Statistica Sinica Preprint No: SS-2026-0052	
Title	Deconfounded-Debiased Distributed Estimation for High-Dimensional Multi-Site Data with Heterogeneous Pervasive Hidden Confounders
Manuscript ID	SS-2026-0052
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202026.0052
Complete List of Authors	Zhaoyang Li, Chen Huang, Guoyou Qin and Zhongyi Zhu
Corresponding Authors	Guoyou Qin
E-mails	gyqin@fudan.edu.cn
Notice: Accepted author version.	

DECONFOUNDED-DEBIASED DISTRIBUTED ESTIMATION FOR HIGH-DIMENSIONAL MULTI-SITE DATA WITH HETEROGENEOUS PERVASIVE HIDDEN CONFOUNDERS

Zhaoyang Li, Chen Huang, Guoyou Qin* and Zhongyi Zhu*

Fudan University and York University

Abstract: In observational studies, high-dimensional multi-site data are subject to heterogeneous pervasive hidden confounders across sites, leading to substantial bias in distributed estimation. To address this challenge, we propose a deconfounded-debiased distributed estimation method, which integrates majority voting for variable selection and the privacy-preserving aggregation of local deconfounded-debiased estimators at the central site. This approach corrects biases from both heterogeneous pervasive hidden confounders and high-dimensional estimation, while also accommodating site-specific heteroscedasticity in the random errors. Theoretically, we prove that local deconfounded-debiased estimators are asymptotically normal, and local individual hypothesis tests are asymptotically valid with an asymptotic lower bound on their power. Furthermore, we establish variable selection consistency and asymptotic normality for the proposed distributed estimator. We also provide finite-sample guarantees for variable selection, including an upper bound on the expected false positive rate and a lower bound on the expected true positive rate. Simulation experiments and an application to a protein dataset from the UK Biobank demonstrate our method's superior finite-sample performance.

Key words and phrases: Distributed estimation, heterogeneous pervasive hidden confounders, majority voting, spectral transformation.

*Corresponding Authors. E-mail: gyqin@fudan.edu.cn; zhuzy@fudan.edu.cn

1. Introduction

In the era of big data, it has become common practice to collect high-dimensional observational data across multiple sites. A conventional approach for analyzing multi-site data is to pool observations from all sites into a single large dataset and then analyze, known as global estimation (Hou et al., 2023). However, as data dimensionality increases dramatically, data pooling often faces serious challenges such as privacy constraints, insufficient memory capacity, or prohibitive computational costs, rendering global estimation infeasible. Distributed estimation methods have thus emerged, leveraging information from all sites without the need to transfer sample-level data (Gao et al., 2022).

Like general observational data, high-dimensional multi-site observational data are inevitably subject to pervasive hidden confounders which are unmeasured and simultaneously affect most predictors and the response. A more complicated issue is that such hidden confounders may differ across sites, referred to as heterogeneous pervasive hidden confounding. For example, our real data application, utilizing data from the UK Biobank (UKB) (Sudlow et al., 2015), focuses on the association between proteins and white matter hyperintensities (WMH). Although the UKB implements rigorous standardization protocols, differences persist across assessment centers in terms of MRI scanner models, imaging parameters, blood collection timings, storage conditions, protein detection batches, among other factors (Bycroft et al., 2018). These factors are difficult to measure and adjust for, thereby posing a risk of heterogeneous pervasive hidden confounding. Analysis of the singular values of predictors at each center provides preliminary evidence for the presence of this confounding form (Figure 6). Consequently, our real data analysis motivates the development of distributed estimation that effectively corrects for bias induced by heterogeneous pervasive

hidden confounding in high-dimensional multi-site observational data.

Distributed estimation methods for high-dimensional multi-site observational data are generally categorized into one-shot and iterative methods (Gao et al., 2022). One-shot methods aggregate local statistics to a central site through averaging (Gu and Chen, 2023; Yu et al., 2022; Tang et al., 2020; Chen and Xie, 2014), whereas iterative methods optimize estimates by transmitting local gradients through multiple communication rounds (Shamir et al., 2014; Jordan et al., 2019; Fan et al., 2023; Liu et al., 2025). When marginal distributions differ across sites, one-shot methods often mitigate distributional bias by weighting local estimates according to each site's information content or degree of heterogeneity (Gu and Chen, 2023; Zhao and Shen, 2025). In contrast, iterative methods rely on multi-round stepwise corrections, enabling site-specific models to adjust mutually and thus approximate the global optimum (Yu et al., 2026; Ogier du Terrail et al., 2025). However, most existing methods are primarily limited to handling heterogeneity in the marginal distributions of observed variables. In high-dimensional multi-site observational data with heterogeneous pervasive hidden confounders, hidden confounding not only induces marginal shifts but also distorts the conditional distribution structure. Consequently, current distributed methods cannot effectively address such data, and their direct application leads to estimation bias.

Pervasive hidden confounders in high-dimensional observational data have motivated the development of numerous methods. Classic approaches, such as instrumental variable regression (Gold et al., 2020), two-stage calibration (Yang and Ding, 2020), and latent confounder adjustment models (Wang et al., 2017), often require prior knowledge of hidden confounders. Spectral transformation methods circumvent this limitation by directly shrinking pervasive hidden confounding without any information about the confounders themselves. Ćevid et al. (2020) apply the lasso method (Tibshirani, 1996) to spectral transformed data, yielding a de-

confounded estimator that nevertheless inherits high-dimensional estimation bias from lasso. Subsequent work by Guo et al. (2022) and Sun et al. (2024)debiases the deconfounded estimator through low-dimensional projections (Zhang and Zhang, 2014) under the sparse precision matrix assumption. They only provide the asymptotic normality for individual coefficient estimates. In contrast, (Li et al., 2025) debiases by inverting the Karush-Kuhn-Tucker conditions of the deconfounded estimator (Javanmard and Montanari, 2014), thereby avoiding the need for precision matrix sparsity. However, its theoretical properties remain to be established. The abovementioned methods are not evidently designed for high-dimensional multi-site observational data with heterogeneous pervasive hidden confounders. This naturally motivates the idea of extending the deconfounded-debiased estimation methods to a distributed framework.

To address this gap, we propose a deconfounded-debiased distributed estimation method through two rounds of communication: (1) Each site computes a deconfounded-debiased local estimator using the generalization of the method in (Li et al., 2025) that accommodates flexible precision matrix estimation. (2) The central site selects significant predictors via majority voting and then aggregates the local estimators using asymptotic confidence distributions. Because deconfounding and debiasing are performed independently at each local site before aggregation, the proposed distributed method effectively corrects biases induced by heterogeneous pervasive hidden confounders and by high-dimensional estimation, while also allowing for site-specific heteroscedasticity of the random errors.

Our work delivers a comprehensive theoretical analysis, covering both local and distributed estimation and characterizing their asymptotic and finite-sample behaviors. First, we fill the theoretical gap left by Li et al. (2025) and give the theoretical properties that hold at each local site. Specifically, we establish the asymptotic normality of local estima-

tors for both each individual predictor and all active predictors. We then demonstrate that local individual hypothesis tests are asymptotically valid and lower-bounded on their power. Building upon these results, we prove the asymptotic and finite-sample properties for our proposed distributed estimator. Specifically, it achieves variable selection consistency and asymptotic normality. To quantify the role of majority voting in variable selection, we derive non-asymptotic bounds on both expected false positive and expected true positive rates. Simulation experiments and an application to the protein dataset from the UKB demonstrate the superior performance of our method for both variable selection and coefficient estimation.

Section 2 introduces notation, describes the high-dimensional multi-site data setting in the presence of pervasive hidden confounders, and presents the corresponding high-dimensional linear regression model with pervasive hidden confounding. Section 3 first reviews the deconfounded-debiased estimation method of Li et al. (2025) and then generalizes it to accommodate flexible precision matrix estimation. Next, we propose the deconfounded-debiased distributed estimation method. Section 4 establishes the theoretical properties of both the local estimators and the distributed estimator. Sections 5 and 6 present simulation studies and a real-data application, respectively. Section 7 concludes the paper with a discussion.

2. Model and notation

We introduce the notation used throughout this paper, describe high-dimensional multi-site data in the presence of pervasive hidden confounders, and give the corresponding high-dimensional linear regression model with pervasive hidden confounders.

2.1 Notation

We start with some notation that will be used throughout the paper. Denote by $[p] = \{1, \ldots, p\}$ for a positive integer p. For $x \in \mathbb{R}$, denote by $\lfloor x \rfloor$ the largest integer no larger than x. Let $I_p \in \mathbb{R}^{p\times p}$ denote the identity matrix. For a set B, denote the cardinality of B by $|B|$. Denote the complement of B by $B^{\complement}$. For a vector $a \in \mathbb{R}^p$, denote by a_j the jth entry. Let $a_B = (a_j)_{j\in B} \in \mathbb{R}^{|B|}$ and $a_{-B} = a_{[p]\setminus B} \in \mathbb{R}^{p-|B|}$ for any index set $B \subset [p]$. Let $\|a\|_0 = |\{j \in [p] : a_j \neq 0\}|$. For a matrix $A \in \mathbb{R}^{n\times p}$, denote by $A_{i,j}$, $A_{i,\cdot}$,and $A_{\cdot,j}$ the (i, j) entry, the ith row, and the jth column, respectively. Let $A_{B,B} = (A_{i,i})_{i\in B} \in \mathbb{R}^{|B|\times|B|}$ for any index set $B \subset [\min\{n, p\}]$. Let $\|A\|_{\max} = \max_{i,j} |A_{i,j}|$. Let $\lambda_i(A)$ be the ith largest singular value of A, and $\lambda_{\min}(A)$, $\lambda_{\max}(A)$ be the smallest and largest singular values, respectively. Let $\|X\|_{\psi_2} = \sup_{q\geq 1} q^{-1/2}(\mathsf{E}|X|^q)^{1/q}$ be the sub-Gaussian norm of a sub-Gaussian random variable X. For two positive sequences $\{a_n\}$ and $\{b_n\}$, write $a_n \lesssim b_n$ if there exists some constant $C > 0$ such that $a_n \leq Cb_n$ for all n, and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Write $a_n \ll b_n$ if $\lim_{n\to\infty} a_n/b_n = 0$. For two positive random sequences $\{X_n\}$ and $\{Y_n\}$, write $X_n \lesssim_{\mathrm{P}} Y_n$ if $X_n \lesssim Y_n$ holds with probability tending to 1 as $n \to \infty$. Write $X_n \xrightarrow{d} X$ if X_n converges to X in distribution.

2.2 Model description

The high-dimensional multi-site data comprises $N = \sum_{k=1}^{K} n_k$ independent observations collected from K distinct sites, where site k contributes n_k samples for $k = 1, ..., K$. At each site k, we observe i.i.d. high-dimensional predictors and responses $\left\{X_{i,\cdot}^{(k)}, Y_i^{(k)}\right\}_{i\in[n_k]}$, as well as unobserved pervasive hidden confounders $\left\{H_{i,\cdot}^{(k)}\right\}_{i\in[n_k]}$. The hidden confounders $\left(H_{i,\cdot}^{(k)}\right)^{\mathrm{T}} \in \mathbb{R}^{q_k}$ jointly affect both the predictors $\left(X_{i,\cdot}^{(k)}\right)^{\mathrm{T}} \in \mathbb{R}^p$ and the response $Y_i^{(k)} \in \mathbb{R}$, potentially inducing spurious associations if not properly accounted for. We allow for covari-

ate shift across sites, i.e., the distributions $P_k(X, H)$ vary, while the conditional distribution $P_k(Y|X, H)$ is identical across all sites. Specifically, the hidden confounders may be heterogeneous across sites in both dimensions q_k and substantive interpretations, though this is not mandatory. We assume that after fully adjusting for all site-specific hidden confounders, the coefficients β of $X_{i,\cdot}^{(k)}$ on $Y_i^{(k)}$ are identical across all sites and are sparse. Random errors are permitted but not forced to exhibit heteroscedasticity across sites.

At site k, we consider a high-dimensional linear regression model with pervasive hidden confounding ($n_k < p$):

$$\begin{aligned} Y_i^{(k)} &= X_{i,\cdot}^{(k)}\beta + H_{i,\cdot}^{(k)}\phi^{(k)} + \xi_i^{(k)}, \\ X_{i,\cdot}^{(k)} &= H_{i,\cdot}^{(k)}\Psi^{(k)} + E_{i,\cdot}^{(k)}, \end{aligned} \tag{2.1}$$

where the idiosyncratic error $\big(E_{i,\cdot}^{(k)}\big)^{\mathrm{T}} \in \mathbb{R}^p$ is independent of $H_{i,\cdot}^{(k)}$ and the random error $\xi_i^{(k)} \in \mathbb{R}$ is independent of both $H_{i,\cdot}^{(k)}$ and $E_{i,\cdot}^{(k)}$. Let $\Sigma_E^{(k)}$ and $\Sigma_X^{(k)}$ denote the population covariance matrices of $E_{i,\cdot}^{(k)}$ and $X_{i,\cdot}^{(k)}$, respectively. Without loss of generality, we assume $\mathsf{E}\big(E_{i,\cdot}^{(k)}\big) = 0$, $\mathsf{E}\big(H_{i,\cdot}^{(k)}\big) = 0$, $\mathsf{Cov}\big(\big(H_{i,\cdot}^{(k)}\big)^{\mathrm{T}}\big) = I_q$, and $\Sigma_X^{(k)} = \big(\Psi^{(k)}\big)^{\mathrm{T}}\Psi^{(k)} + \Sigma_E^{(k)}$. We further assume that the random error $\xi_i^{(k)}$ has zero mean and variance $\big(\sigma_\xi^{(k)}\big)^2$. The coefficients $\phi^{(k)} \in \mathbb{R}^{q_k}$ and $\Psi^{(k)} \in \mathbb{R}^{q_k \times p}$ characterize, respectively, the linear effects of $H_{i,\cdot}^{(k)}$ on $Y_i^{(k)}$ and on $X_{i,\cdot}^{(k)}$, respectively. The coefficient vector of interest $\beta \in \mathbb{R}^p$ captures the true linear effect of $X_{i,\cdot}^{(k)}$ on $Y_i^{(k)}$, which is assumed identical across sites. The number of confounders q_k is taken to be small but may grow with n_k and p for each $k \in [K]$.

At site k, pervasive hidden confounders introduce bias in the estimation of the coefficient vector β. To better understand this issue, we reformulate the model (2.1) as a perturbed high-dimensional linear regression

$$\begin{aligned} Y^{(k)} &= X^{(k)}(\beta + b^{(k)}) + \epsilon^{(k)}, \\ \epsilon^{(k)} &= \xi^{(k)} + H^{(k)}\phi^{(k)} - X^{(k)}b^{(k)}. \end{aligned}$$

The exogeneity condition $\mathsf{Cov}\big(\big(X_{i,\cdot}^{(k)}\big)^{\mathrm{T}}, H_{i,\cdot}^{(k)}\phi^{(k)} - X_{i,\cdot}^{(k)}b^{(k)}\big) = 0$ implies that the perturbed coefficient is given by $b^{(k)} = \big(\Sigma_X^{(k)}\big)^{-1}\big(\Psi^{(k)}\big)^{\mathrm{T}}\phi^{(k)}$. The perturbation term $X^{(k)}b^{(k)}$ accounts for pervasive hidden confounding. Pervasive hidden confounders inflate the leading singular values of $X^{(k)}$, leading to a substantial magnitude in $\|X^{(k)}b^{(k)}\|_2$. Consequently, when $Y^{(k)}$ is regressed on $X^{(k)}$ via the lasso, the resulting estimator captures the composite effect $\beta + b^{(k)}$ due to the non-negligible contribution of $X^{(k)}b^{(k)}$. This estimate thus suffers from both pervasive hidden confounding bias and high-dimensional estimation bias.

A common approach to analyzing multi-site data is to pool observations from all sites and perform a global estimation. However, this strategy faces both practical and statistical challenges. From a practical standpoint, pooling data may be infeasible due to privacy constraints, memory limitations, or prohibitive computational costs. Statistically, heterogeneous pervasive hidden confounders and site-specific heteroscedastic random errors can render naive pooling unreliable, often resulting in severely biased and inconsistent estimates. Furthermore, one-shot distributed methods remain inadequate in this setting. Since each site's data are affected by hidden confounding and high dimensionality, local estimators are inherently biased, and simple aggregation fails to correct this bias. These challenges underscore the need for distributed estimation methods that explicitly incorporate deconfounding and debiasing mechanisms.

3. Deconfounded-debiased distributed estimation

We begin by reviewing the deconfounded-debiased estimation method from (Li et al., 2025) for a single-site high-dimensional dataset with pervasive hidden confounders. The review is accompanied by a slight generalization of the precision matrix estimation technique and the corresponding individual hypothesis testing procedure. Subsequently, we extend the

method to the multi-site setting and propose a deconfounded-debiased distributed estimation method. It effectively corrects biases stemming from heterogeneous pervasive hidden confounders and high-dimensional estimation, accommodates site-specific heteroscedasticity in random errors, and supports variable selection in high-dimensional multi-site data analysis.

3.1 Deconfounded-debiased estimation at one site

In this subsection, we focus on a single-site high-dimensional dataset with pervasive hidden confounders, using site k as an illustrative example. We first review the deconfounded-debiased estimation method proposed in (Li et al., 2025) for model (2.1). We then extend the method to support multiple precision matrix estimation techniques, accommodating diverse dependency structures among the variables. Finally, we develop the corresponding individual hypothesis testing procedure.

Step 1. Deconfounding: We correct the pervasive hidden confounding bias by spectral transformation (Ćevid et al., 2020). Let $X^{(k)} = U^{(k)}\Lambda^{(k)}\big(V^{(k)}\big)^{\mathrm{T}}$ be the singular value decomposition, where $\lambda_1^{(k)} \geq \ldots \geq \lambda_{n_k}^{(k)} \geq 0$ denote the singular values. A spectral transformation matrix is constructed as

$$Q^{(k)} = U^{(k)}S^{(k)}\big(U^{(k)}\big)^{\mathrm{T}},$$
$$S^{(k)} = \mathrm{diag}\big(S_{1,1}^{(k)}, \ldots, S_{n_k,n_k}^{(k)}\big),\ S_{i,i}^{(k)} = \begin{cases} \lambda_{\lfloor \rho_k n_k \rfloor}^{(k)} / \lambda_i^{(k)} & \text{if } i \leq \lfloor \rho_k n_k \rfloor \\ 1 & \text{otherwise} \end{cases}, \tag{3.2}$$

where $\lambda_{\lfloor \rho_k n_k \rfloor}^{(k)}$ represents the ρ_k-quantile of the singular values of $X^{(k)}$ for a pre-specified parameter $\rho_k \in (0, 1)$. The choice of ρ_k requires balancing efficiency against bias correction. Specifically, smaller values of ρ_k improve estimator efficiency through more conservative shrinkage by $Q^{(k)}$. However, for effective deconfounding, $\rho_k n_k$ must substantially exceed

the number of hidden confounders q_k. In practice, $\rho_k = 0.5$ is often chosen to strike a reasonable trade-off. The spectral transformation matrix $Q^{(k)}$ shrinks the dominant singular values of $X^{(k)}$ toward the quantile $\lambda^{(k)}_{\lfloor \rho_k n_k \rfloor}$, ensuring that $\|Q^{(k)}X^{(k)}b^{(k)}\|_2$ becomes sufficiently small to be negligible. Consequently, when regressing $Q^{(k)}Y^{(k)}$ on $Q^{(k)}X^{(k)}$ via the lasso, the pervasive hidden confounding is substantially mitigated due to the diminished contribution of $Q^{(k)}X^{(k)}b^{(k)}$. The resulting deconfounded estimator is therefore given by

$$\hat{\beta}^{(k),\mathrm{dec}} = \underset{\beta\in\mathbb{R}^p}{\arg\min}\left\{ \frac{1}{2n_k}\left\|Q^{(k)}Y^{(k)} - Q^{(k)}X^{(k)}\beta\right\|_2^2 + \lambda_k \sum_{j=1}^{p} \frac{\left\|Q^{(k)}X^{(k)}_{\cdot,j}\right\|_2}{\sqrt{n_k}}|\beta_j| \right\}, \tag{3.3}$$

where λ_k is a regularization parameter selected via cross-validation.

Step 2. Debiasing: To mitigate the high-dimensional estimation bias incurred by the ℓ_1 penalty, we further debias by inverting the Karush-Kuhn-Tucker conditions of $\hat{\beta}^{(k),\mathrm{dec}}$ (Javanmard and Montanari, 2014; Van de Geer et al., 2014). Let $\hat{\Sigma}^{(k)} = \left(X^{(k)}\right)^{\mathrm{T}}\left(Q^{(k)}\right)^2 X^{(k)}/n_k$ denote the empirical covariance matrix, and let $\hat{\Theta}^{(k)} \in \mathbb{R}^{p\times p}$ be an estimator of the precision matrix $\Theta^{(k)} = \left\{\mathrm{Cov}\left(\left(Q^{(k)}X^{(k)}\right)_{i,\cdot}\right)\right\}^{-1}$. The estimation for $\Theta^{(k)}$ accommodates various precision matrix estimation techiniques, such as node-wise regression under sparsity (Van de Geer et al., 2014) or the CLIME-type estimator that do not assume sparsity (Javanmard and Montanari, 2014), to adapt to different dependence structures among the variables. The deconfounded-debiased estimator for β is then defined as

$$\hat{\beta}^{(k),\mathrm{decdeb}} = \hat{\beta}^{(k),\mathrm{dec}} + \frac{1}{n_k}\hat{\Theta}^{(k)}\left(Q^{(k)}X^{(k)}\right)^{\mathrm{T}}\left(Q^{(k)}Y^{(k)} - Q^{(k)}X^{(k)}\hat{\beta}^{(k),\mathrm{dec}}\right). \tag{3.4}$$

The error of $\hat{\beta}^{(k),\mathrm{decdeb}}$ can be decomposed as

$$\hat{\beta}^{(k),\mathrm{decdeb}} - \beta = \frac{1}{n_k}\hat{\Theta}^{(k)}\left(X^{(k)}\right)^{\mathrm{T}}\left(Q^{(k)}\right)^2\epsilon^{(k)} + \hat{\Theta}^{(k)}\hat{\Sigma}^{(k)}b^{(k)} + \left(I_p - \hat{\Theta}^{(k)}\hat{\Sigma}^{(k)}\right)\left(\hat{\beta}^{(k),\mathrm{dec}} - \beta\right).$$

The first term on the right-hand side represents the stochastic variability of $\hat{\beta}^{(k),\mathrm{decdeb}}$, while the last two terms correspond to residual biases, which are asymptotically dominated by

the stochastic variability term under suitable conditions. Thus, the asymptotic variance of $\hat{\beta}^{(k),\text{decdeb}}$ is consistently estimated by

$$\tilde{\Sigma}^{(k)} = \frac{\left(\hat{\sigma}_{\xi}^{(k)}\right)^2}{n_k^2}\hat{\Theta}^{(k)}\left(X^{(k)}\right)^{\mathrm{T}}\left(Q^{(k)}\right)^4 X^{(k)}\left(\hat{\Theta}^{(k)}\right)^{\mathrm{T}} \tag{3.5}$$

where $\left(\hat{\sigma}_{\xi}^{(k)}\right)^2$ is a consistent estimator of $\left(\sigma_{\xi}^{(k)}\right)^2$. For $j \in [p]$, the individual hypothesis testing problem regarding β_j is $H_{0,j} : \beta_j = 0$ v.s. $H_{1,j} : \beta_j \neq 0$. Based on the asymptotic normality of $\hat{\beta}_j^{(k),\text{decdeb}}$ in Theorem 1, we compute the p-value as

$$P_j^{(k)} = 2\Big\{1 - \Phi\Big(\left(\tilde{\Sigma}_{j,j}^{(k)}\right)^{-\frac{1}{2}}\left|\hat{\beta}_j^{(k),\text{decdeb}}\right|\Big)\Big\}, \tag{3.6}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. The null hypothesis $H_{0,j} : \beta_j = 0$ is rejected when $P_j^{(k)} \leq \alpha$, with $\alpha \in (0, 1)$ being a pre-specified significance level.

3.2 Deconfounded-debiased distributed estimation

Next, we extend the deconfounded-debiased estimation method to high-dimensional multi-site data with heterogeneous pervasive hidden confounders and under communication constraints that preclude full sample-level data sharing. We propose a deconfounded-debiased distributed estimation requiring two rounds of communication. It aggregates all sites' sub-vectors of local deconfounded-debiased estimators, each corresponding to predictors selected by majority voting at the central site. The detailed steps of this distributed method are as follows:

Step 1. Local estimation: Each site performs the deconfounded-debiased estimation method to address its own pervasive hidden confounders and high-dimensional predictors. Specifically, we compute the local deconfounded-debiased estimators $\left\{\hat{\beta}^{(k)}\right\}_{k=1}^{K}$ as in (3.4) and their corresponding asymptotic variance estimators $\left\{\tilde{\Sigma}^{(k)}\right\}_{k=1}^{K}$ as in (3.5). (For notational

simplicity, we omit the superscript "decdeb" from $\hat{\beta}^{(k)}$ hereafter.) Each local estimation requires the selection of the spectral transformation parameter ρ_k. Empirical studies indicate that setting $\rho_k = 0.5$ for all $k \in [K]$ is a widely adopted and robust choice that accommodates varying levels of confounding across sites (Guo et al., 2022). Next, each site conducts individual hypothesis tests and computes the p-values $\{P_j^{(k)}\}_{j=1}^p$ as in (3.6) for all $k \in [K]$. Define the set of statistically significant predictors at site k as $\hat{\mathcal{A}}^{(k)} = \{j \in [p] : P_j^{(k)} \leq \alpha\}$. These sets of significant predictors, $\{\hat{\mathcal{A}}^{(k)}\}_{k=1}^K$, are then sent to the central site.

Step 2. Majority voting at the central site: To establish consensus on significant predictors, we aggregate the sets $\{\hat{\mathcal{A}}^{(k)}\}_{k=1}^K$ by majority voting (Chen and Xie, 2014) at the central site:

$$\hat{\mathcal{A}} = \left\{ j \in [p] : \sum_{k=1}^{K} \mathbb{I}\big(j \in \hat{\mathcal{A}}^{(k)}\big) > \omega \right\},$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Predictors that appear in more than ω sites are included in the final significant predictor set $\hat{\mathcal{A}}$ for the distributed estimation. A preassigned threshold $\omega \in [0, K)$ must be chosen such that $\tilde{\Sigma}^{(k)}_{\hat{\mathcal{A}},\hat{\mathcal{A}}}$ remains positive definite for each site, ensuring matrix invertibility in the aggregation step. Moreover, ω critically controls the stringency of selection: larger values eliminate more variables but may also exclude genuinely relevant predictors. This introduces an inherent trade-off between true positive selection and false positive control. In practice, ω should be selected to ensure computational feasibility and to optimize empirical performance. Finally, the consensus set $\hat{\mathcal{A}}$ is sent back to all local sites for the subsequent estimation phase.

Step 3. Aggregation of local estimators: In the second communication round, each site transmits only the local estimation results corresponding to the significant predictors in $\hat{\mathcal{A}}$ to the central site. These include the coefficient estimators $\{\hat{\beta}^{(k)}_{\hat{\mathcal{A}}}\}_{k=1}^K$ and the asymptotic variance estimators $\{\tilde{\Sigma}^{(k)}_{\hat{\mathcal{A}},\hat{\mathcal{A}}}\}_{k=1}^K$. Following Corollary 2 and Theorem 2, we aggregate $\{\hat{\beta}^{(k)}_{\hat{\mathcal{A}}}\}_{k=1}^K$

via asymptotic confidence distributions (Tang et al., 2020; Wang, 2021) to construct the distributed estimator. The site-specific asymptotic confidence density function for $\beta_{\hat{\mathcal{A}}}$ is given by:

$$\hat{f}_k(\beta_{\hat{\mathcal{A}}}) = \Big(\frac{1}{\sqrt{2\pi}}\Big)^{|\hat{\mathcal{A}}|} I_{|\hat{\mathcal{A}}|}^{-\frac{1}{2}} \exp\Big\{-\frac{1}{2}\big(\hat{\beta}_{\hat{\mathcal{A}}}^{(k)} - \beta_{\hat{\mathcal{A}}}\big)^{\mathrm{T}}\Big(\dot{\Sigma}_{\hat{\mathcal{A}},\hat{\mathcal{A}}}^{(k)}\Big)^{-1}\big(\hat{\beta}_{\hat{\mathcal{A}}}^{(k)} - \beta_{\hat{\mathcal{A}}}\big)\Big\},$$

where $\dot{\Sigma}^{(k)} = \big(\sigma_\xi^{(k)}\big)^2\hat{\Theta}^{(k)}\big(X^{(k)}\big)^{\mathrm{T}}\big(Q^{(k)}\big)^4 X^{(k)}\big(\hat{\Theta}^{(k)}\big)^{\mathrm{T}}/n_k^2$. Combining the K confidence density functions and simplifying, we obtain a least-squares-type loss function:

$$L(\beta_{\hat{\mathcal{A}}}) = \sum_{k=1}^{K}\big(\hat{\beta}_{\hat{\mathcal{A}}}^{(k)} - \beta_{\hat{\mathcal{A}}}\big)^{\mathrm{T}}\Big(\dot{\Sigma}_{\hat{\mathcal{A}},\hat{\mathcal{A}}}^{(k)}\Big)^{-1}\big(\hat{\beta}_{\hat{\mathcal{A}}}^{(k)} - \beta_{\hat{\mathcal{A}}}\big).$$

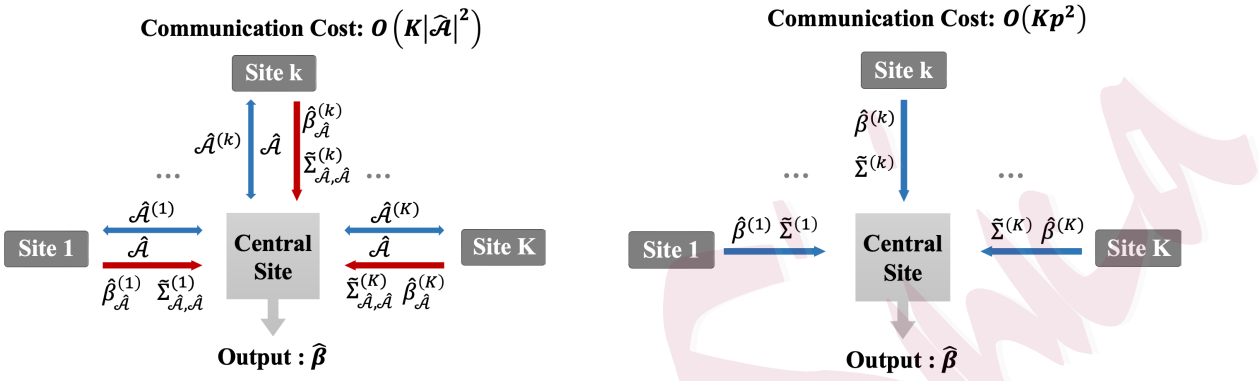
Minimizing $L(\beta_{\hat{\mathcal{A}}})$ and substituting $\dot{\Sigma}_{\hat{\mathcal{A}},\hat{\mathcal{A}}}^{(k)}$ with its consistent estimator $\tilde{\Sigma}_{\hat{\mathcal{A}},\hat{\mathcal{A}}}^{(k)}$ yields the deconfounded-debiased distributed estimator:

$$\hat{\beta}_{\hat{\mathcal{A}}} = \Big\{\sum_{k=1}^{K}\Big(\tilde{\Sigma}_{\hat{\mathcal{A}},\hat{\mathcal{A}}}^{(k)}\Big)^{-1}\Big\}^{-1}\sum_{k=1}^{K}\Big(\tilde{\Sigma}_{\hat{\mathcal{A}},\hat{\mathcal{A}}}^{(k)}\Big)^{-1}\hat{\beta}_{\hat{\mathcal{A}}}^{(k)}, \quad \hat{\beta}_{-\hat{\mathcal{A}}} = 0. \tag{3.7}$$

The proposed deconfounded-debiased distributed estimation effectively corrects biases arising from heterogeneous pervasive hidden confounders and high-dimensional estimation, while accommodating site-specific heteroscedasticity of the random errors. Furthermore, by incorporating majority voting, the method achieves sparsity and enables variable selection in high-dimensional multi-site data analysis.

Remark 1. The proposed distributed estimation involves two communication rounds, as illustrated in Figure 1(a). While a one-round alternative exists (Figure 1(b)), we recommend the two-round design in the high-dimensional setting. Although requiring an additional round to transmit sets $\{\hat{\mathcal{A}}^{(k)}\}_{k=1}^{K}$ and $\hat{\mathcal{A}}$, it significantly reduces communication costs. Majority voting ensures that the number of significant predictors $|\hat{\mathcal{A}}|$ is substantially smaller than the original dimensionality p. Consequently, local estimators sent to the central site decrease from p-dimensional vectors to $|\hat{\mathcal{A}}|$-dimensional vectors and corresponding asymptotic

covariance matrices shrink from $p \times p$ to $|\hat{\mathcal{A}}| \times |\hat{\mathcal{A}}|$. The communication cost for the one-round procedure scales as $O(Kp^2)$, whereas the two-round procedure achieves substantial savings with a reduced order of $O(K|\hat{\mathcal{A}}|^2)$.



(a) Two-round procedure (b) One-round procedure

Figure 1: Communication procedures of the deconfounded-debiased distributed estimation. Blue arrows represent the first round of communication, and red arrows represent the second round.

Remark 2. In the proposed distributed method, neither the lasso (Tibshirani, 1996) nor the deconfounded lasso (Ćevid et al., 2020) can be directly applied for local estimation, as both lack the well-defined asymptotic variance properties required by our approach. An alternative distributed approach based on majority voting (Chen and Xie, 2014) can be adapted to aggregate local lasso or deconfounded lasso estimates using Hessian-weighted averaging, which are precisely the two distributed methods compared in our simulation experiments. However, both alternatives suffer from critical limitations: (i) Chen and Xie (2014)'s distributed estimation incorporating local deconfounded lasso only corrects bias due to heterogeneous pervasive confounders, whereas (ii) Chen and Xie (2014)'s distributed estimation incorporating local lasso offers no bias correction capability.

Remark 3. Recent works that jointly achieve deconfounding and debiasing, such as (Guo et al., 2022) and (Sun et al., 2024), both rely on the sparse precision matrix assumption, and the latter additionally requires consistent estimation of the unknown number of hidden confounders. These requirements limit their applicability in distributed settings. Furthermore, neither of these approaches can be directly used for local estimation in our proposed distributed framework. While they establish asymptotic normality for individual coefficient estimates, merely aggregating local estimates using the asymptotic confidence density function $\hat{f}_k(\beta_j)$ for each β_j would yield $|\hat{\mathcal{A}}|$ distributed estimators $\hat{\beta}_j$ for $j \in \hat{\mathcal{A}}$, each asymptotically normal in isolation. However, this does not guarantee the joint asymptotic normality of the vector $\hat{\beta}_{\hat{\mathcal{A}}}$ (Corollary 2), a property essential for constructing a distributed estimator across all sites.

Remark 4. While the proposed distributed estimation method mitigates information leakage via two-round communication, majority voting, and deconfounding techniques, it cannot entirely eliminate the potential for such leakage. A detailed discussion is provided in Section S8 of Supplementary Materials.

4. Theoretical properties

First, we establish the asymptotic normality of the local deconfounded-debiased estimators, thereby completing the theoretical analysis absent in (Li et al., 2025). We also provide the theoretical properties of the individual hypothesis tests. Furthermore, we prove variable selection consistency and derive the asymptotic normality of the deconfounded-debiased distributed estimator. We also analyze the finite-sample property on variable selection.

4.1 Theoretical properties of local deconfounded-debiased estimators

We now formally establish the asymptotic normality of the local deconfounded-debiased estimators under a set of technical assumptions.

Assumption 1. *$E_{i,\cdot}^{(k)}$ is sub-Gaussian with bounded sub-Gaussian norm $\max_{k\in[K]} \left\|E_{i,\cdot}^{(k)}\right\|_{\psi_2} \leq C_E$ for some constant $C_E > 0$. $\xi_i^{(k)}$ is sub-Gaussian with bounded sub-Gaussian norm $\max_{k\in[K]} \left\|\xi_i^{(k)}\right\|_{\psi_2} \leq C_\xi$ for some constant $C_\xi > 0$. For all $j \in [p]$, $X_{i,j}^{(k)}$ is sub-Gaussian with bounded sub-Gaussian norm $\max_{j\in[p]} \left\|X_{i,j}^{(k)}\right\|_{\psi_2} \leq M_k$, where $1 \lesssim M_k \lesssim \sqrt{q_k \log p}$.*

Assumption 2. *There exist positive constants $c_{\min} > 0$ and $c_{\max} > 0$ such that $0 < c_{\min} \leq \min_{k\in[K]} \lambda_{\min}^{(k)}\left(\Sigma_E^{(k)}\right) \leq \max_{k\in[K]} \lambda_{\max}^{(k)}\left(\Sigma_E^{(k)}\right) \leq c_{\max} < \infty$.*

Assumption 3. *We have $\left\|\phi^{(k)}\right\|_2^2 \lesssim q_k\sqrt{\log p}$.*

Assumption 4. *Let $s = \|\beta\|_0$. For some constant $C > 0$, there exists a constant $\tau_* > 0$ such that the restricted eigenvalue conditions*

$$\mathrm{RE}\left(\hat{\Sigma}^{(k)}\right) = \min_{\substack{B\subseteq[p]\\|B|\leq s}} \min_{\substack{x\in\mathbb{R}^p\\ \|x_{B^\complement}\|_1 \leq C M_k \|x_B\|_1}} \frac{x^{\mathrm{T}}\hat{\Sigma}^{(k)}x}{\|x\|_2^2} \geq \tau_*, \ k \in [K]$$

hold with probability tending to 1.

Assumption 5. *There exists $\Delta_{\max} > 0$ such that $\max_{k\in[K]} \left\|I_p - \hat{\Theta}^{(k)}\hat{\Sigma}^{(k)}\right\|_{\max} \lesssim_{\mathrm{P}} \Delta_{\max}$.*

Assumption 6. *The q_k-th largest singular value of $\Psi^{(k)} \in \mathbb{R}^{q_k\times p}$ satisfies*

$$\lambda_{q_k}^{(k)}\left(\Psi^{(k)}\right) \gg \max\left\{\sqrt{n_k q_k}(\log p)^{\frac{1}{4}}, \sqrt{\Delta_{\max} p q_k} n_k^{\frac{1}{4}} (\log p)^{\frac{1}{4}}\right\}.$$

Remark 5. In Assumption 1, the sub-Gaussian conditions rule out heavy-tailed idiosyncratic errors $E^{(k)}$, random errors $\xi^{(k)}$, and predictors $X^{(k)}$ at each site. Assumption 2 is mild: the eigenvalue condition on $\left\{\Sigma_E^{(k)}\right\}_{k=1}^K$ ensures well-conditioned covariance matrices across

all sites. Assumption 3 restricts the ℓ_2 norm of the hidden confounding coefficient $\phi^{(k)}$ at each site to grow at most at a controlled rate. The restricted eigenvalue conditions in Assumption 4 are widely used in high-dimensional sparse linear regression (Bühlmann and Van De Geer, 2011; Guo et al., 2022; Sun et al., 2024). Their verification follows from the final assumption in Guo et al. (2022) and is therefore omitted here. Assumption 5 specifies the required error rates for the precision matrix estimators at all sites. This condition is mild and can be satisfied by several common precision matrix estimation methods (Van de Geer et al., 2014; Javanmard and Montanari, 2014; Fan et al., 2024). Assumption 6 is a pervasive hidden confounding assumption commonly adopted in factor model literature (Fan et al., 2018; Guo et al., 2022; Fan et al., 2024). It implies that a substantial fraction of predictors are influenced by hidden confounders at each site.

In Theorem 1, we establish the asymptotic normality for each inidividual local deconfounded-debiased estimator $\hat{\beta}_j^{(k)}$ for $j \in [p]$ and $k \in [K]$. The parameters involved in the local estimation include ρ_k in (3.2) and λ_k in (3.3). The conditions on ρ_k, namely $(q_k+1)/n_k$, specifies the minimal amount of shrinkage needed to sufficiently suppress the hidden confounding at each site. The requirement on the regularization parameters λ_k aligns with that given in (Guo et al., 2022).

Theorem 1. *Under Assumptions 1–6, if* $(q_k+1)/n_k \le \rho_k < 1$, $\lambda_k \asymp \sqrt{(\log p)/n_k}$, $\Delta_{\max} < 1$ *and* $s\Delta_{\max} M_k^2 \sqrt{\log p} \to 0$, *then for all* $j \in [p]$ *and* $k \in [K]$, *we have*

$$\left(\dot{\Sigma}_{j,j}^{(k)}\right)^{-\frac{1}{2}} \left(\hat{\beta}_j^{(k)} - \beta_j\right) \xrightarrow{d} \mathcal{N}(0,1).$$

Based on Theorem 1, we establish the theoretical properties of the individual hypothesis tests at each site in Corollary 1. Assumption 7 states that the estimator $\hat{\sigma}_\xi^{(k)}$ is consistent for $\sigma_\xi^{(k)}$ at every site. This condition is satisfied by a number of existing methods,

such as the scaled lasso (Sun and Zhang, 2012), the organic lasso (Yu and Bien, 2019), and estimation based on the residuals (Guo et al., 2022). Define the null and alternative hypothesis parameter spaces for β_j as $\mathcal{H}_{0,j}(s) = \{\beta \in \mathbb{R}^p, \|\beta\|_0 \leq s, \beta_j = 0\}$ and $\mathcal{H}_{1,j}(s,\nu) = \{\beta \in \mathbb{R}^p, \|\beta\|_0 \leq s, |\beta_j| \geq \nu\}$ for some $\nu > 0$. Corollary 1 provides the asymptotic validity of the individual hypothesis tests and derives an asymptotic lower bound on their power, given by $G\big(\alpha, \Delta_n^{(k)}\big)$ for each $k \in [K]$. For any fixed $\alpha \in (0,1)$, the function $G(\alpha, x)$ is continuous and monotonically increasing with respect to x. Consequently, the power of each test increases as $\Delta_n^{(k)}$ grows. If $\Delta_n^{(k)} \to \infty$ as $n_k \to \infty$, the individual individual tests at site k achieve asymptotic power one.

Assumption 7. *We have* $\hat{\sigma}_\xi^{(k)}/\sigma_\xi^{(k)} \xrightarrow{p} 1$.

Corollary 1. *Under Assumptions 1–7, all conditions of Theorem 1 hold.*

- *(Validity). For any given significance level* $\alpha \in (0,1)$*, we have for all* $j \in [p]$ *and* $k \in [K]$

$$\lim_{n_k \to \infty} \sup_{\beta_j \in \mathcal{H}_{0,j}(s)} \mathsf{P}_\beta\big(P_j^{(k)} \leq \alpha\big) \leq \alpha.$$

- *(Power). For any given significance level* $\alpha \in (0,1)$*, if* $\Delta_n^{(k)} \asymp \sqrt{n_k}\nu$*, we have for all* $j \in [p]$ *and* $k \in [K]$

$$\lim_{n_k \to \infty} \inf_{\beta_j \in \mathcal{H}_{1,j}(s,\nu)} \frac{\mathsf{P}_\beta\big(P_j^{(k)} \leq \alpha\big)}{G\big(\alpha, \Delta_n^{(k)}\big)} \geq 1,$$

where, for $x \in \mathbb{R}_+$*, the function* $G(\alpha, x)$ *is defined as follows:*

$$G(\alpha, x) = 1 - \Phi\Big(\Phi^{-1}\Big(1 - \frac{\alpha}{2}\Big) - x\Big) + \Phi\Big(-\Phi^{-1}\Big(1 - \frac{\alpha}{2}\Big) - x\Big).$$

Let $\mathcal{A} = \{j \in [p] : \beta_j \neq 0\}$ and $s = |\mathcal{A}|$. The predictors indexed by $\mathcal{A}$ are referred to as the true active predictors. To establish the asymptotic normality of the local deconfounded-debiased estimator vector for the active predictors, we first require that $\dot{\Sigma}_{\mathcal{A},\mathcal{A}}^{(k)}$ be invertible

in Corollary 2. To this end, we provide Proposition 1, which guarantees that $\dot{\Sigma}^{(k)}_{\mathcal{A},\mathcal{A}}$ is positive definite with probability tending to 1 as $n_k \to \infty$ for each $k \in [K]$, thereby ensuring the existence of its inverse. Based on Proposition 1, we extend Theorem 1 to the local deconfounded-debiased estimator vector for the active predictors, $\hat{\beta}^{(k)}_{\mathcal{A}}$ for $k \in [K]$, and establish the joint asymptotic normality in Corollary 2. Compared with Theorem 1, which deals with individual coefficient estimators, the joint asymptotic normality in Corollary 2 requires stronger conditions: a stricter pervasive hidden confounding assumption (Assumption 8) and a tighter sparsity constraint on β.

Proposition 1. *Under Assumptions 1–6, if* $(q_k+1)/n_k \le \rho_k < 1$ *and* $s \lesssim M_k^{\frac{2}{3}}$, *for* $k \in [K]$, *we have the positive definite* $\dot{\Sigma}^{(k)}_{\mathcal{A},\mathcal{A}}$ *with probability tending to* 1.

Assumption 8. *The* q_k*-th largest singular value of* $\Psi^{(k)} \in \mathbb{R}^{q_k \times p}$ *satisfies*

$$\lambda^{(k)}_{q_k}\big(\Psi^{(k)}\big) \gg \max\Big\{\sqrt{n_k q_k}(\log p)^{\frac{1}{4}}, \sqrt{\Delta_{\max} p q_k}\, n_k^{\frac{1}{4}} (\log p)^{\frac{1}{4}} s^{\frac{1}{4}}\Big\}.$$

Corollary 2. *Under Assumptions 1–5 and 8, if* $(q_k+1)/n_k \le \rho_k < 1$, $\lambda_k \asymp \sqrt{(\log p)/n_k}$, $s\Delta_{\max} < 1$, $s \lesssim M_k^{\frac{2}{3}}$, $s \ll n_k$, *and* $s^{\frac{3}{2}}\Delta_{\max} M_k^2 \sqrt{\log p} \to 0$, *then for* $k \in [K]$ *and any* $u \in \mathbb{R}^s$ *with* $\|u\|_2 = 1$, *we have*

$$u^{\mathrm{T}}\big(\dot{\Sigma}^{(k)}_{\mathcal{A},\mathcal{A}}\big)^{-\frac{1}{2}}\big(\hat{\beta}^{(k)}_{\mathcal{A}} - \beta_{\mathcal{A}}\big) \xrightarrow{d} \mathcal{N}(0,1).$$

4.2 Theoretical properties of the deconfounded-debiased distributed estimator

When the number of sites K and the per-site sample sizes n_k tend to infinity, the deconfounded-debiased distributed estimation achieves variable selection consistency (Theorem 2). The condition on the significant level α is mild: ω/K is a constant for $\omega \in [0, K) \to \infty$ as $K \to \infty$. All remaining conditions are the same as those stated in Theorem 1.

Theorem 2. *Under Assumptions 1–7, if* $(q_k+1)/n_k \le \rho_k < 1$, $\lambda_k \asymp \sqrt{(\log p)/n_k}$, $\Delta_{\max} < 1$, $s\Delta_{\max}\sqrt{\log p}(\max_{k\in[K]} M_k^2) \to 0$, *and* $\alpha < \omega/K$, *we have*

$$\lim_{\substack{n_k\to\infty \\ K\to\infty}} \mathsf{P}(\hat{\mathcal{A}} = \mathcal{A}) = 1.$$

Remark 6. The deconfounded-debiased distributed estimator outperforms its global counterpart estimated by the deconfounded-debiased estimation method (introduced in Section 3.1) in variable selection. First, the global estimator is intrinsically non-sparse and offers no built-in variable selection. Second, even when variable selection is attempted via hypothesis testing as in (3.6), the resulting procedure suffers from asymptotically non-vanishing type I error (Corollary 1), which can produce an unreliable set of selected predictors that may substantially differ from the true active set. In contrast, the distributed method delivers sparse solutions directly and achieves asymptotic selection consistency (Theorem 2).

Based on Theorem 2 and Proposition 1, we establish the asymptotic normality of the deconfounded-debiased distributed estimator for the active predictors $\hat{\beta}_{\mathcal{A}}$ in Theorem 3. Moving from the local to the distributed setting, the pervasive hidden confounding assumption (Assumption 9) and the sparsity constraint on β in Theorem 3 are stronger than those in Theorem 1. The condition on the sample sizes suggests that no single local site dominates the total sample size of the multi-site data.

Assumption 9. *The minimum of the* q_k*-th largest singular values of* $\Psi^{(k)} \in \mathbb{R}^{q_k\times p}$ *across* K *sites satisfies*

$$\min_{k\in[K]} \lambda_{q_k}^{(k)}\big(\Psi^{(k)}\big) \gg \max\left\{\left(\sum_{k=1}^{K} n_k\sqrt{q_k}\right)N^{-\frac{1}{2}}(\log p)^{\frac{1}{4}}, \sqrt{\Delta_{\max} p}\sqrt{\sum_{k=1}^{K} n_k q_k N^{-\frac{1}{4}}(\log p)^{\frac{1}{4}} s^{\frac{1}{4}}}\right\}.$$

Theorem 3. *Under Assumptions 1–5, 7 and 9, if* $(q_k+1)/n_k \le \rho_k < 1$, $\lambda_k \asymp \sqrt{(\log p)/n_k}$, $s\Delta_{\max} < 1$, $s \lesssim \min_{k\in[K]} M_k^{\frac{2}{3}}$, $s^{\frac{3}{2}}\Delta_{\max}\sqrt{\log p}N^{-\frac{1}{2}}(\sum_{k=1}^{K}\sqrt{n_k}M_k^2) \to 0$, $\max_{k\in[K]} n_k/N \to 0$,

and $\alpha < \omega/K$, *for any* $D \in \mathbb{R}^{l\times s}$ *where* $DD^{\mathrm{T}} \in \mathbb{R}^{l\times l}$ *is positive definite, we have*

$$D\left\{\sum_{k=1}^{K}\left(\dot{\Sigma}^{(k)}_{\mathcal{A},\mathcal{A}}\right)^{-1}\right\}^{\frac{1}{2}}\left(\hat{\beta}_{\mathcal{A}}-\beta_{\mathcal{A}}\right)\xrightarrow{d}\mathcal{N}(\mathbf{0},I_l).$$

Next, we analyze the influence of ω on variable selection. Similar to Theorem 3 in Chen and Xie (2014), we give the finite-sample properties about false positive rate (FPR) and true positive rate (TPR). Theorem 4 establishes an upper bound on the expected FPR and a lower bound on the expected TPR. By the trade-off between the upper and lower bounds, the choice of ω should balance FPR and TPR in variable selection for the distributed estimation in practice.

Assumption 10. *The distributions of* $\{\mathbb{I}(j\in\hat{\mathcal{A}}^{(k)}) : j\in\mathcal{A}\}$ *and* $\{\mathbb{I}(j\in\hat{\mathcal{A}}^{(k)}) : j\in\mathcal{A}^{\complement}\}$ *are exchangeable for all* $k\in[K]$, *and* $\mathsf{E}(|\mathcal{A}\cap\hat{\mathcal{A}}^{(k)}|)/\mathsf{E}(|\mathcal{A}^{\complement}\cap\hat{\mathcal{A}}^{(k)}|)\geq|\mathcal{A}|/|\mathcal{A}^{\complement}|$.

Theorem 4. *Let* $s^*=\max_{k\in[K]}\bar{s}_k$ *and* $s_*=\min_{k\in[K]}\bar{s}_k$ *where* $s_k=\mathsf{E}(|\hat{\mathcal{A}}^{(k)}|)$. *Under Assumption 10, if* $\omega\geq Ks^*/p-1$, *we have*

$$\mathsf{E}(\mathrm{FPR})=\mathsf{E}\Big(\frac{|\mathcal{A}^{\complement}\cap\hat{\mathcal{A}}|}{|\mathcal{A}^{\complement}|}\Big)\leq 1-F\big(\omega\mid K,\frac{s^*}{p}\big),$$
$$\mathsf{E}(\mathrm{TPR})=\mathsf{E}\Big(\frac{|\mathcal{A}\cap\hat{\mathcal{A}}|}{|\mathcal{A}|}\Big)\geq 1-F\big(\omega\mid K,\frac{s_*}{p}\big),$$

where $F(\cdot\mid K,p')$ *is the cumulative distribution function of the binomial distribution with* K *trials and success probability* p'.

5. Simulation experiments

We evaluate the finite-sample performance of the deconfounded-debiased distributed estimation method by simulation experiments. The following competing estimation methods are considered:

- Deconfounded-debiased distributed estimation with $\alpha = 0.01$ (DDD);
- Chen and Xie (2014)'s distributed estimation incorporating deconfounded lasso (Ćevid et al., 2020) for local estimation (DLD);
- Chen and Xie (2014)'s distributed estimation incorporating lasso (Tibshirani, 1996) for local estimation (LD);
- Global deconfounded-debiased estimation followed by hypothesis testing with $\alpha = 0.01$ (DDG);
- Global deconfounded lasso estimation (Ćevid et al., 2020) (DLG);
- Global lasso estimation (Tibshirani, 1996) (LG).

All tuning parameters are determined via cross-validation. For DDD and DLD involving the spectral transformation matrix, we set $\rho_k = 0.5$ for local estimation at each site, sufficient to mitigate various pervasive hidden confounding (Guo et al., 2022). The same setting ($\rho = 0.5$) is applied to DDG and DLG. In DDD, the site-specific precision matrix $\Theta^{(k)}$ is estimated via node-wise regression (Van de Geer et al., 2014), and the noise variance $(\sigma_\xi^{(k)})^2$ is estimated by the residuals (Guo et al., 2022). Following the condition $\alpha < \omega/K$ in Theorem 2, we fix the local significance level at $\alpha = 0.01$. DDG adopts identical procedures for precision matrix and noise variance estimation. Since the global deconfounded-debiased estimation yields non-sparse estimates, we conduct additional hypothesis testing to achieve sparsity, maintaining the same significance level ($\alpha = 0.01$). Finally, three distributed estimation methods involve the threshold ω, whose influence on performance will be examined via simulations.

The multi-site data are generated from the model (2.1), where $H_{i,l}^{(k)}$, $E_{i,j}^{(k)}$, $\Psi_{l,j}^{(k)}$, and $\phi_l^{(k)}$ are independently sampled from standard normal distribution $\mathcal{N}(0,1)$ for $i = 1, \ldots, n_k$, $j = 1, \ldots, p$, $l = 1, \ldots, q_k$, and $k = 1, \ldots, K$. We define $\beta = (1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0, \ldots, 0)^{\mathrm{T}}$. We consider two distinct scenarios in the high-dimensional setting:

Scenario 1 (Homogeneous). All sites share identical simulation configurations: (1) equal

sample size, and (2) the same number of pervasive hidden confounders ($q_k = 3$). The random errors $\xi_i^{(k)}$ follow $\mathcal{N}(0, 1)$.

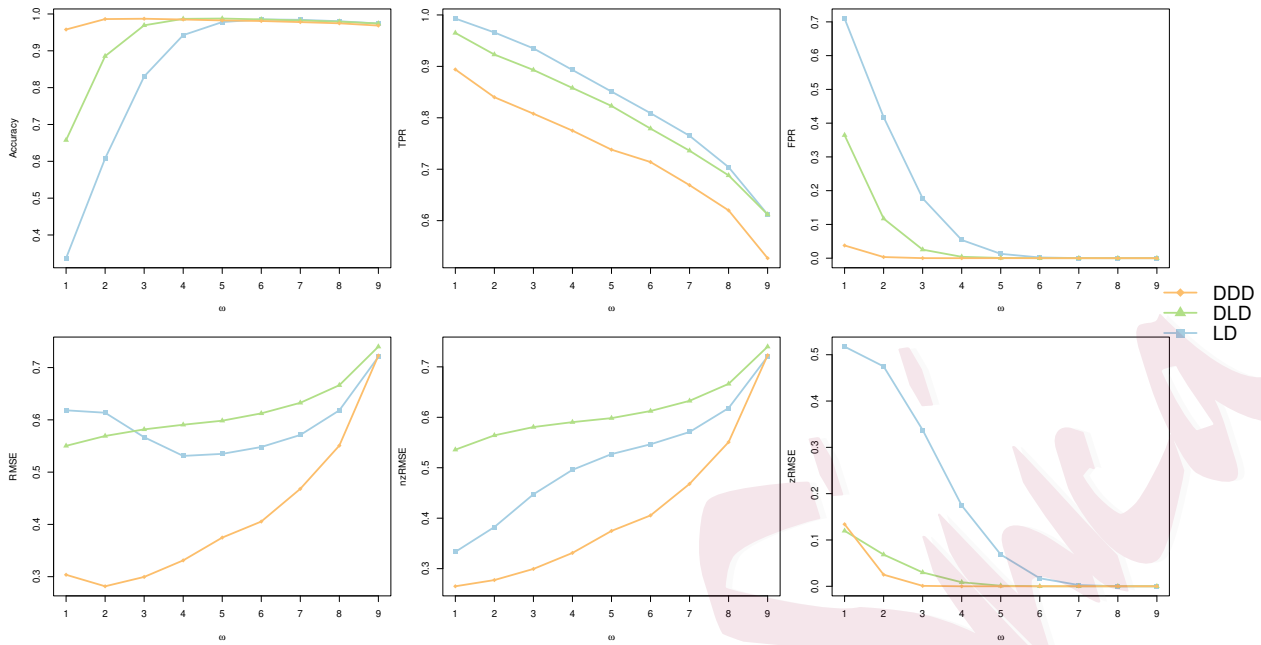
Scenario 2 (Heterogeneous). Local sites feature different simulation configurations: (1) unbalanced sample sizes, and (2) differing numbers of hidden confounders. The random errors $\xi_i^{(k)}$ follow $\mathcal{N}(0, 1)$ at half of the sites and $\mathcal{N}(0, 0.5^2)$ at the remaining sites.

Each experiment is repeated 100 times. The performance in variable selection is comprehensively evaluated using the accuracy, true positive rate (TPR), false positive rate (FPR). To assess estimation accuracy, we compute the root mean square error for all coefficients (RMSE), non-zero coefficients (nzRMSE), and zero coefficients (zRMSE).

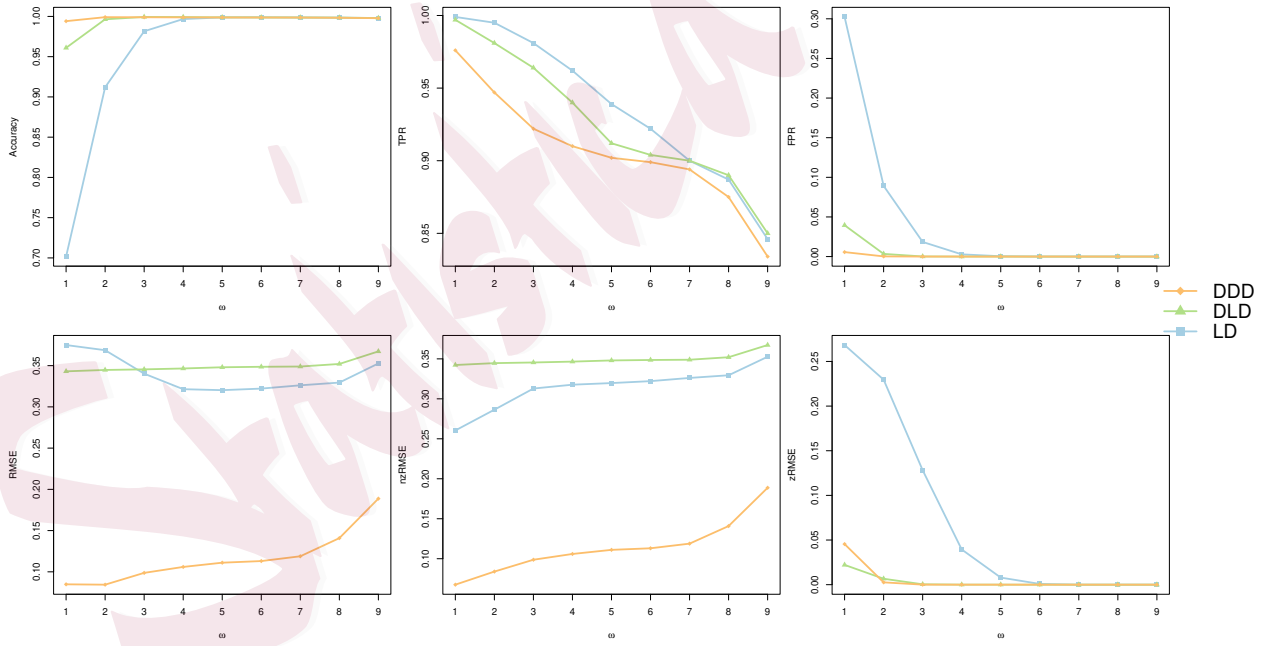
5.1 Impact of the parameter ω

In this part, we investigate the impact of the threshold ω on variable selection and estimation accuracy across three distributed estimation methods under both homogeneous and heterogeneous scenarios. We fix the number of sites at $K = 10$ in simulation experiments, and the threshold ω takes integer values from 1 to 9. We analyze the following four cases: (1) homogeneous small-sample case; (2) homogeneous large-sample case; (3) heterogeneous small-sample case; (4) heterogeneous large-sample case. Detailed simulation configurations and results are provided in Figures 2 and 3.

The threshold ω exhibits consistent effects on variable selection and coefficient estimation across four experimental cases. Specifically, as ω increases, these three methods tend to incorrectly estimate non-zero coefficients to zero, leading to over-selection. This explains the decline in both TPR and FPR, resulting in a inverted U-shaped trend for accuracy (first sharply increasing, then slightly decreasing), which aligns with the theoretical property in Theorem 4. Furthermore, over-selection induced by larger ω values imposes an increased

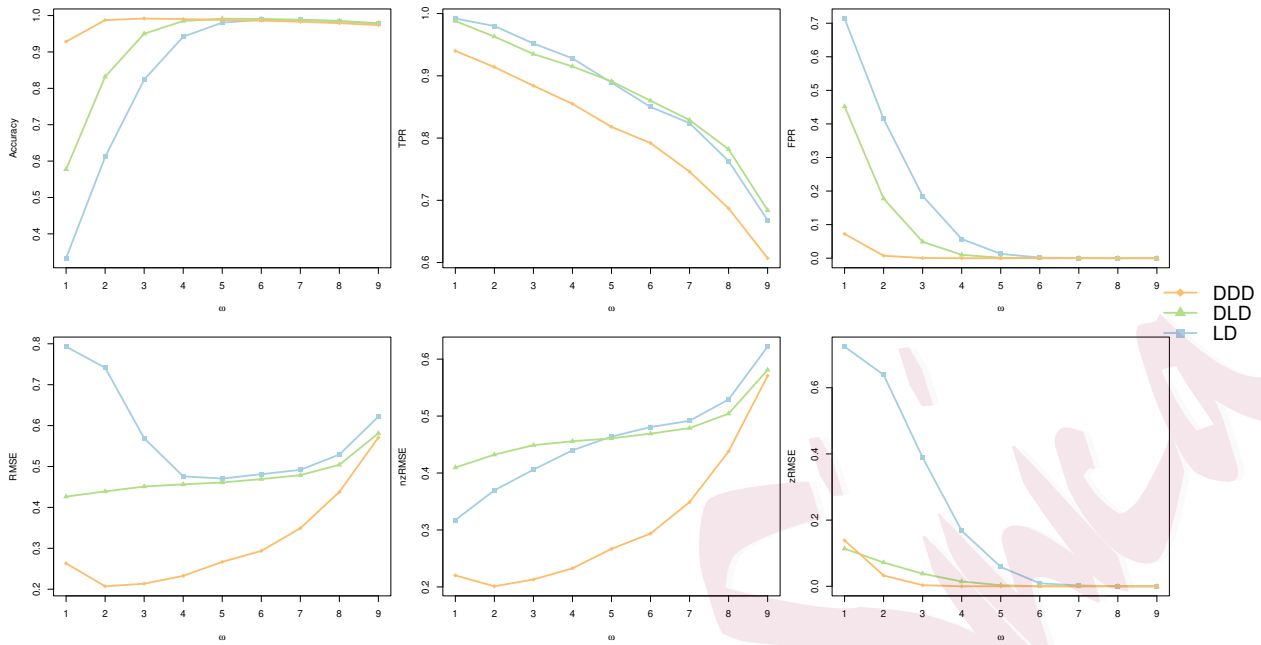


(a) Homogeneous small-sample case: Each local site contains $p = 150$ predictors and $n_k = 100$ samples for $k = 1, \ldots, 10$.

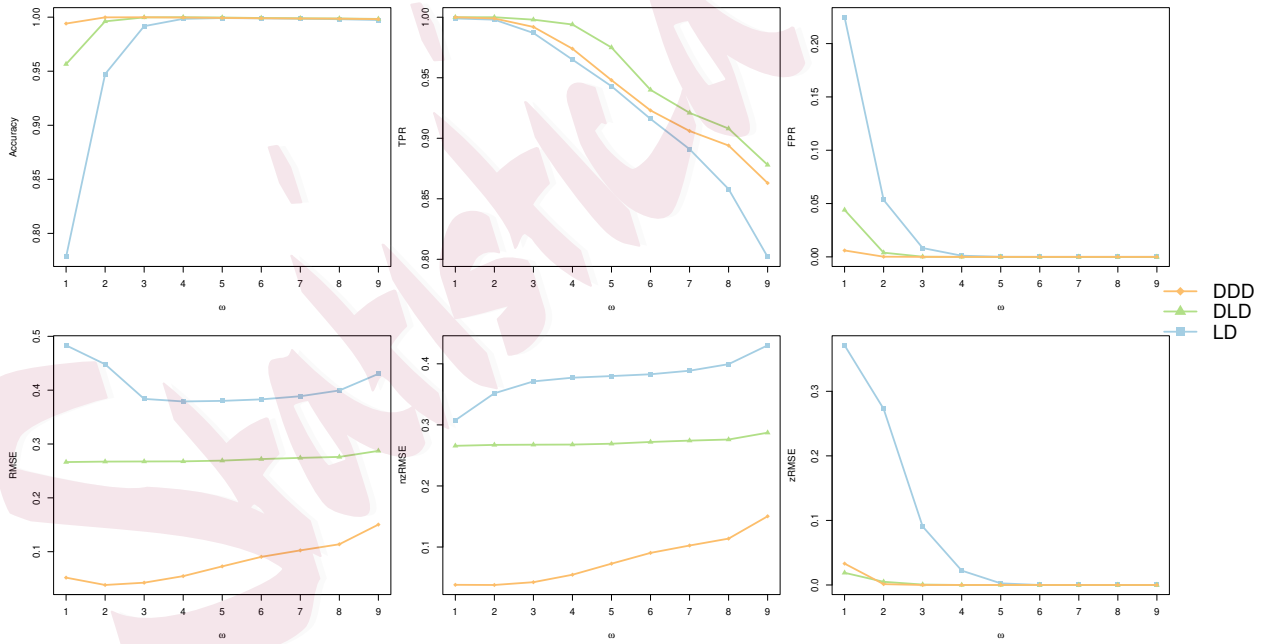


(b) Homogeneous large-sample case: Each local site contains $p = 750$ predictors and $n_k = 500$ samples for $k = 1, \ldots, 10$.

Figure 2: Influence of ω on variable selection and coefficient estimation in the homogeneous scenario. Across all 10 local sites, the number of hidden confounders is $q_k = 3$ for $k = 1, \ldots, 10$.



(a) Heterogeneous small-sample case: Each local site contains $p = 150$ predictors. The local sample sizes n_k are assigned as follows: $n_1, n_2 = 80$, $n_3, n_4 = 90$, $n_5, n_6 = 100$, $n_7, n_8 = 110$, and $n_9, n_{10} = 120$.



(b) Heterogeneous large-sample case: Each local site contains $p = 750$ predictors. The local sample sizes n_k are assigned as follows: $n_1, n_2 = 400$, $n_3, n_4 = 450$, $n_5, n_6 = 500$, $n_7, n_8 = 550$, and $n_9, n_{10} = 600$.

Figure 3: Influence of ω on variable selection and coefficient estimation in the heterogeneous scenario. Across all 10 local sites, site k has $q_k = k$ hidden confounders for $k = 1, \ldots, 10$.

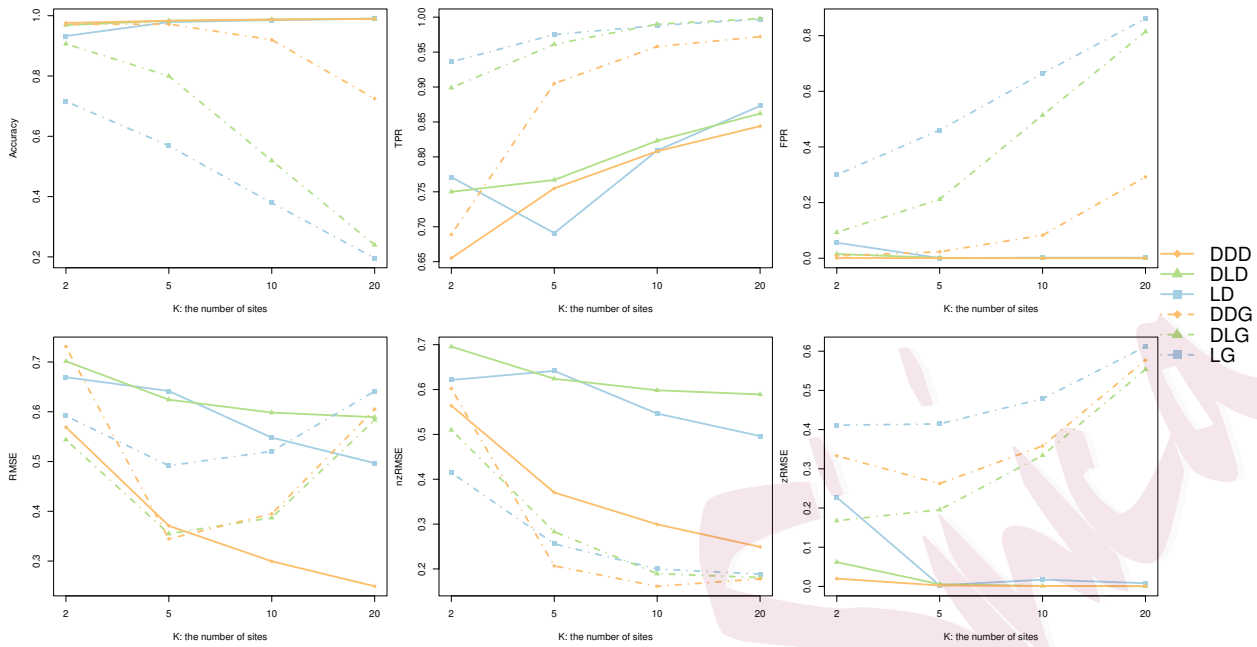
nzRMSE for non-zero coefficients and a reduced zRMSE for zero coefficients. Due to the trade-off between nzRMSE and zRMSE, the RMSE for all coefficients may follow either a U-shaped pattern (as seen in DDD and LD) or an increasing trend (as seen in DLD).

In summary, an optimal threshold ω should simultaneously balance TPR against FPR and nzRMSE versus zRMSE, thereby ensuring robust performance in both variable selection and coefficient estimation. Beyond this theoretical balance, the choice of ω in practice should also consider the specific application goals. For instance, if precise variable identification is prioritized (e.g., in biomarker discovery or causal inference), a larger ω may be preferred.

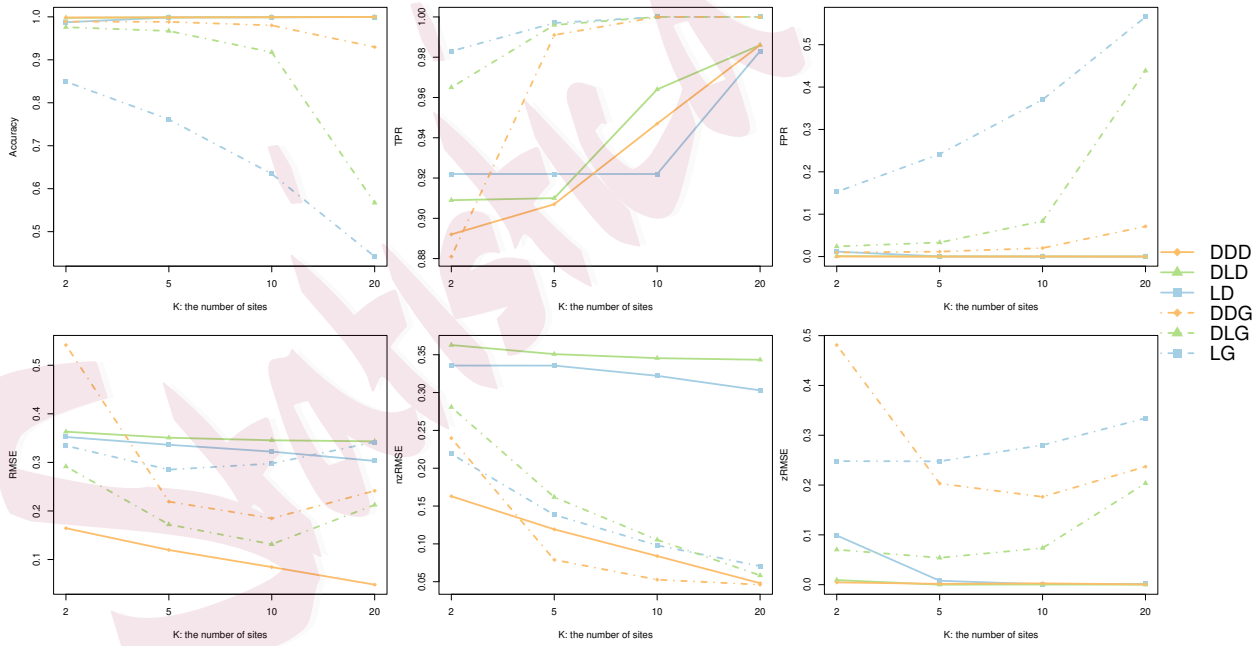
5.2 Varying local sample size and the number of sites

Under both homogeneous and heterogeneous scenarios, we assess the effects of the local sample size n_k and the number of sites K on variable selection and coefficient estimation across all six estimation methods. We consider the following four cases: (1) homogeneous small-sample case; (2) homogeneous large-sample case; (3) heterogeneous small-sample case; (4) heterogeneous large-sample case. For each case, we vary the number of sites $K = 2, 5, 10, 20$. For each K, the threshold ω takes integer values from 1 to $K-1$. Since ω impacts the distributed estimation performance, we determine its optimal value by maximizing the accuracy. All simulation results correspond to their optimal ω values, which are systematically provided in Table 1 (Supplementary Materials, Section S5). Detailed simulation configurations and results are provided in Figures 4 and 5.

In all four simulation cases, DDD demonstrates significant improvements in both variable selection and coefficient estimation as the number of sites K increases. Specifically, TPR increases, FPR decreases, and accuracy converges to 1; nzRMSE, zRMSE, and RMSE all decrease. Furthermore, comparative analysis between small-sample and large-sample cases

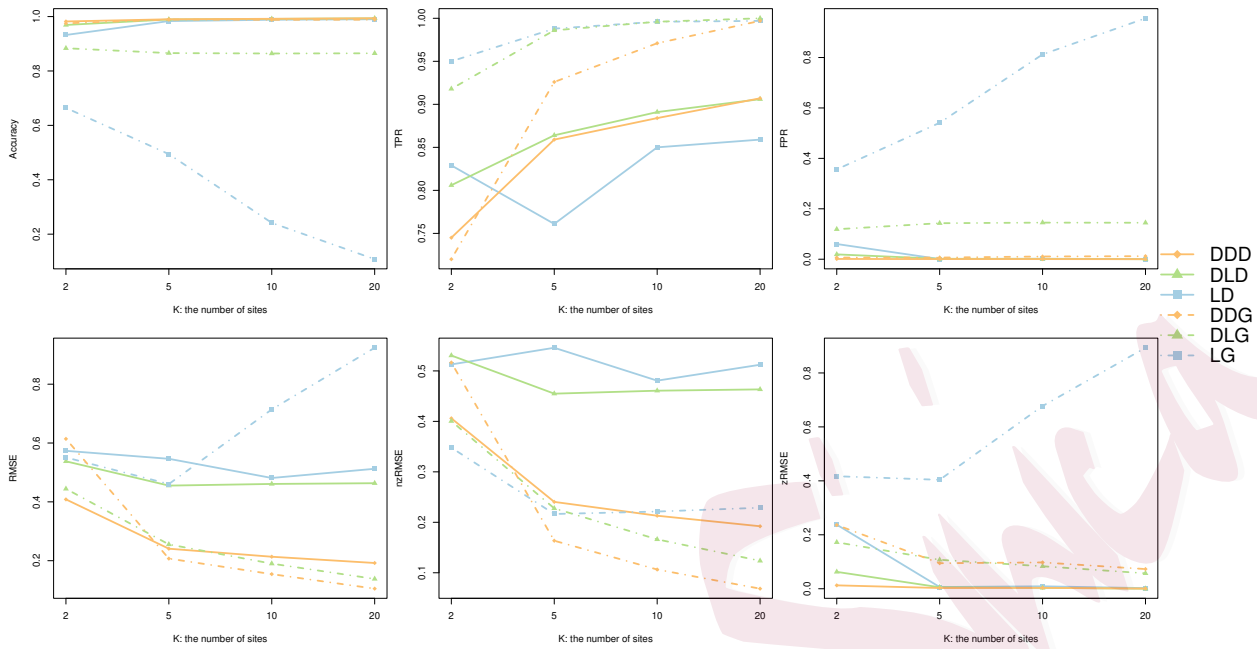


(a) Homogeneous small-sample case: Each local site contains $p = 150$ predictors and $n_k = 100$ samples for $k = 1, \ldots, K$.

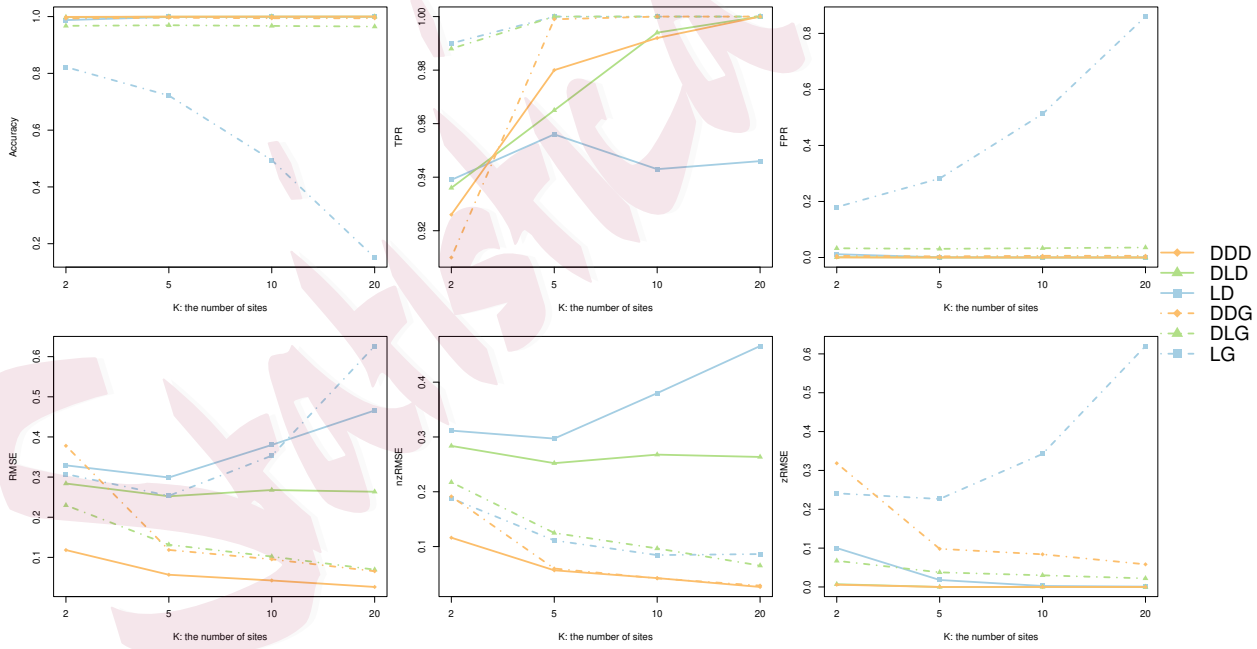


(b) Homogeneous large-sample case: Each local site contains $p = 750$ predictors and $n_k = 500$ samples for $k = 1, \ldots, K$.

Figure 4: Simulation results on variable selection and coefficient estimation in the homogeneous scenario. Across all K local sites, the number of hidden confounders is $q_k = 3$ for $k = 1, \ldots, K$.



(a) Heterogeneous small-sample case: Each local site contains $p = 150$ predictors. The local sample sizes n_k are given in Table 2 (Supplementary Materials, Section S5).



(b) Heterogeneous large-sample case: Each local site contains $p = 750$ predictors. The local sample sizes n_k are given in Table 3 (Supplementary Materials, Section S5).

Figure 5: Simulation results on variable selection and coefficient estimation in the heterogeneous scenario. For $K = 2$, site 1 and 2 have $q_1 = 3$ and $q_2 = 4$ hidden confounders, respectively; for $K \in \{5, 10, 20\}$, site k has $q_k = k$ hidden confounders for $k = 1, \ldots, K$.

reveals that a larger local sample size n_k also enhances the performance of DDD in variable selection and coefficient estimation. These simulation results provide strong empirical validation for the asymptotic properties in Theorems 2 and 3.

Across all four simulation cases, DDD achieves comparable performance in variable selection to both DLD and LD, with marginally reduced TPR and FPR, yet slightly improved accuracy. This stems from majority voting employed by all three distributed estimation methods to control the sparsity of estimators. Notably, DDD exhibits superior performance in coefficient estimation, as evidenced by lower nzRMSE, zRMSE, and RMSE compared to DLD and LD. This advantage arises from the capability of DDD to correct both the heterogeneous pervasive hidden confounding bias and the high-dimensional estimation bias across sites. In contrast, DLD only corrects for the former, while LD fails to mitigate either bias.

Next, we further compare the distributed and global methods across four simulation cases. By incorporating majority voting, the distributed methods perform more stringent variable selection in finite-sample settings. They eliminate irrelevant variables more aggressively, albeit with the risk of excluding some truly relevant variables. This explains why the TPR of DDD is slightly lower than that of DDG. On the other hand, DDG performs variable selection through hypothesis testing. According to Corollary 1, the Type I error rate is nonzero, leading to a higher FPR for DDG than for DDD. Consequently, the two methods deliver comparable accuracy, both nearing unity.

In the four simulation cases, the non-transferability of site-level data inevitably degrades the estimation performance of distributed methods. Specifically, when the sample size is small, DDD exhibits a higher nzRMSE than DDG, but this gap narrows as the sample size increases. Owing to the variable selection performance, DDD achieves a lower zRMSE relative to DDG. Consequently, the RMSE of DDD is comparable to that of DDG for small

sample sizes, and becomes lower than DDG's as the sample size grows.

6. Real data application

We apply the proposed method to a real-world dataset from the UK Biobank (UKB), a large-scale biomedical database. White matter hyperintensities (WMH) are identified as areas of enhanced brightness on T1 and T2 FLAIR magnetic resonance images of the human brain. WMH are common imaging markers of cerebral small vessel disease and neurodegenerative disorders, closely associated with protein-mediated pathological processes (Huang et al., 2025; Festa et al., 2024). For example, hyperphosphorylated tau may indirectly impair white matter integrity through axonal transport dysfunction (Soldan et al., 2020), and matrix metalloproteinases degrade extracellular matrix components, disrupt the blood-brain barrier, and contribute to white matter injury (Egashira et al., 2015). Therefore, we aim to identify WMH-associated proteins and investigate their effects on WMH, facilitating early disease diagnosis, risk prediction, and therapy optimization.

The UKB database provides normalized protein expression data for 2923 plasma proteins of 53014 participants. After excluding 3 proteins with high missingness rates (GLIPR1: 99.7%, NPM1: 74.0%, PCOLCE: 63.6%), we retain 2920 proteins, imputing missing values by k-nearest neighbors ($k = 10$) followed by $\log_2$ transformation. Predictors consist of 2920 proteins and four potential WMH risk factors including sex, age, body mass index (BMI), and Townsend deprivation index, as supported by previous literature (Sliz et al., 2022). The outcome, WHM, is evaluated through total volume of WHM from T1 and T2 FLAIR images. Due to the positive skewness of the distribution of the outcome, a logarithmic transformation is applied. We exclude participants with missing data and any of the following disorders: multiple sclerosis or other demyelinating disorders, cerebral infarction, encephalitis or brain

abscess, brain tumor, or systemic lupus erythematosus on the basis of the codes from the International Classification of Diseases, Tenth Revision (ICD-10) (Wartolowska and Webb, 2021). Finally, the real-world dataset includes 5638 participants recruited from 17 UKB assessment centers. We standardized the dataset to eliminate the influence of differences in scale and magnitude in the data. All UKB data in this study are available to eligible researchers after registration and can be accessed at `https://biobank.ndph.ox.ac.uk`.

Based on UKB assessment centers, 5638 samples can be treated as a multi-site dataset across 17 sites. We calculate singular values of 2924 predictors for each site (Guo et al., 2022). Figure 6 shows that top singular values are larger than the rest, indicating that pervasive hidden confounding still exists in the standardized dataset of each site. The finding supports the use of the proposed distributed method to identify WMH-associated proteins and explore the association between proteins and WMH.

We identify variables associated with WMH and estimate their regression coefficients. The global methods compared with the deconfounded-debiased distributed estimation (DDD) are the global deconfounded-debiased estimation (DDG) and the global lasso estimation (LG) (Tibshirani, 1996), using simulation-consistent parameters. In the deconfounded-debiased distributed estimation, parameters for local estimation align with simulation settings, and the threshold for majority voting is fixed at $\omega = 1$. Our analysis demonstrates that $\omega > 1$ leads to over-selection, an overly stringent variable selection process that excludes potentially relevant variables. Specifically, when $\omega = 2$, the selected variables decrease to 4; $\omega = 3$, only 2 variables remain; and $\omega > 3$, all variables are excluded. Detailed results are available in Table 4 (Supplementary Materials, Section S6).

Variable selection results across the three methods are summarized in Table 1. For complete details on protein overlapping with DDD, refer to Table 5 in Section S6 of Sup-

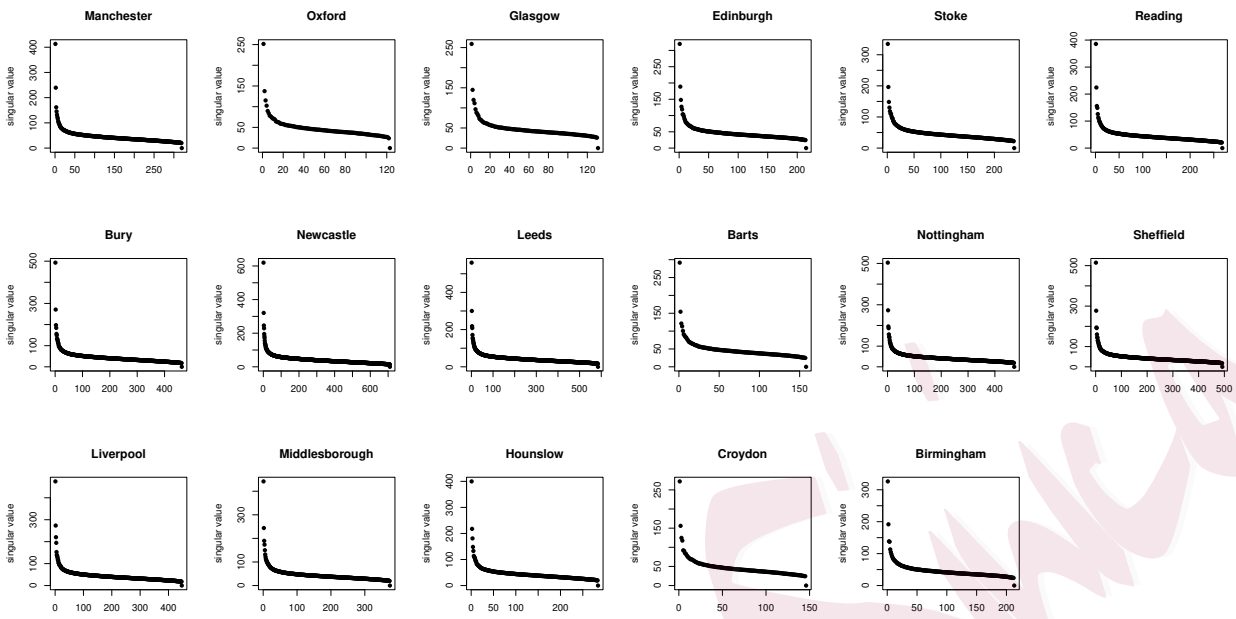


Figure 6: Singular values of 2924 predictors for each site. Each subplot is labeled with the geographical location of its corresponding UKB assessment center.

plementary Materials. The deconfounded-debiased distributed estimation provides more precise selection of WMH-related proteins, whereas the global estimations tend to include potentially irrelevant proteins. This distinction can also be qualitatively demonstrated via enrichment analysis on proteins identified by each method using g:Profiler (Raudvere et al., 2019) with the following data sources: Gene Ontology, Kyoto Encyclopedia of Genes and Genomes, and Reactome Pathway Database. Significant terms were filtered by an adjusted p-value< 0.05, using only annotated genes as the background set. The formation of WMH involves multiple physiological mechanisms including endothelial dysfunction, extracellular matrix remodeling, altered remyelination, inflammation, and response to ischemia (Jickling et al., 2022). As shown in Figure 7, the proteins selected by the proposed distributed estimation are significantly enriched in pathways closely related to these mechanistic themes.

Table 1: Variable selection results of three methods. Numbers in the table represent the number of variables.

Methods	Selected variables	Risk factors	Proteins	Proteins shared with DDD
DDD	58	1 (age)	57	-
DDG	28	2 (age, BMI)	26	3
LG	304	2 (age, BMI)	302	17

However, some pathways enriched in proteins identified via DDG show limited relevance to WMH. For instance, defective binding of VWF variant to GPIb:IX:V describes molecular pathologies of specific monogenic hereditary diseases, with no direct relevance to WMH. Similarly, LG yields proteins enriched in low-relevance pathways such as sulfur compound binding, a metabolic process for which mechanistic links to WMH remains substantially understudied. Enriched pathways for proteins identified by two global estimation methods are provided in Figures 1 and 2 (Supplementary Materials, Section S6).

7. Conclusion

For analyzing high-dimensional multi-site data with heterogeneous pervasive hidden confounders, we propose a deconfounded-debiased distributed estimation method, requiring two rounds of communication between local sites and the central site. Each site applies the the deconfounded-debiased estimation method in (Li et al., 2025) to its own data, thereby simultaneously reducing pervasive hidden confounding bias and high-dimensional estimation bias in the local estimators. The central site aggregates the local estimators corresponding to predictors selected via majority voting. This yields a sparse deconfounded-debiased distributed estimator that naturally performs variable selection. The key strengths of the proposed

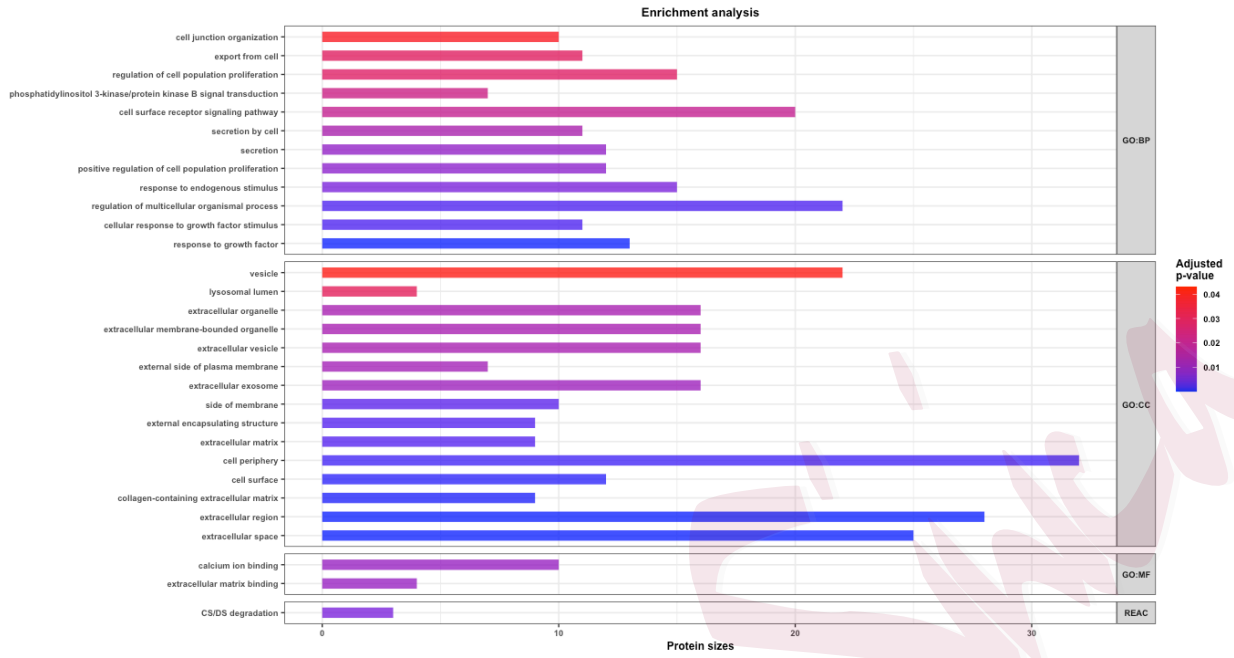


Figure 7: Enriched pathways for proteins identified by the deconfounded-debiased distributed estimation in enrichment analysis.

approach are correcting biases induced by heterogeneous pervasive hidden confounders and by high-dimensional estimation, while also accommodating site-specific heteroscedasticity of the random errors.

Theoretically, we establish the following main results: (1) asymptotic normality of each individual local coefficient estimator and of the local coefficient estimator vector for the active predictors; (2) asymptotic validity and an asymptotic lower bound on the power of individual hypothesis tests at each site; (3) variable selection consistency of the proposed distributed estimator; (4) asymptotic normality of the proposed distributed estimator; (5) upper and lower bounds on the expected false positive rate and the expected true positive rate, respectively, which quantify the influence of majority voting on variable selection in finite samples.

In simulations, we consider two scenarios involving heterogeneous and homogeneous pervasive hidden confounding. The results demonstrate that the proposed distributed estimation outperforms alternative distributed methods and their global counterparts in both variable selection and coefficient estimation. We apply the proposed distributed estimation to a real-world dataset from the UK Biobank, identify proteins associated with white matter hyperintensities (WMH), and investigate their effects on WMH. Compared with global estimation methods, the proposed distributed estimation provides a more precise selection of WMH-related proteins, as further validated through enrichment analysis.

We only consider the linear regression with continuous outcomes at each site. This distributed approach may be extended to generalized linear models for categorical outcomes, Cox models for survival data, and quantile regression models for skewed distributed outcomes. The challenge, however, is that the practical feasibility and theoretical justification of spectral transformation for deconfounding in nonlinear models remain unclear. Moreover, an interesting question for future study is whether the deconfounding technique here can be extended to the communication-efficient surrogate likelihood method, despite its conceptual complexity. Additionally, it is promising to explore the integration of our method with privacy-enhancing technologies to control information leakage risks. For instance, majority voting could be combined with cryptographic protocols, and calibrated Gaussian noise could be injected prior to transmission at local sites. These directions are highly worth investigating and will be the focus of our future work.

Supplementary Materials

All proofs of theorems, corollaries, and the proposition, along with additional tables and figures from simulation experiments and the real data application are available in the online

Supplementary Materials.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No.82473724 to GQ).

References

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Bycroft, C., C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, et al. (2018). The uk biobank resource with deep phenotyping and genomic data. *Nature 562*(7726), 203–209.

Ćevid, D., P. Bühlmann, and N. Meinshausen (2020). Spectral deconfounding via perturbed sparse linear models. *The Journal of Machine Learning Research 21*(1), 9442–9482.

Chen, X. and M. Xie (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica 24*(4), 1655–1684.

Egashira, Y., H. Zhao, Y. Hua, R. F. Keep, and G. Xi (2015). White matter injury after subarachnoid hemorrhage. *Stroke 46*(10), 2909–2915.

Fan, J., Y. Guo, and K. Wang (2023). Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association 118*(542), 1000–1010.

Fan, J., H. Liu, and W. Wang (2018). Large covariance estimation through elliptical factor models. *Annals of statistics 46*(4), 1383.

Fan, J., Z. Lou, and M. Yu (2024). Are latent factor regression and sparse regression adequate? *Journal of the*

American Statistical Association 119(546), 1076–1088.

Festa, L. K., J. B. Grinspan, and K. L. Jordan-Sciutto (2024). White matter injury across neurodegenerative disease. *Trends in Neurosciences 47*(1), 47–57.

Gao, Y., W. Liu, H. Wang, X. Wang, Y. Yan, and R. Zhang (2022). A review of distributed statistical inference. *Statistical Theory and Related Fields 6*(2), 89–99.

Gold, D., J. Lederer, and J. Tao (2020). Inference for high-dimensional instrumental variables regression. *Journal of Econometrics 217*(1), 79–111.

Gu, J. and S. X. Chen (2023). Distributed statistical inference under heterogeneity. *Journal of Machine Learning Research 24*(387), 1–57.

Guo, Z., D. Ćevid, and P. Bühlmann (2022). Doubly debiased lasso: High-dimensional inference under hidden confounding. *Annals of statistics 50*(3), 1320.

Hou, Z., W. Ma, and L. Wang (2023). Sparse and debiased lasso estimation and inference for high-dimensional composite quantile regression with distributed data. *TEST 32*(4), 1230–1250.

Huang, H., W. Song, P. Wang, Y. Zhu, L. Zheng, C. Shen, H. Xu, and J. Qiu (2025). White matter hyperintensities: Cerebral small-vessel diseases and white matter microstructural impairments. *iRADIOLOGY 3*(1), 5–25.

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

Jickling, G. C., B. P. Ander, X. Zhan, B. Stamova, H. Hull, C. DeCarli, and F. R. Sharp (2022). Progression of cerebral white matter hyperintensities is related to leucocyte gene expression. *Brain 145*(9), 3179–3186.

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.

Li, Z., Y. Liu, K. Wei, Y. Yu, G. Qin, and Z. Zhu (2025). Deconfounded and debiased estimation for high-dimensional linear regression under hidden confounding with application to omics data. *Bioinformatics 41*(7), btaf400.

Liu, W., X. Mao, and J. Tu (2025). Communication-efficient distributed sparse learning with oracle property and geometric convergence. *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION 120*(552), 2606–2618.

Ogier du Terrail, J., Q. Klopfenstein, H. Li, I. Mayer, N. Loiseau, M. Hallal, M. Debouver, T. Camalon, T. Fouqueray, J. Arellano Castro, et al. (2025). Fedeca: federated external control arms for causal inference with time-to-event data in distributed settings. *Nature Communications 16*(1), 7496.

Raudvere, U., L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, and J. Vilo (2019). g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research 47*(W1), W191–W198.

Shamir, O., N. Srebro, and T. Zhang (2014). Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pp. 1000–1008. PMLR.

Sliz, E., J. Shin, S. Ahmad, D. M. Williams, S. Frenzel, F. Gauß, S. E. Harris, A.-K. Henning, M. V. Hernandez, Y.-H. Hu, B. Jiménez, M. Sargurupremraj, C. Sudre, R. Wang, K. Wittfeld, Q. Yang, J. M. Wardlaw, H. Völzke, M. W. Vernooij, J. M. Schott, M. Richards, P. Proitsi, M. Nauck, M. R. Lewis, L. Launer, N. Hosten, H. J. Grabe, M. Ghanbari, I. J. Deary, S. R. Cox, N. Chaturvedi, J. Barnes, J. I. Rotter, S. Debette, M. A. Ikram, M. Fornage, T. Paus, S. Seshadri, Z. Pausova, and for the NeuroCHARGE Working Group (2022). Circulating metabolome and white matter hyperintensities in women and men. *Circulation 145*(14), 1040–1052.

Soldan, A., C. Pettigrew, Y. Zhu, M.-C. Wang, A. Moghekar, R. F. Gottesman, B. Singh, O. Martinez, E. Fletcher, C. DeCarli, et al. (2020). White matter hyperintensities and csf alzheimer disease biomarkers in preclinical alzheimer disease. *Neurology 94*(9), e950–e960.

Sudlow, C., J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine 12*(3), e1001779.

Sun, T. and C.-H. Zhang (2012). Scaled sparse linear regression. *Biometrika 99*(4), 879–898.

Sun, Y., L. Ma, and Y. Xia (2024). A decorrelating and debiasing approach to simultaneous inference for high-dimensional confounded models. *Journal of the American Statistical Association 119*(548), 2857–2868.

Tang, L., L. Zhou, and P. X.-K. Song (2020). Distributed simultaneous inference in generalized linear models via confidence distribution. *Journal of Multivariate Analysis 176*, 104567.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology 58*(1), 267–288.

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

Wang, J., Q. Zhao, T. Hastie, and A. B. Owen (2017). Confounder adjustment in multiple hypothesis testing. *Annals of statistics 45*(5), 1863.

Wang, K. (2021). Unified distributed robust regression and variable selection framework for massive data. *Expert Systems with Applications 186*, 115701.

Wartolowska, K. A. and A. J. Webb (2021). Blood pressure determinants of cerebral white matter hyperintensities and microstructural injury: Uk biobank cohort study. *Hypertension 78*(2), 532–539.

Yang, S. and P. Ding (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association 115*(531), 1540–1554.

Yu, G. and J. Bien (2019). Estimating the error variance in a high-dimensional linear model. *Biometrika 106*(3), 533–546.

Yu, J., H. Wang, M. Ai, and H. Zhang (2022). Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association 117*(537), 265–276.

Yu, M., J. Li, and Y. Zhou (2026). Enhancements of communication-efficient distributed statistical inference and its privacy preservation. *Journal of Econometrics 253*, 106125.

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

Zhao, H. and X. Shen (2025). Distributed algorithms for high-dimensional statistical inference and structure learning with heterogeneous data. *Statistica Sinica 38*(1).

Zhaoyang Li

Department of Biostatistics, School of Public Health, Fudan University, Shanghai, China

E-mail: 23111020028@m.fudan.edu.cn

Chen Huang

Department of Mathematics and Statistics, York University, Toronto, Canada

E-mail: huang25@yorku.ca

Guoyou Qin

Department of Biostatistics, School of Public Health, Fudan University, Shanghai, China

E-mail: gyqin@fudan.edu.cn

Zhongyi Zhu

Department of Statistics and Data Science, Fudan University, Shanghai, China

E-mail: zhuzy@fudan.edu.cn