_1)$, $T_2 = X_i(\beta_k^* - \frac{\beta_2+\beta_1}{2})$, and $X_i \sim N(0,\Sigma)$. By Lemma A.8, we have

$$\mathrm{E}_{V|W_i=k}\big\{ 4\exp(-T_2^T T_2/2) \big\} \le C_1/\Delta.$$

for a generic constant C_1. Also note that $\|P_{T_1}(T_2 - T_1/2)\|_2 \le C_b M_b \|T_1\|$, when $C_b \le 1/(4M_2)$,

$$\mathrm{E}_{V|W_i=k}\big\{ 2\exp(-(\|T_1\|_2 - 4\|P_{T_1}(T_2 - T_1/2)\|_2)^2/32) \big\} \le C_2/\Delta.$$

Then, using Lemma A.8 again, we have

$$\begin{aligned}
&\mathrm{E}_{V|W_i=k}\big\{ \exp(-T_1^T T_1/4)|V_{ij}V_{il}|T_2^T T_2 \big\} \\
&\le \mathrm{E}_{V|W_i=k}\big\{ 2\exp(-T_1^T T_1/4)|V_{ij}V_{il}|(T_1^T T_1 + (T_2 - T_1/2)^T(T_2 - T_1/2)) \big\} = C_3/\Delta.
\end{aligned}$$

H.2 Proof of Lemma 7

We prove that $\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\{\|\widehat{\rho}_1(\theta)-\rho_1(\theta)\|_{F,s}\} = O(\sqrt{\frac{s\log(n)^2\log(p)}{n}})$ with probability at least $1-4p^{-1}$. The proof is similar to that for Lemma 3 with some modifications.

Recall that

$$\widehat{\rho}_1(\theta)-\rho_1(\theta) = \frac{1}{n}\sum_{i=1}^{n}\eta_{1,\theta}(X_i,Y_i)X_iY_i^T - \mathrm{E}(\eta_{1,\theta}(X_i,Y_i)X_iY_i^T).$$

Let $\widetilde{W}^{\rho} = \sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\|\widehat{\rho}_1(\theta)-\rho_1(\theta)\|_{F,s}$. By definition, we have

$$\widetilde{W}^{\rho} = \sup_{\mathrm{vec}(\mu)\in\Gamma(s)\cap\mathcal{S}^{pq-1}}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\langle\frac{1}{n}\sum_{i=1}^{n}\eta_{1,\theta}(X_i,Y_i)X_iY_i^T - \mathrm{E}(\eta_{1,\theta}(Y_i,X_i)X_iY_i^T),\mu\rangle_F$$

and

$$W_{\mu}^{\rho} = \sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\langle\frac{1}{n}\sum_{i=1}^{n}\eta_{1,\theta}(X_i,Y_i)X_iY_i - \mathrm{E}(\eta_{1,\theta}(Y_i,X_i)X_iY_i),\mu\rangle_F.$$

Because $\Gamma(s)\cap\mathcal{S}^{pq-1}\subseteq\mathcal{C}(s)\cap\mathcal{S}^{pq-1}$, we will bound

$$W^{\rho} = \sup_{\mu\in\mathcal{C}(s)\cap\mathcal{S}^{pq-1}}W_{\mu}^{\rho} = \sup_{\mu\in\mathcal{C}(s)\cap\mathcal{S}^{pq-1}}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\langle\frac{1}{n}\sum_{i=1}^{n}\eta_{1,\theta}(X_i,Y_i)X_iY_i^T - \mathrm{E}(\eta_{1,\theta}(Y_i,X_i)X_iY_i^T),\mu\rangle_F,$$

instead. Let $\mathrm{vec}(\nu_1),\cdots,\mathrm{vec}(\nu_{M_{net}})$ denote a 1/2-net of $\mathcal{C}(s)\cap\mathcal{S}^{pq-1}$, we have

$$W_{\nu}^{\rho}\le W_{\mu_j}^{\rho}+|W_{\mu_j}^{\rho}-W_{\nu}^{\rho}|\le\max_{j\in[M_{net}]}W_{\mu_j}^{\rho}+W^{\rho}\|\nu-\mu_j\|_F\le\max_{j\in[M_{net}]}W_{\mu_j}^{\rho}+1/2W^{\rho}.$$

Thus $W^{\rho}\le 2\max_{j\in[M_{net}]}W_{\mu_j}^{\rho}$. So, instead of directly bounding the tail of W^{ρ}, we can first get the tail probability for $W_{\mu_j}^{\rho}$ for a fixed j, then using the union bound to get the tail probability for W^{ρ}.

Note that

$$\begin{aligned}&\|\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\langle\eta_{1,\theta}(X_i,Y_i)X_iY_i^T,\mu_j\rangle_F - \mathrm{E}\sup_{\theta\in\mathcal{B}_{con}}\langle\eta_{1,\theta}(X_i,Y_i)X_iY_i^T,\mu_j\rangle_F\|_{\psi_1}\\ &\le c\|\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\langle\eta_{1,\theta}(X_i,Y_i)X_iY_i^T,\mu_j\rangle_F\|_{\psi_1}\\ &\le c\|\langle X_iY_i^T,\mu_j\rangle_F\|_{\psi_1}<\infty,\end{aligned}$$

where we use the fact that $0<\eta_{1,\theta}(X_i,Y_i)<1$ and $\langle X_iY_i^T,\mu_j\rangle_F$ is sub-exponential.

We use Lemma A.7 to get the tail probability for $W_{\mu_j}^{\rho}$. We first bound $\mathrm{E}(W_{\mu_j}^{\rho})$ using similar procedure given in Adamczak (2008). Let $f(X_i,Y_i)=\eta_{1,\theta}(X_i,Y_i)\langle X_iY_i^T,\mu_j\rangle_F - \mathrm{E}(\eta_{1,\theta}(X_i,Y_i)\langle X_iY_i^T,\mu_j\rangle_F)$ for simplicity. We have $W_{\mu_j}^{\rho}\le\frac{1}{n}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}|\sum_{i=1}^{n}f(X_i,Y_i)|$.

Define truncated function and the remaining part of $f(X_i,Y_i)$ as

$$\begin{aligned}f_1 &= f(X_i,Y_i)I(|\langle X_iY_i^T,\mu_j\rangle_F|\le\rho)-\mathrm{E}[f(X_i,Y_i)I(|\langle X_iY_i^T,\mu_j\rangle_F|\le\rho)],\\ f_2 &= f(X_i,Y_i)I(|\langle X_iY_i^T,\mu_j\rangle_F|>\rho)-\mathrm{E}[f(X_i,Y_i)I(|\langle X_iY_i^T,\mu_j\rangle_F|>\rho)],\end{aligned}$$

where $\rho = 8\mathrm{E}\max_i \sup_{\theta\in\mathcal{B}_{con}(\theta^*),\mu\in\Gamma(s)\cap\mathcal{S}^{p-1}} |f(X_i,Y_i)|$.

Let $Q = \max_i |\langle X_i Y_i^T, \mu_j\rangle_F|$. We have

$$\mathrm{E}Q = \int_0^\infty \mathbb{P}(Q > x)dx. \tag{A.15}$$

Note that $\|\langle X_i Y_i^T, \mu_j\rangle_F\|_{\psi_1} \le C$ we have $\mathbb{P}(|\langle X_i Y_i^T, \mu_j\rangle_F| > x\log n) \le \exp(-cx\log n)$. By union bound we have

$$\mathbb{P}(Q > x\log n) \le \sum_{i=1}^n \exp(-cx\log n) = \exp(-xc\log n + \log n).$$

When $\log n > 2$ and $x > 2/c$, we have $\mathbb{P}(Q > x\log n) \le \exp(-cx)$. By (A.15) and $\mathbb{P}(Q > x) < 1$, we have $\mathrm{E}Q \le c\log n$. It follows that $\rho \le C\log(n)$.

Note that

$$\begin{aligned}\sup_{\theta\in\mathcal{B}_{con}(\theta^*),\mu\in\Gamma(s)\cap\mathcal{S}^{p-1}} |\sum_{i=1}^n f(X_i,Y_i)| \le & \sup_{\theta\in\mathcal{B}_{con}(\theta^*),\mu\in\Gamma(s)\cap\mathcal{S}^{p-1}} |\sum_{i=1}^n f_1(X_i,Y_i) - \mathrm{E}f_1(X_i,Y_i)| \\ & + \sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\sum_{i=1}^n f_2(X_i,Y_i) - \mathrm{E}f_2(X_i,Y_i)|,\end{aligned}$$

where we use the fact that $\mathrm{E}(f_1(X_i,Y_i) + f_2(X_i,Y_i)) = 0$. It follows that

$$\begin{aligned}\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\sum_{i=1}^n f(X_i,Y_i)| \le & \mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\sum_{i=1}^n f_1(X_i,Y_i) - \mathrm{E}f_1(X_i,Y_i)| \\ & + 2\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\sum_{i=1}^n f_2(X_i,Y_i)|.\end{aligned} \tag{A.16}$$

By Markov inequality and the definition of $f_2(X_i,Y_i)$, we have

$$\mathbb{P}(\max_{k\le n}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\sum_{i=1}^k f_2(X_i,Y_i)| > 0) \le \mathbb{P}(\max_i |\langle X_i Y_i^T, \mu_j\rangle_F| > \rho) \le 1/8.$$

Then by Hoffmann-Jørgensen inequality (see e.g Ledoux and Talagrand (1991), Proposition 6.8)

$$\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\sum_{i=1}^n f_2(X_i,Y_i)| \le 8\mathrm{E}\max_i \sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |f(X_i,Y_i)| \le \rho. \tag{A.17}$$

Thus, we have

$$\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\frac{1}{n}\sum_{i=1}^n f_2(X_i,Y_i)| \le \frac{C\log n}{n}. \tag{A.18}$$

Next we go back to bound $\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\frac{1}{n}\big(\sum_{i=1}^n f_1(X_i,Y_i) - \mathrm{E}f_1(X_i,Y_i)\big)|$. By Lemme A.5 and Lemma A.6, we have

$$\begin{aligned}&\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\frac{1}{n}\big(\sum_{i=1}^n f_1(X_i,Y_i) - \mathrm{E}f_1(X_i,Y_i)\big)| \\ \le & C\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)} |\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i^T(\beta_2-\beta_1)(Y_i - \frac{(\beta_1+\beta_2)^T}{2}X_i)\langle X_i Y_i^T, \mu_j\rangle_F|,\end{aligned}$$

where $|\langle X_iY_i^T, \mu_j\rangle_F| \le \rho$ for all i. Under the condition $|\langle X_iY_i^T, \mu_j\rangle_F| \le \rho$, $\epsilon_i X_i^T(\beta_2 - \beta_1)(Y_i - \frac{(\beta_1+\beta_2)^T}{2}X_i)\langle X_iY_i^T, \mu_j\rangle_F$ is sub-exponential for any β_1 and β_2. Define set $\mathcal{T} = \{\nu_1 : \nu_1^T = (\beta_1^T, \beta_2^T) \in \mathcal{B}_{con}\}$. We have $D = \mathrm{diam}(\mathcal{T}) \le cC_b\Delta$.

By Corollary 5.2 of Dirksen (2015), we have

$$\begin{aligned}&\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i^T(\beta_2-\beta_1)(Y_i - \frac{(\beta_1+\beta_2)^T}{2}X_i)\langle X_iY_i^T, \mu_j\rangle_F| \\ &\le C_1\big(\frac{1}{\sqrt{n}}\gamma_2(\mathcal{T}, d) + \frac{1}{n}\gamma_1(\mathcal{T}, d)\big) + C_2(\frac{1}{n} + \frac{1}{\sqrt{n}}),\end{aligned} \tag{A.19}$$

where $\gamma_1(\mathcal{T}, d)$ and $\gamma_2(\mathcal{T}, d)$ are Talagrand γ_1 and γ_2 functional (See Dirksen (2015) for details), and d is the ℓ_2-norm.

By Lemma A.3, $\mathcal{T} \in C\mathrm{conv}(\cup_{|J|\le 2c_ds}E_J(2pq) \cap B_2^{2pq})$. From the volumetric argument in Rudelson and Zhou (2012)[Section H.1], we know that the covering number $\mathcal{N}(\mathcal{T}, d, \epsilon)$ satisfies that

$$\log(\mathcal{N}(\mathcal{T}, d, \epsilon)) \le C_4\big(s\log(epq/s) + s\log(1 + 2/\epsilon)\big). \tag{A.20}$$

Then note that

$$\gamma_\alpha(\mathcal{T}, d) \le C\rho\int_0^D \big(\log\mathcal{N}(T, d, \epsilon)\big)^{1/\alpha}d\epsilon,$$

we have

$$\gamma_2(\mathcal{T}, d) = C\rho\int_0^D \sqrt{\log(\mathcal{N}(\mathcal{T}, d, \epsilon))}d\epsilon = C\rho\int_0^D \sqrt{s}\Big(\log(\frac{epq}{s}) + \log(1 + 2/\epsilon)\Big)^{1/2}d\epsilon \le C_1\rho\sqrt{s\log(pq)},$$

and

$$\gamma_1(\mathcal{T}, d) = C\rho\int_0^D \log(\mathcal{N}(\mathcal{T}, d, \epsilon))d\epsilon = C\rho\int_0^D s\Big(\log(\frac{epq}{s}) + \log(1 + 2/\epsilon)\Big)d\epsilon \le C_1\rho s\log(pq).$$

By (A.19), we know that

$$\mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i^T(\beta_2-\beta_1)(Y_i - \frac{(\beta_1+\beta_2)^T}{2}X_i)\langle X_iY_i^T, \mu_j\rangle_F| \le C\log n\sqrt{\frac{s\log(pq)}{n}}.$$

Combine this result and (A.18), we have

$$\mathrm{E}(W_{\mu_j^\rho}) \le \mathrm{E}\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}|\frac{1}{n}\sum_{i=1}^n \frac{1}{n}f(X_i, Y_i)| \le C\Big(\sqrt{\frac{s(\log n)^2\log(pq)}{n}} + \frac{\log n}{n}\Big).$$

Note that $\|\max_i \sup_{\theta\in\mathcal{B}_{con}(\theta^*)} f(X_i, Y_i)\|_{\psi_1} \le \|\max_i |\langle X_iY_i^T, \mu_j\rangle_F|\|_{\psi_1}$ and $\|\max_i |\langle X_iY_i^T, \mu_j\rangle_F|\|_{\psi_1} \le C\log n\|\langle X_iY_i^T, \mu_j\rangle_F\|_{\psi_1}$ (Pisier's inequality (Pisier, 1983)), we have $\|\max_i \sup_{\theta\in\mathcal{B}_{con}(\theta^*)} f(X_i, Y_i)\|_{\psi_1} \le C\log n$. Then by Lemma A.7, we have

$$\mathbb{P}(W_{\mu_j^\rho} \ge C\big(\sqrt{\frac{s(\log n)^2\log(pq)}{n}} + \frac{\log n}{n}\big) + t) \le 4\max\big\{\exp(-C_5nt^2), 3\exp(-\frac{C_6nt}{\log n})\big\}$$

By union bound, we have

$$\mathbb{P}(W^\rho \geq C\big(\sqrt{\frac{s(\log n)^2\log(pq)}{n}}+\frac{\log n}{n}\big)+t) \leq M_{net}\mathbb{P}(W_{\mu_j^\rho}) \\ \leq 4\max\big\{\exp(c_d s\log(pq)-C_5 nt^2), 3\exp(c_d s\log p-\frac{C_6 nt}{\log n})\big\}.$$

Let $t=c\sqrt{\frac{s\log(n)^2\log(pq)}{n}}$ for large enough generic constant c, when $n \gg s\log(pq)$,

$$W^\rho = O(\sqrt{\frac{s\log(n)^2\log(pq)}{n}})$$

with probability at least $1-(pq)^{-1}$. We have proved $\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\{\|\widehat{\rho}_1(\theta)-\rho_1(\theta)\|_{F,s}\} = O(\sqrt{\frac{s\log(n)^2\log(pq)}{n}})$ with probability at least $1-(pq)^{-1}$. The proofs for $\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\{\|(\widehat{\Sigma}_1(\theta)-\rho_1(\theta))\beta_1^*\|_{F,s}\} = O(\sqrt{\frac{s\log(n)^2\log(pq)}{n}})$ and $\sup_{\theta\in\mathcal{B}_{con}(\theta^*)}\{|\widehat{\omega}_1(\theta)-\omega_1(\theta)|\} = O(\sqrt{\frac{s\log(n)^2\log(pq)}{n}})$ are similar. We omit the details.

The proof for Theorem 8 is analogous to that for 4 by replacing 2-norm by Frobenius norm regarding ρ_k, $\Sigma_k\beta_k^*$ and β_k.

References

Radoslaw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 15:2773–2832, 2014.

Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45 (1):77–120, 2017.

Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

T Tony Cai, Weidong Liu, and Xi Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, 2011.

T Tony Cai, Jing Ma, and Linjun Zhang. Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267, 2019.

Antoine Deleforge, Florence Forbes, and Radu Horaud. High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25 (5):893–911, 2015.

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)*, 39(1):1–22, 1977.

Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20: 1–29, 2015.

David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

Sangwon Hyun, Mattias Rolf Cape, Francois Ribalet, and Jacob Bien. Modeling cell populations measured by flow cytometry with covariates using sparse mixture of regressions. *The Annals of Applied Statistics*, 17(1):357 – 377, 2023.

Abbas Khalili and Jiahua Chen. Variable selection in finite mixture of regression models. *Journal of the american Statistical association*, 102(479):1025–1038, 2007.

Jason M Klusowski, Dana Yang, and WD Brinda. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory*, 65(6):3515–3524, 2019.

Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the minimax optimality of the em algorithm for learning two-component mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1405–1413. PMLR, 2021.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer Science & Business Media, 1991.

Friedrich Leisch. Flexmix: A general framework for finite mixture models and latent class regression in r. *Journal of Statistical Software*, 11(8):1–18, 2004.

Qianyun Li, Runmin Shi, and Faming Liang. Drug sensitivity prediction with high-dimensional mixture regression. *PloS one*, 14(2):e0212108, 2019.

Geoffrey J McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, 2004.

Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The annals of statistics*, 37(1):246–270, 2009.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557, 2012.

Gilles Pisier. Some applications of the metric entropy condition to harmonic analysis. In *Banach Spaces, Harmonic Analysis, and Probability Theory*, pages 123–154. Springer, 1983.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. In *Conference on Learning Theory*, pages 10–1. JMLR Workshop and Conference Proceedings, 2012.

Weixing Song, Weixin Yao, and Yanru Xing. Robust mixture regression model fitting by laplace distribution. *Computational Statistics & Data Analysis*, 71:128–137, 2014.

Nicolas Städler, Peter Bühlmann, and Sara Van De Geer. ℓ1-penalization for mixture regression models. *Test*, 19(2):209–256, 2010.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

T Rolf Turner. Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(3):371–384, 2000.

Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional em algorithm: Statistical optimization and asymptotic normality. *Advances in neural information processing systems*, 28:2512–2520, 2015.

Yi Yang and Hui Zou. A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141, 2015.

Weixin Yao, Yan Wei, and Chun Yu. Robust mixture regression using the t-distribution. *Computational Statistics & Data Analysis*, 71:116–127, 2014.

Fei Ye and Cun-Hui Zhang. Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *The Journal of Machine Learning Research*, 11:3519–3540, 2010.

Xinyang Yi and Constantine Caramanis. Regularized em algorithms: A unified framework and statistical guarantees. *Advances in Neural Information Processing Systems*, 28:1567–1575, 2015.

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pages 613–621, 2014.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Linjun Zhang, Rong Ma, T Tony Cai, and Hongzhe Li. Estimation, confidence intervals, and large-scale hypotheses testing for high-dimensional mixed linear regression. *arXiv preprint arXiv:2011.03598*, 2020.

Rongda Zhu, Lingxiao Wang, Chengxiang Zhai, and Quanquan Gu. High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm. In *International Conference on Machine Learning*, pages 4180–4188. PMLR, 2017.