# Estimation, Confidence Intervals, and Large-Scale Hypotheses Testing for High-Dimensional Mixed Linear Regression

Linjun Zhang, Rong Ma, T. Tony Cai, and Hongzhe Li

**Abstract**

This paper studies high-dimensional mixed linear regression (MLR) models where the output variable comes from one of the two linear regression models with an unknown mixing proportion and an unknown covariance structure of the random covariates. Building upon a high-dimensional EM algorithm, we propose an iterative procedure for estimating the two regression vectors and establish their rates of convergence. We further construct debiased estimators and establish their asymptotic normality. Confidence intervals centered at the debiased estimators are constructed for individual coordinates. Furthermore, a large-scale multiple testing procedure is proposed for testing the regression coefficients and is shown to control the false discovery rate (FDR) asymptotically. Simulation studies are carried out to examine the numerical performance of the proposed methods and their superiority over existing methods. The proposed methods are further illustrated through an analysis of a dataset of multiplex image cytometry, which investigates the interaction networks among the cellular phenotypes that include the expression levels of 20 epitopes or combinations of markers.

*KEYWORDS*: Debiasing, EM algorithm, FDR, Iterative estimation, Large-scale multiple testing

[0]Linjun Zhang is Assistant Professor, Department of Statistics, Rutgers University, Piscataway, NJ 08854. (E-mail: *linjun.zhang@rutgers.edu*). Rong Ma is Assistant Professor, Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115 *rongm@pennmedicine.upenn.edu*). T. Tony Cai is Daniel H. Silberberg Professor of Statistics, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail:*tcai@wharton.upenn.edu*). Hongzhe Li is Perelman Professor of Biostatistics, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *hongzhe@upenn.edu*).

## 1 INTRODUCTION

Mixed linear regression (MLR) models are widely used in analyzing heterogeneous data arising from biology, physics, economics, and business (McLachlan and Peel 2004; Grün and Leisch 2007; Netrapalli et al. 2013; Li et al. 2019; Devijver et al. 2020). In many modern applications, the number of the covariates is comparable with, or sometimes far exceeds, the number of the observed samples. In such high-dimensional settings, statistical inference methods designed for the classical low-dimensional setting are often not valid. There is a paucity of methods and fundamental theoretical understanding on statistical estimation and inference for high-dimensional MLR models. This motivates us to develop computationally efficient and theoretically guaranteed statistical methods for high-dimensional MLR models in analyzing large heterogeneous datasets.

We consider the following high-dimensional MLR model where the observed data are $i.i.d.$ draws from one of the two unknown linear models:

$$y_i = \begin{cases} \boldsymbol{x}_i^\top \boldsymbol{\beta}_1^* + \epsilon_i & \text{with probability } \omega^*, \\ \boldsymbol{x}_i^\top \boldsymbol{\beta}_2^* + \epsilon_i & \text{with probability } 1-\omega^*, \end{cases} \qquad i = 1, 2, \ldots, n, \quad \boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^* \in \mathbb{R}^p, \tag{1.1}$$

where the random design variables $\boldsymbol{x}_i \in \mathbb{R}^p$ are $i.i.d.$ samples from $N_p(0, \boldsymbol{\Sigma})$ with the unknown covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}_1^* \neq \boldsymbol{\beta}_2^*$ are latent regression coefficients, $\omega^* \in (0, 1)$ is the unknown mixing proportion, and the noise $\epsilon_i$ is $i.i.d.$ from $N(0, \sigma^{*2})$ for some unknown $\sigma^* > 0$. For identifiability, we assume $\omega^* \in (1/2, 1)$. Under the high-dimensional setting where $p$ is much larger than $n$, we aim to answer the following inference questions:

1. What is an efficient algorithm for estimating the underlying regression vectors $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ where both the mixing proportion $\omega^*$ and the random design covariance matrix $\boldsymbol{\Sigma}$ are unknown? What is the rate of convergence of the estimator?

2. How to construct asymptotically valid tests and confidence intervals for the individual coordinates of the latent regression coefficients $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$, and their difference $\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*$?

3. For simultaneously testing the null hypotheses $H_{0j} : \boldsymbol{\beta}^*_{1j} = \boldsymbol{\beta}^*_{2j} = 0$, $j = 1, ..., p$, how to construct a large-scale multiple testing procedure that controls the false discovery rate (FDR) and false discovery proportion (FDP) asymptotically?

### 1.1 Related Work

In the classical low-dimensional settings, the problems of estimation, hypotheses testing and confidence intervals for MLR have been extensively studied in literature. For example, Zhu and Zhang (2004) considered hypothesis testing and developed an asymptotic theory for both the maximum likelihood and the maximum modified likelihood estimators in MLR. Khalili and Chen (2007) introduced a penalized likelihood approach for variable selection in MLR. Faria and Soromenho (2010) compared three expectation-maximization (EM) algorithms that compute the maximum likelihood estimates of the coefficients of MLR. Chaganty and Liang (2013) developed a computationally efficient algorithm based on the tensor power method, and obtained the rates of convergence for their proposed estimator. Bashir and Carter (2012) proposed a robust model that can achieve high breakdown point in the contaminated data for parameter estimation in MLR. Based on a new initialization step for the EM algorithm, Yi et al. (2014) provided the theoretical guarantees for coefficient estimation in MLR. Moreover, Yao and Song (2015) proposed a deconvolution method to study the MLR with measurement errors. Zhong et al. (2016) introduced a non-convex continuous objective function for solving the general unbalanced $k$-component MLR.

Balakrishnan et al. (2017) developed a general framework for proving rigorous guarantees on the performance of the EM algorithm and applied to some statistical problems, including estimating the coefficients in the symmetric MLR where the mixing proportion is known to be $\omega^* = 1/2$. Li and Liang (2018) presented a fixed parameter algorithm that solves MLR under the Gaussian design in time that is nearly linear in the sample size and the dimension. More recently, building upon the work of Balakrishnan et al. (2017), McLachlan and Peel (2004), Klusowski et al. (2019) and Kwon and Caramanis (2020) introduced new tools for analyzing the convergence rates of the EM algorithm for estimating the coefficients. Shen and Sanghavi (2019) proposed an efficient algorithm, Iterative Least Trimmed Squares, for solving MLR with adversarial corruptions. Chen et al. (2014); Kwon et al. (2021) provided a complete picture of the non-asymptotic behavior of the EM algorithm for low-dimensional symmetric two-component MLE under all regimes of signal-to-noise ratio.

In contrast, statistical inference for MLR in the high-dimensional setting is relatively less studied. Specifically, Städler et al. (2010) proposed an $\ell_1$ penalized estimator and developed an efficient EM algorithm for MLR with provable convergence properties. Wang et al. (2015) and Yi and Caramanis (2015) established a general theory of the EM algorithm for statistical inference in high dimensional latent variable models, including the high-dimensional MLR with the symmetric and spherical assumptions. Zhu et al. (2017) introduced a generic stochastic EM algorithm for the high-dimensional MLR with theoretical guarantees obtained under the symmetric setting ($\omega^* = 1/2$). More recently, Fan et al. (2018) studied the fundamental tradeoffs between statistical accuracy and computational tractability for high-dimensional latent variables models, including testing the global null hypothesis in MLR. However, problems such as statistical inference for the individual regression coefficients and large-scale multiple testing under the general MLR with an unknown mixing proportion and an unknown design covariance matrix have not been addressed in the literature.

### 1.2 Main Contributions

The main contributions of our paper are three-fold.

1. Based on a careful analysis of a high-dimensional EM algorithm, we propose iterative estimators for the regression coefficients $(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$ without the knowledge of the mixing proportions or the design covariance matrix, and obtain explicitly the rates of convergence of the iterative estimators under the $\ell_2$ norm. To the best of our knowledge, this is the first result on the estimation of the high-dimensional MLR with both unknown mixing proportion and unknown design covariance matrix.

2. We construct debiased estimators of the latent regression coefficients, based on the iterative estimators, and establish the asymptotic normality of its individual coordinates. The limiting distribution is then used for constructing confidence intervals and tests for the individual latent regression coefficients.

3. For the problem of large-scale testing of hypotheses $H_{0j} : \boldsymbol{\beta}^*_{1j} = \boldsymbol{\beta}^*_{2j} = 0$, $j = 1, ..., p$, we propose a multiple testing procedure that is shown to control the FDR and FDP at the nominal level asymptotically. Strong numerical results suggest the superior empirical performance of the proposed testing procedure over the existing methods.

### 1.3 Notation and Organization

Throughout our paper, for a vector $\mathbf{a} = (a_1, ..., a_n)^\top \in \mathbb{R}^n$, we define the $\ell_p$ norm $\|\mathbf{a}\|_p = \left(\sum_{i=1}^n a_i^p\right)^{1/p}$, and the $\ell_\infty$ norm $\|\mathbf{a}\|_\infty = \max_{1\le j\le n} |a_i|$. $\mathbf{a}_{-j} \in \mathbb{R}^{n-1}$ stands for the subvector of $\mathbf{a}$ without the $j$-th component. For vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, we denote their inner product $\langle \mathbf{a}, \mathbf{b}\rangle = \sum_{i=1}^n a_i b_i$ and denote the angle by $\angle(\boldsymbol{a}, \boldsymbol{b})$. We use $\boldsymbol{e}_j^{(p)}$ to denote the $p$-dimensional vector with the $j$-th component being one and others being zero. For a matrix $A \in \mathbb{R}^{p\times q}$, $\lambda_i(A)$ stands for the $i$-th largest singular value of $A$ and $\lambda_{\max}(A) = \lambda_1(A)$, $\lambda_{\min}(A) = \lambda_{p\wedge q}(A)$. $\|A\|_1$ denotes the matrix $\ell_1$ norm, and $\|A\|_\infty = \max_{i,j} |A_{ij}|$. In addition, $A_{-i.-j} \in \mathbb{R}^{(p-1)\times(q-1)}$ stands for the submatrix of $A$ without the $i$ th row and $j$-th column. For any positive integer $p$, we denote $[p] = \{1, ..., p\}$. Furthermore, for sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = o(b_n)$ if $\lim_n a_n/b_n = 0$, and write $a_n = O(b_n)$, $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there exists a constant $C$ such that $a_n \le Cb_n$ for all $n$. We also write $a_n = O_P(b_n)$ if there exists a constant $C$ such that $\liminf_{n\to\infty} \mathbb{P}(a_n \le Cb_n) = 1$, and $a_n = o_P(b_n)$ if $a_n/b_n \xrightarrow{p} 1$. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. For a set $A$, we denote $|A|$ as its cardinality. Lastly, $C, C_0, C_1, ...$ are constants that may vary from place to place.

The rest of the paper is organized as follows. We propose in Section 2 the iterative algorithm and the estimators of the latent regression coefficients and study their theoretical properties. Section 3 introduces the debiased estimators of individual regression coefficients and obtains their asymptotic normality and the resulting confidence intervals. In Section 4, by focusing on the problem of testing large-scale simultaneous hypotheses, we present our multiple testing procedure and show that it controls the FDR/FDP asymptotically. In Section 5, the numerical performance of the proposed methods are evaluated through extensive simulations. In Section 6, the proposed procedures are illustrated by an analysis of a multiplex image cytometry dataset. Section 7 discusses extensions to lower signal-to-noise ratio regimes. Further discussions and related problems are discussed in Section 8. The main results are proved in Section 9 and the proofs of other theorems as well as technical lemmas are collected in the Supplementary Materials.

## 2 ITERATIVE ESTIMATION VIA THE EM ALGORITHM

Suppose we have $n$ observations $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ generated independently from the MLR model in (1.1), and wish to estimate and make inference for the coefficient vectors $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$. In the classical setting where $p$ is fixed or much smaller than $n$, the maximum likelihood estimator (MLE) has been shown to perform well under mild conditions (Balakrishnan et al. 2017). The MLE aims to maximize the log-likelihood of the data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, which can be written as

$$l_n(\boldsymbol{\theta}; \boldsymbol{x}, y) = \frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{\omega}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1\rangle)^2}{2\sigma^2}\right\} + \frac{1-\omega}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_2\rangle)^2}{2\sigma^2}\right\}\right], \tag{2.1}$$

where we denote the parameter $\boldsymbol{\theta} = (\omega, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma)$ and the log-likelihood by $l_n(\boldsymbol{\theta})$.

Due to the non-convexity of $l_n(\boldsymbol{\theta}; \boldsymbol{x}, y)$, searching for the MLE is computationally intractable. Moreover, in the high-dimensional setting where the dimension $p$ is much larger than the sample size $n$, the MLE is in general not well defined, unless the models are carefully regularized by sparsity-type assumptions. In this paper, we propose to explore the sparsity of the coefficient vectors. Further, we develop an EM algorithm to address the extra computational challenge for parameter estimation and uncertainty assessment.

### 2.1 High-Dimensional EM Algorithm and the Iterative Estimators

For ease of presentation, let us use $z_i$ to denote the hidden labels of $(\boldsymbol{x}_i, y_i)$, that is, $z_i = 1$ if $(\boldsymbol{x}_i, y_i)$ is drawn from the first model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_1^* + \epsilon_i$, and $z_i = 2$ if the underlying truth is the second model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_2^* + \epsilon_i$. The marginal distribution of $z_i$ is given by $\mathbb{P}(z_i = 1) = 1 - \mathbb{P}(z_i = 2) = \omega$. The EM algorithm is essentially an alternating maximization method, which alternatively optimizes between the identification of hidden labels $\{z_i\}_{i=1}^n$ and the estimation of parameter $\boldsymbol{\theta} = (\omega, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma)$.

Specifically, in the E-step of $(t+1)$-th iteration, given the parameters $\hat{\boldsymbol{\theta}}^{(t)} = (\hat{\omega}^{(t)}, \hat{\boldsymbol{\beta}}_1^{(t)}, \hat{\boldsymbol{\beta}}_2^{(t)}, \hat{\sigma}^{(t)})$ estimated from the previous $t$-th step, the conditional probability of the $i$-th sample in class 1 given the observed data $(\boldsymbol{x}_i, y_i)$ can be calculated as

$$\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{x}_i, y_i) := \mathbb{P}_{\hat{\boldsymbol{\theta}}^{(t)}}(z_i = 1 \mid \boldsymbol{x}_i, y_i) = \frac{\hat{\omega}^{(t)} \exp(-\frac{(y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_1^{(t)} \rangle)^2}{2\hat{\sigma}^{(t)2}})}{\hat{\omega}^{(t)} \exp(-\frac{(y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_1^{(t)} \rangle)^2}{2\hat{\sigma}^{(t)2}}) + (1 - \hat{\omega}^{(t)}) \exp(-\frac{(y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_2^{(t)} \rangle)^2}{2\hat{\sigma}^{(t)2}})}. \tag{2.2}$$

As a result, the conditional expectation of the log-likelihood (2.1), with respect to the conditional distribution given $(\boldsymbol{x}, y)$ under the current estimate of the parameter $\hat{\boldsymbol{\theta}}^{(t)}$, can be calculated as

$$\begin{aligned} Q_n(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) :=& \mathbb{E}_{\hat{\boldsymbol{\theta}}^{(t)}}[l_n(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{x}, y] \qquad (2.3) \\ =& -\frac{1}{2n}\left[\sum_{i=1}^n \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{x}_i, y_i)(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1 \rangle)^2 + \sum_{i=1}^n (1 - \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{x}_i, y_i))(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_2 \rangle)^2\right] \\ &+ \frac{1}{n}\sum_{i=1}^n (1 - \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{x}_i, y_i)) \log(1-\omega) + \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{x}_i, y_i) \log \omega. \end{aligned}$$

Given $\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{x}_i, y_i)$, i.e., the distribution of the latent labels, the M-step is usually proceeded by maximizing $Q_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)})$:

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)}).$$

However, in the high-dimensional setting, such a maximization tends to overfit data. To handle the challenge of high-dimensionality, the key ingredient of our algorithm is to add a regularization term $\|\boldsymbol{\beta}\|_1$ to enforce sparsity. In particular, we write $\gamma_{\hat{\boldsymbol{\theta}},i}^{(t)} = \gamma_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{x}_i, y_i)$, and let

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1^{(t+1)} &= \arg\min_{\boldsymbol{\beta}_1} \frac{1}{2n}\sum_{i=1}^n \gamma_{\boldsymbol{\theta},i}^{(t)}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1\rangle)^2 + \lambda_n^{(t+1)}\|\boldsymbol{\beta}_1\|_1 \\ \hat{\boldsymbol{\beta}}_2^{(t+1)} &= \arg\min_{\boldsymbol{\beta}_2} \frac{1}{2n}\sum_{i=1}^n (1-\gamma_{\boldsymbol{\theta},i}^{(t)})(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_2\rangle)^2 + \lambda_n^{(t+1)}\|\boldsymbol{\beta}_2\|_1,\end{aligned} \tag{2.4}$$

where $\lambda^{(t+1)}$ is a tuning parameter which will also be updated recursively, and will be specified later. We also update $\hat{\omega}^{(t+1)}$ and $\hat{\sigma}^{(t+1)}$ by

$$\hat{\omega}^{(t+1)} = \frac{1}{n}\sum_{i=1}^n \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{x}_i, y_i) \tag{2.5}$$

and

$$(\hat{\sigma}^{(t+1)})^2 = \sum_{i=1}^n y_i^2 - \omega^{(t+1)} \cdot \hat{\boldsymbol{\beta}}_1^{(t+1)\top}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top)\hat{\boldsymbol{\beta}}_1^{(t+1)} - (1-\omega^{(t+1)}) \cdot \hat{\boldsymbol{\beta}}_2^{(t+1)\top}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top)\hat{\boldsymbol{\beta}}_2^{(t+1)}. \tag{2.6}$$

Given a suitable initialization, the proposed high-dimensional EM algorithm then proceeds by iterating between the E-step and the M-step, which is summarized in the following Algorithm 1.

### 2.2 Rate of Convergence

In this section, we give theoretical guarantees for estimating the coefficient vectors $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ using Algorithm 1. To begin with, we introduce the parameter space for $(\omega^*, \boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \sigma^*)$, where we assume that $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ are both sparse vectors, $\omega^*$ is bounded away from 0 or 1 and $\sigma^*$ is of constant order.

$$\Theta(s) = \left\{(\omega, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma) : \omega \in (c_1, 1-c_2), \sigma \in (0, c_2), \|\boldsymbol{\beta}_1\|_0, \|\boldsymbol{\beta}_2\|_0 \le s, \text{ for some } c_1 \in (0, 1/2), c_2 > 0\right\}.$$

Furthermore, we introduce the following regularity conditions on the initialization and signal-to-noise ratio (SNR) strength.

(A1) : Initialization: $\|\boldsymbol{\beta}_1^{(0)} - \boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^{(0)} - \boldsymbol{\beta}_2^*\| + |\sigma^{(0)} - \sigma^*| + |\omega^{(0)} - \omega^*| \le \min\{\omega^*/2, (1-\omega^*)/2, c_l \cdot \Delta^*, c_l \cdot \sigma^*\}$, where $\Delta^* = \sqrt{(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*)}$;

(A2) : SNR strength: $\mathrm{SNR} := \sqrt{(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*)} \ge c_{snr}$,

**Algorithm 1** EM for High-Dimensional MLR

1: **Inputs:** Initializations $\hat{\omega}^{(0)}, \hat{\boldsymbol{\beta}}_1^{(0)}, \hat{\boldsymbol{\beta}}_2^{(0)}, \hat{\sigma}^{(0)}$, maximum number of iterations $T$, and constants $\kappa \in (0,1)$, $C_\lambda > 0$. Split the Dataset into $T$ subsets of size $n/T$. For $i \in [n]$, set

$$\gamma_{\boldsymbol{\theta},i}^{(0)} = \frac{\hat{\omega}^{(0)} \exp(-\frac{(y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_1^{(0)}\rangle)^2}{2\hat{\sigma}^{(0)2}})}{\hat{\omega}^{(0)} \exp(-\frac{(y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_1^{(0)}\rangle)^2}{2\hat{\sigma}^{(0)2}}) + (1-\hat{\omega}^{(0)}) \exp(-\frac{(y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_2^{(0)}\rangle)^2}{2\hat{\sigma}^{(0)2}})}.$$

2: **for** $t = 0, 1, \ldots, T-1$ **do**

3: **E-Step:** Evaluate $Q_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)})$ as defined in (2.3) with the $t$-th data subset.

4: **M-Step:** Update $\hat{\boldsymbol{\beta}}_1^{(t+1)}$ and $\hat{\boldsymbol{\beta}}_2^{(t+1)}$ via

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1^{(t+1)} &= \arg\min_{\boldsymbol{\beta}_1} \frac{1}{2n} \sum_{i=1}^n \gamma_{\boldsymbol{\theta},i}^{(t)} (y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1 \rangle)^2 + \lambda_n^{(t+1)} \|\boldsymbol{\beta}_1\|_1 \\ \hat{\boldsymbol{\beta}}_2^{(t+1)} &= \arg\min_{\boldsymbol{\beta}_2} \frac{1}{2n} \sum_{i=1}^n (1-\gamma_{\boldsymbol{\theta},i}^{(t)}) (y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_2 \rangle)^2 + \lambda_n^{(t+1)} \|\boldsymbol{\beta}_2\|_1, \end{aligned} \tag{2.7}$$

with

$$\lambda_n^{(t+1)} = \kappa_\lambda \lambda_n^{(t)} + C_\lambda \sqrt{\frac{\log p}{n}}. \tag{2.8}$$

Update $\hat{\omega}^{(t+1)}$ and $\hat{\sigma}^{(t+1)}$ via (2.5) and (2.6).

5: **end for**

6: Output $\boldsymbol{\beta}_1^{(T)}$ and $\hat{\boldsymbol{\beta}}_2^{(T)}$.

where $c_l, c_{snr} > 0$ are some universal constants that and do not grow with $n$ or $p$. The (A1) suggests the initialized estimator should be closed to the truth. Such a condition is common in the MLR literature, see Balakrishnan et al. (2017); Yi et al. (2014); Wang et al. (2015); Yi and Caramanis (2015). In practice, our initialization algorithm is discussed in Section 5. Condition (A2) has also been commonly used in the MLR literature (Balakrishnan et al. 2017; Klusowski et al. 2019), and other hidden variable models (Wang et al. 2015; Cai et al. 2019; Ho et al. 2019).

**Theorem 1.** *Suppose $1/M < \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) < M$ and $\|\boldsymbol{\beta}_1^*\|, \|\boldsymbol{\beta}_2^*\| \le M$ for some constant $M > 1$ and conditions (A1) and (A2) hold. If $\frac{s \log p \cdot \log n}{n} = o(1)$, and $c_{snr}$ in the condition (A2) is sufficiently large, that is, $c_{snr} \ge C(\omega^*, M, c_l)$ where $C(\omega^*, M, c_l)$ is a constant depending only on $(\omega^*, M, c_l)$. Then when $T \gtrsim \log n$, we have with probability at least $1 - O(n^{-1})$,*

$$\|\hat{\boldsymbol{\beta}}_1^{(T)} - \boldsymbol{\beta}_1^*\|_2 + \|\hat{\boldsymbol{\beta}}_2^{(T)} - \boldsymbol{\beta}_2^*\|_2 \lesssim \sqrt{\frac{s \log p \cdot \log n}{n}}.$$

**Remark 1.** The condition on the spectrum of $\boldsymbol{\Sigma}$ is standard in the high-dimensional literature. For example, it has been used in Cai et al. (2016); Cai and Zhou (2012) and Javanmard and Montanari (2014a) for estimation of precision matrices, covariance matrices and regression coefficients, respectively. Further, the convergence rate of optimization error is exponentially fast, so we only need that the number of iterations $T \gtrsim \log n$ to make the optimization error negligible comparing to the statistical error.

**Remark 2.** The prior literature on MLR (Wang et al. 2015; Balakrishnan et al. 2017; Klusowski et al. 2019) mostly focuses on the symmetric setting ($\boldsymbol{\beta}_1^* = -\boldsymbol{\beta}_2^*$) and also assumes that the mixing proportion $\omega^*$, the covariate covariance matrix $\Sigma$, and the noise variance $\sigma^{*2}$ are all known. In contrast, our algorithm above allows unknown $\omega^*$, $\Sigma$, and $\sigma^{*2}$, and we establish the convergence of the proposed EM algorithm in a much more general setting. To achieve this, the estimates of these parameters are updated in each iteration of the M-step, and the proofs rely on a careful analysis of the contraction bounds of these updates and the concentration bounds of the corresponding statistical errors.

**Remark 3.** Our method can also be extended to allow different noise level $\sigma$'s for different groups. In particular, let us denote the noise variance for the two groups by $\sigma_1^2$ and $\sigma_2^2$ respectively. In the $t$-th $M$-Step of Algorithm 1, we first estimate them by

$$(\hat{\sigma}_1^{(t+1)})^2 = \frac{1}{n}\sum_{i=1}^{n} \gamma_{\boldsymbol{\theta},i}^{(t)}(y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_1^{(t+1)}\rangle)^2; \quad (\hat{\sigma}_2^{(t+1)})^2 = \frac{1}{n}\sum_{i=1}^{n}(1-\gamma_{\boldsymbol{\theta},i}^{(t)})(y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_2^{(t+1)}\rangle)^2.$$

Correspondingly, we then modify the estimation of $\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{x}_i, y_i)$ to

$$\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{x}_i, y_i) = \mathbb{P}_{\hat{\boldsymbol{\theta}}^{(t)}}(z_i = 1 \mid \boldsymbol{x}_i, y_i) = \frac{\hat{\omega}^{(t)} \exp(-\frac{(y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_1^{(t)}\rangle)^2}{2\hat{\sigma}^{(t)2}})}{\hat{\omega}^{(t)} \exp(-\frac{(y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_1^{(t)}\rangle)^2}{2\hat{\sigma}^{(t)2}}) + (1-\hat{\omega}^{(t)}) \exp(-\frac{(y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_2^{(t)}\rangle)^2}{2\hat{\sigma}^{(t)2}})}.$$

We will discuss more extensions in details in Section 7.

## 3 DEBIASED ESTIMATORS AND THEIR ASYMPTOTIC NORMALITY

The iterative estimators obtained from the high-dimensional EM algorithm (Algorithm 1) enjoys desirable properties in term of squared error, they are however unsuitable to be used directly for statistical inference. In this section, we introduce the debiased estimators for the MLR coefficients $\boldsymbol{\beta}^*_{\ell j}$ with $\ell \in \{1,2\}$ and $j \in [p]$, and obtain their asymptotic normality, which can then be used to perform hypothesis testing and construct confidence intervals for the individual coefficients.

### 3.1 Debiased Estimators

Due to the $\ell_1$ regularization in the M-step, the outputs $\hat{\boldsymbol{\beta}}_1^{(T)}$ and $\hat{\boldsymbol{\beta}}_2^{(T)}$ from the high-dimensional EM algorithm (Algorithm 1) are biased. To facilitate the subsequent statistical inference, we proceed by correcting their biases. Such a de-biased procedure has been used widely in high-dimensional single linear regression models (Javanmard and Montanari 2014a,b; van de Geer et al. 2014; Zhang and Zhang 2014; Ning and Liu 2017), but cannot be directly applied to the EM solutions. In the following, we first present some high-level intuition.

We start with the regression coefficient $\boldsymbol{\beta}^*_1$. Note that in Algorithm 1, $\hat{\boldsymbol{\beta}}_1^{(T)}$ is constructed only based on the $T$-th sample, and the sample size is $n/T$ with $T \asymp \log n$. In the following, for the notational simplicity, we simply write $n_T = n/T$. Firstly, $\hat{\boldsymbol{\beta}}_1^{(T)}$ satisfies the Karush-Kuhn-Tucker (KKT) condition

$$-\frac{1}{n}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta},i}^{(T)}(y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_1^{(T)}\rangle)\boldsymbol{x}_i + \lambda_n^{(T)}\partial||\hat{\boldsymbol{\beta}}_1^{(T)}||_1 = 0, \tag{3.1}$$

where $\partial||\hat{\boldsymbol{\beta}}_1^{(T)}||_1$ is the subgradient of the $\ell_1$ norm $||\cdot||_1$. Letting $\widehat{\Sigma}_{XX} = \frac{1}{n_T}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta},i}^{(T)}\boldsymbol{x}_i\boldsymbol{x}_i^\top$ and $\widehat{\Sigma}_{XY} = \frac{1}{n_T}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta},i}^{(T)}\boldsymbol{x}_i y_i$, equation (3.1) can then be rewritten as

$$\widehat{\Sigma}_{XX}\hat{\boldsymbol{\beta}}_1^{(T)} - \widehat{\Sigma}_{XY} + \lambda_n^{(T)}\partial||\hat{\boldsymbol{\beta}}_1^{(T)}||_1 = 0,$$

and as a result,

$$\widehat{\Sigma}_{XX}(\hat{\boldsymbol{\beta}}_1^{(T)} - \boldsymbol{\beta}_1^*) + \lambda_n^{(T)}\partial||\hat{\boldsymbol{\beta}}_1^{(T)}||_1 = \widehat{\Sigma}_{XY} - \widehat{\Sigma}_{XX}\boldsymbol{\beta}_1^*.$$

Following the debiased Lasso method in Javanmard and Montanari (2014a), suppose one has a good approximation of the "inverse" of $\widehat{\Sigma}_{XX}$, say $M$, then one can multiply $M$ on the left to obtain

$$M\widehat{\Sigma}_{XX}(\hat{\boldsymbol{\beta}}_1^{(T)} - \boldsymbol{\beta}_1^*) + \lambda_n^{(T)}M\partial||\hat{\boldsymbol{\beta}}_1^{(T)}||_1 = M(\widehat{\Sigma}_{XY} - \widehat{\Sigma}_{XX}\boldsymbol{\beta}_1^*).$$

Then it follows

$$(\hat{\boldsymbol{\beta}}_1^{(T)} + \lambda_n^{(T)} M \partial ||\hat{\boldsymbol{\beta}}_1^{(T)}||_1) - \boldsymbol{\beta}_1^* = M(\widehat{\Sigma}_{XY} - \widehat{\Sigma}_{XX}\boldsymbol{\beta}_1^*) + (I - M\widehat{\Sigma}_{XX})(\hat{\boldsymbol{\beta}}_1^{(T)} - \boldsymbol{\beta}_1^*). \tag{3.2}$$

By inspection, if we let $\widehat{\boldsymbol{\beta}}_1^u = \hat{\boldsymbol{\beta}}_1^{(T)} + \lambda_n^{(T)} M \partial ||\hat{\boldsymbol{\beta}}_1^{(T)}||_1 = (I - M\widehat{\Sigma}_{XX})\hat{\boldsymbol{\beta}}_1^{(T)} + M\widehat{\Sigma}_{XY}$, then

$$\begin{aligned}
\sqrt{n}(\widehat{\boldsymbol{\beta}}_1^u - \boldsymbol{\beta}_1^*) =& \sqrt{n}(M\widehat{\Sigma}_{XY} - M\widehat{\Sigma}_{XX}\boldsymbol{\beta}_1^*) + \sqrt{n}(I - M\widehat{\Sigma}_{XX})(\hat{\boldsymbol{\beta}}_1^{(T)} - \boldsymbol{\beta}_1^*) \\
=& \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} \gamma_{\boldsymbol{\theta},i}^{(T)}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1^* \rangle) M \boldsymbol{x}_i\right] + o_P(1),
\end{aligned}$$

where the second equality incorporated the assumption that $M$ approximate the "inverse" of $\widehat{\Sigma}_{XX}$ well and thus $\|(I - M\widehat{\Sigma}_{XX})(\hat{\boldsymbol{\beta}}_1^{(T)} - \boldsymbol{\beta}_1^*)\|_\infty \leq \|I - M\widehat{\Sigma}_{XX}\|_\infty \|\hat{\boldsymbol{\beta}}_1^{(T)} - \boldsymbol{\beta}_1^*\|_1$ is negligible.

Unlike the procedure in Javanmard and Montanari (2014a), our $\hat{\Sigma}_{XX}$ depends on $(\boldsymbol{x}_i, y_i)$ instead of only on $\boldsymbol{x}_i$'s, and therefore solving a direct approximation will mess up with the subsequent asymptotic normality. We propose to solve $M$ by the following two-step procedure. First, let $\tilde{\Sigma}_{XX} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top$, for $j \in [p]$, let $\tilde{\boldsymbol{m}}_j$ be the solution of

$$\begin{aligned}
\underset{\boldsymbol{m}_j \in \mathbb{R}^p}{\text{minimize}} \quad & \boldsymbol{m}_j^\top \tilde{\Sigma}_{XX} \boldsymbol{m}_j \\
\text{subject to} \quad & ||\tilde{\Sigma}_{XX}\boldsymbol{m}_j - \boldsymbol{e}_j^{(p)}||_\infty \leq \mu, \\
& \|\boldsymbol{m}_j\| \leq C\sqrt{\log n}.
\end{aligned} \tag{3.3}$$

where $\mu$ and $C$ are tuning parameters that will be discussed later.

Second we set $\boldsymbol{m}_j = \tilde{\boldsymbol{m}}_j / \hat{\omega}^{(T)}$, with $\hat{\omega}^{(T)} = \frac{1}{n}\sum_{i=1}^n \gamma_{\boldsymbol{\theta},i}^{(T)}$. The denominator is used because $\tilde{\Sigma}_{XX}$ is approximately $\omega^* \cdot \hat{\Sigma}_{XX}$. Although $\hat{\omega}^{(T)}$ still depends on $(\boldsymbol{x}_i, y_i)$, but it is close to $\omega^*$ and the distance is negligible.

Now, to find the asymptotic distribution of $\sqrt{n_T}(\widehat{\boldsymbol{\beta}}_1^u - \boldsymbol{\beta}_1^*)$, let us consider the dominating term

$$\frac{1}{\sqrt{n_T}}\sum_{i=1}^{n_T} \gamma_{\boldsymbol{\theta},i}^{(T)}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1^* \rangle) M \boldsymbol{x}_i. \tag{3.4}$$

By inspection, conditioning on $\boldsymbol{x}$, (3.4) is an approximation of the linear transformation of the score function $\nabla_{\boldsymbol{\theta}^*} l_n(\boldsymbol{\theta}^*; \boldsymbol{x}, y)$. Specifically, straightforward computation yields the score function

$$\nabla_{\boldsymbol{\beta}_1} l_n(\boldsymbol{\theta}^*; \boldsymbol{x}, y) = \frac{1}{n_T}\sum_{i=1}^{n_T} \gamma_{\boldsymbol{\theta}^*,i}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1^* \rangle) \boldsymbol{x}_i. \tag{3.5}$$

By Theorem 1, $\boldsymbol{\theta}^{(T)}$ is close to $\boldsymbol{\theta}^*$, so heuristically, (3.4) is also close to a linear transformation of the score function (3.5) when $n$ is large. Moreover, since the score function $\nabla l_n(\boldsymbol{\theta}; \boldsymbol{x}, y)$ is asymptotically normal at the truth $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ with covariance matrix being the information matrix $I(\boldsymbol{\theta}^*) = -\mathbb{E}_{\boldsymbol{\theta}^*} \nabla^2 l_n(\boldsymbol{\theta}^*; \boldsymbol{x}, y)$, in order to make valid inference about the parameters, it suffices to estimate the information matrix.

The following Lemma 1 provides the an estimator for information matrix and the corresponding asymptotic distribution of (3.4). Let us define $T_n(\boldsymbol{\theta}) = -\big(\nabla^2_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta} \mid \boldsymbol{\theta}') + \nabla^2_{\boldsymbol{\theta},\boldsymbol{\theta}'} Q_n(\boldsymbol{\theta} \mid \boldsymbol{\theta}') \mid_{\boldsymbol{\theta},\boldsymbol{\theta}'=\boldsymbol{\theta}}\big)$. Recall that $\boldsymbol{\theta} = (\omega, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma)$, $T_n(\boldsymbol{\theta}) \in \mathbb{R}^{(2p+2)\times(2p+2)}$ consists of four submatrices whose coordinates corresponding to $\omega, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2$, and $\sigma$ respectively. We denote the submatrix corresponding to $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ by $T_{\beta,n}(\boldsymbol{\theta}) := [T_n(\boldsymbol{\theta})]_{(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2),(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)} \in \mathbb{R}^{2p\times 2p}$, and compute the component corresponding to $\boldsymbol{\beta}_1$ explicitly as

$$(T_{\beta,n}(\hat{\boldsymbol{\theta}}^{(T)}))_{1,1} = \frac{1}{n_T} \sum_{i=1}^{n} \gamma^{(T)}_{\boldsymbol{\theta},i} \boldsymbol{x}_i \boldsymbol{x}_i^\top + \frac{2}{n_T} \sum_{i=1}^{n_T} \frac{(y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_1^{(T)} \rangle)^2}{\eta(\hat{\boldsymbol{\theta}}^{(T)})} \boldsymbol{x}_i \boldsymbol{x}_i^\top, \tag{3.6}$$

where

$$\begin{aligned} \eta(\hat{\boldsymbol{\theta}}^{(T)}) =& \hat{\sigma}^{(T)2} \left[ \hat{\omega}^{(T)} + (1 - \hat{\omega}^{(T)}) \exp\left\{ \frac{(2y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_1^{(T)} + \hat{\boldsymbol{\beta}}_2^{(T)} \rangle) \cdot \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_2^{(T)} - \hat{\boldsymbol{\beta}}_1^{(T)} \rangle}{2\hat{\sigma}^{(T)2}} \right\} \right] \\ & \times \left[ 1 - \hat{\omega}^{(T)} + \hat{\omega}^{(T)} \exp\left\{ -\frac{(2y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_1^{(T)} + \hat{\boldsymbol{\beta}}_2^{(T)} \rangle) \cdot \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_2^{(T)} - \hat{\boldsymbol{\beta}}_1^{(T)} \rangle}{2\hat{\sigma}^{(T)2}} \right\} \right]. \end{aligned} \tag{3.7}$$

**Lemma 1.** *Let $(T_{\beta,n}(\hat{\boldsymbol{\theta}}^{(T)}))_{1,1}$ be defined as in (3.6). Under the conditions of Theorem 1 and additionally assume $\frac{s \log p \log n \cdot (\sqrt{s} \vee \log^2 p)}{\sqrt{n}} = o(1)$, then for $j \in [p]$, conditioning on $\boldsymbol{x}$, we have*

$$\frac{\langle \boldsymbol{m}_j, \frac{1}{\sqrt{n_T}} \sum_{i=1}^{n_T} \gamma^{(T)}_{\boldsymbol{\theta},i} (y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1^* \rangle) \boldsymbol{x}_i \rangle}{\hat{\sigma}^{(T)} \sqrt{\boldsymbol{m}_j^\top (T_{\beta,n}(\hat{\boldsymbol{\theta}}^{(T)}))_{1,1} \boldsymbol{m}_j}} \xrightarrow{d} N(0,1), \quad \text{as } n \to \infty. \tag{3.8}$$

**Remark 4.** The matrix $T_n(\boldsymbol{\theta})$ provides an estimator of the Fisher information at $\boldsymbol{\theta}^*$ for the conditional distribution of the observations given the latent variables (the membership of the components in MLR). In particular, the first term $\nabla^2_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta} \mid \boldsymbol{\theta}')$ is an estimator of the Fisher information for the joint distribution of the observation and latent variables, and the second term $\nabla^2_{\boldsymbol{\theta},\boldsymbol{\theta}'} Q_n(\boldsymbol{\theta} \mid \boldsymbol{\theta}')$ is the negative expectation of the conditional variance. By the law of total variance, $T_n(\boldsymbol{\theta})$ yields an estimator of the variance of the conditional expectation, that is, the Fisher information for the conditional distribution of the observations given the latent variables. Similar result also appears in Wang et al. (2015) for a simple symmetric two-component MLR model via a decorrelated score statistic. Here we develop a different de-biasing algorithm for a much more general model without the knowledge of $\Sigma, \omega, \sigma$.

Similar arguments can be applied to the regression coefficient $\boldsymbol{\beta}_2^*$. In Algorithm 2, we summarize our proposed method for obtaining the debiased estimators $\widehat{\boldsymbol{\beta}}_1^u$ and $\widehat{\boldsymbol{\beta}}_2^u$.

**Algorithm 2** De-biasing EM for High-Dimensional MLR

1: **Inputs:** $\gamma_{\boldsymbol{\theta},i}^{(T)}, \hat{\boldsymbol{\beta}}_1^{(T)}, \hat{\boldsymbol{\beta}}_2^{(T)}$ and $\lambda_n^{(T)}$ from Algorithm 1, $j \in [p]$, tuning parameter $\mu$, and coverage probability $1-\alpha$.

2: **Precision matrix approximation:** Let $\hat{\omega}^{(T)} = \frac{1}{n_T}\sum_{i=1}^{n_T} \gamma_{\boldsymbol{\theta},i}^{(T)}$, and $\tilde{\Sigma}_{XX} = \frac{1}{n_T}\sum_{i=1}^{n_T} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ and for $j \in [p]$, let $\tilde{\boldsymbol{m}}_j$ be the solution of

$$\begin{aligned} \underset{\boldsymbol{m}_j \in \mathbb{R}^p}{\text{minimize}} \quad & \boldsymbol{m}_j^\top \tilde{\Sigma}_{XX} \boldsymbol{m}_j \\ \text{subject to} \quad & ||\tilde{\Sigma}_{XX}\boldsymbol{m}_j - e_j^{(p)}||_\infty \le \mu; \\ & \|\boldsymbol{m}_j\|_1 \le C\sqrt{\log n}. \end{aligned}$$

Let $\boldsymbol{m}_j = \tilde{\boldsymbol{m}}_j / \hat{\omega}^{(T)}$.

3: **De-biasing:** For $\ell = 1, 2$, let $\widehat{\boldsymbol{\beta}}_{\ell j}^u = \hat{\boldsymbol{\beta}}_{\ell j}^{(T)} + \lambda_n^{(T)} \boldsymbol{m}_j^\top \partial ||\hat{\boldsymbol{\beta}}_\ell^{(T)}||_1$.

4: **Variance estimation**: For $\ell = 1, 2$, let

$$\hat{v}_{\ell j} = \boldsymbol{m}_j^\top \left( (T_n(\hat{\boldsymbol{\theta}}^{(T)}))_{\ell,\ell} \right) \boldsymbol{m}_j,$$

where $(T_n(\hat{\boldsymbol{\theta}}))_{\ell,\ell}$ are defined in (3.6) and (3.9).

5: Output $(\widehat{\boldsymbol{\beta}}_{1j}^u, \hat{v}_{1j})$ and $(\widehat{\boldsymbol{\beta}}_{2j}^u, \hat{v}_{2j})$.

The following theorem establishes the asymptotic normality of the the individual components of these debiased estimators, which can be directly used for performing hypotheses testing or constructing confidence intervals.

**Theorem 2.** *Suppose the conditions of Theorem 1 hold. We further assume that* $\|\boldsymbol{\Sigma}\|_2, \|\boldsymbol{\Sigma}^{-1}\|_1 \le L$ *for some* $L > 0$ *and* $\frac{s \log p \log n \cdot (\sqrt{s} \vee \log^2 p)}{\sqrt{n}} = o(1)$*, and the tuning parameters* $C = L$, $\mu = C'\sqrt{\frac{\log p}{n}} \log n$*, for some universal constant* $C' > 0$*. Let* $\sqrt{\hat{v}_{\ell j}} = \hat{\sigma}^{(T)} \sqrt{\boldsymbol{m}_j^\top (T_{\beta,n}(\hat{\boldsymbol{\theta}}^{(T)}))_{\ell,\ell} \boldsymbol{m}_j}$ *for*

$j \in [p]$ *and* $\ell = 1, 2$. *Then conditioning on* $\boldsymbol{x}$, *as* $n \to \infty$,

$$\frac{\sqrt{n_T}\left(\widehat{\boldsymbol{\beta}}^u_{\ell j} - \boldsymbol{\beta}^*_{\ell j}\right)}{\sqrt{\hat{v}_{\ell j}}} \xrightarrow{d} N(0,1).$$

In MLR, it is also of interest to make inference about the differential parameter $(\boldsymbol{\beta}^*_{1j} - \boldsymbol{\beta}^*_{2j})$ for some given $j \in [p]$. To this end, consider its natural estimator $(\widehat{\boldsymbol{\beta}}^u_{1j} - \widehat{\boldsymbol{\beta}}^u_{2j})$. It can be shown that a consistent estimator for the variance of $(\widehat{\boldsymbol{\beta}}^u_{1j} - \widehat{\boldsymbol{\beta}}^u_{2j})$ is given by

$$\tilde{v}_j = \hat{\sigma}^{(T)2} \cdot \boldsymbol{m}_j^\top \left( (T_{\beta,n}(\hat{\boldsymbol{\theta}}^{(T)}))_{1,1} + (T_{\beta,n}(\hat{\boldsymbol{\theta}}^{(T)}))_{2,2} - (T_{\beta,n}(\hat{\boldsymbol{\theta}}^{(T)}))_{1,2} - (T_{\beta,n}(\hat{\boldsymbol{\theta}}^{(T)}))_{2,1} \right) \boldsymbol{m}_j,$$

where $(T_{\beta,n}(\hat{\boldsymbol{\theta}}^{(T)}))_{1,1}$ is given in (3.6) and

$$(T_{\beta,n}(\hat{\boldsymbol{\theta}}^{(T)}))_{2,2} = \frac{1}{n_T}\sum_{i=1}^n (1 - \gamma^{(T)}_{\boldsymbol{\theta},i}) \boldsymbol{x}_i \boldsymbol{x}_i^\top + \frac{2}{n_T}\sum_{i=1}^n \frac{(y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}^{(T)}_2 \rangle)^2}{\eta(\hat{\boldsymbol{\theta}}^{(T)})} \boldsymbol{x}_i \boldsymbol{x}_i^\top, \tag{3.9}$$

$$(T_{\beta,n}(\hat{\boldsymbol{\theta}}^{(T)}))_{2,1} = (T_{\beta,n}(\hat{\boldsymbol{\theta}}^{(T)}))_{1,2} = \frac{2}{n_T}\sum_{i=1}^n \frac{(\langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}^{(T)}_1 \rangle - y_i) \cdot (y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}^{(T)}_2 \rangle)}{\eta(\hat{\boldsymbol{\theta}}^{(T)})} \boldsymbol{x}_i \boldsymbol{x}_i^\top, \tag{3.10}$$

where $\eta(\hat{\boldsymbol{\theta}}^{(T)})$ is computed via (3.7).

Similar to Theorem 2, the following theorem establishes the asymptotic normality of the estimator $(\widehat{\boldsymbol{\beta}}^u_{1j} - \widehat{\boldsymbol{\beta}}^u_{2j})$.

**Theorem 3.** *Suppose the conditions of Theorem 1 hold. We further assume that* $\|\Sigma^{-1}\|_1 \leq L$ *for some* $L > 0$ *and* $\frac{s \log p \log n \cdot (\sqrt{s} \vee \log^2 p)}{\sqrt{n}} = o(1)$. *Then for any* $j \in [p]$, *conditioning on* $\boldsymbol{x}$, *as* $n \to \infty$,

$$\frac{\sqrt{n_T}\left((\widehat{\boldsymbol{\beta}}^u_{1j} - \widehat{\boldsymbol{\beta}}^u_{2j}) - (\boldsymbol{\beta}^*_{1j} - \boldsymbol{\beta}^*_{2j})\right)}{\sqrt{\tilde{v}_j}} \xrightarrow{d} N(0,1).$$

### 3.2 Asymptotic Confidence Intervals

Given the asymptotic normality established in Theorems 2 and 3, we are now ready to present the confidence intervals for the individual coordinates $\beta_{lj}$'s for $j \in [p]$ and $l = 1, 2$ and the differential parameters $(\boldsymbol{\beta}^*_{1j} - \boldsymbol{\beta}^*_{2j})$ for $j \in [p]$. Specifically, let

$$I^{(ind)}_{lj} = [\widehat{\boldsymbol{\beta}}^u_{\ell j} - z_{\alpha/2}\sqrt{\hat{v}_{\ell j}},\ \ \widehat{\boldsymbol{\beta}}^u_{\ell j} + z_{\alpha/2}\sqrt{\hat{v}_{\ell j}}], \quad \text{for } j \in [p] \text{ and } l = 1, 2,$$

and

$$I_j^{(dif)} = [(\widehat{\boldsymbol{\beta}}_{1j}^u - \widehat{\boldsymbol{\beta}}_{2j}^u) - z_{\alpha/2}\sqrt{\tilde{v}_j}, (\widehat{\boldsymbol{\beta}}_{1j}^u - \widehat{\boldsymbol{\beta}}_{2j}^u) + z_{\alpha/2}\sqrt{\tilde{v}_j}], \quad \text{for } j \in [p],$$

where $z_{\alpha/2}$ is the $\alpha/2$-th quantile of a standard normal distribution. The following theorem provides the asymptotic guarantee for the validity of these confidence intervals.

**Theorem 4.** *Under the conditions of Theorem 2, the confidence intervals $I_{lj}^{(ind)}$ and $I_j^{(dif)}$ are asymptotically valid, that is,*

$$\lim_{n\to\infty} \mathbb{P}(\boldsymbol{\beta}_{lj} \in I_{lj}^{(ind)}) = 1 - \alpha, \quad \textit{for } j \in [p] \textit{ and } l = 1, 2;$$
$$\lim_{n\to\infty} \mathbb{P}(\boldsymbol{\beta}_{1j} - \boldsymbol{\beta}_{2j} \in I_j^{(dif)}) = 1 - \alpha, \quad \textit{for } j \in [p].$$

# 4 LARGE-SCALE MULTIPLE TESTING

## 4.1 The Multiple Testing Procedure

In this section, we consider simultaneous testing of the following null hypotheses

$$H_{0j} : \boldsymbol{\beta}_{1j}^* = \boldsymbol{\beta}_{2j}^* = 0, \quad 1 \le j \le p.$$

Apart from identifying as many nonzero coordinates as possible, to obtain results of practical interest, we would also like to control the false discovery rate (FDR) as well as the false discovery proportion (FDP).

Specifically, since each individual hypothesis $H_{0j}$ is a composition of two hypotheses with $H_{0j} = H_{0j}^{(1)} \cap H_{0,j}^{(2)}$ where $H_{0j}^{(\ell)} : \boldsymbol{\beta}_{\ell j}^* = 0$, we can construct statistics

$$T_j^{(\ell)} = \frac{\hat{\boldsymbol{\beta}}_{\ell j}^u}{\hat{v}_{\ell j}/\sqrt{n}}, \quad \text{for } j = 1, ..., p \text{ and } \ell = 1, 2.$$

For a given threshold level $t > 0$, each individual partial hypothesis $H_{0j}^{(\ell)} : \boldsymbol{\beta}_{\ell j}^* = 0$ is rejected if $|T_j^{(\ell)}| \ge t$. Hence if we propose a test statistic

$$T_j = \max\{|T_j^{(1)}|, |T_j^{(2)}|\}$$

for each null hypothesis $H_{0j}$ and we reject $H_{0j}$ whenever $T_j \geq t$, then for each $t$, we can define

$$\text{FDP}(t) = \frac{\sum_{j\in\mathcal{H}_0} I\{T_j \geq t\}}{\max\left\{\sum_{j=1}^p I\{T_j \geq t\}, 1\right\}}, \qquad \text{FDR}(t) = \mathbb{E}[\text{FDP}(t)].$$

In order to control the FDR/FDP at a pre-specified level $0 < \alpha < 1$, we can set the threshold level as

$$\tilde{t}_1 = \inf\left\{0 \leq t \leq b_p : \frac{\sum_{j\in\mathcal{H}_0} I\{T_j \geq t\}}{\max\left\{\sum_{j=1}^p I\{T_j \geq t\}, 1\right\}} \leq \alpha\right\} \tag{4.1}$$

for some $b_p$ to be determined later. In general, the ideal choice $\tilde{t}_1$ is unknown since it depends on the knowledge of the true null $\mathcal{H}_0$. Inspired by the Gaussian approximation idea proposed by Liu (2013), we first substitute the numerator in (4.1) by its upper bound

$$\sum_{j\in\mathcal{H}_0} I\{T_j \geq t\} \leq \sum_{j\in\mathcal{H}_0} I\{|T_j^{(1)}| \geq t\} + \sum_{j\in\mathcal{H}_0} I\{|T_j^{(2)}| \geq t\},$$

and then use Gaussian tails to approximate the counts $\sum_{j\in\mathcal{H}_0} I\{|T_j^{(\ell)}| \geq t\}$ for $\ell = 1, 2$. Specifically, let $G_0^{(\ell)}(t)$ be an estimate of the proportion of the nulls falsely rejected by the test $I\{|T_j^{(\ell)}| \geq t\}$ among all the true nulls at the threshold level $t$, so that

$$G_0^{(\ell)}(t) = \frac{1}{|\mathcal{H}_0|}\sum_{j\in\mathcal{H}_0} I\{|T_j^{(\ell)}| \geq t\}, \qquad \ell = 1, 2. \tag{4.2}$$

Let $G(t) = 2 - 2\Phi(t)$ be the tails of normal distribution. We will show that, asymptotically, we can use $G(t)$ to approximate $G_0^{(\ell)}(t)$ for $\ell = 1, 2$. Therefore, we have the following multiple testing procedure controlling the FDR and the FDP.

**Procedure 1.** *Let $0 < \alpha < 1$, $b_p = \sqrt{2\log p - 2\log\log p}$ and define*

$$\hat{t} = \inf\left\{0 \leq t \leq b_p : \frac{pG(t)}{\max\left\{\sum_{j=1}^p I\{|T_j| \geq t\}, 1\right\}} \leq \alpha/2\right\}. \tag{4.3}$$

*If $\hat{t}$ in (4.3) does not exist, then let $\hat{t} = \sqrt{2\log p}$. We reject $H_{0,j}$ whenever $|T_j| \geq \hat{t}$.*

### 4.2 Theoretical Properties

For $\ell = 1, 2$, let $\Gamma_\ell = M^\top(\mathbb{E}_{\boldsymbol{\theta}^*}[T_n(\boldsymbol{\theta}^*)])_{\ell,\ell}M$ where $M$ has its $j$-th column as $\boldsymbol{m}_j$ and let $D_\ell$ be the diagonal of $\Gamma_\ell$. We define $D_\ell^{-1/2}\Gamma_\ell D_\ell^{-1/2} = (\rho_{jk}^{(\ell)})_{1\leq j,k\leq p}$ and denote $\mathcal{B}_\ell(\delta) = \{(j,k) : |\rho_{jk}^{(\ell)}| \geq \delta, i \neq j\}$ and $\mathcal{A}_\ell(\epsilon) = \mathcal{B}_\ell((\log p)^{-2-\epsilon})$.

(A3). Suppose that for any $\ell = 1, 2$, there is some $\epsilon > 0$ and $q > 0$, such that

$$\sum_{j,k \in \mathcal{H}_0 : (j,k) \in \mathcal{A}_\ell(\epsilon)} p^{\frac{2|\rho_{jk}^{(\ell)}|}{1+|\rho_{jk}^{(\ell)}|}+q} = O(p^2/(\log p)^2).$$

The following theorem shows the asymptotic control of FDR and FDP of our procedure.

**Theorem 5.** *Under the conditions of Theorem 2, if (A3) holds, then for $\hat{t}$ defined in Procedure 1,*

$$\lim_{(n,p)\to\infty} \frac{\mathrm{FDR}(\hat{t})}{\alpha p_0/p} \le 1, \qquad \lim_{(n,p)\to\infty} \mathbb{P}\left( \frac{\mathrm{FDP}(\hat{t})}{\alpha p_0/p} \le 1+\epsilon \right) = 1 \tag{4.4}$$

*for any $\epsilon > 0$.*

# 5 SIMULATION STUDIES

In this section, we evaluate the numerical performance of the proposed methods. For both estimation and large-scale multiple testing, the empirical results in various settings demonstrate the numerical advantages of the proposed procedures over alternative methods.

## 5.1 Estimation

For estimation, we let the dimension of the covariates $p$ range from 600 to 1000, the sparsity $s$ vary from 10 to 30, and set the sample size $n = 400$. We also set the mixture proportion $\omega^* = 0.3$ and the noise level $\sigma^2 = 1$. The design covariates $\boldsymbol{x}_i$'s are generated from a multivariate Gaussian distribution with covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_M$, where $\boldsymbol{\Sigma}_M$ is a $p \times p$ blockwise diagonal matrix of $10$ identical unit diagonal Toeplitz matrices whose off-diagonal entries descend from 0.4 to 0 (see Supplementary Materials for the explicit form). For the two regression coefficients $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$, for some fixed $\rho > 0$, we set $\boldsymbol{\beta}_{1j}^* = \rho \cdot 1\{1 \le j \le s\}$ and $\boldsymbol{\beta}_{2j}^* = -\rho \cdot 1\{p/2 + 1 \le j \le p/2 + s\}$ so that each of the coefficient vectors is $s$-sparse.

In particular, for our proposed methods, as of practical interest, we assume that the noise level $\sigma^2$ is unknown and also needs to be estimated at each iteration. Throughout, we set $T = 30$, $\kappa = 0.3$ and $C = 0.8$.

We consider two initializations for our proposed algorithm. We start with fitting a Lasso to the mixed samples, which results to a coarse but useful variable screening. Combining the response variable $y$ and the Lasso selected covariates, we use one of the following high-dimensional clustering methods to divide the samples. In particular, our two initialization corresponds to the `emgm` function in the R package `xLLiM`, and the `hddc` algorithm in the R package `HDclassif`. Once we obtain an initial two-group clustering of samples, we can fit the Elastic Net separately using the samples within each group. The resulting regression coefficients will be used as initial values $\hat{\boldsymbol{\beta}}_1^{(0)}$ and $\hat{\boldsymbol{\beta}}_2^{(0)}$, respectively. For the above Elastic Net algorithm (Zou and Hastie 2005), we set the elasticnet mixing parameter as $0.5$.

We evaluate and compare the empirical performances of 1) GLLiM: the Gaussian Locally Linear Mapping EM algorithm proposed by Deleforge et al. (2015), which is implemented by the `gllim` function in the R package `xLLiM`; 2) Initial1: fit Elastic Net separately to the clusters determined by Lasso+`emgm`; 3) Initial2: fit Elastic Net separately to the clusters determined by Lasso+`hddc`; 4) MIREM1 : our proposed algorithm based on initialization Initial1; and 5) MIREM2: our proposed algorithm based on initialization Initial2.

The estimation performance is evaluated using the empirical mean-squared error (EMSE): for $N$ rounds of simulations and estimators $(\hat{\boldsymbol{\beta}}_1^r, \hat{\boldsymbol{\beta}}_2^r)$ obtained in the $r$-th round, we define

$$\text{EMSE} = \min\left\{\frac{1}{N}\sum_{r=1}^{N}[\|\hat{\boldsymbol{\beta}}_1^r - \boldsymbol{\beta}_1^*\|_2 + \|\hat{\boldsymbol{\beta}}_2^r - \boldsymbol{\beta}_2^*\|_2], \frac{1}{N}\sum_{r=1}^{N}[\|\hat{\boldsymbol{\beta}}_1^r - \boldsymbol{\beta}_2^*\|_2 + \|\hat{\boldsymbol{\beta}}_2^r - \boldsymbol{\beta}_1^*\|_2]\right\}.$$

In Table 1, we show the EMSEs calculated from $N = 500$ rounds of simulations. We observe that both MIREM1 and MIREM2 outperform the other three methods across almost all the settings. As dimension $p$, the sparsity $s$, or the signal magnitude $\rho$ increases, all the methods show increased estimation errors. In addition, comparing our proposed methods MIREM1 and MIREM2, we find that MIREM1 has better performance than MIREM2 in almost all the settings. The EM based GLLiM method, with 100 iterations, performs slightly better than our initializations, but our proposed MIREM algorithms, with only $T = 30$ iterations, have superior performance, suggesting significant improvement upon the initial estimators.

Table 1: Comparison of empirical mean-squared error (EMSE) of different methods with $\omega^* = 0.3$ and $n = 400$

| | $\rho = 0.45$ | | | | | $\rho = 0.85$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $p =$ | 600 | 700 | 800 | 900 | 1000 | 600 | 700 | 800 | 900 | 1000 |
| $s = 10$ | | | | | | | | | | |
| GLLiM | 2.86 | 2.95 | 3.17 | 3.20 | 3.26 | 5.09 | 5.11 | 5.12 | 5.13 | 5.16 |
| Initial1 | 3.19 | 3.00 | 3.12 | 3.12 | 3.27 | 6.07 | 5.94 | 6.00 | 6.06 | 6.27 |
| Initial2 | 2.43 | 2.43 | 2.50 | 2.41 | 2.59 | 5.08 | 5.20 | 5.22 | 5.06 | 4.76 |
| MIREM1 | 1.40 | 1.40 | 1.42 | 1.42 | 1.43 | 1.18 | 1.18 | 1.18 | 1.21 | 1.23 |
| MIREM2 | 1.73 | 1.81 | 1.75 | 1.79 | 1.81 | 2.85 | 2.17 | 2.29 | 2.61 | 2.66 |
| $s = 15$ | | | | | | | | | | |
| GLLiM | 3.16 | 3.21 | 3.27 | 3.30 | 3.34 | 6.26 | 6.29 | 6.35 | 6.31 | 6.37 |
| Initial1 | 4.19 | 4.21 | 4.21 | 4.21 | 4.40 | 9.04 | 9.02 | 9.01 | 8.99 | 8.97 |
| Initial2 | 3.21 | 3.15 | 3.25 | 3.16 | 3.08 | 8.68 | 7.66 | 8.35 | 8.00 | 7.71 |
| MIREM1 | 1.42 | 1.47 | 1.45 | 1.47 | 1.49 | 1.26 | 1.31 | 1.62 | 1.57 | 1.56 |
| MIREM2 | 1.92 | 1.98 | 2.04 | 2.02 | 1.94 | 3.36 | 3.65 | 3.48 | 3.75 | 4.03 |
| $s = 20$ | | | | | | | | | | |
| GLLiM | 3.63 | 3.66 | 3.66 | 3.68 | 3.73 | 7.31 | 7.32 | 7.34 | 7.32 | 7.35 |
| Initial1 | 5.45 | 5.38 | 5.68 | 5.62 | 5.52 | 12.18 | 12.11 | 12.09 | 12.15 | 11.89 |
| Initial2 | 4.35 | 3.92 | 4.11 | 4.39 | 4.18 | 11.80 | 10.39 | 10.61 | 10.81 | 10.59 |
| MIREM1 | 1.57 | 1.58 | 1.60 | 1.56 | 1.61 | 1.77 | 2.43 | 2.15 | 1.77 | 2.14 |
| MIREM2 | 2.49 | 2.45 | 2.32 | 2.34 | 2.41 | 5.14 | 4.91 | 5.19 | 5.23 | 5.31 |
| $s = 25$ | | | | | | | | | | |
| GLLiM | 4.08 | 4.08 | 4.14 | 4.11 | 4.14 | 8.23 | 8.23 | 8.24 | 8.24 | 8.24 |
| Initial1 | 6.96 | 7.04 | 6.91 | 6.99 | 6.88 | 15.53 | 15.29 | 15.22 | 15.08 | 15.24 |
| Initial2 | 5.68 | 6.02 | 5.74 | 5.57 | 5.78 | 13.74 | 13.85 | 13.86 | 13.48 | 13.35 |
| MIREM1 | 1.82 | 1.71 | 1.73 | 1.74 | 1.75 | 3.60 | 4.41 | 3.93 | 4.81 | 5.70 |
| MIREM2 | 3.00 | 2.88 | 2.84 | 3.10 | 2.99 | 7.98 | 7.26 | 6.71 | 7.40 | 7.02 |
| $s = 30$ | | | | | | | | | | |
| GLLiM | 4.51 | 4.52 | 4.53 | 4.58 | 4.56 | 9.06 | 9.07 | 9.04 | 9.05 | 9.08 |
| Initial1 | 8.35 | 8.30 | 8.55 | 8.59 | 8.55 | 18.63 | 18.46 | 18.17 | 18.35 | 17.77 |
| Initial2 | 7.61 | 7.30 | 7.50 | 7.35 | 7.75 | 16.41 | 17.10 | 16.79 | 16.24 | 15.44 |
| MIREM1 | 1.99 | 1.94 | 1.95 | 1.91 | 1.95 | 5.36 | 8.34 | 6.76 | 10.50 | 8.79 |
| MIREM2 | 3.64 | 3.50 | 3.48 | 3.68 | 3.89 | 8.32 | 9.26 | 9.65 | 9.66 | 9.00 |

## 5.2 Large-scale Multiple Testing and FDR Control

In this section, the empirical performance of the proposed multiple testing procedure is evaluated under different settings. Specifically, we vary the number of covariates $p$ from 800 to 1000, the sparsity level $s$ from 10, 15 to 20, and set the sample size $n$ as 300 or 400. The two regression coefficients $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$, the design covariates, the mixing proportion $\omega^*$ and the number of iterations

$T$ are the same as previous simulations with $\rho = 0.45$. About our proposed method, in light of the results from the previous section, we will focus on MIREM1 instead of MIREM2 for its superior performance across most settings. To the best of our knowledge, there is no existing method for multiple testing in MLR models. So we compare the empirical FDRs and powers of our proposed testing procedure to the Benjamini-Yekutieli (B-Y) procedure (Benjamini and Yekutieli 2001) applied to our proposed test statistics for individual tests. To illustrate the necessity of fitting an MLR model when the underlying model is indeed a mixture, we also evaluate the performance of the multiple testing procedure based on ordinary debiased Lasso estimators (Javanmard and Javadi 2019), denoted as dLasso, designed for the linear regression models.

Table 2: Empirical powers and FDRs with $\alpha = 0.1$, $\omega^* = 0.3$ and $n = 300$

| | Powers | | | | | FDRs | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $p =$ | 800 | 850 | 900 | 950 | 1000 | 800 | 850 | 900 | 950 | 1000 |
| | | | | | $s = 10$ | | | | | |
| MIREM1 | 0.459 | 0.459 | 0.430 | 0.465 | 0.441 | 0.009 | 0.009 | 0.014 | 0.012 | 0.025 |
| B-Y | 0.344 | 0.344 | 0.290 | 0.346 | 0.332 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| dLasso | 0.934 | 0.934 | 0.930 | 0.918 | 0.896 | 0.958 | 0.958 | 0.960 | 0.963 | 0.965 |
| | | | | | $s = 15$ | | | | | |
| MIREM1 | 0.582 | 0.592 | 0.563 | 0.623 | 0.609 | 0.022 | 0.024 | 0.028 | 0.028 | 0.030 |
| B-Y | 0.510 | 0.530 | 0.484 | 0.550 | 0.551 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| dLasso | 0.897 | 0.922 | 0.916 | 0.914 | 0.901 | 0.946 | 0.948 | 0.951 | 0.953 | 0.956 |
| | | | | | $s = 20$ | | | | | |
| MIREM1 | 0.724 | 0.744 | 0.723 | 0.756 | 0.778 | 0.088 | 0.086 | 0.071 | 0.089 | 0.110 |
| B-Y | 0.621 | 0.635 | 0.617 | 0.657 | 0.672 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| dLasso | 0.882 | 0.909 | 0.897 | 0.894 | 0.896 | 0.936 | 0.937 | 0.942 | 0.945 | 0.947 |

From Tables 2 and 3, we find that the B-Y procedure and our proposed multiple testing procedure are both able to control the FDR below or around the nominal level $\alpha = 0.1$, whereas the dLasso fails to control the FDR, as a consequence of its inability to capture the mixture structure. In particular, the empirical FDRs of our proposed test procedure are closer to the nominal level $\alpha$ in comparison to the rather conservative B-Y procedure, yielding improved empirical powers of our proposed method across all the settings. In particular, by inspecting the intermediate steps of the dLasso method, we found that, due to the failure to account for the mixture components, dLasso

Table 3: Empirical powers and FDRs with $\alpha = 0.1$, $\omega^* = 0.3$ and $n = 400$

| | Powers | | | | | FDRs | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $p =$ | 800 | 850 | 900 | 950 | 1000 | 800 | 850 | 900 | 950 | 1000 |
| | | | | | $s = 10$ | | | | | |
| MIREM1 | 0.864 | 0.805 | 0.774 | 0.796 | 0.846 | 0.046 | 0.017 | 0.036 | 0.015 | 0.066 |
| B-Y | 0.849 | 0.779 | 0.748 | 0.768 | 0.833 | 0.001 | <0.001 | <0.001 | <0.001 | 0.001 |
| dLasso | 0.977 | 0.973 | 0.975 | 0.985 | 0.994 | 0.943 | 0.957 | 0.957 | 0.961 | 0.962 |
| | | | | | $s = 15$ | | | | | |
| MIREM1 | 0.847 | 0.863 | 0.859 | 0.859 | 0.877 | 0.044 | 0.040 | 0.028 | 0.052 | 0.044 |
| B-Y | 0.825 | 0.842 | 0.843 | 0.834 | 0.857 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| dLasso | 0.964 | 0.968 | 0.968 | 0.965 | 0.973 | 0.945 | 0.949 | 0.952 | 0.954 | 0.956 |
| | | | | | $s = 20$ | | | | | |
| MIREM1 | 0.933 | 0.914 | 0.935 | 0.911 | 0.920 | 0.105 | 0.125 | 0.109 | 0.104 | 0.112 |
| B-Y | 0.905 | 0.873 | 0.900 | 0.877 | 0.889 | 0.001 | 0.002 | 0.001 | 0.001 | 0.001 |
| dLasso | 0.983 | 0.969 | 0.969 | 0.974 | 0.975 | 0.932 | 0.941 | 0.942 | 0.947 | 0.947 |

significantly underestimates the standard errors of the debiased Lasso estimators and therefore the individual $p$-values, which explains the anticonservativeness of the dLasso method.

# 6 ANALYSIS OF A MULTIPLEX IMAGE CYTOMETRY DATASET

In this section, we apply the proposed methods to analyze a multiplex image cytometry dataset studied by Schapiro et al. (2017). Specifically, our dataset contains cellular phenotypes visualized by the imaging mass cytometry (IMC). By pairing classic immunohistochemistry staining, high-resolution tissue laser ablation, and mass cytometry, IMC can measure abundances of more than 40 unique metal-isotope-labeled tissue-bound antibodies simultaneously at a resolution comparable to that of fluorescence microscopy. Schapiro et al. (2017) analyzed images collected from 49 diverse breast cancer samples and 3 matched normal tissues, with each image containing cells whose number varies from 266 to 1,454. Among the 30 cellular phenotypes, there are expression levels of 20 different epitopes (e.g., vimentin; and CD68) or combinations of markers (e.g., proliferative Ki-67+ and phospho-S6+). An initial analysis of our image cytometry datasets using tSNE indicates strong evidences of population heterogeneity among the cells within each of the images (see Figure 1 for some examples).

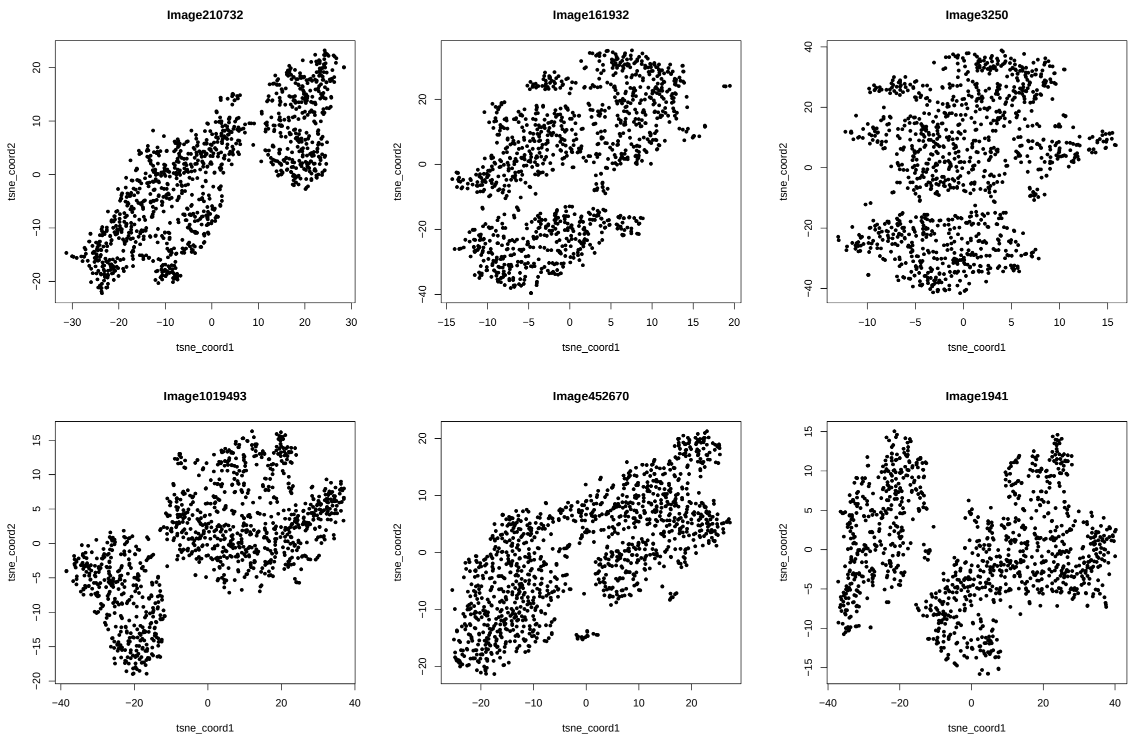


Figure 1: tSNE plots of cells in six randomly selected images/samples based on the expression levels of 30 epitopes or protein markers.

We focus on analyzing the conditional dependence network among these 30 protein epitopes and markers based on the single cell data for each of the samples or images. It is well known that the conditional dependence network can be modeled by Gaussian graphical model, which can be obtained using node-wise regression (Meinshausen and Bühlmann 2006; Yuan and Lin 2007). In other words, to obtain the dependence between two variables $X$ and $Y$ conditioning on all the other variables $(Z_1, ..., Z_d)$, it suffices to perform a linear regression between $(X, Z_1, ..., Z_d)$ against $Y$ and assess the coefficient of $X$. The construction of the conditional dependence network thus requires fitting such linear regressions over all the possible variable configurations. However, in our image cytometry data, heterogeneity among different cell-types may induce a mixture of different dependence structures. To address such an issue, we apply our proposed methods based on the sparse MLR model instead of the ordinary sparse linear regression model for network construction.

As an example, we first focus on the Image 210732. In addition to the global heterogeneity, we also observed that the marginal associations between many pairs of epitopes or markers contain a two-class mixture pattern, as shown in Figure 2. To obtain a conditional dependence network based on the 1,151 cells in this image, we fitted node-wise MLRs and for each of them performed the proposed multiple testing procedure with FDR $< 10\%$. The final network (Figure, 3, top left) was constructed such that the edges indicate the identified associations from at least one of such node-wise regressions. To better illustrate the effects of mixture, the widths of the edges were set to be proportional to the $\ell_2$ distances between the two mixed regression coefficients, so that a thicker edge indicates a larger discrepancy between the two mixtures. As a comparison, we also obtained a network (Figure 3, top right) based on the standard node-wise Lasso and the multiple testing procedure of Javanmard and Javadi (2019) with the same FDR level. Similar to our simulation results, the standard Lasso-based methods tend to report many more associations than our proposed method, due to its failure to account for the underlying mixtures and the resulting underestimated $p$-values for the individual tests. In particular, we found that many heterogeneous associations shown by scatter plots in Figure 2 were indeed captured by our methods as the thicker edges in the network estimated by our mixture model.

Naturally, the above analysis can be conducted similarly for each of the images. Here we present the results for Image 1941 and Image 452670, as two additional examples. With the global heterogeneity shown in Figure 1, again we obtained much denser networks from the standard Lasso based method and sparser networks from our proposed method (both with FDR $< 10\%$ for the node-wise regressions). Moreover, many marginal associations (Figure 2) with heterogeneous associations had thicker edges of the networks based on our proposed method. Our analysis also suggests that the naive application of Lasso to heterogeneous datasets can lead to false associations.

# 7 EXTENSION

In Section 2.2, we establish the convergence rate of the proposed algorithm by assuming that the SNR $= \sqrt{(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*)}$ is a sufficiently large constant (high-SNR regime). The convergence of EM algorithms under all regimes of SNR in the classic low-dimensional setting has been studied in Kwon et al. (2021), where a symmetric two-component MLR model is studied. In

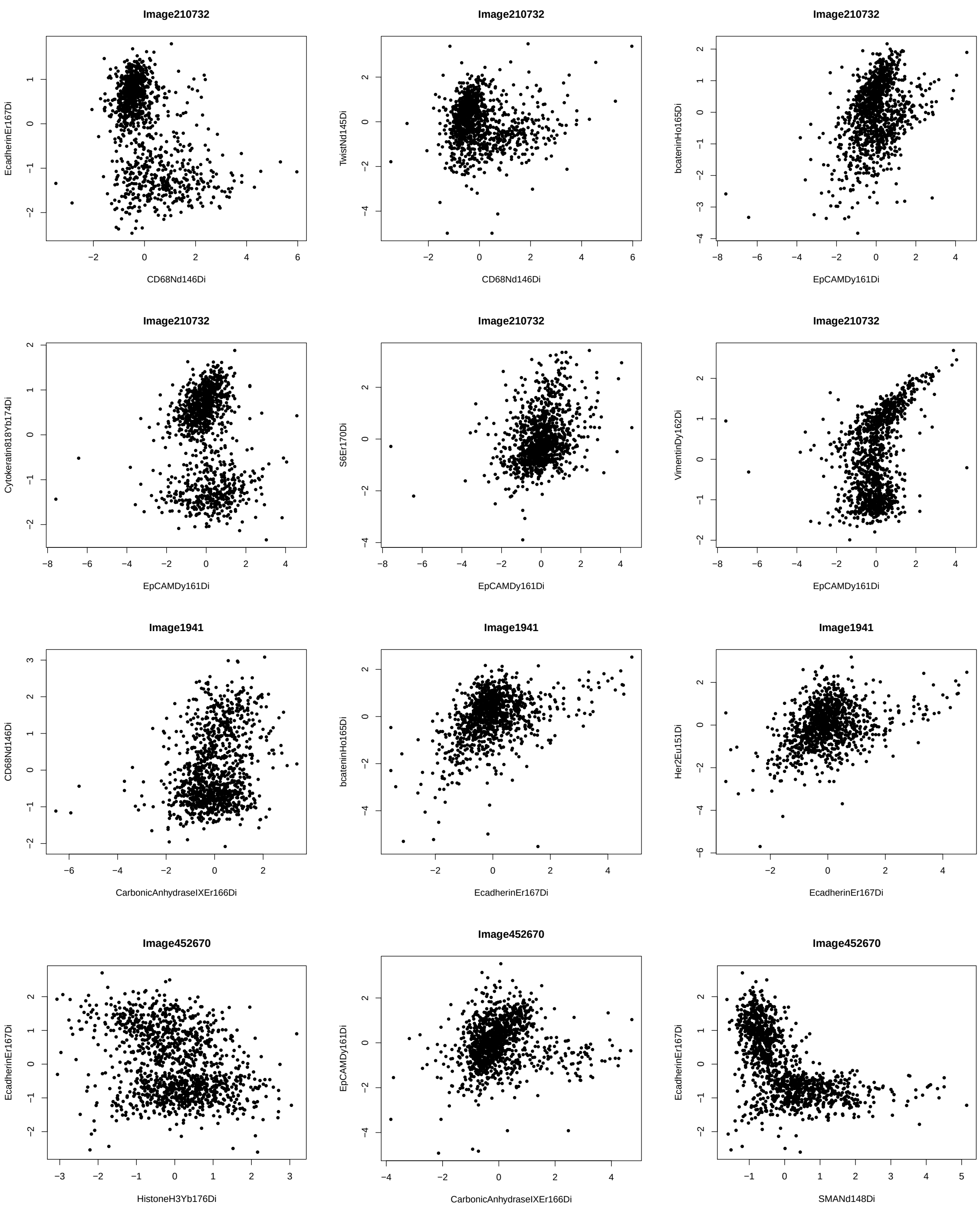


Figure 2: Pairwise scatter plots of epitope expressions for cells on different images, showing mixture of associations.

the high-dimensional setting, however, the results under all regimes of SNR is not well established.

In this section, we address this problem by investigating the following model

$$y_i = \begin{cases} \boldsymbol{x}_i^\top \boldsymbol{\beta}_1^* + \epsilon_i & \text{with probability } 1/2, \\ \boldsymbol{x}_i^\top \boldsymbol{\beta}_2^* + \epsilon_i & \text{with probability } 1/2, \end{cases} \quad i = 1, 2, \ldots, n, \quad \boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^* \in \mathbb{R}^p, \tag{7.1}$$

where the random design variables $\boldsymbol{x}_i \in \mathbb{R}^p$ are $i.i.d.$ samples from $N_p(0, I)$, $\boldsymbol{\beta}_1^* \neq \boldsymbol{\beta}_2^*$ are latent regression coefficients with $\|\boldsymbol{\beta}_1^*\|_0, \|\boldsymbol{\beta}_2^*\|_0 \leq s$, and the noise $\epsilon_i$ is $i.i.d.$ from $N(0, \sigma^2)$ for some known $\sigma > 0$. In addition to the high-SNR regime studied in Section 2.2, we consider in this section the convergence rates in the medium-SNR regime where the SNR is above $O((\frac{s \log p \log^2 n}{n})^{1/6})$ and below a constant, and the low-SNR regime where the SNR is below $O((\frac{s \log p \log^2 n}{n})^{1/6})$. We will use the term *medium-to-high SNR regime* to denote the regime where the SNR is above $O((\frac{s \log p \log^2 n}{n})^{1/6})$.

We remark here that we consider the case where the mixing weights are assumed to be balanced. As noted in Kwon et al. (2021), even in the simpler classic low-dimensional setting, the unknown mixing weights would completely change the population landscape of two-component MLR, and makes the analysis much more challenging in the medium- and low-SNR regimes.

Under model (7.1), we propose the following high-dimensional EM algorithm by combining a symmetrization step and the framework developed in Section 2. Without loss of generality, we assume the sample size is $2n$. We first randomly split the dataset into two subsets of size $n$, and use the first half data to estimate $\boldsymbol{\mu}^* := (\boldsymbol{\beta}_1^* + \boldsymbol{\beta}_2^*)/2$ by solving the following optimization

$$\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \boldsymbol{x}_i, \boldsymbol{\mu} \rangle)^2 + \tilde{\lambda} \|\boldsymbol{\mu}\|_1.$$

**Proposition 1.** *Suppose* $\|\boldsymbol{\beta}_1^*\|, \|\boldsymbol{\beta}_2^*\|_0 \leq s$, $\sigma^* < C_1$, $\|\boldsymbol{\beta}_1^*\|, \|\boldsymbol{\beta}_2^*\| \leq C_2$ *for some constants* $C_1, C_2 > 0$. *If* $\frac{s \log p}{n} = o(1)$, *and* $\tilde{\lambda} \in (c_1 \sqrt{\frac{\log p}{n}}, c_2 \sqrt{\frac{\log p}{n}})$ *for sufficiently large* $c_1, c_2$, *we have with probability at least* $1 - O(n^{-1})$,

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\| \lesssim \sqrt{\frac{s \log p}{n}}.$$

Given $\hat{\boldsymbol{\mu}}$, we then symmetrize the second half of data by using the following transformation:

$$\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_{n+i}; \tilde{y}_i = y_{n+i} - \boldsymbol{x}_{n+i}^\top \hat{\boldsymbol{\mu}}, \text{for } i = 1, 2, ..., n.$$

By Proposition 1, such a transformation makes the two components of linear regression symmetric with negligible errors. In the following, for the simplicity of notation, we will write $(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)$ as $(\boldsymbol{x}_i, y_i)$, and use the following EM algorithm to estimate $(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*)/2$.

**Algorithm 3** EM for High-Dimensional MLR with balanced mixing weights

1: **Inputs:** Initializations $\hat{\boldsymbol{\beta}}^{(0)}$, maximum number of iterations $T$, and constants $\kappa \in (0, 1)$, $C_\lambda > 0$. Split the dataset into two subsets of size $n/2$.
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: **E-Step:** For $i \in [n]$, set

$$\gamma_{\boldsymbol{\theta},i}^{(t)} = \frac{1}{1 + \exp(-\frac{2y_i\langle \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}^{(t)}\rangle}{\sigma^2})}.$$

4: **M-Step:** Update $\hat{\boldsymbol{\beta}}_1^{(t+1)}$ and $\hat{\boldsymbol{\beta}}_2^{(t+1)}$ via

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^n \gamma_{\boldsymbol{\theta},i}^{(t)}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)^2 + \frac{1}{2n}\sum_{i=1}^n (1 - \gamma_{\boldsymbol{\theta},i}^{(t)})(y_i + \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)^2 + \lambda_n^{(t+1)}\|\boldsymbol{\beta}\|_1 \tag{7.2}$$

with

$$\lambda_n^{(t+1)} = \kappa_\lambda \lambda_n^{(t)} + C_\lambda \sqrt{\frac{\log p}{n}}. \tag{7.3}$$

5: **end for**
6: Output $\hat{\boldsymbol{\beta}}_1^{(T)} = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\beta}}^{(T)}$ and $\hat{\boldsymbol{\beta}}_2^{(T)} = \hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\beta}}^{(T)}$.

We then have the following convergence results for the full regimes of SNR.

**Theorem 6.** *Suppose the conditions in Proposition 1 hold. Given any* $\delta \in (0, 1)$.

1. *In the medium-to-high SNR regime where* $\mathrm{SNR} \geq C_1(s\log p \log(n/\delta)^2/n)^{1/6}$ *for some* $C_1 > 0$, *suppose the initialization* $\hat{\boldsymbol{\beta}}^{(0)}$ *satisfies* $\|\hat{\boldsymbol{\beta}}^{(0)} - (\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*)/2\| \leq 0.1 \cdot \|(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*)/2\|$ *and* $\cos(\angle(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*, \hat{\boldsymbol{\beta}}^{(0)})) \geq 0.96$. *Then for any* $\delta > 0$, *there exist universal constants* $C_2 > 0$ *such that with probability at least* $1 - \delta$,

$$|\hat{\boldsymbol{\beta}}_1^{(T)} - \boldsymbol{\beta}_1^*\| + \|\hat{\boldsymbol{\beta}}_2^{(T)} - \boldsymbol{\beta}_2^*\| \lesssim \max\{1, \|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|^{-2}\}\sqrt{\frac{s\log p \cdot \log^2(n/\delta)}{n}},$$

*for* $T \geq C_2 \max\{1, \|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|^{-2}\}\log n$.

2. *In the low-SNR regime when* $\mathrm{SNR} \leq C(s\log p \log(n/\delta)^2/n)^{1/6}$, *suppose the initialization* $\hat{\boldsymbol{\beta}}^{(0)}$ *satisfies* $\|\hat{\boldsymbol{\beta}}^{(0)}\| \leq 0.19$. *Then for any* $\delta > 0$, *there exists universal constant* $C_3 > 0$ *such*

*that with probability at least* $1-\delta$*,*

$$|\hat{\boldsymbol{\beta}}_1^{(T)} - \boldsymbol{\beta}_1^*\| + \|\hat{\boldsymbol{\beta}}_2^{(T)} - \boldsymbol{\beta}_2^*\| \lesssim (s \log p \log(n/\delta)^2/n)^{1/6},$$

*for* $T \geq C_3 \log(\log n) \cdot \sqrt{n/(s \log p \log^2(n/\delta))}$.

**Remark 5.** The division of medium-SNR and low-SNR regimes in the above theorem is different from the classic low-dimensional MLR model (Kwon et al. 2021), where the transition is at $O((p/n)^{1/4})$. It would be interesting to study if this transition order is optimal in the high-dimensional setting.

**Remark 6.** By using a similar proof as that of Theorem 2, the same asymptotic results can be established in the medium-SNR regime when $(\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|^{-4} \vee 1) \cdot \frac{(\sqrt{s} \vee \log^2 p) s \log p \cdot \log^2 n}{\sqrt{n}} = o(1)$.

## 8 DISCUSSION

The present paper introduced an iterative estimation procedure using an EM algorithm, a debiased approach for individual coefficient inference based on the EM solutions, and a multiple testing procedure based on the debiased estimators for the high-dimensional MLR. This paper focuses on the two-class mixed regression model. It is interesting to extend the proposed algorithms to the general $k$-class mixed regression models and analyze its performance, especially when $k$ is unknown.

In addition to the estimation, individual coefficient inference and multiple testing problems considered in the current paper, there are several other interesting and related problems that are worth investigating. One such related problem is testing a single regression model against a mixed regression model. This involves, for example, the construction and analysis of a goodness-of-fit test. Finally, a natural generalization of the MLR model is the mixed generalized linear models (MGLM), where the outcome variables are allowed to be categorical. Estimation and multiple testing for high-dimensional MGLM are important and challenging problems that we leave for future research.

## 9 PROOFS

We present in this section the proofs of Theorems 2 and 5, the results on the individual coordinate inference and multiple testing. Theorem 3 can be proved by using the same derivation as that in Theorem 2, and the proofs Theorem 1 and other technical lemmas are given in the Supplementary Materials.

### 9.1 Proof of Theorem 2

We first state the following lemmas.

**Lemma 2.** *Under the same conditions as in Theorem 1. For any given vector $\boldsymbol{m}^* \in \mathbb{R}^p$, there exists a constant $C > 0$ such that*

$$\|\mathbb{E}[\frac{1}{n_T}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta},i}^{(T)}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{m}^*] - \mathbb{E}[\frac{1}{n_T}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta}^*,i}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{m}^*]\|_2 \leq C\|\boldsymbol{m}^*\|_2\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(T)}\|_2;$$

$$\|\mathbb{E}[\frac{1}{n_T}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta},i}^{(T)}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{m}^*] - \frac{1}{n_T}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta},i}^{(T)}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{m}^*\|_\infty = \|\boldsymbol{m}^*\|_2 \cdot O_p(\sqrt{\frac{\log p}{n_T}}).$$

**Lemma 3.** *Under the same conditions as in Theorem 2. There exists a constant $c > 0$ such that*

$$\boldsymbol{m}_j^\top(T_n(\hat{\boldsymbol{\theta}}^{(T)}))_{1,1}\boldsymbol{m}_j \geq c, j \in [p].$$

Given the lemmas, we now proceed to proving Theorem 2. By symmetry, in the following, we only consider the case where $l = 1$.

Recall that $n_T = n/T$ with $T \asymp \log n$, $\tilde{\Sigma}_{XX} = \frac{1}{n_T}\sum_{i=1}^{n_T}\boldsymbol{x}_i\boldsymbol{x}_i^\top$. We first verify that for $\mu = C\sqrt{\frac{\log p \log n}{n}}$ with sufficiently large constant $C$, the optimization of $\boldsymbol{m}_j$ is feasible, that is, there exits $\boldsymbol{m}_j^* \in \mathbb{R}^p$, such that $||\tilde{\Sigma}_{XX}\boldsymbol{m}_j^* - \boldsymbol{e}_j^{(p)}||_\infty \leq \mu$ and $\|\boldsymbol{m}_j^*\|_1 \leq C\sqrt{\log n}$.

Take $\boldsymbol{m}_j^* = (\boldsymbol{\Sigma}^{-1})_j$ and use the fact that $\|\boldsymbol{\Sigma}^{-1}\| \leq L$, we have some $\|\boldsymbol{m}_j^*\|_1 \leq C\sqrt{\log n}$.

Further, since $\|\boldsymbol{m}_j^*\|_2 \leq C\sqrt{\log n}$ and $n/T \asymp n/\log n$, by the Bernstein inequality and union bound, we get

$$\|\mathbb{E}[\frac{1}{n_T}\sum_{i=1}^{n_T}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{m}_j^*] - \frac{1}{n_T}\sum_{i=1}^{n_T}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{m}_j^*\|_\infty = O_p(\sqrt{\frac{\log p}{n}}\log n).$$

Therefore, the optimization is feasible, and recall that the solution is denoted as $\tilde{\boldsymbol{m}}_j$. Then we proceed to showing the asymptotic normality.

For a given $j$, by (3.2), we have

$$\sqrt{n_T}(\widehat{\boldsymbol{\beta}}^u_{1j} - \boldsymbol{\beta}^*_{1j}) =\sqrt{n_T}(\boldsymbol{m}_j^\top \widehat{\Sigma}_{XY} - \boldsymbol{m}_j^\top \widehat{\Sigma}_{XX}\boldsymbol{\beta}_1^*) + \sqrt{n_T}(\boldsymbol{e}_j^\top - \boldsymbol{m}_j^\top \widehat{\Sigma}_{XX})(\hat{\boldsymbol{\beta}}_1^{(T)} - \boldsymbol{\beta}_1^*) \tag{9.1}$$
$$=\sqrt{n_T}\left[\frac{1}{n_T}\sum_{i=1}^{n}\gamma_{\boldsymbol{\theta},i}^{(T)}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1^*\rangle)\boldsymbol{m}_j^\top \boldsymbol{x}_i\right] + \sqrt{n_T}\|\boldsymbol{e}_j^\top - \boldsymbol{m}_j^\top \widehat{\Sigma}_{XX}\|_\infty \cdot \|\hat{\boldsymbol{\beta}}_1^{(T)} - \boldsymbol{\beta}_1^*\|_1.$$

We then show that for $\boldsymbol{m}_j = \tilde{\boldsymbol{m}}_j/\hat{\omega}^{(T)}$, $\|\boldsymbol{e}_j^\top - \boldsymbol{m}_j^\top \widehat{\Sigma}_{XX}\|_\infty$ is small.

Recall that $\widehat{\Sigma}_{XX} = \frac{1}{n_T}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta},i}^{(T)}\boldsymbol{x}_i\boldsymbol{x}_i^\top$. By Lemma 2, we get

$$\|\mathbb{E}[\frac{1}{n_T}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta},i}^{(T)}\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{m}_j] - \mathbb{E}[\frac{1}{n_T}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta}^*,i}\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{m}_j]\|_2 \lesssim \|\boldsymbol{m}_j\|_2 \cdot \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(T)}\|_2 = O_p(\sqrt{\frac{s\log p}{n}}\log n),$$

and

$$\|\mathbb{E}[\frac{1}{n_T}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta},i}^{(T)}\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{m}_j^*] - \frac{1}{n_T}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta},i}^{(T)}\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{m}_j^*\|_\infty = O_p(\sqrt{\frac{\log p}{n}}\log n).$$

Let $z_i$ be the class for the pair of data $(\boldsymbol{x}_i, y_i)$, we obtain

$$\mathbb{E}[\frac{1}{n_T}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta}^*,i}\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{m}_j^*] = \mathbb{E}[\frac{1}{n_T}\sum_{i=1}^{n_T}\mathbb{E}_{\hat{\boldsymbol{\theta}}^*}[1(z_i = 1)\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{m}_j^* \mid \boldsymbol{x}_i, y_i]] = \omega^*\boldsymbol{\Sigma}.$$

Therefore, we have

$$\begin{aligned}
&||\hat{\Sigma}_{XX}\boldsymbol{m}_j - \boldsymbol{e}_j^{(p)}||_\infty \\
=&||\mathbb{E}[\frac{1}{n_T}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta},i}^{(T)}\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{m}_j] - \frac{1}{n_T}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta},i}^{(T)}\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{m}_j||_\infty \\
&+ ||\mathbb{E}[\frac{1}{n_T}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta},i}^{(T)}\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{m}_j] - \mathbb{E}[\frac{1}{n_T}\sum_{i=1}^{n_T}\gamma_{\boldsymbol{\theta}^*,i}\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{m}_j]||_\infty + ||\mathbb{E}[\frac{1}{n_T}\sum_{i=1}^{n}\gamma_{\boldsymbol{\theta}^*,i}\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{m}_j] - \boldsymbol{e}_j^{(p)}||_\infty \\
=&O_p(\sqrt{\frac{\log p}{n}}\log n) + O_p(\sqrt{\frac{s\log p}{n}}\log n) + \|\omega^*\boldsymbol{\Sigma}\boldsymbol{m}_j - \boldsymbol{e}_j^{(p)}\|_\infty \\
\leq&O_p(\sqrt{\frac{s\log p}{n}}\log n) + \|\omega^*\boldsymbol{\Sigma}\boldsymbol{m}_j - \hat{\omega}^{(T)}\tilde{\Sigma}_{XX}\boldsymbol{m}_j\|_\infty + \|\hat{\omega}^{(T)}\tilde{\Sigma}_{XX}\boldsymbol{m}_j - \boldsymbol{e}_j^{(p)}\|_\infty \\
\leq&O_p(\sqrt{\frac{s\log p}{n}}\log n) + \|\omega^*\boldsymbol{\Sigma} - \hat{\omega}^{(T)}\tilde{\Sigma}_{XX}\|_\infty \cdot \|\boldsymbol{m}_j\|_1 + \|\tilde{\Sigma}_{XX}\tilde{\boldsymbol{m}}_j - \boldsymbol{e}_j^{(p)}\|_\infty \\
\leq&O_p(\sqrt{\frac{s\log p}{n}}\log n).
\end{aligned}$$

Then (9.1) becomes

$$\begin{aligned}
&\sqrt{n_T}(\widehat{\boldsymbol{\beta}}^u_{1j} - \boldsymbol{\beta}^*_{1j}) \\
=&\sqrt{n_T}\left[\frac{1}{n_T}\sum_{i=1}^{n}\gamma^{(T)}_{\boldsymbol{\theta},i}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}^*_1\rangle)\boldsymbol{m}_j^\top \boldsymbol{x}_i\right] + \sqrt{n_T}\|\boldsymbol{e}_j^\top - \boldsymbol{m}_j^\top \widehat{\Sigma}_{XX}\|_\infty \cdot \|\hat{\boldsymbol{\beta}}^{(T)}_1 - \boldsymbol{\beta}^*_1\|_1 \qquad (9.2)\\
=&\langle \boldsymbol{m}_j, \frac{1}{\sqrt{n_T}}\sum_{i=1}^{n_T}\gamma^{(T)}_{\boldsymbol{\theta},i}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}^*_1\rangle)\boldsymbol{x}_i\rangle + O_P(\frac{s^{3/2}\log p \log n}{\sqrt{n}}) \\
=&\langle \boldsymbol{m}_j, \frac{1}{\sqrt{n_T}}\sum_{i=1}^{n_T}\gamma^{(T)}_{\boldsymbol{\theta},i}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}^*_1\rangle)\boldsymbol{x}_i\rangle + o_P(1).
\end{aligned}$$

Then, by Lemma 3, we have

$$\frac{\sqrt{n_T}\left(\widehat{\boldsymbol{\beta}}^u_{1j} - \boldsymbol{\beta}^*_{1j}\right)}{\sqrt{\hat{v}_{1j}}} = \frac{\langle \boldsymbol{m}_j, \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\gamma^{(T)}_{\boldsymbol{\theta},i}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}^*_1\rangle)\boldsymbol{x}_i\rangle}{\sqrt{\hat{v}_{1j}}} + o_P(1).$$

Finally, using Lemma 1, we obtain the desired result.

$$\frac{\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^u_{1j} - \boldsymbol{\beta}^*_{1j}\right)}{\sqrt{\hat{v}_{1j}}} \xrightarrow{d} N(0,1).$$

## 9.2 Proof of Theorem 5

We first consider the case when $\hat{t}$, given by (4.3), does not exist. In this case, we have $\hat{t} = \sqrt{2\log p}$. Note that for $j \in \mathcal{H}_0$, we have

$$T^{(1)}_j = \frac{\sqrt{n}\hat{\boldsymbol{\beta}}^u_{1,j}}{\hat{v}^{(1)}_j} = \frac{\langle \boldsymbol{m}_j, \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\gamma_{\boldsymbol{\theta}^*,i}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}^*_1\rangle)\boldsymbol{x}_i\rangle}{\hat{v}^{(1)}_j} + \frac{\sqrt{n}Rem_1}{\hat{v}^{(1)}_j},$$

where $Rem_1 = o_P(1/\sqrt{n})$, and a similar expression holds for $T^{(2)}_j$. Then we have

$$\begin{aligned}
&\mathbb{P}\Bigg(\sum_{j\in\mathcal{H}_0} I\big(|T_j| \geq \sqrt{2\log p}\big) \geq 1\Bigg) \\
&\leq \mathbb{P}\Bigg(\sum_{j\in\mathcal{H}_0} I\big(|T^{(1)}_j| \geq \sqrt{2\log p}\big) \geq 1\Bigg) + \mathbb{P}\Bigg(\sum_{j\in\mathcal{H}_0} I\big(|T^{(2)}_j| \geq \sqrt{2\log p}\big) \geq 1\Bigg) \\
&\leq \mathbb{P}\Bigg(\sum_{j\in\mathcal{H}_0} I\Bigg(\frac{\langle \boldsymbol{m}_j, \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\gamma_{\boldsymbol{\theta}^*,i}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}^*_1\rangle)\boldsymbol{x}_i\rangle}{\hat{v}^{(1)}_j} + \frac{\sqrt{n}Rem_1}{\hat{v}^{(1)}_j} \geq \sqrt{2\log p}\Bigg) \geq 1\Bigg) \\
&\quad + \mathbb{P}\Bigg(\sum_{j\in\mathcal{H}_0} I\Bigg(\frac{\langle \boldsymbol{m}_j, \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\gamma_{\boldsymbol{\theta}^*,i}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}^*_1\rangle)\boldsymbol{x}_i\rangle}{\hat{v}^{(1)}_j} + \frac{\sqrt{n}Rem_1}{\hat{v}^{(1)}_j} \leq -\sqrt{2\log p}\Bigg) \geq 1\Bigg)
\end{aligned}$$

$$
+\mathbb{P}\bigg(\sum_{j\in\mathcal{H}_0} I\bigg(\frac{\langle \boldsymbol{m}_j, \frac{1}{\sqrt{n}}\sum_{i=1}^n (1-\gamma_{\boldsymbol{\theta}^*,i})(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_2^*\rangle)\boldsymbol{x}_i\rangle}{\hat{v}_j^{(2)}} + \frac{\sqrt{n}Rem_2}{\hat{v}_j^{(2)}} \geq \sqrt{2\log p}\bigg) \geq 1\bigg)
$$

$$
+\mathbb{P}\bigg(\sum_{j\in\mathcal{H}_0} I\bigg(\frac{\langle \boldsymbol{m}_j, \frac{1}{\sqrt{n}}\sum_{i=1}^n (1-\gamma_{\boldsymbol{\theta}^*,i})(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_2^*\rangle)\boldsymbol{x}_i\rangle}{\hat{v}_j^{(2)}} + \frac{\sqrt{n}Rem_2}{\hat{v}_j^{(2)}} \leq -\sqrt{2\log p}\bigg) \geq 1\bigg). \tag{9.3}
$$

Define $(v_j^{(1)})^2 = \mathrm{Var}(\langle \boldsymbol{m}_j, \frac{1}{\sqrt{n}}\sum_{i=1}^n \gamma_{\boldsymbol{\theta}^*,i}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1^*\rangle)\boldsymbol{x}_i\rangle)$. For any $\epsilon > 0$, we can bound the first term by

$$
\begin{aligned}
&\mathbb{P}\bigg(\sum_{j\in\mathcal{H}_0} I\bigg(\frac{\langle \boldsymbol{m}_j, \frac{1}{\sqrt{n}}\sum_{i=1}^n \gamma_{\boldsymbol{\theta}^*,i}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1^*\rangle)\boldsymbol{x}_i\rangle}{\hat{v}_j^{(1)}} + \frac{\sqrt{n}Rem_1}{\hat{v}_j^{(1)}} \geq \sqrt{2\log p}\bigg) \geq 1\bigg) \\
&= \mathbb{P}\bigg(\sum_{j\in\mathcal{H}_0} I\bigg(\frac{\langle \boldsymbol{m}_j, \frac{1}{\sqrt{n}}\sum_{i=1}^n \gamma_{\boldsymbol{\theta}^*,i}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1^*\rangle)\boldsymbol{x}_i\rangle}{v_j^{(1)}} \geq \frac{\hat{v}_j^{(1)}}{v_j^{(1)}}\sqrt{2\log p} - \frac{\sqrt{n}Rem_1}{v_j^{(1)}}\bigg) \geq 1\bigg) \\
&\leq \mathbb{P}\bigg(\sum_{j\in\mathcal{H}_0} I\bigg(\frac{\langle \boldsymbol{m}_j, \frac{1}{\sqrt{n}}\sum_{i=1}^n \gamma_{\boldsymbol{\theta}^*,i}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1^*\rangle)\boldsymbol{x}_i\rangle}{v_j^{(1)}} \geq (1-\epsilon)\sqrt{2\log p} - \epsilon\bigg) \geq 1\bigg) \\
&\quad + \mathbb{P}\bigg(\max_{j\in\mathcal{H}_0}\bigg|\frac{\sqrt{n}Rem_1}{v_j^{(1)}}\bigg| \geq \epsilon\bigg) + \mathbb{P}\bigg(\bigg|\frac{\hat{v}_j^{(1)}}{v_j^{(1)}} - 1\bigg| \geq \epsilon\bigg) \\
&\leq p\max_{j\in\mathcal{H}_0}\mathbb{P}\bigg(\frac{\langle \boldsymbol{m}_j, \frac{1}{\sqrt{n}}\sum_{i=1}^n \gamma_{\boldsymbol{\theta}^*,i}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1^*\rangle)\boldsymbol{x}_i\rangle}{v_j^{(1)}} \geq (1-\epsilon)\sqrt{2\log p} - \epsilon\bigg) + \mathbb{P}\bigg(\max_{j\in\mathcal{H}_0}\bigg|\frac{\sqrt{n}Rem_1}{v_j^{(1)}}\bigg| \geq \epsilon\bigg) \\
&\quad + \mathbb{P}\bigg(\bigg|\frac{\hat{v}_j^{(1)}}{v_j^{(1)}} - 1\bigg| \geq \epsilon\bigg).
\end{aligned}
$$

By the proof of Theorem 2, we know that

$$
\mathbb{P}\bigg(\max_{j\in\mathcal{H}_0}\bigg|\frac{\sqrt{n}Rem_1}{v_j^{(1)}}\bigg| \geq \epsilon\bigg) \to 0, \qquad \mathbb{P}\bigg(\bigg|\frac{\hat{v}_j^{(1)}}{v_j^{(1)}} - 1\bigg| \geq \epsilon\bigg) \to 0.
$$

In addition, for $j \in \mathcal{H}_0$, let

$$
T_{0j}^{(1)} = \frac{\langle \boldsymbol{m}_j, \frac{1}{\sqrt{n}}\sum_{i=1}^n \gamma_{\boldsymbol{\theta}^*,i}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1^*\rangle)\boldsymbol{x}_i\rangle}{v_j^{(1)}}.
$$

where $\mathbb{E}\langle \boldsymbol{m}_j, \gamma_{\boldsymbol{\theta}^*,i}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1^*\rangle)\boldsymbol{x}_i\rangle/v_j^{(1)} = 0$ and $\mathrm{Var}(\mathbb{E}\langle \boldsymbol{m}_j, \gamma_{\boldsymbol{\theta}^*,i}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1^*\rangle)\boldsymbol{x}_i\rangle/v_j^{(1)}) = 1$. Conditional on $X$ by Lemma 6.1 of Liu (2013), we have

$$
\sup_{0\leq t\leq 4\sqrt{\log p}}\bigg|\frac{P(|T_{0j}^{(1)}| \geq t)}{G(t)} - 1\bigg| \leq C(\log p)^{-1}. \tag{9.4}
$$

Hereafter, unless explicitly noted, all of our discussion will be conditional on $\{\boldsymbol{x}_i\}_{i=1}^n$. Now let $t = (1-\epsilon)\sqrt{2\log p} - 2\epsilon$, we have

$$\mathbb{P}\Big(T_{0j}^{(1)} \geq (1-\epsilon)\sqrt{2\log p} - 2\epsilon\Big) \leq G((1-\epsilon)\sqrt{2\log p} - 2\epsilon) + C\frac{G((1-\epsilon)\sqrt{2\log p} - 2\epsilon)}{\log p}.$$

Hence

$$p \max_{j\in\mathcal{H}_0} \mathbb{P}\Big(T_{0j} \geq (1-\epsilon)\sqrt{2\log p} - \epsilon\Big) \leq CpG((1-\epsilon)\sqrt{2\log p} - 2\epsilon) + O(p^{-c}),$$

which goes to zero as $(n,p)\to\infty$. By symmetry, we know that the rest three terms in (9.3) also goes to 0. Therefore we have proved the theorem when $\hat{t} = \sqrt{2\log p}$.

Now consider the case when $0 \leq \hat{t} \leq b_p$ holds. We have

$$\text{FDP}(\hat{t}) = \frac{\sum_{j\in\mathcal{H}_0} I\{|T_j| \geq \hat{t}\}}{\max\big\{\sum_{j=1}^p I\{|T_j| \geq \hat{t}\}, 1\big\}} \leq \frac{\sum_{j\in\mathcal{H}_0} I\{|T_j^{(1)}| \geq \hat{t}\} + \sum_{j\in\mathcal{H}_0} I\{|T_j^{(2)}| \geq \hat{t}\}}{\max\big\{\sum_{j=1}^p I\{|T_j| \geq \hat{t}\}, 1\big\}}.$$

Note that for $\ell = 1, 2$,

$$\frac{\sum_{j\in\mathcal{H}_0} I\{|T_j^{(\ell)}| \geq \hat{t}\}}{\max\big\{\sum_{j=1}^p I\{|T_j| \geq \hat{t}\}, 1\big\}} \leq \frac{p_0 G(\hat{t})}{\max\big\{\sum_{j=1}^p I\{|T_j| \geq \hat{t}\}, 1\big\}}(1 + A_p^{(\ell)})$$

where

$$A_p^{(\ell)} = \sup_{0\leq t\leq b_p} \left| \frac{\sum_{j\in\mathcal{H}_0} I\{|T_j^{(\ell)}| \geq t\}}{p_0 G(t)} - 1 \right|.$$

Note that by definition

$$\frac{p_0 G(\hat{t})}{\max\big\{\sum_{j=1}^p I\{|T_j| \geq \hat{t}\}, 1\big\}} \leq \frac{p_0\alpha}{p}.$$

The proof is complete if $A_p^{(\ell)} \to 0$ in probability. The rest of the proof is devoted to it. We first show that

$$|T_j^{(\ell)} - T_{0j}^{(\ell)}| = o_P(1/\sqrt{\log p}). \tag{9.5}$$

To see this, we notice that, under the sparsity condition $s = o\big(\frac{n^{1/2}}{\log^{3/2} p \log n}\big)$, with probability at least $1 - O(p^{-c})$,

$$|T_j^{(\ell)} - T_{0j}^{(\ell)}| \leq \left|\frac{\sqrt{n}Rem_\ell}{v_j^{(\ell)}}\right| \cdot \left|\frac{v_j^{(\ell)}}{\hat{v}_j^{(\ell)}}\right| + \left|T_{0j}^{(\ell)}(v_j^{(\ell)}/\hat{v}_j^{(\ell)} - 1)\right| = o\Big(\frac{1}{\sqrt{\log p}}\Big).$$

By the fact that $G(t+o(1/\sqrt{\log p}))/G(t) = 1+o(1)$ uniformly in $0 \le t \le \sqrt{2\log p}$, it suffices to show that

$$\sup_{0\le t\le b_p}\left|\frac{\sum_{j\in\mathcal{H}_0} I\{|T_{0j}^{(\ell)}|\ge t\}}{p_0G(t)} - 1\right| \to 0 \quad \text{in probability.} \tag{9.6}$$

Let $z_0 < z_1 < \ldots < z_{d_p} \le 1$ and $t_i = G^{-1}(z_i)$, where $z_0 = G(b_p)$, $z_i = c_p/p + c_p^{2/3}e^{i^\delta}/p$ with $c_p = pG(b_p)$, and $d_p = [\log((p-c_p)/c_p^{2/3})]^{1/\delta}$ and $0<\delta<1$, which will be specified later. We have $G(t_i)/G(t_{i+1}) = 1+o(1)$ uniformly in $i$, and $t_0/\sqrt{2\log(p/c_p)} = 1+o(1)$. Note that uniformly for $1\le j\le m$, $G(t_i)/G(t_{i-1}) \to 1$ as $p\to\infty$. The proof of (9.6) reduces to show that

$$\max_{0\le i\le d_p}\left|\frac{\sum_{j\in\mathcal{H}_0} I\{|T_{0j}^{(\ell)}|\ge t_i\}}{p_0G(t_i)} - 1\right| \to 0 \tag{9.7}$$

in probability. Hereafter, we omit the dependence on the index $\ell$ for simplicity. In fact, for each $\epsilon>0$, we have

$$\mathbb{P}\left(\max_{0\le i\le d_p}\left|\frac{\sum_{j\in\mathcal{H}_0}[I\{|T_{0j}|\ge t_i\} - G(t_i)]}{p_0G(t_i)}\right| \ge \epsilon\right) \le \sum_{j=0}^{d_p}\mathbb{P}\left(\left|\frac{\sum_{j\in\mathcal{H}_0}[I\{|T_{0j}|\ge t_i\} - G(t_i)]}{p_0G(t_i)}\right| \ge \epsilon/2\right).$$

Set $I(t) = \frac{\sum_{j\in\mathcal{H}_0}[I\{|T_{0j}|\ge t\} - P(|T_{0j}|\ge t)]}{p_0G(t)}$. By Markov's inequality $P(|I(t_i)|\ge\epsilon/2) \le \frac{\mathbb{E}[I(t_i)]^2}{\epsilon^2/4}$, and it suffices to show $\sum_{j=0}^{d_p}\mathbb{E}[I(t_i)]^2 = o(1)$. To see this, by (9.4),

$$\begin{aligned}
\mathbb{E}I^2(t) &= \frac{\sum_{j\in\mathcal{H}_0}[P(|T_{0j}|\ge t) - P^2(|T_{0j}|\ge t)]}{p_0^2G^2(t)} \\
&\quad + \frac{\sum_{j,k\in\mathcal{H}_0,k\ne j}[P(|T_{0k}|\ge t, |T_{0j}|\ge t) - P(|T_{0k}|\ge t)P(|T_{0j}|\ge t)]}{p_0^2G^2(t)} \\
&\le \frac{C}{p_0G(t)} + \frac{1}{p_0^2}\sum_{(j,k)\in\mathcal{A}(\epsilon)\cap\mathcal{H}_0}\frac{P(|T_{0k}|\ge t, |T_{0j}|\ge t)}{G^2(t)} \\
&\quad + \frac{1}{p_0^2}\sum_{(j,k)\in\mathcal{A}(\epsilon)^c\cap\mathcal{H}_0}\left[\frac{P(|T_{0k}|\ge t, |T_{0j}|\ge t)}{G^2(t)} - 1\right] \\
&= \frac{C}{p_0G(t)} + I_{11}(t) + I_{12}(t).
\end{aligned}$$

For $(j,k)\in\mathcal{A}(\epsilon)^c\cap\mathcal{H}_0$, applying Lemma 6.1 in Liu (2013), we have $I_{12}(t) \le C(\log p)^{-1-\xi}$ for some $\xi>0$ uniformly in $0<t<\sqrt{2\log p}$. By Lemma 6.2 in Liu (2013), for $(j,k)\in\mathcal{A}(\epsilon)\cap\mathcal{H}_0$, we have

$$P(|T_{0k}|\ge t, |T_{0j}|\ge t) \le C(t+1)^{-2}\exp\left(-\frac{t^2}{1+|\rho_{jk}|}\right).$$

So that

$$I_{11}(t) \leq C \frac{1}{p_0^2} \sum_{(j,k)\in\mathcal{A}(\epsilon)\cap\mathcal{H}_0} (t+1)^{-2} \exp\Big(-\frac{t^2}{1+|\rho_{jk}|}\Big) G^{-2}(t) \leq C \frac{1}{p_0^2} \sum_{(j,k)\in\mathcal{A}(\epsilon)\cap\mathcal{H}_0} [G(t)]^{-\frac{2|\rho_{jk}|}{1+|\rho_{jk}|}}.$$

Note that for $0 \leq t \leq b_p$, we have $G(t) \geq G(b_p) = c_p/p$, so that by assumption (A3) it follows that for some $\epsilon, q > 0$,

$$I_{11}(t) \leq C \sum_{(j,k)\in\mathcal{A}(\epsilon)\cap\mathcal{H}_0} p^{\frac{2|\rho_{jk}|}{1+|\rho_{jk}|}+q-2} = O(1/(\log p)^2).$$

By the above inequalities, we can prove (9.7) by choosing $0 < \delta < 1$ so that

$$\begin{aligned}
\sum_{i=0}^{d_p} \mathbb{E}[I(t_i)]^2 &\leq C \sum_{i=0}^{d_p} (pG(t_i))^{-1} + Cd_p[(\log p)^{-1-\delta} + (\log p)^{-2}] \\
&\leq C \sum_{i=0}^{d_p} \frac{1}{c_p + c_p^{2/3} e^{i^\delta}} + o(1) \\
&= o(1).
\end{aligned}$$

Lately, as all the above arguments are conditional on $\{\boldsymbol{x}_i\}_{i=1}^n$, the statements of Theorem 5 follow by averaging over the probability measure of $\{\boldsymbol{x}_i\}_{i=1}^n$ . □

## FUNDING

This research was supported by NIH grants R01GM123056 and R01GM129781 and NSF grant DMS-1712735.

## SUPPLEMENTARY MATERIALS

In the Supplemental Materials, we prove all the main theorems and the technical lemmas.

## References

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

Bashir, S. and E. Carter (2012). Robust mixture of linear regression models. *Communications in Statistics-Theory and Methods 41*(18), 3371–3388.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics 29*, 1165–1188.

Cai, T. T., W. Liu, and H. H. Zhou (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics 44*(2), 455–488.

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

Cai, T. T. and H. H. Zhou (2012). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics 40*(5), 2389–2420.

Chaganty, A. T. and P. Liang (2013). Spectral experts for estimating mixtures of linear regressions. In *International Conference on Machine Learning*, pp. 1040–1048.

Chen, Y., X. Yi, and C. Caramanis (2014). A convex formulation for mixed regression with two components: Minimax optimal rates. In *Conference on Learning Theory*, pp. 560–604. PMLR.

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

Devijver, E., Y. Goude, and J.-M. Poggi (2020). Clustering electricity consumers using high-dimensional regression mixture models. *Applied Stochastic Models in Business and Industry 36*(1), 159–177.

Fan, J., H. Liu, Z. Wang, and Z. Yang (2018). Curse of heterogeneity: Computational barriers in sparse mixture models and phase retrieval. *arXiv preprint arXiv:1808.06996*.

Faria, S. and G. Soromenho (2010). Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation 80*(2), 201–225.

Grün, B. and F. Leisch (2007). Applications of finite mixtures of regression models. *http://cran.r-project.org/web/packages/flexmix/vignettes/regression-examples.pdf 2007*, 1–26.

Ho, N., C.-Y. Yang, and M. I. Jordan (2019). Convergence rates for gaussian mixtures of experts. *arXiv preprint arXiv:1907.04377*.

Javanmard, A. and H. Javadi (2019). False discovery rate control via debiased lasso. *Electronic Journal of Statistics 13*(1), 1212–1253.

Javanmard, A. and A. Montanari (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research 15*(1), 2869–2909.

Javanmard, A. and A. Montanari (2014b). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory 60*(10), 6522–6554.

Khalili, A. and J. Chen (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association 102*(479), 1025–1038.

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

Kwon, J. and C. Caramanis (2020). Em converges for a mixture of many linear regressions. In *International Conference on Artificial Intelligence and Statistics*, pp. 1727–1736. PMLR.

Kwon, J., N. Ho, and C. Caramanis (2021). On the minimax optimality of the em algorithm for learning two-component mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 1405–1413. PMLR.

Li, Q., R. Shi, and F. Liang (2019). Drug sensitivity prediction with high-dimensional mixture regression. *PloS one 14*(2), 1–18.

Li, Y. and Y. Liang (2018). Learning mixtures of linear regressions with nearly optimal complexity. In *Conference On Learning Theory*, pp. 1125–1144.

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics 41*(6), 2948–2978.

McLachlan, G. J. and D. Peel (2004). *Finite mixture models*. John Wiley & Sons.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*, 1436–1462.

Netrapalli, P., P. Jain, and S. Sanghavi (2013). Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pp. 2796–2804.

Ning, Y. and H. Liu (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics 45*(1), 158–195.

Schapiro, D., H. W. Jackson, S. Raghuraman, J. R. Fischer, V. R. Zanotelli, D. Schulz, C. Giesen, R. Catena, Z. Varga, and B. Bodenmiller (2017). histocat: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature Methods 14*(9), 873.

Shen, Y. and S. Sanghavi (2019). Iterative least trimmed squares for mixed linear regression. *arXiv preprint arXiv:1902.03653*.

Städler, N., P. Bühlmann, and S. van de Geer (2010). $\ell_1$-penalization for mixture regression models. *Test 19*(2), 209–256.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202.

Wang, Z., Q. Gu, Y. Ning, and H. Liu (2015). High dimensional EM algorithm: Statistical optimization and asymptotic normality. In *Advances in neural information processing systems*, pp. 2521–2529.

Yao, W. and W. Song (2015). Mixtures of linear regression with measurement errors. *Communications in Statistics-Theory and Methods 44*(8), 1602–1614.

Yi, X. and C. Caramanis (2015). Regularized em algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, pp. 1567–1575.

Yi, X., C. Caramanis, and S. Sanghavi (2014). Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pp. 613–621.

Yuan, M. and Y. Lin (2007). Model selection and estimation in the gaussian graphical model. *Biometrika 94*(1), 19–35.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 217–242.

Zhong, K., P. Jain, and I. S. Dhillon (2016). Mixed linear regression with multiple components. In *Advances in Neural Information Processing Systems*, pp. 2190–2198.

Zhu, H.-T. and H. Zhang (2004). Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(1), 3–16.

Zhu, R., L. Wang, C. Zhai, and Q. Gu (2017). High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 4180–4188. JMLR. org.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.

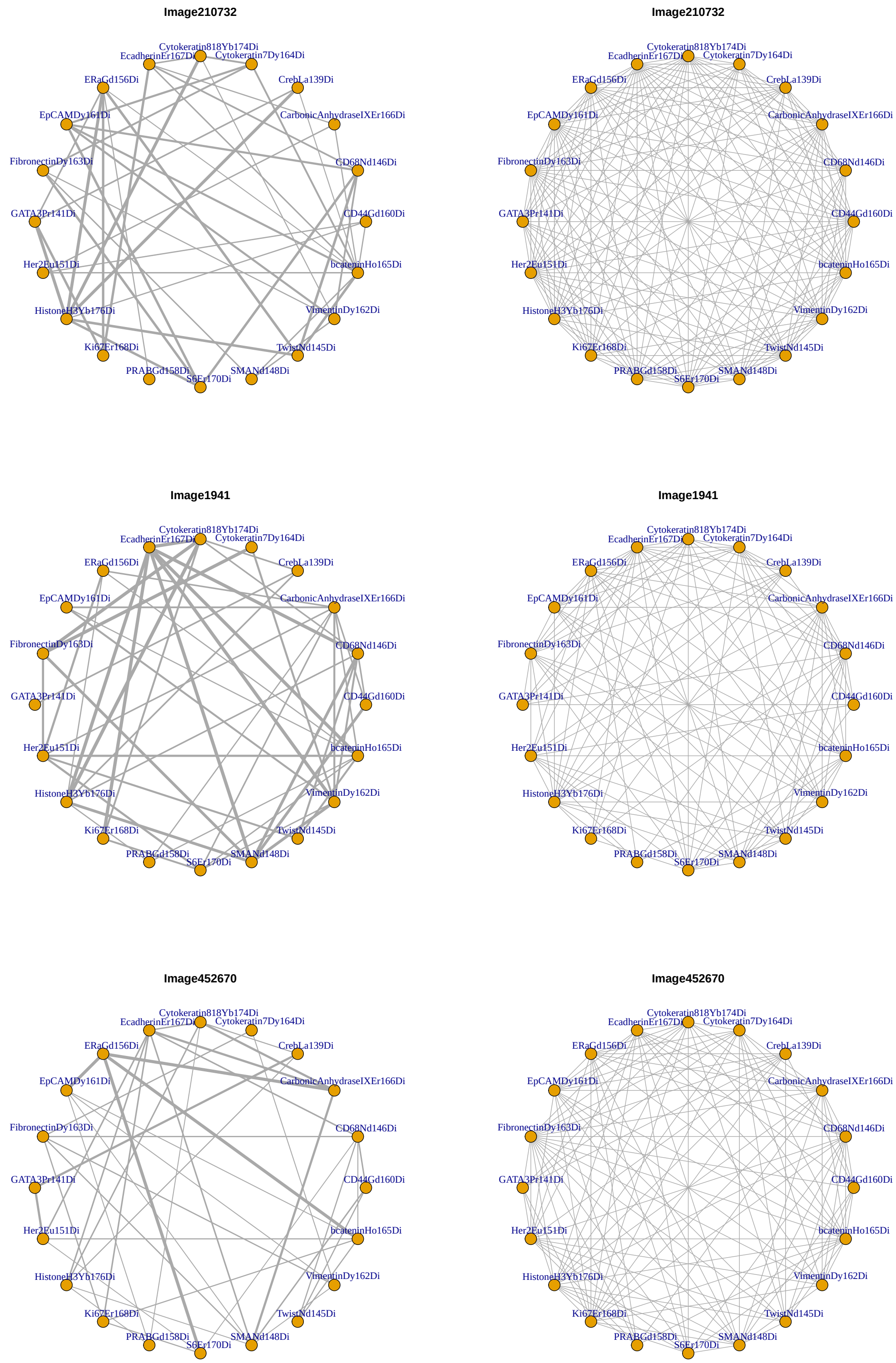


Figure 3: Networks of conditional dependence generated from node-wise regressions based on our proposed methods (left) and the standard Lasso based methods (right), both with FDR$< 10\%$ for three different images. In the networks based on MLR, the widths of the edges were set to be proportional to the $\ell_2$ distances between the two mixed regression coefficients.