Statistica Sinica Preprint No: SS-2025-0087	
Title	Distributed Algorithms for High-Dimensional Statistical Inference and Structure Learning with Heterogeneous Data
Manuscript ID	SS-2025-0087
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0087
Complete List of Authors	Hongru Zhao and Xiaotong Shen
Corresponding Authors	Hongru Zhao
E-mails	zhao1118@umn.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

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

Hongru Zhao and **Xiaotong Shen**
School of Statistics, University of Minnesota, Twin Cities

Abstract: This paper addresses critical data-sharing issues encountered when disseminating individual-level data across multiple sites, particularly under stringent privacy constraints and site heterogeneity. In many multi-site clinical trials, for example, privacy concerns restrict sharing to site-specific summary statistics rather than raw data, complicating the analysis of global effects relative to individual or site-specific effects. Our contribution offers a robust distributed framework for high-dimensional, heterogeneous data analysis that overcomes these limitations. We develop a heterogeneous model that integrates both global and site-specific effects, employing nonconvex regularization via difference of convex programming under an ℓ_0 constraint to ensure selection consistency. Although the underlying optimization problem is worst-case NP-hard, our method converges to the global minimizer in polynomial time with high probability under realistic conditions. Moreover, by applying ℓ_0 penalization exclusively to nuisance parameters while leaving hypothesized parameters unpenalized, our approach yields valid statistical inference. This work not only advances methodological research but also directly addresses the challenges of data sharing in distributed data environments.

Key words and phrases: Multi-site studies, Inference regularization, Asymptotic analysis.

1. Introduction

Multicenter research, especially with clinical data, provides advantages over single-center studies, including larger sample sizes for enhanced generalizability and collaborative resource-sharing (Sidransky et al., 2009; Cheng

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

et al., 2017). However, privacy regulations often restrict access to individual-level data, complicating efforts to pool data across centers (Barrows Jr and Clayton, 1996). Consequently, there is a pressing need for efficient statistical tools that synthesize evidence while maintaining privacy.

In parallel, federated learning (Konečnỳ et al., 2016; McMahan et al., 2017) aims to train machine learning models on decentralized data without explicitly sharing them. Many recent studies extend federated algorithms to handle heterogeneous data and improve stability (Yu et al., 2024; Khaled et al., 2020; Wang et al., 2019; Han et al., 2025; Guo et al., 2025; Yu et al., 2025).

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., 2019). These studies propose communication-efficient distributed algorithms for optimization and regression, underlining the statistical inference complexities in decentralized settings. Yet, they

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

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 shown in Theorem 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 genera-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

tion 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 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. In Section 4, we demonstrate the convergence and consistency of our proposed algorithm in a general linear model setting. Section 5 establishes the theoretical properties of our estimator and the con-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

strained likelihood ratio test, including the generalized Wilks' phenomenon. Finally, Section 6 summarizes our findings and discusses the implications of our work.

1.1 Our Contribution

Our main contributions are fourfold.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

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 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 Regression Models

In this section, we formally introduce our heterogeneous regression framework—covering both linear and logistic cases (and, by extension, other GLM-type models)—and establish the notation used throughout. We aim

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

to develop heterogeneous regression methods that account for heterogeneity across K sites, facilitating structure learning and distinguishing global vs. site-specific effects. 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

$$L(\boldsymbol{\beta}) = L_{pooled}(\boldsymbol{\beta}) := \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j), \boldsymbol{\beta}^T = [\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_K^T], \tag{2.1}$$

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 $\boldsymbol{\beta}^0$.

We now give two examples of $L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j)$: one under the heterogeneous linear model and one under logistic regression.

Remark 1. Under the heterogeneous linear model, the responses at each site j can be written in matrix form as $\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta}_0 + \mathbf{W}_j\boldsymbol{\beta}_j + \varepsilon_j, \varepsilon_j \sim N_{n_j}\left(0, \sigma^2 I_{n_j}\right)$, where $\mathbf{Y}_j \in \mathbb{R}^{n_j}$ is the response vector, $\mathbf{X}_j \in \mathbb{R}^{n_j \times p_0}$ is the global design matrix, $\mathbf{W}_j \in \mathbb{R}^{n_j \times p_j}$ is the site-specific design matrix, $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_0}$ is the central parameter vector, $\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}$ is the nuisance parameter vector, and ε_j is Gaussian noise with variance σ^2. The squared

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

error loss at site j is $L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) = \frac{1}{2}\left\|\mathbf{Y}_j - \mathbf{X}_j\boldsymbol{\beta}_0 - \mathbf{W}_j\boldsymbol{\beta}_j\right\|_2^2$. A discussion of site-specific heteroskedastic error variances $\varepsilon_j \sim N_{n_j}\left(0, \sigma_j^2 I_{n_j}\right)$ appears in Supplementary Material S1.

Remark 2. Suppose the response at site j is $\mathbf{Y}_j \in \{0,1\}^{n_j}$. For a logistic model, we introduce $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_0}, \boldsymbol{\beta}_j \in \mathbb{R}^{p_j}$, respectively. The negative log-likelihood at site j becomes $L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) = -\mathbf{Y}_j^T(\mathbf{X}_j\boldsymbol{\beta}_0 + \mathbf{W}_j\boldsymbol{\beta}_j) + \mathbf{1}_{n_j}^T \log\{1 + \exp(\mathbf{X}_j\boldsymbol{\beta}_0 + \mathbf{W}_j\boldsymbol{\beta}_j)\}$.

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, $\widehat{\boldsymbol{\beta}}^{ol} = \left(\widehat{\boldsymbol{\beta}}^{ol}_{A^0}, \mathbf{0}\right)^T$ supported on A^0. The following optimization problems have been described in Shen et al. (2013).

Constrained ℓ_0-method

Consider the ℓ_0-constrained regression problem

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & S(\boldsymbol{\beta}) = \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) \\ \text{subj to:} \quad & \sum_{(k,j)\in\mathcal{S}} I\left(\beta_{kj} \neq 0\right) \leq \kappa, \end{aligned} \tag{3.1}$$

where $\kappa > 0$ is an integer-valued tuning parameter. Denote the global minimizer of (3.1) as $\widehat{\boldsymbol{\beta}}^{\ell_0} = \left(\widehat{\boldsymbol{\beta}}^{\ell_0}_{\widehat{A}^{\ell_0}}, \mathbf{0}\right)^T$. Theorem 2 in Shen et al. (2013)

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

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 (3.1).

3.1 Algorithm

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

$$\widetilde{\mathbf{\Gamma}}^{[t+1]} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}; \widetilde{\mathbf{\Gamma}}^{[t]}), \tag{3.2}$$

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 (3.2) at the t-th iteration. The DC algorithm terminates at $\widetilde{\mathbf{\Gamma}} = \widetilde{\mathbf{\Gamma}}^{[t]}$ if $S(\widetilde{\mathbf{\Gamma}}^{[t]}; \widetilde{\mathbf{\Gamma}}^{[t]}) \le S(\widetilde{\mathbf{\Gamma}}^{[t+1]}; \widetilde{\mathbf{\Gamma}}^{[t]}) +$ machine tolerance, or if t reaches a large pre-specified maximum number of iterations. We obtain the solution $\widehat{\mathbf{\Gamma}}$ of (3.1) 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

Algorithm 1 Constrained minimization via DC programming & ℓ_0 projection

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

2: **Weighted Lasso Update:** Use a weighted Lasso solver to solve (3.2).

3: **Check Convergence:** 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: **Identify the Top-κ Indices:** Let

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

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: **ℓ_0-Projected Estimator:** Compute the ℓ_0 projection estimator $\widehat{\mathbf{\Gamma}}$:

$$\widehat{\mathbf{\Gamma}} = \arg\min_{\boldsymbol{\beta}} \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) \text{ s.t. } \beta_{kj} = 0 \text{ for } (k, j) \in \mathcal{S}\setminus C. \tag{3.3}$$

6: **Output:** The ℓ_0-projected estimator $\widehat{\mathbf{\Gamma}}$.

For the weighted Lasso problem (3.2) 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

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

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

$$\text{central server update} : \widehat{\boldsymbol{\beta}}_0^{(l+1)} = \text{solver}\left\{\widehat{\boldsymbol{\beta}}_0^{(l)}, \sum_{j=1}^{K} \frac{\partial S_j(\widehat{\boldsymbol{\beta}}_0^{(l)}, \widehat{\boldsymbol{\beta}}_j^{(l)})}{\partial \boldsymbol{\beta}_0}; \boldsymbol{w}^0\right\}, \text{ and} \tag{3.4}$$

$$\text{site server update} : \widehat{\boldsymbol{\beta}}_j^{(l+1)} = \text{solver}\left\{\widehat{\boldsymbol{\beta}}_j^{(l)}, \frac{\partial S_j(\widehat{\boldsymbol{\beta}}_0^{(l)}, \widehat{\boldsymbol{\beta}}_j^{(l)})}{\partial \boldsymbol{\beta}_j}; \boldsymbol{w}^j\right\}, j \in [K], \tag{3.5}$$

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

At the fixed t-th iteration in Algorithm 1, the weight for the l-th iteration of the weighted Lasso solver (step 2) is given by

$$\boldsymbol{w} = \left\{\lambda\tau \cdot I\left(\left|\widetilde{\Gamma}_{kj}^{[t]}\right| \leq \tau\right) : (k,j) \in \mathcal{S}\right\},$$

where $\widetilde{\boldsymbol{\Gamma}}^{[t]}$ is the solution at the t-th iteration.

For the central server and each site $j \in \{0, 1, \cdots, K\}$, the corresponding

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

weights are

$$\boldsymbol{w}^j = \left\{ \lambda\tau \cdot I\left(\left| \widetilde{\Gamma}^{[t]}_{k'j'} \right| \leq \tau \right) : (k', j') \in \mathcal{S}, j' = j \right\}.$$

Identifying the top-κ indices in Step 4 of Algorithm 1 might appear to require transmitting all site-specific nuisance parameters to the central server for ranking. However, a threshold-based selection algorithm (see Algorithm S2 in Supplementary Material, Section S5) enables this step to be performed in a fully distributed manner: each site sends only a summarized count to the central server, thereby avoiding the need to share individual parameter estimates.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

Algorithm 2 Constrained Minimization in the Distributed Algorithm Setting

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

2: For each $j = 0, \ldots, K$, let $\boldsymbol{w}^j = \big\{\lambda\tau\, I\big(|\widetilde{\Gamma}^{[t]}_{k'j'}| \leq \tau\big) : (k', j') \in \mathcal{S},\, j' = j\big\}$, set the inner iteration counter $l = 0$, and initialize $\widehat{\boldsymbol{\beta}}_j^{(l)}$.

3: **Site-by-Site Parameter Updates:**

4: **for** $j = 1$ to K **do**

- Update the site-specific parameter $\widehat{\boldsymbol{\beta}}_j^{(l)}$ using the weighted Lasso solver in (3.5).
- Pass $\frac{\partial S_j(\widehat{\boldsymbol{\beta}}_0^{(l)}, \widehat{\boldsymbol{\beta}}_j^{(l)})}{\partial \boldsymbol{\beta}_0}$ to the central server.
- The central server updates $\widehat{\boldsymbol{\beta}}_0^{(l)}$ according to (3.4).

5: **end for**

6: **Check Convergence of Inner Iterations:** If

$$\max_{1\leq j\leq K}\left\|\frac{\partial S_j(\widehat{\boldsymbol{\beta}}_{l,0}, \widehat{\boldsymbol{\beta}}_{l,j})}{\partial \boldsymbol{\beta}_j}\right\|_2 \quad \text{and} \quad \left\|\sum_{j=1}^{K}\frac{\partial S_j(\widehat{\boldsymbol{\beta}}_{l,0}, \widehat{\boldsymbol{\beta}}_{l,j})}{\partial \boldsymbol{\beta}_0}\right\|_2$$

are below prespecified tolerances, proceed; otherwise set $l \leftarrow l + 1$ and return to Step 3.

7: **Update Overall Parameter Estimates:** Set $\widetilde{\mathbf{\Gamma}}^{[t+1]} \leftarrow \widehat{\boldsymbol{\beta}}_l$. If

$$S\big(\widetilde{\mathbf{\Gamma}}^{[t]}, \widetilde{\mathbf{\Gamma}}^{[t]}\big) - S\big(\widetilde{\mathbf{\Gamma}}^{[t+1]}, \widetilde{\mathbf{\Gamma}}^{[t]}\big)$$

has not converged, set $t \leftarrow t + 1$ and return to Step 2.

8: **Identify Top-κ Indices:** Apply the threshold-based selection algorithm (see Algorithm S2 in Supplementary Material, Section S5) to $\widetilde{\mathbf{\Gamma}}^{[t]}$, obtaining a set $C \subset \mathcal{S}$ such that $|C| = \kappa$.

9: **ℓ_0-Projected Estimator:**

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

10: **Output:** The ℓ_0-projected estimator $\widehat{\mathbf{\Gamma}}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

Remark 3. A single weighted Lasso iteration broadcasts the current coefficient vector from the center to all K sites and uploads the site–specific gradients back, incurring $2Kp_0$ scalar transmissions. Across T_{DC} outer DC iterations, each containing T_{Lasso} weighted Lasso steps, the total traffic is $2Kp_0T_{\mathrm{DC}}T_{\mathrm{Lasso}}$ scalars, where T_{DC} denotes the number of outer DC iterations and T_{Lasso} the number of weighted Lasso iterations per DC step. Algorithm S2 communicates only site–level counts: every bisection round involves one scalar upload and one scalar broadcast per site, i.e., $2K$ scalars, for a total of $2KT_\kappa$ scalars over T_κ rounds. Hence, the combined communication cost of these two stages is $2Kp_0T_{\mathrm{DC}}T_{\mathrm{Lasso}} + 2KT_\kappa$. Theorem 1 further shows that T_{DC} is upper-bounded by a quantity logarithmic in the sparsity level.

Remark 4. The initial $\widetilde{\mathbf{\Gamma}}^{[0]}$ needs to be sparse, such as $\mathbf{0}$ or a sparse estimator obtained through penalized methods.

4. Convergence and Consistency Results

Truncated ℓ_1 penalties furnish valid post-selection inference through score-based, constrained-likelihood, debiased, and adaptive testing procedures (Kim et al., 2014; Shen et al., 2013; Zhu et al., 2020; Wu et al., 2020). They likewise attain near-oracle risk and support recovery for regression, network estimation, clustering, and longitudinal modeling (Shen et al., 2012; Kim et al., 2013; Pan et al., 2013; Wu et al., 2016; Austin et al., 2020).

In this section we analyze support recovery, while Section 5 develops the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

constrained likelihood-ratio test.

4.1 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

$$H_0 : \boldsymbol{\beta}_B = 0 \text{ versus } H_1 : \boldsymbol{\beta}_B \neq 0. \tag{4.1}$$

Here, $\boldsymbol{\beta}_B = 0$ if and only if $\beta_i = 0$ for any $i \in B$.

	Index Set	**Parameter Dimension**	**DC Algorithm**
Non-distributed	$i \in [p]$	$\boldsymbol{\beta} \in \mathbb{R}^p \quad (p = \sum_{j=0}^K p_j)$	Algorithm 1
Distributed	$(k, j) \in \mathcal{S}$	$\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}$, $0 \leq j \leq K$	Algorithm 2

In this section, we present the non-distributed version of our DC programming and ℓ_0-projection procedure in Algorithm 3. It builds on the framework of Algorithm 1. A corresponding distributed version, analogous to Algorithm 2, can be derived with straightforward extensions; hence, we omit the distributed counterpart of Algorithm 3.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

4.2 Computational Algorithm

Consider the constrained optimization problem regression for H_0 in (4.1)

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & S(\boldsymbol{\beta}) = \sum_{i=1}^{n} -\log\{f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})\} \\ \text{subj to:} \quad & \sum_{i\in[p]\backslash B} I\left(\beta_i \neq 0\right) \leq \kappa, \beta_B = 0 \end{aligned} \tag{4.2}$$

and for H_1 in (4.1)

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & S(\boldsymbol{\beta}) = \sum_{i=1}^{n} -\log\{f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})\} \\ \text{subj to:} \quad & \sum_{i\in[p]\backslash B} I\left(\beta_i \neq 0\right) \leq \kappa, \end{aligned} \tag{4.3}$$

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

$$\widehat{\boldsymbol{\beta}}^{ol}_{H_0} = \arg \min_{\boldsymbol{\beta}:\boldsymbol{\beta}_{(A^0_{H_0})^c}=\mathbf{0}} L(\boldsymbol{\beta}) \text{ with } \kappa^0_{H_0} = |A^0_{H_0}|, \text{ and} \tag{4.4}$$

$$\widehat{\boldsymbol{\beta}}^{ol}_{H_1} = \arg \min_{\boldsymbol{\beta}:\boldsymbol{\beta}_{(A^0_{H_1})^c}=\mathbf{0}} L(\boldsymbol{\beta}), \text{ respectively,} \tag{4.5}$$

where $A^0_{H_0} = \{i \in [p]\backslash B; \beta^0_i \neq 0\}$ and $A^0_{H_1} = \{i \in [p]\backslash B; \beta^0_i \neq 0\} \cup B$.

For the given hypothesis set B, at the $(t+1)$-th iteration, we solve the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

following weighted Lasso problems, corresponding to H_0 and H_1 respectively:

$$\widetilde{\mathbf{\Gamma}}^{[t+1]} = \arg\min_{\boldsymbol{\beta}:\boldsymbol{\beta}_B=\mathbf{0}} S(\boldsymbol{\beta}; \widetilde{\Gamma}^{[t]}), \text{ and} \tag{4.6}$$

$$\widetilde{\mathbf{\Gamma}}^{[t+1]} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}; \widetilde{\Gamma}^{[t]}), \tag{4.7}$$

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

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

and $\widetilde{\mathbf{\Gamma}}^{[t]}$ is the solution of (4.6) or (4.7), respectively, at the t-th iteration. The DC algorithm terminates at $\widetilde{\mathbf{\Gamma}} = \widetilde{\mathbf{\Gamma}}^{[t]}$ such that $S(\widetilde{\mathbf{\Gamma}}^{[t]}; \widetilde{\mathbf{\Gamma}}^{[t]}) \le S(\widetilde{\mathbf{\Gamma}}^{[t+1]}; \widetilde{\mathbf{\Gamma}}^{[t]}) +$ machine epsilon, or if t reaches a large pre-specified maximum number of iterations (or $\text{supp}\{\widetilde{\mathbf{\Gamma}}^{[t]}\}\setminus B = \text{supp}\{\widetilde{\mathbf{\Gamma}}^{[t+1]}\}\setminus B$). We then obtain the approximate solution $\widehat{\mathbf{\Gamma}}$ to (4.2) or (4.3), respectively, by projecting $\widetilde{\mathbf{\Gamma}}_{B^c}$ onto the ℓ_0-constrained set $\{\|\mathbf{\Gamma}_{B^c}\|_0 \le \kappa\}$.

4.3 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:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon}, \tag{4.8}$$

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

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

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 (4.6) for H_0 or (4.7) 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:

$$H_0 : \widehat{\mathbf{\Gamma}}_{H_0} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) \text{ s.t. } \boldsymbol{\beta}_i = 0, \text{ for } i \in [p]\backslash C, \text{ or} \tag{4.9}$$

$$H_1 : \widehat{\mathbf{\Gamma}}_{H_1} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) \text{ s.t. } \boldsymbol{\beta}_i = 0, \text{ for } i \in [p]\backslash(B \cup C). \tag{4.10}$$

on n. Consider $B \subset [p]$ such that

$$\frac{\sqrt{|B|}(|A^0| + |B|)}{n} \to 0. \tag{4.11}$$

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

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

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 $c_1 > 0$,

$$\min_{A:|A\setminus B|\leq 2\kappa_{max}} \min_{\boldsymbol{\xi}:\|\boldsymbol{\xi}_{A^c}\|_1 \leq 3\|\boldsymbol{\xi}_A\|_1} \frac{\|\mathbf{X}\boldsymbol{\xi}\|_2^2}{n\,\|\boldsymbol{\xi}\|_2^2} \geq c_1, \tag{4.12}$$

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 $c_2, c_3 > 0$,

$$\begin{aligned} &\max_{1\leq i\leq p} \frac{1}{n}\{\mathbf{X}^T(I - P_A)\mathbf{X}\}_{ii} \leq c_2^2, \\ &\max_{1\leq i\leq p} n\{(\mathbf{X}_A^T\mathbf{X}_A)^\dagger\}_{ii} \leq c_3^2, \end{aligned} \tag{4.13}$$

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)**.**

$$\min_{\beta_i^0 \neq 0, i\notin B} \frac{|\beta_i^0|}{\sigma} \geq \frac{50c_3}{3}\sqrt{\frac{\log p}{n} + \frac{\log n}{n}}. \tag{4.14}$$

Assumption 4 (Degree of separation)**.**

$$\begin{aligned} C_{\min} = C_{\min}(\boldsymbol{\beta}^0, \mathbf{X}) &:= \min_{A:|A|\leq|A^0| \text{ and } A\neq A^0} \inf_{\boldsymbol{\beta}} \frac{\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}_{A\cup B}\boldsymbol{\beta}_{A\cup B}\|_2^2}{n\,|A^0\setminus A|} \\ &\geq 72\sigma^2\frac{\log p + \log n}{n}. \end{aligned} \tag{4.15}$$

Assumption 1 is a common condition related to restricted eigenvalues,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

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).

4.4 Correct Identification

The theory presented extends the correct identification result for structure learning, as found in Theorem 14 of Li et al. (2023), to include selection consistency.

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

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

then the following statements hold:

- *under both* H_0 *and* H_1, $\widehat{\mathbf{\Gamma}}$ *in Algorithm 3 yields the oracle estimators* (4.4) *and* (4.5), *as well as the global minimizer of* (4.2) *and* (4.3),

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

respectively, almost surely; the DC algorithm almost surely converges in at most $\lceil \log(2\,\kappa_{\max}) / \log 4 \rceil$ *iterations, where* $\kappa_{\max} = \max\{\kappa, \kappa_1\}$ *and* $\kappa_1 = \Big|\{i \in [p] \backslash B : |\widetilde{\Gamma}_i^{[0]}| \geq \tau\}\Big|$.

Moreover, by replacing condition 2 with $\kappa \geq |A_{H_0}^0|$, *Algorithm 3 ensures:*

- *under both* H_0 *and* H_1, *the supports of the oracle estimators* (4.4) *and* (4.5) *are subsets of* $\mathrm{supp}(\widehat{\mathbf{\Gamma}}) \backslash B$ *and* $\mathrm{supp}(\widehat{\mathbf{\Gamma}}) \cup B$, *respectively, almost surely; the DC algorithm almost surely converges in at most* $\lceil \log(2\,\kappa_{\max}) / \log 4 \rceil$ *iterations.*

The first part of Theorem 1 establishes the result of almost sure subset recovery, while the second part confirms the almost sure selection consistency for Algorithm 3.

Remark 5. Building on the foundation established by Algorithm 3 and Theorem 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 as $n, p \to \infty$. 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.

5. Sampling Distribution and Hypothesis Testing

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

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

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

$$H_0 : \boldsymbol{\beta}_B = 0 \text{ versus } H_1 : \boldsymbol{\beta}_B \neq 0, \tag{5.1}$$

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

5.1 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). As an example, we illustrate our approach using a simple heterogeneous linear regression setting. A heterogeneous linear regression model can be summarized as $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$, with the log-likelihood

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

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 $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{W}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{0} & \mathbf{W}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_K & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{W}_K \end{pmatrix}$. The constrained log-likelihood ratio, corresponding to the test (5.1), is defined as $2\left\{\mathcal{L}_n(\widehat{\boldsymbol{\beta}}^1, \widehat{\sigma}^1) - \mathcal{L}_n(\widehat{\boldsymbol{\beta}}^0, \widehat{\sigma}^0)\right\}$, 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

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

full spaces of the hypothesis test, respectively, that is,

$$\widehat{\boldsymbol{\beta}}^0 = \arg \min_{\|\boldsymbol{\beta}\|_0 \le \kappa, \boldsymbol{\beta}_B = \mathbf{0}} \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j), \text{ and} \tag{5.2}$$

$$\widehat{\boldsymbol{\beta}}^1 = \arg \min_{\|\boldsymbol{\beta}\|_{0,B} \le \kappa} \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j), \tag{5.3}$$

where $\|\boldsymbol{\beta}\|_{0,B} = \sum_{(k,j)\in\mathcal{S}} I(\beta_{kj} \neq 0) I((k,j) \notin B)$ and

$$L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) = \left\| \boldsymbol{Y}_j - \mathbf{X}_j \boldsymbol{\beta}_0 - \mathbf{W}_j \boldsymbol{\beta}_j \right\|_2^2. \tag{5.4}$$

To conduct the hypothesis test (5.1) in this heterogeneous linear regression setting, we replace the non-convex MLEs in (5.2)–(5.3) by the solutions $\widehat{\boldsymbol{\Gamma}}_{H_0}$ and $\widehat{\boldsymbol{\Gamma}}_{H_1}$ from our DC programming and ℓ_0-projection Algorithm 3.

The resulting constrained maximum likelihood ratio (CMLR) statistic $\Lambda_n(B) := 2\left\{\mathcal{L}_n(\widehat{\boldsymbol{\Gamma}}_{H_1}, \widehat{\sigma}^1) - \mathcal{L}_n(\widehat{\boldsymbol{\Gamma}}_{H_0}, \widehat{\sigma}^0)\right\}$, compares these constrained estimators in a manner analogous to a traditional log-likelihood ratio test, where $(\widehat{\sigma}^l)^2 = \frac{1}{n}\left\|\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\Gamma}}_{H_l}\right\|^2$ for $l \in \{0, 1\}$.

Under suitable conditions on κ, τ, λ and the set $|B|$, Theorem 2 establishes Wilks' phenomenon in both the fixed-dimensional and increasing-dimensional regimes for $|B|$. Specifically, the theorem shows that the distribution of $\Lambda_n(B)$ converges to a chi-square distribution for fixed $|B|$ and to a normal distribution (after appropriate centering and scaling) for $|B| \to \infty$.

Theorem 2. *Suppose* $\frac{\sqrt{|B|}\left(|A^0|+|B|\right)}{n} \longrightarrow 0$. *Under Assumptions 1–4, if there exist tuning parameters* (κ, τ, λ) *satisfying the three conditions in*

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

Theorem 1, with $\kappa = |A^0_{H_0}|$*, then, under the null hypothesis* $H_0 : \boldsymbol{\beta}_B = 0$ *(i.e.,* $|A^0| = |A^0_{H_0}|$*), the following hold:*

1. ***Wilks' phenomenon.*** *If* $\beta_{k,j} = 0$ *for all* $(k, j) \in B$ *and* $|B|$ *is fixed, then* $\Lambda_n(B) \xrightarrow{d} \chi^2_{|B|}$ *as* $n \to \infty$.

2. ***Generalized Wilks' phenomenon.*** *If* $\beta_{k,j} = 0$ *for all* $(k, j) \in B$ *and* $|B| \to \infty$*, then* $(2|B|)^{-\frac{1}{2}}\{\Lambda_n(B) - |B|\} \xrightarrow{d} N(0, 1)$ *as* $n \to \infty$.

Remark 6. In practice, the CMLR test can be unstable if the H_0 and H_1 fits are not nested. This may occur when the data–driven sparsity level κ fails to recover the true sparsity excluding B (i.e., $|A^0_{H_0}|$), or when the H_1 solution $\widehat{\boldsymbol{\beta}}^1$ omits coordinates selected under H_0, so that $\text{supp}(\widehat{\boldsymbol{\beta}}^1) \not\supseteq \text{supp}(\widehat{\boldsymbol{\beta}}^0)$. In such cases, the $\chi^2_{|B|}$ reference is invalid. A simple repair is to enforce nesting: (i) compute $\widehat{\boldsymbol{\beta}}^0$ by solving (5.2); (ii) set $\widetilde{B} := B \cup \text{supp}(\widehat{\boldsymbol{\beta}}^0)$; and (iii) obtain the H_1 fit by solving (5.3) on the design restricted to $\widetilde{B}$. Then $\text{supp}(\widehat{\boldsymbol{\beta}}^0) \subseteq \text{supp}(\widehat{\boldsymbol{\beta}}^1)$ by construction, ensuring a valid test statistic. By Theorem 1, if $\kappa \geq |A^0_{H_0}|$ the H_0 fit contains the oracle support with high probability (and if $\kappa = |A^0_{H_0}|$ it exactly recovers it), so $\widetilde{B}$ contains $A^0_{H_0}$; in this exact-recovery case, the Wilks-type limits in Theorem 2 apply. When $\kappa > |A^0_{H_0}|$, only subset recovery is guaranteed, so the $\chi^2_{|B|}$ calibration is *heuristic*; in applications, report the χ^2 p-value alongside a parametric-bootstrap p-value computed under H_0 using the nested refit.

In the context of linear regression, a straightforward asymptotic result can be derived.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

Theorem 3. *Under the same setting as Theorem 2, let B be fixed. Assume further that the Moore–Penrose inverse* $\sigma^2\left(\frac{1}{n}\mathbf{X}_{A^0\cup B}^T\mathbf{X}_{A^0\cup B}\right)^{\dagger}_{B,B}$ *converges in distribution to a positive semidefinite matrix* Σ. *Also assume that* $\{\boldsymbol{\xi}\in\mathbb{R}^{A^0\cup B}:\boldsymbol{\xi}_i=0 \text{ for all } i\notin B\}\subset\mathcal{R}(\mathbf{X}_{A^0\cup B}^T)$, *where* $\mathcal{R}(\mathbf{X}_{A^0\cup B}^T)$ *denotes the column space of* $\mathbf{X}_{A^0\cup B}^T$. *Then*

$$\sqrt{n}(\widehat{\boldsymbol{\Gamma}}_B^{(1)}-\boldsymbol{\beta}_B^0)\xrightarrow{d}N(0,\Sigma). \tag{5.5}$$

6. Simulation Studies

This section uses simulation studies to investigate the support recovery via ROC curves and evaluate the size and power of the proposed CMLR test.

6.1 Support Recovery

For support recovery, we examine both the linear model setting of Remark 1 and the logistic regression setting of Remark 2, each with $K=3$ sites. At every site j ($j=1,2,3$) the global and site-specific covariate vectors are drawn i.i.d. from a centered multivariate Gaussian distribution whose equicorrelation covariance matrix has ones on the diagonal and ρ on the off-diagonal entries. Sparsity is imposed by first selecting s_0 global indices and, independently for each site, s_j local indices; the active coefficients are then sampled from the uniform distribution $\texttt{Unif}\big([\beta_{\min},2\beta_{\min}]\cup[-2\beta_{\min},-\beta_{\min}]\big)$.

Linear model. We fix $n_j=200$, $p_0=p_j=50$, and $s_0=s_j=10$ for all sites and use a common Gaussian error variance σ^2. Each simulated sample

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

is randomly partitioned into equally sized training and validation sets. An ℓ_0-penalized regression is then fitted on the training set across a grid of 100 penalty values tuned by MSE.

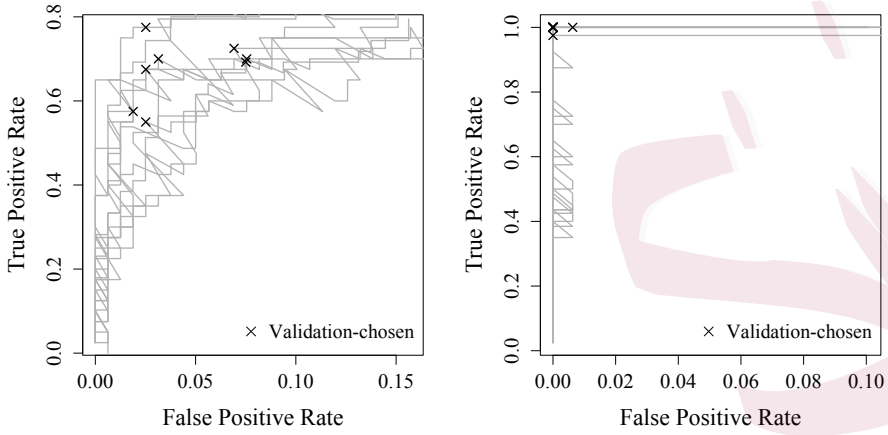

Figure 1: Ten ROC curves (gray) from 10 Monte Carlo replications, with the corresponding validation-chosen models marked by crosses. Left panel—high correlation and weaker signals: $\rho = 0.6$, $\beta_{\min} = 0.4$, $\sigma^2 = 2$. Right panel—moderate correlation and stronger signals: $\rho = 0.3$, $\beta_{\min} = 0.6$, $\sigma^2 = 1$.

Logistic regression. Set $n_j = 500$, $p_0 = p_j = 30$, and $s_0 = 10$, $s_j = 5$ for all sites. Each of the 10 Monte Carlo replications is split evenly into training and validation subsets. An ℓ_0-penalized logistic model is trained over 60 penalty values and tuned by cross-entropy on the validation set. For every tuning point we record the true positive rate (TPR) and false positive rate (FPR); plotting FPR against TPR gives the ROC curve, with the validation-selected model marked by crosses. Under the stronger signals (right panels of Figures 1 and 2) the curves cluster near the upper-left corner, and some replicates recover the exact support—consistent with The-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

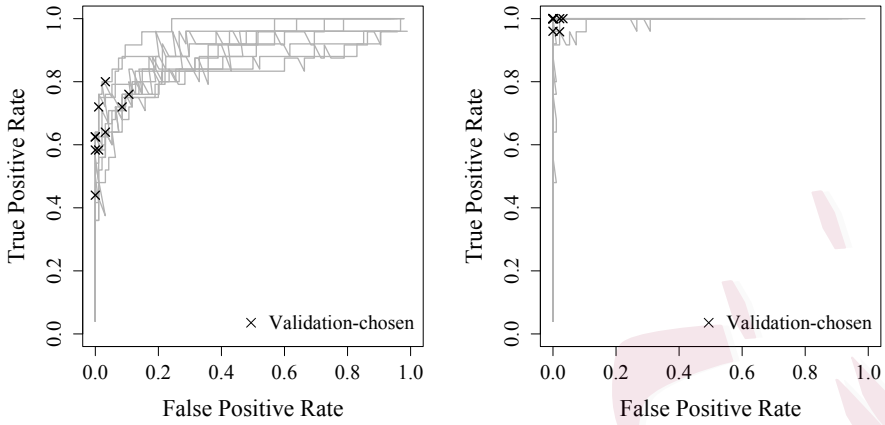

Figure 2: Ten ROC curves (gray) from 10 Monte Carlo replications, with the corresponding validation-chosen models marked by crosses. Left panel—high correlation and weaker signals: $\rho = 0.6$, $\beta_{\min} = 0.4$. Right panel—low correlation and stronger signals: $\rho = 0.2$, $\beta_{\min} = 1$.

orem 1.

6.2 Size and power

We simulate a linear model with $K = 3$ sites, each contributing $n_j = 100$ observations ($n = 300$ total). Covariates are i.i.d. Gaussian with equicorrelation $\rho = 0.5$; the intercept is the first global entry. The total dimension varies over $p = p_0 + 30 \in \{50, 100, 500, 1000\}$ with every site-specific block of size $p_j = 10$. The global vector is set to $\boldsymbol{\beta}_0 = (1, 2, 3, \delta, 0, \ldots, 0)^T \in \mathbb{R}^{p_0}$, and the site–specific vectors are $\boldsymbol{\beta}_j = (j, 0, \ldots, 0)^T \in \mathbb{R}^{p_j}$ for $j = 1, 2, 3$. The block under test is $B = \{4, \ldots, 3 + |B|\}$ with $|B| \in \{1, 5, 10\}$ and $\boldsymbol{\beta}_B = (\delta, 0, \ldots, 0)^T$. When $\delta = 0$ the block coefficients vanish, so the data are generated under the null $H_0 : \boldsymbol{\beta}_B = 0$; positive values $\delta \in \{0.1, 0.2, 0.3\}$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

generate alternative scenarios for the power study.

Each of the 1000 Monte Carlo replications tunes an ℓ_0-penalized model for both H_0 and H_1 (see Remark 6) by 5-fold cross-validation along a regularization path $\{(\lambda_t, \kappa_t)\}_{t=1}^{T}$ with $\lambda_{t+1} < \lambda_t$ and $\kappa_{t+1} \geq \kappa_t$. For simplicity, we fix the truncation parameter at $\tau = 0.3\sqrt{\log(p)/n}$. After the tuning pair $(\widehat{\lambda}, \widehat{\kappa})$ is selected, the replication computes the likelihood ratio statistic and compares it with either the $\chi^2_{|B|}$ reference distribution (χ^2) or the normal approximation (NORMAL) given in Theorem 2. To benchmark CMLR against a widely used alternative, we implement the screen-and-clean procedure (Wasserman and Roeder, 2009): one half of the sample is used to select nuisance covariates (excluding B) with Lasso, and the other half is reserved for classical unpenalized inference. The results appear in Table 1.

7. Real Data Analysis

To demonstrate our distributed framework on a multi-site problem, we analyze the UCI Heart Disease Collection from the UC Irvine Machine Learning Repository. The archive contains data from four hospitals (Cleveland, Hungarian, Switzerland, and VA Long Beach), we retain $n = 734$ complete observations. The seven global predictors (`age`, `sex`, `chol`, `trestbps`, `thalach`, `oldpeak`, `cp`) expand to nine design columns after one-hot encoding and share common effects across hospitals. The six site-specific predictors (`fbs`, `restecg`, `exang`, `slope`, `thal`, `ca`) generate 36 encoded columns and are allowed to vary by hospital. For the non-Cleveland hos-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

Table 1: Empirical size (level $\alpha = 0.05$) and power of the two CMLR tests and a screen-and-clean benchmark based on 1000 Monte Carlo replications. Power is reported for signal strengths $\delta \in \{0.1, 0.2, 0.3\}$.

$\lvert B\rvert$	n	p	Method	Size	Power
1	300	50	χ^2	0.048	(0.251, 0.765, 0.980)
			NORMAL	0.057	(0.297, 0.810, 0.988)
			SCREEN-AND-CLEAN	0.056	(0.153, 0.483, 0.775)
		100	χ^2	0.048	(0.272, 0.763, 0.967)
			NORMAL	0.062	(0.314, 0.798, 0.976)
			SCREEN-AND-CLEAN	0.057	(0.182, 0.478, 0.745)
		500	χ^2	0.045	(0.250, 0.719, 0.952)
			NORMAL	0.065	(0.303, 0.758, 0.969)
			SCREEN-AND-CLEAN	0.059	(0.174, 0.417, 0.769)
		1000	χ^2	0.054	(0.262, 0.742, 0.962)
			NORMAL	0.067	(0.306, 0.776, 0.972)
			SCREEN-AND-CLEAN	0.067	(0.174, 0.412, 0.739)
5	300	50	χ^2	0.053	(0.176, 0.544, 0.894)
			NORMAL	0.069	(0.216, 0.602, 0.914)
			SCREEN-AND-CLEAN	0.064	(0.136, 0.317, 0.572)
		100	χ^2	0.049	(0.158, 0.536, 0.878)
			NORMAL	0.080	(0.194, 0.596, 0.910)
			SCREEN-AND-CLEAN	0.085	(0.114, 0.300, 0.570)
		500	χ^2	0.059	(0.143, 0.524, 0.871)
			NORMAL	0.083	(0.185, 0.582, 0.893)
			SCREEN-AND-CLEAN	0.084	(0.139, 0.305, 0.582)
		1000	χ^2	0.072	(0.149, 0.508, 0.879)
			NORMAL	0.102	(0.191, 0.567, 0.896)
			SCREEN-AND-CLEAN	0.080	(0.125, 0.294, 0.569)
10	300	50	χ^2	0.061	(0.120, 0.441, 0.828)
			NORMAL	0.083	(0.158, 0.499, 0.871)
			SCREEN-AND-CLEAN	0.083	(0.119, 0.266, 0.528)
		100	χ^2	0.059	(0.139, 0.411, 0.800)
			NORMAL	0.075	(0.175, 0.461, 0.830)
			SCREEN-AND-CLEAN	0.085	(0.147, 0.246, 0.490)
		500	χ^2	0.063	(0.127, 0.433, 0.790)
			NORMAL	0.083	(0.150, 0.476, 0.818)
			SCREEN-AND-CLEAN	0.097	(0.155, 0.278, 0.496)
		1000	χ^2	0.078	(0.140, 0.417, 0.775)
			NORMAL	0.094	(0.171, 0.473, 0.823)
			SCREEN-AND-CLEAN	0.115	(0.161, 0.298, 0.513)

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

pitals, missing values in `slope` and `thal` are recoded as a separate factor level representing the missing state, and `ca` is dropped because it is entirely missing. We fit the heterogeneous logistic regression model described in Remark 2 and select the tuning parameters (λ, κ) via 5–fold cross–validation. We test global predictors while treating site-specific coefficients as nuisance parameters. Table 2 summarizes the CMLR results: `sex`, `oldpeak`, and `cp` are strongly associated with heart disease accounting for site heterogeneity.

Variable	**Explanation**	**df**	**Statistic**	**p-value**
`age`	Age (years)	1	1.81	1.79×10^{-1}
`sex`	Sex (1=male, 0=female)	1	17.7	2.58×10^{-5}
`chol`	Serum cholesterol (mg/dl)	1	2.72	9.91×10^{-2}
`trestbps`	Resting blood pressure (mm Hg)	1	0.992	3.19×10^{-1}
`thalach`	Max heart rate achieved	1	1.48	2.24×10^{-1}
`oldpeak`	ST depression (exercise vs rest)	1	14.5	1.43×10^{-4}
`cp`	Chest pain type (4 types)	3	40.0	1.08×10^{-8}

Table 2: Variables, brief explanations, and constrained LRT statistics.

8. 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 an ℓ_0 penalty for estimation and inference.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

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 selection consistency and support recovery when the tuning parameter for the ℓ_0-constraint equals the true sparsity level. Moreover, 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 ℓ_0 penalty enables simultaneous variable selection and parameter estimation, leading to more interpretable models.

Supplementary Material

In the Supplementary Material, we present the threshold-based selection Algorithm S2, the proofs of Theorems 1, 2, and 3. In addition, Section S1 extends the heterogeneous linear model in Remark 1 to site-specific heteroskedastic errors and introduces a block coordinate-descent estimator for jointly estimating coefficients and variances.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

Acknowledgments

The authors thank the reviewers for their careful reading and constructive comments, which have helped improve the quality and clarity of the manuscript. This work was supported in part by NSF grant DMS-2513668 and NIH grants R01AG069895, R01AG065636, R01AG074858, U01AG073079.

References

Austin, E., W. Pan, and X. Shen (2020). A new semiparametric approach to finite mixture of regressions using penalized regression via fusion. *Statistica Sinica 30*(2), 783–807.

Barrows Jr, R. C. and P. D. Clayton (1996). Privacy, confidentiality, and electronic medical records. *Journal of the American Medical Informatics Association 3*(2), 139–148.

Battey, H., J. Fan, H. Liu, J. Lu, and Z. Zhu (2018). Distributed testing and estimation under sparse high dimensional models. *Annals of statistics 46*(3), 1352–1382.

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

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Chen, Y., D. Ge, M. Wang, Z. Wang, Y. Ye, and H. Yin (2017). Strong np-hardness for sparse optimization with concave penalty functions. In *International Conference on Machine Learning*, pp. 740–747. PMLR.

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

Cheng, A., D. Kessler, R. Mackinnon, T. P. Chang, V. M. Nadkarni, E. A. Hunt, J. Duval-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

Arnould, Y. Lin, M. Pusic, and M. Auerbach (2017). Conducting multicenter research in healthcare simulation: Lessons learned from the inspire network. *Advances in Simulation 2*(1), 1–14.

Daubechies, I., M. Defrise, and C. De Mol (2004). 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.

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

Fan, J., L. Xue, and H. Zou (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of statistics 42*(3), 819–849.

Guo, Z., X. Li, L. Han, and T. Cai (2025). Robust inference for federated meta-learning. *Journal of the American Statistical Association 120*(551), 1695–1710.

Han, L., J. Hou, K. Cho, R. Duan, and T. Cai (2025). Federated adaptive causal estimation (face) of target treatment effects. *Journal of the American Statistical Association 120*(551), 1503–1516.

Jordan, M. I., J. D. Lee, and Y. Yang (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association 114*(526), 668–681.

Khaled, A., K. Mishchenko, and P. Richtárik (2020). Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4519–4529. PMLR.

Kim, S., W. Pan, and X. Shen (2013). Network-based penalized regression with application to genomic data. *Biometrics 69*(3), 582–593.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

Kim, S., W. Pan, and X. Shen (2014). Penalized regression approaches to testing for quantitative trait-rare variant association. *Frontiers in genetics 5*, Article 121.

Konečnỳ, J., H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

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

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

McMahan, B., E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR.

Pan, W., X. Shen, and B. Liu (2013). Cluster analysis: Unsupervised learning via supervised learning with a non-convex penalty. *Journal of machine learning research 14*(7), 1865–1889.

Shen, X., W. Pan, and Y. Zhu (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association 107*(497), 223–232.

Shen, X., W. Pan, Y. Zhu, and H. Zhou (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics 65*(5), 807–832.

Shi, C., R. Song, Z. Chen, and R. Li (2019). Linear hypothesis testing for high dimensional generalized linear models. *Annals of statistics 47*(5), 2671–2703.

Sidransky, E., M. A. Nalls, J. O. Aasly, J. Aharon-Peretz, G. Annesi, E. R. Barbosa, A. Bar-Shira, D. Berg, J. Bras, A. Brice, et al. (2009). Multicenter analysis of glucocerebrosidase

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

mutations in parkinson's disease. *New England Journal of Medicine 361*(17), 1651–1661.

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

Wang, J., M. Kolar, N. Srebro, and T. Zhang (2017). Efficient distributed learning with sparsity. In *International conference on machine learning*, pp. 3636–3645. PMLR.

Wang, S., T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan (2019). Adaptive federated learning in resource constrained edge computing systems. *IEEE journal on selected areas in communications 37*(6), 1205–1221.

Wasserman, L. and K. Roeder (2009). High dimensional variable selection. *Annals of statistics 37*(5A), 2178–2201.

Wu, C., S. Kwon, X. Shen, and W. Pan (2016). A new algorithm and theory for penalized regression-based clustering. *Journal of Machine Learning Research 17*(188), 1–25.

Wu, C., G. Xu, X. Shen, and W. Pan (2020). A regularization-based adaptive test for high-dimensional glms. *Journal of Machine Learning Research 21*(128), 1–67.

Yu, S., G. Wang, and L. Wang (2025). Distributed heterogeneity learning for generalized partially linear models with spatially varying coefficients. *Journal of the American Statistical Association 120*(550), 779–793.

Yu, T., S. Ye, and R. Wang (2024). High-dimensional variable selection accounting for heterogeneity in regression coefficients across multiple data sources. *Canadian Journal of Statistics 52*(3), 900–923.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0087

Hongru Zhao School of Statistics, University of Minnesota, Twin Cities, MN 55455, U.S.A.

E-mail: zhao1118@umn.edu

Xiaotong Shen School of Statistics, University of Minnesota, Twin Cities, MN 55455, U.S.A.

E-mail: xshen@umn.edu