Journal of Machine Learning Research 25 (2024) 1-60 Submitted 1/23; Revised 2/24; Published 6/24

Heterogeneity-aware Clustered Distributed Learning for Multi-source Data Analysis

Yuanxing Chen yxchen_research@163.com
*Department of Statistics and Data Science, School of Economics
Xiamen University
Xiamen, 361005, China*

Qingzhao Zhang qzzhang@xmu.edu.cn
*Department of Statistics and Data Science, School of Economics
The Wang Yanan Institute for Studies in Economics
Xiamen University
Xiamen, 361005, China*

Shuangge Ma shuangge.ma@yale.edu
*Department of Biostatistics
Yale University
New Haven, CT 06520, USA*

Kuangnan Fang[*] xmufkn@163.com
*Department of Statistics and Data Science, School of Economics
Xiamen University
Xiamen, 361005, China*

Editor: Dan Alistarh

Abstract

In diverse fields ranging from finance to omics, it is increasingly common that data is distributed with multiple individual sources (referred to as "clients" in some studies). Integrating raw data, although powerful, is often not feasible, for example, when there are considerations on privacy protection. Distributed learning techniques have been developed to integrate summary statistics as opposed to raw data. In many existing distributed learning studies, it is stringently assumed that all the clients have the same model. To accommodate data heterogeneity, some federated learning methods allow for client-specific models. In this article, we consider the scenario that clients form clusters, those in the same cluster have the same model, and different clusters have different models. Further considering the clustering structure can lead to a better understanding of the "interconnections" among clients and reduce the number of parameters. To this end, we develop a novel penalization approach. Specifically, group penalization is imposed for regularized estimation and selection of important variables, and fusion penalization is imposed to automatically cluster clients. An effective ADMM algorithm is developed, and the estimation, selection, and clustering consistency properties are established under mild conditions. Simulation and data analysis further demonstrate the practical utility and superiority of the proposed approach.

[*] Kuangnan Fang is the corresponding author.

Keywords: high dimensionality, data heterogeneity, clustering structure, sparsity, penalization

1. Introduction

In diverse fields, it is increasingly common that data is distributed with multiple individual sources (referred to as "clients" in this article and some published studies). For example, in financial studies, it is common that data is with, for example, multiple individual bank branches. In omics studies, it is common that multiple independent studies have generated their own data and address the same scientific question. The power of integrating data from multiple sources has been well identified (Liu et al., 2014). A family of studies/methods integrate raw data (Tang and Song, 2016; Huang et al., 2017). Such integrative analysis methods, although effective, are not always feasible due to the need for privacy protection.

Since the huge financial loss of Facebook due to its privacy breach (Kelleher, 2018), privacy issues have once again attracted widespread attention from both the industry and academia. The innovations in a wide range of industries, such as smart healthcare, financial technology, and surveillance systems, rely on newly developing machine learning methods, and then, the development of machine learning methods needs to take privacy protection into full consideration (Liu et al., 2021). In machine learning, privacy protection can be roughly divided into three mechanisms, including homomorphic encryption, obfuscation, and aggregation (Liu et al., 2021). Homomorphic encryption facilitates the processing of the encrypted data without the need to access the raw data. Such technique has been successfully applied to regression (Chen et al., 2018), classification (Bost et al., 2015), and deep neural networks (Gilad-Bachrach et al., 2016). Obfuscation mechanism can be achieved by adding noises to the model parameters or the original data set, and differential privacy (DP) (Dwork, 2006; Agarwal et al., 2018) is the most popular scheme in obfuscation. Aggregation organizes multiple parties to join a machine learning task while avoiding the transmission of the raw data (Zhang et al., 2021).

Distributed learning (DL), as the most famous framework in aggregation, aims to train the global model by aggregating the summary statistics from all clients without sharing the raw data. The existing works can be divided into two categories based on whether the number of iterations is once or multiple times (Zhou et al., 2024). One-shot approaches require just one communication round between the local clients and the central server, in which the divide-and-conquer (DC) strategy is the most popular one designed to reduce communication burden and improve feasibility and performance in the analysis of big data (Lee et al., 2017; Battey et al., 2018). Although one-shot approaches have the lowest communication costs, to obtain the same convergence rate with a centralized estimator, a sufficient sample size of each client relative to the number of clients is necessary (Wang et al., 2017). To further relax this constraint, communication-iterative approaches, such as distributed approximate Newton-type method (Shamir et al., 2014) and communication-efficient surrogate likelihood (Jordan et al., 2019), have been developed. Additionally, when data is large, sending raw data from individual clients to a central server and constructing a statistical model with the pooled data may lead to considerable computational cost (Bhowmick et al., 2018). However, the transmission and aggregation of summary statistics in DL can alleviate communication and computation burden simultaneously.

The DL methods described above all assume that the individual clients share the same data generation model. This assumption, although convenient, may be overly stringent. In raw data-based integrative analysis (Tang and Song, 2016; Huang et al., 2017), it has been well established that data may be heterogeneous and demand different models. The issue of data heterogeneity, known as non-i.i.d data, in the distributed learning setting has attracted widespread attention. For example, different hospitals usually store electronic health records (EHR) in their local sites and are unwilling to share their raw data with others. The data heterogeneity (reflected by heterogeneous outcome-covariate relationships) due to different patient populations should be further considered (Liu et al., 2022; Duan et al., 2022). Besides, Yu et al. (2020) further confirmed that incorrectly borrowing information from other sites with large heterogeneity leads to unreliable inferences and/or low prediction power. As the data heterogeneity can be caused by differences in sample characteristics, data collection techniques, and multiple other factors (Ghosh et al., 2020), the personalized DL methods that borrow strength from similar individual clients have attracted growing interest (Smith et al., 2017).

Federated learning (FL), as a popular DL paradigm, pays more attention to the problem of data islands in a collaborative manner compared to general distributed learning (Kaissis et al., 2020; Liu et al., 2021). It provides a novel method to build personalized models without violating user privacy (Zhang et al., 2021). It typically involves multiple rounds of communication between the central server and local clients to obtain the final model estimates. Clustered federated learning (CFL), as a special case of FL, aims to classify clients into multiple clusters such that clients in the same cluster share the same model, and different clusters have different models (Ghosh et al., 2020; Marfoq et al., 2021). This type of heterogeneity analysis may have been more popular in computer science than statistics and is crucial in applications such as recommendation systems and personalized advertisement placement (Ghosh et al., 2020). Intuitively, assuming and identifying a clustering structure can lead to a better understanding of the “interconnections” among clients (those in the same cluster are more alike and can be more closely related to each other) and a smaller number of model parameters. For instance, mobile phone users (clients) may focus on different clusters of news, like politics, sports, or fashion. Besides, different groups of customers are interested in different categories of ads. Thus, having a deeper understanding of the “interconnections” within a cluster can benefit more accurate personalized recommendations.

A common limitation shared by the existing CFL methods is that it is usually challenging to determine the number of clusters. For example, Ghosh et al. (2020) and Marfoq et al. (2021) first pre-specified the number of clusters and then alternately updated the cluster membership of each client and model parameters for each cluster. Similar to classic clustering analysis, results can be sensitive to the number of clusters, and in practice, usually, there is not enough information to accurately specify this number. Specifically, Ghosh et al. (2020) mentioned in one experiment that by setting a larger number of clusters, their algorithm can identify the correct number by emptying the excess clusters. However, in most cases (refer to the numerical results in Section 4), the ultimately estimated number of clusters remains unchanged, staying at the initially pre-specified number. In addition, since the membership updating algorithm in Ghosh et al. (2020) is similar to K-means, the final clustering structure is highly sensitive to the initial clustering segmentation, which further leads to unstable estimation results (refer to the numerical results in Section 4). Moreover, both Ghosh et al. (2020) and Marfoq et al. (2021) focused on the dense setting with all parameters being nonzero. Therefore, to deal with high-dimensional scenarios with sparsity settings, we should develop a new clustered distributed learning method, which can generate stable and sparse estimates (for interpretation) and identify the true number of clusters in a data-driven way. Of course, this method should also accommodate data heterogeneity and privacy protection at the same time.

In the statistical literature, there are also a few heterogeneous distributed learning methods that allow for client-specific models. For example, Zhao et al. (2016) proposed a heterogeneous distributed learning method with a partially linear model, under which the nonparametric parameter is assumed to be shared by all clients, while the parametric parameters are allowed to be client-specific. Duan et al. (2022) extended the surrogate likelihood function approach to allow client-specific nuisance parameters by adopting a surrogate estimating equation technique. It is noted that these two (and some other) studies are limited to low-dimensional settings. Cai et al. (2022) further studied the high-dimensional heterogeneous setting by aggregating local summary statistics under a generalized linear model. As recognized in Tang et al. (2021), allowing all clients to have individual models may lead to a large number of redundant parameters, negatively affecting estimation and inference.

In this article, we consider the integrative analysis of multi-source data under privacy protection. We utilize the summary statistics instead of the raw individual-level data to avoid privacy breaches while learning parametric models based on the distributed learning framework. Here the summary statistics can contain initial parameter estimates, gradient vectors, hessian matrices, and so on. To sufficiently accommodate data heterogeneity, captured by different model parameters, clients are allowed to have different models. Specifically, motivated by the success of clustered federated learning, we consider the scenario where clients form clusters, and the models are cluster-specific. Besides, to address the high-dimensional issues with sparsity assumption, we focus on the scenario where those models have the same sparsity structure (set of important variables) and note that the proposed strategy can be extended to accommodate different sparsity structures.

To achieve the goal of simultaneous estimation, variable selection, and clustering, we develop an integrative clustered regression (ICR) method, which may advance from the existing literature in multiple important ways. First, compared to methods that assume homogeneity (Lee et al., 2017), it is more flexible and can effectively accommodate data heterogeneity. Second, compared to methods that allow for client-specific models (Zhao et al., 2016; Duan et al., 2022), it can lead to a better understanding of the similarity/differences among data sets and a smaller number of parameters (and hence improved estimation). Third, compared to the existing CFL methods (Ghosh et al., 2020; Marfoq et al., 2021), it can data-dependently and conveniently determine the number of clusters, and the estimated cluster memberships are not sensitive to the initial cluster partition, with the assistance of penalized fusion. Fourth, compared to the dense setting applied to neutral network (Ghosh et al., 2020), the sparse parameter estimates due to sparsity penalty can facilitate model interpretability as well as efficient inference and training in neutral network (Hoefler et al., 2021). Last but not least, it can accommodate multiple types of data/models, and our computational and theoretical developments can shed broader insights.

The rest of the article is organized as follows. In Section 2, we introduce the data/model settings, the proposed approach, and an effective proximal ADMM algorithm. In Section 3, we rigorously establish that the proposed approach enjoys the estimation, variable selection, and clustering consistency properties. Numerical studies, including simulation in Section 4 and data analysis in Section 5, demonstrate the practical utilization and superiority of the proposed approach. Brief discussions are provided in Section 6. The proofs of theoretical results and additional numerical results are relegated to the Appendix.

2. Methods

In this section, we first introduce the integrative clustered model in a distributed setup and then develop a proximal ADMM algorithm to obtain the ICR estimator.

2.1 Integrative Analysis under Privacy Constraints

Suppose that there are K independent clients, and for the kth client, there are n_k observations. The total sample size is $N = \sum_{k=1}^{K} n_k$. For the kth client, let $y_i^{(k)}$ and $\boldsymbol{x}_i^{(k)} = (x_{i1}^{(k)}, \ldots, x_{ip}^{(k)})^\top \in \mathbb{R}^p$ be the response and covariate vector of the ith observation, respectively, where the first element of $\boldsymbol{x}_i^{(k)}$ is fixed as $x_{i1}^{(k)} \equiv 1$ to accommodate intercept. Accordingly, let $\mathbf{X}^{(k)} = (\boldsymbol{x}_1^{(k)}, \ldots, \boldsymbol{x}_{n_k}^{(k)})^\top$ and $\mathbf{Y}^{(k)} = (y_1^{(k)}, \ldots, y_{n_k}^{(k)})^\top$ denote the design matrix and response vector of the kth client, respectively. Let $f(\cdot)$ be the pre-specified twice-differentiable loss function, and define the true population coefficients as

$$\boldsymbol{\theta}^{*(k)} = \underset{\boldsymbol{\theta}^{(k)} \in \mathbb{R}^p}{\arg\min}\ \mathcal{L}_k(\boldsymbol{\theta}^{(k)}) \quad \text{and} \quad \mathcal{L}_k(\boldsymbol{\theta}^{(k)}) = \mathbb{E}\big[f(\boldsymbol{\theta}^{(k)\top}\boldsymbol{x}_i^{(k)}, y_i^{(k)})\big],\ k \in [K],$$

where $\boldsymbol{\theta}^{(k)} = (\theta_1^{(k)}, \ldots, \theta_p^{(k)})^\top$ is the p-dimensional coefficient vector, and $[d]$ denotes the index set $\{1, \ldots, d\}$ for an integer d. Accordingly, the empirical local and global loss functions are defined as

$$\widehat{\mathcal{L}}_k(\boldsymbol{\theta}^{(k)}) = \frac{1}{n_k}\sum_{i=1}^{n_k} f\big(\boldsymbol{x}_i^{(k)\top}\boldsymbol{\theta}^{(k)}, y_i^{(k)}\big),\ k \in [K] \quad \text{and} \quad \widehat{\mathcal{L}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{k=1}^{K} n_k \widehat{\mathcal{L}}_k(\boldsymbol{\theta}^{(k)}),$$

respectively, where $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(K)})$ is a $p \times K$ coefficient matrix with the jth row $\boldsymbol{\theta}_j = (\theta_j^{(1)}, \cdots, \theta_j^{(K)})^\top$. Assume that $\mathcal{G} = \{\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(M)}\}$ forms a non-overlapping partition of $\{1, \ldots, K\}$, and that clients from the same cluster share the same coefficient vector. That is, given $m \in [M]$, for any $k \in \mathcal{G}^{(m)}$, $\boldsymbol{\theta}^{*(k)} = \boldsymbol{\psi}^{*(m)}$, where $\boldsymbol{\psi}^{*(m)}$ is the cluster-specific coefficient vector for cluster m. Additionally, for each covariate, its coefficients across the K clients can be viewed as a group (Cai et al., 2022), leading to p groups corresponding to the covariates.

For simultaneous regularized estimation, variable selection, and identification of the clustering structure of clients, we propose the objective function with the ideal pooling (IP) strategy

$$\begin{aligned}\widehat{\mathcal{Q}}_{\mathrm{IP}}(\boldsymbol{\theta}) &= \widehat{\mathcal{L}}(\boldsymbol{\theta}) + \mathcal{P}_{\lambda_1}(\boldsymbol{\theta}) + \mathcal{P}_{\lambda_2}(\boldsymbol{\theta}) \\ &= \frac{1}{N}\sum_{k=1}^{K} n_k \widehat{\mathcal{L}}_k(\boldsymbol{\theta}^{(k)}) + \sum_{j=2}^{p} p_\tau\left(\|\boldsymbol{\theta}_j\|_2, \lambda_1\right) + \sum_{k<k'} p_\tau\left(\left\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k')}\right\|_2, \lambda_2\right),\end{aligned} \tag{1}$$

where penalty $\mathcal{P}_{\lambda_1}(\boldsymbol{\theta})$ is mainly for regularized estimation and variable selection, and penalty $\mathcal{P}_{\lambda_2}(\boldsymbol{\theta})$ is mainly for clustering. Here $p_\tau(,)$ is a penalty function with concavity parameter τ, $\|\cdot\|_2$ is the L_2 norm, and λ_1, λ_2 are two non-negative tuning parameters.

With the privacy-preservation constraints, raw data of the individual client is not available, and hence objective function $\widehat{\mathcal{Q}}_{\mathrm{IP}}(\boldsymbol{\theta})$ in (1) cannot be directly implemented. To tackle this problem, we adopt the least-square approximation (LSA) of He et al. (2016) and Zhu et al. (2021), which leads to the objective function

$$\widehat{\mathcal{Q}}_1(\boldsymbol{\theta}) = \frac{1}{N}\sum_{k=1}^{K} n_k(\boldsymbol{\theta}^{(k)} - \widetilde{\boldsymbol{\theta}}^{(k)})^\top \widetilde{\mathbf{V}}^{(k)}(\boldsymbol{\theta}^{(k)} - \widetilde{\boldsymbol{\theta}}^{(k)}) + \mathcal{P}_{\lambda_1}(\boldsymbol{\theta}) + \mathcal{P}_{\lambda_2}(\boldsymbol{\theta}), \tag{2}$$

where $\widetilde{\boldsymbol{\theta}}^{(k)}$ is the local estimator of the kth client, and $\widetilde{\mathbf{V}}^{(k)} = \partial^2 \widehat{\mathcal{L}}_k(\widetilde{\boldsymbol{\theta}}^{(k)})/\partial\boldsymbol{\theta}^{(k)}\partial\boldsymbol{\theta}^{(k)\top}$ is the Hessian matrix of $\widehat{\mathcal{L}}_k(\boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}^{(k)}$ at $\widetilde{\boldsymbol{\theta}}^{(k)}$. He et al. (2016) recommended adopting ordinary least square (OLS) estimates as the local estimators when $p < n_k$. Under high-dimensional settings, OLS estimates are not available, and a "straightforward" approach is to replace the OLS estimates with the Lasso estimates. However, the computationally efficient ordinary Lasso estimates are usually biased, and the debiased Lasso estimates (van de Geer et al., 2014) are often computationally expensive. Inspired by Cai et al. (2022), we propose the ICR estimator $\widehat{\boldsymbol{\theta}}$ by minimizing the following objective function

$$\widehat{\mathcal{Q}}_{\mathrm{ICR}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{k=1}^{K} n_k\Big(\boldsymbol{\theta}^{(k)\top}\widetilde{\mathbf{V}}^{(k)}\boldsymbol{\theta}^{(k)} - 2\boldsymbol{\theta}^{(k)\top}\widetilde{\boldsymbol{\zeta}}^{(k)}\Big) + \mathcal{P}_{\lambda_1}(\boldsymbol{\theta}) + \mathcal{P}_{\lambda_2}(\boldsymbol{\theta}), \tag{3}$$

where $\widetilde{\boldsymbol{\zeta}}^{(k)} = \widetilde{\mathbf{V}}^{(k)}\widetilde{\boldsymbol{\theta}}^{(k)} - \widetilde{\mathbf{g}}^{(k)}$ and $\widetilde{\mathbf{g}}^{(k)} = \partial\widehat{\mathcal{L}}_k(\widetilde{\boldsymbol{\theta}}^{(k)})/\partial\boldsymbol{\theta}^{(k)}$ is the gradient of $\widehat{\mathcal{L}}_k(\boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}^{(k)}$ at $\widetilde{\boldsymbol{\theta}}^{(k)}$. Here we use Lasso estimators as local estimators $\widetilde{\boldsymbol{\theta}}^{(k)}$. For the penalty function, viable choices include SCAD (Fan and Li, 2001), MCP (Zhang, 2010), and others. We adopt MCP in our numerical studies. Note that, with the first loss term in (3), we can achieve debiasing without actually resorting to the computationally expensive debiased estimates (we refer to Cai et al., 2022 for more details). The overall analysis approach is schematically presented in Figure 1. It consists of generating individual estimates based on raw data by individual clients, sending summary estimates from local clients to a central server, conducting the proposed estimation, and outputting the final estimators to guide downstream analysis/actions.

This approach has been motivated by the following considerations. In $\widehat{\mathcal{Q}}_{\mathrm{ICR}}(\boldsymbol{\theta})$, we only make use of four summary statistics, namely the initial local estimators $\{\widetilde{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K$, corresponding gradient vectors $\{\widetilde{\mathbf{g}}^{(k)}\}_{k=1}^K$, Hessian matrices $\{\widetilde{\mathbf{V}}^{(k)}\}_{k=1}^K$, and local sample sizes $\{n_k\}_{k=1}^K$. That is, the proposed approach and estimate are fully based on the summary statistics as opposed to the raw data – data privacy protection is thus achieved. In (3), the first term measures lack-of-fit, and similar forms have been considered in the literature (Zhu et al., 2021; Cai et al., 2022). When higher-order estimation properties are not of interest, the estimates and Hessian matrices from the local clients contain sufficient information. The first penalty determines which covariates have overall nonzero effects, under the assumptions that one covariate may have different effects/coefficients for different clients, but the effects are either all nonzero or all zero. It is possible to replace it with more complex penalties, for example, those that can conduct two-level selection (Huang et al., 2017), to

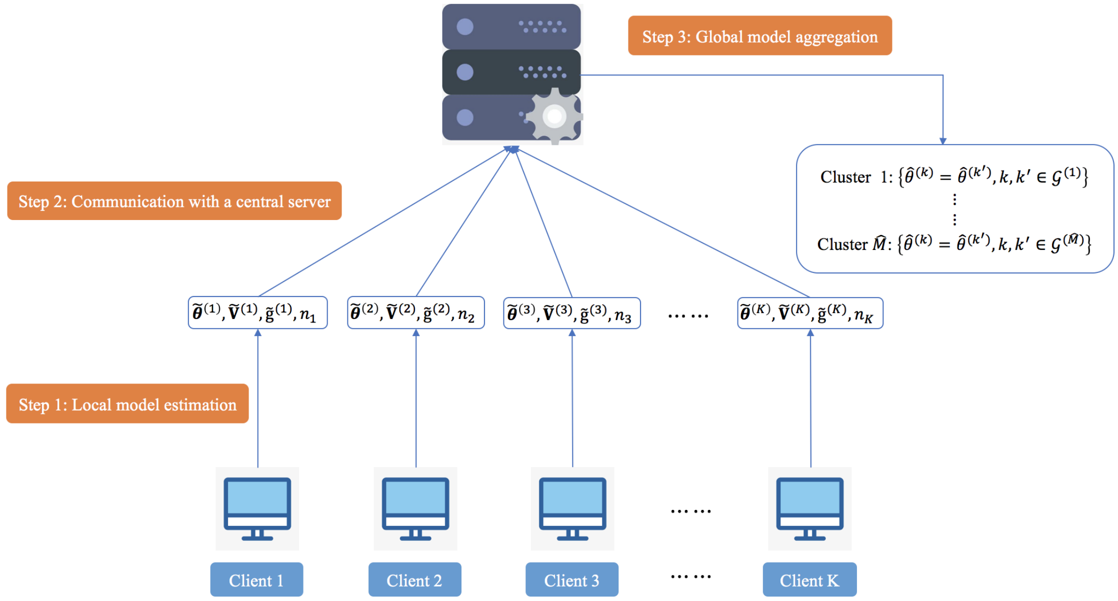

Step 3: Global model aggregation
Step 2: Communication with a central server
Cluster 1: $\{\hat{\theta}^{(k)} = \hat{\theta}^{(k')}, k, k' \in \mathcal{G}^{(1)}\}$
⋮
Cluster $\widehat{M}$: $\{\hat{\theta}^{(k)} = \hat{\theta}^{(k')}, k, k' \in \mathcal{G}^{(\widehat{M})}\}$
$\widetilde{\boldsymbol{\theta}}^{(1)}, \widetilde{\mathbf{V}}^{(1)}, \tilde{\mathrm{g}}^{(1)}, n_1$
$\widetilde{\boldsymbol{\theta}}^{(2)}, \widetilde{\mathbf{V}}^{(2)}, \tilde{\mathrm{g}}^{(2)}, n_2$
$\widetilde{\boldsymbol{\theta}}^{(3)}, \widetilde{\mathbf{V}}^{(3)}, \tilde{\mathrm{g}}^{(3)}, n_3$
……
$\widetilde{\boldsymbol{\theta}}^{(K)}, \widetilde{\mathbf{V}}^{(K)}, \tilde{\mathrm{g}}^{(K)}, n_K$
Step 1: Local model estimation
……
……
Client 1
Client 2
Client 3
Client K

Figure 1: Scheme of the proposed analysis.

obtain "more subtle" information. The second is a fusion penalty (Ma and Huang, 2017), with which some clients may have exactly equal estimates. Clients k and k' are clustered together if and only if their estimates are equal. For identifying clustering structures, fusion penalization has been recognized to have multiple unique advantages and has been popular in the recent literature (Ma and Huang, 2017; Yang et al., 2019; Chen et al., 2021). For example, it translates clustering to an "easier" estimation problem and can more conveniently determine the number/structure of clusters (by examining the estimates). It is worth noting that most of the existing studies, such as Ma and Huang (2017) and Chen et al. (2021), focus on heterogeneity analysis with a single data set, while here we study the subgrouping structure of multiple clients (data sets). The identification of subgrouping can facilitate "personalized" analysis and improve individual analysis by reducing the number of parameters/increasing sample size. Additionally, different from Yang et al. (2019) and others that demand raw data, we can achieve the goal of privacy protection and reduce computational cost by analyzing summary statistics.

Remark 1 *We claim that taking into account the clustering structure can reduce the number of parameters and we are going to further explain this from two perspectives. From the theoretical perspective, after assuming a latent cluster partition within K clients and the clients from the same cluster share the same parameters, the true number of model parameters depends on the number of clusters. To see this, note that if we allow client-specific parameters for all K clients, there are a total of Kp parameters that need to be estimated. However, the existing clustering structure leads to M cluster-specific parameters, which results in Mp parameters being estimated. Since M is usually much smaller than K, the proposed ICR method can utilize samples from all clients belonging to a cluster to estimate cluster-specific parameters, thus improving the theoretical convergence rate of estimation er-*

rors (see Section 3 for more details). From the computational perspective, although there are still Kp parameters involved in the estimation process, through penalized fusion, parameters belonging to the same cluster tend to be the same. As a result, the number of distinct parameters is greatly reduced. Based on this, for subsequent observations generated by clients from cluster m, predictions can be made based on the estimated mth cluster-specific parameters.

Remark 2 *It is worth emphasizing that the local estimators $\{\widetilde{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K$, obtained by solving the corresponding local penalized loss functions, serve only as a part of summary statistics for obtaining the final ICR estimators $\{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K$. For the local estimators, the sparsity structures vary across different clients and there is no clustering structure among clients. On the contrary, the proposed ICR estimators share the same sparsity structure and clients from the same cluster share the same estimated parameters. This difference leads to better variable selection performance and higher estimation accuracy for the ICR estimators compared to the local estimators (see the numerical results in Section 4).*

2.2 Computational Algorithm

We use local linear approximation — LLA (Zou and Li, 2008) to approximate the fused penalty and propose an iterative algorithm. Specifically, in the tth iteration, we update the coefficients by solving

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^{p\times K}}{\arg\min}\ \frac{1}{N}\sum_{k=1}^{K} n_k\Big(\boldsymbol{\theta}^{(k)\top}\widetilde{\mathbf{V}}^{(k)}\boldsymbol{\theta}^{(k)}-2\boldsymbol{\theta}^{(k)\top}\widetilde{\boldsymbol{\zeta}}^{(k)}\Big)+\sum_{k<k'}\omega_{kk'}^{t-1}\Big\|\boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}^{(k')}\Big\|_2+\sum_{j=2}^{p}p_\tau(\|\boldsymbol{\theta}_j\|_2,\lambda_1),$$

where $\omega_{kk'}^{t-1}=p_\tau'(\|\boldsymbol{\theta}^{(k),t-1}-\boldsymbol{\theta}^{(k'),t-1}\|_2,\lambda_2)$ denotes the weight and $p_\tau'(x,\lambda)$ is the derivative of $p_\tau(x,\lambda)$ with respect to x. The above minimization problem can be reformulated as a constrained minimization problem

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^{p\times K}}{\arg\min}\left\{\ell(\boldsymbol{\theta}):=\overbrace{\frac{1}{N}\sum_{k=1}^{K} n_k\Big(\boldsymbol{\theta}^{(k)\top}\widetilde{\mathbf{V}}^{(k)}\boldsymbol{\theta}^{(k)}-2\boldsymbol{\theta}^{(k)\top}\widetilde{\boldsymbol{\zeta}}^{(k)}\Big)}^{g(\boldsymbol{\theta})}\right.$$
$$\left.+\underbrace{\sum_{k<k'}\omega_{kk'}^{t-1}\|\boldsymbol{\alpha}_{kk'}\|_2}_{h_1(\boldsymbol{\alpha})}+\underbrace{\sum_{j=2}^{p}p_\tau(\|\boldsymbol{\theta}_j\|_2,\lambda_1)}_{h_2(\boldsymbol{\theta})}\right\},$$
$$\text{subject to}\quad \boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}^{(k')}=\boldsymbol{\alpha}_{kk'},\ 1\le k<k'\le K,$$

where $\boldsymbol{\alpha}=(\boldsymbol{\alpha}_{12},\ldots,\boldsymbol{\alpha}_{(K-1)K})$ is a $p\times K(K-1)/2$ matrix composed of the auxiliary variables. This optimization problem is equivalent to the minimization of the augmented Lagrangian

$$\ell_\nu(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\xi})=\ell(\boldsymbol{\theta})+\sum_{k<k'}\boldsymbol{\xi}_{kk'}^\top(\boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}^{(k')}-\boldsymbol{\alpha}_{kk'})+\frac{\nu}{2}\sum_{k<k'}\Big\|\boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}^{(k')}-\boldsymbol{\alpha}_{kk'}\Big\|_2^2, \tag{4}$$

where $\boldsymbol{\xi}=(\boldsymbol{\xi}_{12},\ldots,\boldsymbol{\xi}_{(K-1)K})$ is a $p\times K(K-1)/2$ matrix composed of the dual variables. ν is a small positive constant. Following Shimmura and Suzuki (2022), we can minimize objective function $\ell_\nu(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\xi})$ in (4) via the following iterations

$$\begin{aligned} (\boldsymbol{\theta}^t, \boldsymbol{\alpha}^t) &= \arg\min_{\boldsymbol{\theta}, \boldsymbol{\alpha}} \ \ell_\nu(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\xi}^{t-1}), \\ \boldsymbol{\xi}_{kk'}^t &= \boldsymbol{\xi}_{kk'}^{t-1} + \nu(\boldsymbol{\theta}^{(k),t} - \boldsymbol{\theta}^{(k'),t} - \boldsymbol{\alpha}_{kk'}^t), \ 1 \le k < k' \le K. \end{aligned} \tag{5}$$

To update $(\boldsymbol{\theta}, \boldsymbol{\alpha})$ via (4), we minimize $\eta(\boldsymbol{\theta})$ defined by

$$\begin{aligned} & \eta(\boldsymbol{\theta}) \\ := & \min_{\boldsymbol{\alpha}} \ \ell_\nu(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\xi}^{t-1}) \\ = & \min_{\boldsymbol{\alpha}} \left\{ \overbrace{\sum_{k<k'} \left[\omega_{kk'}^{t-1} \|\boldsymbol{\alpha}_{kk'}\|_2 + \boldsymbol{\xi}_{kk'}^{t-1\top} (\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k')} - \boldsymbol{\alpha}_{kk'}) + \frac{\nu}{2} \left\| \boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k')} - \boldsymbol{\alpha}_{kk'} \right\|_2^2 \right]}^{\eta_1(\boldsymbol{\theta}, \boldsymbol{\alpha})} \right\} \\ & + g(\boldsymbol{\theta}) + h_2(\boldsymbol{\theta}) \\ = & \ \eta_2(\boldsymbol{\theta}) + h_2(\boldsymbol{\theta}). \end{aligned} \tag{6}$$

Following Chi and Lange (2015), we define the proximal map with respect to $\Omega(\mathbf{v})$ as

$$\text{prox}_{\sigma\Omega}(\mathbf{u}) = \arg\min_{\mathbf{v}} \left[\sigma\Omega(\mathbf{v}) + \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|_2^2 \right].$$

Besides, the conjugate function of $\Omega(\mathbf{v})$ is defined by $\Omega^*(\mathbf{u}) = \sup_{\mathbf{v}}[\mathbf{u}^\top \mathbf{v} - \Omega(\mathbf{v})]$. Then, it is easy to show that $\eta_1(\boldsymbol{\theta}, \boldsymbol{\alpha})$ is minimized when

$$\boldsymbol{\alpha}(\boldsymbol{\theta}) = \text{prox}_{\nu^{-1} h_1}(\boldsymbol{\theta}\mathbf{A} + \nu^{-1}\boldsymbol{\xi}^{t-1}), \tag{7}$$

where $\mathbf{A} = (\mathbf{e}_{12}, \ldots, \mathbf{e}_{(K-1)K})$ is a $K \times K(K-1)/2$ matrix and $\mathbf{e}_{kk'} = \mathbf{e}_k - \mathbf{e}_{k'}$, in which $\mathbf{e}_k$ is a $K \times 1$ vector whose kth element is 1 and the remaining elements are 0. Plugging (7) into $\eta_1(\boldsymbol{\theta}, \boldsymbol{\alpha})$ in (6) and combining the results of Theorem 1 and Lemmas 1—2 of Shimmura and Suzuki (2022), we can show that $\eta_2(\boldsymbol{\theta})$ is differentiable, and the gradient of $\eta_2(\boldsymbol{\theta})$ is

$$\frac{\partial \eta_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{\theta}_g + \left[\text{prox}_{\nu h_1^*}(\nu\boldsymbol{\theta}\mathbf{A} + \boldsymbol{\xi}^{t-1}) \right] \mathbf{A}^\top,$$

where $\boldsymbol{\theta}_g = 2/N\big[n_1(\widetilde{\mathbf{V}}^{(1)}\boldsymbol{\theta}^{(1)} - \widetilde{\boldsymbol{\zeta}}^{(1)}), \ldots, n_K(\widetilde{\mathbf{V}}^{(K)}\boldsymbol{\theta}^{(K)} - \widetilde{\boldsymbol{\zeta}}^{(K)})\big]$ is a $p \times K$ matrix and h_1^* is the conjugate function of h_1. Then, we can adopt one proximal gradient technique, called Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) (Parikh et al., 2014), to obtain the solution of the first minimization problem in (5). Further, similar to equations (26) and (27) in Shimmura and Suzuki (2022), the second iterative step in (5) can be reformulated as

$$\boldsymbol{\xi}^t = \left[\text{prox}_{\nu h_1^*}(\nu\boldsymbol{\theta}^t\mathbf{A} + \boldsymbol{\xi}^{t-1}) \right].$$

The proposed proximal ADMM algorithm is summarized as follows.

Step 1. Obtain the initial estimates with $(\boldsymbol{\theta}^0, \boldsymbol{\xi}^0)$.

Step 2. At iteration $t, t = 1, 2, \ldots,$ update $\boldsymbol{\theta}^t$ as follows.

Step 2.1. Initialize $\boldsymbol{u}^{t-1,0} = \boldsymbol{\theta}^{t-1,0} = \boldsymbol{\theta}^{t-1}$ and $\rho_0 = 1$.

Step 2.2. At iteration $s, s = 1, 2, \ldots$, compute

$$\omega_{kk'}^{t-1,s} \leftarrow p'_\tau\left(\left\|\boldsymbol{u}^{(k),t-1,s-1} - \boldsymbol{u}^{(k'),t-1,s-1}\right\|_2, \lambda_2\right), \quad 1 \le k < k' \le K,$$

$$\boldsymbol{\theta}^{t-1,s} \leftarrow \operatorname{prox}_{\varsigma h_2}\left[\boldsymbol{u}^{t-1,s-1} - \varsigma \frac{\partial \eta_2(\boldsymbol{u}^{t-1,s-1})}{\partial \boldsymbol{\theta}}\right],$$

$$\rho_s \leftarrow \frac{1 + \sqrt{1 + 4\rho_{s-1}^2}}{2}, \quad \boldsymbol{u}^{t-1,s} \leftarrow \boldsymbol{\theta}^{t-1,s} + \frac{\rho_{s-1} - 1}{\rho_s}(\boldsymbol{\theta}^{t-1,s} - \boldsymbol{\theta}^{t-1,s-1}).$$

Step 2.3. Repeat Step 2.2 until convergence, and set $\boldsymbol{\theta}^t \leftarrow \boldsymbol{\theta}^{t-1,s}$.

Step 3. For $1 \le k < k' \le K$, update $\omega_{kk'}^t \leftarrow p'_\tau(\|\boldsymbol{\theta}^{(k),t} - \boldsymbol{\theta}^{(k'),t}\|_2, \lambda_2)$.

Step 4. Update $\boldsymbol{\xi}^t \leftarrow \operatorname{prox}_{\nu h_1^*}(\nu \boldsymbol{\theta}^t \mathbf{A} + \boldsymbol{\xi}^{t-1})$.

Step 5. Repeat Steps 2—4 until convergence, and set $\boldsymbol{\alpha}^t \leftarrow \operatorname{prox}_{\nu^{-1} h_1}(\boldsymbol{\theta}^t \mathbf{A} + \nu^{-1} \boldsymbol{\xi}^t)$.

In the above calculation, we conclude convergence if the absolute difference of estimates from two consecutive iterations is smaller than a predefined cutoff.

Remark 3 *There exist closed-form solutions for the proximal maps of νh_1^*, $\nu^{-1} h_1$, and ςh_2. Specifically, the proximal map of νh_1^* is a projection function. And the proximal maps of $\nu^{-1} h_1$ and ςh_2 can be easily derived as in Ma and Huang (2017) with $\tau > \varsigma$. In Step 2.2, ς denotes the step size. As in Shimmura and Suzuki (2022), we can derive the Lipschitz constant of $\eta_2(\boldsymbol{\theta})$, denoted by $L_\eta = 1 + 2\nu \max_{k \in [K]} (\mathbf{A}\mathbf{A}^\top)_{k,k}$, and then set $\varsigma = L_\eta^{-1}$. With $\nu = 1$ and $\tau = 3$, $\tau > \varsigma$ since $\varsigma \le 1/3$. Here, although superlinear convergence may be achieved if $\nu \to \infty$ (Rockafellar, 1976), it is difficult to prove the convergence of ADMM when ν varies by iteration (see Boyd et al., 2011, Section 3.4.1). Therefore, while there may be improvements in convergence rate, varying ν may also lead to the algorithm failing to converge. In practice, $\nu = 1$ is widely adopted and achieves good convergence in the implementation of the ADMM algorithm for extensive studies (Ma and Huang, 2017; Zhu and Qu, 2018; Ren et al., 2023).*

Remark 4 *The basic framework of our algorithm is ADMM (Boyd et al., 2011), of which one variant accommodates a differential loss function plus nonconvex penalties, and its convergence properties have been studied in Ma and Huang (2017) and Tang et al. (2021). The "standard" method for minimizing the first problem in (5) involves inverting a matrix (Ma and Huang, 2017; Zhu and Qu, 2018), which can be computationally difficult if p and K are large. A novelty in our algorithm is to replace this with the proximal gradient method, and convergence of the FISTA technique has been studied in Beck and Teboulle (2009). Therefore, by combining the convergence properties of Beck and Teboulle (2009) and Tang et al. (2021), the proposed algorithm is also expected to have satisfactory convergence.*

Tuning parameter selection Following the literature, we set $\nu = 1$ and the concavity related parameter $\tau = 3$. Following Yang et al. (2019), we select λ_1 and λ_2 by minimizing the modified BIC defined as

$$\mathrm{mBIC}(\lambda_1,\lambda_2) = \frac{1}{N}\sum_{k=1}^{K} n_k \left(\left[\widehat{\boldsymbol{\theta}}^{(k)}(\lambda_1,\lambda_2)\right]^\top \widetilde{\mathbf{V}}^{(k)} \left[\widehat{\boldsymbol{\theta}}^{(k)}(\lambda_1,\lambda_2)\right] - 2\left[\widehat{\boldsymbol{\theta}}^{(k)}(\lambda_1,\lambda_2)\right]^\top \widetilde{\boldsymbol{\zeta}}^{(k)} \right) + C_N \frac{\log N}{N}\widehat{q}(\lambda_1,\lambda_2),$$

where $\widehat{q}(\lambda_1,\lambda_2)$ is the number of nonzero distinct coefficient vectors, and C_N is a positive constant depending on N. Following Ma and Huang (2017), we adopt $C_N = \log(\log(Kp))$, which can automatically adapt to a diverging number of parameters.

3. Theoretical Properties

Here we establish that the proposed ICR estimator has the well-desired estimation consistency, model selection consistency, and clustering consistency properties. Although sharing some similar spirit with the existing studies, with a significantly different problem and penalized estimation, our theoretical development can have a unique value.

3.1 Notations and Definitions

For a vector $\mathbf{z} = (z_1,\ldots,z_p) \in \mathbb{R}^p$, and $1 \le l < \infty$, define $\|\mathbf{z}\|_l = (\sum_{j=1}^p |z_j|^l)^{1/l}$ and $\|\mathbf{z}\|_\infty = \max_{j\in[p]}|z_j|$. Given an index set $\mathcal{S}$, let $\mathbf{z}_{\mathcal{S}}$ denote the subvector of $\mathbf{z}$ corresponding to the elements of $\mathcal{S}$. For a matrix $\mathbf{Z}_{s\times p}$, let $\|\mathbf{Z}\|_2 = \sup_{\mathbf{v}\in\mathbb{R}^p,\|\mathbf{v}\|_2=1}\|\mathbf{Z}\mathbf{v}\|_2$, $\|\mathbf{Z}\|_\infty = \max_{1\le i\le s}\sum_{j=1}^p |Z_{ij}|$, $\|\mathbf{Z}\|_{\max} = \max_{1\le i\le s,1\le j\le p}|Z_{ij}|$, and $\|\mathbf{Z}\|_F = \sqrt{\sum_{i=1}^s\sum_{j=1}^p Z_{ij}^2}$. For two index sets $\mathcal{S}_1$ and $\mathcal{S}_2$, let $\mathbf{Z}_{\mathcal{S}_1\mathcal{S}_2}$ denote the submatrix of $\mathbf{Z}$ corresponding to the rows in $\mathcal{S}_1$ and columns in $\mathcal{S}_2$, and $\mathbf{Z}_{\mathcal{S}_1}$ denote the submatrix of $\mathbf{Z}$ corresponding to the rows in $\mathcal{S}_1$. For a vector $\mathbf{v}_0 \in \mathbb{R}^p$, let $\mathcal{B}_r(\mathbf{v}_0) = \{\mathbf{v}\in\mathbb{R}^p : \|\mathbf{v}-\mathbf{v}_0\|_2 \le r\}$ be the ℓ_2-ball around $\mathbf{v}_0$ with radius $r > 0$. For a random variable X, its sub-Gaussian norm is defined by $\|X\|_{\psi_2} = \sup_{s\ge1} s^{-1/2}(\mathbb{E}|X|^s)^{1/s}$. For a random vector $\mathbf{z}\in\mathbb{R}^p$, its sub-Gaussian norm is defined by $\|\mathbf{z}\|_{\psi_2} = \sup_{\mathbf{v}\in\mathcal{B}_1(\mathbf{0})}\|\mathbf{v}^\top\mathbf{z}\|_{\psi_2}$. For a symmetric matrix $\mathbf{H}$, its maximum and minimum eigenvalues are denoted by $\Lambda_{\max}(\mathbf{H})$ and $\Lambda_{\min}(\mathbf{H})$, respectively. For two sequences of real numbers $\{a_n\} \ge 1$ and $\{b_n\} \ge 1$, $b_n \ll a_n$ (or $b_n = o(a_n)$) means that $\limsup_{n\to\infty} b_n/a_n = 0$, $b_n \lesssim a_n$ (or $b_n = O(a_n)$) means that $\exists C > 0$ such that $b_n \le Ca_n$ for all n, and we use $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Similarly, we let $o_p(\cdot)$ and $O_p(\cdot)$ represent each of the corresponding rates with probability approaching 1 as $n\to\infty$. Let $f'(a,y) = \partial f(a,y)/\partial a$ and $f''(a,y) = \partial^2 f(a,y)/\partial a^2$, where $\partial f(a,y)/\partial a$ and $\partial^2 f(a,y)/\partial a^2$ denote the first and second order derivatives of $f(a,y)$ with respect to a, respectively.

Let $\mathcal{M}_{\mathcal{G}}$ be a subspace of $\mathbb{R}^{p\times K}$ defined as

$$\mathcal{M}_{\mathcal{G}} = \left\{\boldsymbol{\theta}\in\mathbb{R}^{p\times K} : \boldsymbol{\theta}^{(k)} = \boldsymbol{\psi}^{(m)}, \text{for any } k\in\mathcal{G}^{(m)}, 1\le m\le M\right\},$$

where $\boldsymbol{\psi}^{(m)}$ is the distinct coefficient vector for the mth cluster. Further, we define the $p\times M$ common coefficient matrix $\boldsymbol{\psi} = (\boldsymbol{\psi}^{(1)},\ldots,\boldsymbol{\psi}^{(M)}) = (\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_p)^\top$,where $\boldsymbol{\psi}^{(m)} = (\psi_1^{(m)},\ldots,\psi_p^{(m)})^\top$ and $\boldsymbol{\psi}_j = (\psi_j^{(1)},\ldots,\psi_j^{(M)})^\top$. Let $\boldsymbol{\theta}^*$ and $\boldsymbol{\psi}^*$ be the true coefficient matrices corresponding to $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, respectively. Without loss of generality, assume the first q groups of covariates have nonzero effects, and the rest $(p-q)$ have zero effects. Let $\mathcal{A} = \{1, \ldots, q\}$ and $\mathcal{A}^c = \{q+1, \ldots, p\}$. Further, denote $d_1 = \min_{j \in \mathcal{A}} \|\boldsymbol{\psi}_j^*\|_2$ and $d_2 = \min_{m, m' \in [M], m \neq m'} \|\boldsymbol{\psi}_{\mathcal{A}}^{*(m)} - \boldsymbol{\psi}_{\mathcal{A}}^{*(m')}\|_2$.

Let $\mathbf{V}^{(k)}(\boldsymbol{\theta}^{(k)}) = \partial^2 \widehat{\mathcal{L}}_k(\boldsymbol{\theta}^{(k)}) / \partial \boldsymbol{\theta}^{(k)} \partial \boldsymbol{\theta}^{(k)\top}$ and $\mathbf{g}^{(k)}(\boldsymbol{\theta}^{(k)}) = \partial \widehat{\mathcal{L}}_k(\boldsymbol{\theta}^{(k)}) / \partial \boldsymbol{\theta}^{(k)}$. We further denote $\widetilde{\mathbf{V}}^{(k)} = \mathbf{V}^{(k)}(\widetilde{\boldsymbol{\theta}}^{(k)})$, $\mathbf{V}^{*(k)} = \mathbf{V}^{(k)}(\boldsymbol{\theta}^{*(k)})$, $\widetilde{\mathbf{g}}^{(k)} = \mathbf{g}^{(k)}(\widetilde{\boldsymbol{\theta}}^{(k)})$ and $\mathbf{g}^{*(k)} = \mathbf{g}^{(k)}(\boldsymbol{\theta}^{*(k)})$ for simplicity. Let $\varphi^{(k)} = \left\| \mathbb{E}(\mathbf{V}^{*(k)}_{\mathcal{A}^c \mathcal{A}}) \left[\mathbb{E}(\mathbf{V}^{*(k)}_{\mathcal{A}\mathcal{A}})\right]^{-1} \right\|_\infty$ for any $k \in [K]$ and $\varphi_{\max} = \max_{k \in [K]} \varphi^{(k)}$. Let $N_m = \sum_{k \in \mathcal{G}^{(m)}} n_k$, $N_{\max} = \max_{m \in [M]} N_m$, and $N_{\min} = \min_{m \in [M]} N_m$. Let $|\mathcal{G}^{(m)}|$ be the cardinality of index set $\mathcal{G}^{(m)}$ with $m \in [M]$, and denote $|\mathcal{G}_{\max}| = \max_{m \in [M]} |\mathcal{G}^{(m)}|$ and $|\mathcal{G}_{\min}| = \min_{m \in [M]} |\mathcal{G}^{(m)}|$.

When the underlying true clustering structure $\mathcal{G} = \{\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(M)}\}$ is known, we can define the cluster-oracle objective function for $\boldsymbol{\theta}$ by

$$\underset{\boldsymbol{\theta} \in \mathcal{M}_{\mathcal{G}}}{\arg\min} \left\{ \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{k=1}^{K} n_k \Big(\boldsymbol{\theta}^{(k)\top} \widetilde{\mathbf{V}}^{(k)} \boldsymbol{\theta}^{(k)} - 2 \boldsymbol{\theta}^{(k)\top} \widetilde{\boldsymbol{\zeta}}^{(k)} \Big) + \sum_{j=2}^{p} p_\tau(\|\boldsymbol{\theta}_j\|_2, \lambda_1) \right\}. \tag{8}$$

Accordingly, the cluster-oracle objective function for the common coefficient matrix $\boldsymbol{\psi}$ is

$$\begin{aligned} \mathcal{L}^{\mathcal{G}}(\boldsymbol{\psi}) = \frac{1}{N} \sum_{m=1}^{M} &\left[\boldsymbol{\psi}^{(m)\top} \left(\sum_{k \in \mathcal{G}^{(m)}} n_k \widetilde{\mathbf{V}}^{(k)} \right) \boldsymbol{\psi}^{(m)} - 2 \boldsymbol{\psi}^{(m)\top} \left(\sum_{k \in \mathcal{G}^{(m)}} n_k \widetilde{\boldsymbol{\zeta}}^{(k)} \right) \right] \\ &+ \sum_{j=2}^{p} p_\tau \left(\sqrt{\sum_{m=1}^{M} \left(|\mathcal{G}^{(m)}|^{1/2} \psi_j^{(m)} \right)^2}, \lambda_1 \right). \end{aligned} \tag{9}$$

3.2 Asymptotic Properties

Assume that $n_k \asymp N/K$ for $k \in [K]$, and we denote $n^* \asymp n_k$. We further assume the following mild conditions.

(C1) For each $k \in [K]$ and $i \in [n_k]$, $\{\boldsymbol{x}_i^{(k)}, y_i^{(k)}\}$'s are independent and identically distributed. There exists a constant $C_x > 0$ such that $\max_{k \in [K], i \in [n_k]} \|\boldsymbol{x}_i^{(k)}\|_\infty \leq C_x$ and $\max_{\mathbf{x} \in \mathcal{B}_1(\mathbf{0})} \mathbb{E}(\mathbf{x}^\top \boldsymbol{x}_i^{(k)})^2 \leq C_x^2$.

(C2) For each $k \in [K]$ and $i \in [n_k]$, $f'(\boldsymbol{\theta}^{*(k)\top} \boldsymbol{x}_i^{(k)}, y_i^{(k)})$'s are sub-Gaussian. That is, there exists a constant $\kappa_x > 0$ such that $\|f'(\boldsymbol{\theta}^{*(k)\top} \boldsymbol{x}_i^{(k)}, y_i^{(k)})\|_{\psi_2} \leq \kappa_x$.

(C3) For each $k \in [K]$, there exist two constants $C_{\min}$ and $C_{\max}$ such that $0 < C_{\min} \leq \Lambda_{\min}\big(\mathbb{E}(\mathbf{V}^{*(k)}_{\mathcal{A}\mathcal{A}})\big) \leq \Lambda_{\max}\big(\mathbb{E}(\mathbf{V}^{*(k)}_{\mathcal{A}\mathcal{A}})\big) \leq C_{\max}$.

(C4) For each $k \in [K]$, if $\delta = o(1)$, then there exists a constant $C_L > 0$ such that

$$\left| f''(\boldsymbol{\theta}^{(k)\top} \boldsymbol{x}_i^{(k)}, y_i^{(k)}) \right| \leq C_L, \quad \text{for all } \boldsymbol{\theta}^{(k)} \in \mathcal{B}_\delta(\boldsymbol{\theta}^{*(k)}).$$

Further, the second-order derivatives are Lipschitz continuous. That is,

$$\left| f''(a, y) - f''(b, y) \right| \leq C_L |a - b|, \quad \text{for any } a, b, y \in \mathbb{R}.$$

(C5) The local estimators satisfy

$$\max_{k\in[K]}\left\|\widetilde{\boldsymbol{\theta}}^{(k)}-\boldsymbol{\theta}^{*(k)}\right\|_2 \asymp \max_{k\in[K]} n_k^{-1/2}\left\|\mathbf{X}^{(k)}(\widetilde{\boldsymbol{\theta}}^{(k)}-\boldsymbol{\theta}^{*(k)})\right\|_2 = O_p\left(\sqrt{\frac{q\log p}{n^*}}\right).$$

(C6) The penalty function $p_\tau(t,\lambda)$ is non-decreasing and concave in t for $t\in[0,\infty)$. For $\tau>0$, $\lambda^{-1}p_\tau(t,\lambda)$ is a constant for all $t\geq\tau\lambda$, and $p_\tau(0,\lambda)=0$. In addition, $p'_\tau(t,\lambda)$ exists and is continuous except for a finite number of t values and $\lambda^{-1}p'_\tau(0+,\lambda)=1$.

(C7) $(|\mathcal{G}_{\max}|/|\mathcal{G}_{\min}|)^2q^4\log p \ll n^*$ and $K\ll p$.

Condition (C1) assumes that all covariates are uniformly bounded. Similar conditions have been commonly assumed in the literature, especially including Cai et al. (2022). It is satisfied under many practical scenarios. Condition (C2) controls the tail behavior of $x_{ij}^{(k)}f'(a,y)$ and bounds the random error $\mathbf{g}^{*(k)}$. Condition (C3) has been commonly assumed to ensure that the eigenvalues of $\mathbb{E}(\mathbf{V}_{\mathcal{A}\mathcal{A}}^{*(k)})$ are bounded above and below. The first part of Condition (C4) assumes that the second-order derivatives of the loss function are bounded, and the second part is a Lipschitz condition to ensure that the loss function is sufficiently smooth. Condition (C5) provides the error bounds for the local estimators and similar conditions have been assumed in Cai et al. (2022) and Battey et al. (2018). It is noted that such error bounds have been established in Negahban et al. (2012). Condition (C6) is commonly assumed under high-dimensional settings, and it can be easily verified that both MCP and SCAD satisfy this condition.

It is noted that in Cai et al. (2022) and Jordan et al. (2019), restricted strong convexity has been assumed to derive the upper bound of distributed estimators with full p dimensions. Different from such studies, we follow another framework designed for nonconvex penalties (Fan and Lv, 2011) to study the upper bound of the proposed ICR estimator constrained on the true q-dimensional variables and achieve sparsity by the KKT conditions.

Theorem 1 *Suppose that Conditions (C1)-(C7) hold. If $\lambda_1 \gg |\mathcal{G}_{\max}|^{1/2}r_{1N}+\varphi_{\max}r_{2N}$, $|\mathcal{G}_{\min}|^{1/2}d_1>\tau\lambda_1$ and $r_{1N}=o(1)$, then there exists a strictly local minimizer $\widehat{\boldsymbol{\psi}}^{or}$ of $\mathcal{L}^{\mathcal{G}}(\boldsymbol{\psi})$ in (9) such that*

$$\left\|\widehat{\boldsymbol{\psi}}^{or}_{\mathcal{A}}-\boldsymbol{\psi}^*_{\mathcal{A}}\right\|_F = O_p(r_{1N}), \qquad P(\widehat{\boldsymbol{\psi}}^{or}_{\mathcal{A}^c}=\mathbf{0})\to 1 \quad \text{as} \quad N\to\infty,$$

where

$$r_{1N}=\sqrt{\frac{(K/|\mathcal{G}_{\min}|)q}{N_{\min}}}+\frac{|\mathcal{G}_{\max}|M^{1/2}q^{3/2}\log p}{N_{\min}},$$
$$r_{2N}=\sqrt{\frac{(|\mathcal{G}_{\max}|/|\mathcal{G}_{\min}|)M\log p}{KN}}+\frac{(|\mathcal{G}_{\max}|/|\mathcal{G}_{\min}|^{1/2})M^{1/2}q\log p}{N}.$$

Theorem 1 establishes the estimation consistency and model selection consistency of the cluster-oracle estimator $\widehat{\boldsymbol{\psi}}^{or}$. Note that the second term of r_{1N} is the additional error due to the aggregation of summary statistics as opposed to raw data. If $M|\mathcal{G}_{\max}| = o(\sqrt{N/[q^2(\log p)^2]})$, the second term in the error bound can be dominated by the first term, which means that the additional errors are asymptotically negligible. Furthermore, if M is fixed and $|\mathcal{G}_m| \asymp K/M$ for $m \in [M]$, then r_{1N} turns to be $\sqrt{q/N} + Kq^{3/2}\log p/N$. Similarly, the additional errors are asymptotically negligible if $K = o(\sqrt{N/q^2(\log p)^2})$.

Similar constraints regarding the number of clients and sample sizes of clients are widely recognized in the existing literature (Jordan et al., 2019; Cai et al., 2022). In theory, one-shot algorithms require a sufficient sample size for each client to achieve the same statistical accuracy as centralized algorithms (Battey et al., 2018). However, this condition can be relaxed in communication-iterative algorithms via multiple rounds of communication (Jordan et al., 2019). As a result, if the sample size of each client is not large enough, compared with communication-iterative algorithms with a sufficient number of rounds, one-shot algorithms may yield poorer estimation results. In this paper, the proposed ICR method belongs to one-shot algorithms and then the asymptotic equivalence between ICP and IP holds if such constraint can be satisfied.

Based on Theorem 1 and the equivalence of $\mathcal{L}(\boldsymbol{\theta})$ in (8) and $\mathcal{L}^{\mathcal{G}}(\boldsymbol{\psi})$ in (9), we can similarly construct a strictly local minimizer $\widehat{\boldsymbol{\theta}}^{or}$ of $\mathcal{L}(\boldsymbol{\theta})$ such that $\|\widehat{\boldsymbol{\theta}}^{or}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}\|_F = O_p(|\mathcal{G}_{\max}|^{1/2} r_{1N})$ and $P(\widehat{\boldsymbol{\theta}}^{or}_{\mathcal{A}^c} = \mathbf{0}) \to 1$. Furthermore, in the following Theorem 2, we will show that $\widehat{\boldsymbol{\theta}}^{or}$ is a strictly local minimizer of $\widehat{\mathcal{Q}}_{\mathrm{ICR}}(\boldsymbol{\theta})$ in (3) with probability approaching 1. Consequently, with probability approaching 1, the ICR estimator $\widehat{\boldsymbol{\theta}}$ is equal to the cluster-oracle estimator $\widehat{\boldsymbol{\theta}}^{or}$, which indicates that $\widehat{\boldsymbol{\theta}}$ also possesses estimation consistency with the same convergence rate as $\widehat{\boldsymbol{\theta}}^{or}$ and model selection consistency. As clustering is based on estimation, we can obtain $P(\widehat{M} = M) \to 1$ and $P(\widehat{\mathcal{G}} = \mathcal{G}) \to 1$, which indicates clustering consistency.

Theorem 2 *Suppose that the conditions of Theorem 1 hold. If* $\lambda_1 \gg \varphi_{\max}(\log p/N)^{1/2}$, $d_2 > \tau\lambda_2$ *and* $\lambda_2 \gg |\mathcal{G}_{\max}|^{1/2} r_{1N}$, *then there exists a strictly local minimizer* $\widehat{\boldsymbol{\theta}}$ *of* $\widehat{\mathcal{Q}}_{\mathrm{ICR}}(\boldsymbol{\theta})$ *in (3) such that*

$$P(\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}^{or}) \to 1 \quad \text{as} \quad N \to \infty.$$

It is noted that, compared to the local estimators with convergence rate $O_p(\sqrt{q\log p/n^*})$ as shown in Condition (C5), each $\widehat{\boldsymbol{\theta}}^{(k)}$ possesses a much faster convergence attributable to information aggregation across clients sharing common coefficients. To see this, note that if $|\mathcal{G}_m| \asymp K/M$ for $m \in [M]$ and $M \asymp \log p$, r_{1N} turns to be $\sqrt{q\log p/N_{\min}}$. This rate is the same as that of the cluster-oracle estimator $\widehat{\boldsymbol{\theta}}^{or(k)}$, which has a $\sqrt{|\mathcal{G}_{\min}|}$ times faster convergence compared to the local estimators.

4. Simulation Study

We conduct abundant simulations to gauge the performance of the proposed approach. For benchmarking, we consider three classes of alternatives, which include local methods, homogeneous integrative methods (homoIM), and heterogeneous integrative methods (heterIM). A straightforward local method is (a) the Local estimator obtained by minimizing a local penalized loss function for each client separately. For the class of homoIM, we consider (b) the distributed least square approximation (DLSA) estimator (Zhu et al., 2021); and (c) the weighted one-shot distributed ridge (WONDER) estimator (Dobriban and Sheng, 2020). For the class of heterIM, we consider (d) the Sparse K-means (SK) estimator obtained by applying the sparse K-means approach (Witten and Tibshirani, 2010) to the local estimators in (a), of which the process can be achieved via R package *sparcl*; (e) the clustered federated learning (CFL) estimator (Ghosh et al., 2020); (f) the data-Shielding High dimensional Integrative Regression (SHIR) estimator Cai et al. (2022); and (g) the Sparse Meta-Analysis (SMA) estimator (He et al., 2016) obtained after executing the sure independent screening procedure (Fan and Lv, 2008) to reduce dimension to $n/(3\log n)$ as recommended by He et al. (2016). For the SK estimator, we adopt two criteria, namely the Hartigan statistic (Hartigan, 1975) and gap statistic (Tibshirani et al., 2001), to choose the number of clusters – this is realized using R package *NbClust*. The corresponding two variants are referred to as SK(har) and SK(gap), respectively. For the CFL estimator, we separately analyze one-shot CFL (OCFL) and iterative CFL with multiple rounds (ICFL), where the number of clusters is specified as the true value for them. Here, both ICFL and OCFL correspond to Algorithm 2 of Ghosh et al. (2020), but the former sets the number of communication rounds as $R = 100$, while the latter sets $R = 1$. Since WONDER and CFL methods all focus on dense estimation, for comparison in variable selection, we apply a hard threshold of 0.1 to their dense estimators for obtaining sparse estimates. For the alternatives, tuning parameters are chosen in a way compatible with the proposed.

In addition to these alternatives, we also consider two ideal golden methods, including (h) the Oracle estimator obtained by minimizing objective function $\mathcal{L}^{or,\mathcal{G}}(\boldsymbol{\psi})$ in (A.3); and (i) the ideal pooling (IP) estimator obtained by minimizing objective function (1). Note that the Oracle method is not realistic in practice and the IP method is not feasible in a distributed framework. Here, they serve as the ideal targets and help to verify the theoretical properties established in Section 3.

4.1 Simulation Settings

In this subsection, we design six examples to observe the performance of the proposed method and the other alternatives. Examples 1-2 and 5-6 are on logistic regression and logistic loss, and Examples 3-4 are on linear regression and squared loss. The true number of clusters is $M = 2$ in Examples 1 and 3, and $M = 4$ in Examples 2, 4 and 5. For Examples 1-4, we let $n_1 = \cdots = n_K = n$. Example 5 is an imbalanced setting with varying n_k. To match the real data of the anomaly detection study (used in Section 5), we further design Example 6, in which both the sample sizes and the cluster sizes are imbalanced, and the proportion of anomalies can vary significantly across clients. More specific settings are as follows.

Example 1. $\boldsymbol{\psi}^{(1)} = (0.4\times\mathbf{1}_8^\top, \mathbf{0}_{p-8}^\top)^\top$ and $\boldsymbol{\psi}^{(2)} = (-0.4\times\mathbf{1}_8^\top, \mathbf{0}_{p-8}^\top)^\top$. We generate $\boldsymbol{x}_{i,-1}^{(k)}, i \in [n_k], k \in [K]$, where $\boldsymbol{x}_{i,-1}^{(k)} = (x_{i2}^{(k)}, \ldots, x_{ip}^{(k)})^\top$, from a multivariate normal distribution with mean $\mathbf{0}$ and $\text{cov}(X_w, X_t) = \rho^{|w-t|}$ for $w, t \in \{2, \cdots, p\}$ and $\rho = 0.5$. Given $\mathbf{X}^{(k)}$, we generate $\mathbf{Y}^{(k)}$ from a logistic model. We set the number of clients in each cluster as $|\mathcal{G}_1| = |\mathcal{G}_2| = K/2$. We further set $n = 200$ and consider $K \in \{16, 32, 64\}$ and $p \in \{100, 500\}$.

Example 2. $\boldsymbol{\psi}^{(1)} = (0.6 \times \mathbf{1}_4^\top, -0.6 \times \mathbf{1}_4^\top, \mathbf{0}_{p-8}^\top)^\top$, $\boldsymbol{\psi}^{(2)} = (0.6 \times \mathbf{1}_2^\top, -0.6 \times \mathbf{1}_2^\top, 0.6 \times \mathbf{1}_2^\top, -0.6 \times \mathbf{1}_2^\top, \mathbf{0}_{p-8}^\top)^\top$, $\boldsymbol{\psi}^{(3)} = (-0.6 \times \mathbf{1}_2^\top, 0.6 \times \mathbf{1}_2^\top, -0.6 \times \mathbf{1}_2^\top, 0.6 \times \mathbf{1}_2^\top, \mathbf{0}_{p-8}^\top)^\top$, and $\boldsymbol{\psi}^{(4)} = (-0.6 \times \mathbf{1}_4^\top, 0.6 \times \mathbf{1}_4^\top, \mathbf{0}_{p-8}^\top)^\top$. We set the number of clients in each cluster as $|\mathcal{G}_1| = |\mathcal{G}_2| = |\mathcal{G}_3| = |\mathcal{G}_4| = K/4$. $\mathbf{X}^{(k)}$ and $\mathbf{Y}^{(k)}$ are generated in a similar manner as in Example 1. We consider $K \in \{64, 128\}$ and $n \in \{200, 400, 800\}$ and set $p = 100$.

Example 3. The data generation is the same as in Example 1. The difference is that the response is generated from a linear regression model, where the random error has a normal distribution $\mathcal{N}(0, \sigma^2)$. We consider $K \in \{16, 32, 64\}$ and $\sigma \in \{1, 2\}$ and set $p = 100$ and $n = 100$.

Example 4. The data generation is the same as in Example 2. The difference is that the response is generated from a linear regression model, where the random error has a normal distribution $\mathcal{N}(0, \sigma^2)$. We consider $\sigma \in \{1, 2\}$ and $n \in \{100, 200, 400\}$ and set $p = 100$ and $K = 64$.

Example 5. The data generation is the same as in Example 2. We sample each n_k from $\{n_0 \cdot a : a \in [U]\}$ to allow imbalanced sample sizes across different clients. Here, larger U indicates greater imbalance. We consider $U \in \{5, 10\}$ and $p \in \{200, 500, 800\}$ as well as fixed $K = 40$ and $n_0 = 100$.

Example 6. We consider $M = 6$ clusters with coefficients $\boldsymbol{\psi}^{(1)} = (1.5, \boldsymbol{\beta}^{(1)\top}, \mathbf{0}_{p-11}^\top)^\top$, $\boldsymbol{\psi}^{(2)} = (1.0, \boldsymbol{\beta}^{(2)\top}, \mathbf{0}_{p-11}^\top)^\top$, $\boldsymbol{\psi}^{(3)} = (0.5, \boldsymbol{\beta}^{(3)\top}, \mathbf{0}_{p-11}^\top)^\top$, $\boldsymbol{\psi}^{(4)} = (-0.5, \boldsymbol{\beta}^{(4)\top}, \mathbf{0}_{p-11}^\top)^\top$, $\boldsymbol{\psi}^{(5)} = (-1.0, \boldsymbol{\beta}^{(5)\top}, \mathbf{0}_{p-11}^\top)^\top$, and $\boldsymbol{\psi}^{(6)} = (-1.5, \boldsymbol{\beta}^{(6)\top}, \mathbf{0}_{p-11}^\top)^\top$, where $\boldsymbol{\beta}^{(m)} = (\beta_1^{(m)}, \ldots, \beta_{10}^{(m)})^\top$. Here, for all $m \in [M]$, we let $\beta_j^{(m)} = Z_1 \cdot \mathrm{sign}(W)$ if $j \in [5]$, and $\beta_j^{(m)} = Z_2 \cdot \mathrm{sign}(W)$ otherwise, where Z_1, Z_2 are normally distributed with $Z_1 \sim \mathcal{N}(0.4, 0.1^2)$ and $Z_2 \sim \mathcal{N}(0.8, 0.1^2)$, and W is uniform distributed with $W \sim \mathcal{U}(-0.5, 0.5)$. We set the number of clients in each cluster as $|\mathcal{G}_1| = |\mathcal{G}_2| = 10$, $|\mathcal{G}_3| = |\mathcal{G}_4| = 15$ as well as $|\mathcal{G}_5| = |\mathcal{G}_6| = 20$. We generate $\mathbf{X}^{(k)}$ and $\mathbf{Y}^{(k)}$ in a similar manner as in Example 1, and each n_k is sampled from $\{n_0 \cdot a : a \in [U]\}$. We consider $n_0 \in \{200, 400\}$ and set $p = 100$ and $U = 5$.

For each example, we generate 100 replicates. We first observe that the proposed computational algorithm has satisfactory convergence properties. With all of our simulated data sets, convergence is achieved within 100 iterations. Additionally, the proposed approach is computationally affordable. For example, the analysis of one simulated data set under Example 1 with $K = 32$, $p = 100$, and 25 candidate tuning parameter values takes about 3 minutes using a desktop with standard configurations – here we note that penalized fusion estimation is in general computationally more expensive. For evaluation and comparison, we comprehensively consider the following measures. Denote the set of selected variables as $\widehat{\mathcal{A}} = \{j : \widehat{\boldsymbol{\theta}}_j \neq 0\}$. For evaluating variable selection accuracy, we consider (1) TPR, the percentage of correctly identified important variables across the K studies; (2) FPR, the percentage of falsely identified important variables across the K studies; and (3) MS, the model size defined by $\mathrm{MS} = \sum_{j=1}^{p} \widehat{q}_j$, where $\widehat{q}_j$ is the number of distinct nonzero coefficients of the jth variable. For evaluating clustering accuracy, we consider: (4) $\widehat{M}$, the number of identified clusters; (5) Per, the percentage of fully accurate identification; (6) RI, the Rand Index defined as $\mathrm{RI} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN})$, where TP (true positive) is the number of pairs of data sets from the same cluster classified into the same cluster, and TN (true negative), FP (false positive), and FN (false negative) are defined accordingly. Since Rand index tends to be large even under random clusterings, we also adopt (7) ARI, the adjusted Rand index defined by $\mathrm{ARI} = (\mathrm{RI} - \mathbb{E}(\mathrm{RI}))/(\max(\mathrm{RI}) - \mathbb{E}(\mathrm{RI}))$, where $\mathbb{E}(\mathrm{RI})$ and $\max(\mathrm{RI})$ are the expected value and maximum value of Rand index, respectively. Rand index ranges from 0 to 1, adjusted Rand index ranges from -1 to 1, and higher values indicate a better agreement between the identified and true clustering structures. For

Table 1: The clustering accuracy: mean (sd) based on 100 replicates in Example 1.

		$K = 16$				$K = 32$				$K = 64$			
	Method	$\widehat{M}$	Per	RI	ARI	$\widehat{M}$	Per	RI	ARI	$\widehat{M}$	Per	RI	ARI
$p = 100$	ICR	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	IP	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	ICFL	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	OCFL	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	SK(har)	4.640	0.000	0.779	0.539	5.230	0.000	0.759	0.508	5.220	0.000	0.734	0.462
		(1.508)	(-)	(0.097)	(0.208)	(1.847)	(-)	(0.098)	(0.202)	(1.495)	(-)	(0.065)	(0.133)
	SK(gap)	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
$p = 500$	ICR	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	IP	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	ICFL	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	OCFL	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	SK(har)	4.580	0.000	0.793	0.570	5.440	0.000	0.754	0.499	5.150	0.000	0.766	0.528
		(1.505)	(-)	(0.101)	(0.216)	(1.684)	(-)	(0.097)	(0.199)	(1.720)	(-)	(0.098)	(0.199)
	SK(gap)	2.010	0.990	0.999	0.999	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**
		(0.100)	(-)	(0.006)	(0.012)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)

evaluating estimation, we consider (8) RMSE, the root mean squared error of $\widehat{\boldsymbol{\theta}}$ defined as $\text{RMSE} = \sqrt{\sum_{j \in \mathcal{A}} \|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^*\|_2^2 / K}$.

4.2 Simulations Results

Results for Example 1 are summarized in Table 1, Table 2, and Figure 2. It is observed that the proposed ICR approach tends to have larger TPR and smaller FPR values as K increases. It has an average MS value of around 16, which is the true model size, while the alternatives (in the categories of heterIM or local) generate much larger models and DLSA (in the categories of homoIM) generates much smaller models. When K is sufficiently large, it outperforms the alternatives (except for IP) in variable selection. Compared to IP, ICR has slightly worse performance in variable selection and estimation accuracy, which is a reasonable result. Figure 2 suggests that the estimation accuracy of ICR is very close to that of Oracle, especially when K is large, which is consistent with our theoretical results. Compared with ICFL and OCFL (clustered heterIM) with a pre-specified true number of clusters, ICR shows the same performance in the estimation of the number of clusters with $\widehat{M} = 2$ (the true number of clusters) and clustering accuracy and comparable performance in estimation accuracy. Compared to Local, ICR shows much better performance in both variable selection and estimation accuracy. This further illustrates why we need to inte-

Table 2: The variable selection accuracy: mean (sd) based on 100 replicates in Example 1.

		$K = 16$			$K = 32$			$K = 64$		
	Method	TPR	FPR	MS	TPR	FPR	MS	TPR	FPR	MS
$p = 100$	ICR	0.990	**0.000**	**15.840**	**1.000**	**0.000**	**16.000**	**1.000**	**0.000**	**16.000**
		(0.058)	(0.000)	(0.929)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	IP	0.990	**0.000**	**15.840**	**1.000**	**0.000**	**16.000**	**1.000**	**0.000**	**16.000**
		(0.038)	(0.000)	(0.615)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	ICFL	**1.000**	0.257	63.320	**1.000**	0.093	33.100	**1.000**	0.017	19.040
		(0.000)	(0.050)	(9.137)	(0.000)	(0.034)	(6.204)	(0.000)	(0.013)	(2.470)
	OCFL	**1.000**	0.337	78.020	**1.000**	0.175	48.260	**1.000**	0.056	26.360
		(0.000)	(0.046)	(8.502)	(0.000)	(0.040)	(7.378)	(0.000)	(0.021)	(3.912)
	SHIR	**1.000**	0.010	128.900	**1.000**	0.008	256.730	**1.000**	0.007	512.670
		(0.000)	(0.012)	(1.087)	(0.000)	(0.009)	(0.863)	(0.000)	(0.010)	(0.922)
	SMA	**1.000**	0.008	128.780	**1.000**	0.007	256.660	**1.000**	0.003	512.920
		(0.000)	(0.011)	(0.991)	(0.000)	(0.008)	(0.742)	(0.000)	(0.007)	(6.402)
	Local	0.889	0.102	264.180	0.893	0.100	524.470	0.891	0.102	1056.990
		(0.026)	(0.020)	(29.586)	(0.018)	(0.013)	(37.805)	(0.013)	(0.009)	(54.532)
	SK(har)	0.985	0.350	173.690	0.996	0.516	268.430	**1.000**	0.728	379.450
		(0.051)	(0.161)	(57.414)	(0.021)	(0.210)	(87.220)	(0.000)	(0.183)	(104.023)
	SK(gap)	**1.000**	0.571	121.000	**1.000**	0.817	166.400	**1.000**	0.964	193.300
		(0.000)	(0.109)	(20.013)	(0.000)	(0.075)	(13.775)	(0.000)	(0.025)	(4.613)
	DLSA	0.125	0.001	1.060	0.125	**0.000**	1.020	0.125	**0.000**	1.010
		(0.000)	(0.003)	(0.239)	(0.000)	(0.002)	(0.141)	(0.000)	(0.001)	(0.100)
$p = 500$	ICR	0.921	**0.000**	14.740	0.999	**0.000**	**16.000**	**1.000**	**0.000**	16.340
		(0.100)	(0.000)	(1.599)	(0.013)	(0.000)	(0.284)	(0.000)	(0.001)	(0.945)
	IP	0.958	**0.000**	**15.320**	**1.000**	**0.000**	**16.000**	**1.000**	**0.000**	**16.000**
		(0.084)	(0.000)	(1.340)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	ICFL	**1.000**	0.447	455.520	**1.000**	0.190	203.120	**1.000**	0.036	51.680
		(0.000)	(0.028)	(27.275)	(0.000)	(0.021)	(20.981)	(0.000)	(0.009)	(9.197)
	OCFL	**1.000**	0.017	32.660	**1.000**	0.002	17.960	**1.000**	0.001	17.440
		(0.000)	(0.006)	(5.498)	(0.000)	(0.001)	(1.449)	(0.000)	(0.001)	(0.903)
	SHIR	0.999	0.004	129.730	0.999	0.004	257.730	**1.000**	0.002	513.100
		(0.013)	(0.003)	(2.206)	(0.013)	(0.003)	(3.681)	(0.000)	(0.002)	(1.185)
	SMA	0.999	0.004	129.580	**1.000**	0.003	257.230	**1.000**	0.001	512.610
		(0.013)	(0.003)	(2.180)	(0.000)	(0.002)	(1.230)	(0.000)	(0.002)	(0.886)
	Local	0.849	0.034	373.360	0.846	0.033	741.420	0.844	0.034	1493.910
		(0.029)	(0.006)	(49.342)	(0.021)	(0.004)	(71.624)	(0.016)	(0.003)	(107.543)
	SK(har)	0.984	0.119	281.850	0.993	0.181	487.520	**1.000**	0.354	865.340
		(0.052)	(0.067)	(121.622)	(0.043)	(0.106)	(211.896)	(0.000)	(0.180)	(357.825)
	SK(gap)	**1.000**	0.235	248.190	**1.000**	0.418	427.740	**1.000**	0.662	667.620
		(0.000)	(0.061)	(59.641)	(0.000)	(0.061)	(59.943)	(0.000)	(0.050)	(49.399)
	DLSA	0.477	0.540	269.717	0.449	0.531	264.800	0.443	0.510	254.370
		(0.163)	(0.055)	(27.438)	(0.171)	(0.050)	(25.376)	(0.157)	(0.049)	(24.475)

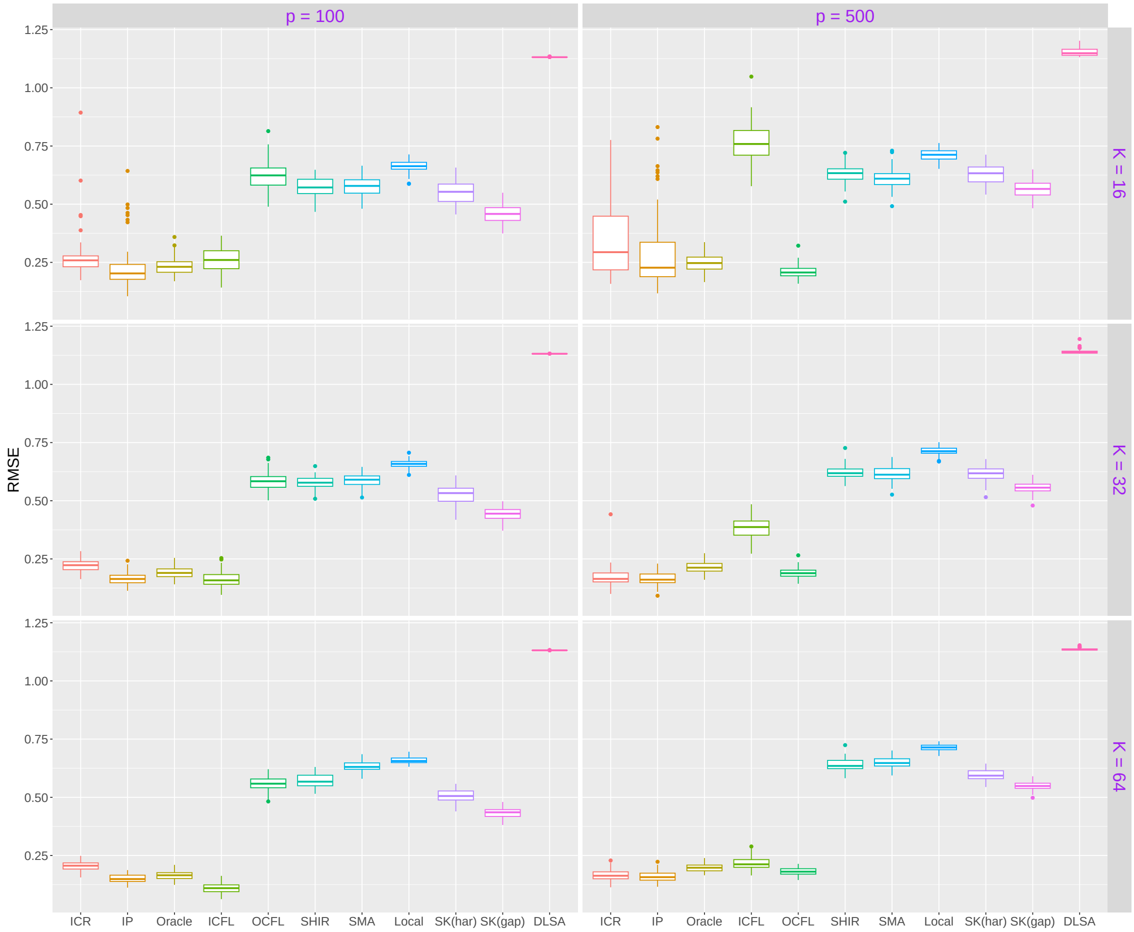


Figure 2: Boxplots of RMSE in Example 1.

grate clustering structure into distributed learning. Besides, compared to the SK(har) and SK(gap) (two-step clustered heterIM), ICR has better variable selection performance. ICR, ICFL, and OCFL all perform better than SK(har) in clustering and estimation accuracy since SK(har) often overestimates the number of clusters.

Results for Example 2 are summarized in Table 4, Table 5, and Figure 5 (Appendix B). It is observed that ICR also outperforms the other alternatives (except for IP) in variable selection and clustering accuracy and the performance improves as n increases. The performance of ICR approaches IP for larger n, which further verifies the asymptotic equivalence between ICR and IP. In this setting, although both ICFL and OCFL set the correct number of clusters, they show much worse performance in clustering accuracy compared with ICR, which leads to extremely unstable estimation results. Besides, for larger p in Example 5, compared to ICR, both ICFL and OCFL have poorer and more unstable performance (see Figure 8), since the estimation of important variables is significantly influenced by a large number of abundant parameters due to dense estimation. These results explain why we should develop a new clustered distributed learning method to address sparsity issues in high-dimensional data. Compared to Local, SHIR, and SMA have worse estimation performance, which suggests that inappropriate data integration may not help. Table 5 shows that, with ICR and SK(har), the identified number of clusters is close to the true. Further, the RI and ARI values are close to 1, indicating satisfactory clustering performance. In comparison, SK(gap) usually underestimates the number of clusters and has much lower clustering accuracy. Thus, SK methods are very sensitive to the number of clusters. This further illustrates how crucial it is to automatically select the correct value for M. Finally, due to model misspecification, DLSA (in the category of homoIM) has the worst variable selection and estimation performance in both Examples 1 and 2. Since WONDER (also in the category of homoIM), is only feasible in linear regression, so its performance can be observed in Examples 3-4. Similar to DLSA, it also shows much worse performance compared with alternatives in the category of heterIM.

Results for Examples 3-6 are summarized in Tables 6-15 and Figures 6-9 (Appendix B). The overall findings are very similar to Examples 1-2.

5. Data Application

With the emergence of technological innovations, cyberattacks (generally carried out by abnormal requests) are becoming increasingly serious and may hinder enterprise operations or interrupt critical infrastructure systems. Web logs, which are generated by systems to record detailed access information, have been widely used to detect abnormal requests in system monitoring and intrusion detection (also called anomaly detection) systems (Hu et al., 2017; Guo et al., 2021; Ünal and Dağ, 2022). Large-scale web logs are usually stored with distributed clients, and the transmission of raw logs from local clients to a central server is often infeasible. As discussed in Guo et al. (2021), on one hand, only a small part of raw logs contains useful information, and hence the full transmission of raw logs, which is very time-consuming, is not necessary. On the other hand, to facilitate log analysis, raw logs are often transferred to third-party analytic services, which increases the risks of privacy leakage. To tackle this problem, Guo et al. (2021) resorted to federated learning for anomaly detection under distributed settings. A limitation of this study is that homogeneity among clients is assumed. Hu et al. (2017) and Ünal and Dağ (2022) pointed out the heterogeneity among clients and constructed client-specific models. This study can be limited with too many redundant parameters. In this section, we apply the proposed method, which takes into consideration both multi-source heterogeneity and estimation efficiency, to a bank website logs data, which is stored in multiple interfaces (clients).

In this analysis, the request type is the binary response and takes values “normal” and “abnormal”. Our goal is to distinguish the abnormal requests from the normal ones based on the log contents, which poses a binary classification problem. There are a total of $K = 123$ URL interfaces, and the sample sizes range from 60 to 25,552. The total sample size is $N = 306,377$, and the overall percentage of abnormal requests is 21.6%. Among the 123 interfaces, 76 have the percentages of abnormal requests equal to 50%, and for the rest 47 interfaces, the percentages range from 1.2% to 75.9%. In Figure 10 (Appendix B), we present the percentages of abnormal requests for these 47 URL interfaces. The significant differences across interfaces suggest the possibility of heterogeneity.

The collected request logs can only be observed in the form of character strings. In particular, each request log contains the interface name and two submitted parameters from "POST" and "GET" access, respectively. In addition, each parameter from POST or GET access consists of a series of key-value pairs separated by "&". For demonstration, in Table 17, we present one representative record of the initial request logs from URL interface "ajaxNoSessionGetSmsAction". The unstructured parameters can be difficult to model, and we extract statistical features from the submitted parameters as follows. First, we generate two features, namely Gnum and Pnum, which are defined as the number of GET and POST key-value pairs, respectively. Second, we generate two features, namely Glen and Plen, which are defined as the total length of the GET and POST parameters, respectively. Third, we generate a series of features to measure the lengths of some key-value pairs in the GET and POST parameters, denoted by Glx and Plx, which are defined as the lengths of the (x+1)-th key-value pairs, respectively. Finally, we generate a series of features to measure the types of some key-value pairs in the GET and POST parameters, denoted by Gtx and Ptx, which are defined as the types of the (x+1)-th key-value pairs, respectively. There are three types of key-value pairs, namely "na", "str", and "num", which indicate that the key-value pair is missing, string and numeric, respectively. Take the record in Table 17 (Appendix B) as an example. We can obtain the following feature values: Gnum = 1, Glen = 9, Pnum = 4, Plen = 72, $Gl0 = 9$, $Gl1 = \cdots = Gl19 = 0$, $Pl0 = 19$, $Pl1 = 16$, $Pl2 = 10$, $Pl3 = 24$, $Pl4 = \cdots = Pl19 = 0$, $Gt0 =$"str", $Gt1 = \cdots = Gt19 =$"na", $Pt0 = \cdots = Pt2 =$"str", $Pt3 =$"num", $Pt4 = \cdots = Pt19 =$"na". Since the values of $Gl1, \dots, Gl19$ are 0 for all requests, we delete these features. To further utilize the character strings, we concatenate $Gt0 - Gt19$ and $Pt0 - Pt19$ consecutively into a sequence of strings and train the Skip-gram model (which is a popular model of word2vec) to obtain an 80-dimensional continuous word vector features, denoted by $GPw1, \dots, GPw80$. Overall, there are 105 features available for analysis.

Prior to analysis, we standardize the $p = 105$ continuous variables to have means 0 and variances 1. The proposed method identifies five nontrivial clusters (with sizes larger than one and denoted as $\mathrm{ICR}^{(1)}, \dots, \mathrm{ICR}^{(5)}$), which have sizes 37, 30, 11, 7, and 4, respectively. Additionally, there are 34 interfaces forming their own individual clusters. In Table 3, we present important variables identified by the proposed method and/or the four integrative analysis alternatives ICFL, SHIR, SMA, and DLSA, where the DLSA method identifies another 46 important variables (due to space limitation, they are not listed in Table 3). For the ICFL method, we separately pre-specify the number of clusters as 5 or 10 and the corresponding estimators are denoted by ICFL(5) and ICFL(10). Besides, due to space limitation, we only present the cluster-specific parameters $\mathrm{ICFL}_5^{(1)}, \dots, \mathrm{ICFL}_5^{(5)}$ (with cluster sizes $47, 39, 25, 9$ and 3) corresponding to 5 clusters in the ICFL(5) method, and the estimated parameters $\mathrm{ICFL}_{10}^{(1)}, \dots, \mathrm{ICFL}_{10}^{(9)}$ (with cluster sizes $42, 25, 23, 14, 6, 5, 4, 2$ and 2) in the ICFL(10) method are reported in Table 16 (Appendix B). The ICFL(10) method identifies a cluster with 0 members, which leads to 9 final clusters. It is observed that different methods lead to quite different identification and selection results. Note that, for the proposed method, we only present estimates for the nontrivial clusters. The results for the trivial clusters are omitted and available from the authors. Besides, to obtain sparse estimates in two ICFL methods, we similarly introduce a hard threshold of 0.1, as in the simulation.

At first, both our method and the two variants of ICFL result in highly imbalanced clustering structures, which partially contribute to the heterogeneity. From Table 3, we can

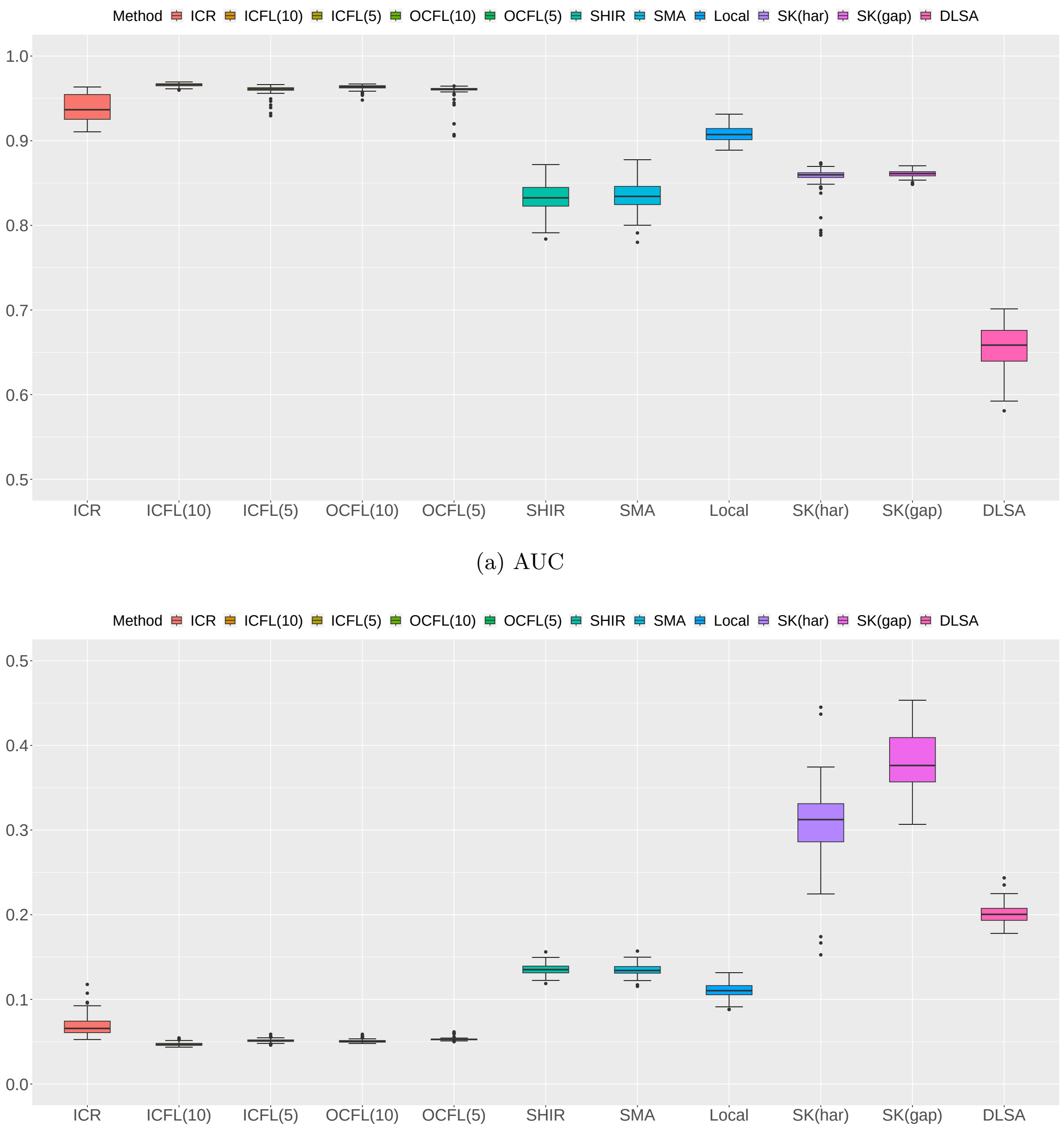
Table 3: The identified important variables and their estimates using the five integrative analysis methods in data analysis. For the proposed method, only estimates for the nontrivial clusters are shown.

Variable	ICR$^{(1)}$	ICR$^{(2)}$	ICR$^{(3)}$	ICR$^{(4)}$	ICR$^{(5)}$	ICFL$_5^{(1)}$	ICFL$_5^{(2)}$	ICFL$_5^{(3)}$	ICFL$_5^{(4)}$	ICFL$_5^{(5)}$	SHIR	SMA	DLSA
Intercept	3.321	-0.303	7.530	-1.003	8.345	1.266	1.572	-2.545	-0.254	0.981	-1.651	-1.764	-0.362
Gnum	-0.437	0.002	-0.025	0.184	-0.093	-0.213	-0.213	0.142	0.819	-0.189	–	–	0.213
Glen	-0.145	0.009	-0.281	0.008	-0.156	-0.233	-0.184	0.117	-0.175	-0.271	-0.057	-0.059	-0.362
Pnum	0.096	0.066	-0.040	-0.114	0.213	0.102	–	-0.120	0.455	–	0.044	0.004	0.139
Plen	-0.087	0.123	-0.507	-0.834	-0.348	-0.358	-0.171	-0.338	0.274	–	-0.034	–	0.070
$Gl0$	1.686	0.034	1.304	0.622	1.834	1.252	1.535	0.594	-0.528	2.021	0.270	0.159	0.134
$Pl0$	1.393	0.131	2.266	1.731	2.840	1.661	0.404	1.026	0.452	2.882	0.261	0.179	-0.081
$Pl1$	4.115	0.338	8.460	4.488	11.026	2.420	2.486	1.778	1.611	1.083	0.575	0.503	0.167
$Pl2$	3.376	0.264	7.736	6.376	3.162	2.117	2.720	1.712	1.031	0.224	0.478	0.411	0.107
$Pl3$	0.013	0.013	0.089	0.087	0.061	0.142	–	–	0.112	–	–	–	–
$Pl4$	0.018	0.030	0.010	-0.031	0.007	–	–	0.101	0.787	–	0.080	0.062	–
$Pl5$	0.005	0.013	0.004	0.175	0.067	–	0.104	–	0.332	–	0.035	0.024	–
$Pl6$	0.011	-0.002	0.013	0.041	-0.082	–	–	–	–	–	–	–	–
$Pl7$	–	–	–	–	–	–	–	–	–	–	–	–	-0.008
$Pl8$	–	–	–	–	–	–	–	–	-0.685	–	–	–	–
$Pl9$	–	–	–	–	–	–	–	–	0.257	–	0.017	0.028	–
$Pl10$	–	–	–	–	–	–	–	–	-0.161	–	0.016	0.004	–
$Pl11$	–	–	–	–	–	–	–	–	0.123	–	–	–	–
$Pl13$	–	–	–	–	–	–	–	–	0.143	–	–	–	–
$Pl14$	–	–	–	–	–	–	–	–	0.134	–	0.013	0.008	-0.011
$Pl19$	–	–	–	–	–	–	–	–	–	–	–	–	-0.035
$GPw1$	-0.032	0.017	0.069	0.119	-0.001	–	–	–	–	–	–	–	-0.028
$GPw4$	–	–	–	–	–	–	–	–	0.113	–	–	–	–
$GPw6$	0.004	0.011	0.135	-0.028	-0.241	–	–	–	–	–	–	–	0.044
$GPw11$	–	–	–	–	–	–	–	–	0.140	–	–	–	–
$GPw15$	–	–	–	–	–	–	–	–	–	–	-0.007	–	-0.024
$GPw22$	–	–	–	–	–	–	–	–	-0.114	–	–	–	–
$GPw26$	–	–	–	–	–	–	–	–	–	–	-0.009	–	–
$GPw27$	–	–	–	–	–	–	–	–	–	–	–	-0.029	–
$GPw32$	–	–	–	–	–	–	–	–	–	–	0.017	–	-0.036
$GPw33$	–	–	–	–	–	–	–	–	–	–	-0.010	–	-0.013
$GPw34$	–	–	–	–	–	–	–	–	–	–	0.006	–	0.049
$GPw38$	–	–	–	–	–	–	–	–	-0.103	–	–	–	–
$GPw39$	–	–	–	–	–	–	–	0.118	–	–	–	–	–
$GPw40$	–	–	–	–	–	–	–	–	–	–	-0.015	–	0.016
$GPw43$	–	–	–	–	–	–	–	–	-0.150	–	–	–	–
$GPw50$	–	–	–	–	–	–	–	–	-0.112	–	–	–	–
$GPw53$	–	–	–	–	–	–	–	–	-0.128	–	–	–	–
$GPw54$	–	–	–	–	–	–	–	–	-0.103	–	–	–	–
$GPw59$	–	–	–	–	–	–	–	–	0.112	–	–	–	–
$GPw63$	–	–	–	–	–	–	–	–	–	–	0.009	–	0.093
$GPw64$	–	–	–	–	–	–	–	–	–	–	-0.010	–	0.063
$GPw68$	–	–	–	–	–	–	–	0.111	-0.113	–	–	–	–
$GPw74$	–	–	–	–	–	–	–	-0.112	–	–	–	–	–

see that, with the proposed method, $Gl0$, $Pl0$, $Pl1$, and $Pl2$ all have strong positive effects for the five identified clusters, while Gnum, Glen, Pnum, and Plen have heterogeneous effects across the five clusters. For example, Pnum has negative effects in clusters $\text{ICR}^{(3)}, \text{ICR}^{(4)}$ and positive effects in clusters $\text{ICR}^{(1)}, \text{ICR}^{(2)}, \text{ICR}^{(5)}$. This suggests that requests with a longer length of the first key-value pair in GET and longer lengths of the first three key-value pairs in POST are more likely to be abnormal for most of the interfaces. This can potentially lead to a general security rule for the initial screening of abnormal requests. The heterogeneous security rules for the specific interfaces should be derived cluster-by-cluster. Here we note that the traditional anomaly detection for logs is to extract security rules from samples (El Hadj et al., 2018). Moreover, the numbers of important variables identified by the proposed ICR and the other two heterogeneous integrative methods (SHIR and SMA) are much less than that by the homogeneous DLSA method. Too many selected variables can lead to poor prediction performance, which can be further observed by the latter analysis. Additionally, the proposed analysis can reduce the number of models to 39 (5 clustered ones and 34 individual ones), which corresponds to a lower cost of maintaining models than the client-specific modeling methods (with 123 models).

With practical data, it is difficult to objectively evaluate identification and estimation results. To support our findings, we conduct a prediction evaluation. Specifically, we randomly select 4/5 of the samples and form the training data. In this selection, the normal: abnormal ratio is retained. The remaining samples form the testing data. Estimation is conducted using the training data, and we evaluate prediction performance on the testing data via several accuracy measures, which include the area under the receiver operating characteristic curve (AUC), the Brier score defined as the mean squared residuals, as well as the F_1-score at threshold value chosen to attain a false positive rate of 0.1‰ and 0.5‰ (denoted by $\text{F}_{0.1‰}$ and $\text{F}_{0.5‰}$). The F_1-score is defined as the harmonic mean of the sensitivity and positive predictive value. Note that AUC and the Brier score measure the prediction accuracy across the entire range of class distributions and error costs, while F_1-score is used to evaluate the prediction accuracy under a deterministic class distribution, which is usually obtained after using a cutoff for the predicted probabilities to enable the separation of the positive and negative classes. For comparison, we also consider the OCFL method with the pre-specified number of clusters 5 or 10, denoted by OCFL(5) and OCFL(10), respectively. This process is repeated 100 times, and the results are summarized in Figures 3 and 4.

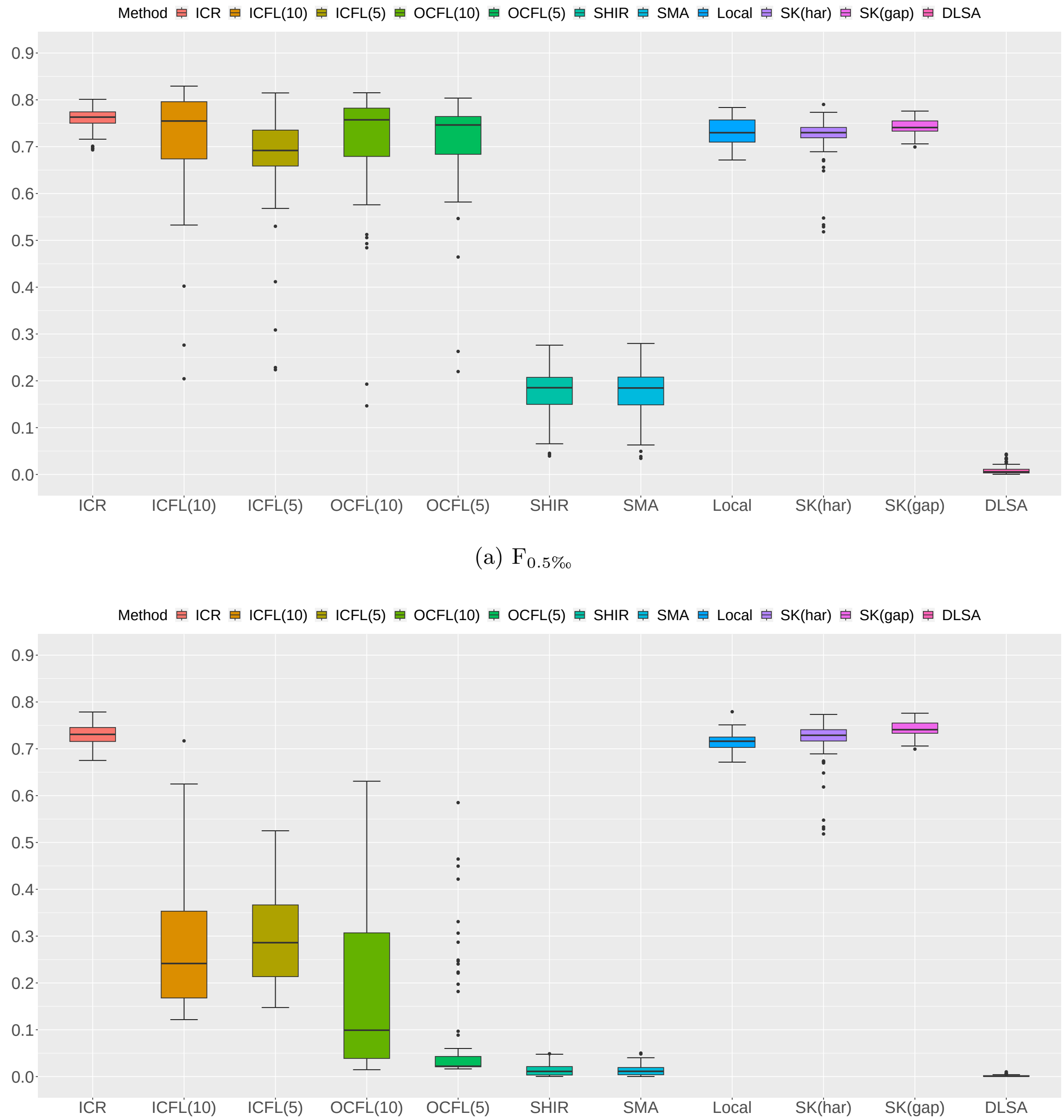
It is observed that the proposed method outperforms all alternatives except for the two variants of ICFL and OCFL (referred to as CFLs), in terms of AUC, brier score as well as two F_1-scores. From the AUC and Brier score, the CFLs show slightly better performance than the proposed method. Nevertheless, two F_1-scores reveal a completely different phenomenon. Specifically, by Figure 4(a), the proposed method outperforms the CFLs in terms of $\text{F}_{0.5‰}$, meanwhile, the CFLs exhibit significant volatility, which is consistent with their simulation results. Moreover, Figure 4(b) shows that this situation appears much more severe on the $\text{F}_{0.1‰}$. However, the proposed method shows much higher and more stable F_1-scores even if the FPR is controlled in 0.1‰. Therefore, in terms of overall prediction performance, the proposed method also beats the CFLs. Additionally, the Local method also has competitive performance, which suggests that inappropriate integration may lead to inferior prediction. Compared to the Local method, the proposed one can have better interpretability and prediction performance.



(a) AUC

(b) Brier Score

Figure 3: Boxplots of (a) AUC and (b) Brier Score based on 100 random splits in data analysis.



(a) $F_{0.5‰}$

(b) $F_{0.1‰}$

Figure 4: Boxplots of (a) $F_{0.5‰}$ and (b) $F_{0.1‰}$ based on 100 random splits in data analysis.

6. Conclusion

In this article, we have developed a new integrative data analysis method that is based on summary statistics and hence can sufficiently protect the privacy of individual clients' data. The most significant advancement is that it allows for data/model heterogeneity and can automatically identify the underlying clustering structure. Our rigorous theoretical investigation has shown that the proposed method has multiple much-desired consistency properties. Additionally, simulation and data analysis have shown its competitive numerical performance.

This study can be potentially extended in multiple directions. The same as the existing one-shot methods, the proposed analysis only demands one communication round between the local clients and the central server. Our Theorem 1 suggests that the additional error due to the aggregation of summary statistics is asymptotically negligible when we properly restrict the divergence rate of K. If we allow multiple communication rounds (which may lead to higher computational costs), this condition can be relaxed, and there is also a possibility of further improving numerical performance. The proposed method demands mild conditions on the lack-of-fit function. In numerical studies, we have investigated the logistic and linear regressions, which are involved in the framework of generalized linear models. The proposed strategy can be potentially applied to much broader models/loss functions. For example, given the loss function specified as $f(\boldsymbol{\theta}^{(k)\top}\boldsymbol{x}_i^{(k)}, y_i^{(k)})$, we can further extend it to survival data with Cox proportional hazards model (Li et al., 2023) and others. Besides, if we consider a more general loss function $f(\boldsymbol{\theta}^{(k)}, \boldsymbol{x}_i^{(k)}, y_i^{(k)})$, we can extend the proposed method to generalized additive model (Wood, 2017) or generalized additive partial linear model (Wang et al., 2011) for heterogeneity identification.

Another possible extension, as previously mentioned, is to consider different sparsity structures. For example, two-level penalized selection (Huang et al., 2017) can be conducted to allow different sparsity structures for multiple clients. Besides, for clients in which the observed variables have a spatial or temporal order, the same sparsity for a group of adjacent variables within a client can be further assumed (Li and Sang, 2019; Park et al., 2023). These aforementioned extensions will be postponed for future research.

Acknowledgments

Qingzhao Zhang's research is supported by the National Bureau of Statistics of China [2022LZ34] and the National Natural Science Foundation of China [11971404]. Shuangge Ma's research is supported by NSF 2209685 and NIH CA204120. Kuangnan Fang's research is supported by the National Natural Science Foundation of China (72071169, 72233002), the National Social Science Foundation of China (21&ZD146), and the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD910001). The authors are very grateful to the action editor and five anonymous referees for their constructive comments and suggestions that substantially improve the quality of the paper.

Appendix A. Proofs

This section includes two lemmas and the proofs of Theorems 1 and 2.

A.1 Auxiliary Lemmas

Lemma 1 *Suppose that $z_1, \dots, z_n \in \mathbb{R}$ are independent and centered sub-Gaussian random variables. Let $\mathbf{z} = (z_1, \dots, z_n)^\top$ and $\kappa = \max_{i\in[n]} \|z_i\|_{\psi_2}$. Then for any $\mathbf{a} = (a_1, \dots, a_n)^\top \in \mathbb{R}^n$ and $t > 0$, there exists a constant $C_1 > 0$ such that*

$$P(|\mathbf{a}^\top \mathbf{z}| \geq t) \leq 2\exp\bigg(-\frac{C_1 t^2}{\kappa^2 \|\mathbf{a}\|_2^2}\bigg).$$

Proof. Lemma 1 follows directly from Lemma 14.3, Chapter 14.2.2 of Bühlmann and Van De Geer (2011). □

Lemma 2 *Under Conditions (C1), (C4), (C5) and (C7), we have*

$$\max_{k\in[K]} \Big\|\widetilde{\mathbf{V}}^{(k)} - \mathbb{E}(\mathbf{V}^{*(k)})\Big\|_{\max} = O_p\left(\sqrt{\frac{q\log p}{n^*}}\right).$$

Proof. Note that,

$$\begin{aligned} &\max_{k\in[K]} \Big\|\widetilde{\mathbf{V}}^{(k)} - \mathbb{E}(\mathbf{V}^{*(k)})\Big\|_{\max} \\ \leq\ &\underbrace{\max_{k\in[K]} \Big\|\widetilde{\mathbf{V}}^{(k)} - \mathbb{E}(\widetilde{\mathbf{V}}^{(k)})\Big\|_{\max}}_{I_1} + \underbrace{\max_{k\in[K]} \Big\|\mathbb{E}(\widetilde{\mathbf{V}}^{(k)}) - \mathbb{E}(\mathbf{V}^{*(k)})\Big\|_{\max}}_{I_2}. \end{aligned} \tag{A.1}$$

At first, we derive the upper bound of I_1. Note that, under Conditions (C5) and (C7), $\max_{k\in[K]} \|\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)}\|_2 = o_p(1)$. Then by Conditions (C1) and (C4), for all $k \in [K]$ and $j_1, j_2 \in [p]$, with probability approaching 1,

$$\Big|\widetilde{\mathbf{V}}^{(k)}_{j_1 j_2}\Big| = \Big|n_k^{-1}\sum_{i=1}^{n_k} f''(\widetilde{\boldsymbol{\theta}}^{(k)\top}\boldsymbol{x}_i^{(k)}, y_i^{(k)}) x_{ij_1} x_{ij_2}\Big| \leq C_L C_x^2.$$

Then, for any given $t > 0$, Hoeffding's inequality and the union bound yield that

$$P(I_1 > t) \leq 2Kp^2 \exp\bigg\{-\frac{\min_{k\in[K]} n_k t^2}{2C_x^4 C_L^2}\bigg\},$$

which leads to $I_1 = O_p(\sqrt{\log p/n^*})$.

Next, we derive the bound of I_2. For all $k \in [K]$ and $j_1, j_2 \in [p]$, under Conditions (C1) and (C4),

$$\begin{aligned} \Big|\mathbb{E}(\widetilde{\mathbf{V}}^{(k)}_{j_1 j_2}) - \mathbb{E}(\mathbf{V}^{*(k)}_{j_1 j_2})\Big| &\leq \mathbb{E}\Big[\Big|f''(\widetilde{\boldsymbol{\theta}}^{(k)\top}\boldsymbol{x}_i^{(k)}, y_i^{(k)}) x_{ij_1} x_{ij_2} - f''(\boldsymbol{\theta}^{*(k)\top}\boldsymbol{x}_i^{(k)}, y_i^{(k)}) x_{ij_1} x_{ij_2}\Big|\Big] \\ &\leq C_L C_x^2 \mathbb{E}\Big[\Big|\boldsymbol{x}_i^{(k)\top}(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)})\Big|\Big] \\ &\leq C_L C_x^2 \bigg\{\max_{\boldsymbol{v}\in\mathcal{B}_1(\mathbf{0})} \mathbb{E}\big[(\boldsymbol{v}^\top \boldsymbol{x}_i)^2\big] \Big\|\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)}\Big\|_2^2\bigg\}^{1/2} \\ &\leq C_L C_x^3 \Big\|\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)}\Big\|_2. \end{aligned}$$

Thus,

$$I_2 \le C_L C_x^3 \max_{k\in[K]} \left\| \widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)} \right\|_2 = O_p\left(\sqrt{\frac{q\log p}{n^*}}\right).$$

Combining the bounds of I_1 and I_2 as well as (A.1), we can prove the result. □

A.2 Proof of Theorem 1

Recall the definitions of $\mathcal{L}(\boldsymbol{\theta})$ in (8) and $\mathcal{L}^{\mathcal{G}}(\boldsymbol{\psi})$ in (9), we are going to define another two objective functions. If both the true clustering structure and important covariate set are known, we can define the oracle estimator $\widehat{\boldsymbol{\theta}}^{or} = (\widehat{\boldsymbol{\theta}}_{\mathcal{A}}^{or\top}, \mathbf{0}_{(p-q)\times K}^\top)^\top$ for $\boldsymbol{\theta}$ as

$$\mathop{\arg\min}_{\boldsymbol{\theta}\in\mathcal{M}_{\mathcal{G}}, \boldsymbol{\theta}_{\mathcal{A}^c}=\mathbf{0}} \mathcal{L}^{or}(\boldsymbol{\theta}) := \frac{1}{N}\sum_{k=1}^K n_k \Big(\boldsymbol{\theta}^{(k)\top}\widetilde{\mathbf{V}}^{(k)}\boldsymbol{\theta}^{(k)} - 2\boldsymbol{\theta}^{(k)\top}\widetilde{\boldsymbol{\zeta}}^{(k)}\Big). \tag{A.2}$$

Accordingly, the oracle estimator $\widehat{\boldsymbol{\psi}}^{or} = (\widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or\top}, \mathbf{0}_{(p-q)\times M}^\top)^\top$ for $\boldsymbol{\psi}$ can be defined as

$$\mathop{\arg\min}_{\boldsymbol{\psi}\in\mathbb{R}^{p\times M}, \boldsymbol{\psi}_{\mathcal{A}^c}=\mathbf{0}} \mathcal{L}^{or,\mathcal{G}}(\boldsymbol{\psi}) := \frac{1}{N}\sum_{m=1}^M \left[\boldsymbol{\psi}^{(m)\top}\left(\sum_{k\in\mathcal{G}^{(m)}} n_k\widetilde{\mathbf{V}}^{(k)}\right)\boldsymbol{\psi}^{(m)} - 2\boldsymbol{\psi}^{(m)\top}\left(\sum_{k\in\mathcal{G}^{(m)}} n_k\widetilde{\boldsymbol{\zeta}}^{(k)}\right)\right]. \tag{A.3}$$

The results in Theorem 1 can be proved via two steps. In Step 1, we want to show that $\|\widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or} - \boldsymbol{\psi}_{\mathcal{A}}^*\|_2 = O_p(r_{1N})$, where r_{1N} is defined in (A.15). In Step 2, we further show that $\widehat{\boldsymbol{\psi}}^{or}$ is a strictly local minimizer of $\mathcal{L}^{\mathcal{G}}(\boldsymbol{\psi})$ with probability approaching 1. As a result, combining Steps 1 and 2, the sparsity and upper bound of estimation error for the nonzero coefficients can be naturally obtained.

Step 1: Let $\boldsymbol{\psi}_{\mathcal{A}} = (\boldsymbol{\psi}_{\mathcal{A}}^{(1)}, \ldots, \boldsymbol{\psi}_{\mathcal{A}}^{(M)})$ with $\boldsymbol{\psi}_{\mathcal{A}}^{(m)} = (\psi_1^{(m)}, \ldots, \psi_q^{(m)})^\top$. Based on the definition of $\mathcal{L}^{or,\mathcal{G}}(\boldsymbol{\psi})$, we can further define

$$\mathcal{L}_{\mathcal{A}}^{or,\mathcal{G}}(\boldsymbol{\psi}_{\mathcal{A}}) = \frac{1}{N}\sum_{m=1}^M \left[\boldsymbol{\psi}_{\mathcal{A}}^{(m)\top}\left(\sum_{k\in\mathcal{G}^{(m)}} n_k\widetilde{\mathbf{V}}_{\mathcal{A}\mathcal{A}}^{(k)}\right)\boldsymbol{\psi}_{\mathcal{A}}^{(m)} - 2\boldsymbol{\psi}_{\mathcal{A}}^{(m)\top}\left(\sum_{k\in\mathcal{G}^{(m)}} n_k\widetilde{\boldsymbol{\zeta}}_{\mathcal{A}}^{(k)}\right)\right]. \tag{A.4}$$

The solution of (A.4) is denoted by $\widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or} = (\widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or(1)}, \ldots, \widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or(M)})$, and the corresponding true coefficient matrix is denoted by $\boldsymbol{\psi}_{\mathcal{A}}^* = (\boldsymbol{\psi}_{\mathcal{A}}^{*(1)}, \ldots, \boldsymbol{\psi}_{\mathcal{A}}^{*(M)})$. Then we have $\mathcal{L}_{\mathcal{A}}^{or,\mathcal{G}}(\widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or}) \le \mathcal{L}_{\mathcal{A}}^{or,\mathcal{G}}(\boldsymbol{\psi}_{\mathcal{A}}^*)$, and accordingly,

$$\sum_{m=1}^M \Big(\widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or(m)\top}\widetilde{\mathbf{V}}_{\mathcal{A}\mathcal{A}}^{\mathcal{G}(m)}\widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or(m)} - 2\widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or(m)\top}\widetilde{\boldsymbol{\zeta}}_{\mathcal{A}}^{\mathcal{G}(m)}\Big) \le \sum_{m=1}^M \Big(\boldsymbol{\psi}_{\mathcal{A}}^{*(m)\top}\widetilde{\mathbf{V}}_{\mathcal{A}\mathcal{A}}^{\mathcal{G}(m)}\boldsymbol{\psi}_{\mathcal{A}}^{*(m)} - 2\boldsymbol{\psi}_{\mathcal{A}}^{*(m)\top}\widetilde{\boldsymbol{\zeta}}_{\mathcal{A}}^{\mathcal{G}(m)}\Big), \tag{A.5}$$

where

$$\widetilde{\mathbf{V}}^{\mathcal{G}(m)} = N^{-1}\sum_{k\in\mathcal{G}^{(m)}} n_k\widetilde{\mathbf{V}}^{(k)}, \quad \widetilde{\boldsymbol{\zeta}}^{\mathcal{G}(m)} = N^{-1}\sum_{k\in\mathcal{G}^{(m)}} n_k\widetilde{\boldsymbol{\zeta}}^{(k)}.$$

Motivated by Cai et al. (2022), for a vector or matrix $A(t)$ whose (i,j)th entry $A_{ij}(t)$ is a function of a scalar $t\in[0,1]$, we define $\int_0^1 A(t)dt$ as the vector or matrix with its (i,j)th entry being $\int_0^1 A_{ij}(t)dt$. Then, we can transform the term $-\widetilde{\boldsymbol{\zeta}}^{(k)}$ in (A.5) into

$$\begin{aligned}\widetilde{\mathbf{g}}^{(k)} - \widetilde{\mathbf{V}}^{(k)}\widetilde{\boldsymbol{\theta}}^{(k)} &= \mathbf{g}^{*(k)} - \widetilde{\mathbf{V}}^{(k)}\boldsymbol{\theta}^{*(k)} \\ &+ \int_0^1 \Big\{\mathbf{V}^{(k)}\Big(\big[\boldsymbol{\theta}^{*(k)} + t(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)})\big]\Big) - \widetilde{\mathbf{V}}^{(k)}\Big\}(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)})\ dt.\end{aligned} \tag{A.6}$$

Plugging (A.6) into (A.5), we have

$$\begin{aligned}&\sum_{m=1}^{M}\bigg[\Big(\widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or(m)} - \boldsymbol{\psi}_{\mathcal{A}}^{*(m)}\Big)^{\top}\widetilde{\mathbf{V}}_{\mathcal{A}\mathcal{A}}^{\mathcal{G}(m)}\Big(\widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or(m)} - \boldsymbol{\psi}_{\mathcal{A}}^{*(m)}\Big)\bigg] \\ &\leq 2\sum_{m=1}^{M}\Big(\boldsymbol{\psi}_{\mathcal{A}}^{*(m)} - \widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or(m)}\Big)^{\top}\bigg[N^{-1}\sum_{k\in\mathcal{G}^{(m)}} n_k\mathbf{g}^{*(k)}\bigg]_{\mathcal{A}} + 2\sum_{m=1}^{M}\Big(\boldsymbol{\psi}_{\mathcal{A}}^{*(m)} - \widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or(m)}\Big)^{\top} \\ &\quad\times\bigg[N^{-1}\sum_{k\in\mathcal{G}^{(m)}} n_k\int_0^1\Big\{\mathbf{V}^{(k)}\Big(\big[\boldsymbol{\theta}^{*(k)} + t(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)})\big]\Big) - \widetilde{\mathbf{V}}^{(k)}\Big\}(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)})\ dt\bigg]_{\mathcal{A}}.\end{aligned} \tag{A.7}$$

Let $\boldsymbol{\alpha} = \big[\mathrm{vec}(\widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or}) - \mathrm{vec}(\boldsymbol{\psi}_{\mathcal{A}}^{*})\big]$ and $\widetilde{\mathbf{V}}^{(\mathcal{G},\mathcal{A})} = \mathrm{bdiag}(\widetilde{\mathbf{V}}_{\mathcal{A}\mathcal{A}}^{\mathcal{G}(1)},\ldots,\widetilde{\mathbf{V}}_{\mathcal{A}\mathcal{A}}^{\mathcal{G}(M)})$, where $\mathrm{vec}(\mathbf{A})$ is a vectorization of the matrix $\mathbf{A}$ by columns and $\mathrm{bdiag}(\mathbf{A}_1,\ldots,\mathbf{A}_M)$ denotes the block diagonal matrix with the diagonal elements being $\mathbf{A}_1,\ldots,\mathbf{A}_M$. Besides, we denote $\boldsymbol{\xi} = \big(\boldsymbol{\xi}^{(1)},\ldots,\boldsymbol{\xi}^{(M)}\big)$ and $\boldsymbol{\eta} = \big(\boldsymbol{\eta}^{(1)},\ldots,\boldsymbol{\eta}^{(M)}\big)$, where

$$\boldsymbol{\xi}^{(m)} = \bigg[N^{-1}\sum_{k\in\mathcal{G}^{(m)}} n_k\mathbf{g}^{*(k)}\bigg],$$

$$\boldsymbol{\eta}^{(m)} = \bigg[N^{-1}\sum_{k\in\mathcal{G}^{(m)}} n_k\int_0^1\Big\{\mathbf{V}^{(k)}\Big(\big[\boldsymbol{\theta}^{*(k)} + t(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)})\big]\Big) - \widetilde{\mathbf{V}}^{(k)}\Big\}(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)})\ dt\bigg].$$

Then by (A.7), we have

$$\boldsymbol{\alpha}^{\top}\widetilde{\mathbf{V}}^{(\mathcal{G},\mathcal{A})}\boldsymbol{\alpha} \leq \big|\boldsymbol{\alpha}^{\top}\mathrm{vec}(\boldsymbol{\xi}_{\mathcal{A}})\big| + \big|\boldsymbol{\alpha}^{\top}\mathrm{vec}(\boldsymbol{\eta}_{\mathcal{A}})\big|. \tag{A.8}$$

Since $\boldsymbol{\alpha}^{\top}\widetilde{\mathbf{V}}^{(\mathcal{G},\mathcal{A})}\boldsymbol{\alpha} \geq \Lambda_{\min}(\widetilde{\mathbf{V}}^{(\mathcal{G},\mathcal{A})})\|\boldsymbol{\alpha}\|_2^2$, (A.8) turns to be

$$\|\boldsymbol{\alpha}\|_2^2\Lambda_{\min}(\widetilde{\mathbf{V}}^{(\mathcal{G},\mathcal{A})}) \leq \big|\boldsymbol{\alpha}^{\top}\mathrm{vec}(\boldsymbol{\xi}_{\mathcal{A}})\big| + \big|\boldsymbol{\alpha}^{\top}\mathrm{vec}(\boldsymbol{\eta}_{\mathcal{A}})\big|. \tag{A.9}$$

To obtain the lower bound of $\Lambda_{\min}(\widetilde{\mathbf{V}}_{\mathcal{A}\mathcal{A}}^{(k)})$, noting that for each $k\in[K]$, we have

$$\begin{aligned}\Lambda_{\min}(\widetilde{\mathbf{V}}_{\mathcal{A}\mathcal{A}}^{(k)}) &\geq \Lambda_{\min}\Big[\widetilde{\mathbf{V}}_{\mathcal{A}\mathcal{A}}^{(k)} - \mathbb{E}(\mathbf{V}_{\mathcal{A}\mathcal{A}}^{*(k)})\Big] + \Lambda_{\min}[\mathbb{E}(\mathbf{V}_{\mathcal{A}\mathcal{A}}^{*(k)})] \\ &\geq\ -\Big\|\widetilde{\mathbf{V}}_{\mathcal{A}\mathcal{A}}^{(k)} - \mathbb{E}(\mathbf{V}_{\mathcal{A}\mathcal{A}}^{*(k)})\Big\|_F + \Lambda_{\min}[\mathbb{E}(\mathbf{V}_{\mathcal{A}\mathcal{A}}^{*(k)})].\end{aligned} \tag{A.10}$$

By Lemma 2, since $q^3\log p \ll n^*$ by Condition (C7), for all $k\in[K]$,

$$\begin{aligned}\Big\|\widetilde{\mathbf{V}}_{\mathcal{A}\mathcal{A}}^{(k)} - \mathbb{E}(\mathbf{V}_{\mathcal{A}\mathcal{A}}^{*(k)})\Big\|_F &\leq \bigg(q^2\cdot\max_{k\in[K]}\Big\|\widetilde{\mathbf{V}}^{(k)} - \mathbb{E}(\mathbf{V}^{*(k)})\Big\|_{\max}^2\bigg)^{1/2} \\ &= O_p\left(\sqrt{q^3\log p/n^*}\right) = o_p(1).\end{aligned} \tag{A.11}$$

With (A.10), (A.11), and Condition (C3), for all $k \in [K]$, with probability approaching 1, $\Lambda_{\min}(\widetilde{\mathbf{V}}^{(k)}_{\mathcal{A}\mathcal{A}}) \geq C_{\min}/2$. Consequently, from (A.9) and the Cauchy-Schwarz inequality,

$$\frac{C_{\min} N_{\min}}{2N} \|\boldsymbol{\alpha}\|_2^2 \leq \|\boldsymbol{\alpha}\|_2 \left\|\mathrm{vec}(\boldsymbol{\xi}_{\mathcal{A}})\right\|_2 + \|\boldsymbol{\alpha}\|_2 \left\|\mathrm{vec}(\boldsymbol{\eta}_{\mathcal{A}})\right\|_2 . \tag{A.12}$$

Now we prepare to get the upper bounds of $\|\mathrm{vec}(\boldsymbol{\xi}_{\mathcal{A}})\|_2$ and $\|\mathrm{vec}(\boldsymbol{\eta}_{\mathcal{A}})\|_2$, respectively. For $\|\mathrm{vec}(\boldsymbol{\xi}_{\mathcal{A}})\|_2$, note that the product of a sub-Gaussian random variable and a bounded random variable is also sub-Gaussian and the fact of $\mathbb{E}(n_k \mathbf{g}^{*(k)}) = \mathbb{E}(\mathbf{X}^{(k)\top}\Phi^{(k)})$ with $\Phi^{(k)} = (f'(\boldsymbol{x}_1^{(k)\top}\boldsymbol{\theta}^{*(k)}, y_1^{(k)}), \ldots, f'(\boldsymbol{x}_{n_k}^{(k)\top}\boldsymbol{\theta}^{*(k)}, y_{n_k}^{(k)}))^\top$. Then, by Conditions (C1), (C2), and Lemma 1, for all $m \in [M], j \in [p]$ and any $t > 0$, we have

$$P\left(\frac{1}{\sqrt{N_m}}\left|\sum_{k \in \mathcal{G}^{(m)}} n_k \mathbf{g}_j^{*(k)}\right| \geq t\right) \leq 2\exp\left(-\frac{C_1 t^2}{C_x^2 \kappa_x^2}\right).$$

Accordingly, there exists a constant $C_2 > 0$ such that

$$\mathbb{E}\left[\left(\sum_{k \in \mathcal{G}^{(m)}} n_k \mathbf{g}_j^{*(k)}\right)^2\right] \leq C_2 N_m.$$

Then, by Markov's inequality,

$$\begin{aligned} P\left(\|\mathrm{vec}(\boldsymbol{\xi}_{\mathcal{A}})\|_2^2 \geq t^2\right) &\leq \frac{\sum_{m=1}^M \mathbb{E}(\|\boldsymbol{\xi}_{\mathcal{A}}^{(m)}\|_2^2)}{t^2} \\ &\leq \frac{\sum_{m=1}^M \sum_{j \in \mathcal{A}} \mathbb{E}\left[\left(\sum_{k \in \mathcal{G}^{(m)}} n_k \mathbf{g}_j^{*(k)}\right)^2\right]}{N^2 t^2} \leq \frac{C_2 q}{N t^2}, \end{aligned} \tag{A.13}$$

which leads to $\|\mathrm{vec}(\boldsymbol{\xi}_{\mathcal{A}})\|_2 = O_p(\sqrt{q/N})$.

For $\|\mathrm{vec}(\boldsymbol{\eta}_{\mathcal{A}})\|_2$, note that, for all $k \in [K]$ and any $t \in [0, 1]$, by Conditions (C1), (C4) and (C5), we have

$$\begin{aligned} &\left\|\left\{\mathbf{V}^{(k)}\left(\left[\boldsymbol{\theta}^{*(k)} + t(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)})\right]\right) - \widetilde{\mathbf{V}}^{(k)}\right\}(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)})\right\|_\infty \\ =& \left\|\frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{x}_i^{(k)}\boldsymbol{x}_i^{(k)\top}\left\{f''\left([\boldsymbol{\theta}^{*(k)} + t(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)})]^\top \boldsymbol{x}_i^{(k)}, y_i^{(k)}\right) - f''(\widetilde{\boldsymbol{\theta}}^{(k)\top}\boldsymbol{x}_i^{(k)}, y_i^{(k)})\right\}(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)})\right\|_\infty \\ \leq& \frac{\max_{i,j,k}|x_{ij}^{(k)}|}{n_k}\sum_{i=1}^{n_k}\left|(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)})^\top \boldsymbol{x}_i^{(k)}\right| \cdot C_L\left|(1-t)(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)})^\top \boldsymbol{x}_i^{(k)}\right| \\ \leq& \frac{C_L C_x}{n_k}\left\|\mathbf{X}^{(k)}(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)})\right\|_2^2 = O_p\left(\frac{q\log p}{n^*}\right). \end{aligned}$$

Thus, we have

$$\begin{aligned}\|\mathrm{vec}(\boldsymbol{\eta}_{\mathcal{A}})\|_2 &= \sqrt{\sum_{m=1}^{M}\left\|\boldsymbol{\eta}_{\mathcal{A}}^{(m)}\right\|_2^2} \\ &\leq \sqrt{\sum_{m=1}^{M}(\sqrt{q}\|\boldsymbol{\eta}^{(m)}\|_\infty)^2} \\ &= O_p\left(\sqrt{\frac{\sum_{m=1}^{M}|\mathcal{G}^{(m)}|^2 q^3(\log p)^2}{N^2}}\right) \\ &= O_p\left(\frac{M^{1/2}|\mathcal{G}_{\max}|q^{3/2}\log p}{N}\right).\end{aligned} \tag{A.14}$$

Combining (A.12)–(A.14), we have

$$\|\boldsymbol{\alpha}\|_2 \leq O_p(r_{1N}), \tag{A.15}$$

where

$$r_{1N} = \sqrt{\frac{(K/|\mathcal{G}_{\min}|)q}{N_{\min}}} + \frac{|\mathcal{G}_{\max}|M^{1/2}q^{3/2}\log p}{N_{\min}}.$$

Step 2: Let $\widehat{\boldsymbol{\psi}}^{or} = (\widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or\top}, \mathbf{0}_{(p-q)\times M}^{\top})^{\top}$ and define

$$\mathcal{L}_1^{\mathcal{G}}(\boldsymbol{\psi}) = \frac{1}{N}\sum_{m=1}^{M}\left[\boldsymbol{\psi}^{(m)\top}\left(\sum_{k\in\mathcal{G}^{(m)}} n_k\widetilde{\mathbf{V}}^{(k)}\right)\boldsymbol{\psi}^{(m)} - 2\boldsymbol{\psi}^{(m)\top}\left(\sum_{k\in\mathcal{G}^{(m)}} n_k\widetilde{\boldsymbol{\zeta}}^{(k)}\right)\right].$$

We show that $\widehat{\boldsymbol{\psi}}^{or}$ is a local minimizer of $\mathcal{L}^{\mathcal{G}}(\boldsymbol{\psi})$ in (9) through verifying the KKT conditions

$$\frac{\partial\mathcal{L}_1^{\mathcal{G}}(\widehat{\boldsymbol{\psi}}^{or})}{\partial\boldsymbol{\psi}_j} + \partial p_\tau\left(\sqrt{\sum_{m=1}^{M}\left(|\mathcal{G}^{(m)}|^{1/2}\widehat{\psi}_j^{or(m)}\right)^2}, \lambda_1\right)/\partial\boldsymbol{\psi}_j = \mathbf{0}, \qquad j\in\mathcal{A}, \tag{A.16}$$

$$\left\|\frac{\partial\mathcal{L}_1^{\mathcal{G}}(\widehat{\boldsymbol{\psi}}^{or})}{\partial\boldsymbol{\psi}_j}\right\|_2 \leq p_\tau'(0+,\lambda_1)\cdot|\mathcal{G}_{\min}|^{1/2}, \qquad j\in\mathcal{A}^c. \tag{A.17}$$

Note that $\widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or}$ is the solution of (A.4), then (A.16) holds if we can show that

$$\partial p_\tau\left(\sqrt{\sum_{m=1}^{M}\left(|\mathcal{G}^{(m)}|^{1/2}\widehat{\psi}_j^{or(m)}\right)^2}, \lambda_1\right)/\partial\boldsymbol{\psi}_j = \mathbf{0}.$$

By the properties of penalty function in Condition (C6), it suffices to show

$$\sqrt{\sum_{m=1}^{M}\left(|\mathcal{G}^{(m)}|^{1/2}\widehat{\psi}_j^{or(m)}\right)^2} > \tau\lambda_1, \qquad j\in\mathcal{A}.$$

As a result, the above KKT conditions hold if the following conditions hold

$$|\mathcal{G}_{\min}|^{1/2}\left\|\widehat{\boldsymbol{\psi}}_j^{or}\right\|_2 > \tau\lambda_1, \qquad j \in \mathcal{A}, \tag{A.18}$$

$$\left\|\frac{\partial \mathcal{L}_1^{\mathcal{G}}(\widehat{\boldsymbol{\psi}}^{or})}{\partial \boldsymbol{\psi}_j}\right\|_2 \le \lambda_1 |\mathcal{G}_{\min}|^{1/2}, \qquad j \in \mathcal{A}^c. \tag{A.19}$$

The similar proof logic has also been employed in Huang et al. (2010) and Fan and Lv (2011). At first, we show that Condition (A.18) is satisfied with probability approaching 1. By the triangle inequality and (A.15), when N is sufficiently large,

$$\begin{aligned} |\mathcal{G}_{\min}|^{1/2}\min_{j\in\mathcal{A}}\left\|\widehat{\boldsymbol{\psi}}_j^{or}\right\|_2 &\ge |\mathcal{G}_{\min}|^{1/2}\left(\min_{j\in\mathcal{A}}\|\boldsymbol{\psi}_j^*\|_2 - \max_{j\in\mathcal{A}}\|\widehat{\boldsymbol{\psi}}_j^{or}-\boldsymbol{\psi}_j^*\|_2\right)\\ &\ge |\mathcal{G}_{\min}|^{1/2}\left(\min_{j\in\mathcal{A}}\|\boldsymbol{\psi}_j^*\|_2 - \|\boldsymbol{\alpha}\|_2\right) \ge |\mathcal{G}_{\min}|^{1/2}(d_1 - Cr_{1N}) > \tau\lambda_1, \end{aligned}$$

where C is a constant. The last inequality is satisfied since $|\mathcal{G}_{\min}|^{1/2}d_1 > \tau\lambda_1$ and $\lambda_1 \gg |\mathcal{G}_{\min}|^{1/2}r_{1N}$. Accordingly, (A.18) is satisfied with probability approaching 1 when $N \to \infty$.

Second, we show that Condition (A.19) is satisfied with probability approaching 1. Note that $\left\|\partial\mathcal{L}_1^{\mathcal{G}}(\widehat{\boldsymbol{\psi}}^{or})/\partial\boldsymbol{\psi}_j\right\|_2 = \sqrt{\sum_{m=1}^M(\partial\mathcal{L}_1^{\mathcal{G}}(\widehat{\boldsymbol{\psi}}^{or})/\partial\boldsymbol{\psi}_j^{(m)})^2}$. Then Condition (A.19) holds if

$$\left\|\frac{\partial \mathcal{L}_1^{\mathcal{G}}(\widehat{\boldsymbol{\psi}}^{or})}{\partial \boldsymbol{\psi}_{\mathcal{A}^c}^{(m)}}\right\|_\infty \le \lambda_1(|\mathcal{G}_{\min}|/M)^{1/2}, \quad m \in [M]. \tag{A.20}$$

Since for each $m \in [M]$, $\partial\mathcal{L}^{or,\mathcal{G}}(\widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or})/\partial\boldsymbol{\psi}_{\mathcal{A}}^{(m)} = \mathbf{0}$, we have

$$\widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or(m)} - \boldsymbol{\psi}_{\mathcal{A}}^{*(m)} = \left(\widetilde{\mathbf{V}}_{\mathcal{A}\mathcal{A}}^{\mathcal{G}(m)}\right)^{-1}(\boldsymbol{\xi}_{\mathcal{A}}^{(m)} + \boldsymbol{\eta}_{\mathcal{A}}^{(m)}), \quad m \in [M]. \tag{A.21}$$

Thus, combining (A.20) and (A.21) leads to

$$\begin{aligned} \frac{1}{2}\left\|\frac{\partial \mathcal{L}_1^{\mathcal{G}}(\widehat{\boldsymbol{\psi}}^{or})}{\partial \boldsymbol{\psi}_{\mathcal{A}^c}^{(m)}}\right\|_\infty &= \left\|\widetilde{\mathbf{V}}_{\mathcal{A}^c\mathcal{A}}^{\mathcal{G}(m)}(\widehat{\boldsymbol{\psi}}_{\mathcal{A}}^{or(m)} - \boldsymbol{\psi}_{\mathcal{A}}^{*(m)}) + (\boldsymbol{\xi}_{\mathcal{A}^c}^{(m)} + \boldsymbol{\eta}_{\mathcal{A}^c}^{(m)})\right\|_\infty\\ &\le \left\|\widetilde{\mathbf{V}}_{\mathcal{A}^c\mathcal{A}}^{\mathcal{G}(m)}\left(\widetilde{\mathbf{V}}_{\mathcal{A}\mathcal{A}}^{\mathcal{G}(m)}\right)^{-1}(\boldsymbol{\xi}_{\mathcal{A}}^{(m)} + \boldsymbol{\eta}_{\mathcal{A}}^{(m)})\right\|_\infty + \left\|\boldsymbol{\xi}_{\mathcal{A}^c}^{(m)} + \boldsymbol{\eta}_{\mathcal{A}^c}^{(m)}\right\|_\infty \\ &\le \left(\left\|\widetilde{\mathbf{V}}_{\mathcal{A}^c\mathcal{A}}^{\mathcal{G}(m)}\left(\widetilde{\mathbf{V}}_{\mathcal{A}\mathcal{A}}^{\mathcal{G}(m)}\right)^{-1}\right\|_\infty + 1\right)\left(\left\|\boldsymbol{\xi}^{(m)}\right\|_\infty + \left\|\boldsymbol{\eta}^{(m)}\right\|_\infty\right). \end{aligned} \tag{A.22}$$

Following Xue et al. (2012), we can derive the upper bound of $\|\widetilde{\mathbf{V}}_{\mathcal{A}^c\mathcal{A}}^{\mathcal{G}(m)}\left(\widetilde{\mathbf{V}}_{\mathcal{A}\mathcal{A}}^{\mathcal{G}(m)}\right)^{-1}\|_\infty, m \in [M]$. Following the definition of $\widetilde{\mathbf{V}}^{\mathcal{G}(m)}(m = 1, \ldots, M)$, we can also define $\mathbf{V}^{*\mathcal{G}(m)}(m = 1, \ldots, M)$ accordingly. By Condition (C3), we define

$$c_m := \left\|\left[\mathbb{E}(\mathbf{V}_{\mathcal{A}\mathcal{A}}^{*\mathcal{G}(m)})\right]^{-1}\right\|_\infty \le \sqrt{q}\left\|\left[\mathbb{E}(\mathbf{V}_{\mathcal{A}\mathcal{A}}^{*\mathcal{G}(m)})\right]^{-1}\right\|_2 \le \frac{\sqrt{q}K}{|\mathcal{G}_{\min}|C_{\min}}. \tag{A.23}$$

Furthermore, we define

$$\begin{aligned}
\phi_m &= \left\|\widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}^c\mathcal{A}}\big(\widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)^{-1} - \mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}^c\mathcal{A}}\big)\big[\mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)\big]^{-1}\right\|_\infty, \\
\phi_{1m} &= \left\|\big(\widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)^{-1} - \big[\mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)\big]^{-1}\right\|_\infty, \\
\phi_{2m} &= \left\|\widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}} - \mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)\right\|_\infty, \qquad \phi_{3m} = \left\|\widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}^c\mathcal{A}} - \mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}^c\mathcal{A}}\big)\right\|_\infty.
\end{aligned}$$

Then by definition,

$$\begin{aligned}
\phi_m = \Big\| &\Big[\widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}^c\mathcal{A}} - \mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}^c\mathcal{A}}\big)\Big]\Big[\big(\widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)^{-1} - \big[\mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)\big]^{-1}\Big] \\
&\quad + \mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}^c\mathcal{A}}\big)\big[\mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)\big]^{-1}\Big[-\widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}} + \mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)\Big]\big(\widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)^{-1} \\
&\qquad + \Big[\widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}^c\mathcal{A}} - \mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}^c\mathcal{A}}\big)\Big]\big[\mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)\big]^{-1}\Big\|_\infty \\
&\le \phi_{3m}\phi_{1m} + \varphi^{\mathcal{G}(m)}\phi_{2m}\left\|\big(\widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)^{-1}\right\|_\infty + \phi_{3m}c_m \\
&\le \phi_{3m}\phi_{1m} + \varphi^{\mathcal{G}(m)}\phi_{2m}(c_m + \phi_{1m}) + \phi_{3m}c_m.
\end{aligned}$$

Besides, ϕ_{1m} can be reformulated as

$$\begin{aligned}
\phi_{1m} = &\left\|\big(\widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)^{-1}\Big[\mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big) - \widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\Big]\big[\mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)\big]^{-1}\right\|_\infty \\
\le &\left\|\big(\widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)^{-1}\right\|_\infty \cdot \left\|\mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big) - \widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\right\|_\infty \cdot \left\|\big[\mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)\big]^{-1}\right\|_\infty \\
\le &\ (c_m + \phi_{1m})\phi_{2m}c_m.
\end{aligned}$$

Thus, as long as $\phi_{2m}c_m < 1$, we have $\phi_{1m} \le \phi_{2m}c_m^2/(1-\phi_{2m}c_m)$, which yields

$$\phi_m \le (\phi_{3m} + \varphi^{\mathcal{G}(m)}\phi_{2m})\frac{c_m}{1-\phi_{2m}c_m}. \tag{A.24}$$

Then, by Lemma 2,

$$\begin{aligned}
\phi_{2m} = \left\|\widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}} - \mathbb{E}\big(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\big)\right\|_\infty &\le \frac{q\sum_{k\in\mathcal{G}^{(m)}} n_k \left\|\widetilde{\mathbf{V}}^{(k)}_{\mathcal{A}\mathcal{A}} - \mathbb{E}\big(\mathbf{V}^{*(k)}_{\mathcal{A}\mathcal{A}}\big)\right\|_{\max}}{N} \\
&= O_p\left(\frac{|\mathcal{G}_{\max}|}{K}\cdot\sqrt{\frac{q^3\log p}{n^*}}\right).
\end{aligned} \tag{A.25}$$

Combining (A.23), (A.25), and Condition (C7), we have

$$\phi_{2m}c_m = O_p\left(\frac{|\mathcal{G}_{\max}|}{|\mathcal{G}_{\min}|}\cdot\sqrt{\frac{q^4\log p}{n^*}}\right) = o_p(1). \tag{A.26}$$

Following the proof of (A.26), we also have

$$\phi_{3m}c_m = O_p\left(\frac{|\mathcal{G}_{\max}|}{|\mathcal{G}_{\min}|}\cdot\sqrt{\frac{q^4\log p}{n^*}}\right) = o_p(1). \tag{A.27}$$

Combining (A.24), (A.26) and (A.27), with probability approaching 1, for all $m \in [M]$ $\phi_m = o_p(\varphi_{\max})$,

$$\begin{aligned}\max_{m\in[M]} \left\|\widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}^c\mathcal{A}}\left(\widetilde{\mathbf{V}}^{\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}}\right)^{-1}\right\|_\infty &\leq \max_{m\in[M]} \phi_m + \max_{m\in[M]} \left\|\mathbb{E}(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}^c\mathcal{A}})\left[\mathbb{E}(\mathbf{V}^{*\mathcal{G}(m)}_{\mathcal{A}\mathcal{A}})\right]^{-1}\right\|_\infty \\ &\leq 2\varphi_{\max}.\end{aligned} \tag{A.28}$$

Now we consider $\|\boldsymbol{\xi}^{(m)}\|_\infty, m \in [M]$. By Conditions (C1), (C2), Lemma 1, and the union bound

$$\begin{aligned}&P\left(\max_{m\in[M]}\max_{j\in[p]}\left|\sum_{k\in\mathcal{G}^{(m)}} n_k \mathbf{g}_j^{*(k)}\right| \geq t\sqrt{N_{\max}\log p}\right) \\ &\leq 2pM\exp\left(-\frac{C_1 t^2 \log p}{C_x^2\kappa_x^2}\right) \leq 2p^{2-C_3},\end{aligned}$$

where $C_3 = C_1t^2/(C_x^2\kappa_x^2)$. When t is sufficiently large and $2 - C_3 < 0$, $2p^{2-C_3} \to 0$ when $p \to \infty$. Hence,

$$\max_{m\in[M]}\left\|\boldsymbol{\xi}^{(m)}\right\|_\infty = O_p\left(\sqrt{\frac{|\mathcal{G}_{\max}|\log p}{KN}}\right). \tag{A.29}$$

Following the proof of (A.14), we have

$$\max_{m\in[M]}\left\|\boldsymbol{\eta}^{(m)}\right\|_\infty = O_p\left(\frac{|\mathcal{G}_{\max}|q\log p}{N}\right). \tag{A.30}$$

Combining (A.22), (A.28), (A.29), (A.30), since $\varphi_{\max}r_{2N} \ll \lambda_1$, where

$$r_{2N} = \sqrt{\frac{(|\mathcal{G}_{\max}|/|\mathcal{G}_{\min}|)M\log p}{KN}} + \frac{(|\mathcal{G}_{\max}|/|\mathcal{G}_{\min}|^{1/2})M^{1/2}q\log p}{N},$$

we have verified that (A.20) is satisfied with probability approaching 1. Therefore, the KKT conditions have been verified, and the proof of Step 2 is completed. □

A.3 Proof of Theorem 2

Define

$$\mathcal{Q}(\boldsymbol{\theta}) = \underbrace{\mathcal{L}_1(\boldsymbol{\theta}) + \mathcal{P}_1(\boldsymbol{\theta})}_{\mathcal{L}(\boldsymbol{\theta})} + \mathcal{P}_2(\boldsymbol{\theta}), \qquad \mathcal{Q}^{\mathcal{G}}(\boldsymbol{\psi}) = \underbrace{\mathcal{L}_1^{\mathcal{G}}(\boldsymbol{\psi}) + \mathcal{P}_1^{\mathcal{G}}(\boldsymbol{\psi})}_{\mathcal{L}^{\mathcal{G}}(\boldsymbol{\psi})} + \mathcal{P}_2^{\mathcal{G}}(\boldsymbol{\psi}),$$

where

$$
\begin{aligned}
\mathcal{L}_1(\boldsymbol{\theta}) &= \frac{1}{N}\sum_{k=1}^{K} n_k\Big(\boldsymbol{\theta}^{(k)\top}\widetilde{\mathbf{V}}^{(k)}\boldsymbol{\theta}^{(k)} - 2\boldsymbol{\theta}^{(k)\top}\widetilde{\boldsymbol{\zeta}}^{(k)}\Big),\\
\mathcal{P}_1(\boldsymbol{\theta}) &= \sum_{j=2}^{p} p_{\mathcal{T}}\big(\|\boldsymbol{\theta}_j\|_2, \lambda_1\big),\\
\mathcal{P}_2(\boldsymbol{\theta}) &= \sum_{k<k'} p_{\mathcal{T}}\left(\left\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k')}\right\|_2, \lambda_2\right),\\
\mathcal{L}_1^{\mathcal{G}}(\boldsymbol{\psi}) &= \frac{1}{N}\sum_{m=1}^{M}\left[\boldsymbol{\psi}^{(m)\top}\left(\sum_{k\in\mathcal{G}^{(m)}} n_k\widetilde{\mathbf{V}}^{(k)}\right)\boldsymbol{\psi}^{(m)} - 2\boldsymbol{\psi}^{(m)\top}\left(\sum_{k\in\mathcal{G}^{(m)}} n_k\widetilde{\boldsymbol{\zeta}}^{(k)}\right)\right],\\
\mathcal{P}_1^{\mathcal{G}}(\boldsymbol{\psi}) &= \sum_{j=2}^{p} p_{\mathcal{T}}\left(\left[\sum_{m=1}^{M}|\mathcal{G}^{(m)}|\psi_j^{(m)2}\right]^{1/2}, \lambda_1\right),\\
\mathcal{P}_2^{\mathcal{G}}(\boldsymbol{\psi}) &= \sum_{m<m'} \big|\mathcal{G}^{(m)}\big|\big|\mathcal{G}^{(m')}\big| p_{\mathcal{T}}\left(\left\|\boldsymbol{\psi}^{(m)} - \boldsymbol{\psi}^{(m')}\right\|_2, \lambda_2\right).
\end{aligned}
$$

Define two mapping functions $\mathcal{T}^{\mathcal{G}}(\cdot)$ and $\mathcal{T}(\cdot)$. Specifically, let $\mathcal{T}^{\mathcal{G}} : \mathcal{M}_{\mathcal{G}} \to \mathbb{R}^{p\times M}$ be the function such that $\mathcal{T}^{\mathcal{G}}(\boldsymbol{\theta})$ is the $p \times M$ matrix whose mth column equals the common coefficient vector of $\boldsymbol{\theta}^{(k)}$ for $k \in \mathcal{G}^{(m)}$. Additionally, let $\mathcal{T} : \mathbb{R}^{p\times K} \to \mathbb{R}^{p\times M}$ be the function such that $\mathcal{T}(\boldsymbol{\theta}) = \big\{|\mathcal{G}^{(m)}|^{-1}\sum_{k\in\mathcal{G}^{(m)}}\boldsymbol{\theta}^{(k)}\big\}_{m=1}^{M}$. Obviously, if $\boldsymbol{\theta} \in \mathcal{M}_{\mathcal{G}}$, $\mathcal{T}^{\mathcal{G}}(\boldsymbol{\theta}) = \mathcal{T}(\boldsymbol{\theta})$.

For every $\boldsymbol{\theta} \in \mathcal{M}_{\mathcal{G}}$, we have $\mathcal{P}_2(\boldsymbol{\theta}) = \mathcal{P}_2^{\mathcal{G}}(\mathcal{T}^{\mathcal{G}}(\boldsymbol{\theta}))$. Similarly, for every $\boldsymbol{\psi} \in \mathbb{R}^{p\times M}$, we have $\mathcal{P}_2(\mathcal{T}^{\mathcal{G}^{-1}}(\boldsymbol{\psi})) = \mathcal{P}_2^{\mathcal{G}}(\boldsymbol{\psi})$. Hence,

$$
\mathcal{Q}(\boldsymbol{\theta}) = \mathcal{Q}^{\mathcal{G}}(\mathcal{T}^{\mathcal{G}}(\boldsymbol{\theta})), \qquad \mathcal{Q}^{\mathcal{G}}(\boldsymbol{\psi}) = \mathcal{Q}(\mathcal{T}^{\mathcal{G}^{-1}}(\boldsymbol{\psi})). \tag{A.31}
$$

Consider the neighborhood of $\boldsymbol{\theta}^*$, denoted by Θ, which is defined as

$$
\Theta_1 = \left\{\boldsymbol{\theta} \in \mathbb{R}^{p\times K} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_F \le C|\mathcal{G}_{\max}|^{1/2} r_{1N}\right\}.
$$

Define the event

$$
E_1 = \left\{\|\widehat{\boldsymbol{\theta}}^{or} - \boldsymbol{\theta}^*\|_F \le C|\mathcal{G}_{\max}|^{1/2} r_{1N}\right\}.
$$

Then, by the result in Theorem 1, for any $\epsilon > 0$, there exists a constant $C_\epsilon > 0$ such that for any $C \ge C_\epsilon$, $P(E_1) \ge 1 - \epsilon$. Accordingly, $\widehat{\boldsymbol{\theta}}^{or} \in \Theta_1$ with probability at least $1 - \epsilon$. Furthermore, we define another neighborhood

$$
\Theta_2 = \left\{\boldsymbol{\theta} \in \mathbb{R}^{p\times K} : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^{or}\|_F \le t_N\right\},
$$

where t_N is a positive sequence. For any $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathcal{A}}^\top, \boldsymbol{\theta}_{\mathcal{A}^c}^\top)^\top \in \Theta_1$, let $\breve{\boldsymbol{\theta}} = (\boldsymbol{\theta}_{\mathcal{A}}^\top, \mathbf{0}_{(p-q)\times K}^\top)^\top$ and $\breve{\boldsymbol{\theta}}^{\mathcal{G}} = \mathcal{T}^{\mathcal{G}^{-1}}(\mathcal{T}(\breve{\boldsymbol{\theta}}))$. Then we show that $\widehat{\boldsymbol{\theta}}^{or}$ is a strictly local minimizer of objective function (3) with probability approaching 1 through the following two steps

(a) On event E_1, $\mathcal{Q}(\breve{\boldsymbol{\theta}}^{\mathcal{G}}) > \mathcal{Q}(\widehat{\boldsymbol{\theta}}^{or})$ for any $\boldsymbol{\theta} \in \Theta_1$ and $\boldsymbol{\theta}^{\mathcal{G}} \ne \widehat{\boldsymbol{\theta}}^{or}$;

(b) On event E_1, $\mathcal{Q}(\boldsymbol{\theta}) \geq \mathcal{Q}(\breve{\boldsymbol{\theta}}) \geq \mathcal{Q}(\breve{\boldsymbol{\theta}}^{\mathcal{G}})$ for any $\boldsymbol{\theta} \in \Theta_1 \cap \Theta_2$ for a sufficiently large N.

For any $\boldsymbol{\theta} \in \Theta_1$, let $\mathcal{T}(\breve{\boldsymbol{\theta}}) = (\breve{\boldsymbol{\psi}}^{(1)}, \ldots, \breve{\boldsymbol{\psi}}^{(M)})$. Note that,

$$\mathcal{P}_2^{\mathcal{G}}(\mathcal{T}(\breve{\boldsymbol{\theta}})) = \sum_{m<m'} |\mathcal{G}^{(m)}||\mathcal{G}^{(m')}| p_\tau \left(\left\| \breve{\boldsymbol{\psi}}^{(m)} - \breve{\boldsymbol{\psi}}^{(m')} \right\|_2, \lambda_2 \right),$$

and, for any $m, m' \in [M]$ and $m \neq m'$, we have

$$\begin{aligned}
\left\| \breve{\boldsymbol{\psi}}^{(m)} - \breve{\boldsymbol{\psi}}^{(m')} \right\|_2 &\geq \min_{m,m' \in [M], m \neq m'} \left\| \boldsymbol{\psi}^{*(m)} - \boldsymbol{\psi}^{*(m')} \right\|_2 - 2 \cdot \max_{m \in [M]} \left\| \breve{\boldsymbol{\psi}}^{(m)} - \boldsymbol{\psi}^{*(m)} \right\|_2 \\
&\geq d_2 - 2 \cdot \max_{m \in [M]} \left\| |\mathcal{G}^{(m)}|^{-1} \sum_{k \in \mathcal{G}^{(m)}} (\breve{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)}) \right\|_2 \\
&\geq d_2 - 2 \cdot \max_{k \in [K]} \left\| \breve{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)} \right\|_2 \\
&\geq d_2 - 2C |\mathcal{G}_{\max}|^{1/2} r_{1N} \\
&> \tau \lambda_2,
\end{aligned} \tag{A.32}$$

where C is a constant and the last inequality follows from $d_2 > \tau\lambda_2$ and $\lambda_2 \gg |\mathcal{G}_{\max}|^{1/2} r_{1N}$. Consequently, for any $\boldsymbol{\theta} \in \Theta_1$, $\mathcal{P}_2^{\mathcal{G}}(\mathcal{T}(\breve{\boldsymbol{\theta}})) = C_N$, and $C_N > 0$ is a constant. By the result of Theorem 1, $\widehat{\boldsymbol{\psi}}^{or}$ is the local minimizer of objective function $\mathcal{L}^{\mathcal{G}}(\boldsymbol{\psi})$ with probability approaching 1. Thus we have $\mathcal{L}^{\mathcal{G}}(\widehat{\boldsymbol{\psi}}^{or}) < \mathcal{L}^{\mathcal{G}}(\mathcal{T}(\breve{\boldsymbol{\theta}}))$ for any $\boldsymbol{\theta} \in \Theta_1$ and $\mathcal{T}(\breve{\boldsymbol{\theta}}) \neq \widehat{\boldsymbol{\psi}}^{or}$. Combining this with $\mathcal{P}_2^{\mathcal{G}}(\mathcal{T}(\breve{\boldsymbol{\theta}})) = C_N$ for $\boldsymbol{\theta} \in \Theta_1$, we have $\mathcal{Q}^{\mathcal{G}}(\widehat{\boldsymbol{\psi}}^{or}) < \mathcal{Q}^{\mathcal{G}}(\mathcal{T}(\breve{\boldsymbol{\theta}}))$. By (A.31), we have

$$\mathcal{Q}^{\mathcal{G}}(\widehat{\boldsymbol{\psi}}^{or}) = \mathcal{Q}(\mathcal{T}^{\mathcal{G}^{-1}}(\widehat{\boldsymbol{\psi}}^{or})) = \mathcal{Q}(\widehat{\boldsymbol{\theta}}^{or}), \qquad \mathcal{Q}^{\mathcal{G}}(\mathcal{T}(\breve{\boldsymbol{\theta}})) = \mathcal{Q}(\mathcal{T}^{\mathcal{G}^{-1}}(\mathcal{T}(\breve{\boldsymbol{\theta}}))) = \mathcal{Q}(\breve{\boldsymbol{\theta}}^{\mathcal{G}}).$$

Accordingly, we have $\mathcal{Q}(\widehat{\boldsymbol{\theta}}^{or}) < \mathcal{Q}(\breve{\boldsymbol{\theta}}^{\mathcal{G}})$ for any $\boldsymbol{\theta} \in \Theta_1$ and $\breve{\boldsymbol{\theta}}^{\mathcal{G}} \neq \widehat{\boldsymbol{\theta}}^{or}$. This finishes the proof of the result in (a).

Next, we show that the result in (b) holds with probability approaching 1. First, we show that, on event E_1, $\mathcal{Q}(\boldsymbol{\theta}) \geq \mathcal{Q}(\breve{\boldsymbol{\theta}})$ for any $\boldsymbol{\theta} \in \Theta_1 \cap \Theta_2$. By the Taylor series expansion, we have

$$\mathcal{Q}(\boldsymbol{\theta}) - \mathcal{Q}(\breve{\boldsymbol{\theta}}) = \boldsymbol{\Omega}_{11} + \boldsymbol{\Omega}_{12} + \boldsymbol{\Omega}_{13}, \tag{A.33}$$

where

$$\begin{aligned}
\boldsymbol{\Omega}_{11} &= \sum_{j=1}^{p} \frac{\partial \mathcal{L}_1(\overline{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_j^\top} (\boldsymbol{\theta}_j - \breve{\boldsymbol{\theta}}_j), \qquad \boldsymbol{\Omega}_{12} = \sum_{j=2}^{p} \frac{\partial \mathcal{P}_1(\overline{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_j^\top} (\boldsymbol{\theta}_j - \breve{\boldsymbol{\theta}}_j), \\
\boldsymbol{\Omega}_{13} &= \sum_{k<k'} \left[p_\tau \left(\left\| \boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k')} \right\|_2, \lambda_2 \right) - p_\tau \left(\left\| \breve{\boldsymbol{\theta}}^{(k)} - \breve{\boldsymbol{\theta}}^{(k')} \right\|_2, \lambda_2 \right) \right],
\end{aligned}$$

in which $\overline{\boldsymbol{\theta}} = \delta_1 \boldsymbol{\theta} + (1 - \delta_1) \breve{\boldsymbol{\theta}}$ for some $\delta_1 \in (0, 1)$. Note that, for any $j \in \mathcal{A}$, $\boldsymbol{\theta}_j = \breve{\boldsymbol{\theta}}_j$, and for any $j \in \mathcal{A}^c$, $\breve{\boldsymbol{\theta}}_j = \mathbf{0}$ and $\overline{\boldsymbol{\theta}}_j = \delta_1 \boldsymbol{\theta}_j$. Since $p_\tau(t, \lambda_2)$ is a nondecreasing function of t with

$t \in [0, \infty)$, $\mathbf{\Omega}_{13} \geq 0$. Besides,

$$\begin{aligned}\mathbf{\Omega}_{12} = & \sum_{j \in \mathcal{A}^c} p'_\tau \left(\left\|\overline{\boldsymbol{\theta}}_j\right\|_2, \lambda_1\right) \frac{\overline{\boldsymbol{\theta}}_j^\top (\boldsymbol{\theta}_j - \breve{\boldsymbol{\theta}}_j)}{\left\|\overline{\boldsymbol{\theta}}_j\right\|_2} \\ = & \sum_{j \in \mathcal{A}^c} p'_\tau(\delta_1 \|\boldsymbol{\theta}_j\|_2, \lambda_1) \|\boldsymbol{\theta}_j\|_2 \\ \geq & \sum_{j \in \mathcal{A}^c} p'_\tau(t_N, \lambda_1) \|\boldsymbol{\theta}_j\|_2,\end{aligned} \tag{A.34}$$

where the last inequality follows from the concavity of $p_\tau(t, \lambda_1)$. Furthermore,

$$\begin{aligned}|\mathbf{\Omega}_{11}| & = \left| \sum_{j \in \mathcal{A}^c} \frac{\partial \mathcal{L}_1(\overline{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_j^\top} (\boldsymbol{\theta}_j - \breve{\boldsymbol{\theta}}_j) \right| \\ & \leq \sum_{j \in \mathcal{A}^c} \left\| \frac{\partial \mathcal{L}_1(\overline{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_j^\top} \right\|_2 \|\boldsymbol{\theta}_j\|_2 \\ & \leq \sqrt{K} \max_{k \in [K]} \left\| \frac{\partial \mathcal{L}_1(\overline{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_{\mathcal{A}^c}^{(k)}} \right\|_\infty \sum_{j \in \mathcal{A}^c} \|\boldsymbol{\theta}_j\|_2.\end{aligned} \tag{A.35}$$

Combining (A.33), (A.34), and (A.35), we have

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta}) - \mathcal{Q}(\breve{\boldsymbol{\theta}}) & = \mathbf{\Omega}_{11} + \mathbf{\Omega}_{12} + \mathbf{\Omega}_{13} \\ & \geq \sum_{j \in \mathcal{A}^c} \left[p'_\tau(t_N, \lambda_1) - \sqrt{K} \max_{k \in [K]} \left\| \partial \mathcal{L}_1(\overline{\boldsymbol{\theta}}) / \partial \boldsymbol{\theta}_{\mathcal{A}^c}^{(k)} \right\|_\infty \right] \|\boldsymbol{\theta}_j\|_2.\end{aligned} \tag{A.36}$$

Following the proof from (A.21) to (A.30), we have

$$\sqrt{K} \max_{k \in [K]} \left\| \partial \mathcal{L}_1(\overline{\boldsymbol{\theta}}) / \partial \boldsymbol{\theta}_{\mathcal{A}^c}^{(k)} \right\|_\infty = O_p \left(\varphi_{\max} \left[\sqrt{\log p / N} + K^{1/2} q \log p / N \right] + (Kq)^{1/2} t_N \right).$$

In addition, let $t_N = o(1)$, and then $p'_\tau(t_N, \lambda_1) \to \lambda_1$. Furthermore, let $(Kq)^{1/2} t_N \ll \lambda_1$, and then

$$\lambda_1 \gg \varphi_{\max} \left[r_{2N} + \sqrt{\log p / N} \right]$$

leads to

$$\lambda_1 \gg \varphi_{\max} \left[\sqrt{\log p / N} + K^{1/2} q \log p / N \right] + (Kq)^{1/2} t_N. \tag{A.37}$$

Then, by (A.36) and (A.37), when N is sufficiently large, with probability approaching 1,

$$\mathcal{Q}(\boldsymbol{\theta}) - \mathcal{Q}(\breve{\boldsymbol{\theta}}) \geq 0.$$

Next, we show that, on event E_1, $\mathcal{Q}(\breve{\boldsymbol{\theta}}) \geq Q(\breve{\boldsymbol{\theta}}^{\mathcal{G}})$ for any $\boldsymbol{\theta} \in \Theta_1 \cap \Theta_2$. By the Taylor series expansion, we have

$$\mathcal{Q}(\breve{\boldsymbol{\theta}}) - \mathcal{Q}(\breve{\boldsymbol{\theta}}^{\mathcal{G}}) = \mathbf{\Omega}_{21} + \mathbf{\Omega}_{22} + \mathbf{\Omega}_{23},$$

where

$$\boldsymbol{\Omega}_{21} = \sum_{k=1}^{K} \frac{\partial \mathcal{L}_1(\breve{\boldsymbol{\theta}}^{\hbar})}{\partial \boldsymbol{\theta}^{(k)\top}} (\breve{\boldsymbol{\theta}}^{(k)} - \breve{\boldsymbol{\theta}}^{\mathcal{G}(k)}), \qquad \boldsymbol{\Omega}_{22} = \sum_{j=2}^{p} \frac{\partial \mathcal{P}_1(\breve{\boldsymbol{\theta}}^{\hbar})}{\partial \boldsymbol{\theta}_j^{\top}} (\breve{\boldsymbol{\theta}}_j - \breve{\boldsymbol{\theta}}_j^{\mathcal{G}}),$$

$$\boldsymbol{\Omega}_{23} = \sum_{k<k'} \frac{\partial \mathcal{P}_2(\breve{\boldsymbol{\theta}}^{\hbar})}{\partial \boldsymbol{\theta}^{(k)\top}} (\breve{\boldsymbol{\theta}}^{(k)} - \breve{\boldsymbol{\theta}}^{\mathcal{G}(k)}),$$

in which $\breve{\boldsymbol{\theta}}^{\hbar} = \delta_2 \breve{\boldsymbol{\theta}} + (1-\delta_2)\breve{\boldsymbol{\theta}}^{\mathcal{G}}$ for some $\delta_2 \in (0,1)$. Next, we bound $\boldsymbol{\Omega}_{21}$, $\boldsymbol{\Omega}_{22}$, and $\boldsymbol{\Omega}_{23}$. Recall that, for any $j \in \mathcal{A}^c$, $\breve{\boldsymbol{\theta}}^{\hbar}_{\mathcal{A}^c} = \breve{\boldsymbol{\theta}}_{\mathcal{A}^c} = \breve{\boldsymbol{\theta}}^{\mathcal{G}}_{\mathcal{A}^c} = \mathbf{0}$. Then,

$$\boldsymbol{\Omega}_{22} = \sum_{j=2}^{q} \frac{\partial \mathcal{P}_1(\breve{\boldsymbol{\theta}}^{\hbar})}{\partial \boldsymbol{\theta}_j^{\top}} (\breve{\boldsymbol{\theta}}_j - \breve{\boldsymbol{\theta}}_j^{\mathcal{G}}) = \sum_{j=2}^{q} p'_\tau\left(\left\| \breve{\boldsymbol{\theta}}_j^{\hbar} \right\|_2, \lambda_1 \right) \frac{\breve{\boldsymbol{\theta}}_j^{\hbar\top} (\breve{\boldsymbol{\theta}}_j - \breve{\boldsymbol{\theta}}_j^{\mathcal{G}})}{||\breve{\boldsymbol{\theta}}_j^{\hbar}||_2}. \tag{A.38}$$

Note that,

$$\begin{aligned}
\left\| \breve{\boldsymbol{\theta}}_j^{\mathcal{G}} - \boldsymbol{\theta}_j^* \right\|_2 &= \sqrt{\sum_{m=1}^{M} |\mathcal{G}^{(m)}| \left(\frac{\sum_{k\in\mathcal{G}^{(m)}} (\breve{\theta}_j^{(k)} - \psi_j^{*(m)})}{|\mathcal{G}^{(m)}|} \right)^2} \\
&\leq \sqrt{\sum_{m=1}^{M} |\mathcal{G}^{(m)}| \times \frac{|\mathcal{G}^{(m)}| \Big[\sum_{k\in\mathcal{G}^{(m)}} \left(\breve{\theta}_j^{(k)} - \psi_j^{*(m)} \right)^2 \Big]}{|\mathcal{G}^{(m)}|^2}} \\
&= \left\| \breve{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^* \right\|_2.
\end{aligned}$$

Besides, since $\breve{\boldsymbol{\theta}}_j^{\hbar} = \delta_2 \breve{\boldsymbol{\theta}}_j + (1-\delta_2)\breve{\boldsymbol{\theta}}_j^{\mathcal{G}}$,

$$\left\| \breve{\boldsymbol{\theta}}_j^{\hbar} - \boldsymbol{\theta}_j^* \right\|_2 \leq \delta_2 \left\| \breve{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^* \right\|_2 + (1-\delta_2) \left\| \breve{\boldsymbol{\theta}}_j^{\mathcal{G}} - \boldsymbol{\theta}_j^* \right\|_2 \leq \left\| \breve{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^* \right\|_2.$$

Hence, for any $j \in \mathcal{A}$, by the triangle inequality,

$$\begin{aligned}
\left\| \breve{\boldsymbol{\theta}}_j^{\hbar} \right\|_2 \geq\ & \left\| \boldsymbol{\theta}_j^* \right\|_2 - \left\| \breve{\boldsymbol{\theta}}_j^{\hbar} - \boldsymbol{\theta}_j^* \right\|_2 \\
\geq\ & \min_{j\in\mathcal{A}} \left\| \boldsymbol{\theta}_j^* \right\|_2 - \max_{j\in\mathcal{A}} \left\| \breve{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^* \right\|_2 \\
\geq\ & |\mathcal{G}_{\min}|^{1/2} d_1 - C |\mathcal{G}_{\max}|^{1/2} r_{1N} \\
>\ & \tau\lambda_1,
\end{aligned} \tag{A.39}$$

where the last inequality follows from $|\mathcal{G}_{\min}|^{1/2} d_1 > \tau\lambda_1 \gg |\mathcal{G}_{\max}|^{1/2} r_{1N}$. Combining (A.38) and (A.39), since $p_\tau(t, \lambda_1)$ is a constant for $t \geq \tau\lambda_1$, we have $\boldsymbol{\Omega}_{22} = 0$.

Now consider $\boldsymbol{\Omega}_{23}$. Recall that, for any $j \in \mathcal{A}^c$, $\breve{\boldsymbol{\theta}}^{\hbar}_{\mathcal{A}^c} = \breve{\boldsymbol{\theta}}_{\mathcal{A}^c} = \breve{\boldsymbol{\theta}}^{\mathcal{G}}_{\mathcal{A}^c} = \mathbf{0}$. Then,

$$\begin{aligned}
\boldsymbol{\Omega}_{23} =& \sum_{k<k'} \Bigg\{ p'_\tau\left(\left\| \breve{\boldsymbol{\theta}}^{\hbar(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{\hbar(k')}_{\mathcal{A}} \right\|_2, \lambda_2 \right) \left\| \breve{\boldsymbol{\theta}}^{\hbar(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{\hbar(k')}_{\mathcal{A}} \right\|_2^{-1} \\
& \qquad \times \left(\breve{\boldsymbol{\theta}}^{\hbar(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{\hbar(k')}_{\mathcal{A}} \right)^\top \left[\left(\breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{\mathcal{G}(k)}_{\mathcal{A}} \right) - \left(\breve{\boldsymbol{\theta}}^{(k')}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{\mathcal{G}(k')}_{\mathcal{A}} \right) \right] \Bigg\} \\
=& \sum_{m=1}^{M} \sum_{k,k' \in \mathcal{G}^{(m)}, k<k'} p'_\tau\left(\left\| \breve{\boldsymbol{\theta}}^{\hbar(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{\hbar(k')}_{\mathcal{A}} \right\|_2, \lambda_2 \right) \left\| \breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{(k')}_{\mathcal{A}} \right\|_2 \\
& + \sum_{m<m'} \sum_{k \in \mathcal{G}^{(m)}, k' \in \mathcal{G}^{(m')}} \Bigg\{ p'_\tau\left(\left\| \breve{\boldsymbol{\theta}}^{\hbar(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{\hbar(k')}_{\mathcal{A}} \right\|_2, \lambda_2 \right) \left\| \breve{\boldsymbol{\theta}}^{\hbar(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{\hbar(k')}_{\mathcal{A}} \right\|_2^{-1} \\
& \qquad \times \left(\breve{\boldsymbol{\theta}}^{\hbar(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{\hbar(k')}_{\mathcal{A}} \right)^\top \left[\left(\breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{\mathcal{G}(k)}_{\mathcal{A}} \right) - \left(\breve{\boldsymbol{\theta}}^{(k')}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{\mathcal{G}(k')}_{\mathcal{A}} \right) \right] \Bigg\},
\end{aligned} \tag{A.40}$$

where the first term of the second equality follows from the fact that, when $k, k' \in \mathcal{G}^{(m)}$, $\breve{\boldsymbol{\theta}}^{\mathcal{G}(k)}_{\mathcal{A}} = \breve{\boldsymbol{\theta}}^{\mathcal{G}(k')}_{\mathcal{A}}$ and $\breve{\boldsymbol{\theta}}^{\hbar(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{\hbar(k')}_{\mathcal{A}} = \delta_2(\breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{(k')}_{\mathcal{A}})$. Note that, for any $m \in [M]$, $k \in \mathcal{G}^{(m)}$,

$$\left\| \breve{\boldsymbol{\theta}}^{\mathcal{G}(k)}_{\mathcal{A}} - \boldsymbol{\theta}^{*(k)}_{\mathcal{A}} \right\|_2 = \left\| \frac{\sum_{k \in \mathcal{G}^{(m)}} \breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}}}{|\mathcal{G}^{(m)}|} - \boldsymbol{\psi}^{*(m)}_{\mathcal{A}} \right\|_2 \leq \max_{k \in [K]} \left\| \breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \boldsymbol{\theta}^{*(k)}_{\mathcal{A}} \right\|_2.$$

And then for any $k \in [K]$, we have $\|\breve{\boldsymbol{\theta}}^{\hbar(k)}_{\mathcal{A}} - \boldsymbol{\theta}^{*(k)}_{\mathcal{A}}\|_2 \leq \max_{k \in [K]} \|\breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \boldsymbol{\theta}^{*(k)}_{\mathcal{A}}\|_2$. Hence, similar to (A.32), for any $m < m'$, $k \in \mathcal{G}^{(m)}$, $k' \in \mathcal{G}^{(m')}$,

$$\begin{aligned}
\left\| \breve{\boldsymbol{\theta}}^{\hbar(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{\hbar(k')}_{\mathcal{A}} \right\|_2 \geq& \min_{k \in \mathcal{G}^{(m)}, k' \in \mathcal{G}^{(m')}} \left\| \boldsymbol{\theta}^{*(k)}_{\mathcal{A}} - \boldsymbol{\theta}^{*(k')}_{\mathcal{A}} \right\|_2 - 2 \max_{k \in [K]} \left\| \breve{\boldsymbol{\theta}}^{\hbar(k)}_{\mathcal{A}} - \boldsymbol{\theta}^{*(k)}_{\mathcal{A}} \right\|_2 \\
\geq& \min_{k \in \mathcal{G}^{(m)}, k' \in \mathcal{G}^{(m')}} \left\| \boldsymbol{\theta}^{*(k)}_{\mathcal{A}} - \boldsymbol{\theta}^{*(k')}_{\mathcal{A}} \right\|_2 - 2 \max_{k \in [K]} \left\| \breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \boldsymbol{\theta}^{*(k)}_{\mathcal{A}} \right\|_2 \\
\geq& d_2 - 2C|\mathcal{G}_{\max}|^{1/2} r_{1N} > \tau\lambda_2.
\end{aligned} \tag{A.41}$$

Combining (A.40) and (A.41), since that

$$\left\| \breve{\boldsymbol{\theta}}^{\hbar(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{\hbar(k')}_{\mathcal{A}} \right\|_2 \leq 2 \max_{k \in [K]} \left\| \breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \boldsymbol{\theta}^{*(k)}_{\mathcal{A}} \right\|_2 \leq 2C|\mathcal{G}_{\max}|^{1/2} r_{1N},$$

we have

$$\begin{aligned}
\boldsymbol{\Omega}_{23} =& \sum_{m=1}^{M} \sum_{k,k' \in \mathcal{G}^{(m)}, k<k'} p'_\tau\left(\left\| \breve{\boldsymbol{\theta}}^{\hbar(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{\hbar(k')}_{\mathcal{A}} \right\|_2, \lambda_2 \right) \left\| \breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{(k')}_{\mathcal{A}} \right\|_2 \\
\geq& \sum_{m=1}^{M} \sum_{k,k' \in \mathcal{G}^{(m)}, k<k'} p'_\tau\left(2C|\mathcal{G}_{\max}|^{1/2} r_N, \lambda_2 \right) \left\| \breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{(k')}_{\mathcal{A}} \right\|_2 \\
\geq& \sum_{m=1}^{M} \sum_{k,k' \in \mathcal{G}^{(m)}, k<k'} \frac{\lambda_2}{2} \left\| \breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{(k')}_{\mathcal{A}} \right\|_2,
\end{aligned} \tag{A.42}$$

where the last inequality follows from $2C|\mathcal{G}_{\max}|^{1/2}r_{1N} \to 0$ when $N \to \infty$.

We now consider the bound of $\boldsymbol{\Omega}_{21}$. For $k \in [K]$, we define

$$
\begin{aligned}
\boldsymbol{\omega}^{(k)} &:= \frac{\partial \mathcal{L}_1(\breve{\boldsymbol{\theta}}^{\hbar})}{\partial \boldsymbol{\theta}^{(k)}_{\mathcal{A}}} = 2\Big\{\big[(n_k/N)\widetilde{\mathbf{V}}^{(k)}_{\mathcal{A}\mathcal{A}}\big]\breve{\boldsymbol{\theta}}^{\hbar(k)}_{\mathcal{A}} - (n_k/N)\widetilde{\boldsymbol{\zeta}}^{(k)}_{\mathcal{A}}\Big\} \\
&= 2\Bigg\{\big[(n_k/N)\widetilde{\mathbf{V}}^{(k)}_{\mathcal{A}\mathcal{A}}\big](\breve{\boldsymbol{\theta}}^{\hbar(k)}_{\mathcal{A}} - \boldsymbol{\theta}^{*(k)}_{\mathcal{A}}) + (n_k/N)\mathbf{g}^{*(k)}_{\mathcal{A}} \\
&\qquad + (n_k/N)\int_0^1 \Big\{\mathbf{V}^{(k)}\Big(\big[\boldsymbol{\theta}^{*(k)} + t(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)})\big]\Big) - \widetilde{\mathbf{V}}^{(k)}\Big\}(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{*(k)})\ dt\Bigg\} \\
&:= \boldsymbol{\omega}^{(k)}_1 + \boldsymbol{\omega}^{(k)}_2 + \boldsymbol{\omega}^{(k)}_3.
\end{aligned}
$$

Then,

$$
\begin{aligned}
\boldsymbol{\Omega}_{21} &= \sum_{k=1}^{K} \boldsymbol{\omega}^{(k)\top}(\breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{\mathcal{G}(k)}_{\mathcal{A}}) = \sum_{m=1}^{M}\sum_{k,k' \in \mathcal{G}^{(m)}} \frac{\boldsymbol{\omega}^{(k)\top}(\breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{(k')}_{\mathcal{A}})}{|\mathcal{G}^{(m)}|} \\
&= \sum_{m=1}^{M}\sum_{k,k' \in \mathcal{G}^{(m)}} \frac{\boldsymbol{\omega}^{(k')\top}(\breve{\boldsymbol{\theta}}^{(k')}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}})}{2|\mathcal{G}^{(m)}|} + \sum_{m=1}^{M}\sum_{k,k' \in \mathcal{G}^{(m)}} \frac{\boldsymbol{\omega}^{(k)\top}(\breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{(k')}_{\mathcal{A}})}{2|\mathcal{G}^{(m)}|} \\
&= \sum_{m=1}^{M}\sum_{k,k' \in \mathcal{G}^{(m)}} \frac{(\boldsymbol{\omega}^{(k)} - \boldsymbol{\omega}^{(k')})^\top(\breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{(k')}_{\mathcal{A}})}{2|\mathcal{G}^{(m)}|} \\
&= \sum_{m=1}^{M}\sum_{k,k' \in \mathcal{G}^{(m)}, k<k'} \frac{(\boldsymbol{\omega}^{(k)} - \boldsymbol{\omega}^{(k')})^\top(\breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{(k')}_{\mathcal{A}})}{|\mathcal{G}^{(m)}|}.
\end{aligned}
$$

Following the proof from (A.10) to (A.14) in Theorem 1, we can show that

$$
\begin{aligned}
\max_{k\in[K]} \left\|\boldsymbol{\omega}^{(k)}_1\right\|_2 &= O_p\big(|\mathcal{G}_{\max}|^{1/2} r_{1N}/K\big), \\
\max_{k\in[K]} \left\|\boldsymbol{\omega}^{(k)}_2\right\|_2 &= O_p\big(\sqrt{q/(KN)}\big), \\
\max_{k\in[K]} \left\|\boldsymbol{\omega}^{(k)}_3\right\|_2 &= O_p\big(q^{3/2}\log p/N\big).
\end{aligned}
$$

Then,

$$
|\boldsymbol{\Omega}_{21}| \le \sum_{m=1}^{M}\sum_{k,k' \in \mathcal{G}^{(m)}, k<k'} \frac{2\max_{k\in[K]}||\boldsymbol{\omega}^{(k)}||_2}{|\mathcal{G}_{\min}|}\left\|\breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{(k')}_{\mathcal{A}}\right\|_2. \tag{A.43}
$$

Combining (A.42) and (A.43), since $\lambda_2 \gg |\mathcal{G}_{\max}|^{1/2}r_{1N}$, we have

$$
\lambda_2 \gg \frac{|\mathcal{G}_{\max}|^{1/2} r_{1N}}{K|\mathcal{G}_{\min}|} + \sqrt{\frac{q}{KN|\mathcal{G}_{\min}|^2}} + \frac{q^{3/2}\log p}{N|\mathcal{G}_{\min}|},
$$

which leads to

$$
\mathcal{Q}(\breve{\boldsymbol{\theta}}) - \mathcal{Q}(\breve{\boldsymbol{\theta}}^{\mathcal{G}}) \ge \sum_{m=1}^{M}\sum_{k,k' \in \mathcal{G}^{(m)}, k<k'} \left\{\frac{\lambda_2}{2} - \frac{2\max_{k\in[K]}\|\boldsymbol{\omega}^{(k)}\|_2}{|\mathcal{G}_{\min}|}\right\}\left\|\breve{\boldsymbol{\theta}}^{(k)}_{\mathcal{A}} - \breve{\boldsymbol{\theta}}^{(k')}_{\mathcal{A}}\right\|_2 \ge 0,
$$

for a sufficiently large N with probability approaching 1. Thus, we have proved the result in (b). This finishes all the proofs of Theorem 2. □

Appendix B. Additional Numerical Results

This section contains simulation results for Examples 2–6 and additional data application results.

Table 4: The variable selection accuracy: mean (sd) based on 100 replicates in Example 2.

		$n = 200$			$n = 400$			$n = 800$		
	Method	TPR	FPR	MS	TPR	FPR	MS	TPR	FPR	MS
$K = 64$	ICR	**1.000**	**0.000**	**33.920**	**1.000**	**0.000**	**32.000**	**1.000**	**0.000**	**32.000**
		(0.000)	(0.000)	(3.959)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	IP	**1.000**	**0.000**	34.880	**1.000**	**0.000**	**32.000**	**1.000**	**0.000**	**32.000**
		(0.000)	(0.000)	(5.514)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	ICFL	0.903	0.099	65.280	0.896	0.016	34.680	0.900	0.002	29.360
		(0.193)	(0.062)	(27.846)	(0.203)	(0.024)	(12.777)	(0.201)	(0.006)	(7.083)
	OCFL	0.909	0.137	79.360	0.895	0.018	35.120	0.900	0.001	29.000
		(0.183)	(0.067)	(29.144)	(0.205)	(0.024)	(12.771)	(0.201)	(0.003)	(6.639)
	SHIR	0.760	0.001	389.220	**1.000**	0.001	512.700	**1.000**	0.004	533.770
		(0.042)	(0.004)	(21.718)	(0.000)	(0.003)	(6.400)	(0.000)	(0.005)	(30.463)
	SMA	0.763	0.001	390.500	**1.000**	0.001	513.330	**1.000**	0.003	528.000
		(0.042)	(0.004)	(21.367)	(0.000)	(0.003)	(9.002)	(0.000)	(0.005)	(27.852)
	Local	0.853	0.148	1309.750	0.987	0.198	1673.740	**1.000**	0.218	1794.750
		(0.020)	(0.010)	(63.847)	(0.005)	(0.010)	(56.069)	(0.001)	(0.011)	(63.438)
	SK(har)	0.945	0.526	222.880	0.996	0.934	375.480	**1.000**	0.967	387.880
		(0.098)	(0.456)	(169.339)	(0.028)	(0.137)	(51.331)	(0.000)	(0.020)	(7.335)
	SK(gap)	**1.000**	0.995	199.000	**1.000**	0.999	199.840	**1.000**	0.987	273.200
		(0.000)	(0.008)	(1.518)	(0.000)	(0.003)	(0.615)	(0.000)	(0.020)	(92.285)
	DLSA	0.125	**0.000**	1.030	0.125	**0.000**	1.000	0.125	**0.000**	1.000
		(0.000)	(0.002)	(0.171)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$K = 128$	ICR	**1.000**	**0.000**	**34.800**	**1.000**	**0.000**	32.080	**1.000**	**0.000**	**32.000**
		(0.000)	(0.000)	(5.005)	(0.000)	(0.000)	(0.800)	(0.000)	(0.000)	(0.000)
	IP	**1.000**	**0.000**	36.560	**1.000**	**0.000**	**32.000**	**1.000**	**0.000**	**32.000**
		(0.000)	(0.000)	(7.391)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	ICFL	0.915	0.033	41.240	0.940	0.005	31.760	0.890	**0.000**	28.600
		(0.189)	(0.046)	(19.750)	(0.163)	(0.016)	(8.372)	(0.208)	(0.002)	(6.760)
	OCFL	0.915	0.056	49.880	0.940	0.006	32.240	0.890	**0.000**	28.600
		(0.189)	(0.062)	(25.732)	(0.163)	(0.018)	(8.932)	(0.208)	(0.002)	(6.760)
	SHIR	0.784	**0.000**	802.560	**1.000**	0.005	1077.770	**1.000**	0.011	1152.000
		(0.068)	(0.000)	(70.120)	(0.000)	(0.005)	(63.485)	(0.000)	(0.000)	(0.000)
	SMA	0.781	**0.000**	800.000	**1.000**	0.004	1067.530	**1.000**	0.011	1152.000
		(0.074)	(0.000)	(75.835)	(0.000)	(0.005)	(60.954)	(0.000)	(0.000)	(0.000)
	Local	0.851	0.147	2602.260	0.987	0.199	3351.110	**1.000**	0.219	3599.230
		(0.015)	(0.007)	(94.466)	(0.003)	(0.007)	(84.503)	(0.000)	(0.007)	(85.526)
	SK(har)	0.899	0.505	216.400	**1.000**	0.988	394.320	**1.000**	0.980	392.440
		(0.121)	(0.497)	(189.078)	(0.000)	(0.100)	(38.757)	(0.000)	(0.141)	(52.333)
	SK(gap)	**1.000**	1.000	199.980	**1.000**	1.000	200.000	**1.000**	0.990	299.240
		(0.000)	(0.001)	(0.200)	(0.000)	(0.000)	(0.000)	(0.000)	(0.100)	(102.996)
	DLSA	0.125	**0.000**	1.000	0.125	**0.000**	1.000	0.125	**0.000**	1.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

Table 5: The clustering accuracy: mean (sd) based on 100 replicates in Example 2.

		$n = 200$				$n = 400$				$n = 800$			
	Method	$\widehat{M}$	Per	RI	ARI	$\widehat{M}$	Per	RI	ARI	$\widehat{M}$	Per	RI	ARI
$K = 64$	ICR	4.240	0.790	**0.998**	**0.995**	**4.000**	**1.000**	**1.000**	**1.000**	**4.000**	**1.000**	**1.000**	**1.000**
		(0.495)	(-)	(0.004)	(0.010)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	IP	4.360	0.740	0.997	0.992	**4.000**	**1.000**	**1.000**	**1.000**	**4.000**	**1.000**	**1.000**	**1.000**
		(0.689)	(-)	(0.005)	(0.015)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	ICFL	**4.000**	**1.000**	0.944	0.865	**4.000**	**1.000**	0.951	0.881	**4.000**	**1.000**	0.941	0.858
		(0.000)	(-)	(0.075)	(0.181)	(0.000)	(-)	(0.072)	(0.174)	(0.000)	(-)	(0.076)	(0.183)
	OCFL	**4.000**	**1.000**	0.944	0.865	**4.000**	**1.000**	0.951	0.881	**4.000**	**1.000**	0.941	0.858
		(0.000)	(-)	(0.075)	(0.181)	(0.000)	(-)	(0.072)	(0.174)	(0.000)	(-)	(0.076)	(0.183)
	SK(har)	3.980	0.980	0.994	0.984	**4.000**	**1.000**	**1.000**	**1.000**	**4.000**	**1.000**	**1.000**	**1.000**
		(0.141)	(-)	(0.020)	(0.048)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	SK(gap)	2.000	0.000	0.746	0.487	2.000	0.000	0.746	0.488	2.780	0.390	0.845	0.688
		(0.000)	(-)	(0.002)	(0.003)	(0.000)	(-)	(0.000)	(0.000)	(0.980)	(-)	(0.124)	(0.251)
$K = 128$	ICR	4.350	0.640	0.991	0.978	4.010	0.990	**1.000**	**1.000**	**4.000**	**1.000**	**1.000**	**1.000**
		(0.626)	(-)	(0.030)	(0.070)	(0.100)	(-)	(0.000)	(0.001)	(0.000)	(-)	(0.000)	(0.000)
	IP	4.570	0.620	**0.998**	**0.994**	**4.000**	**1.000**	**1.000**	**1.000**	**4.000**	**1.000**	**1.000**	**1.000**
		(0.924)	(-)	(0.003)	(0.010)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	ICFL	**4.000**	**1.000**	0.944	0.866	**4.000**	**1.000**	0.952	0.885	**4.000**	**1.000**	0.935	0.844
		(0.000)	(-)	(0.075)	(0.179)	(0.000)	(-)	(0.072)	(0.172)	(0.000)	(-)	(0.077)	(0.184)
	OCFL	**4.000**	**1.000**	0.944	0.866	**4.000**	**1.000**	0.952	0.885	**4.000**	**1.000**	0.935	0.844
		(0.000)	(-)	(0.075)	(0.179)	(0.000)	(-)	(0.072)	(0.172)	(0.000)	(-)	(0.077)	(0.184)
	SK(har)	**4.000**	0.960	0.994	0.984	3.980	0.980	0.997	0.994	3.990	0.990	0.999	0.997
		(0.201)	(-)	(0.019)	(0.044)	(0.141)	(-)	(0.018)	(0.041)	(0.100)	(-)	(0.013)	(0.029)
	SK(gap)	2.000	0.000	0.748	0.494	2.000	0.000	0.748	0.494	3.030	0.510	0.878	0.754
		(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)	(1.000)	(-)	(0.126)	(0.253)

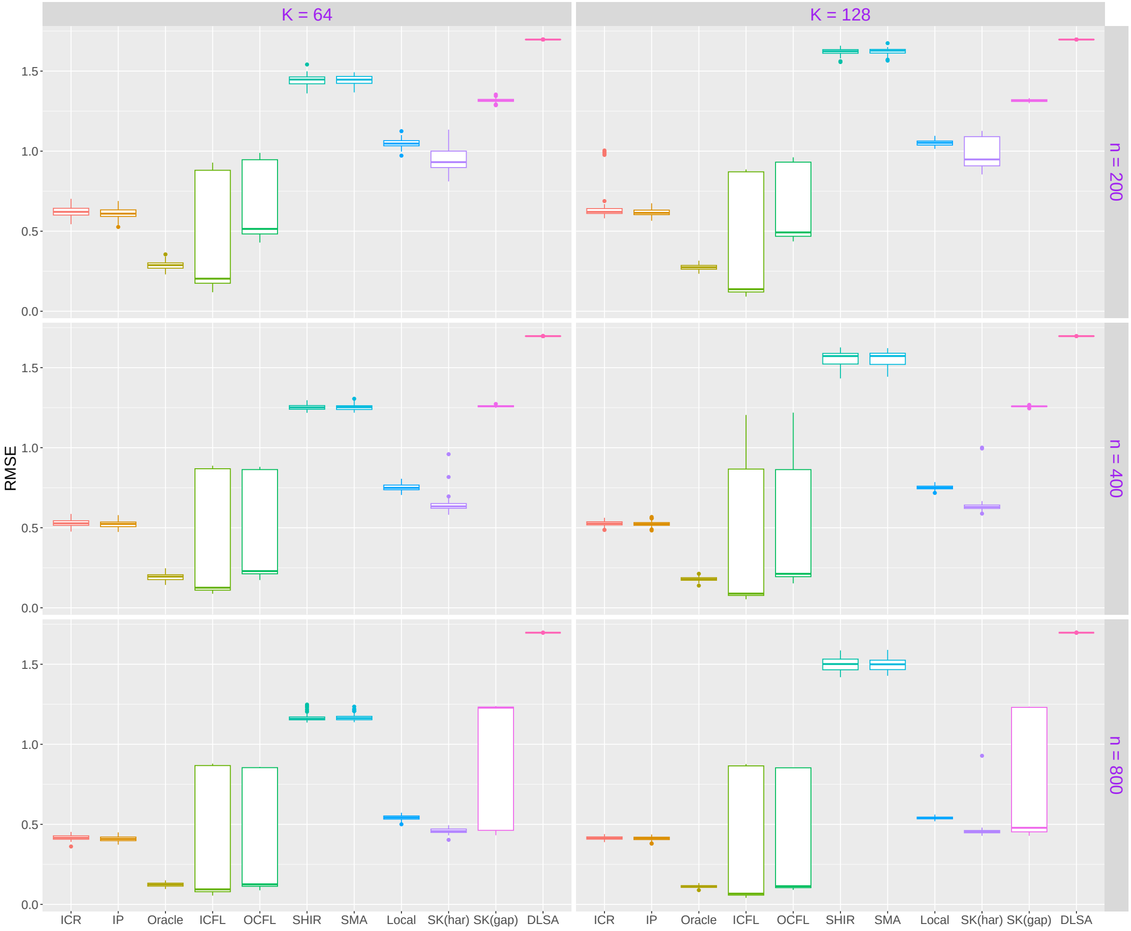


Figure 5: Boxplots of RMSE in Example 2.

Table 6: The variable selection accuracy: mean (sd) based on 100 replicates in Example 3.

		$K = 16$			$K = 32$			$K = 64$		
	Method	TPR	FPR	MS	TPR	FPR	MS	TPR	FPR	MS
$\sigma = 1$	ICR	**1.000**	**0.000**	**16.000**	**1.000**	**0.000**	**16.000**	**1.000**	**0.000**	16.060
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.343)
	IP	**1.000**	**0.000**	**16.000**	**1.000**	**0.000**	**16.000**	**1.000**	**0.000**	**16.000**
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	ICFL	0.980	0.070	28.640	0.990	0.007	17.140	**1.000**	**0.000**	16.040
		(0.141)	(0.027)	(5.832)	(0.100)	(0.010)	(2.366)	(0.000)	(0.002)	(0.281)
	OCFL	0.980	0.156	44.300	0.991	0.042	23.580	**1.000**	0.003	16.580
		(0.141)	(0.046)	(9.344)	(0.088)	(0.024)	(4.942)	(0.000)	(0.006)	(1.112)
	SHIR	**1.000**	0.019	129.920	**1.000**	0.022	258.030	**1.000**	0.015	513.410
		(0.000)	(0.016)	(2.347)	(0.000)	(0.015)	(1.374)	(0.000)	(0.014)	(1.248)
	SMA	**1.000**	0.001	128.410	**1.000**	0.002	256.190	**1.000**	0.003	512.240
		(0.000)	(0.003)	(2.257)	(0.000)	(0.005)	(0.419)	(0.000)	(0.005)	(0.495)
	Local	0.990	0.116	297.450	0.989	0.117	598.030	0.990	0.115	1184.390
		(0.008)	(0.022)	(32.006)	(0.007)	(0.015)	(43.364)	(0.004)	(0.011)	(62.616)
	SK(har)	0.994	0.382	198.580	0.998	0.571	300.420	**1.000**	0.802	383.890
		(0.033)	(0.183)	(74.091)	(0.018)	(0.201)	(95.568)	(0.000)	(0.142)	(90.112)
	SK(gap)	**1.000**	0.646	134.820	**1.000**	0.861	174.340	**1.000**	0.982	196.700
		(0.000)	(0.104)	(19.080)	(0.000)	(0.062)	(11.353)	(0.000)	(0.019)	(3.416)
	DLSA	0.445	0.431	43.250	0.424	0.402	40.370	0.395	0.372	37.420
		(0.141)	(0.104)	(10.011)	(0.195)	(0.109)	(10.977)	(0.161)	(0.102)	(9.995)
	WONDER	0.124	0.113	11.390	0.073	0.066	6.640	0.026	0.029	2.880
		(0.251)	(0.231)	(23.132)	(0.202)	(0.186)	(18.659)	(0.127)	(0.125)	(12.483)
$\sigma = 2$	ICR	0.996	**0.000**	**16.190**	**1.000**	**0.000**	**16.000**	**1.000**	**0.000**	16.060
		(0.021)	(0.003)	(1.522)	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.343)
	IP	0.996	**0.000**	16.200	**1.000**	**0.000**	**16.000**	**1.000**	**0.000**	**16.000**
		(0.021)	(0.001)	(1.435)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	ICFL	0.899	0.344	77.300	0.951	0.195	49.970	**1.000**	0.057	26.440
		(0.251)	(0.069)	(15.793)	(0.199)	(0.100)	(17.030)	(0.000)	(0.024)	(4.500)
	OCFL	0.823	0.480	101.120	0.858	0.333	73.870	0.990	0.166	46.300
		(0.254)	(0.100)	(21.438)	(0.294)	(0.116)	(23.646)	(0.068)	(0.072)	(13.353)
	SHIR	**1.000**	0.135	228.320	**1.000**	0.144	436.990	**1.000**	0.139	762.890
		(0.000)	(0.064)	(81.849)	(0.000)	(0.054)	(116.607)	(0.000)	(0.034)	(134.836)
	SMA	**1.000**	0.111	283.240	**1.000**	0.110	515.970	**1.000**	0.088	772.130
		(0.000)	(0.036)	(55.266)	(0.000)	(0.057)	(182.428)	(0.000)	(0.048)	(266.909)
	Local	0.816	0.099	250.050	0.818	0.100	504.280	0.819	0.099	1000.380
		(0.034)	(0.020)	(31.474)	(0.022)	(0.015)	(44.000)	(0.016)	(0.011)	(65.717)
	SK(har)	0.986	0.319	168.58	0.996	0.485	255.770	**1.000**	0.701	365.340
		(0.053)	(0.166)	(71.974)	(0.038)	(0.203)	(97.201)	(0.000)	(0.204)	(118.276)
	SK(gap)	**1.000**	0.555	118.370	**1.000**	0.809	165.630	**1.000**	0.967	193.840
		(0.000)	(0.150)	(26.845)	(0.000)	(0.084)	(15.860)	(0.000)	(0.026)	(4.745)
	DLSA	0.648	0.657	65.630	0.590	0.622	61.900	0.649	0.638	63.860
		(0.178)	(0.076)	(7.498)	(0.170)	(0.075)	(7.132)	(0.160)	(0.073)	(7.027)
	WONDER	0.141	0.137	13.690	0.088	0.093	9.250	0.028	0.033	3.290
		(0.249)	(0.233)	(23.212)	(0.211)	(0.212)	(21.085)	(0.112)	(0.110)	(10.931)

Table 7: The clustering accuracy: mean (sd) based on 100 replicates in Example 3.

		$K = 16$				$K = 32$				$K = 64$			
	Method	$\widehat{M}$	Per	RI	ARI	$\widehat{M}$	Per	RI	ARI	$\widehat{M}$	Per	RI	ARI
$\sigma = 1$	ICR	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	IP	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	ICFL	**2.000**	**1.000**	0.990	0.980	**2.000**	**1.000**	0.995	0.990	**2.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.074)	(0.141)	(0.000)	(-)	(0.051)	(0.100)	(0.000)	(-)	(0.000)	(0.000)
	OCFL	**2.000**	**1.000**	0.990	0.980	**2.000**	**1.000**	0.995	0.990	**2.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.074)	(0.141)	(0.000)	(-)	(0.051)	(0.100)	(0.000)	(-)	(0.000)	(0.000)
	SK(har)	4.900	0.000	0.752	0.483	5.340	0.000	0.739	0.468	4.790	0.000	0.744	0.484
		(1.418)	(-)	(0.085)	(0.182)	(1.821)	(-)	(0.088)	(0.182)	(1.274)	(-)	(0.065)	(0.132)
	SK(gap)	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
$\sigma = 2$	ICR	2.020	0.980	**0.999**	**0.998**	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**
		(0.141)	(-)	(0.008)	(0.017)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	IP	2.030	0.970	0.998	0.996	**2.000**	**1.000**	**1.000**	**1.000**	**2.000**	**1.000**	**1.000**	**1.000**
		(0.171)	(-)	(0.013)	(0.026)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	ICFL	**1.990**	**0.990**	0.926	0.860	1.970	0.970	0.964	0.930	**2.000**	**1.000**	**1.000**	**1.000**
		(0.100)	(-)	(0.183)	(0.349)	(0.171)	(-)	(0.132)	(0.256)	(0.000)	(-)	(0.000)	(0.000)
	OCFL	**1.990**	**0.990**	0.827	0.665	1.970	0.970	0.887	0.779	**2.000**	**1.000**	0.983	0.965
		(0.100)	(-)	(0.225)	(0.433)	(0.171)	(-)	(0.199)	(0.389)	(0.000)	(-)	(0.074)	(0.149)
	SK(har)	4.800	0.000	0.764	0.507	5.180	0.000	0.754	0.497	5.200	0.000	0.736	0.466
		(1.589)	(-)	(0.097)	(0.207)	(1.850)	(-)	(0.095)	(0.195)	(1.589)	(-)	(0.077)	(0.156)
	SK(gap)	2.020	0.980	0.996	0.993	2.010	0.990	0.998	0.997	**2.000**	**1.000**	0.999	0.999
		(0.141)	(-)	(0.019)	(0.039)	(0.100)	(-)	(0.013)	(0.026)	(0.000)	(-)	(0.004)	(0.009)

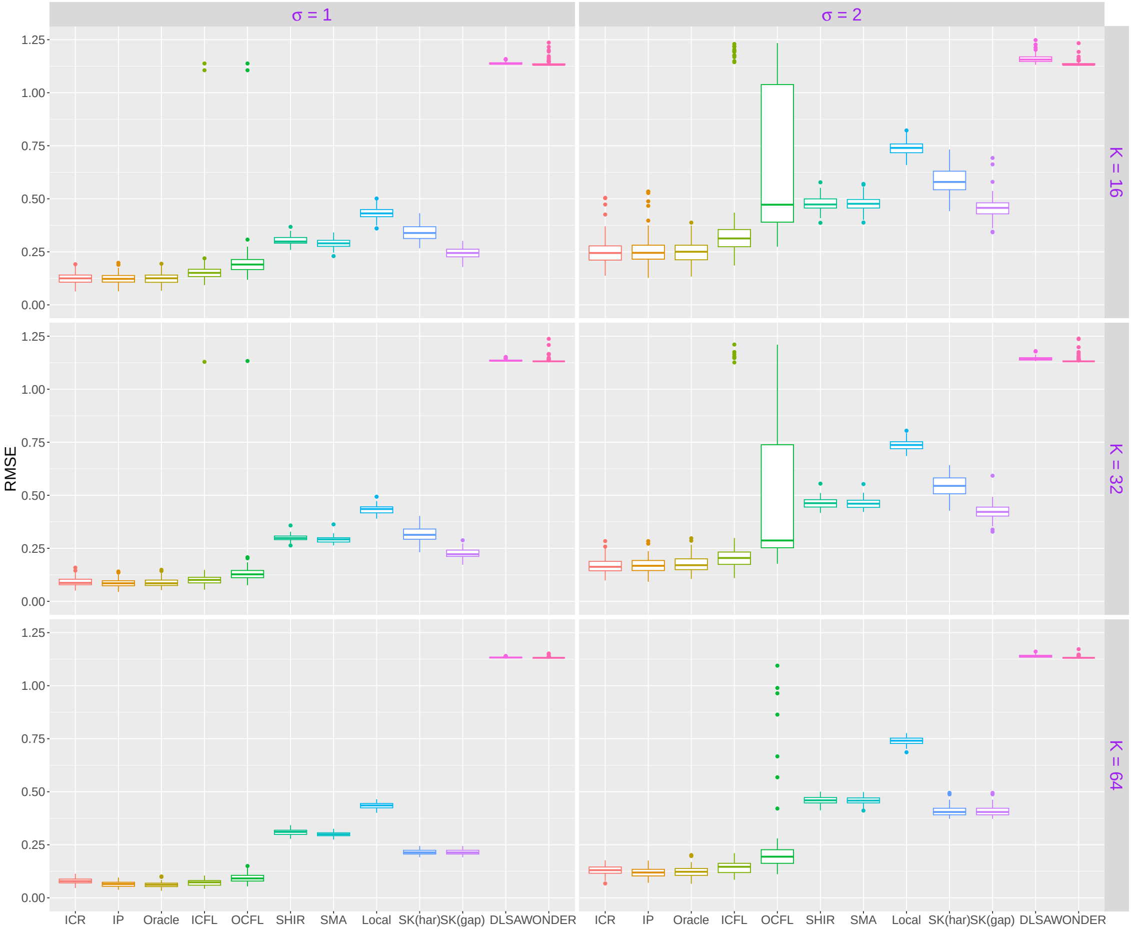


Figure 6: Boxplots of RMSE in Example 3.

Table 8: The variable selection accuracy: mean (sd) based on 100 replicates in Example 4.

		$n = 100$			$n = 200$			$n = 400$		
	Method	TPR	FPR	MS	TPR	FPR	MS	TPR	FPR	MS
$\sigma = 1$	ICR	**1.000**	**0.000**	32.040	**1.000**	**0.000**	**32.000**	**1.000**	**0.000**	**32.000**
		(0.000)	(0.001)	(0.400)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	IP	**1.000**	**0.000**	**32.000**	**1.000**	**0.000**	**32.000**	**1.000**	**0.000**	**32.000**
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	ICFL	0.915	0.012	33.250	0.885	**0.000**	28.440	0.900	**0.000**	28.880
		(0.189)	(0.057)	(22.316)	(0.211)	(0.002)	(6.891)	(0.201)	(0.002)	(6.522)
	OCFL	0.918	0.040	43.440	0.885	0.001	28.600	0.900	**0.000**	28.880
		(0.185)	(0.050)	(21.119)	(0.211)	(0.005)	(7.200)	(0.201)	(0.002)	(6.522)
	SHIR	**1.000**	0.077	531.640	**1.000**	0.081	524.450	**1.000**	0.072	523.670
		(0.000)	(0.029)	(31.566)	(0.000)	(0.030)	(17.815)	(0.000)	(0.030)	(17.975)
	SMA	**1.000**	0.060	539.570	**1.000**	0.078	524.870	**1.000**	0.073	525.040
		(0.000)	(0.026)	(38.983)	(0.000)	(0.030)	(19.013)	(0.000)	(0.028)	(21.384)
	Local	0.989	0.204	1709.120	**1.000**	0.199	1685.360	**1.000**	0.195	1658.000
		(0.005)	(0.012)	(70.651)	(0.000)	(0.011)	(63.342)	(0.000)	(0.011)	(64.948)
	SK(har)	0.999	0.932	375.080	**1.000**	0.933	375.160	**1.000**	0.937	376.680
		(0.013)	(0.137)	(50.832)	(0.000)	(0.098)	(36.120)	(0.000)	(0.036)	(13.237)
	SK(gap)	**1.000**	0.998	203.380	**1.000**	0.967	301.720	**1.000**	0.959	320.040
		(0.000)	(0.010)	(24.503)	(0.000)	(0.036)	(89.220)	(0.000)	(0.039)	(83.909)
	DLSA	0.359	0.305	30.950	0.203	0.069	7.930	0.168	0.025	3.660
		(0.167)	(0.116)	(11.400)	(0.123)	(0.067)	(6.753)	(0.076)	(0.035)	(3.514)
	WONDER	0.006	0.012	1.120	0.000	**0.000**	0.000	0.000	**0.000**	0.000
		(0.037)	(0.072)	(6.896)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$\sigma = 2$	ICR	**1.000**	**0.000**	32.400	**1.000**	**0.000**	**32.000**	**1.000**	**0.000**	32.040
		(0.000)	(0.000)	(1.752)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.400)
	IP	**1.000**	**0.000**	**32.240**	**1.000**	**0.000**	**32.000**	**1.000**	**0.000**	**32.000**
		(0.000)	(0.000)	(1.372)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	ICFL	0.916	0.175	91.030	0.915	0.048	47.080	0.905	0.008	31.840
		(0.179)	(0.101)	(40.681)	(0.189)	(0.042)	(18.872)	(0.197)	(0.030)	(13.465)
	OCFL	0.915	0.340	150.420	0.915	0.065	53.200	0.905	0.009	32.120
		(0.163)	(0.119)	(49.251)	(0.189)	(0.049)	(21.846)	(0.197)	(0.031)	(13.709)
	SHIR	**1.000**	0.270	1542.320	**1.000**	0.229	1246.210	**1.000**	0.197	1084.500
		(0.000)	(0.080)	(486.815)	(0.000)	(0.076)	(367.035)	(0.000)	(0.068)	(350.924)
	SMA	**1.000**	0.263	1575.730	**1.000**	0.230	1246.960	**1.000**	0.188	1058.500
		(0.000)	(0.0840)	(469.041)	(0.000)	(0.076)	(391.310)	(0.000)	(0.068)	(374.399)
	Local	0.719	0.127	1116.360	0.944	0.176	1519.300	0.998	0.194	1650.780
		(0.030)	(0.013)	(85.809)	(0.014)	(0.012)	(75.027)	(0.002)	(0.011)	(64.969)
	SK(har)	0.908	0.285	140.730	0.994	0.868	351.280	0.999	0.928	373.640
		(0.106)	(0.410)	(161.999)	(0.033)	(0.223)	(83.034)	(0.013)	(0.101)	(37.494)
	SK(gap)	0.998	0.980	196.360	**1.000**	0.998	199.580	**1.000**	0.991	230.660
		(0.025)	(0.100)	(18.750)	(0.000)	(0.005)	(0.955)	(0.000)	(0.022)	(68.584)
	DLSA	0.614	0.608	60.830	0.368	0.228	23.920	0.248	0.044	6.070
		(0.191)	(0.086)	(8.705)	(0.159)	(0.109)	(10.723)	(0.132)	(0.048)	(5.127)
	WONDER	0.036	0.045	4.440	0.000	**0.000**	0.000	0.000	**0.000**	0.000
		(0.136)	(0.135)	(13.444)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

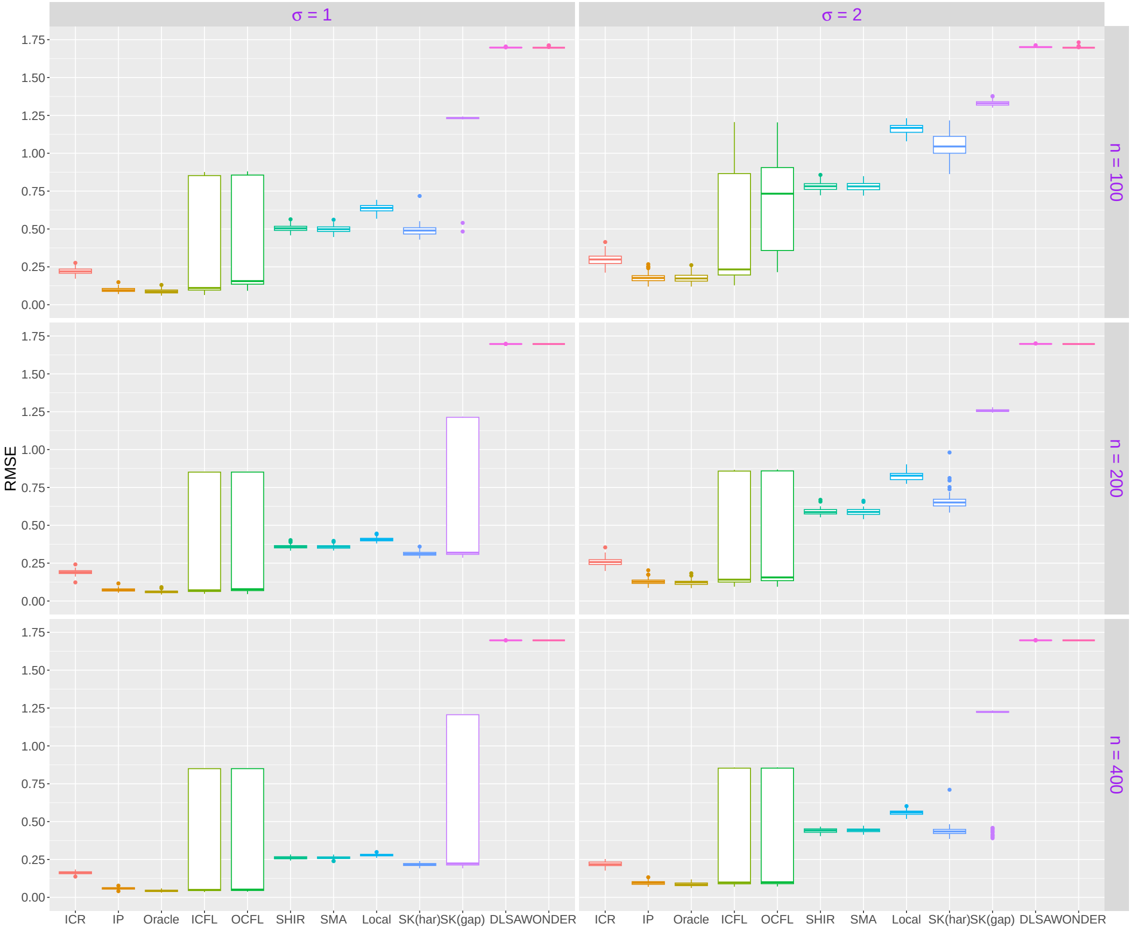


Figure 7: Boxplots of RMSE in Example 4.

Table 9: The clustering accuracy: mean (sd) based on 100 replicates in Example 4.

		$n = 100$				$n = 200$				$n = 400$			
	Method	$\widehat{M}$	Per	RI	ARI	$\widehat{M}$	Per	RI	ARI	$\widehat{M}$	Per	RI	ARI
$\sigma = 1$	ICR	**4.000**	**1.000**	**1.000**	**1.000**	**4.000**	**1.000**	**1.000**	**1.000**	**4.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	IP	**4.000**	**1.000**	**1.000**	**1.000**	**4.000**	**1.000**	**1.000**	**1.000**	**4.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	ICFL	3.930	0.930	0.956	0.894	**4.000**	**1.000**	0.943	0.863	**4.000**	**1.000**	0.944	0.865
		(0.256)	(-)	(0.067)	(0.160)	(0.000)	(-)	(0.075)	(0.181)	(0.000)	(-)	(0.075)	(0.181)
	OCFL	3.930	0.930	0.956	0.894	**4.000**	**1.000**	0.943	0.863	**4.000**	**1.000**	0.944	0.865
		(0.256)	(-)	(0.067)	(0.160)	(0.000)	(-)	(0.075)	(0.181)	(0.000)	(-)	(0.075)	(0.181)
	SK(har)	**4.000**	**1.000**	**1.000**	**1.000**	**4.000**	**1.000**	**1.000**	**1.000**	**4.000**	**1.000**	**1.000**	**1.000**
		(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	SK(gap)	2.040	0.020	0.751	0.498	3.140	0.570	0.891	0.780	3.350	0.670	0.917	0.833
		(0.281)	(-)	(0.036)	(0.072)	(0.995)	(-)	(0.126)	(0.255)	(0.936)	(-)	(0.119)	(0.240)
$\sigma = 2$	ICR	4.050	0.950	**1.000**	**0.999**	**4.000**	**1.000**	**1.000**	**1.000**	**4.000**	**1.000**	**1.000**	**1.000**
		(0.219)	(-)	(0.002)	(0.005)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	IP	**4.030**	**0.970**	**1.000**	**0.999**	**4.000**	**1.000**	**1.000**	**1.000**	**4.000**	**1.000**	**1.000**	**1.000**
		(0.171)	(-)	(0.001)	(0.004)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	ICFL	3.880	0.890	0.946	0.872	**4.000**	**1.000**	0.949	0.877	**4.000**	**1.000**	0.949	0.876
		(0.356)	(-)	(0.072)	(0.168)	(0.000)	(-)	(0.073)	(0.176)	(0.000)	(-)	(0.073)	(0.177)
	OCFL	3.880	0.890	0.930	0.832	**4.000**	**1.000**	0.949	0.877	**4.000**	**1.000**	0.949	0.876
		(0.356)	(-)	(0.073)	(0.173)	(0.000)	(-)	(0.073)	(0.176)	(0.000)	(-)	(0.073)	(0.177)
	SK(har)	4.150	0.860	0.977	0.935	**4.000**	**1.000**	**1.000**	**1.000**	**4.000**	**1.000**	**1.000**	**1.000**
		(0.386)	(-)	(0.021)	(0.058)	(0.000)	(-)	(0.000)	(0.000)	(0.000)	(-)	(0.000)	(0.000)
	SK(gap)	2.000	0.000	0.745	0.485	2.000	0.000	0.746	0.488	2.340	0.170	0.789	0.575
		(0.000)	(-)	(0.005)	(0.010)	(0.000)	(-)	(0.000)	(0.000)	(0.755)	(-)	(0.096)	(0.193)

Table 10: The variable selection accuracy: mean (sd) under 100 replicates in Example 5.

		$p = 200$			$p = 500$			$p = 800$		
	Method	TPR	FPR	MS	TPR	FPR	MS	TPR	FPR	MS
$U = 5$	ICR	**1.000**	**0.000**	32.560	**1.000**	**0.000**	32.420	**1.000**	**0.000**	32.160
		(0.000)	(0.000)	(2.051)	(0.000)	(0.000)	(2.128)	(0.000)	(0.000)	(1.126)
	IP	**1.000**	**0.000**	**32.400**	**1.000**	**0.000**	**32.240**	**1.000**	**0.000**	**32.000**
		(0.000)	(0.000)	(1.752)	(0.000)	(0.000)	(1.372)	(0.000)	(0.000)	(0.000)
	ICFL	0.934	0.184	170.840	0.964	0.249	520.680	0.934	0.250	822.080
		(0.161)	(0.098)	(79.131)	(0.126)	(0.089)	(177.677)	(0.170)	(0.086)	(275.921)
	OCFL	0.934	0.103	109.160	0.963	0.016	61.840	0.926	0.002	37.040
		(0.162)	(0.045)	(37.966)	(0.121)	(0.024)	(47.272)	(0.175)	(0.004)	(14.345)
	SHIR	**1.000**	0.065	332.450	**1.000**	0.039	339.050	**1.000**	0.026	340.210
		(0.000)	(0.025)	(4.719)	(0.000)	(0.013)	(6.475)	(0.000)	(0.008)	(6.217)
	SMA	**1.000**	0.064	332.210	**1.000**	0.032	335.700	**1.000**	0.020	335.470
		(0.000)	(0.027)	(5.149)	(0.000)	(0.010)	(5.098)	(0.000)	(0.005)	(4.270)
	Local	0.838	0.100	1038.480	0.769	0.048	1199.800	0.733	0.032	1263.060
		(0.035)	(0.010)	(88.302)	(0.042)	(0.006)	(132.189)	(0.036)	(0.004)	(119.201)
	SK(har)	0.969	0.432	373.770	0.944	0.193	429.890	0.948	0.127	468.470
		(0.082)	(0.305)	(242.129)	(0.104)	(0.194)	(397.009)	(0.096)	(0.145)	(503.149)
	SK(gap)	**1.000**	0.881	355.890	**1.000**	0.624	632.980	**1.000**	0.486	786.440
		(0.000)	(0.046)	(24.108)	(0.000)	(0.094)	(95.989)	(0.000)	(0.061)	(96.642)
	DLSA	0.128	0.008	2.510	0.355	0.373	186.350	0.591	0.589	471.300
		(0.018)	(0.017)	(3.274)	(0.141)	(0.067)	(33.167)	(0.189)	(0.060)	(48.197)
$U = 10$	ICR	**1.000**	**0.000**	32.560	**1.000**	**0.000**	**32.080**	**1.000**	**0.000**	**32.000**
		(0.000)	(0.000)	(2.346)	(0.000)	(0.000)	(0.800)	(0.000)	(0.000)	(0.000)
	IP	**1.000**	**0.000**	**32.320**	**1.000**	**0.000**	**32.080**	**1.000**	**0.000**	**32.000**
		(0.000)	(0.000)	(1.576)	(0.000)	(0.000)	(0.800)	(0.000)	(0.000)	(0.000)
	ICFL	0.959	0.098	106.240	0.946	0.117	260.600	0.964	0.131	446.760
		(0.133)	(0.101)	(79.170)	(0.154)	(0.072)	(145.067)	(0.144)	(0.070)	(223.048)
	OCFL	0.959	0.066	81.240	0.958	0.020	69.400	0.966	0.008	55.360
		(0.134)	(0.076)	(59.372)	(0.128)	(0.019)	(39.399)	(0.137)	(0.036)	(114.654)
	SHIR	**1.000**	0.074	335.320	**1.000**	0.046	342.410	**1.000**	0.022	337.390
		(0.000)	(0.024)	(9.045)	(0.000)	(0.015)	(7.429)	(0.000)	(0.008)	(6.334)
	SMA	**1.000**	0.070	334.280	**1.000**	0.045	342.740	**1.000**	0.020	335.550
		(0.000)	(0.022)	(7.721)	(0.000)	(0.015)	(8.864)	(0.000)	(0.006)	(4.659)
	Local	0.919	0.121	1223.780	0.883	0.061	1488.800	0.865	0.043	1639.920
		(0.028)	(0.009)	(75.231)	(0.031)	(0.005)	(97.978)	(0.037)	(0.005)	(160.497)
	SK(har)	0.983	0.480	402.570	0.973	0.331	702.860	0.976	0.225	765.640
		(0.053)	(0.313)	(242.951)	(0.072)	(0.203)	(413.880)	(0.068)	(0.163)	(531.936)
	SK(gap)	**1.000**	0.913	366.800	**1.000**	0.719	723.220	**1.000**	0.587	945.320
		(0.000)	(0.096)	(36.036)	(0.000)	(0.043)	(42.432)	(0.000)	(0.081)	(128.451)
	DLSA	0.130	**0.000**	1.090	0.133	0.033	17.240	0.300	0.293	234.810
		(0.030)	(0.002)	(0.379)	(0.035)	(0.039)	(19.086)	(0.153)	(0.075)	(59.824)

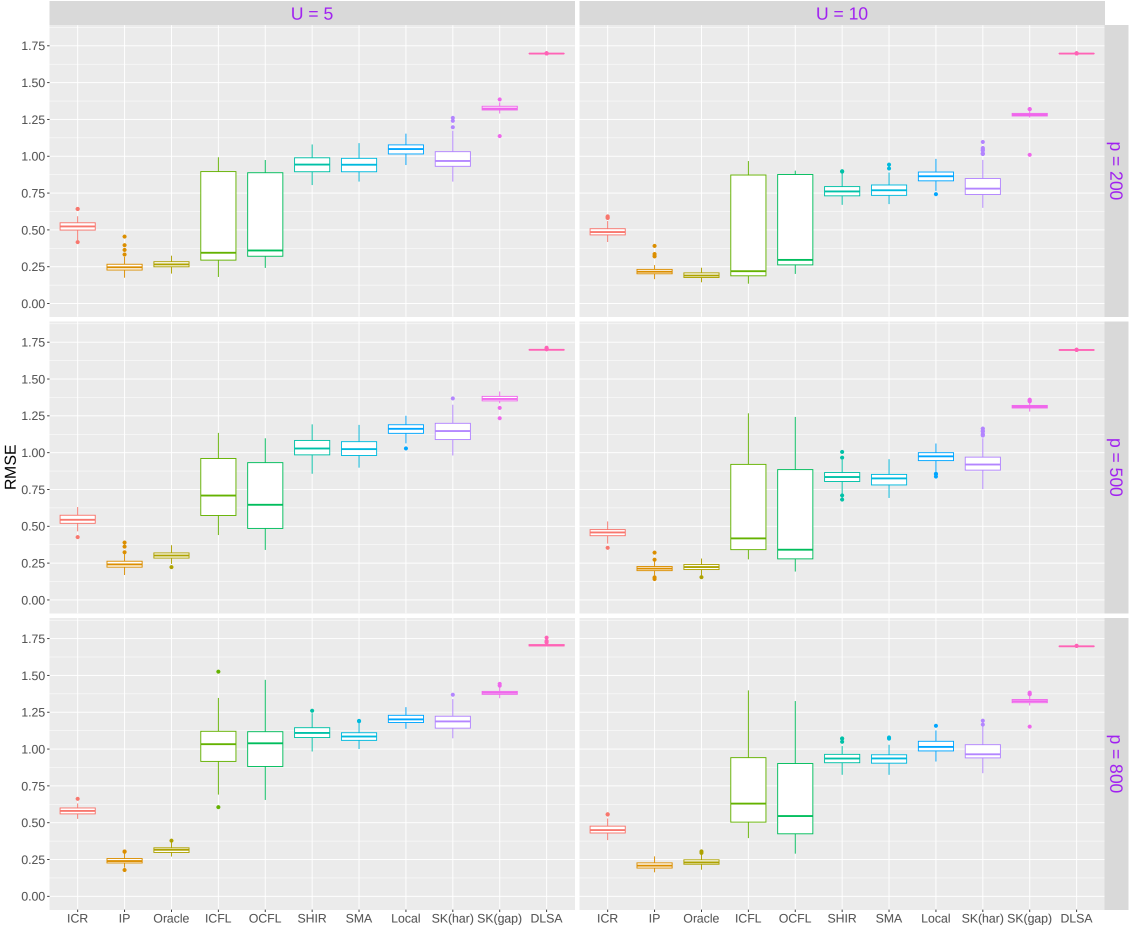


Figure 8: Boxplots of RMSE in Example 5.

Table 11: The clustering accuracy: mean (sd) based on 100 replicates in Example 5.

		$p = 200$				$p = 500$				$p = 800$			
	Method	$\widehat{M}$	Per	RI	ARI	$\widehat{M}$	Per	RI	ARI	$\widehat{M}$	Per	RI	ARI
$U = 5$	ICR	4.070	0.930	**0.999**	**0.998**	4.040	0.960	**1.000**	**0.999**	4.020	0.980	**1.000**	0.999
		(0.256)	(-)	(0.003)	(0.008)	(0.197)	(-)	(0.002)	(0.007)	(0.141)	(-)	(0.002)	(0.005)
	IP	4.050	0.950	**0.999**	**0.998**	4.030	0.970	**1.000**	**0.999**	**4.000**	**1.000**	**1.000**	**1.000**
		(0.219)	(-)	(0.003)	(0.007)	(0.171)	(-)	(0.002)	(0.006)	(0.000)	(-)	(0.000)	(0.000)
	ICFL	**4.000**	**1.000**	0.955	0.890	**4.000**	**1.000**	0.936	0.839	**4.000**	**1.000**	0.876	0.696
		(0.000)	(-)	(0.070)	(0.170)	(0.000)	(-)	(0.070)	(0.169)	(0.000)	(-)	(0.081)	(0.184)
	OCFL	**4.000**	**1.000**	0.955	0.890	**4.000**	**1.000**	0.935	0.836	**4.000**	**1.000**	0.876	0.695
		(0.000)	(-)	(0.070)	(0.170)	(0.000)	(-)	(0.070)	(0.169)	(0.000)	(-)	(0.081)	(0.185)
	SK(har)	4.180	0.830	0.964	0.901	4.250	0.780	0.939	0.833	4.340	0.750	0.936	0.820
		(0.796)	(-)	(0.038)	(0.103)	(1.029)	(-)	(0.048)	(0.127)	(1.165)	(-)	(0.042)	(0.116)
	SK(gap)	2.010	0.000	0.744	0.481	2.020	0.000	0.743	0.477	2.000	0.000	0.742	0.476
		(0.100)	(-)	(0.014)	(0.023)	(0.141)	(-)	(0.016)	(0.027)	(0.000)	(-)	(0.007)	(0.013)
$U = 10$	ICR	4.070	0.940	0.999	0.998	4.010	0.990	**1.000**	**1.000**	**4.000**	**1.000**	**1.000**	**1.000**
		(0.293)	(-)	(0.003)	(0.010)	(0.100)	(-)	(0.001)	(0.003)	(0.000)	(-)	(0.000)	(0.000)
	IP	4.040	0.960	**1.000**	**0.999**	4.010	0.990	**1.000**	**1.000**	**4.000**	**1.000**	**1.000**	**1.000**
		(0.197)	(-)	(0.002)	(0.007)	(0.100)	(-)	(0.001)	(0.003)	(0.000)	(-)	(0.000)	(0.000)
	ICFL	**4.000**	**1.000**	0.958	0.898	**4.000**	**1.000**	0.943	0.860	**4.000**	**1.000**	0.942	0.855
		(0.000)	(-)	(0.065)	(0.157)	(0.000)	(-)	(0.075)	(0.185)	(0.000)	(-)	(0.073)	(0.178)
	OCFL	**4.000**	**1.000**	0.958	0.898	**4.000**	**1.000**	0.943	0.860	**4.000**	**1.000**	0.942	0.854
		(0.000)	(-)	(0.065)	(0.157)	(0.000)	(-)	(0.075)	(0.185)	(0.000)	(-)	(0.073)	(0.178)
	SK(har)	4.050	0.900	0.981	0.950	4.130	0.890	0.974	0.926	4.100	0.910	0.967	0.907
		(0.575)	(-)	(0.033)	(0.081)	(0.800)	(-)	(0.032)	(0.090)	(0.541)	(-)	(0.033)	(0.090)
	SK(gap)	2.010	0.000	0.744	0.481	2.000	0.000	0.743	0.479	2.010	0.000	0.744	0.480
		(0.100)	(-)	(0.013)	(0.023)	(0.000)	(-)	(0.003)	(0.007)	(0.100)	(-)	(0.012)	(0.019)

Table 12: The Computational time: mean (sd) based on 100 replicates in Example 5. For methods (ICR, IP, SHIR and SMA), computation time refers to the average computation time for each tuning parameter based on a set of tuning parameters.

		Method	ICR	IP	ICFL	OCFL	SHIR	SMA
$U = 5$	$p = 200$	Time (seconds)	20.18	288.51	2.15	0.03	112.05	120.77
			(4.62)	(48.07)	(0.29)	(0.01)	(20.15)	(21.23)
	$p = 500$	Time (seconds)	77.72	1248.41	8.34	0.12	425.79	307.55
			(13.77)	(136.72)	(1.40)	(0.02)	(64.02)	(68.99)
	$p = 800$	Time (seconds)	151.32	2199.70	13.55	0.19	222.01	111.53
			(18.14)	(331.44)	(1.93)	(0.02)	(41.48)	(25.39)
$U = 10$	$p = 200$	Time (seconds)	13.80	293.61	6.31	0.08	130.55	180.83
			(3.80)	(36.41)	(1.24)	(0.01)	(13.36)	(23.52)
	$p = 500$	Time (seconds)	71.21	954.14	23.44	0.26	599.04	497.45
			(32.37)	(79.07)	(6.22)	(0.04)	(91.95)	(76.09)
	$p = 800$	Time (seconds)	91.09	3284.76	42.68	0.46	390.31	206.13
			(14.66)	(272.30)	(12.91)	(0.11)	(66.77)	(36.91)

Table 13: The variable selection accuracy: mean (sd) under 100 replicates in Example 6.

	$n_0 = 200$			$n_0 = 400$		
Method	TPR	FPR	MS	TPR	FPR	MS
ICR	**1.000**	**0.000**	**70.400**	**1.000**	**0.000**	**66.220**
	(0.000)	(0.000)	(11.274)	(0.000)	(0.000)	(3.823)
IP	0.999	**0.000**	70.560	**1.000**	**0.000**	**66.220**
	(0.009)	(0.000)	(11.654)	(0.000)	(0.000)	(3.823)
ICFL	0.891	0.054	87.900	0.867	0.013	64.380
	(0.164)	(0.086)	(50.671)	(0.175)	(0.061)	(35.674)
OCFL	0.895	0.046	83.580	0.875	0.009	62.520
	(0.149)	(0.079)	(46.077)	(0.171)	(0.060)	(34.506)
SHIR	0.963	**0.000**	826.750	0.990	0.002	942.920
	(0.058)	(0.002)	(114.240)	(0.039)	(0.005)	(86.582)
SMA	0.805	**0.000**	480.570	0.987	0.002	936.620
	(0.121)	(0.001)	(83.700)	(0.047)	(0.004)	(97.650)
Local	0.889	0.239	2794.920	0.961	0.285	3236.110
	(0.026)	(0.015)	(135.467)	(0.014)	(0.014)	(119.034)
SK(har)	0.987	0.870	470.700	**1.000**	0.934	561.980
	(0.070)	(0.262)	(171.349)	(0.000)	(0.155)	(120.517)
SK(gap)	0.975	0.960	193.320	0.936	0.870	177.460
	(0.129)	(0.197)	(39.317)	(0.177)	(0.338)	(66.105)
DLSA	0.704	**0.000**	7.740	0.777	**0.000**	8.550
	(0.159)	(0.000)	(1.750)	(0.149)	(0.000)	(1.641)

Table 14: The clustering accuracy: mean (sd) based on 100 replicates in Example 6.

	$n_0 = 200$				$n_0 = 400$			
Method	$\widehat{M}$	Per	RI	ARI	$\widehat{M}$	Per	RI	ARI
ICR	6.400	0.680	**0.993**	0.976	6.020	0.910	**0.977**	**0.992**
	(1.025)	(-)	(0.019)	(0.058)	(0.348)	(-)	(0.013)	(0.040)
IP	6.420	0.680	**0.993**	**0.977**	6.020	0.910	**0.977**	**0.992**
	(1.056)	(-)	(0.018)	(0.056)	(0.348)	(-)	(0.013)	(0.040)
ICFL	**6.000**	**1.000**	0.941	0.828	**6.000**	**1.000**	0.947	0.839
	(0.000)	(-)	(0.060)	(0.152)	(0.000)	(-)	(0.046)	(0.123)
OCFL	**6.000**	**1.000**	0.940	0.824	**6.000**	**1.000**	0.947	0.838
	(0.000)	(-)	(0.060)	(0.152)	(0.000)	(-)	(0.046)	(0.123)
SK(har)	5.240	0.420	0.940	0.837	5.970	0.510	0.977	0.932
	(1.334)	(-)	(0.073)	(0.178)	(0.989)	(-)	(0.043)	(0.115)
SK(gap)	2.010	0.000	0.652	0.320	2.020	0.000	0.651	0.321
	(0.100)	(-)	(0.021)	(0.024)	(0.141)	(-)	(0.035)	(0.038)

Table 15: The computational time: mean (sd) based on 100 replicates in Example 6. For methods (ICR, IP, SHIR and SMA), computation time refers to the average computation time for each tuning parameter based on a set of tuning parameters.

	Method	ICR	IP	ICFL	OCFL	SHIR	SMA
$n_0 = 200$	Time (seconds)	22.63	347.57	5.83	0.09	125.66	99.63
		(4.25)	(43.96)	(0.45)	(0.01)	(54.08)	(33.50)
$n_0 = 400$	Time (seconds)	19.65	657.10	20.88	0.26	84.15	59.21
		(3.44)	(84.94)	(3.11)	(0.03)	(23.17)	(11.94)

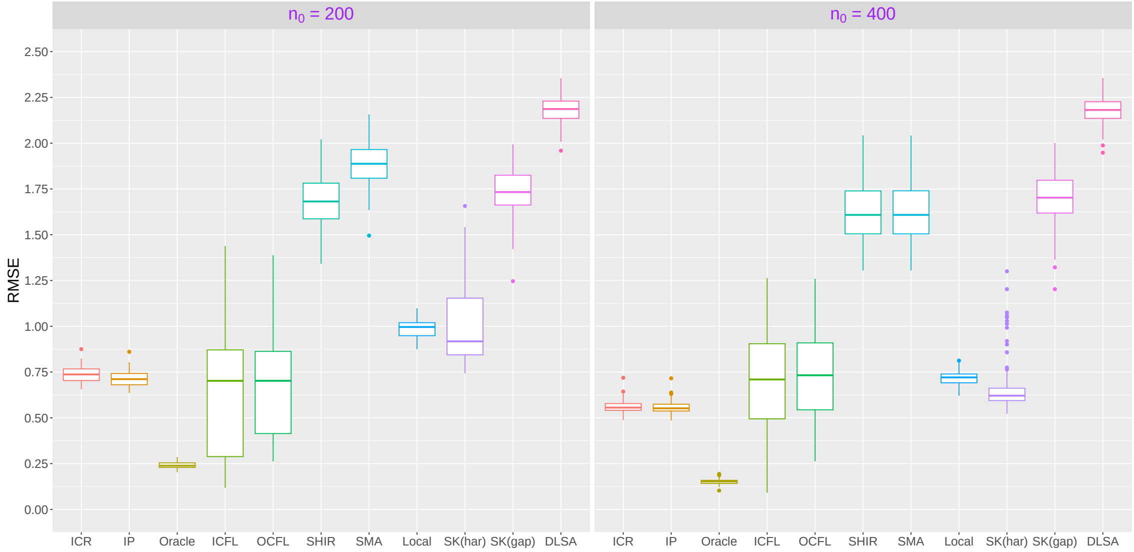


Figure 9: Boxplots of RMSE in Example 6.

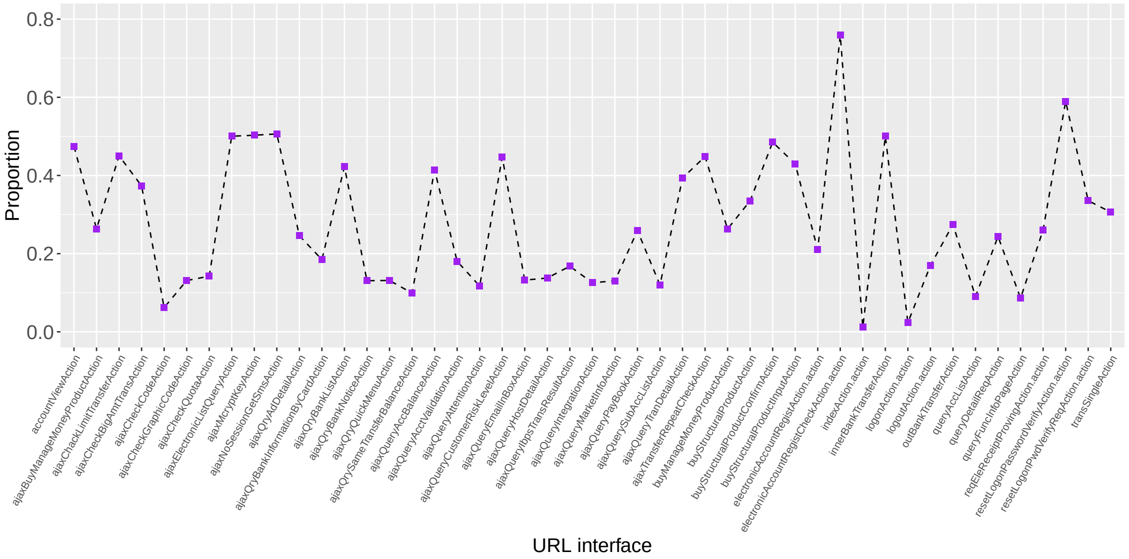


Figure 10: The abnormal proportions among the 47 URL interfaces in data analysis.

Table 16: The identified important variables and their estimates using the five integrative analysis methods in data analysis. For the proposed method, only estimates for the nontrivial clusters are shown.

Variable	$\text{ICFL}_{10}^{(1)}$	$\text{ICFL}_{10}^{(2)}$	$\text{ICFL}_{10}^{(3)}$	$\text{ICFL}_{10}^{(4)}$	$\text{ICFL}_{10}^{(5)}$	$\text{ICFL}_{10}^{(6)}$	$\text{ICFL}_{10}^{(7)}$	$\text{ICFL}_{10}^{(8)}$	$\text{ICFL}_{10}^{(9)}$
Intercept	1.401	-2.545	1.199	1.437	1.149	-0.970	0.926	0.868	1.406
Gnum	-0.181	0.142	-0.205	-0.239	–	0.288	-0.284	-0.139	0.606
Glen	-0.191	0.117	-0.231	-0.210	–	–	-0.316	–	-0.523
Pnum	–	-0.120	0.159	–	-0.207	0.903	0.323	–	-0.714
Plen	-0.380	-0.338	-0.334	–	–	0.417	–	2.115	0.149
$Gl0$	1.534	0.594	1.032	1.895	–	-0.272	1.372	2.800	-0.867
$Pl0$	1.031	1.026	1.730	–	0.394	0.641	2.282	0.431	–
$Pl1$	2.494	1.778	2.359	2.118	2.770	1.831	2.237	0.202	0.533
$Pl2$	2.369	1.712	2.103	2.730	2.840	0.797	0.376	–	0.363
$Pl3$	–	–	0.196	–	–	0.390	–	–	0.110
$Pl4$	0.106	0.101	0.101	–	0.228	0.276	–	–	2.439
$Pl5$	–	–	–	–	0.113	–	–	–	1.205
$Pl6$	–	–	–	–	–	–	–	–	0.622
$Pl7$	–	–	–	–	–	0.151	–	–	0.579
$Pl8$	–	–	–	–	–	–	–	–	-1.714
$Pl9$	–	–	–	–	0.118	–	–	–	0.288
$Pl10$	–	–	–	–	0.137	–	–	–	0.219
$Pl11$	–	–	–	–	–	–	–	–	0.168
$Pl12$	–	–	–	–	–	–	–	–	0.169
$Pl13$	–	–	–	–	–	–	–	–	0.139
$Pl14$	–	–	–	–	–	0.107	–	–	0.157
$Pl15$	–	–	–	–	–	0.110	–	–	0.175
$Pl17$	–	–	–	-0.131	–	–	–	–	–
$Pl18$	–	–	–	–	–	–	-0.104	–	–
$Pl19$	–	–	–	0.129	–	–	–	–	–
$GPw6$	–	–	–	–	-0.103	–	–	–	–
$GPw39$	–	0.118	–	–	–	–	–	–	–
$GPw68$	–	0.111	–	–	–	–	–	–	–
$GPw74$	–	-0.112	–	–	–	–	–	–	–

Table 17: One record of the initial request logs in data analysis.

URL interface	GET Parameter	POST Parameter

ajaxNoSessionGetSmsAction	s=captcha	_method=__construct&filter[]=phpinfo&
		method=get&server[REQUEST_METHOD]=1

References

Naman Agarwal, Ananda Theertha Suresh, Felix Xinnan X Yu, Sanjiv Kumar, and Brendan McMahan. cpsgd: Communication-efficient and differentially-private distributed sgd. *Advances in Neural Information Processing Systems*, 31, 2018.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3): 1352–1382, 2018.

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

Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

Raphael Bost, Raluca Ada Popa, Stephen Tu, and Shafi Goldwasser. Machine learning classification over encrypted data. In *Proceedings 2015 Network and Distributed System Security Symposium*. Internet Society, 2015.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

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

Tianxi Cai, Molei Liu, and Yin Xia. Individual data protected integrative regression analysis of high-dimensional heterogeneous data. *Journal of the American Statistical Association*, 117(540):2105–2119, 2022.

Jingxiang Chen, Quoc Tran-Dinh, Michael R. Kosorok, and Yufeng Liu. Identifying heterogeneous effect using latent supervised clustering with adaptive fusion. *Journal of Computational and Graphical Statistics*, 30(1):43–54, 2021.

Yi-Ruei Chen, Amir Rezapour, and Wen-Guey Tzeng. Privacy-preserving ridge regression on distributed data. *Information Sciences*, 451:34–49, 2018.

Eric C Chi and Kenneth Lange. Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013, 2015.

Edgar Dobriban and Yue Sheng. Wonder: Weighted one-shot distributed ridge regression in high dimensions. *Journal of Machine Learning Research*, 21(66):1–52, 2020.

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

Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

Maryem Ait El Hadj, Mohammed Erradi, Ahmed Khoumsi, and Yahya Benkaouz. Validation and correction of large security policies: a clustering and access log based approach. *2018 IEEE international conference on big Data (Big Data)*, pages 5330–5332, 2018.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(5): 849–911, 2008.

Jianqing Fan and Jinchi Lv. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484, 2011.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.

Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *International conference on machine learning*, pages 201–210. PMLR, 2016.

Yalan Guo, Yulei Wu, Yanchao Zhu, Bingqiang Yang, and Chunjing Han. Anomaly detection using distributed log data: A lightweight federated learning approach. *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.

John A Hartigan. *Clustering algorithms.* Wiley, New York, 1975.

Qianchuan He, Hao Helen Zhang, Christy L. Avery, and D. Y. Lin. Sparse meta-analysis with high-dimensional data. *Biostatistics*, 17(2):205–220, 2016.

Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *The Journal of Machine Learning Research*, 22(1):10882–11005, 2021.

Qiaona Hu, Baoming Tang, and Derek Lin. Anomalous user activity detection in enterprise multi-source logs. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 797–803, 2017.

Jian Huang, Joel L Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *The Annals of Statistics*, 38(4):2282–2313, 2010.

Yuan Huang, Qingzhao Zhang, Sanguo Zhang, Jian Huang, and Shuangge Ma. Promoting similarity of sparsity structures in integrative analysis with penalization. *Journal of the American Statistical Association*, 112(517):342–350, 2017.

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

Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.

K Kelleher. Facebook loses around $13 billion in value after data breach affects 50 million of its users. *Retrieved*, 12:2019, 2018.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *The Journal of Machine Learning Research*, 18(1):115–144, 2017.

Dongdong Li, Wenbin Lu, Di Shu, Sengwee Toh, and Rui Wang. Distributed cox proportional hazards regression using summary-level information. *Biostatistics*, 24(3):776–794, 2023.

Furong Li and Huiyan Sang. Spatial homogeneity pursuit of regression coefficients for large datasets. *Journal of the American Statistical Association*, 114(527):1050–1062, 2019.

Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.

Jin Liu, Jian Huang, Yawei Zhang, Qing Lan, Nathaniel Rothman, Tongzhang Zheng, and Shuangge Ma. Integrative analysis of prognosis data on multiple cancer subtypes. *Biometrics*, 70(3):480–488, 2014.

Xiaokang Liu, Rui Duan, Chongliang Luo, Alexis Ogdie, Jason H Moore, Henry R Kranzler, Jiang Bian, and Yong Chen. Multisite learning of high-dimensional heterogeneous data with applications to opioid use disorder study of 15,000 patients across 5 clinical sites. *Scientific reports*, 12(1):11073, 2022.

Shujie Ma and Jian Huang. A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, 112(517):410–423, 2017.

Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni, and Richard Vidal. Federated multi-task learning under a mixture of distributions. *Advances in Neural Information Processing Systems*, 34:15434–15447, 2021.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of -estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.

Seyoung Park, Eun Ryung Lee, and Hyokyoung G Hong. Varying-coefficients for regional quantile via knn-based lasso with applications to health outcome study. *Statistics in Medicine*, 42(22):3903–3918, 2023.

Mingyang Ren, Sanguo Zhang, and Junhui Wang. Consistent estimation of the number of communities via regularized network embedding. *Biometrics*, 79(3):2404–2416, 2023.

R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pages 1000–1008. PMLR, 2014.

Ryosuke Shimmura and Joe Suzuki. Converting admm to a proximal gradient for efficient sparse estimation. *Japanese Journal of Statistics and Data Science*, pages 1–21, 2022.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Lu Tang and Peter XK Song. Fused lasso approach in regression coefficients clustering: learning parameter heterogeneity in data integration. *The Journal of Machine Learning Research*, 17(1):3915–3937, 2016.

Xiwei Tang, Fei Xue, and Annie Qu. Individualized multidirectional variable selection. *Journal of the American Statistical Association*, 116(535):1280–1296, 2021.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

Uğur Ünal and Hasan Dağ. Anomalyadapters: Parameter-efficient multi-anomaly task detection. *IEEE Access*, 10:5635–5646, 2022.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

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

Li Wang, Xiang Liu, Hua Liang, and Raymond J Carroll. Estimation and variable selection for generalized additive partial linear models. *The Annals of Statistics*, 39(4):1827–1851, 2011.

Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

Simon N Wood. *Generalized additive models: an introduction with R.* CRC press, 2017.

Lingzhou Xue, Hui Zou, and Tianxi Cai. Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *The Annals of Statistics*, 40(3):1403–1429, 2012.

Xinfeng Yang, Xiaodong Yan, and Jian Huang. High-dimensional integrative analysis with homogeneity and sparsity recovery. *Journal of Multivariate Analysis*, 174:104529, 2019.

Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*, 2020.

Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

Tianqi Zhao, Guang Cheng, and Han Liu. A partially linear framework for massive heterogeneous data. *The Annals of Statistics*, 44(4):1400–1437, 2016.

Ling Zhou, Ziyang Gong, and Pengcheng Xiang. Distributed computing and inference for big data. *Annual Review of Statistics and Its Application*, 11, 2024.

Xiaolu Zhu and Annie Qu. Cluster analysis of longitudinal profiles with subgroups. *Electronic Journal of Statistics*, 12(1):171–193, 2018.

Xuening Zhu, Feng Li, and Hansheng Wang. Least-square approximation for a distributed system. *Journal of Computational and Graphical Statistics*, 30(4):1004–1018, 2021.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533, 2008.