Sparse Mixed Linear Regression with Guarantees: Taming an Intractable Problem with Invex Relaxation

Adarsh Barik [*1] **Jean Honorio** [*1]

Abstract

In this paper, we study the problem of sparse mixed linear regression on an unlabeled dataset that is generated from linear measurements from two different regression parameter vectors. Since the data is unlabeled, our task is not only to figure out a good approximation of the regression parameter vectors but also to label the dataset correctly. In its original form, this problem is NP-hard. The most popular algorithms to solve this problem (such as Expectation-Maximization) have a tendency to stuck at local minima. We provide a novel invex relaxation for this intractable problem which leads to a solution with provable theoretical guarantees. This relaxation enables exact recovery of data labels. Furthermore, we recover a close approximation of the regression parameter vectors which match the true parameter vectors in support and sign. Our formulation uses a carefully constructed primal dual witnesses framework for the invex problem. Furthermore, we show that the sample complexity of our method is only logarithmic in terms of the dimension of the regression parameter vectors.

1. Introduction

In this paper, we study sparse mixed linear regression where the measurements come from one of the two regression models depending upon the unknown label $z_i^* \in \{0, 1\}$. The observation model can be described as follows:

$$y_i = z_i^* \langle X_i, \beta_1^* \rangle + (1 - z_i^*) \langle X_i, \beta_2^* \rangle + e_i, \forall i \in \{1, \cdots, n\} , \tag{1}$$

where $X_i \in \mathbb{R}^d, y_i \in \mathbb{R}$ and $e_i \in \mathbb{R}$ is independent additive noise. The regression parameter vectors $\beta_1^* \in \mathbb{R}^d$, $\beta_2^* \in \mathbb{R}^d$ are sparse vectors with possibly non-overlapping supports.

Mixed linear regression models have been extensively used in a wide range of applications (Grün et al., 2007) which include but are not limited to behavioral health-care (Deb & Holmes, 2000), market segmentation (Wedel & Kamakura, 2000), music perception studies (Viele & Tong, 2002) and vehicle merging (Li et al., 2019). The main task of the problem is to estimate the regression parameter vectors and the unknown labels accurately from linear measurements. However, the problem is NP-hard without any assumptions (Yi et al., 2014). Being such a difficult problem, it also lends itself to be used as a benchmark for many non-convex optimization algorithms (Chaganty & Liang, 2013; Klusowski et al., 2019).

Related Work. There have been many approaches to solve the mixed linear regression problem after it was introduced by (Wedel & DeSarbo, 1995). The most popular and natural approach has been to use Expectation-minimization (EM) based alternate minimization algorithms (see Ghosh & Kannan (2020) and references therein). More broadly, the problem can be modeled under the hierarchical mixtures of experts model (Jordan & Jacobs, 1994) and solved using EM based algorithms. All these methods run the risk of getting stuck at local minima (Wu, 1983) without good initialization. (Yi et al., 2014) provides a good initialization for the noiseless case under strict technical conditions, however their method does not provide any guarantees for the noisy case. Based on the recent work of (Anandkumar et al., 2014; Hsu & Kakade, 2013), (Chaganty & Liang, 2013) have proposed an approach which uses a third order moment method based on tensor decomposition. Their approach suffers from high sample complexity (up to $\mathcal{O}(d^6)$) due to tensor decomposition. (Städler et al., 2010) proposed an ℓ_1-regularized approach for the sparse case and showed the existence of a local minimizer with correct support but there are no guarantees that EM achieves this local minima. (Chen et al., 2014) provided a convex relaxation involving nuclear norms for the problem. They do not focus on providing guarantees for exact label recovery and their results only hold for bounded noise and require balanced samples (almost equal number of samples for both labels). Besides, the optimization problems involving nuclear norms are computationally heavy

[*]Equal contribution [1]Department of Computer Science, Purdue University, West Lafayette, Indiana, USA. Correspondence to: Adarsh Barik <abarik@purdue.edu>, Jean Honorio <jhonorio@purdue.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

and slow. The mixed linear regression problem can also be modeled as a subspace clustering problem. But typically these problems require $\mathcal{O}(d^2)$ measurements to have a unique solution (Vidal et al., 2005; Elhamifar & Vidal, 2013).

Contribution. Broadly, we can categorize our contribution in the following points:

- **A Combinatorial Problem**: We view the problem as a combinatorial version of a mixture of sparse linear regressions. The exact label recovery is as important for us as the recovery of regression vectors. This added exact label recovery guarantee comes at no extra cost in terms of the performance.
- **Invex Relaxation**: We solve a non-convex problem which is known to be intractable. We propose a novel relaxation of the combinatorial problem and formally show that this relaxation is invex.
- **Theoretical Guarantees**: Our method solves two sparse linear regressions and a label recovery problem simultaneously with theoretical guarantees. To that end, we recover the true labels and sparse regression parameter vectors which are correct up to the sign of entries with respect to the true parameter vectors. As a side product, we propose a novel primal-dual witness construction for our invex problem and provide theoretical guarantees for recovery. The sample complexity of our method only varies logarithmically with respect to dimension of the regression parameter vector.
- **A Novel Framework**: It should be noted that we are providing a novel framework (not an algorithm) to solve the problem. This opens the door for many algorithms to be used for this problem.

2. Problem Setup

In this section, we collect the notations used throughout the paper and define our problem formally. We consider a problem where measurements come from a mixture of two linear regression problem. Let $y_i \in \mathbb{R}$ be the response variable and $X_i \in \mathbb{R}^d$ be the observed attributes. Let $z_i^* \in \{0,1\}$ denote the unknown label associated with measurement i. The response y_i is generated using the observation model (1) where $e_i \in \mathbb{R}$ is an independent noise term. We collect a total of n linear measurements with n_1 measurements belonging to label 1 and n_2 measurements belonging to label 0. Clearly, $n = n_1 + n_2$. We take $\|\beta_1^*\|_1 \leqslant b_1$ and $\|\beta_2^*\|_1 \leqslant b_2$.

Let $[d]$ denote the set $\{1, 2, \cdots, d\}$. We assume $X_i \in \mathbb{R}^d$ to be a zero mean sub-Gaussian random vector (Hsu et al., 2012) with covariance $\Sigma \in \mathbb{S}_+^d$, i.e., there exists a $\rho > 0$, such that for all $\tau \in \mathbb{R}^d$ the following holds: $\mathbb{E}(\exp(\tau^\intercal X_i)) \leqslant \exp(\frac{\|\tau\|_2^2 \rho^2}{2})$. By simply taking $\tau_j = r$ and $\tau_k = 0, \forall k \neq j$, it follows that each entry of X_i is sub-Gaussian with parameter ρ. In particular, we will assume that $\forall j \in [d]\,, \frac{X_{ij}}{\sqrt{\Sigma_{jj}}}$ is a sub-Gaussian random variable with parameter $\sigma > 0$. It follows trivially that $\max_{j\in[d]} \sqrt{\Sigma_{jj}}\sigma \leqslant \rho$. We will further assume that e_i is zero mean independent sub-Gaussian noise with variance σ_e. Our setting works with a variety of random variables as the class of sub-Gaussian random variable includes for instance Gaussian variables, any bounded random variable (e.g., Bernoulli, multinomial, uniform), any random variable with strictly log-concave density, and any finite mixture of sub-Gaussian variables.

The parameter vectors $\beta_1^* \in \mathbb{R}^d$ and $\beta_2^* \in \mathbb{R}^d$ are s_1-sparse and s_2-sparse respectively, i.e., at most s_1 entries of β_1^* are non-zero whereas at most s_2 entries of β_2^* are non-zero. We receive n i.i.d. samples of $X_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ and collect them in $X \in \mathbb{R}^{n\times d}$ and $y \in \mathbb{R}^n$ respectively. Similarly, $z^* \in \{0,1\}^n$ collects all the labels. Our goal is to recover β_1^*, β_2^* and z^* using the samples (X, y).

We denote a matrix $A \in \mathbb{R}^{p\times q}$ restricted to the columns and rows in $P \subseteq [p]$ and $Q \subseteq [q]$ respectively as A_{PQ}. Similarly, a vector $v \in \mathbb{R}^p$ restricted to entries in P is denoted as v_P. We use $\mathrm{eig}_i(A)$ to denote the i-th eigenvalue (1st being the smallest) of matrix A. Similarly, $\mathrm{eig}_{\max}(A)$ denotes the maximum eigenvalue of matrix A. We use $\mathrm{diag}(A)$ to denote a vector containing the diagonal element of matrix A. By overriding the same notation, we use $\mathrm{diag}(v)$ to denote a diagonal matrix with its diagonal being the entries in vector v. We denote the inner product between two matrices (or vectors) A and B by $\langle A, B\rangle$, i.e., $\langle A, B\rangle = \mathrm{trace}(A^\intercal B)$, where trace denotes the trace of a matrix. The notation $A \succeq B$ denotes that $A - B$ is a positive semidefinite matrix. Similarly, $A \succ B$ denotes that $A - B$ is a positive definite matrix. For vectors, $\|v\|_p$ denotes the ℓ_p-vector norm of vector $v \in \mathbb{R}^d$, i.e., $\|v\|_p = (\sum_{i=1}^d |v_i|^p)^{\frac{1}{p}}$. If $p = \infty$, then we define $\|v\|_\infty = \max_{i=1}^d |v_i|$. As is the tradition, we used $\|v\|_0$ to denote number of non-zero entries on vector v. It should be remembered that ℓ_0 is not a proper vector norm. For matrices, $\|A\|_p$ denotes the induced ℓ_p-matrix norm for matrix $A \in \mathbb{R}^{p\times q}$. In particular, $\|A\|_2$ denotes the spectral norm of A and $\|A\|_\infty \triangleq \max_{i\in[p]} \sum_{j=1}^q |A_{ij}|$. For a matrix $A \in \mathbb{R}^{p\times q}$, $A(:) \in \mathbb{R}^{pq}$ denotes a vector which collects all entries of the matrix A. We define an operator $\mathrm{sign}(A)$ for a matrix(or vector) A, which returns a matrix (or a vector) with entries being the sign of the entries of A. A function $f(x)$ is of order $\Omega(g(x))$ and denoted by $f(x) = \Omega(g(x))$, if there exists a constant $C > 0$ such that for big enough x_0, $f(x) \geqslant Cg(x), \forall x \geqslant x_0$. Similarly, a function $f(x)$ is of order $\mathcal{O}(g(x))$ and denoted by $f(x) = \mathcal{O}(g(x))$, if there exists a constant $C > 0$ such that for big enough x_0, $f(x) \leqslant Cg(x), \forall x \geqslant x_0$. For brevity in our notations, we

treat any quantity independent of d, s and n as constant. Detailed proofs for lemmas and theorems are available in the supplementary material.

3. A Novel Invex Relaxation

In this section, we introduce a combinatorial formulation for mixed linear regression (MLR) and propose a novel invex relaxation for this problem. Since the measurements come from a true observation model (1), we can write the following optimization problem to estimate β_1^*, β_2^* and z^*.

Definition 3.1 (Standard MLR)**.**

$$\begin{aligned} \min_{\beta_1\in\mathbb{R}^d,\beta_2\in\mathbb{R}^d,z\in\{0,1\}^n} \quad & l(z,\beta_1,\beta_2) \\ \text{such that} \quad & \|\beta_1\|_0 = s_1,\ \|\beta_2\|_0 = s_2 \end{aligned} \tag{2}$$

where $l(z,\beta_1,\beta_2) = \frac{1}{n}\sum_{i=1}^n z_i(y_i - X_i^\intercal\beta_1)^2 + (1-z_i)(y_i - X_i^\intercal\beta_2)^2$.

Even without constraints, optimization problem (2) is a non-convex NP-hard problem (Yi et al., 2014) in its current form. In fact, a continuous relaxation of $z \in [0,1]^n$ does not help and it still remains a non-convex problem (See Appendix A). Furthermore, the sparsity constraints make it even difficult to solve. To deal with this intractability, we come up with a novel invex relaxation of the problem.

For ease of notation, we define the following quantities:

$$S_i = \begin{bmatrix} X_i \\ -y_i \end{bmatrix}\begin{bmatrix} X_i^\intercal & -y_i \end{bmatrix} = \begin{bmatrix} X_iX_i^\intercal & -X_iy_i \\ -y_iX_i^\intercal & y_i^2 \end{bmatrix}, \tag{3}$$

We provide the following invex relaxation to the optimization problem (2).

Definition 3.2 (Invex MLR)**.**

$$\begin{aligned} \min_{t,W,U} \quad & f(t,W,U) + \lambda_1 g(t,W,U) + \lambda_2 h(t,W,U) \\ \text{such that} \quad & W \succeq \mathbf{0},\ U \succeq \mathbf{0} \\ & W_{d+1,d+1} = 1,\ U_{d+1,d+1} = 1 \\ & \|t\|_\infty \leqslant 1 \end{aligned} \tag{4}$$

where $f(t,W,U) = \sum_{i=1}^n \frac{1}{2}\langle S_i, W + U\rangle + \sum_{i=1}^n \frac{1}{2}t_i\langle S_i, W - U\rangle, g(t,W,U) = \|W(:)\|_1$ and $h(t,W,U) = \|U(:)\|_1$ and λ_1 and λ_2 are positive regularizers.

To get an intuition behind this formulation, one can think of W and U as two rank-1 matrices which are defined as follows:

$$W = \begin{bmatrix} \beta_1 \\ 1 \end{bmatrix}\begin{bmatrix} \beta_1^\intercal & 1 \end{bmatrix},\ U = \begin{bmatrix} \beta_2 \\ 1 \end{bmatrix}\begin{bmatrix} \beta_2^\intercal & 1 \end{bmatrix} \tag{5}$$

The variable t is simply a replacement of variable z, i.e., $z_i = \frac{t_i+1}{2}$. An analogous transformation exists between z_i^* and t_i^*. Then after substituting t, W and U in $f(t,W,U)$, we get back $l(z,\beta_1,\beta_2)$. The ℓ_1-regularization of $W(:)$ and $U(:)$ helps us ensure sparsity. Note that for fixed t, optimization problem (4) is continuous and convex with respect to W and U. Specifically, it merges two independent regularized semidefinite programs. Unfortunately, problem (4) is not jointly convex on t, W and U, and thus, it might still remain difficult to solve. Next, we will provide arguments that despite being non-convex, optimization problem (4) belongs to a particular class of non-convex functions namely "invex" functions. The "invexity" of functions can be defined as a generalization of convexity (Hanson, 1981). Invexity has been recently used by (Barik & Honorio, 2021) to solve fair sparse regression problem with clustering. While, we borrow some definitions from their work to suit our context, we should emphasize that our problem is fundamentally different than their problem. They use two groups in sparse regression which have different means and they try to achieve fairness. While here, we have two groups with the same mean and there is no unfairness in the problem. We also model our parameter vectors with positive semidefinite matrices which is fundamentally different from their approach.

Definition 3.3 (Invex function (Barik & Honorio, 2021))**.** Let $\phi(t)$ be a function defined on a set C. Let η be a vector valued function defined in $C\times C$ such that $\eta(t_1,t_2)^\intercal\nabla\phi(t_2)$, is well defined $\forall t_1,t_2 \in C$. Then, $\phi(t)$ is a η-invex function if $\phi(t_1) - \phi(t_2) \geqslant \eta(t_1,t_2)^\intercal\nabla\phi(t_2),\ \forall t_1,t_2 \in C$.

Note that convex functions are η-invex for $\eta(t_1,t_2) = t_1 - t_2$. (Hanson, 1981) showed that if the objective function and constraints are both η-invex with respect to same η defined in $C \times C$, then Karush-Kuhn-Tucker (KKT) conditions are sufficient for optimality, while it is well-known that KKT conditions are necessary. (Ben-Israel & Mond, 1986) showed a function is invex if and only if each of its stationarity point is a global minimum.

In the next lemma, we show that the relaxed optimization problem (4) is indeed η-invex for a particular η defined in $C \times C$ and a well defined set C. Let $C = \{(t,W,U) \mid t \in [-1,1]^n, W \succeq \mathbf{0}, U \succeq \mathbf{0}, W_{d+1,d+1} = 1, U_{d+1,d+1} = 1\}$.

Lemma 3.4. *For $(t,W,U) \in C$, the functions $f(t,W,U) = \sum_{i=1}^n \frac{1}{2}\langle S_i, W + U\rangle + \sum_{i=1}^n \frac{1}{2}t_i\langle S_i, W - U\rangle, g(t,W,U) = \|W(:)\|_1$ and $h(t,W,U) = \|U(:)\|_1$ are η-invex for $\eta(t,\tilde{t},W,\widetilde{W},U,\widetilde{U}) \triangleq \begin{bmatrix} \eta_t \\ \eta_W \\ \eta_U \end{bmatrix}$, where $\eta_t = \mathbf{0} \in \mathbb{R}^n, \eta_W = -\widetilde{W}$ and $\eta_U = -\widetilde{U}$. We abuse the vector/matrix notation (by ignoring the dimensions) for clarity of presentation, and avoid the vectorization of matrices.*

Now that we have established that optimization problem (4) is invex, we are ready to discuss our main results in the next section.

4. Main Results

In this section, we present our main results along with the technical assumptions. Our main goal is to show that the solution to optimization problem (4) recovers the labels t^* exactly and also recovers a good approximation of β_1^* and β_2^*. In that, we will show that the recovered β_1 and β_2 have the same support and sign as β_1^* and β_2^* respectively and are close to the true vectors in ℓ_2-norm. But before that, we will describe a set of technical assumptions which will help us in our analysis.

4.1. Assumptions

Our first assumption ensures that each sample can be assigned only one label. Formally,

Assumption 4.1 (Identifiability)**.** For $i \in [n]$, $-\frac{1}{2}(y_i - X_i^\intercal \beta_1^*)^2 + \frac{1}{2}(y_i - X_i^\intercal \beta_2^*)^2 \geqslant \epsilon$ if $z_i^* = 1$ and $\frac{1}{2}(y_i - X_i^\intercal \beta_1^*)^2 - \frac{1}{2}(y_i - X_i^\intercal \beta_2^*)^2 \geqslant \epsilon$ if $z_i^* = 0$ for some $\epsilon > 0$.

Clearly, if Assumption 4.1 does not hold for sample i, then we can reverse the label of sample i without increasing objective function of optimization problem (2). Another equivalent way of expressing Assumption 4.1 is as following: for $i \in [n]$, $\langle S_i, W^* \rangle < \langle S_i, U^* \rangle$ if $z_i^* = 1$ and $\langle S_i, W^* \rangle > \langle S_i, U^* \rangle$ if $z_i^* = 0$ where,

$$W^* = \begin{bmatrix} \beta_1^* \\ 1 \end{bmatrix} \begin{bmatrix} \beta_1^{*\intercal} & 1 \end{bmatrix}, \; U^* = \begin{bmatrix} \beta_2^* \\ 1 \end{bmatrix} \begin{bmatrix} \beta_2^{*\intercal} & 1 \end{bmatrix}. \quad (6)$$

Let P denote the support of β_1^*, i.e., $P = \{i \,|\, \beta_{1\,i}^* \neq 0, i \in [d]\}$ and let Q denote the support of β_2^*, i.e., $Q = \{i \,|\, \beta_{2\,i}^* \neq 0, i \in [d]\}$. Similarly, we define their complement as $P^c = \{i \,|\, \beta_{1\,i}^* = 0, i \in [d]\}$ and $Q^c = \{i \,|\, \beta_{2\,i}^* = 0, i \in [d]\}$. We take $|P| = s_1, |P^c| = d - s_1, |Q| = s_2$ and $|Q^c| = d - s_2$. For ease of notation, we define $H \triangleq \mathbb{E}(X_i X_i^\intercal) \forall i \in [n]$. Let $\mathcal{I}_1 \triangleq \{i \,|\, z_i^* = 1, i \in [n]\}$ and $\mathcal{I}_2 \triangleq \{i \,|\, z_i^* = 0, i \in [n]\}$. We define $\widehat{H}_1 \triangleq \frac{1}{n_1} \sum_{i \in \mathcal{I}_1} X_i X_i^\intercal$ and $\widehat{H}_2 \triangleq \frac{1}{n_2} \sum_{i \in \mathcal{I}_2} X_i X_i^\intercal$. As our next assumption, we need the minimum eigenvalue of the population covariance matrix of X restricted to rows and columns in P (similarly in Q) to be greater than zero.

Assumption 4.2 (Positive Definiteness of Hessian)**.** $H_{PP} \succ \mathbf{0}$ and $H_{QQ} \succ \mathbf{0}$ or equivalently $\min(\text{eig}_{\min}(H_{PP}), \text{eig}_{\min}(H_{QQ})) = C_{\min} > 0$. We also assume that $\text{eig}_{\max}(H) = C_{\max} > 0$. Note that $\max(\text{eig}_{\max}(H_{PP}), \text{eig}_{\max}(H_{QQ})) \leqslant C_{\max}$.

In practice, we only deal with finite samples and not populations. In the next lemma, we will show that with a sufficient number of samples, a condition similar to Assumption 4.2 holds with high probability in the finite-sample setting.

Lemma 4.3. *If Assumption 4.2 holds and* $n_1 = \Omega(\frac{s_1 + \log d}{C_{\min}^2})$ *and* $n_2 = \Omega(\frac{s_2 + \log d}{C_{\min}^2})$ *, then* $\min(\text{eig}_{\min}(\widehat{H}_{1_{PP}}), \text{eig}_{\min}(\widehat{H}_{2_{QQ}})) \geqslant \frac{C_{\min}}{2}$ *and* $\max(\text{eig}_{\max}(\widehat{H}_{1_{PP}}), \text{eig}_{\max}(\widehat{H}_{2_{QQ}})) \leqslant \frac{3C_{\max}}{2}$ *with probability at least* $1 - \mathcal{O}(\frac{1}{d})$.

As the third assumption, we will need to ensure that the variates outside the support of β_1^* and β_2^* do not exert lot of influence on the variates in the support of β_1^* and β_2^* respectively. For this, we use a technical condition commonly known as the mutual incoherence condition. It has been previously used in many problems related to regularized regression such as compressed sensing (Wainwright, 2009b), Markov random fields (Ravikumar et al., 2010), non-parametric regression (Ravikumar et al., 2007), diffusion networks (Daneshmand et al., 2014), among others.

Assumption 4.4 (Mutual Incoherence)**.** $\max(\|H_{P^cP} H_{PP}^{-1}\|_\infty, \|H_{Q^cQ} H_{QQ}^{-1}\|_\infty) \leqslant 1 - \xi$ for some $\xi \in (0, 1]$.

Again, we will show that with a sufficient number of samples, a condition similar to Assumption 4.4 holds in the finite-sample setting with high probability.

Lemma 4.5. *If Assumption 4.4 holds and* $n_1 = \Omega(\frac{s_1^3(\log s_1 + \log d)}{\tau(C_{\min}, \xi, \sigma, \Sigma)})$ *and* $n_2 = \Omega(\frac{s_2^3(\log s_2 + \log d)}{\tau(C_{\min}, \xi, \sigma, \Sigma)})$*, then* $\max(\|\widehat{H}_{P^cP} \widehat{H}_{PP}^{-1}\|_\infty, \|\widehat{H}_{Q^cQ} \widehat{H}_{QQ}^{-1}\|_\infty) \leqslant 1 - \frac{\xi}{2}$ *with probability at least* $1 - \mathcal{O}(\frac{1}{d})$ *where* $\tau(C_{\min}, \xi, \sigma, \Sigma)$ *is a constant independent of* n_1, n_2, d, s_1 *and* s_2.

4.2. Main Theorem

Now we are ready to state our main result.

Theorem 4.6. *If Assumptions 4.1, 4.2 and 4.4 hold,* $\lambda_1 \geqslant \frac{64\rho\sigma_e}{\xi}\sqrt{n_1 \log d}$, $\lambda_2 \geqslant \frac{64\rho\sigma_e}{\xi}\sqrt{n_2 \log d}$ *and* $n_1 = \Omega(\frac{s_1^3 \log^2 d}{\tau_0(C_{\min}, \xi, \sigma, \Sigma, \rho)})$ *and* $n_2 = \Omega(\frac{s_2^3 \log^2 d}{\tau_0(C_{\min}, \xi, \sigma, \Sigma, \rho)})$*, then with probability at least* $1 - \mathcal{O}(\frac{1}{d})$ *the solution to the optimization problem* (4) *satisfies the following properties:*

1. *The labels are recovered exactly, i.e.,*

$$t_i = t_i^*, \; \forall i \in [n] \quad (7)$$

2. *The regression parameter vectors are close to the true vectors. Formally,*

$$W = \begin{bmatrix} \beta_1 \\ 1 \end{bmatrix} \begin{bmatrix} \beta_1^\intercal & 1 \end{bmatrix}, \; U = \begin{bmatrix} \beta_2 \\ 1 \end{bmatrix} \begin{bmatrix} \beta_2^\intercal & 1 \end{bmatrix} \quad (8)$$

such that $\beta_1 = \begin{bmatrix} \beta_{1_P} & \mathbf{0}_{P^c} \end{bmatrix}^\intercal$ *and* $\beta_2 = \begin{bmatrix} \beta_{2_Q} & \mathbf{0}_{Q^c} \end{bmatrix}^\intercal$

and

$$\begin{aligned}\|\beta_1 - \beta_1^*\|_2 &\leqslant (2+b_1)\frac{2\lambda_1\sqrt{s_1}}{C_{\min}n_1}\\ \|\beta_2 - \beta_2^*\|_2 &\leqslant (2+b_2)\frac{2\lambda_2\sqrt{s_2}}{C_{\min}n_2}\end{aligned}. \tag{9}$$

In order to prove Theorem 4.6, we will have to show that the labels are recovered exactly. We will also need to show that W and U are rank-1 matrices with eigenvectors $\begin{bmatrix}\beta_1 & 1\end{bmatrix}^\intercal$ and $\begin{bmatrix}\beta_2 & 1\end{bmatrix}^\intercal$ respectively. Moreover, we will also need to ensure that their supports match supports of the true vectors and they are close to true vectors in ℓ_2-norm.

5. Theoretical Analysis

We use primal-dual witness approach to show our results. The primal-dual witness approach was developed by (Wainwright, 2009a) for linear regression problem which has been later used in many convex problems such as Markov random fields (Ravikumar et al., 2010), non-parametric regression (Ravikumar et al., 2007), diffusion networks (Daneshmand et al., 2014) etc. The main idea is to start with a potential solution with certain properties and then later show that these properties are indeed consistent with the final solution. We extend this idea to our invex problem. To that end, we start our proof with a potential solution which has certain "consistency certificate".

5.1. Consistency Certificate

We start by taking solutions W and U with the following properties which we call consistency certificates:

C1. W and U are sparse. In particular, they have the following sparsity structure:

$$\begin{aligned}W &= \begin{bmatrix}W_{PP} & \mathbf{0}_{PP^c} & W_{Pd+1}\\ \mathbf{0}_{P^cP} & \mathbf{0}_{P^cP^c} & \mathbf{0}_{P^cd+1}\\ W_{d+1P} & \mathbf{0}_{d+1P^c} & W_{d+1,d+1}\end{bmatrix}\\ U &= \begin{bmatrix}U_{QQ} & \mathbf{0}_{QQ^c} & U_{Qd+1}\\ \mathbf{0}_{Q^cQ} & \mathbf{0}_{Q^cQ^c} & \mathbf{0}_{Q^cd+1}\\ U_{d+1Q} & \mathbf{0}_{d+1Q^c} & U_{d+1,d+1}\end{bmatrix}\end{aligned} \tag{10}$$

We collect all the non-zero entries of W and U in $\overline{W} \in \mathbb{R}^{s_1+1,s_1+1}$ and $\overline{U} \in \mathbb{R}^{s_2+1,s_2+1}$.

It should be noted that the consistency certificate C1 is not another assumption. In that, eventually we will have to show that it holds in the final solution. We can prove that C1 is consistent with the final solution by showing strict dual feasibility for both W and U which we do in subsection 5.7.

5.2. A Modified Compact Invex Problem

Once we substitute W and U from C1 in optimization problem (4), we get a low dimensional optimization problem.

Definition 5.1 (Compact Invex MLR)**.**

$$\begin{aligned}\min_{t,\overline{W},\overline{U}} \quad & \bar{f}(t,\overline{W},\overline{U}) + \lambda_1\bar{g}(t,\overline{W},\overline{U}) + \lambda_2\bar{h}(t,\overline{W},\overline{U})\\ \text{such that} \quad & \overline{W} \geq \mathbf{0},\ \overline{U} \geq \mathbf{0}\\ & \overline{W}_{s_1+1,s_1+1} = 1,\ \overline{U}_{s_2+1,s_2+1} = 1\\ & \|t\|_\infty \leqslant 1\end{aligned} \tag{11}$$

where $\bar{f}(t,\overline{W},\overline{U}) = \sum_{i=1}^n \frac{1}{2}(\langle \overline{S}_i^P, \overline{W}\rangle + \langle \overline{S}_i^Q, \overline{U}\rangle) + \sum_{i=1}^n \frac{1}{2}t_i(\langle \overline{S}_i^P, \overline{W}\rangle - \langle \overline{S}_i^Q, \overline{U}\rangle)$, $\bar{g}(t,\overline{W},\overline{U}) = \|\overline{W}(:)\|_1$, $\bar{h}(t,\overline{W},\overline{U}) = \|\overline{U}(:)\|_1$ and λ_1 and λ_2 are positive regularizers.

Note that

$$\overline{S}_i^P = \begin{bmatrix}S_{i_{P,P}} & S_{i_{P,d+1}}\\ S_{i_{d+1,P}} & S_{i_{d+1,d+1}}\end{bmatrix},\ \overline{S}_i^Q = \begin{bmatrix}S_{i_{Q,Q}} & S_{i_{Q,d+1}}\\ S_{i_{d+1,Q}} & S_{i_{d+1,d+1}}\end{bmatrix}. \tag{12}$$

For clarity, we will drop the superscripts from $\overline{S}_i$ when the context is clear. Next, we list down the necessary and sufficient conditions to solve optimization problem (11).

5.3. Necessary and Sufficient KKT Conditions

First, we write the Lagrangian $L(\Theta)$ for fixed $\lambda_1 > 0$ and $\lambda_2 > 0$, where $\Theta = (t,\overline{W},\overline{U};\Pi,\Lambda,\alpha,\gamma,\nu,\mu)$ is a collection of parameters.

$$\begin{aligned}L(\Theta) =& \bar{f}(t,\overline{W},\overline{U}) + \lambda_1\bar{g}(t,\overline{W},\overline{U}) + \lambda_2\bar{h}(t,\overline{W},\overline{U})\\ & - \langle\Pi,\overline{W}\rangle - \langle\Lambda,\overline{U}\rangle + \alpha(\overline{W}_{s_1+1,s_1+1} - 1) +\\ & \gamma(\overline{U}_{s_2+1,s_2+1} - 1) - \sum_{i=1}^n \nu_i(t_i+1) + \sum_{i=1}^n \mu_i(t_i - 1)\end{aligned} \tag{13}$$

Here $\Pi \geq \mathbf{0}, \Lambda \geq \mathbf{0}, \alpha \in \mathbb{R}, \gamma \in \mathbb{R}, \nu_i > 0$ and $\mu_i > 0$ are the dual variables (of appropriate dimensions) for optimization problem (11). Using this Lagrangian, the KKT conditions at the optimum can be written as:

1. Stationarity conditions:

$$\sum_{i=1}^n \frac{t_i+1}{2}\overline{S}_i^P + \lambda_1 Z - \Pi + I_\alpha = \mathbf{0} \tag{14}$$

where Z is an element of the subgradient set of $\|\overline{W}(:)\|_1$, i.e., $Z \in \frac{\partial\|\overline{W}(:)\|_1}{\partial\overline{W}}$ and $\|Z(:)\|_\infty \leqslant 1$ and $I_\alpha \in \mathbb{R}^{s_1+1,s_1+1}$ has all zero entries except (s_1+1, s_1+1) entry which is α.

$$\sum_{i=1}^n \frac{1-t_i}{2}\overline{S}_i^Q + \lambda_2 V - \Lambda + I_\gamma = \mathbf{0} \tag{15}$$

where V is an element of the subgradient set of $\|\overline{U}(:)\|_1$, i.e., $V \in \frac{\partial\|\overline{U}(:)\|_1}{\partial\overline{U}}$ and $\|V(:)\|_\infty \leqslant 1$ and $I_\gamma \in \mathbb{R}^{s_2+1,s_2+1}$ has all zero entries except (s_2+1, s_2+1) entry which is γ. For all $i \in [n]$, it holds that

$$\frac{1}{2}\langle \overline{S}_i^P, \overline{W}\rangle - \frac{1}{2}\langle \overline{S}_i^Q, \overline{U}\rangle - \nu_i + \mu_i = 0 \tag{16}$$

2. Complementary Slackness conditions:

$$\langle \Pi, \overline{W}\rangle = 0, \quad \langle \Lambda, \overline{U}\rangle = 0 \tag{17}$$
$$\nu_i(t_i+1) = 0,\ \mu_i(t_i-1) = 0\ \forall i \in [n] \tag{18}$$

3. Dual Feasibility conditions:

$$\Pi \succeq \mathbf{0}, \quad \Lambda \succeq \mathbf{0} \tag{19}$$
$$\nu_i \geqslant 0,\ \mu_i \geqslant 0\ \forall i \in [n] \tag{20}$$

4. Primal Feasibility conditions:

$$\overline{W} \succeq \mathbf{0}, \quad \overline{U} \succeq \mathbf{0}$$
$$\overline{W}_{s_1+1,s_1+1} = 1,\ \overline{U}_{s_2+1,s_2+1} = 1, \quad \|t\|_\infty \leqslant 1 \tag{21}$$

Next, we will provide a setting for primal and dual variables which satisfies all the KKT conditions.

5.4. Construction of Primal and Dual Variables

In this subsection, we will provide a construction of primal and dual variables which satisfies the KKT conditions for optimization problem (11). To that end, we provide our first main result.

Theorem 5.2 (Primal Dual Variables Construction)**.** *If Assumptions 4.1, 4.2 and 4.4 hold, $\lambda_1 \geqslant \frac{64\rho\sigma_e}{\xi}\sqrt{n_1 \log d}$, $\lambda_2 \geqslant \frac{64\rho\sigma_e}{\xi}\sqrt{n_2 \log d}$ and $n_1 = \Omega(\frac{s_1^3 \log^2 d}{\tau_0(C_{\min},\xi,\sigma,\Sigma,\rho)})$ and $n_2 = \Omega(\frac{s_2^3 \log^2 d}{\tau_0(C_{\min},\xi,\sigma,\Sigma,\rho)})$, then the following setting of primal and dual variables:*

- *Primal Variables:*

$$t_i = t_i^*,\ \forall i \in [n]$$
$$\overline{W} = \begin{bmatrix}\tilde{\beta}_1 \\ 1\end{bmatrix}\begin{bmatrix}\tilde{\beta}_1^\intercal & 1\end{bmatrix}, \quad \overline{U} = \begin{bmatrix}\tilde{\beta}_2 \\ 1\end{bmatrix}\begin{bmatrix}\tilde{\beta}_2^\intercal & 1\end{bmatrix}$$

where

$$\tilde{\beta}_1 = \arg\min_{\beta\in\mathbb{R}^{s_1}} \sum_{i=1}^n \frac{t_i^*+1}{2}(y_i - X_{i_P}^\intercal\beta)^2 + \lambda_1(\|\beta\|_1+1)^2$$

and

$$\tilde{\beta}_2 = \arg\min_{\beta\in\mathbb{R}^{s_2}} \sum_{i=1}^n \frac{1-t_i^*}{2}(y_i - X_{i_Q}^\intercal\beta)^2 + \lambda_2(\|\beta\|_1+1)^2$$

- *Dual Variables:*

$$\nu_i = 0, \mu_i = -\frac{1}{2}\langle \overline{S}_i^P, \overline{W}\rangle + \frac{1}{2}\langle \overline{S}_i^Q, \overline{U}\rangle\ \forall i \in \mathcal{I}_1$$
$$\mu_i = 0, \nu_i = \frac{1}{2}\langle \overline{S}_i^P, \overline{W}\rangle - \frac{1}{2}\langle \overline{S}_i^Q, \overline{U}\rangle\ \forall i \in \mathcal{I}_2$$
$$\Pi = \sum_{i=1}^n \frac{t_i^*+1}{2}\overline{S}_i^P + \lambda_1 Z + I_\alpha$$
$$\Lambda = \sum_{i=1}^n \frac{1-t_i^*}{2}\overline{S}_i^Q + \lambda_2 V + I_\gamma$$
$$\alpha = -\langle \sum_{i=1}^n \frac{t_i+1}{2}\overline{S}_i^P + \lambda_1 Z, \overline{W}\rangle$$
$$\gamma = -\langle \sum_{i=1}^n \frac{1-t_i}{2}\overline{S}_i^Q + \lambda_2 V, \overline{U}\rangle$$

satisfies all the KKT conditions for optimization problem (11) with probability at least $1 - \mathcal{O}(\frac{1}{d})$, where $\tau_0(C_{\min}, \alpha, \sigma, \Sigma, \rho, \gamma)$ is a constant independent of s_1, s_2, d, n_1 and n_2 and thus, the primal variables are a globally optimal solution for (11). Furthermore, the above solution is also unique.

Proof Sketch. The main idea behind our proofs is to verify that the setting of primal and dual variables in Theorem 5.2 satisfies all the KKT conditions described in subsection 5.3. We do this by proving multiple lemmas in subsequent subsections. The outline of the proof is as follows:

- It can be trivially verified that the primal feasibility condition (21) holds. The stationarity conditions (14) and (15) holds by construction of Π and Λ respectively. Similarly, the stationarity condition (16) holds by choice of ν_i and μ_i. Choice of t, ν_i, μ_i, α and γ ensure that complementary slackness conditions (17) and (18) also hold.
- In subsection 5.5, we use Lemmas 5.3, 5.5 and 5.6 to verify that the dual feasibility conditions (20) and (19) hold. We will also show in subsection subsection 5.5 that our solution is also unique.

5.5. Verifying Dual Feasibility

To verify dual feasibility, first we will show that $\mu_i \geqslant 0, \forall i \in \mathcal{I}_1$, $\nu_i \geqslant 0, \forall i \in \mathcal{I}_2$. We define $\Delta_1 \triangleq \tilde{\beta}_1 - \beta^*_{1_P}$ and $\Delta_2 \triangleq \tilde{\beta}_2 - \beta^*_{2_Q}$. Then, the following lemma holds true.

Lemma 5.3. *If Assumptions 4.1, 4.2 and 4.4 hold, and $\lambda_1 \geqslant 8\rho\sigma_e\sqrt{n_1 \log d}$, $\lambda_2 \geqslant 8\rho\sigma_e\sqrt{n_2 \log d}$, $n_1 = \Omega(\frac{s_1^3 \log d}{\tau(C_{\min},\xi,\sigma,\Sigma)})$, and $n_2 = \Omega(\frac{s_2^3 \log d}{\tau(C_{\min},\xi,\sigma,\Sigma)})$ then $\|\Delta_1\|_2 \leqslant (2+b_1)\frac{2\lambda_1\sqrt{s_1}}{C_{\min}n_1}$ and $\|\Delta_2\|_2 \leqslant (2+b_2)\frac{2\lambda_2\sqrt{s_2}}{C_{\min}n_2}$*

with probability at least $1-\mathcal{O}(\frac{1}{d})$ *where* $\tau(C_{\min},\xi,\sigma,\Sigma)$ *is a constant independent of* s_1, s_2, d, n_1 *or* n_2.

Using the result of Lemma 5.3, we are going to prove that the settings for dual variables μ_i and ν_i works with high probability.

Lemma 5.4. *If Assumptions 4.1, 4.2 and 4.4 hold, and* $\lambda_1 \geqslant 8\rho\sigma_e\sqrt{n_1\log d}$, $\lambda_2 \geqslant 8\rho\sigma_e\sqrt{n_2\log d}$, $n_1 = \Omega(\frac{s_1^3\log d}{\tau(C_{\min},\xi,\sigma,\Sigma)})$, *and* $n_2 = \Omega(\frac{s_2^3\log d}{\tau(C_{\min},\xi,\sigma,\Sigma)})$ *then* $\mu_i \geqslant 0, \forall i \in \mathcal{I}_1$ *and* $\nu_i \geqslant 0, \forall i \in \mathcal{I}_2$ *with probability at least* $1-\mathcal{O}(\frac{1}{d})$ *where* $\tau(C_{\min},\xi,\sigma,\Sigma)$ *is a constant independent of* s_1, s_2, d, n_1 *or* n_2.

Now we will show that $\Pi \succeq \mathbf{0}$ and $\Lambda \succeq \mathbf{0}$. We will do this in two steps. The first step is to show that both Π and Λ have a zero eigenvalue. In particular,

Lemma 5.5. *Both* Π *and* Λ *have zero eigenvalues corresponding to eigenvectors* $\begin{bmatrix}\tilde{\beta}_1\\1\end{bmatrix}$ *and* $\begin{bmatrix}\tilde{\beta}_2\\1\end{bmatrix}$ *respectively.*

Next, we show that all the other eigenvalues of both Π and Λ are strictly positive.

Lemma 5.6. *If Assumption 4.2 holds and* $n_1 = \Omega(\frac{s_1+\log d}{C_{\min}^2})$ *and* $n_2 = \Omega(\frac{s_2+\log d}{C_{\min}^2})$, *then the second eigenvalues of* Π *and* Λ *are strictly positive with probability at least* $1-\mathcal{O}(\frac{1}{d})$, *i.e.,* $\mathrm{eig}_2(\Pi) > 0$ *and* $\mathrm{eig}_2(\Lambda) > 0$.

On the one hand, Lemma 5.6 ensures that $\Pi \succeq 0$ and $\Lambda \succeq 0$, but on the other it also forces $\overline{W}$ and $\overline{U}$ to be rank-1 and unique as both Π and $\overline{W}$ have to be positive semidefinite and Π has exactly one vector in its nullspace (same with Λ and $\overline{U}$).

5.6. Going back to Invex MLR

Now that we have the setting of $t_i, \overline{W}$ and $\overline{U}$ for Compact Invex MLR problem (11), we can extend these to the original Invex MLR problem (4). Notice that all the other entries of W and U are zeros, thus it readily follows that

$$W = \begin{bmatrix}\beta_1\\1\end{bmatrix}\begin{bmatrix}\beta_1^\intercal & 1\end{bmatrix},\ U = \begin{bmatrix}\beta_2\\1\end{bmatrix}\begin{bmatrix}\beta_2^\intercal & 1\end{bmatrix} \tag{22}$$

where $\beta_1 = \begin{bmatrix}\tilde{\beta}_1\\\mathbf{0}\end{bmatrix}$ and $\beta_2 = \begin{bmatrix}\tilde{\beta}_2\\\mathbf{0}\end{bmatrix}$. Furthermore, result from Lemma 5.3 extends directly and gives us

$$\begin{aligned}\|\beta_1-\beta_1^*\|_2 &\leqslant (2+b_1)\frac{2\lambda_1\sqrt{s_1}}{C_{\min}n_1}\\ \|\beta_2-\beta_2^*\|_2 &\leqslant (2+b_2)\frac{2\lambda_2\sqrt{s_2}}{C_{\min}n_2}\end{aligned}. \tag{23}$$

The last remaining thing is to show that consistency certificate C1 indeed holds which we will do in next subsection.

5.7. Validating Consistency Certificate

Observe that once we substitute $t_i = t_i^*$ in optimization problem (4), it decouples into two independent convex optimization problems involving W and U respectively. Furthermore, since we established that W and U are rank-1, we can rewrite these independent problems in terms of β_1 and β_2. Our task is to show that $\beta_{1_{P^c}} = \mathbf{0}$ and $\beta_{2_{Q^c}} = \mathbf{0}$. It suffices to show it for β_1 as arguments for β_2 are the same. Below, we consider the simplified optimization problem in terms of β_1:

$$\beta_1 = \arg\min_{\beta\in\mathbb{R}^d}\sum_{i\in\mathcal{I}_1}(X_i^\intercal\beta - y_i)^2 + \lambda_1(\|\beta\|_1+1)^2 \tag{24}$$

Since we are only dealing with measurements in $\mathcal{I}_1$, we can substitute $y_i = X_i^\intercal\beta^* + e_i$. Furthermore, β_1 must satisfy stationarity KKT condition which can be written as:

$$\begin{aligned}&\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}X_iX_i^\intercal(\beta_1-\beta^*) - \frac{1}{n_1}\sum_{i\in\mathcal{I}_1}X_ie_i +\\ &\frac{1}{n_1}\lambda_1(\|\beta_1\|_1+1)\zeta = \mathbf{0},\end{aligned} \tag{25}$$

where ζ is in the subdifferential set of $\|\beta_1\|_1$ and $\|\zeta\|_\infty \leqslant 1$. Specifically, $\zeta_i = \mathrm{sign}(\beta_1), \forall i \in P$ and $\zeta_i \in [-1,1], \forall i \in P^c$. Our task is to show that ζ fulfills strict dual feasibility, i.e., $\|\zeta_{P^c}\|_\infty < 1$. We decompose equation (25) into two parts – one corresponding to entries in P and the other corresponding to entries in P^c. For entries in P, we have

$$\begin{aligned}&\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}X_{i_P}X_{i_P}^\intercal(\beta_{1_P}-\beta_{1_P}^*) - \frac{1}{n_1}\sum_{i\in\mathcal{I}_1}X_{i_P}e_i\\ &+\frac{1}{n_1}\lambda_1(\|\beta_1\|_1+1)\zeta_P = \mathbf{0}\end{aligned} \tag{26}$$

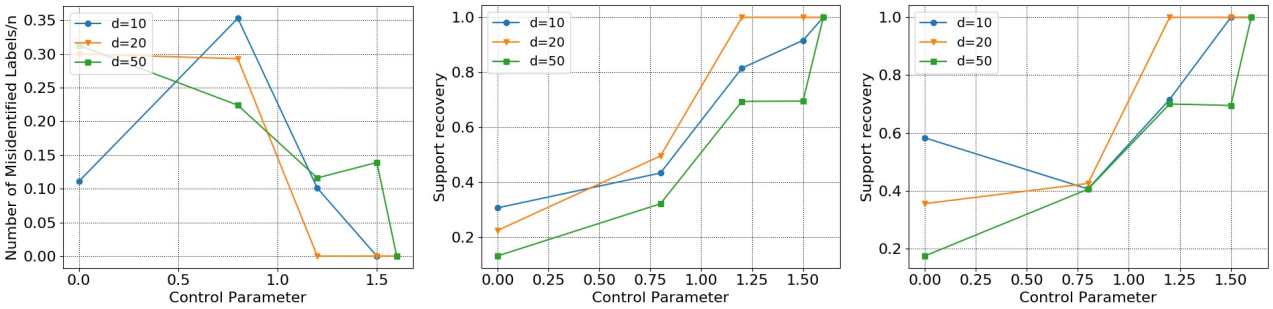
Similarly, for entries in P^c, we have

$$\begin{aligned}&\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}X_{i_{P^c}}X_{i_P}^\intercal(\beta_{1_P}-\beta_{1_P}^*) - \frac{1}{n_1}\sum_{i\in\mathcal{I}_1}X_{i_{P^c}}e_i\\ &+\frac{1}{n_1}\lambda_1(\|\beta_1\|_1+1)\zeta_{P^c} = \mathbf{0}\end{aligned}$$

After rearranging the terms and substituting for $(\beta_{1_P}-\beta_{1_P}^*)$ from equation (26), we get

$$\begin{aligned}&\frac{\lambda_1}{n_1}(1+\|\beta_1\|_1)\zeta_{P^c} = -\widehat{H}_{P^cP}\widehat{H}_{PP}^{-1}(\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}X_{i_P}e_i -\\ &\frac{1}{n_1}\lambda_1(\|\beta_1\|_1+1)\zeta_P) + \frac{1}{n_1}\sum_{i\in\mathcal{I}_1}X_{i_{P^c}}e_i\end{aligned}$$

Let $\bar{\lambda}_1 = \frac{\lambda_1}{n_1}$ and note that $\|\beta_1\|_1 \geqslant 0$, using norm inequali-



(a) Misidentified labels (in ratio to n) (b) Support recovery of β_1^* (c) Support recovery of β_2^*

Figure 1. Label and support recovery with control parameter C_p.

ties we can rewrite the above equation as:

$$\|\zeta_{P^c}\|_\infty \leqslant \|\widehat{H}_{P^cP}\widehat{H}_{PP}^{-1}\|_\infty(\|\frac{1}{\bar{\lambda}_1}\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{i_P}e_i\|_\infty + \|\zeta_P\|_\infty)$$
$$+ \|\frac{1}{\bar{\lambda}_1}\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{i_{P^c}}e_i\|_\infty$$

We know that $\|\widehat{H}_{P^cP}\widehat{H}_{PP}^{-1}\|_\infty \leqslant (1-\frac{\xi}{2})$ for some $\xi \in (0,1]$. The following lemma provides bounds on $\|\frac{1}{\bar{\lambda}_1}\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{i_P}e_i\|_\infty$ and $\|\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{i_{P^c}}e_i\|_\infty$.

Lemma 5.7. *Let $\lambda_1 \geqslant \frac{64\rho\sigma_e}{\xi}\sqrt{n_1\log d}$. Then the following holds true:*

$$\mathbb{P}(\|\frac{1}{\bar{\lambda}_1}\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{i_P}e_i\|_\infty \geqslant \frac{\xi}{8-4\xi}) \leqslant \mathcal{O}(\frac{1}{d}),$$
$$\mathbb{P}(\|\frac{1}{\bar{\lambda}_1}\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{i_{P^c}}e_i\|_\infty \geqslant \frac{\xi}{8}) \leqslant \mathcal{O}(\frac{1}{d})$$

It follows that $\|\zeta_{P^c}\|_\infty \leqslant 1-\frac{\xi}{4}$ with probability at least $1-\mathcal{O}(\frac{1}{d})$. Thus, the consistency certificate C1 indeed holds with high probability.

6. Experimental Validation

Note that we are not proposing any new algorithm in our paper. However, to validate our theoretical results we performed experiments on synthetic data. We generated response y using Gaussian random variables X and chose regression parameter β_1^* (or β_2^*) based on the label of the samples. We fixed the sparsity $s_1 = s_2 = 4$, however supports were not necessarily the same for both the regression parameter vectors. We varied n_1 and n_2 according to our theorems, i.e., both were varied with $10^{C_p}\log^2 d$ for $d = 10, 20$ and 50 where C_p is a control parameters. The regularizers were kept according to our theorem and were varied as $\mathcal{O}(\sqrt{n_1\log d})$ and $\mathcal{O}(\sqrt{n_2\log d})$. We measured performance of our algorithm based on the label recovery (in ratio to supplied n) and support recovery for both parameter vectors. The experiments were run three times independently. Note how we make zero mistakes in label recovery as we increase number of samples. Similarly, support recovery (ratio of intersection and union with correct support) for both parameter vectors goes to 1 as we increase sample size. It should be noted that we do not propose any new algorithm but our method is free of any initialization requirement. As for specific algorithm for empirical verification, we use a projected subgradient method (Duchi & Singer, 2009) to check convergence for our problem which is achieved without any requirement on initialization. In fact, any algorithm which converges to a stationary point should work for our framework.

7. Concluding Remarks

We provide a novel formulation of invex MLR. We show that invexity of our optimization problem allows for a tractable solution. We provide provable theoretical guarantees for our solution. The sample complexity of our method is polynomial in terms of sparsity and logarithmic in terms of the dimension of the true parameter. Our method helps to identify labels exactly and recovers regression parameter vectors with correct support and correct sign. It would be interesting to think about extending our ideas to mixture of more than two groups of regressions in future.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2134209-DMS.

References

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 15:2773–2832, 2014.

Barik, A. and Honorio, J. Fair sparse regression with clustering: An invex relaxation for a combinatorial problem. *Advances in neural information processing systems*, 2021.

Ben-Israel, A. and Mond, B. What is invexity? *The ANZIAM Journal*, 28(1):1–9, 1986.

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

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

Daneshmand, H., Gomez-Rodriguez, M., Song, L., and Schoelkopf, B. Estimating Diffusion Network Structures: Recovery Conditions, Sample Complexity & Soft-Thresholding Algorithm. In *International Conference on Machine Learning*, pp. 793–801, 2014.

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

Duchi, J. and Singer, Y. Efficient online and batch learning using forward backward splitting. *The Journal of Machine Learning Research*, 10:2899–2934, 2009.

Elhamifar, E. and Vidal, R. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11): 2765–2781, 2013.

Ghosh, A. and Kannan, R. Alternating minimization converges super-linearly for mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 1093–1103. PMLR, 2020.

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

Hanson, M. A. On sufficiency of the kuhn-tucker conditions. *Journal of Mathematical Analysis and Applications*, 80 (2):545–550, 1981.

Haynsworth, E. V. Determination of the inertia of a partitioned hermitian matrix. *Linear algebra and its applications*, 1(1):73–81, 1968.

Horn, R. A. and Johnson, C. R. *Matrix Analysis*. Cambridge university press, 2012.

Hsu, D. and Kakade, S. M. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pp. 11–20, 2013.

Hsu, D., Kakade, S., Zhang, T., et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

Jordan, M. I. and Jacobs, R. A. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2): 181–214, 1994.

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

Li, G., Pan, Y., Yang, Z., and Ma, J. Modeling vehicle merging position selection behaviors based on a finite mixture of linear regression models. *IEEE Access*, 7: 158445–158458, 2019.

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. Spam: Sparse Additive Models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 1201–1208. Curran Associates Inc., 2007.

Ravikumar, P., Wainwright, M. J., Lafferty, J. D., et al. High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Ravikumar, P., Wainwright, M. J., Raskutti, G., Yu, B., et al. High-dimensional Covariance Estimation by Minimizing L1-Penalized Log-Determinant Divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

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

Vershynin, R. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012.

Vidal, R., Ma, Y., and Sastry, S. Generalized principal component analysis (gpca). *IEEE transactions on pattern analysis and machine intelligence*, 27(12):1945–1959, 2005.

Viele, K. and Tong, B. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12(4):315–330, 2002.

Wainwright, M. J. Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Info. Theory*, 55:5728–5741, December 2009a.

Wainwright, M. J. Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using L1-Constrained Quadratic Programming (Lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009b.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wedel, M. and DeSarbo, W. S. A mixture likelihood approach for generalized linear models. *Journal of classification*, 12(1):21–55, 1995.

Wedel, M. and Kamakura, W. A. *Market segmentation: Conceptual and methodological foundations*. Springer Science & Business Media, 2000.

Wu, C. J. On the convergence properties of the em algorithm. *The Annals of statistics*, pp. 95–103, 1983.

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

Supplementary Material: Sparse Mixed Linear Regression with Guarantees: Taming an Intractable Problem with Invex Relaxation

A. Continuous Relaxation of Standard MLR is non-convex

It suffices to prove that the objective function $l(z, \beta_1, \beta_2)$ of optimization problem (2) is non-convex when z_i is allowed to be between 0 and 1. We note that

$$l(z, \beta_1, \beta_2) = \sum_{i=1}^{n} z_i(y_i - X_i^\intercal \beta_1)^2 + (1 - z_i)(y_i - X_i^\intercal \beta_2)^2 \tag{27}$$

where $\beta_1 \in \mathbb{R}^d, \beta_2 \in \mathbb{R}^d$ and $z_i \in [0, 1], \forall i \in [n]$. Let $\Theta = (z, \beta_1, \beta_2, \bar{z}, \bar{\beta}_1, \bar{\beta}_2)$. We consider the following quantity:

$$F(\Theta) = f(z, \beta_1, \beta_2) - f(\bar{z}, \bar{\beta}_1, \bar{\beta}_2) - \sum_{i=1}^{n} \frac{\partial f}{\partial \bar{z}_i}(z_i - \bar{z}_i) - \frac{\partial f}{\partial \bar{\beta}_1}^\intercal (\beta_1 - \bar{\beta}_1) - \frac{\partial f}{\partial \bar{\beta}_2}^\intercal (\beta_2 - \bar{\beta}_2) \tag{28}$$

where

$$\begin{aligned} \frac{\partial f}{\partial \bar{z}_i} &= (y_i - X_i^\intercal \bar{\beta}_1)^2 - (y_i - X_i^\intercal \bar{\beta}_2)^2, \quad \forall i \in [n] \\ \frac{\partial f}{\partial \bar{\beta}_1} &= \sum_{i=1}^{n} -2\bar{z}_i X_i (y_i - X_i^\intercal \bar{\beta}_1) \\ \frac{\partial f}{\partial \bar{\beta}_2} &= \sum_{i=1}^{n} -2(1 - \bar{z}_i) X_i (y_i - X_i^\intercal \bar{\beta}_2) \end{aligned} \tag{29}$$

It suffices to show that $F(\Theta)$ changes sign for different feasible values of Θ. We choose the following variables:

$$\begin{aligned} z_i = 0, \bar{z}_i = \frac{1}{2} \quad \forall i \in [n] \\ \beta_{1_k} = u_1, \beta_{1_j} = 0, \forall j \neq k \\ \beta_{2_l} = u_2, \beta_{2_j} = 0, \forall j \neq l \\ \bar{\beta}_{1_k} = w_1, \bar{\beta}_{1_j} = 0, \forall j \neq k \\ \bar{\beta}_{2_l} = w_2, \bar{\beta}_{2_j} = 0, \forall j \neq l \\ u_1 = w_1 - (u_2 - w_2) \end{aligned} \tag{30}$$

Note that choice of w_1, u_2 and w_2 can be arbitrary. This simplifies $F(\Theta)$:

$$F(\Theta) = \sum_{i=1}^{n} (y_i - u_2 X_{il})^2 - \sum_{i=1}^{n} (y_i - w_2 X_{il})^2 \tag{31}$$

Consider the case when $X_{il} > 0, \forall i \in [n]$. Then choosing $u_2 < w_2$ makes $F(\Theta) > 0$ while choosing $u_2 > w_2$ makes $F(\Theta) < 0$. This proves our claim.

B. Proof of Lemma 3.4

Lemma 3.4 *For $(t, W, U) \in C$, the functions $f(t, W, U) = \sum_{i=1}^{n} \frac{1}{2}\langle S_i, W + U\rangle + \sum_{i=1}^{n} \frac{1}{2} t_i \langle S_i, W - U\rangle, g(t, W, U) = \|W(:)\|_1$ and $h(t, W, U) = \|U(:)\|_1$ are η-invex for $\eta(t, \tilde{t}, W, \widetilde{W}, U, \widetilde{U}) \triangleq \begin{bmatrix} \eta_t \\ \eta_W \\ \eta_U \end{bmatrix}$, where $\eta_t = \mathbf{0} \in \mathbb{R}^n, \eta_W = -\widetilde{W}$ and $\eta_U = -\widetilde{U}$. We abuse the vector/matrix notation (by ignoring the dimensions) for clarity of presentation, and avoid the vectorization of matrices.*

Proof. We know $f(t,W,U) = \sum_{i=1}^{n} \frac{t_i+1}{2}\langle S_i, W\rangle + \frac{1-t_i}{2}\langle S_i, U\rangle$. Then,

$$\begin{aligned}\frac{\partial f}{\partial t_i} &= \frac{1}{2}\langle S_i, W-U\rangle \\ \frac{\partial f}{\partial W} &= \sum_{i=1}^{n}\frac{t_i+1}{2}S_i \\ \frac{\partial f}{\partial U} &= \sum_{i=1}^{n}\frac{1-t_i}{2}S_i\end{aligned} \tag{32}$$

To prove that $f(t,W,U)$ is invex, we need to show that

$$f(t,W,U) - f(\tilde{t},\widetilde{W},\widetilde{U}) - \sum_{i=1}^{n}\eta_{t_i}\frac{\partial f}{\partial \tilde{t}_i} - \langle \eta_W, \frac{\partial f}{\partial \widetilde{W}}\rangle - \langle \eta_U, \frac{\partial f}{\partial \widetilde{U}}\rangle \geqslant 0 \tag{33}$$

We take $\eta_t = \mathbf{0} \in \mathbb{R}^n, \eta_W = -\widetilde{W}$ and $\eta_U = -\widetilde{U}$ and expand LHS of equation (33) as follows:

$$\begin{aligned}&\sum_{i=1}^{n}\frac{t_i+1}{2}\langle S_i, W\rangle + \frac{1-t_i}{2}\langle S_i, U\rangle - \sum_{i=1}^{n}\frac{\tilde{t}_i+1}{2}\langle S_i, \widetilde{W}\rangle - \frac{1-\tilde{t}_i}{2}\langle S_i, \widetilde{U}\rangle + \langle \widetilde{W}, \sum_{i=1}^{n}\frac{\tilde{t}_i+1}{2}S_i\rangle + \langle \widetilde{U}, \sum_{i=1}^{n}\frac{1-\tilde{t}_i}{2}S_i\rangle \\ &= \sum_{i=1}^{n}\frac{t_i+1}{2}\langle S_i, W\rangle + \frac{1-t_i}{2}\langle S_i, U\rangle \\ &\geqslant 0\end{aligned} \tag{34}$$

The last inequality holds because S_i, W and U are all positive semidefinite and $t_i \in [-1,1]$.

Similarly,

$$\begin{aligned}&g(t,W,U) - g(\tilde{t},\widetilde{W},\widetilde{U}) - \sum_{i=1}^{n}\eta_{t_i}\frac{\partial g}{\partial \tilde{t}_i} - \langle \eta_W, \frac{\partial g}{\partial \widetilde{W}}\rangle - \langle \eta_U, \frac{\partial g}{\partial \widetilde{U}}\rangle \\ &= \|W(:)\|_1 - \|\widetilde{W}(:)\|_1 + \|\widetilde{W}(:)\|_1 \geqslant 0\end{aligned} \tag{35}$$

and

$$\begin{aligned}&h(t,W,U) - h(\tilde{t},\widetilde{W},\widetilde{U}) - \sum_{i=1}^{n}\eta_{t_i}\frac{\partial h}{\partial \tilde{t}_i} - \langle \eta_W, \frac{\partial h}{\partial \widetilde{W}}\rangle - \langle \eta_U, \frac{\partial h}{\partial \widetilde{U}}\rangle \\ &= \|U(:)\|_1 - \|\widetilde{U}(:)\|_1 + \|\widetilde{U}(:)\|_1 \geqslant 0\end{aligned} \tag{36}$$

□

C. Proof of Lemma 4.3

Lemma 4.3 *If Assumption 4.2 holds and* $n_1 = \Omega(\frac{s_1+\log d}{C_{\min}^2})$ *and* $n_2 = \Omega(\frac{s_2+\log d}{C_{\min}^2})$ *, then* $\min(\text{eig}_{\min}(\widehat{H}_{1_{PP}}), \text{eig}_{\min}(\widehat{H}_{2_{QQ}})) \geqslant \frac{C_{\min}}{2}$ *and* $\max(\text{eig}_{\max}(\widehat{H}_{1_{PP}}), \text{eig}_{\max}(\widehat{H}_{2_{QQ}})) \leqslant \frac{3C_{\max}}{2}$ *with probability at least* $1 - \mathcal{O}(\frac{1}{d})$.

Proof. We prove the Lemma for a general support S and samples n. The results follow when we substitute S by P and Q and n by n_1 or n_2 based on the context. By the Courant-Fischer variational representation (Horn & Johnson, 2012):

$$\begin{aligned}\text{eig}_{\min}(\mathbb{E}(X_iX_i^\intercal)_{SS}) = \min_{\|y\|_2=1} y^\intercal\mathbb{E}(X_iX_i^\intercal)_{SS}y &= \min_{\|y\|_2=1} y^\intercal(\mathbb{E}(X_iX_i^\intercal)_{SS} - \frac{1}{n}X_S^\intercal X_S + \frac{1}{n}X_S^\intercal X_S)y \\ &\leqslant y^\intercal(\mathbb{E}(X_iX_i^\intercal)_{SS} - \frac{1}{n}X_S^\intercal X_S + \frac{1}{n}X_S^\intercal X_S)y \\ &= y^\intercal(\mathbb{E}(X_iX_i^\intercal)_{SS} - \frac{1}{n}X_S^\intercal X_S)y + y^\intercal\frac{1}{n}X_S^\intercal X_S y\end{aligned} \tag{37}$$

It follows that

$$\text{eig}_{\min}(\frac{1}{n}X_S^\intercal X_S) \geqslant C_{\min} - \|\mathbb{E}(X_i X_i^\intercal)_{SS} - \frac{1}{n}X_S^\intercal X_S\|_2 \tag{38}$$

The term $\|\mathbb{E}(X_i X_i^\intercal)_{SS} - \frac{1}{n}X_S^\intercal X_S\|_2$ can be bounded using Proposition 2.1 in (Vershynin, 2012) for sub-Gaussian random variables. In particular,

$$\mathbb{P}(\|\mathbb{E}(X_i X_i^\intercal)_{SS} - \frac{1}{n}X_S^\intercal X_S\|_2 \geqslant \epsilon) \leqslant 2\exp(-c\epsilon^2 n + s) \tag{39}$$

for some constant $c > 0$. Taking $\epsilon = \frac{C_{\min}}{2}$, we show that $\text{eig}_{\min}(\frac{1}{n}X_S^\intercal X_S) \geqslant \frac{C_{\min}}{2}$ with probability at least $1 - 2\exp(-\frac{cC_{\min}^2 n}{4} + |S|)$. The specific results for n_1 and n_2 follow directly.

Remark: Similarly, it can be shown that $\text{eig}_{\max}(\frac{1}{n}X_S^\intercal X_S) \leqslant \frac{3C_{\max}}{2}$ with probability at least $1 - 2\exp(-\frac{cC_{\max}^2 n}{4} + |S|)$. □

D. Proof of Lemma 4.5

Lemma 4.5 *If Assumption 4.4 holds and* $n_1 = \Omega(\frac{s_1^3(\log s_1 + \log d)}{\tau(C_{\min}, \xi, \sigma, \Sigma)})$ *and* $n_2 = \Omega(\frac{s_2^3(\log s_2 + \log d)}{\tau(C_{\min}, \xi, \sigma, \Sigma)})$, *then* $\max(\|\widehat{H}_{P^cP}\widehat{H}_{PP}^{-1}\|_\infty, \|\widehat{H}_{Q^cQ}\widehat{H}_{QQ}^{-1}\|_\infty) \leqslant 1 - \frac{\xi}{2}$ *with probability at least* $1 - \mathcal{O}(\frac{1}{d})$ *where* $\tau(C_{\min}, \xi, \sigma, \Sigma)$ *is a constant independent of* n_1, n_2, d, s_1 *and* s_2.

Proof. We prove the Lemma for a general support S (and corresponding non-support S^c) and samples n. The results follow when we substitute S by P and Q and n by n_1 or n_2 based on the context. Let $|S| = s$ and $|S^c| = d - s$. Before we prove the result of Lemma 4.5, we will prove a helper lemma.

Lemma D.1. *If Assumption 4.4 holds then for some* $\delta > 0$, *the following inequalities hold:*

$$\begin{aligned}
&\mathbb{P}(\|\widehat{H}_{S^cS} - H_{S^cS}\|_\infty \geqslant \delta) \leqslant 4(d-s)s\exp(-\frac{n\delta^2}{128s^2(1+4\sigma^2)\max_l \Sigma_{ll}^2}) \\
&\mathbb{P}(\|\widehat{H}_{SS} - H_{SS}\|_\infty \geqslant \delta) \leqslant 4s^2\exp(-\frac{n\delta^2}{128s^2(1+4\sigma^2)\max_l \Sigma_{ll}^2}) \\
&\mathbb{P}(\|(\widehat{H}_{SS})^{-1} - (H_{SS})^{-1}\|_\infty \geqslant \delta) \leqslant 2\exp(-\frac{c\delta^2 C_{\min}^4 n}{4s} + s) + 2\exp(-\frac{cC_{\min}^2 n}{4} + s)
\end{aligned} \tag{40}$$

Proof. Let A_{ij} be (i,j)-th entry of $\widehat{H}_{S^cS} - H_{S^cS}$. Clearly, $\mathbb{E}(A_{ij}) = 0$. By using the definition of the $\|\cdot\|_\infty$ norm, we can write:

$$\begin{aligned}
\mathbb{P}(\|\widehat{H}_{S^cS} - H_{S^cS}\|_\infty \geqslant \delta) &= \mathbb{P}(\max_{i \in S^c}\sum_{j \in S}|A_{ij}| \geqslant \delta) \\
&\leqslant (d-s)\mathbb{P}(\sum_{j \in S}|A_{ij}| \geqslant \delta) \\
&\leqslant (d-s)s\mathbb{P}(|A_{ij}| \geqslant \frac{\delta}{s})
\end{aligned} \tag{41}$$

where the second last inequality comes as a result of the union bound across entries in S^c and the last inequality is due to the union bound across entries in S. Recall that $X_i, i \in [d]$ are zero mean random variables with covariance Σ and each $\frac{X_i}{\sqrt{\Sigma_{ii}}}$ is a sub-Gaussian random variable with parameter σ. Using the results from Lemma 1 of (Ravikumar et al., 2011), for some $\delta \in (0, s\max_l \Sigma_{ll} 8(1+4\sigma^2))$, we can write:

$$\mathbb{P}(|A_{ij}| \geqslant \frac{\delta}{s}) \leqslant 4\exp(-\frac{n\delta^2}{128s^2(1+4\sigma^2)\max_l \Sigma_{ll}^2}) \tag{42}$$

Therefore,

$$\mathbb{P}(\|\widehat{H}_{S^cS} - H_{S^cS}\|_\infty \geqslant \delta) \leqslant 4(d-s)s\exp(-\frac{n\delta^2}{128s^2(1+4\sigma^2)\max_l \Sigma_{ll}^2}) \tag{43}$$

Similarly, we can show that

$$\mathbb{P}(\|\widehat{H}_{SS} - H_{SS}\|_\infty \geqslant \delta) \leqslant 4s^2 \exp(-\frac{n\delta^2}{128s^2(1+4\sigma^2)\max_l \Sigma^2_{ll}}) \tag{44}$$

Next, we will show that the third inequality in (40) holds. Note that

$$\begin{aligned}\|(\widehat{H}_{S^cS})^{-1} - (H_{S^cS})^{-1}\|_\infty &= \|(H_{SS})^{-1}(H_{SS} - \widehat{H}_{SS})(\widehat{H}_{SS})^{-1}\|_\infty \\ &\leqslant \sqrt{s}\|(H_{SS})^{-1}(H_{SS} - \widehat{H}_{SS})(\widehat{H}_{SS})^{-1}\|_2 \\ &\leqslant \sqrt{s}\|(H_{SS})^{-1}\|_2\|(H_{SS} - \widehat{H}_{SS})\|_2\|(\widehat{H}_{SS})^{-1}\|_2\end{aligned} \tag{45}$$

Note that $\|H_{SS}\|_2 \geqslant C_{\min}$, thus $\|(H_{SS})^{-1}\|_2 \leqslant \frac{1}{C_{\min}}$. Similarly, $\|H_{SS}\|_2 \geqslant \frac{C_{\min}}{2}$ with probability at least $1 - 2\exp(-\frac{cC^2_{\min}n}{4} + s)$. We also have $\|(H_{SS} - \widehat{H}_{SS})\|_2 \leqslant \epsilon$ with probability at least $1 - 2\exp(-c\epsilon^2 n + s)$. Taking $\epsilon = \delta\frac{C^2_{\min}}{2\sqrt{s}}$, we get

$$\mathbb{P}(\|(H_{SS} - \widehat{H}_{SS})\|_2 \geqslant \delta\frac{C^2_{\min}}{2\sqrt{s}}) \leqslant 2\exp(-\frac{c\delta^2 C^4_{\min}n}{4s} + s) \tag{46}$$

It follows that $\|(\widehat{H}_{SS})^{-1} - (H_{SS})^{-1}\|_\infty \leqslant \delta$ with probability at least $1 - 2\exp(-\frac{c\delta^2C^4_{\min}n}{4s} + s) - 2\exp(-\frac{cC^2_{\min}n}{4} + s)$. □

Now we are ready to show that the statement of Lemma 4.5 holds using the results from Lemma D.1. We will rewrite $\widehat{H}_{S^cS}(\widehat{H}_{SS})^{-1}$ as the sum of four different terms:

$$\widehat{H}_{S^cS}(\widehat{H}_{SS})^{-1} = T_1 + T_2 + T_3 + T_4, \tag{47}$$

where

$$\begin{aligned}T_1 &\triangleq \widehat{H}_{S^cS}((\widehat{H}_{SS})^{-1} - (H_{SS})^{-1}) \\ T_2 &\triangleq (\widehat{H}_{S^cS} - H_{S^cS})(H_{SS})^{-1} \\ T_3 &\triangleq (\widehat{H}_{S^cS} - H_{S^cS})((\widehat{H}_{SS})^{-1} - (H_{SS})^{-1}) \\ T_4 &\triangleq H_{S^cS}(H_{SS})^{-1}.\end{aligned} \tag{48}$$

Then it follows that $\|\widehat{H}_{S^cS}(\widehat{H}_{SS})^{-1}\|_\infty \leqslant \|T_1\|_\infty + \|T_2\|_\infty + \|T_3\|_\infty + \|T_4\|_\infty$. Now, we will bound each term separately. First, recall that Assumption 4.4 ensures that $\|T_4\|_\infty \leqslant 1 - \xi$.

Controlling T_1. We can rewrite T_1 as,

$$T_1 = -H_{S^cS}(H_{SS})^{-1}(\widehat{H}_{SS} - H_{SS})(\widehat{H}_{SS})^{-1} \tag{49}$$

then,

$$\begin{aligned}\|T_1\|_\infty &= \|H_{S^cS}(H_{SS})^{-1}(\widehat{H}_{SS} - H_{SS})(\widehat{H}_{SS})^{-1}\|_\infty \\ &\leqslant \|H_{S^cS}(H_{SS})^{-1}\|_\infty\|(\widehat{H}_{SS} - H_{SS})\|_\infty\|(\widehat{H}_{SS})^{-1}\|_\infty \\ &\leqslant (1-\xi)\|(\widehat{H}_{SS} - H_{SS})\|_\infty\sqrt{s}\|(\widehat{H}_{SS})^{-1}\|_2 \\ &\leqslant (1-\xi)\|(\widehat{H}_{SS} - H_{SS})\|_\infty\frac{2\sqrt{s}}{C_{\min}} \\ &\leqslant \frac{\xi}{6}\end{aligned} \tag{50}$$

The last inequality holds with probability at least $1 - 2\exp(-\frac{cC^2_{\min}n}{4} + s) - 4s^2\exp(-\frac{nC^2_{\min}\xi^2}{18432(1-\xi)^2s^3(1+4\sigma^2)\max_l\Sigma^2_{ll}})$ by taking $\delta = \frac{C_{\min}\xi}{12(1-\xi)\sqrt{s}}$.

Controlling T_2. Recall that $T_2 = (\widehat{H}_{S^cS} - H_{S^cS})(H_{SS})^{-1}$. Thus,

$$\begin{aligned}\|T_2\|_\infty &\leqslant \sqrt{s}\|(H_{SS})^{-1}\|_2\|(\widehat{H}_{S^cS} - H_{S^cS})\|_\infty \\ &\leqslant \frac{\sqrt{s}}{C_{\min}}\|(\widehat{H}_{S^cS} - H_{S^cS})\|_\infty \\ &\leqslant \frac{\xi}{6}\end{aligned} \tag{51}$$

The last inequality holds with probability at least $1 - 4(d-s)s\exp(-\frac{nC_{\min}^2\xi^2}{4608s^3(1+4\sigma^2)\max_l \Sigma_{ll}^2})$ by choosing $\delta = \frac{C_{\min}\xi}{6\sqrt{s}}$.

Controlling T_3. Note that,

$$\begin{aligned}\|T_3\|_\infty &\leqslant \|(\widehat{H}_{S^cS} - H_{S^cS})\|_\infty\|((\widehat{H}_{SS})^{-1} - (H_{SS})^{-1})\|_\infty \\ &\leqslant \frac{\xi}{6}\end{aligned} \tag{52}$$

The last inequality holds with probability at least $1 - 4(d-s)s\exp(-\frac{n\xi}{768s^2(1+4\sigma^2)\max_l \Sigma_l l^2}) - 2\exp(-\frac{c\xi C_{\min}^4 n}{24s} + s) - 2\exp(-\frac{cC_{\min}^2 n}{4} + s)$ by choosing $\delta = \sqrt{\frac{\xi}{6}}$ in the first and third inequality of equation (40). By combining all the above results, we prove Lemma 4.5. The specific results for n_1 and n_2 follow directly. □

E. Proof of Lemma 5.3

Lemma 5.3. *If Assumptions 4.1, 4.2 and 4.4 hold, and $\lambda_1 \geqslant 8\rho\sigma_e\sqrt{n_1\log d}$, $\lambda_2 \geqslant 8\rho\sigma_e\sqrt{n_2\log d}$, $n_1 = \Omega(\frac{s_1^3\log d}{\tau(C_{\min},\xi,\sigma,\Sigma)})$, and $n_2 = \Omega(\frac{s_2^3\log d}{\tau(C_{\min},\xi,\sigma,\Sigma)})$ then $\|\Delta_1\|_2 \leqslant (2+b_1)\frac{2\lambda_1\sqrt{s_1}}{C_{\min}n_1}$ and $\|\Delta_2\|_2 \leqslant (2+b_2)\frac{2\lambda_2\sqrt{s_2}}{C_{\min}n_2}$ with probability at least $1 - \mathcal{O}(\frac{1}{d})$ where $\tau(C_{\min},\xi,\sigma,\Sigma)$ is a constant independent of s_1, s_2, d, n_1 or n_2.*

Proof. It suffices to prove the result for Δ_1 as the result for Δ_2 follows in the same way. Note,

$$\begin{aligned}\tilde{\beta}_1 &= \arg\min_{\beta\in\mathbb{R}^{s_1}} \sum_{i=1} \frac{t_i^*+1}{2}(y_i - X_{i_P}^\intercal\beta)^2 + \lambda_1(\|\beta\|_1+1)^2 \\ &= \arg\min_{\beta\in\mathbb{R}^{s_1}} \sum_{i\in\mathcal{I}_1} (y_i - X_{i_P}^\intercal\beta)^2 + \lambda_1(\|\beta\|_1+1)^2\end{aligned}$$

The optimal $\tilde{\beta}_1$ must satisfy stationarity KKT condition at the optimum, i.e.,

$$\sum_{i\in\mathcal{I}_1} X_{i_P}(-y_i + X_{i_P}^\intercal\tilde{\beta}_1) + z\lambda_1(z^\intercal\tilde{\beta}_1 + 1) = \mathbf{0}$$

where $\|\tilde{\beta}_1\|_1 = z^\intercal\tilde{\beta}_1$ and z is in the subdifferential set of $\|\tilde{\beta}_1\|_1$ and $\|z\|_\infty \leqslant 1$. Since $i \in \mathcal{I}_1$, we can substitute $y_i = X_{i_P}^\intercal\beta_{1_P}^* + e_i$.

$$(\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{i_P}X_{i_P}^\intercal + \frac{1}{n_1}\lambda_1 zz^\intercal)(\beta_{1_P}^* - \tilde{\beta}_{1_P}) + \frac{1}{n_1}(\sum_{i\in\mathcal{I}_1} X_{i_P}e_i) + \frac{1}{n_1}\lambda_1 zz^\intercal\beta_{1_P}^* + \frac{1}{n_1}\lambda_1 z = \mathbf{0}$$

Note that $\widehat{H}_{1_{PP}} = \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{i_P}X_{i_P}^\intercal$. Using norm-inequalities:

$$\|\Delta_1\|_2 \leqslant \|(\widehat{H}_{1_{PP}} + \frac{\lambda_1}{n_1}zz^\intercal)^{-1}\|_2(\|\frac{1}{n_1}(\sum_{i\in\mathcal{I}_1} X_{i_P}e_i)\|_2 + \|\frac{1}{n_1}\lambda_1 zz^\intercal\beta_{1_P}^*\|_2 + \|\frac{1}{n_1}\lambda_1 z\|_2) \tag{53}$$

Using Lemma 4.3, $\text{eig}_{\min}(\widehat{H}_{1_{PP}}) \geqslant \frac{C_{\min}}{2}$, and using Weyl's inequality $\text{eig}_{\min}(\widehat{H}_{1_{PP}} + \frac{\lambda_1}{n_1}zz^\intercal) \geqslant \frac{C_{\min}}{2}$. It follows that $\|(\widehat{H}_{1_{PP}} + \frac{\lambda_1}{n_1}zz^\intercal)^{-1}\|_2 \leqslant \frac{2}{C_{\min}}$.

$$\|\Delta_1\|_2 \leqslant \frac{2}{C_{\min}}(\|\frac{1}{n_1}(\sum_{i\in\mathcal{I}_1} X_{i_P}e_i)\|_2 + \frac{\lambda_1\sqrt{s_1}}{n_1}\|\beta_{1_P}^*\|_1 + \frac{\lambda_1\sqrt{s_1}}{n_1}) \tag{54}$$

We know that $\|\beta_1^*\|_1 \leqslant b_1$. Thus,

$$\|\Delta_1\|_2 \leqslant \frac{2}{C_{\min}}(\|\frac{1}{n_1}(\sum_{i\in\mathcal{I}_1} X_{i_P} e_i)\|_2 + \frac{\lambda_1\sqrt{s_1}}{n_1}b_1 + \frac{\lambda_1\sqrt{s_1}}{n_1}) \tag{55}$$

It only remains to bound $\|\frac{1}{n_1}(\sum_{i\in\mathcal{I}_1} X_{i_P} e_i)\|_2$ which we do in the following lemma.

Lemma E.1. *If $\lambda_1 \geqslant 8\rho\sigma_e\sqrt{n_1 \log d}$, then $\|\frac{1}{n_1}(\sum_{i\in\mathcal{I}_1} X_{i_P} e_i)\|_2 \leqslant \sqrt{s_1}\frac{\lambda_1}{n_1}$ with probability at least $1 - \mathcal{O}(\frac{1}{d})$*

Thus, it follows that

$$\|\Delta_1\|_\infty \leqslant \|\Delta_1\|_2 \leqslant (2 + b_1)\frac{2\lambda_1\sqrt{s_1}}{C_{\min} n_1} \tag{56}$$

□

F. Proof of Lemma E.1

Lemma E.1 *If $\lambda_1 \geqslant 8\rho\sigma_e\sqrt{n_1 \log d}$, then $\|\frac{1}{n_1}(\sum_{i\in\mathcal{I}_1} X_{i_P} e_i)\|_2 \leqslant \sqrt{s_1}\frac{\lambda_1}{n_1}$ with probability at least $1 - \mathcal{O}(\frac{1}{d})$.*

Proof. We will start with $\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{i_P} e_i$. We take the i-th entry of $\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{i_P} e_i$ for some $i \in P$, i.e., $|\frac{1}{n}\sum_{j\in\mathcal{I}_1} X_{ji}\mathbf{e}_j|$. Recall that X_{ji} is a sub-Gaussian random variable with parameter ρ^2 and e_j is a sub-Gaussian random variable with parameter σ_e^2. Then, $\frac{X_{ji}}{\rho}\frac{e_j}{\sigma_e}$ is a sub-exponential random variable with parameters $(4\sqrt{2}, 2)$. Using the concentration bounds for the sum of independent sub-exponential random variables (Wainwright, 2019), we can write:

$$\mathbb{P}(|\frac{1}{n_1}\sum_{j\in\mathcal{I}_1}\frac{X_{ji}}{\rho}\frac{e_j}{\sigma_e}| \geqslant t) \leqslant 2\exp(-\frac{n_1 t^2}{64}),\ 0 \leqslant t \leqslant 8 \tag{57}$$

Taking a union bound across $i \in P$:

$$\begin{aligned}&\mathbb{P}(\exists i \in P \mid |\frac{1}{n_1}\sum_{j\in\mathcal{I}_1}\frac{X_{ji}}{\rho}\frac{e_j}{\sigma_e}| \geqslant t) \leqslant 2s_1\exp(-\frac{n_1 t^2}{64})\\ &0 \leqslant t \leqslant 8\end{aligned} \tag{58}$$

It follows that $\|\|\frac{1}{n_1}(\sum_{i\in\mathcal{I}_1} X_{i_P} e_i)\|_2\|_2 \leqslant \sqrt{s}t$ with probability at least $1 - 2s\exp(-\frac{nt^2}{64\rho^2\sigma_e^2})$ for some $0 \leqslant t \leqslant 8\rho\sigma_e$. Taking $t = \frac{\lambda_1}{n}$, we get the desired result. □

G. Proof of Lemma 5.4

Lemma 5.4 *If Assumptions 4.1, 4.2 and 4.4 hold, and $\lambda_1 \geqslant 8\rho\sigma_e\sqrt{n_1 \log d}$, $\lambda_2 \geqslant 8\rho\sigma_e\sqrt{n_2 \log d}$, $n_1 = \Omega(\frac{s_1^3 \log^2 d}{\tau(C_{\min},\xi,\sigma,\Sigma)})$, and $n_2 = \Omega(\frac{s_2^3 \log^2 d}{\tau(C_{\min},\xi,\sigma,\Sigma)})$ then $\mu_i \geqslant 0, \forall i \in \mathcal{I}_1$ and $\nu_i \geqslant 0, \forall i \in \mathcal{I}_2$ with probability at least $1 - \mathcal{O}(\frac{1}{d})$ where $\tau(C_{\min}, \xi, \sigma, \Sigma)$ is a constant independent of s_1, s_2, d, n_1 or n_2.*

Proof. We start with the setting of μ_i when i is in $\mathcal{I}_1$.

$$\begin{aligned}\mu_i &= -\frac{1}{2}\langle \overline{S}_i^P, \overline{W}\rangle + \frac{1}{2}\langle \overline{S}_i^Q, \overline{U}\rangle\\ &= -\frac{1}{2}(y_i - X_{i_P}^\intercal\tilde{\beta}_1)^2 + \frac{1}{2}(y_i - X_{i_Q}^\intercal\tilde{\beta}_2)^2\\ &= -\frac{1}{2}(y_i - X_{i_P}^\intercal\beta_{1_P}^* + X_{i_P}^\intercal(\tilde{\beta}_1 - \beta_{1_P}^*))^2 + \frac{1}{2}(y_i - X_{i_Q}^\intercal\beta_{2_Q}^* + X_{i_Q}^\intercal(\tilde{\beta}_2 - \beta_{2_Q}^*))^2\\ &= -\frac{1}{2}((y_i - X_{i_P}^\intercal\beta_{1_P}^*)^2 + (\tilde{\beta}_1 - \beta_{1_P}^*)^\intercal X_{i_P}X_{i_P}^\intercal(\tilde{\beta}_1 - \beta_{1_P}^*) + 2(y_i - X_{i_P}^\intercal\beta_{1_P}^*)X_{i_P}^\intercal(\tilde{\beta}_1 - \beta_{1_P}^*)) +\\ &\frac{1}{2}((y_i - X_{i_Q}^\intercal\beta_{2_Q}^*)^2 + (\tilde{\beta}_2 - \beta_{2_Q}^*)^\intercal X_{i_Q}X_{i_Q}^\intercal(\tilde{\beta}_2 - \beta_{2_Q}^*) + 2(y_i - X_{i_Q}^\intercal\beta_{2_Q}^*)X_{i_Q}^\intercal(\tilde{\beta}_2 - \beta_{2_P}^*))\end{aligned} \tag{59}$$

Since $i \in \mathcal{I}_1$, we can substitute $y_i = X_{i_P}^\intercal \beta_{1_P}^* + e_i$.

$$\begin{aligned}\mu_i =& -\frac{1}{2}(y_i - X_{i_P}^\intercal \beta_{1_P}^*)^2 + \frac{1}{2}(y_i - X_{i_Q}^\intercal \beta_{2_Q}^*)^2 - \frac{1}{2}\Delta_1^\intercal X_{i_P} X_{i_P}^\intercal \Delta_1 + \frac{1}{2}\Delta_2^\intercal X_{i_Q} X_{i_Q}^\intercal \Delta_2 - e_i X_{i_P}^\intercal \Delta_1 + \\ & (X_{i_P}^\intercal \beta_{1_P}^* + e_i - X_{i_Q}^\intercal \beta_{2_Q}^*) X_{i_Q}^\intercal \Delta_2 \\ =& -\frac{1}{2}(y_i - X_{i_P}^\intercal \beta_{1_P}^*)^2 + \frac{1}{2}(y_i - X_{i_Q}^\intercal \beta_{2_Q}^*)^2 - \frac{1}{2}\Delta_1^\intercal X_{i_P} X_{i_P}^\intercal \Delta_1 + \frac{1}{2}\Delta_2^\intercal X_{i_Q} X_{i_Q}^\intercal \Delta_2 - e_i X_{i_P}^\intercal \Delta_1 + \\ & e_i X_{i_Q}^\intercal \Delta_2 + (\beta_1^* - \beta_2^*)^\intercal X_i X_i^\intercal (\beta_2 - \beta_2^*)\end{aligned} \tag{60}$$

Using bounds on the eigenvalue of data matrix, Assumption 4.1 and bounds on $\|\Delta_1\|_2$ and $\|\Delta_2\|_2$, we can place a bound on μ_i.

$$\mu_i \geqslant \epsilon - \frac{n_1 3 C_{\max}}{2}\|\Delta_1\|_2^2 + \frac{n_2 C_{\min}}{2}\|\Delta_2\|_2^2 - |e_i X_{i_P}^\intercal \Delta_1| - |e_i X_{i_Q}^\intercal \Delta_2| - \frac{n 3 C_{\max}}{2}\|(\beta_1^* - \beta_2^*)\|_2 \|\Delta_2\|_2 \tag{61}$$

We still need bound to bound $|e_i X_{i_P}^\intercal \Delta_1|$ and $|e_i X_{i_Q}^\intercal \Delta_2|$ which we do in the following lemma.

Lemma G.1. *The following holds:*

1. *For fixed* $\|\Delta_1\|_2$, $\mathbb{P}(|e_i X_{i_P}^\intercal \Delta_1| \leqslant \frac{\epsilon}{4})$ *with probability at least* $1 - \mathcal{O}(\frac{1}{d})$.
2. *For fixed* $\|\Delta_2\|_2$, $\mathbb{P}(|e_i X_{i_Q}^\intercal \Delta_2| \leqslant \frac{\epsilon}{4})$ *with probability at least* $1 - \mathcal{O}(\frac{1}{d})$.

Proof. Recall that $X_{i_P}^\intercal \Delta_1$ is a sub-Gaussian random variable with parameter $\rho^2\|\Delta_1\|_2^2$ and e_i is a sub-Gaussian random variable with parameter σ_e^2. Then, $\frac{X_{i_P}^\intercal \Delta_1}{\rho\|\Delta_1\|_2}\frac{e_i}{\sigma_e}$ is a sub-exponential random variable with parameters $(4\sqrt{2}, 2)$. Using the concentration bounds for the sum of independent sub-exponential random variables (Wainwright, 2019), we can write:

$$\mathbb{P}(|\frac{X_{i_P}^\intercal \Delta_1}{\rho\|\Delta_1\|_2}\frac{e_i}{\sigma_e}| \geqslant t) \leqslant 2\exp(-\frac{t^2}{64}),\ 0 \leqslant t \leqslant 8 \tag{62}$$

Taking $t = \frac{t}{\rho\|\Delta_1\|_2\sigma_e}$, we get

$$\mathbb{P}(|e_i X_{i_P}^\intercal \Delta_1| \geqslant t) \leqslant 2\exp(-\frac{t^2}{64\rho^2\|\Delta_1\|_2^2\sigma_e^2}),\ 0 \leqslant t \leqslant 8\rho\|\Delta_1\|_2\sigma_e \tag{63}$$

We take $t = \frac{\epsilon}{4}$, then

$$\mathbb{P}(|e_i X_{i_P}^\intercal \Delta_1| \geqslant \frac{\epsilon}{4}) \leqslant 2\exp(-\frac{\epsilon^2}{16 \times 64\rho^2\|\Delta_1\|_2^2\sigma_e^2}),\ 0 \leqslant \epsilon \leqslant 32\rho\|\Delta_1\|_2\sigma_e \tag{64}$$

Since $\|\Delta_1\|_2$ is upper bounded with $\mathcal{O}(\frac{\lambda_1}{n_1}\sqrt{s_1})$ and n_1 is of order $\mathcal{O}(s_1^3 \log^2 d)$, thus $\mathbb{P}(|e_i X_{i_P}^\intercal \Delta_1| \leqslant \frac{\epsilon}{4})$ with probability at least $1 - \mathcal{O}(\frac{1}{d})$. Similarly, $\mathbb{P}(|e_i X_{i_Q}^\intercal \Delta_2| \leqslant \frac{\epsilon}{4})$ with probability at least $1 - \mathcal{O}(\frac{1}{d})$. □

Till now, we have considered $\|\Delta_1\|_2$ and $\|\Delta_2\|_2$ to be fixed quantity, however they are also upper bounded by $\mathcal{O}(\frac{\lambda_1}{n_1}\sqrt{s_1})$ with probability at least $1 - \mathcal{O}(\frac{1}{d})$, thus the overall probability that $u_i \geqslant 0, i \in \mathcal{I}_1$ is at least $1 - \mathcal{O}(\frac{1}{d})$ as long as

$$\epsilon \geqslant 3n_1 C_{\max}\|\Delta_1\|_2^2 - n_2 C_{\min}\|\Delta_2\|_2^2 + 3nC_{\max}\|\beta_1^* - \beta_2^*\|_2\|\Delta_2\|_2 \tag{65}$$

We need to take a union bound across entries in $\mathcal{I}_1$ which changes the probability to at least $1 - \mathcal{O}(\exp(-\log d + \log n_1))$ which is still dominated by $1 - \mathcal{O}(\frac{1}{d})$.

□

H. Proof of Lemma 5.5

Lemma 5.5 *Both Π and Λ have zero eigenvalues corresponding to eigenvectors $\begin{bmatrix}\tilde{\beta}_1\\1\end{bmatrix}$ and $\begin{bmatrix}\tilde{\beta}_2\\1\end{bmatrix}$ respectively.*

Proof. It suffices to prove the result for Π as the result for Λ follows in the same way. Note,

$$\begin{aligned}\tilde{\beta}_1 &= \arg\min_{\beta\in\mathbb{R}^{s_1}} \sum_{i=1} \frac{t_i^*+1}{2}(y_i - X_{i_P}^\intercal\beta)^2 + \lambda_1(\|\beta\|_1+1)^2\\ &= \arg\min_{\beta\in\mathbb{R}^{s_1}} \sum_{i\in\mathcal{I}_1} (y_i - X_{i_P}^\intercal\beta)^2 + \lambda_1(\|\beta\|_1+1)^2\end{aligned}$$

The optimal $\tilde{\beta}_1$ must satisfy stationarity KKT condition at the optimum, i.e.,

$$\sum_{i\in\mathcal{I}_1} X_{i_P}(-y_i + X_{i_P}^\intercal\tilde{\beta}_1) + z\lambda_1(z^\intercal\tilde{\beta}_1+1) = \mathbf{0}$$

By little algebraic manipulation, we can rewrite the above as following:

$$(\sum_{i=1}^{n} \frac{t_i^*+1}{2}\bar{S}_i^P + \lambda_1 Z + I_\alpha)\begin{bmatrix}\tilde{\beta}_1\\1\end{bmatrix} = \mathbf{0}$$

where $Z = \begin{bmatrix}z\\1\end{bmatrix}\begin{bmatrix}z^\intercal & 1\end{bmatrix} = \text{sign}(\overline{W})$. Clearly,

$$\Pi\begin{bmatrix}\tilde{\beta}_1\\1\end{bmatrix} = \mathbf{0}$$

Similarly, we can show

$$\Lambda\begin{bmatrix}\tilde{\beta}_2\\1\end{bmatrix} = \mathbf{0}$$

□

I. Proof of Lemma 5.6

Lemma 5.6 *If Assumption 4.2 holds and $n_1 = \Omega(\frac{s_1+\log d}{C_{\min}^2})$ and $n_2 = \Omega(\frac{s_2+\log d}{C_{\min}^2})$, then the second eigenvalues of Π and Λ are strictly positive with probability at least $1-\mathcal{O}(\frac{1}{d})$, i.e., $\text{eig}_2(\Pi) > 0$ and $\text{eig}_2(\Lambda) > 0$.*

Proof. It suffices to prove the result for Π as similar arguments can be used to prove the result for Λ. We know

$$\begin{aligned}\Pi &= \sum_{i\in\mathcal{I}_1} S_i^P + \lambda_1 Z + I_\alpha\\ &= \sum_{i\in\mathcal{I}_1}\begin{bmatrix}X_iX_i^\intercal & -X_iy_i\\ -y_iX_i^\intercal & y_i^2\end{bmatrix} + \lambda_1\begin{bmatrix}zz^\intercal & z\\ z^\intercal & 1\end{bmatrix} + I_\alpha\\ &= \begin{bmatrix}\sum_{i\in\mathcal{I}_1} X_iX_i^\intercal + \lambda_1 zz^\intercal & \sum_{i\in\mathcal{I}_1} -X_iy_i + \lambda_1 z\\ \sum_{i\in\mathcal{I}_1} -y_iX_i^\intercal + \lambda_1 z^\intercal & \sum_{i\in\mathcal{I}_1} y_i^2 + \lambda_1 + \alpha\end{bmatrix}\end{aligned} \tag{66}$$

Also note that $\alpha = -\langle\sum_{i\in\mathcal{I}_1} S_i^P + \lambda_1 Z, \overline{W}\rangle = -\sum_{i\in\mathcal{I}_1}(y_i - X_{i_P}^\intercal\tilde{\beta}_1)^2 + \lambda_1(\|\tilde{\beta}_1\|_1+1)^2$. We also know that $\tilde{\beta}_1$ satisfies the stationarity KKT condition, i.e.,

$$\sum_{i\in\mathcal{I}_1} X_{i_P}(-y_i + X_{i_P}^\intercal\tilde{\beta}_1) + z\lambda_1(z^\intercal\tilde{\beta}_1+1) = \mathbf{0}$$

$$\tilde{\beta}_1 = -(\sum_{i\in\mathcal{I}_1} X_iX_i^\intercal + \lambda_1 zz^\intercal)^{-1}(\sum_{i\in\mathcal{I}_1} -X_iy_i + \lambda_1 z)$$

Using the stationarity KKT condition, we can simplify objective function value of optimization problem (11) at $\tilde{\beta}_1$ to $\sum_{i\in\mathcal{I}_1} y_i^2 + (\sum_{i\in\mathcal{I}_1} -y_i X_i^\intercal + \lambda_1 z^\intercal)\tilde{\beta}_1 + \lambda_1$. Now, we invoke Haynesworth's inertia additivity formula (Haynsworth, 1968) to prove our claim. Let R be a block matrix of the form $R = \begin{bmatrix} A & B \\ B^\intercal & C \end{bmatrix}$, then inertia of matrix R, denoted by $\text{In}(R)$, is defined as the tuple $(|\text{eig}_+(R)|, |\text{eig}_-(R)|, |\text{eig}_0(R)|)$ where $|\text{eig}_+(R)|$ is the number of positive eigenvalues, $|\text{eig}_-(R)|$ is the number of negative eigenvalues and $|\text{eig}_0(R)|$ is the number of zero eigenvalues of matrix R. Haynesworth's inertia additivity formula is given as:

$$\text{In}(R) = \text{In}(A) + \text{In}(C - B^\intercal A^{-1} B) \tag{67}$$

We take $A = \sum_{i\in\mathcal{I}_1} X_i X_i^\intercal + \lambda_1 z z^\intercal$, $B = \sum_{i\in\mathcal{I}_1} -X_i y_i + \lambda_1 z$ and $C = \sum_{i\in\mathcal{I}_1} y_i^2 + \lambda_1 + \alpha$. It should be noted that $C - B^\intercal A^{-1} B$ evaluates to zero. Thus,

$$\text{In}(\Pi) = \text{In}(\sum_{i\in\mathcal{I}_1} X_i X_i^\intercal + \lambda_1 z z^\intercal) + \text{In}(0) \tag{68}$$

We note that 0 has precisely one zero eigenvalue and no other eigenvalues. Moreover, from Lemma 4.3 and Weyl's inequality:

$$\text{eig}_{\min}(\sum_{i\in\mathcal{I}_1} X_i X_i^\intercal + \lambda_1 z z^\intercal) \geqslant \frac{C_{\min}}{2} > 0 \tag{69}$$

with probability at least $1 - \mathcal{O}(\frac{1}{d})$ as long as $n_1 = \Omega(\frac{s_1 + \log d}{C_{\min}^2})$. It follows that the second eigenvalue of Π is strictly positive. Similar, arguments can be made for Λ. □

J. Proof of Lemma 5.7

Lemma 5.7 *Let $\lambda_1 \geqslant \frac{64\rho\sigma_e}{\xi}\sqrt{n_1 \log d}$. Then the following holds true:*

$$\mathbb{P}(\|\frac{1}{\bar{\lambda}_1}\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{i_P} e_i\|_\infty \geqslant \frac{\xi}{8 - 4\xi}) \leqslant \mathcal{O}(\frac{1}{d}),$$
$$\mathbb{P}(\|\frac{1}{\bar{\lambda}_1}\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{i_{P^c}} e_i\|_\infty \geqslant \frac{\xi}{8}) \leqslant \mathcal{O}(\frac{1}{d})$$

Proof. We will start with $\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{i_P} e_i$. We take the i-th entry of $\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{i_P} e_i$ for some $i \in P$, i.e., $|\frac{1}{n}\sum_{j\in\mathcal{I}_1} X_{ji}\mathbf{e}_j|$. Recall that X_{ji} is a sub-Gaussian random variable with parameter ρ^2 and e_j is a sub-Gaussian random variable with parameter σ_e^2. Then, $\frac{X_{ji}}{\rho}\frac{e_j}{\sigma_e}$ is a sub-exponential random variable with parameters $(4\sqrt{2}, 2)$. Using the concentration bounds for the sum of independent sub-exponential random variables (Wainwright, 2019), we can write:

$$\mathbb{P}(|\frac{1}{n_1}\sum_{j\in\mathcal{I}_1}\frac{X_{ji}}{\rho}\frac{e_j}{\sigma_e}| \geqslant t) \leqslant 2\exp(-\frac{n_1 t^2}{64}),\ 0 \leqslant t \leqslant 8 \tag{70}$$

Taking a union bound across $i \in P$:

$$\mathbb{P}(\exists i \in P \mid |\frac{1}{n_1}\sum_{j\in\mathcal{I}_1}\frac{X_{ji}}{\rho}\frac{e_j}{\sigma_e}| \geqslant t) \leqslant 2s_1\exp(-\frac{n_1 t^2}{64})$$
$$0 \leqslant t \leqslant 8 \tag{71}$$

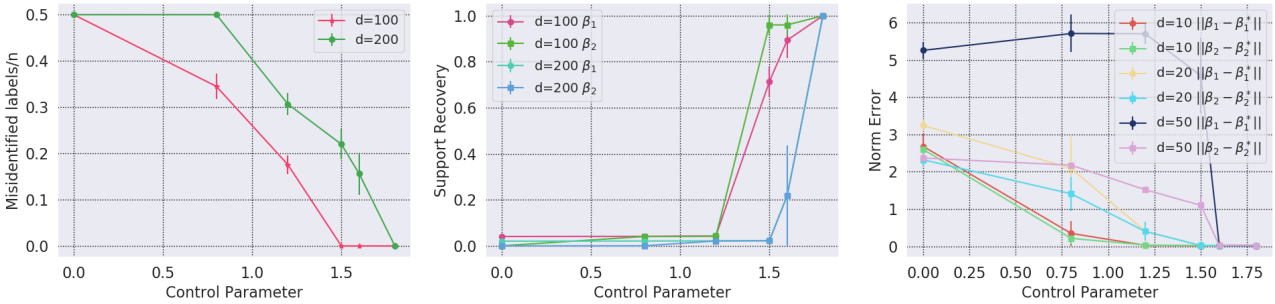
Taking $t = \frac{\bar{\lambda}_1 t}{\rho\sigma_e}$, we get:

$$\mathbb{P}(\exists i \in P \mid |\frac{1}{\bar{\lambda}_1}\frac{1}{n_1}\sum_{j=1}^{n} X_{ji} e_j| \geqslant t) \leqslant 2s_1\exp(-\frac{n_1\bar{\lambda}_1^2 t^2}{64\rho^2\sigma_e^2})$$
$$0 \leqslant t \leqslant 8\frac{\rho\sigma_e}{\bar{\lambda}_1} \tag{72}$$

It follows that $\|\frac{1}{\bar{\lambda}_1}\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{i_P} e_i\|_\infty \leqslant t$ with probability at least $1 - 2s_1\exp(-\frac{n\bar{\lambda}_1^2 t^2}{64\rho^2\sigma_e^2})$.

Using a similar argument, we can show that $\|\frac{1}{\bar{\lambda}_1}\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{i_{P^c}} e_i\|_\infty \leqslant t$ with probability at least $1-2(d-s_1)\exp(-\frac{n_1\bar{\lambda}_1^2 t^2}{64\rho^2\sigma_e^2})$. Taking $t = \frac{\xi}{8-4\xi}$ and $\frac{\xi}{8}$ in the first and second inequality of Lemma 5.7 and choosing the provided setting of λ_1 and n_1 completes our proof. □

K. Additional Experiments



(a) Misidentified labels (b) Support recovery (c) ℓ_2 Norm error

Figure 2. Label and support recovery with control parameter C_p for high dimensional case $d = 100, 200$). We also show that the norm error indeed goes towards 0.

Following the setting mentioned in Section 6, we conduct further high dimensional ($d = 100, 200$) experiments to validate our theoretical results. We observe a similar trend, i.e., as we increase number of samples, we make zero mistakes in label recovery and achieve 100% correct support recovery for both parameter vectors. Additionally, we also show that the norm error, i.e., $\|\beta_j - \beta_j^*\|_2, j \in \{1, 2\}$ goes towards zero in our experiments.