Journal of Machine Learning Research ? (2024) 1-20 Submitted 4/24; Revised ?/24; Published ?/24

Distributed Algorithms for High-Dimensional Statistical Inference and Structure Learning with Heterogeneous Data

Hongru Zhao ZHAO1118@UMN.EDU
School of Statistics
University of Minnesota,
Minneapolis, MN, USA

Xiaotong Shen XSHEN@UMN.EDU
School of Statistics
University of Minnesota,
Minneapolis, MN, USA

Editor: My editor

Abstract

This paper addresses the challenge of sharing individual-level data across multiple sites in studies, hindered by privacy and legal concerns, and site diversity. We introduce a novel statistical framework for the distributed processing of heterogeneous data, enabling comprehensive analysis using nonconvex regularization techniques. Specifically, we develop linear regression methods that handle heterogeneous data, facilitate structure learning, and identify global and site-specific effects. We provide a theoretical foundation for these methods, designing efficient algorithms for our framework, particularly through regularization with an ℓ_0 constraint. We demonstrate that our approach, which employs Difference of Convex (DC) programming and the ℓ_0 projection algorithm, can approximate the global minimizer in polynomial time with probability tending to one under the data generation distribution. This is significant given that, in the worst-case scenario, it has been shown that no algorithm can resolve this nonconvex minimization problem in polynomial time. Additionally, our method suggests applying an ℓ_0 constraint on nuisance parameters to encourage sparsity while unregularizing hypothesized parameters in testing. We demonstrate the validity of the inference methods for global effects by suitably identifying local effects, contingent upon precise alignment of the tuning parameter with the sparsity level.

Keywords: Multi-site studies, global effect, inference regularization, asymptotic analysis.

1. Introduction

Multicenter research, particularly for clinical data, offers significant advantages over single-center studies, including the ability to collect larger sample sizes, which enhances the generalizability of findings and to pool resources, expertise, and ideas across collaborating sites (Sidransky et al., 2009). Refer to Cheng et al. (2017) for more information. However, privacy regulations frequently limit access to individual-level data, complicating the data aggregation process from multiple sites (Barrows Jr and Clayton, 1996). There is a high demand for efficient tools to synthesize evidence across clinical sites.

Federated learning, as discussed in Konecnỳ et al. (2016) and McMahan et al. (2017), represents a distributed machine learning approach that permits the training of models on

©2024 Zhao and Shen.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v?/21-0000.html.

decentralized data without direct data sharing. Various federated learning frameworks have been available to manage heterogeneous data distributions and outliers, enhancing model performance and stability (Khaled et al., 2020; Wang et al., 2019; Han et al., 2021; Guo et al., 2023).

A key challenge in distributed computation is integrating statistical inference to manage uncertainty with heterogeneous data across different sites. Duan et al. (2022) introduces a distributed algorithm that considers heterogeneous distributions by including site-specific nuisance parameters essential for reflecting site-specific variations. However, this approach relies on the efficient score function to mitigate the impact of inaccurate estimations of these parameters, which may falter when the number of nuisance parameters exceeds the sample size. Due to the complexities of multiple sites and limited sample sizes at each site, previous research often utilizes regularization to prevent overfitting (Wang et al., 2017; Battey et al., 2018; Jordan et al., 2018). These studies propose communication-efficient distributed algorithms for optimization and regression, underlining the statistical inference complexities in decentralized settings. Yet, they do not account for site-specific nuisance parameters crucial for depicting heterogeneity across sites. Our paper addresses this gap by integrating site-specific nuisance parameters and regularization in a high-dimensional context, facilitating the management of overparametrized settings where the number of parameters substantially exceeds the sample size.

This paper will focus on statistical inference for distributed algorithms in linear models to assimilate heterogeneous data involving regularization. This exploration addresses the crucial requirement for integrating inference with distributed computation, enhancing the precision and reliability of statistical methods within distributed environments. Our approach distinguishes itself from existing methods by employing a likelihood approach for higher efficacy rather than relying on surrogate methods. Specifically, we introduce a linear regression framework designed to estimate the global effect across heterogeneous data sets by integrating data from multiple sites while managing site-specific effects individually. This integration is achieved through the application of regularization techniques. By pooling information from multiple sites to estimate a global effect, the overall sample size increases, leading to more efficient estimation and improved inference quality compared to using data from individual sites alone. Furthermore, we develop algorithms to execute this process utilizing nonlinear regularization via an ℓ_0-constraint. As showed in Theorem 5.1, our constrained Difference of Convex (DC) algorithm with the ℓ_0 projection attains a global minimizer in polynomial time, with probability tending to one under the data generation distribution. This result is in contrast to a negative result that in the worst case scenario there does not exist an algorithm that can resolve this nonconvex minimization in polynomial time (Chen et al., 2017, 2019).

In the context of composite hypotheses, we present a hypothesis test that preserves the parameters of interest without regularization, while applying an ℓ_0-constraint on nuisance parameters, such as numerous site-specific parameters, to enhance the power of the test. We derive the asymptotic distribution of the global effect for inference. Additionally, we establish a theoretical guarantee of the validity of the proposed algorithms. Our key result demonstrates that the algorithm achieves over-selection consistency, ensuring that the supports of the oracle estimators are subsets of the estimated supports with high probability. Moreover, when the sparsity tuning parameter precisely aligns with the true sparsity level,

our estimator achieves support recovery, guaranteeing accurate identification of the true model structure. These theoretical findings underscore the effectiveness of our methodology in high-dimensional settings.

The rest of the paper is organized as follows. Section 2 introduces the heterogeneous linear model and establishes the necessary notation. Section 3 presents the constrained optimization approach using the ℓ_0-constraint and provides the general computational algorithm and the distributed version of the algorithm. Section 4 establishes the theoretical properties of our estimator and the constrained likelihood ratio test, including the generalized Wilks' phenomenon. In Section 5, we demonstrate the convergence and consistency of our proposed algorithm in general linear model setting. Finally, Section 6 summarizes our findings and discusses the implications of our work.

1.1 Our Contribution

Our main contributions are four-folded.

1. We introduce a new statistical framework specifically designed for the distributed processing of heterogeneous data, enabling comprehensive global analysis through nonconvex regularization techniques. Our research is dedicated to developing linear regression methods that effectively handle heterogeneous data, facilitating structure learning, and distinguishing between global and site-specific effects. By aggregating information from multiple sites to ascertain a global effect, we increase the overall sample size. This leads to more efficient estimation and superior inference quality compared to analyzing data from individual sites alone.

2. We develop efficient algorithms to execute the proposed methodology, utilizing nonlinear regularization with an ℓ_0-constraint. Although finding an approximately optimal solution for our optimization problem has been shown to be NP-hard in the worst-case scenario, we demonstrate that our constrained minimization approach using DC programming and the ℓ_0 projection algorithm can obtain the global minimizer with probability tending to one under the data generation distribution.

3. We present a hypothesis testing strategy for composite hypotheses that preserves the parameters of interest without regularization, while applying an ℓ_0-constraint on other parameters, such as numerous site-specific parameters, to ensure adequate control of their sparsity. We establish the asymptotic properties of the constrained likelihood ratio test, including the generalized Wilks' phenomenon, facilitating accurate inference in high-dimensional settings.

4. We demonstrate the convergence and consistency of our proposed algorithm in a general linear model setting. Our key result shows that the algorithm achieves over-selection consistency, ensuring that the supports of the oracle estimators are subsets of the estimated supports with high probability. Moreover, when the sparsity tuning parameter aligns precisely with the true sparsity level, our estimator attains support recovery, guaranteeing the accurate identification of the true model structure. These theoretical findings highlight the effectiveness of our methodology in high-dimensional settings.

2. Heterogeneous Linear Model

In this section, we introduce the heterogeneous linear model and establish the necessary notation for the rest of the paper. Our work focuses on developing linear regression methods tailored for heterogeneous data, enabling structural learning and distinguishing between global and site-specific influences. Assume we have access to K independent training datasets. At each site j, we consider loss function $L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j)$, where $\boldsymbol{\beta}_0$ denotes the global effect parameter vector and $\boldsymbol{\beta}_j$ denotes the site-specific effect nuisance parameter vector.

If we pool all patient-level data together, the combined loss function is given by

$$L(\boldsymbol{\beta}) = L_{pooled}(\boldsymbol{\beta}) := \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j), \boldsymbol{\beta}^T = [\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_K^T], \tag{1}$$

where unknown central server parameter $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_0}$ and site-specific nuisance parameters $\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}, j = 1, \cdots, K$.

Let $\mathcal{S} = \{(k, j) : 1 \leq k \leq p_j, 0 \leq j \leq K\}$ denote the index set of parameter vector $\boldsymbol{\beta}$. Define the true parameters as $\boldsymbol{\beta}_j^0 = (\beta_{1j}^0, \beta_{2j}^0, \ldots, \beta_{p_j j}^0)^T$ for $j = 0, 1, 2, \ldots, K$. Let $A^0 = \{(k, j) \in \mathcal{S} : \beta_{kj}^0 \neq 0\}$ represent the support of the true parameter vector $\boldsymbol{\beta}^0$.

3. Constrained Optimization Approach

To address the challenge of heterogeneous data in high-dimensional settings, we propose a constrained optimization approach using the ℓ_0 penalty. We aim to reconstruct the oracle estimator–the least squares estimator $\widehat{\boldsymbol{\beta}}^{ol} = \left(\widehat{\boldsymbol{\beta}}_{A^0}^{ol}, \mathbf{0}\right)^T$ supported on A^0.

The following optimization problems have been described in Shen et al. (2013).

Constrained ℓ_0-method
Consider the constrained least squares regression

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & S(\boldsymbol{\beta}) = \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) \\ \text{subj to:} \quad & \sum_{(k,j)\in\mathcal{S}} I\left(\beta_{kj} \neq 0\right) \leq \kappa, \end{aligned} \tag{2}$$

where $\kappa > 0$ is an integer-valued tuning parameter. Denote the global minimizer of (2) as $\widehat{\boldsymbol{\beta}}^{\ell_0} = \left(\widehat{\boldsymbol{\beta}}_{\widehat{A}^{\ell_0}}^{\ell_0}, \mathbf{0}\right)$. Theorem 2 in Shen et al. (2013) demonstrates that the global minimizer consistently reconstructs the oracle estimator at a degree of separation level slightly higher than the minimum required.

Inspired by the works of Shen et al. (2013), Shi et al. (2019), and Zhu et al. (2020), we employ a constrained minimization algorithm via DC programming and ℓ_0 projection to address the ℓ_0 optimization problem as formulated in (2).

3.1 Algorithm

Set the tuning parameters $(\lambda, \tau, \kappa) \in \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{N} \cup \{0\}$. At $(t+1)$th iteration, we solve a weighted Lasso problem,

$$\widetilde{\mathbf{\Gamma}}^{[t+1]} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}; \widetilde{\mathbf{\Gamma}}^{[t]}), \tag{3}$$

where

$$S(\boldsymbol{\beta}; \boldsymbol{\beta}^{[t]}) = \frac{1}{n} \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) + \lambda\tau \sum_{(k,j)\in\mathcal{S}} I\left(\left|\beta_{kj}^{[t]}\right| \le \tau\right) |\beta_{kj}|,$$

$\lambda > 0$ is a tuning parameter and $\widetilde{\mathbf{\Gamma}}^{[t]}$ is the solution of (3) at the tth iteration. The DC algorithm terminates at $\widetilde{\mathbf{\Gamma}} = \widetilde{\mathbf{\Gamma}}^{[t]}$ such that $S(\widetilde{\mathbf{\Gamma}}^{[t]}; \widetilde{\mathbf{\Gamma}}^{[t]}) \le S(\widetilde{\mathbf{\Gamma}}^{[t+1]}; \widetilde{\mathbf{\Gamma}}^{[t]})$. Then, we obtain the solution $\widehat{\mathbf{\Gamma}}$ of (2) by projection $\widetilde{\mathbf{\Gamma}}$ onto the ℓ_0-constrained set $\{\|\mathbf{\Gamma}\|_0 \le \kappa\}$, where $\|\mathbf{\Gamma}\|_0 = \sum_{(k,j)\in\mathcal{S}} I(\Gamma_{kj} \neq 0)$. We summarize the general constrained minimization via DC programming and ℓ_0 projection algorithm in Algorithm 1.

Algorithm 1 Constrained minimization via DC programming & ℓ_0 projection

1: Specify $\lambda > 0$, $\tau > 0$, and $\kappa \ge 1$. Set $t = 0$. Initialize $\widetilde{\mathbf{\Gamma}}^{[0]} = \left\{\widetilde{\Gamma}_{kj}^{[0]}\right\}_{(k,j)\in\mathcal{S}}$.

2: Use a weighted Lasso solver to solve (3).

3: If $S(\widetilde{\mathbf{\Gamma}}^{[t]}; \widetilde{\mathbf{\Gamma}}^{[t]}) - S(\widetilde{\mathbf{\Gamma}}^{[t+1]}; \widetilde{\mathbf{\Gamma}}^{[t]})$ has not converged, set $t \leftarrow t+1$ and return to line 2.

4: (ℓ_0-projection) Let

$$C = \left\{ (k', j') \in \mathcal{S} : \sum_{(k,j)\in\mathcal{S}} I\left(|\widetilde{\Gamma}_{kj}^{[t]}| \ge |\widetilde{\Gamma}_{k'j'}^{[t]}|\right) \le \kappa \right\}.$$

Without loss of generality (WLOG), assume $|C| = \kappa$. Otherwise, if $|C| < \kappa$, then select $\kappa - |C|$ more elements from $\arg\max_{(k,j)\in\mathcal{S}\setminus C} \left|\widetilde{\Gamma}_{kj}^{[t]}\right|$.

5: Compute the ℓ_0 projection estimator $\widehat{\mathbf{\Gamma}}$:

$$\widehat{\mathbf{\Gamma}} = \arg\min_{\boldsymbol{\beta}} \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) \text{ s.t. } \beta_{kj} = 0 \text{ for } (k,j) \in \mathcal{S}\setminus C. \tag{4}$$

For the weighted Lasso problem (3) in step 2 of Algorithm 1, we can consider a first-order iterative algorithm, such as ISTA Daubechies et al. (2004) and FISTA Beck and Teboulle (2009). Denote the first order iterative solver with weights $\boldsymbol{w} = \{w_{k,j}; (k,j) \in \mathcal{S}\}$,

$$\widehat{\boldsymbol{\beta}}_{l+1} = \text{solver}\left(\widehat{\boldsymbol{\beta}}_l, \frac{\partial S(\widehat{\boldsymbol{\beta}}_l)}{\partial \boldsymbol{\beta}}; \boldsymbol{w}\right).$$

In multicenter research, individual-level data are often protected and cannot be shared across sites. Therefore, it is essential that our weighted Lasso solver is designed to operate under these constraints. Specifically, the central server parameter $\boldsymbol{\beta}_0$ from the previous iteration and its partial derivative $\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_0}$ should be communicated to the central server.

Meanwhile, the site-specific nuisance parameters at the jth site, $\boldsymbol{\beta}_j$, from the previous iteration and their partial derivatives $\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j}$ should remain local to the jth site.

Define the central server weight and the site weights $\boldsymbol{w}^j = \{w_{k',j'} \in \mathcal{S}; j' = j\}, j = 0, 1, \cdots, K$. The central server solver and the site solvers are given by

$$\text{central server update}: \widehat{\boldsymbol{\beta}}_{l+1,0} = \text{solver}\left(\widehat{\boldsymbol{\beta}}_{l,0}, \sum_{j=1}^{K} \frac{\partial S_j(\widehat{\boldsymbol{\beta}}_{l,0}, \widehat{\boldsymbol{\beta}}_{l,j})}{\partial \boldsymbol{\beta}_0}; \boldsymbol{w}^0\right), \text{ and} \tag{5}$$

$$\text{site server update}: \widehat{\boldsymbol{\beta}}_{l+1,j} = \text{solver}\left(\widehat{\boldsymbol{\beta}}_{l,j}, \frac{\partial S_j(\widehat{\boldsymbol{\beta}}_{l,0}, \widehat{\boldsymbol{\beta}}_{l,j})}{\partial \boldsymbol{\beta}_j}; \boldsymbol{w}^j\right), 1 \le j \le K, \tag{6}$$

where for any $j = 1, \cdots, K$, $S_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) = L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j)$.

We summarize the constrained minimization algorithm, which employs DC programming and ℓ_0 projection, in the distributed algorithm setting, as presented in Algorithm 2.

Algorithm 2 Constrained Minimization in the distributed algorithm setting

1: Specify $\lambda > 0$, $\tau > 0$, and $\kappa \ge 1$. Set $t = 0$. Initialize $\widetilde{\boldsymbol{\Gamma}}^{[0]} = \left\{\widetilde{\Gamma}_{kj}^{[0]}\right\}_{(k,j)\in\mathcal{S}}$.

2: Let $\boldsymbol{w}^j = \left\{\lambda\tau \cdot I\left(\left|\widetilde{\Gamma}_{kj}^{[t]}\right| \le \tau\right)\right\}_{k;(k,j)\in\mathcal{S}}$, $j = 0, \cdots, K$. Set $l = 0$. Initialize $\widehat{\boldsymbol{\beta}}_{0,j}, 0 \le j \le K$.

3: **for** Sites $j = 1$ to $j = K$ **do**

4: Update jth site-specific nuisance parameters according to (6).

5: Pass $\frac{\partial S_j(\widehat{\boldsymbol{\beta}}_{l,0}, \widehat{\boldsymbol{\beta}}_{l,j})}{\partial \boldsymbol{\beta}_0}$ to central server.

6: Update central server parameter according to (5).

7: **end for**

8: If convergence has not been achieved at all sites and the central server, based on $\left\|\frac{\partial S_j(\widehat{\boldsymbol{\beta}}_{l,0}, \widehat{\boldsymbol{\beta}}_{l,j})}{\partial \boldsymbol{\beta}_j}\right\|_2, 1 \le j \le K$ and $\left\|\sum_{j=1}^{K} \frac{\partial S_j(\widehat{\boldsymbol{\beta}}_{l,0}, \widehat{\boldsymbol{\beta}}_{l,j})}{\partial \boldsymbol{\beta}_0}\right\|_2$, set $l \leftarrow l + 1$ and return to line 3.

9: Set $\widetilde{\boldsymbol{\Gamma}}^{[t+1]} = \widehat{\boldsymbol{\beta}}_l$.

10: If $S(\widetilde{\boldsymbol{\Gamma}}^{[t]}; \widetilde{\boldsymbol{\Gamma}}^{[t]}) - S(\widetilde{\boldsymbol{\Gamma}}^{[t+1]}; \widetilde{\boldsymbol{\Gamma}}^{[t]})$ has not converged, set $t \leftarrow t + 1$ and return to line 2.

11: (ℓ_0-projection) Let

$$C = \left\{(k', j') \in \mathcal{S} : \sum_{(k,j)\in\mathcal{S}} I(|\widetilde{\Gamma}_{kj}^{[t]}| \ge |\widetilde{\Gamma}_{k'j'}^{[t]}|) \le \kappa\right\}.$$

WLOG, assume $|C| = \kappa$. Otherwise, if $|C| < \kappa$, then select $\kappa - |C|$ more elements from $\arg\max_{(k,j)\in\mathcal{S}\setminus C} \left|\widetilde{\Gamma}_{kj}^{[t]}\right|$.

12: Compute the ℓ_0 projection estimator $\widehat{\boldsymbol{\Gamma}}$:

$$\widehat{\boldsymbol{\Gamma}} = \arg\min_{\boldsymbol{\beta}} \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) \text{ s.t. } \beta_{kj} = 0 \text{ for } (k, j) \in \mathcal{S}\setminus C. \tag{7}$$

Remark 1. The initial $\widetilde{\mathbf{\Gamma}}^{[0]}$ needs to be sparse, such as $\mathbf{0}$ or a sparse estimator obtained through penalized methods.

4. Theoretical Results

In this section, we establish the theoretical properties of our proposed estimator, including its sampling distribution under various conditions.

For a hypothesized parameter subset $B \subset \mathcal{S}$, we consider the hypothesis testing

$$H_0 : \boldsymbol{\beta}_B = 0 \text{ versus } H_1 : \boldsymbol{\beta}_B \neq 0, \tag{8}$$

where $\boldsymbol{\beta}_B = 0$ if and only if $\beta_{kj} = 0$ for all $(k, j) \in B$.

4.1 Degree of separation

Consider the following measure of the level of difficulty for feature selection (see Zhu et al. (2020)):

$$C_{\min} = C_{\min}(\boldsymbol{\beta}^0, \mathbf{X}) \equiv \min_{A:|A|\leq|A^0| \text{ and } A\neq A^0} \inf_{\boldsymbol{\beta}} \frac{\left\|\mathbf{X}\boldsymbol{\beta}^0 - \mathbf{X}_{A\cup B}\boldsymbol{\beta}_{A\cup B}\right\|_2^2}{n\,|A^0\backslash A|} \geq d_1\sigma^2\frac{\log p + \log n}{n}, \tag{9}$$

where $\mathbf{X}_A$ and $\boldsymbol{\beta}_A$ are the design matrix for subset A of predictors and the regression coefficient vector over A, $n = \sum_{j=0}^{K} n_j$, and B is the index set of the hypothesized parameters.

4.2 Constrained likelihood ratio testing

The problem of constructing a constrained likelihood ratio with a sparsity constraint on nuisance parameters has been discussed in Zhu et al. (2020) and Shi et al. (2019). In this section, we illustrate our approach using a simple heterogeneous linear regression setting as an example. In Appendix A, we derived a heterogeneous linear regression model, which can be summarized as

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon,$$

with the log-likelihood

$$\mathcal{L}_n(\boldsymbol{\beta}, \sigma) = -\frac{1}{2\sigma^2}\sum_{j=1}^{K}\|\boldsymbol{Y}_j - \mathbf{X}_j\boldsymbol{\beta}_0 - \mathbf{W}_j\boldsymbol{\beta}_j\|_2^2 - \frac{n}{2}\log(2\pi\sigma^2),$$

where $\|\cdot\|_2$ denotes the Euclidean norm and the relationship between $\mathbf{X}$ and $\{\mathbf{X}_j, \mathbf{W}_j\}_{j=1}^K$ is given by (30). The constrained log-likelihood ratio, corresponding to the test (8), is defined as

$$\Lambda_n(B) := 2\left(\mathcal{L}_n(\widehat{\boldsymbol{\beta}}^1, \widehat{\sigma}^1) - \mathcal{L}_n(\widehat{\boldsymbol{\beta}}^0, \widehat{\sigma}^0)\right),$$

where $(\widehat{\boldsymbol{\beta}}^0, \widehat{\sigma}^0)$ and $(\widehat{\boldsymbol{\beta}}^1, \widehat{\sigma}^1)$ are the constrained maximum likelihood estimators (CMLE) based on the null and full spaces of the hypothesis test, respectively, that is,

$$\widehat{\boldsymbol{\beta}}^0 = \arg\min_{\|\boldsymbol{\beta}\|_0\leq\kappa, \boldsymbol{\beta}_B=\mathbf{0}} \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j),$$

and

$$\widehat{\boldsymbol{\beta}}^1 = \arg \min_{\|\boldsymbol{\beta}\|_{0,B} \leq \kappa} \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j),$$

where $\|\boldsymbol{\beta}\|_{0,B} = \sum_{(k,j)\in\mathcal{S}} I(\beta_{kj} \neq 0) I((k,j) \notin B)$ and $L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) = \|\boldsymbol{Y}_j - \mathbf{X}_j\boldsymbol{\beta}_0 - \mathbf{W}_j\boldsymbol{\beta}_j\|_2^2$.

Theorem 2 in Zhu et al. (2020) establishes the sampling distribution of $\Lambda_n(B)$. For completeness, we adapt Theorem 2 in Zhu et al. (2020) to our specific setting as follows (without proof).

Theorem 4.1. *Assume* $\frac{\sqrt{|B|}(|A^0|+|B|)}{n} \to 0$. *Under the degree of separation condition, there exist optimal tuning parameters* (κ, τ) *with* $\kappa = |A^0|$ *and* $0 < \tau \leq \sigma\sqrt{\frac{6}{(n+2)p\cdot\lambda_{\max}(\mathbf{X}^\top\mathbf{X})}}$ *such that under* H_0

1. *Wilks' phenomenon: If* $\beta_{kj} = 0$ *for* $(k,j) \in B$ *with* $|B|$ *fixed, then*

$$\Lambda_n(B) \xrightarrow{d} \chi^2_{|B|} \text{ as } n \to \infty. \tag{10}$$

2. *Generalized Wilks' phenomenon: If* $\beta_{kj} = 0$ *for* $(k,j) \in B$ *with* $|B| \to \infty$, *then*

$$(2|B|)^{-1/2}\left(\Lambda_n(B) - |B|\right) \xrightarrow{d} N(0,1) \text{ as } n \to \infty. \tag{11}$$

In the context of linear regression, a straightforward asymptotic result can be derived for CMLE. This is motivated by Proposition 2 from Zhu et al. (2020).

Theorem 4.2. *Under Assumptions 1-4, assume that for fixed* B, *the Moore–Penrose inverse* $\sigma^2\left(\frac{1}{n}\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B}\right)^\dagger_{B,B}$ *converges in distribution to some positive semi-definite matrix* Σ *and* $\{\boldsymbol{\xi} \in \mathbb{R}^{A^0\cup B}; \boldsymbol{\xi}_i = 0, i \notin B\} \subset \mathcal{R}(\mathbf{X}^T_{A^0\cup B})$, *where* $\mathcal{R}(\mathbf{X}^T_{A^0\cup B})$ *denotes the columns space of* $\mathbf{X}^T_{A^0\cup B}$. *If the tuning parameters* (κ, τ) *satisfy* $\kappa = |A^0|$ *and* $\tau \leq \sigma\sqrt{\frac{6}{(n+2)p\cdot\lambda_{max}(X^TX)}}$, *then*

$$\sqrt{n}(\widehat{\boldsymbol{\Gamma}}^{(1)}_B - \boldsymbol{\beta}^0_B) \xrightarrow{d} N(0, \Sigma). \tag{12}$$

Here, Assumptions 1-4 will be introduced in Section 5.3, under a bijection between the index sets $\mathcal{S}$ and $[p] = \{1, 2, \cdots, p\}$. The local power analysis, exemplified by Theorem 4 in Zhu et al. (2020), is applicable. The detailed statement is omitted for brevity.

5. Convergence and Consistency Results

5.1 Problem Setup and Notations

Before presenting our main theoretical results, we first introduce the linear model setup and necessary notation. Assume the training data are given by $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n, \boldsymbol{\beta} \in \mathbb{R}^p$ and $y_1, \cdots, y_n \in \mathbb{R}$. Furthermore, assume that y_i, given $\boldsymbol{x}_i$, has density $f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})$. For $B \subset [p]$, consider hypothesis testing

$$H_0 : \boldsymbol{\beta}_B = 0 \text{ versus } H_1 : \boldsymbol{\beta}_B \neq 0. \tag{13}$$

5.2 Computational Algorithm

Consider the constrained optimization problem regression for H_0 in (13)

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & S(\boldsymbol{\beta}) = \sum_{i=1}^{n} -\log(f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})) \\ \text{subj to:} \quad & \sum_{i\in[p]\backslash B} I\left(\beta_i \neq 0\right) \leq \kappa, \beta_B = 0 \end{aligned} \tag{14}$$

and for H_1 in (13)

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & S(\boldsymbol{\beta}) = \sum_{i=1}^{n} -\log(f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})) \\ \text{subj to:} \quad & \sum_{i\in[p]\backslash B} I\left(\beta_i \neq 0\right) \leq \kappa, \end{aligned} \tag{15}$$

where $\kappa > 0$ is an integer-valued tuning parameter. Set $L(\boldsymbol{\beta}) = \sum_{i=1}^{n} -\log(f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}))$.

The oracle estimators corresponding to H_0 and H_1 are given by

$$\widehat{\boldsymbol{\beta}}^{ol}_{H_0} = \arg \min_{\boldsymbol{\beta}:\boldsymbol{\beta}_{(A^0_{H_0})^c}=\mathbf{0}} L(\boldsymbol{\beta}) \text{ with } A^0_{H_0} = \{i \in [p]\backslash B; \beta^0_i \neq 0\} \text{ and } \kappa^0_{H_0} = |A^0_{H_0}|, \tag{16}$$

and

$$\widehat{\boldsymbol{\beta}}^{ol}_{H_1} = \arg \min_{\boldsymbol{\beta}:\boldsymbol{\beta}_{(A^0_{H_1})^c}=\mathbf{0}} L(\boldsymbol{\beta}) \text{ with } A^0_{H_1} = \{i \in [p]\backslash B; \beta^0_i \neq 0\} \cup B. \tag{17}$$

For the given hypothesis set B, at $(t+1)$th iteration, we solve the following weighted Lasso problems, corresponding to H_0 and H_1 respectively:

$$\widetilde{\boldsymbol{\Gamma}}^{[t+1]} = \arg \min_{\boldsymbol{\beta}:\boldsymbol{\beta}_B=\mathbf{0}} S(\boldsymbol{\beta}; \widetilde{\Gamma}^{[t]}), \text{ and} \tag{18}$$

$$\widetilde{\boldsymbol{\Gamma}}^{[t+1]} = \arg \min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}; \widetilde{\Gamma}^{[t]}), \tag{19}$$

where $\lambda > 0$ is a tuning parameter,

$$S(\boldsymbol{\beta}; \widetilde{\Gamma}^{[t]}) = \frac{1}{n} L(\boldsymbol{\beta}) + \lambda\tau \sum_{i\in[p]\backslash B} I\left(\left|\widetilde{\Gamma}^{[t]}_i\right| \leq \tau\right) |\beta_i|,$$

and $\widetilde{\boldsymbol{\Gamma}}^{[t]}$ is the solution of (18) or (19), respectively, at the tth iteration. The DC algorithm terminates at $\widetilde{\boldsymbol{\Gamma}} = \widetilde{\boldsymbol{\Gamma}}^{[t]}$ such that $S(\widetilde{\boldsymbol{\Gamma}}^{[t]}; \widetilde{\boldsymbol{\Gamma}}^{[t]}) \leq S(\widetilde{\boldsymbol{\Gamma}}^{[t+1]}; \widetilde{\boldsymbol{\Gamma}}^{[t]})$ (or $\text{supp}\{\widetilde{\boldsymbol{\Gamma}}^{[t]}\}\backslash B = \text{supp}\{\widetilde{\boldsymbol{\Gamma}}^{[t+1]}\}\backslash B$). Then, we obtain the solution $\widehat{\boldsymbol{\Gamma}}$ of (14) or (15), respectively, by projecting $\widetilde{\boldsymbol{\Gamma}}_{B^c}$ onto the ℓ_0-constrained set $\{\|\boldsymbol{\Gamma}_{B^c}\|_0 \leq \kappa\}$.

Algorithm 3 Constrained minimization via DC programming & ℓ_0 projection

1: Specify $\lambda > 0$, $\tau > 0$, and $\kappa \geq 1$. Set $t = 0$. Initialize $\widetilde{\mathbf{\Gamma}}^{[0]} = \left\{\widetilde{\Gamma}_i^{[0]}\right\}_{i\in[p]}$.

2: Use a weighted Lasso solver to solve (18) for H_0 or (19) for H_1 .

3: If $S(\widetilde{\mathbf{\Gamma}}^{[t]}; \widetilde{\mathbf{\Gamma}}^{[t]}) - S(\widetilde{\mathbf{\Gamma}}^{[t+1]}; \widetilde{\mathbf{\Gamma}}^{[t]})$ has not converged, set $t \leftarrow t + 1$ and return to line 2.

4: (ℓ_0-projection) Let

$$C = \left\{ i' \in [p]\backslash B; \sum_{i\in[p]\backslash B} I(|\widetilde{\Gamma}_i^{[t]}| \geq |\widetilde{\Gamma}_{i'}^{[t]}|) \leq \kappa, i' \in [p]\backslash B \right\}.$$

WLOG, assume $|C| = \kappa$. Otherwise, if $|C| < \kappa$, then select $\kappa - |C|$ more elements from $\arg\max_{i\in[p]\backslash(B\cup C)} \left|\widetilde{\Gamma}_i^{[t]}\right|$ into C.

5: Compute the ℓ_0 projection estimators $\widehat{\mathbf{\Gamma}} = \widehat{\mathbf{\Gamma}}_{H_0}$ or $\widehat{\mathbf{\Gamma}} = \widehat{\mathbf{\Gamma}}_{H_1}$, respectively, according to:

$$H_0 : \widehat{\mathbf{\Gamma}}_{H_0} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) \text{ s.t. } \boldsymbol{\beta}_i = 0, \text{ for } i \in [p]\backslash C, \text{ or} \tag{20}$$

$$H_1 : \widehat{\mathbf{\Gamma}}_{H_1} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) \text{ s.t. } \boldsymbol{\beta}_i = 0, \text{ for } i \in [p]\backslash(B \cup C). \tag{21}$$

5.3 Assumptions

To derive the convergence and consistency results of Algorithm 3, we will focus exclusively on the least squares regression setting from this point forward in the section. We begin by considering the linear model:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon}, \tag{22}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\beta}^0 \in \mathbb{R}^p$, $\boldsymbol{Y} \in \mathbb{R}^n$, $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I_n)$, and σ^2 might depend on n. Consider $B \subset [p]$ such that

$$\frac{\sqrt{|B|}(|A^0| + |B|)}{n} \to 0. \tag{23}$$

Without loss of generality, we can set $S(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$.

Let $\kappa \geq |A^0\backslash B|$, and $\kappa_{max} = \max\left\{\kappa, |\{i \in [p]\backslash B; |\widetilde{\Gamma}_i^{[0]}| \geq \tau\}|\right\}$, where $\widetilde{\mathbf{\Gamma}}^{[0]}$ denotes the initial estimator used in Algorithm 3. Without loss of generality, we can assume that $\widetilde{\mathbf{\Gamma}}_B^{[0]} = \mathbf{0}$.

To derive the statistical and computational properties of Algorithm 3 in least squares regression setting, we introduce the following technical assumptions which generalized the convergence and consistency of structure learning assumptions from Li et al. (2023).

Assumption 1 (Restricted eigenvalues). For a constant $c_1 > 0$,

$$\min_{A:|A\backslash B|\leq 2\kappa_{max}} \min_{\boldsymbol{\xi}:\|\boldsymbol{\xi}_{A^c}\|_1 \leq 3\|\boldsymbol{\xi}_A\|_1} \frac{\|\mathbf{X}\boldsymbol{\xi}\|_2^2}{n\|\boldsymbol{\xi}\|_2^2} \geq c_1, \tag{24}$$

where $\boldsymbol{\xi}_A \in \mathbb{R}^{|A|}$ is the projection of $\boldsymbol{\xi} \in \mathbb{R}^p$ onto coordinates in A.

Assumption 2. For constants $c_2, c_3 > 0$,

$$\begin{aligned} &\max_{1\le i\le p} \frac{1}{n}(\mathbf{X}^T(I-P_A)\mathbf{X})_{ii} \le c_2^2, \\ &\max_{1\le i\le p} n((\mathbf{X}_A^T\mathbf{X}_A)^\dagger)_{ii} \le c_3^2, \end{aligned} \tag{25}$$

where $P_A = \mathbf{X}_A(\mathbf{X}_A^T\mathbf{X}_A)^\dagger\mathbf{X}_A^T$, and $A \in \{A_{H_0}^0, A_{H_1}^0\}$.

Assumption 3 (Nuisance signals).

$$\min_{\beta_i^0\neq 0, i\notin B} \frac{|\beta_i^0|}{\sigma} \ge \frac{50c_3}{3}\sqrt{\frac{\log p}{n} + \frac{\log n}{n}}. \tag{26}$$

Assumption 4 (Degree of separation).

$$C_{\min} = C_{\min}(\boldsymbol{\beta}^0, \mathbf{X}) \equiv \min_{A:|A|\le|A^0| \text{ and } A\neq A^0} \inf_{\boldsymbol{\beta}} \frac{\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}_{A\cup B}\boldsymbol{\beta}_{A\cup B}\|_2^2}{n\,|A^0\backslash A|} \ge 72\sigma^2\frac{\log p + \log n}{n}. \tag{27}$$

Assumption 1 is a common condition related to restricted eigenvalues, as discussed in Bickel et al. (2009) and Wainwright (2019). Assumption 2 generalizes from the lower eigenvalue and mutual incoherence conditions found in Section 7.5.1 of Wainwright (2019). Assumption 3 specifies the minimal signal strength across the support, which is used to establish high-dimensional variable selection consistency, as seen in Fan et al. (2014) and Loh and Wainwright (2017). Finally, Assumption 4 is a commonly recognized condition for the degree of separation in feature selection, according to Shen et al. (2013) and Zhu et al. (2020).

5.4 Selection Consistency

The theory presented below extends the convergence and consistency results for structure learning, as found in Theorem 14 of Li et al. (2023), to include over-selection consistency.

Theorem 5.1. *Under Assumptions 1, 2, 3, and 4, if the tuning parameters* (κ, τ, λ) *of Algorithm 3 in least squares regression setting satisfy:*

1. $\sqrt{32\sigma^2c_3^2\left(\frac{\log p}{n} + \frac{\log n}{n}\right)} \le \tau \le \min_{i\in B^c, \beta_i^0\neq 0}|\beta_i^0|$,
2. $\kappa = |A_{H_0}^0|$,
3. $\frac{1}{\tau}\sqrt{32\sigma^2c_2^2\left(\frac{\log p}{n} + \frac{\log n}{n}\right)} \le \lambda \le c_1/6$,

then, under both H_0 *and* H_1, $\widehat{\mathbf{\Gamma}}$ *in Algorithm 3 yields the oracle estimators* (16) *and* (17) *as well as the global minimizer of* (14) *and* (15)*, respectively, in at most* $\lceil \log(2\kappa_{max})/\log 4\rceil$ *DC iterations almost surely, where* $\kappa_1 = \left|\{i \in [p]\backslash B; |\widetilde{\Gamma}_i^{[0]}| \ge \tau\}\right|$ *and* $\kappa_{max} = \max\{\kappa, \kappa_1\}$.

By replacing the second condition with $\kappa \ge |A_{H_0}^0|$*, under both* H_0 *and* H_1*, Algorithm 3 ensures that the supports of the oracle estimators* (16) *and* (17) *are subsets of* $\mathrm{supp}(\widehat{\mathbf{\Gamma}})\backslash B$ *and* $\mathrm{supp}(\widehat{\mathbf{\Gamma}}) \cup B$ *almost surely, respectively. Additionally, the DC algorithm almost surely converges in at most* $\lceil \log(2\kappa_{\max})/\log 4\rceil$ *iterations.*

The first part of Theorem 5.1 establishes the result of almost sure subset recovery, while the second part confirms the almost sure over-selection consistency for Algorithm 3.

Remark 2. Building on the foundation established by Algorithm 3 and Theorem 5.1, our constrained DC algorithm, incorporating the ℓ_0 projection, is capable of reaching a global minimum within polynomial time, with the probability approaching 1 given the data generation process. This outcome starkly contrasts with previous findings, such as those reported by Chen et al. (2017, 2019), which state that no algorithm can consistently solve such nonconvex minimization problems in polynomial time under worst-case conditions.

6. Summary

In this paper, we have proposed a novel approach for handling heterogeneous data in high-dimensional statistical inference and structure learning problems. The proposed framework utilizes a parametric likelihood setting and introduces a truncated lasso penalty (TLP) for variable selection and parameter estimation.

For hypothesis testing, we have developed a procedure that leaves the parameters of interest unregularized while imposing an ℓ_0-constraint on the nuisance parameters to control their sparsity. Under a degree of separation condition and suitable choices of the tuning parameters, we have established the asymptotic properties of the constrained likelihood ratio statistic.

In terms of parameter estimation, we have proposed a constrained optimization approach using DC programming and ℓ_0 projection. We have established the theoretical properties of the resulting estimator, including its over-selection consistency and support recovery when the tuning parameter for the ℓ_0-constraint equals the true sparsity level. Moreover, we have shown that the estimator attains the oracle property and global minimizer of the constrained optimization problem within a logarithmic number of iterations.

The proposed methodology offers several advantages in the context of distributed learning with heterogeneous data. By allowing for site-specific nuisance parameters, our approach can effectively account for the inherent heterogeneity across different data sources. The use of the truncated lasso penalty enables simultaneous variable selection and parameter estimation, leading to more interpretable models. Furthermore, the communication-efficient nature of the algorithm makes it well suited for distributed algorithm settings where data cannot be directly shared across sites.

References

Randolph C Barrows Jr and Paul D Clayton. Privacy, confidentiality, and electronic medical records. *Journal of the American medical informatics association*, 3(2):139–148, 1996.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *Annals of statistics*, 46(3):1352, 2018.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705 – 1732, 2009.

Yichen Chen, Dongdong Ge, Mengdi Wang, Zizhuo Wang, Yinyu Ye, and Hao Yin. Strong np-hardness for sparse optimization with concave penalty functions. In *International Conference on Machine Learning*, pages 740–747. PMLR, 2017.

Yichen Chen, Yinyu Ye, and Mengdi Wang. Approximation hardness for a class of sparse optimization problems. *Journal of Machine Learning Research*, 20(38):1–27, 2019.

Adam Cheng, David Kessler, Ralph Mackinnon, Todd P Chang, Vinay M Nadkarni, Elizabeth A Hunt, Jordan Duval-Arnould, Yiqun Lin, Martin Pusic, and Marc Auerbach. Conducting multicenter research in healthcare simulation: Lessons learned from the inspire network. *Advances in Simulation*, 2(1):1–14, 2017.

Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

Rui Duan, Yang Ning, and Yong Chen. Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika*, 109(1):67–83, 2022.

Jianqing Fan, Lingzhou Xue, and Hui Zou. Strong oracle optimality of folded concave penalized estimation. *Annals of statistics*, 42(3):819, 2014.

Zijian Guo, Xiudi Li, Larry Han, and Tianxi Cai. Robust inference for federated meta-learning. *arXiv preprint arXiv:2301.00718*, 2023.

Larry Han, Jue Hou, Kelly Cho, Rui Duan, and Tianxi Cai. Federated adaptive causal estimation (face) of target treatment effects. *arXiv preprint arXiv:2112.09313*, 2021.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 2018.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529. PMLR, 2020.

Jakub Konecnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 8, 2016.

Gue Myung Lee and Kwang Baik Lee. Vector variational inequalities for nondifferentiable convex vector optimization problems. *Journal of Global Optimization*, 32:597–612, 2005.

Chunlin Li, Xiaotong Shen, and Wei Pan. Inference for a large directed acyclic graph with unspecified interventions. *Journal of Machine Learning Research*, 24(73):1–48, 2023.

Po-Ling Loh and Martin J. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Xiaotong Shen, Wei Pan, Yunzhang Zhu, and Hui Zhou. On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65(5): 807–832, 2013.

Chengchun Shi, Rui Song, Zhao Chen, and Runze Li. Linear hypothesis testing for high dimensional generalized linear models. *Annals of statistics*, 47(5):2671, 2019.

Ellen Sidransky, Michael A Nalls, Jan O Aasly, Judith Aharon-Peretz, Grazia Annesi, Egberto R Barbosa, Anat Bar-Shira, Daniela Berg, Jose Bras, Alexis Brice, et al. Multicenter analysis of glucocerebrosidase mutations in parkinson's disease. *New England Journal of Medicine*, 361(17):1651–1661, 2009.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient distributed learning with sparsity. In *International conference on machine learning*, pages 3636–3645. PMLR, 2017.

Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive federated learning in resource constrained edge computing systems. *IEEE journal on selected areas in communications*, 37(6):1205–1221, 2019.

Yunzhang Zhu, Xiaotong Shen, and Wei Pan. On high-dimensional constrained maximum likelihood inference. *Journal of the American Statistical Association*, 115(529):217–230, 2020.

A. Design Matrix of Heterogeneous Linear Model

I will explore a simple heterogeneous linear regression setting below. Assume the ith observation at the jth site follows

$$Y_{ij} = \boldsymbol{\beta}_0^T \boldsymbol{x}_i^{(j)} + \boldsymbol{\beta}_j^T \boldsymbol{w}_i^{(j)} + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma_j^2); i = 1, \cdots, n_j, \tag{28}$$

where vectors $\boldsymbol{\beta}_0, \boldsymbol{x}_i^{(j)} \in \mathbb{R}^{p_0}$, $\boldsymbol{\beta}_j, \boldsymbol{w}_i^{(j)} \in \mathbb{R}^{p_j}, j = 1, 2, \cdots, K$, and $\{\boldsymbol{x}_i^{(j)}, \boldsymbol{w}_i^{(j)}\}_{1 \le j \le K, 1 \le i \le n_j}$ are independent of error $\{\varepsilon_{ij}\}_{1 \le j \le K, 1 \le i \le n_j}$. Let $p = p_0 + \sum_{j=1}^K p_j$. Assume that we already know some positive numbers $r_1, \cdots, r_K$, such that $\sigma_j^2 = r_j^2 \sigma^2, j = 1, 2 \cdots, K$. Assume $\sigma^2 = \max_{1 \le j \le K} \sigma_j^2$.

Replacing $Y_{ij}, \boldsymbol{x}_i^{(j)}, \boldsymbol{w}_i^{(j)}, \varepsilon_{ij}$ in (28) by $Y_{ij} r_j, \boldsymbol{x}_i^{(j)} r_j, \boldsymbol{w}_i^{(j)} r_j, \varepsilon_{ij} r_j$, respectively, we have

$$Y_{ij} = \boldsymbol{\beta}_0^T \boldsymbol{x}_i^{(j)} + \boldsymbol{\beta}_j^T \boldsymbol{w}_i^{(j)} + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2); i = 1, \cdots, n_j. \tag{29}$$

Rewrite linear model (29) as $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, that is

$$\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_K \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{W}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{0} & \mathbf{W}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_K & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{W}_K \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_K \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_K \end{pmatrix}, \tag{30}$$

where error terms $\boldsymbol{\varepsilon}_j \sim N_{n_j}(0, \sigma^2 I_{n_j})$, unknown central server parameter $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_0}$ and site-specific nuisance parameters $\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}, j = 1, \cdots, K$, response vectors $\boldsymbol{Y}_j = [Y_{1j}, Y_{2j}, \cdots, Y_{n_j j}]^T$, design matrices $\mathbf{X}_j^T = [\boldsymbol{x}_1^{(j)}, \boldsymbol{x}_2^{(j)}, \cdots, \boldsymbol{x}_{n_j}^{(j)}]$, and $\mathbf{W}_j^T = [\boldsymbol{w}_1^{(j)}, \boldsymbol{w}_2^{(j)}, \cdots, \boldsymbol{w}_{n_j}^{(j)}]$, $j = 1, 2, \cdots, K$.

The jth site log-likelihood function is given by

$$L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j, \sigma^2) = -\frac{1}{2\sigma^2} \left\| \boldsymbol{Y}_j - \mathbf{X}_j\boldsymbol{\beta}_0 - \mathbf{W}_j\boldsymbol{\beta}_j \right\|_2^2 - \frac{n_j}{2} \log(2\pi\sigma^2),$$

and $\|\cdot\|_2$ denotes the Euclidean norm.

B. Proof of Theorem 4.2

Proof of Theorem 4.2 By Theorem 5.1,

$$\lim_{n\to\infty} \mathbb{P}(\widehat{\boldsymbol{\Gamma}}^{(1)} = \widehat{\boldsymbol{\beta}}^{ls}) = 1, \tag{31}$$

where $\widehat{\boldsymbol{\Gamma}}^{(1)}$ is the CMLE over the full spaces of the test, and the support of $\widehat{\boldsymbol{\beta}}^{ls}$ is restricted on $A^0 \cup B$.

Thus, to show (12), it suffices to show

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_B^{ls} - \boldsymbol{\beta}_B^0) \xrightarrow{d} N(0, \Sigma), \tag{32}$$

as $n \to \infty$.

Let $t \in \mathbb{R}^B$ and $u \in \mathbb{R}^{A^0 \cup B}$ such that $u_B = t$ and $u_{B^c} = 0$. Note that

$$\widehat{\boldsymbol{\beta}}_{A^0\cup B}^{ls} = (\mathbf{X}_{A^0\cup B}^T \mathbf{X}_{A^0\cup B})^\dagger (\mathbf{X}_{A^0\cup B}^T \mathbf{X}_{A^0\cup B}) \boldsymbol{\beta}_{A^0\cup B}^0 + (\mathbf{X}_{A^0\cup B}^T \mathbf{X}_{A^0\cup B})^\dagger \mathbf{X}_{A^0\cup B}^T \boldsymbol{\varepsilon}, \tag{33}$$

where $\mathbf{X}_{A^0\cup B}$ denotes the sub-matrix with columns of $A^0 \cup B$.

Due to $u \in \{\boldsymbol{\xi} \in \mathbb{R}^{A^0\cup B}; \boldsymbol{\xi}_i = 0, i \notin B\} \subset \mathcal{R}(\mathbf{X}_{A^0\cup B}^T)$, we know that there exists vector v such that $u = \mathbf{X}_{A^0\cup B}^T v$ and

$$(\mathbf{X}_{A^0\cup B}^T \mathbf{X}_{A^0\cup B})(\mathbf{X}_{A^0\cup B}^T \mathbf{X}_{A^0\cup B})^\dagger u = \mathbf{X}_{A^0\cup B}^T \Big(\mathbf{X}_{A^0\cup B} (\mathbf{X}_{A^0\cup B}^T \mathbf{X}_{A^0\cup B})^\dagger \mathbf{X}_{A^0\cup B}^T \Big) v = \mathbf{X}_{A^0\cup B}^T v = u.$$

By direct calculation of the characteristic function of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_B^{ls} - \boldsymbol{\beta}_B^0)$, we know that

$$\begin{aligned} &\mathbb{E} e^{it^T \sqrt{n}(\widehat{\boldsymbol{\beta}}_B^{ls} - \boldsymbol{\beta}_B^0)} = \mathbb{E} e^{iu^T \sqrt{n}(\widehat{\boldsymbol{\beta}}_{A^0\cup B}^{ls} - \boldsymbol{\beta}_{A^0\cup B}^0)} \\ =&\mathbb{E}_{\mathbf{X}} \mathbb{E}_{\boldsymbol{\varepsilon}} e^{iu^T \sqrt{n}(\mathbf{X}_{A^0\cup B}^T \mathbf{X}_{A^0\cup B})^\dagger \mathbf{X}_{A^0\cup B}^T \boldsymbol{\varepsilon}} \\ =&\mathbb{E}_{\mathbf{X}} \exp\{-1/2 \cdot \sigma^2 u^T n (\mathbf{X}_{A^0\cup B}^T \mathbf{X}_{A^0\cup B})^\dagger \mathbf{X}_{A^0\cup B}^T \mathbf{X}_{A^0\cup B} (\mathbf{X}_{A^0\cup B}^T \mathbf{X}_{A^0\cup B})^\dagger u\} \\ =&\mathbb{E}_{\mathbf{X}} \exp\{-1/2 \cdot \sigma^2 u^T (1/n \cdot \mathbf{X}_{A^0\cup B}^T \mathbf{X}_{A^0\cup B})^\dagger u\} \\ =&\mathbb{E}_{\mathbf{X}} \exp\{-1/2 \cdot \sigma^2 t^T (1/n \cdot \mathbf{X}_{A^0\cup B}^T \mathbf{X}_{A^0\cup B})_{B,B}^\dagger t\}. \end{aligned} \tag{34}$$

Under the assumption of Moore–Penrose inverse $\sigma^2 \left(\frac{1}{n}\mathbf{X}_{A^0\cup B}^T\mathbf{X}_{A^0\cup B}\right)_{B,B}^{\dagger}$ converges in distribution to some positive semi-definite matrix Σ, and by applying the continuous mapping theorem, we obtain that

$$\mathbb{E}_{\mathbf{X}} \exp\{-1/2 \cdot \sigma^2 t^T (1/n \cdot \mathbf{X}_{A^0\cup B}^T \mathbf{X}_{A^0\cup B})_{B,B}^{\dagger} t\} \to e^{-1/2 t^T \Sigma t}, \tag{35}$$

as $n \to \infty$, for any $t \in \mathbb{R}^B$.

By Lévy's continuity theorem, we complete the proof of weak convergence (32). Therefore, the proof of Theorem 4.2 is completed. □

C. Proof of Theorem 5.1

Proof of Theorem 5.1 We will show that if $\kappa \geq |A_{H_0}^0|$, then $\{i \in [p]\backslash B : \beta_i^0 \neq 0\} \subset \{i \in [p]\backslash B : \widehat{\Gamma}_i \neq 0\}$ almost surely, where the estimators $\widehat{\mathbf{\Gamma}}_{H_0}$ and $\widehat{\mathbf{\Gamma}}_{H_1}$ are obtained from Algorithm 3. For H_0, let $A^0 = \{i \in [p]\backslash B : \beta_i^0 \neq 0\}$ and $A^{[t]} = \{i \in [p]\backslash B : |\widetilde{\Gamma}_i^{[t]}| \geq \tau\}$. For H_1, let $A^0 = \{i \in [p]\backslash B : \beta_i^0 \neq 0\} \cup B$ and $A^{[t]} = \{i \in [p]\backslash B : |\widetilde{\Gamma}_i^{[t]}| \geq \tau\} \cup B$. Set $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{ol}$, where $\widehat{\boldsymbol{\beta}}^{ol}$ is the oracle estimate that minimizes $\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ under the constraint $\boldsymbol{\beta}_{(A^0)^c} = 0$. Set $E = \left\{\left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n\right\|_\infty \leq 0.5\lambda\tau\right\} \bigcap \left\{\left\|\boldsymbol{\beta}^0 - \widehat{\boldsymbol{\beta}}^{ol}\right\|_\infty \leq 0.5\tau\right\}$.

We will show that $A^0 \triangle A^{[t]}$ is eventually empty set on event E, which has a probability tending to 1, where $\triangle$ denotes the symmetric difference. By the optimality criterion Lee and Lee (2005) for (18) and (19), we have

$$\langle \widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\mathbf{\Gamma}}^{[t]}, -\mathbf{X}^T(\boldsymbol{Y} - \mathbf{X}\widetilde{\mathbf{\Gamma}}^{[t]})/n + \lambda\tau\nabla \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1 \rangle \geq 0. \tag{36}$$

Rearranging the terms, we have

$$\begin{aligned}
&\left\|\mathbf{X}(\widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\mathbf{\Gamma}}^{[t]})\right\|_2^2/n \\
\leq &\langle \widetilde{\mathbf{\Gamma}}^{[t]} - \widehat{\boldsymbol{\beta}}^{ol}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1 \rangle \\
= &\langle \widetilde{\mathbf{\Gamma}}^{[t]}_{A^0\backslash A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0\backslash A^{[t-1]}}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1 \rangle \\
&+ \langle \widetilde{\mathbf{\Gamma}}_{A^{[t-1]}\backslash A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]}\backslash A^0}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1 \rangle \\
&+ \langle \widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1 \rangle \\
&+ \langle \widetilde{\mathbf{\Gamma}}^{[t]}_{A^{[t-1]}\cap A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]}\cap A^0}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1 \rangle.
\end{aligned} \tag{37}$$

Let $\mathbf{X}_{A^0}$ denote the matrix, whose A^0 columns are the same as the A^0 columns of $\mathbf{X}$ and all other columns are zeros. Similarly, we can define $\mathbf{X}_{(A^0)^c}$. Notice that $\mathbf{X} = \mathbf{X}_{A^0} + \mathbf{X}_{(A^0)^c}$, $\widehat{\boldsymbol{\beta}}^{ol}_{(A^0)^c} = \mathbf{0}$, $\widehat{\boldsymbol{\beta}}^{ol} = (\mathbf{X}_{A^0}^T\mathbf{X}_{A^0})^{\dagger}\mathbf{X}_{A^0}^T\boldsymbol{Y}$, $\widehat{\boldsymbol{Y}} = \mathbf{X}_{A^0}(\mathbf{X}_{A^0}^T\mathbf{X}_{A^0})^{\dagger}\mathbf{X}_{A^0}^T\boldsymbol{Y}$, $P_{A^0} = \mathbf{X}_{A^0}(\mathbf{X}_{A^0}^T\mathbf{X}_{A^0})^{\dagger}\mathbf{X}_{A^0}^T$, as well as $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (I - P_{A^0})\boldsymbol{Y} = (I - P_{A^0})\boldsymbol{\varepsilon}$. Note that $\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}} = \mathbf{X}_{(A^0)^c}^T\widehat{\boldsymbol{\varepsilon}}$.

Thus, we know that the fourth term in the last equation of (37) vanishes, since

$$\begin{aligned}
&\langle \widetilde{\mathbf{\Gamma}}^{[t]}_{A^{[t-1]}\cap A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]}\cap A^0}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1\rangle \\
=&\langle \widetilde{\mathbf{\Gamma}}^{[t]}_{A^{[t-1]}\cap A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]}\cap A^0}, \mathbf{X}^T_{(A^0)^c}\widehat{\boldsymbol{\varepsilon}}/n\rangle \\
=&\langle \mathbf{X}_{(A^0)^c}(\widetilde{\mathbf{\Gamma}}^{[t]}_{A^{[t-1]}\cap A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]}\cap A^0}), \widehat{\boldsymbol{\varepsilon}}/n\rangle = 0.
\end{aligned} \tag{38}$$

Note that the third term in the last equation of (37) satisfies

$$\langle \widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}, \nabla \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1\rangle = \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}\right\|_1, \tag{39}$$

because $\widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c} = 0$.

Recalling (37), we have

$$\begin{aligned}
0 \le& \left\|\mathbf{X}(\widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\mathbf{\Gamma}}^{[t]})\right\|_2^2/n \\
\le& \langle \widetilde{\mathbf{\Gamma}}^{[t]}_{A^0\setminus A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0\setminus A^{[t-1]}}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{A^0\setminus A^{[t-1]}}\right\|_1\rangle \\
&+ \langle \widetilde{\mathbf{\Gamma}}_{A^{[t-1]}\setminus A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]}\setminus A^0}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n\rangle \\
&+ \langle \widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c}\right\|_1\rangle \\
\le& \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{A^0\triangle A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0\triangle A^{[t-1]}}\right\|_1 \cdot (\left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}\right\|_\infty/n + \lambda\tau) \\
&+ \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}\right\|_1 \cdot (\left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}\right\|_\infty/n - \lambda\tau).
\end{aligned} \tag{40}$$

Rearranging inequality (40) implies that

$$\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}\right\|_1 \cdot (\lambda\tau - \left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}\right\|_\infty/n) \le \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{A^0\triangle A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0\triangle A^{[t-1]}}\right\|_1 \cdot (\left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}\right\|_\infty/n + \lambda\tau). \tag{41}$$

Over event E, we have

$$\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}\right\|_1 \le 3\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{A^0\triangle A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0\triangle A^{[t-1]}}\right\|_1 \le 3\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{A^0\cup A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0\cup A^{[t-1]}}\right\|_1. \tag{42}$$

Recall that $\kappa_1 = \left|\{i \in [p]\setminus B; |\widetilde{\Gamma}^{[0]}_i| \ge \tau\}\right|$ and $\kappa_{max} = \max\{\kappa, \kappa_1\}$. Without loss of generality, we can assume that $\widetilde{\mathbf{\Gamma}}^{[0]}_B = \mathbf{0}$. For the base case, we know that

$$\left|A^0\triangle A^{[0]}\right| \le |A^0\setminus B| + |A^{[0]}\setminus A^0| \le 2\kappa_{max}.$$

For the induction step, assume $\left|A^0\triangle A^{[t-1]}\right| \le 2\kappa_{max}$ on event E, for $t \ge 1$. We aim to show that $\left|A^0\triangle A^{[t]}\right| \le 2\kappa_{max}$ over event E.

Applying Assumption 1 and (40), over event E, we have

$$\begin{aligned} & c_1 \left\| \widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\boldsymbol{\Gamma}}^{[t]} \right\|_2^2 \leq \left\| \mathbf{X}(\widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\boldsymbol{\Gamma}}^{[t]}) \right\|_2^2 / n \\ \leq & \left\| \widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^0 \triangle A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0 \triangle A^{[t-1]}} \right\|_1 \cdot \left(\left\| \mathbf{X}^T \widehat{\boldsymbol{\varepsilon}} \right\|_\infty / n + \lambda\tau \right) \\ & + \left\| \widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]} \cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]} \cup A^0)^c} \right\|_1 \cdot \left(\left\| \mathbf{X}^T \widehat{\boldsymbol{\varepsilon}} \right\|_\infty / n - \lambda\tau \right) \\ \leq & \frac{3}{2} \lambda\tau \left\| \widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^0 \triangle A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0 \triangle A^{[t-1]}} \right\|_1 . \end{aligned} \tag{43}$$

By Cauchy-Schwarz inequality,

$$\left\| \widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^0 \triangle A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0 \triangle A^{[t-1]}} \right\|_1^2 \leq \left| A^0 \triangle A^{[t-1]} \right| \left\| \widetilde{\boldsymbol{\Gamma}}^{[t]} - \widehat{\boldsymbol{\beta}}^{ol} \right\|_2^2 . \tag{44}$$

Combining (43), (44) and the third condition in Theorem 5.1,

$$\left\| \widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\boldsymbol{\Gamma}}^{[t]} \right\|_2 \leq \frac{3\lambda\tau}{2c_1} \sqrt{\left| A^0 \triangle A^{[t-1]} \right|} \leq \frac{\tau}{4} \sqrt{\left| A^0 \triangle A^{[t-1]} \right|}. \tag{45}$$

Applying (45) and induction assumption $\left| A^0 \triangle A^{[t-1]} \right| \leq 2\kappa_{max}$, we have

$$\left\| \widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\boldsymbol{\Gamma}}^{[t]} \right\|_2 / \tau \leq \frac{1}{4} \sqrt{2\kappa_{max}} \leq \frac{1}{2} \sqrt{\kappa_{max}} \leq \sqrt{\kappa_{max}}. \tag{46}$$

Because for any $i \in A^{[t]} \backslash A^0$, $|\widetilde{\Gamma}_i^{[t]} - \widehat{\beta}_i^{ol}| = |\widetilde{\Gamma}_i^{[t]}| \geq \tau$, we have

$$\sqrt{|A^{[t]} \backslash A^0|} \leq \left\| \widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\boldsymbol{\Gamma}}^{[t]} \right\|_2 / \tau \leq \sqrt{\kappa_{\max}}.$$

Thus,

$$\left| A^0 \triangle A^{[t]} \right| \leq |A^0 \backslash B| + |A^{[t]} \backslash A^0| \leq 2\kappa_{max}.$$

By induction, we already showed that over event E, $\left| A^0 \triangle A^{[t]} \right| \leq 2\kappa_{max}$ for any $0 \leq t \leq t_{max}$, where t_{max} denotes the total number of the iteration of Algorithm 3.

Note that Algorithm 3 is terminated at t if $\text{supp}\{\widetilde{\boldsymbol{\Gamma}}^{[t]}\} \backslash B = \text{supp}\{\widetilde{\boldsymbol{\Gamma}}^{[t-1]}\} \backslash B$.

We will show that over event E, for any $t \geq 1$, $\sqrt{\left| A^0 \triangle A^{[t]} \right|} \leq \frac{1}{2} \sqrt{\left| A^0 \triangle A^{[t-1]} \right|}$.

If $i \in A^0 \backslash A^{[t]}$, then we know that $\beta_i^0 \neq 0$ and $\widetilde{\Gamma}_i^{[t]} = 0$. Over event E, $|\beta_i^0 - \widehat{\beta}_i^{ol}| \leq 0.5\tau$. According to assumption $\kappa \geq |A^0_{H_0}|$ in Theorem 5.1, we know that

$$|\widetilde{\Gamma}_i^{[t]} - \widehat{\beta}_i^{ol}| \geq |\widetilde{\Gamma}_i^{[t]} - \beta_i^0| - |\beta_i^0 - \widehat{\beta}_i^{ol}| \geq \tau - 0.5\tau = 0.5\tau. \tag{47}$$

If $i \in A^{[t]} \backslash A^0$, then we know that $\beta_i^0 = \widehat{\beta}_i^{ol} = 0$ and $|\widetilde{\Gamma}_i^{[t]}| \geq \tau$, which implies $|\widetilde{\Gamma}_i^{[t]} - \widehat{\beta}_i^{ol}| = |\widetilde{\Gamma}_i^{[t]}| \geq \tau$.

Over event E, we obtain that for any $i \in A^0 \triangle A^{[t]}$, $|\widetilde{\Gamma}_i^{[t]} - \widehat{\beta}_i^{ol}| \geq \tau - 0.5\tau = 0.5\tau$. Combining (45), we obtain

$$\sqrt{\left| A^0 \triangle A^{[t]} \right|} \leq \frac{1}{0.5\tau} \left\| \widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\boldsymbol{\Gamma}}^{[t]} \right\|_2 \leq \frac{1}{2} \sqrt{\left| A^0 \triangle A^{[t-1]} \right|}. \tag{48}$$

In conclusion, over event E, we obtain

$$\sqrt{\left|A^0 \triangle A^{[t]}\right|} \leq \frac{1}{2^t}\sqrt{2\kappa_{max}}. \tag{49}$$

If $t \geq \lceil \frac{\log(2\kappa_{max})}{\log 4} \rceil$, $\sqrt{\left|A^0 \triangle A^{[t]}\right|} < 1$, i.e., $A^0 \triangle A^{[t]} = \emptyset$.

Set $t_{max} = \lceil \frac{\log(2\kappa_{max})}{\log 4} \rceil$. We already showed that over event E,

$$\{i \in [p]\backslash B : |\widetilde{\Gamma}_i^{[t_{max}]}| \geq \tau\} = \{i \in [p]\backslash B : \beta_i^0 \neq 0\}.$$

This means that for any $j \notin \{i \in [p]\backslash B : |\widetilde{\Gamma}_i^{[t_{max}]}| \geq \tau\}$, $|\widetilde{\Gamma}_j^{[t_{max}]}| < \tau$, and for any $k \in \{i \in [p]\backslash B : \widetilde{\Gamma}_i^{[t_{max}]} \geq \tau\}$, $|\widetilde{\Gamma}_k^{[t_{max}]}| \geq \tau$. Thus, over the event E

$$\text{supp}\{\boldsymbol{\beta}^0\}\backslash B = \{i \in [p]\backslash B : \beta_i^0 \neq 0\} = \{i \in [p]\backslash B : \left|\widetilde{\Gamma}_i^{[t_{max}]}\right| \geq \tau\} \subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B. \tag{50}$$

Next, we aim to bound the probability of event E. Let $\boldsymbol{a} := \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}} = (a_1, \cdots, a_p)^T$. It is obvious that $\boldsymbol{a} = \mathbf{X}^T(I - P_{A^0})\boldsymbol{\varepsilon} = \mathbf{X}^T_{(A^0)^c}(I - P_{A^0})\boldsymbol{\varepsilon}$. Recall the error $\boldsymbol{\varepsilon} \sim N(0, \Sigma)$ with $\Sigma = \sigma^2 I$. By Assumption 2, for any $1 \leq l \leq p$,

$$\text{Var}(a_l) = \boldsymbol{e}_l^T\mathbf{X}^T(I - P_{A^0})\Sigma(I - P_{A^0})\mathbf{X}\boldsymbol{e}_l \leq \sigma^2(\mathbf{X}^T(I - P_{A^0})\mathbf{X})_{ll} \leq n\sigma^2 c_2^2. \tag{51}$$

By the upper-tail inequality for sub-gaussian distribution and the third condition in Theorem 5.1, we have

$$\begin{aligned} &\mathbb{P}(\left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}\right\|_\infty / n > 0.5\lambda\tau) = \mathbb{P}(\|\boldsymbol{a}\|_\infty / n > 0.5\lambda\tau) \\ \leq &\sum_{l=1}^p \mathbb{P}(|\boldsymbol{a}|_l > 0.5n\lambda\tau) \leq 2p\exp\left(-\frac{1}{8}\frac{n\lambda^2\tau^2}{\sigma^2 c_2^2}\right) \leq \frac{2}{p^3 n^4}. \end{aligned} \tag{52}$$

Set $\boldsymbol{b} := \widehat{\boldsymbol{\beta}}^{ol} - \boldsymbol{\beta}^0$. A direct calculation implies $\boldsymbol{b} = (\mathbf{X}^T_{A^0}\mathbf{X}_{A^0})^\dagger\mathbf{X}^T_{A^0}\boldsymbol{\varepsilon} = (b_1, \cdots, b_p)^T$. By Assumption 2, for any $1 \leq l \leq p$,

$$\text{Var}(b_l) = \boldsymbol{e}_l^T(\mathbf{X}^T_{A^0}\mathbf{X}_{A^0})^\dagger\mathbf{X}^T_{A^0}\Sigma\mathbf{X}_{A^0}(\mathbf{X}^T_{A^0}\mathbf{X}_{A^0})^\dagger\boldsymbol{e}_l \leq \sigma^2\big((\mathbf{X}^T_{A^0}\mathbf{X}_{A^0})^\dagger\big)_{ll} \leq \frac{1}{n}\sigma^2 c_3^2 I(l \in A^0). \tag{53}$$

By Assumptions 2 and (23), we have

$$\begin{aligned} &\mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}^{ol} - \boldsymbol{\beta}^0\right\|_\infty > 0.5\tau\right) \leq \sum_{i \in A^0} \mathbb{P}(|\boldsymbol{b}|_i > 0.5\tau) \\ \leq &2|A^0|\exp\left(-\frac{1}{8}\frac{n\tau^2}{\sigma^2 c_3^2}\right) \leq \frac{2(\kappa^0_{H_0} + |B|)}{p^4 n^4} = o\Big(\frac{1}{p^4 n^3}\Big). \end{aligned} \tag{54}$$

Case 1: $\kappa = |A^0_{H_0}|$. In this case, (50) implies that over the event E

$$\text{supp}\{\boldsymbol{\beta}^0\}\backslash B = \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B. \tag{55}$$

Thus, under the requirement for (τ, λ) in Theorem 5.1,

$$\begin{aligned}&\mathbb{P}(\text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B \neq A^0\backslash B)\\ =&\mathbb{P}(\{\text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B \neq A^0\backslash B\}\cap E) + \mathbb{P}(\{\text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B \neq A^0\backslash B\}\cap E^c).\end{aligned} \tag{56}$$

Notice that $E \subset \{\text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B = A^{[t_{max}]}\backslash B\}$, and $E \subset \{A^{[t_{max}]}\backslash B = A^0\backslash B\}$. Thus,

$$\mathbb{P}(\{\text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B \neq A^0\backslash B\}\cap E) = 0, \text{ and}$$

$$\mathbb{P}(\text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B \neq A^0\backslash B) \leq \mathbb{P}(E^c) \leq \frac{C'}{n^3}, \tag{57}$$

where C' is a absolute constant.

Note that $\{\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\beta}}^{ol}\} = \{\text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B = A^0\backslash B\}$. By Borel-Cantelli lemma, we have $\{\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\beta}}^{ol}\}$ almost surely as $n \to \infty$.

It remains to show that $\widehat{\boldsymbol{\beta}}^{ol}$ is a global minimizer of (14) or (15) with high probability.

Applying Theorem 2 of Shen et al. (2013) and its proof, with the degree of separation condition 4, we obtain

$$\mathbb{P}\left(\widehat{\boldsymbol{\beta}}^{\ell_0} \neq \widehat{\boldsymbol{\beta}}^{ol}\right) \leq \frac{e+1}{e-1}\exp\left(-\frac{n}{18\sigma^2}\left(C_{\min} - 36\frac{\log p + \log n}{n}\sigma^2\right)\right) \leq \frac{e+1}{e-1}\frac{1}{p^2n^2}, \tag{58}$$

where $\widehat{\boldsymbol{\beta}}^{\ell_0}$ denotes the global minimizer of (14) or (15). By Borel-Cantelli lemma, we have $\{\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\beta}}^{ol} = \widehat{\boldsymbol{\beta}}^{\ell_0}\}$ almost surely as $n \to \infty$.

Case 2: $\kappa \geq |A^0_{H_0}|$. Thus, under the requirement for (τ, λ) in Theorem 5.1,

$$\begin{aligned}&\mathbb{P}(\text{supp}\{\boldsymbol{\beta}^0\}\backslash B \not\subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B)\\ =&\mathbb{P}(\{\text{supp}\{\boldsymbol{\beta}^0\}\backslash B \not\subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\}\cap E) + \mathbb{P}(\{\text{supp}\{\boldsymbol{\beta}^0\}\backslash B \not\subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\}\cap E^c).\end{aligned} \tag{59}$$

By (50), we know that $E \subset \{\text{supp}\{\boldsymbol{\beta}^0\}\backslash B \subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\}$. Thus,

$$\mathbb{P}(\{\text{supp}\{\boldsymbol{\beta}^0\}\backslash B \not\subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\}\cap E) = 0 \text{ and}$$

$$\mathbb{P}\big(\text{supp}\{\boldsymbol{\beta}^0\}\backslash B \not\subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\big) \leq \mathbb{P}(E^c) \leq \frac{C'}{n^3}. \tag{60}$$

By the Borel-Cantelli lemma, we have $\{\text{supp}\{\boldsymbol{\beta}^0\}\backslash B \subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\}$ almost surely as $n \to \infty$.

Under H_0, we know that $\text{supp}\{\widehat{\boldsymbol{\beta}}^{ol}_{H_0}\} \subset \text{supp}\{\boldsymbol{\beta}^0\}\backslash B$, which implies that $\{\text{supp}\{\widehat{\boldsymbol{\beta}}^{ol}_{H_0}\} \subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\}$ almost surely as $n \to \infty$.

Under H_1, we know that $\text{supp}\{\widehat{\boldsymbol{\beta}}^{ol}_{H_1}\} \subset \text{supp}\{\boldsymbol{\beta}^0\}\cup B$, which implies that $\{\text{supp}\{\widehat{\boldsymbol{\beta}}^{ol}_{H_1}\} \subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\cup B\}$ almost surely as $n \to \infty$. □