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

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 (Konecnỳ 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., 2024b; Khaled et al., 2020; Wang et al., 2019; Han et al., 2025; Guo et al., 2025; Yu et al., 2024a).

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

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 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-

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 general linear model setting. Section 5 establishes the theoretical properties of our estimator and the con-

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 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 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 formally introduce our heterogeneous linear model and the notation used throughout. We aim to develop linear 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 $\boldsymbol{\beta}^0$.

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

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & S(\boldsymbol{\beta}) = \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) \\ \text{subj to:} \quad & \sum_{(k,j)\in\mathcal{S}} I\left(\beta_{kj} \neq 0\right) \leq \kappa, \end{aligned} \tag{3.2}$$

where $\kappa > 0$ is an integer-valued tuning parameter. Denote the global minimizer of (3.2) 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) 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 (3.2).

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,

$$\widetilde{\boldsymbol{\Gamma}}^{[t+1]} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}; \widetilde{\boldsymbol{\Gamma}}^{[t]}), \tag{3.3}$$

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 $\widetilde{\boldsymbol{\Gamma}}^{[t]}$ is the solution of (3.3) at the t-th iteration. The DC algorithm terminates at $\widetilde{\boldsymbol{\Gamma}} = \widetilde{\boldsymbol{\Gamma}}^{[t]}$ if $S(\widetilde{\boldsymbol{\Gamma}}^{[t]};\widetilde{\boldsymbol{\Gamma}}^{[t]}) \le S(\widetilde{\boldsymbol{\Gamma}}^{[t+1]};\widetilde{\boldsymbol{\Gamma}}^{[t]}) +$ Machine tolerance, or if t reaches a large pre-specified maximum number of iterations. Then, we obtain the solution $\widehat{\boldsymbol{\Gamma}}$ of (3.2) by projection $\widetilde{\boldsymbol{\Gamma}}$ onto the ℓ_0-constrained set $\{\|\boldsymbol{\Gamma}\|_0 \le \kappa\}$, where $\|\boldsymbol{\Gamma}\|_0 = \sum_{(k,j)\in\mathcal{S}} I(\Gamma_{kj} \neq 0)$. We summarize the general constrained minimization via DC programming and ℓ_0 projection algorithm in Algorithm 1.

For the weighted Lasso problem (3.3) 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,

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 (3.3).

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 $\widehat{\mathbf{\Gamma}}$:

$$\widehat{\mathbf{\Gamma}} = \arg\min_{\boldsymbol{\beta}} \sum_{j=1}^{K} L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) \text{ s.t. } \beta_{kj} = 0 \text{ for } (k, j) \in \mathcal{S}\setminus C. \tag{3.4}$$

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

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

$$\text{central server update} : \widehat{\boldsymbol{\beta}}_0^{(l+1)} = \text{solver}\left(\widehat{\boldsymbol{\beta}}_0^{(l)}, \sum_{j=1}^{K} \frac{\partial S_j(\widehat{\boldsymbol{\beta}}_0^{(l)}, \widehat{\boldsymbol{\beta}}_j^{(l)})}{\partial \boldsymbol{\beta}_0}; \boldsymbol{w}^0\right), \text{ and} \tag{3.5}$$

$$\text{site server update}: \widehat{\boldsymbol{\beta}}_j^{(l+1)} = \text{solver}\left(\widehat{\boldsymbol{\beta}}_j^{(l)}, \frac{\partial S_j(\widehat{\boldsymbol{\beta}}_0^{(l)}, \widehat{\boldsymbol{\beta}}_j^{(l)})}{\partial \boldsymbol{\beta}_j}; \boldsymbol{w}^j\right), 1 \le j \le K, \tag{3.6}$$

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

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

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 weights are

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

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 S1 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.

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 $j = 0, \ldots, K$, define $\boldsymbol{w}^j = \left\{\lambda\tau \cdot I\left(\left|\widetilde{\Gamma}^{[t]}_{k'j'}\right| \leq \tau\right); (k', j') \in \mathcal{S}, j' = j\right\}$. Set the inner iteration counter $l = 0$, and initialize $\widehat{\boldsymbol{\beta}}_j^{(l)}$ for all $0 \leq j \leq K$.

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 (3.6).
- 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 (3.5).

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 (Threshold-Based Selection):** Apply the threshold-based selection algorithm (see Algorithm S1 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 $\widehat{\mathbf{\Gamma}}$.

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. Convergence and Consistency Results

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

$$H_0 : \boldsymbol{\beta}_B = 0 \text{ versus } H_1 : \boldsymbol{\beta}_B \neq 0. \tag{4.7}$$

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.

4.2 Computational Algorithm

Consider the constrained optimization problem regression for H_0 in (4.7)

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & S(\boldsymbol{\beta}) = \sum_{i=1}^{n} -\log(f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})) \\ \text{subj to:} \quad & \sum_{i\in[p]\backslash B} I\left(\beta_i \neq 0\right) \leq \kappa, \beta_B = 0 \end{aligned} \tag{4.8}$$

and for H_1 in (4.7)

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & S(\boldsymbol{\beta}) = \sum_{i=1}^{n} -\log(f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})) \\ \text{subj to:} \quad & \sum_{i\in[p]\backslash B} I\left(\beta_i \neq 0\right) \leq \kappa, \end{aligned} \tag{4.9}$$

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

$$\widehat{\boldsymbol{\beta}}^{ol}_{H_0} = \arg \min_{\boldsymbol{\beta}:\boldsymbol{\beta}_{(A^0_{H_0})^c}=\mathbf{0}} L(\boldsymbol{\beta}) \text{ with } \kappa^0_{H_0} = |A^0_{H_0}|, \text{ and} \tag{4.10}$$

$$\widehat{\boldsymbol{\beta}}^{ol}_{H_1} = \arg \min_{\boldsymbol{\beta}:\boldsymbol{\beta}_{(A^0_{H_1})^c}=\mathbf{0}} L(\boldsymbol{\beta}), \text{ respectively,} \tag{4.11}$$

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 $(t+1)$-th iteration, we solve the fol-

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

$$\widetilde{\mathbf{\Gamma}}^{[t+1]} = \arg\min_{\boldsymbol{\beta}:\boldsymbol{\beta}_B=\mathbf{0}} S(\boldsymbol{\beta}; \widetilde{\Gamma}^{[t]}), \text{ and} \tag{4.12}$$

$$\widetilde{\mathbf{\Gamma}}^{[t+1]} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}; \widetilde{\Gamma}^{[t]}), \tag{4.13}$$

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 (4.12) or (4.13), 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 approximated solution $\widehat{\mathbf{\Gamma}}$ to (4.8) or (4.9), 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:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon}, \tag{4.16}$$

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 (4.12) for H_0 or (4.13) 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:

$$H_0 : \widehat{\mathbf{\Gamma}}_{H_0} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) \text{ s.t. } \boldsymbol{\beta}_i = 0, \text{ for } i \in [p]\backslash C, \text{ or} \tag{4.14}$$

$$H_1 : \widehat{\mathbf{\Gamma}}_{H_1} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) \text{ s.t. } \boldsymbol{\beta}_i = 0, \text{ for } i \in [p]\backslash(B \cup C). \tag{4.15}$$

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

$$\frac{\sqrt{|B|}(|A^0| + |B|)}{n} \to 0. \tag{4.17}$$

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 $c_1 > 0$,

$$\min_{A:|A\setminus B|\leq 2\kappa_{max}} \min_{\boldsymbol{\xi}:\|\boldsymbol{\xi}_{A^c}\|_1 \leq 3\|\boldsymbol{\xi}_A\|_1} \frac{\|\mathbf{X}\boldsymbol{\xi}\|_2^2}{n\,\|\boldsymbol{\xi}\|_2^2} \geq c_1, \tag{4.18}$$

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 $c_2, c_3 > 0$,

$$\begin{aligned} &\max_{1\leq i\leq p} \frac{1}{n}(\mathbf{X}^T(I - P_A)\mathbf{X})_{ii} \leq c_2^2, \\ &\max_{1\leq i\leq p} n((\mathbf{X}_A^T\mathbf{X}_A)^\dagger)_{ii} \leq c_3^2, \end{aligned} \tag{4.19}$$

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)**.**

$$\min_{\beta_i^0\neq 0, i\notin B} \frac{|\beta_i^0|}{\sigma} \geq \frac{50c_3}{3}\sqrt{\frac{\log p}{n} + \frac{\log n}{n}}. \tag{4.20}$$

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

$$\begin{aligned} C_{\min} = C_{\min}(\boldsymbol{\beta}^0, \mathbf{X}) &:= \min_{A:|A|\leq|A^0| \text{ and } A\neq A^0} \inf_{\boldsymbol{\beta}} \frac{\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}_{A\cup B}\boldsymbol{\beta}_{A\cup B}\|_2^2}{n\,|A^0\setminus A|} \\ &\geq 72\sigma^2\frac{\log p + \log n}{n}. \end{aligned} \tag{4.21}$$

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).

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

then the following statements hold:

- *under both* H_0 *and* H_1, $\widehat{\mathbf{\Gamma}}$ *in Algorithm 3 yields the oracle estimators*

(4.10) *and* (4.11)*, as well as the global minimizer of* (4.8) *and* (4.9)*, respectively;*

- *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* (4.10) *and* (4.11) *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 2. 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 hypothesized parameter subset $B \subset \mathcal{S}$, we consider the hypothesis testing

$$H_0 : \boldsymbol{\beta}_B = 0 \text{ versus } H_1 : \boldsymbol{\beta}_B \neq 0, \tag{5.22}$$

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). In this section, we illustrate our approach using a simple heterogeneous linear regression setting as an example. In the Supplementary Material section S1, we derived a heterogeneous linear regression model, which can be summarized as

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

with the log-likelihood

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

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 (S1.3) in Supplementary Material. The constrained log-likelihood ratio, corresponding to the test (5.22), 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 full spaces of the hypothesis test, respectively, that is,

$$\widehat{\boldsymbol{\beta}}^0 = \arg \min_{\|\boldsymbol{\beta}\|_0 \le \kappa, \boldsymbol{\beta}_B = \mathbf{0}} \sum_{j=1}^K L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j), \text{ and} \tag{5.23}$$

$$\widehat{\boldsymbol{\beta}}^1 = \arg \min_{\|\boldsymbol{\beta}\|_{0,B} \le \kappa} \sum_{j=1}^K L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j), \tag{5.24}$$

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

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

To conduct the hypothesis test (5.22) in this heterogeneous linear regression setting, we replace the non-convex MLEs in (5.23)–(5.24) 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 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 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*

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

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

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

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

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*

$$\left\{\boldsymbol{\xi}\in\mathbb{R}^{A^0\cup B}:\boldsymbol{\xi}_i=0 \text{ for all } i\notin B\right\}\ \subset\ \mathcal{R}\big(\mathbf{X}_{A^0\cup B}^T\big),$$

where $\mathcal{R}(\mathbf{X}_{A^0\cup B}^T)$ *denotes the column space of* $\mathbf{X}_{A^0\cup B}^T$*. Then*

$$\sqrt{n}(\widehat{\boldsymbol{\Gamma}}_B^{(1)}-\boldsymbol{\beta}_B^0)\xrightarrow{d}N(0,\Sigma). \tag{5.25}$$

Since the literature Shen et al. (2012); Zhu et al. (2020); Shen et al. (2013); Kim et al. (2013); Pan et al. (2013); Wu et al. (2020, 2016); Austin et al. (2020) has already systematically studied TLP inference under a wide range of settings, we omit the simulation study on inference performance.

6. Real Data Analysis

In this section, we evaluate the predictive performance of our distributed high-dimensional framework for heterogeneous data using the METABRIC breast cancer dataset from Kaggle. The dataset contains RNA expression profiles, mutation data, and detailed clinical annotations, making it a rich resource for modeling high-dimensional survival outcomes with potential heterogeneity.

6.1 Data Preprocessing

The raw METABRIC data were processed in R, following the pipeline available on our GitHub repository. A detailed dataset description is available on Kaggle dataset webpage. The primary response variable, denoted by `overall_survival_months`, represents the duration from the intervention to the occurrence of death. Among clinical covariates, the `cohort` variable (ranging from 1 to 5) is notable because it naturally partitions subjects into distinct subpopulations, analogous to multiple sites in a multicenter study.

We separate the predictor variables into two categories: (1) clinical attributes, which we further label as 'global' (e.g., baseline demographics, tumor characteristics) or 'site-specific' (e.g., type of surgical intervention, post-chemotherapy cellularity, and treatment indicators) that may differ by cohort, and (2) genetic attributes (RNA expression features), treated as global. In total, these variables yield a high-dimensional design matrix with both global and site-specific components.

To evaluate predictive performance, we consider two modeling strategies. First, a heterogeneous linear model is fitted, integrating both global and site-specific predictors within a unified likelihood framework. Second, independent linear models are estimated separately for each cohort. The results demonstrate that the heterogeneous model, which jointly models global and site-specific effects, exhibits superior predictive accuracy compared to the disjoint cohort-specific models.

6.2 Model Estimation and Evaluation

We implement a heterogeneous linear model that jointly fits global and site-specific predictors via a truncated Lasso penalty (TLP), using the `glmtlp` R package. Given the skewness of the survival response, we consider the log-transformed outcome variable defined as

$$\log(\texttt{overall\_survival\_months} + 1).$$

A 70/30 train-test split is performed, and hyperparameter tuning is conducted via 5-fold cross-validation. To ensure stability in evaluation, the process is repeated over 1000 iterations, and the mean squared error (MSE) on the test set is computed for each repetition. The average test MSE and standard deviation for both the heterogeneous model and the cohort-specific model are summarized in Table 1.

For comparison, we also fit separate TLP-regularized linear models for each cohort individually, treating them as if they were independent analyses. In that case, we compute the test MSE for each cohort, then take a weighted average (weighted by cohort size). Table 1 shows the average test MSE and standard deviation across the 1000 repetitions.

Table 1: Comparison of Test MSE Across Model Frameworks

Model	Average Test MSE	Standard Deviation
Heterogeneous Model	0.6368	0.0508
Cohort-Specific Model	0.6650	0.0527

Table 1 shows that the heterogeneous linear model—which leverages

both global and cohort-specific effects in a unified framework—achieves a lower average test MSE of 0.6368 than the cohort-specific approach (0.6650), with both standard errors no greater than 0.00167. This difference suggests that jointly modeling global features and site-specific terms can yield more accurate predictions, presumably because the global parameters borrow information across cohorts.

6.3 Discussion

Overall, the METABRIC example demonstrates that our distributed, high-dimensional framework can effectively combine global and site-specific predictors to enhance predictive performance. Even when cohorts are naturally heterogeneous, joint estimation methods improve accuracy by pooling information across sites. This advantage highlights the promise of distributed approaches that incorporate nonconvex regularization.

7. 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 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.

Supplementary Materials

In the Supplementary Material, we present the threshold-based selection Algorithm S1, the proofs of Theorems 1, 2, and 3, as well as the precise formulation of the heterogeneous linear regression model in Section 5.1.

Acknowledgements

This work was supported in part by NSF grant DMS-1952539 and NIH grants R01AG069895, R01AG065636, R01AG074858, U01AG073079.

References

Austin, E., Pan, W., and Shen, X. (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 Clayton, P. D. (1996). Privacy, confidentiality, and electronic medical records. *Journal of the American medical informatics association*, 3(2):139–148.

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

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

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

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

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

Cheng, A., Kessler, D., Mackinnon, R., Chang, T. P., Nadkarni, V. M., Hunt, E. A., Duval-Arnould, J., Lin, Y., Pusic, M., and Auerbach, M. (2017). Conducting multicenter research in healthcare simulation: Lessons learned from the inspire network. *Advances in Simula-*

tion, 2(1):1–14.

Daubechies, I., Defrise, M., and De Mol, C. (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., Ning, Y., and Chen, Y. (2022). Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika*, 109(1):67–83.

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

Guo, Z., Li, X., Han, L., and Cai, T. (2025). Robust inference for federated meta-learning. *Journal of the American Statistical Association*, 0(0):1–16.

Han, L., Hou, J., Cho, K., Duan, R., and Cai, T. (2025). Federated adaptive causal estimation (face) of target treatment effects. *Journal of the American Statistical Association*, 0(ja):1–25.

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

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

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

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

arXiv:1610.05492, 8.

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

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

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

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

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

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

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

Sidransky, E., Nalls, M. A., Aasly, J. O., Aharon-Peretz, J., Annesi, G., Barbosa, E. R., Bar-Shira, A., Berg, D., Bras, J., Brice, A., et al. (2009). Multicenter analysis of glucocerebrosidase 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*, volume 48. Cambridge university press.

Wang, J., Kolar, M., Srebro, N., and Zhang, T. (2017). Efficient distributed learning with

sparsity. In *International conference on machine learning*, pages 3636–3645. PMLR.

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

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

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

Yu, S., Wang, G., and Wang, L. (2024a). Distributed heterogeneity learning for generalized partially linear models with spatially varying coefficients. *Journal of the American Statistical Association*, pages 1–15.

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

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

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