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

Yuanxing Chen  
Department of Statistics, School of Economics, Xiamen University  
Qingzhao Zhang  
MOE Key Laboratory of Econometrics, Department of Statistics, School of Economics, Wang Yanan Institute for Studies in Economics, and Fujian Key Lab of Statistics, Xiamen University  
Shuangge Ma  
Department of Biostatistics, Yale University  
Kuangnan Fang  
Department of Statistics, School of Economics, Xiamen University

November 8, 2022

Abstract

In diverse fields ranging from finance to omics, it is increasingly common that data is distributed and 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 of the 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.

Keywords: High dimensionality; Data heterogeneity; Clustering structure; Sparsity; Penalization.
1 Introduction

In diverse fields, it is increasingly common that data is distributed and 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. For example, with financial and omics data, there are often strong considerations on data privacy protection (Abbe et al., 2012; Cai et al., 2021). 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).

To avoid sharing raw data, distributed learning techniques have been developed, with which all clients share summary statistics (which have no privacy concerns and much smaller sizes) as opposed to raw data. It is noted that this is also closely related to the divide-and-conquer (DC) strategy, which has been designed to reduce cost and improve feasibility and performance in the analysis of big data (Lee et al., 2017; Battey et al., 2018). Some distributed learning methods demand multiple communications between the local clients and central server (Jordan et al., 2019). There are also one-shot methods (Dobriban and Sheng, 2020), with which the local clients only need to communicate once with the central server. Such methods may enjoy lower cost and higher convenience. In most of the existing studies, it has been assumed that all the clients have 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. Here, heterogeneity can be caused by differences in sample characteristics, data collection techniques, and multiple other factors (Ghosh et al. 2020).

To accommodate data heterogeneity, some federated learning methods allow client-specific models (Smith et al. 2017). Of the most relevance to this study is clustered federated learning, under which clients form clusters, those in the same cluster share the same model, and different clusters have different model (Ghosh et al. 2020; Marfoq et al. 2021). This type of analysis may have been more popular in computer science than statistics. Intuitively, assuming and identifying a clustering structure can lead to a better understanding of the “interconnections” among clients (for example, those in the same cluster are more alike and can be more closely related to each other) and a smaller number of model parameters. A common limitation shared by the existing clustered federated learning 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.

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 like-
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. (2021) 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, where only summary statistics can be shared. To sufficiently accommodate data heterogeneity, clients are allowed to have different models. 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. Further, motivated by the success of clustered federated learning, we consider the scenario where clients form clusters, and the models are cluster specific. For 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 datasets and a smaller number of parameters (and hence improved estimation). Third, compared to the existing clustered federated learning methods (Ghosh et al., 2020; Marfoq et al., 2021), it can data-dependently and conveniently determine the number of clusters, with the assistance of penalized fusion. 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. Proofs of the theoretical results and additional numerical results are provided in the Appendix.

2 Methods

2.1 Integrative Analysis under Privacy Constraints

Suppose that there are $K$ independent clients, and for the $k$th client, there are $n_k$ observations. The total sample size is $N = \sum_{k=1}^{K} n_k$. For the $k$th client, let $y_{i}^{(k)}$ and $\mathbf{x}_{i}^{(k)} = (x_{i1}^{(k)}, \ldots, x_{ip}^{(k)})^\top \in \mathbb{R}^p$ be the response and covariate vector of the $i$th observation, respectively, where the first element of $\mathbf{x}_{i}^{(k)}$ is fixed as $x_{i1}^{(k)} \equiv 1$ to accommodate intercept. Accordingly, let $\mathbf{X}^{(k)} = (\mathbf{x}_{1}^{(k)}, \ldots, \mathbf{x}_{n_k}^{(k)})^\top$ and $\mathbf{Y}^{(k)} = (y_{1}^{(k)}, \ldots, y_{n_k}^{(k)})^\top$ denote the design matrix and response vector of the $k$th client, respectively. Let $f(\cdot)$ be the pre-specified twice-differentiable loss function and $\mathbf{\theta}^{(k)} = (\theta_{1}^{(k)}, \ldots, \theta_{p}^{(k)})^\top$ be the $p$-dimensional coefficient vector. The empirical local and global loss functions are defined as

$$
\hat{L}_k(\mathbf{\theta}^{(k)}) = \frac{1}{n_k} \sum_{i=1}^{n_k} f(\mathbf{x}_{i}^{(k)}\mathbf{\theta}^{(k)}, y_{i}^{(k)}), \quad k \in [K] \quad \text{and} \quad \hat{L}(\mathbf{\theta}) = \frac{1}{N} \sum_{k=1}^{K} n_k \hat{L}_k(\mathbf{\theta}^{(k)}),
$$

respectively, where $\mathbf{\theta} = (\mathbf{\theta}^{(1)}, \ldots, \mathbf{\theta}^{(K)})$ is a $p \times K$ coefficient matrix with the $j$th row $\mathbf{\theta}_j = (\theta_{j1}, \ldots, \theta_{jk})^\top$, and $[d]$ denotes the index set $\{1, \ldots, d\}$ for an integer $d$. Assume
that the indexes of $\theta$ can be classified into $M$ non-overlapping subsets $\{G^{(1)}, \ldots, G^{(M)}\}$.

Additionally, for each covariate, its coefficients across the $K$ datasets can be viewed as a group (Cai et al., 2021), leading to $p$ groups corresponding to the covariates.

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

$$
\hat{Q}_{IP}(\theta) = \hat{L}(\theta) + \mathcal{P}_{\lambda_1}(\theta) + \mathcal{P}_{\lambda_2}(\theta)
$$

$$
= \frac{1}{N} \sum_{k=1}^{K} n_k \hat{L}_k(\theta^{(k)}) + \sum_{j=2}^{p} p_{r}(||\theta_j||_2, \lambda_1) + \sum_{k<k'} p_{r}(||\theta^{(k)} - \theta^{(k')}||_2, \lambda_2),
$$

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

With the privacy-preservation constraints, raw data of the individual datasets is not available, and hence objective function $\hat{Q}_{IP}(\theta)$ 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

$$
\hat{Q}_1(\theta) = \frac{1}{N} \sum_{k=1}^{K} n_k (\theta^{(k)} - \bar{\theta}^{(k)})^\top \hat{V}^{(k)} (\theta^{(k)} - \bar{\theta}^{(k)}) + \mathcal{P}_{\lambda_1}(\theta) + \mathcal{P}_{\lambda_2}(\theta),
$$

where $\bar{\theta}^{(k)}$ is the local estimator of the $k$th client, and $\hat{V}^{(k)} = \partial^2 \hat{L}_k(\hat{\theta}^{(k)})/\partial \theta^{(k)} \partial \theta^{(k)}$ is the Hessian matrix of $\hat{L}_k(\theta^{(k)})$ with respect to $\theta^{(k)}$. 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, in general,
the Lasso estimates are biased. Motivated by Cai et al. (2021), we adopt the debiased Lasso estimators (van de Geer et al., 2014) and propose the ICR estimator by minimizing the following objective function

$$\hat{Q}_{ICR}(\theta) = \frac{1}{N} \sum_{k=1}^{K} n_k \left( \theta^{(k)} \top \tilde{V}^{(k)} \theta^{(k)} - 2 \theta^{(k)} \top \tilde{\zeta}^{(k)} \right) + P_{\lambda_1}(\theta) + P_{\lambda_2}(\theta),$$

(3)

where $\tilde{\zeta}^{(k)} = \tilde{V}^{(k)} \tilde{\theta}^{(k)} - \tilde{g}^{(k)}$ and $\tilde{g}^{(k)} = \partial \hat{L}_k(\tilde{\theta}^{(k)})/\partial \theta^{(k)}$ is the gradient of $\hat{L}_k(\theta^{(k)})$ with respect to $\theta^{(k)}$ at $\tilde{\theta}^{(k)}$. For the penalty function, viable choices include SCAD (Fan and Li, 2001), MCP (Zhang, 2010), and others. We adopt MCP in our numerical studies. 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.

**Figure 1:** Scheme of the proposed analysis.

This approach has been motivated by the following considerations. In $\hat{Q}_{ICR}(\theta)$, we only
make use of four summary statistics, namely the initial local estimators \( \{ \tilde{\theta}^{(k)} \}_{k=1}^K \), corresponding gradient vectors \( \{ \tilde{g}^{(k)} \}_{k=1}^K \), Hessian matrices \( \{ \tilde{V}^{(k)} \}_{k=1}^K \), 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., 2021). 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 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 been popular in the recent literature (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).

### 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 \( t \)th iteration, we update
the coefficients by solving

\[
\begin{align*}
\arg\min_{\theta} \ & \frac{1}{N} \sum_{k=1}^{K} n_k \left( \theta^{(k)\top} \tilde{V}^{(k)} \theta^{(k)} - 2 \theta^{(k)\top} \tilde{\zeta}^{(k)} \right) + \sum_{k<k'} \omega_{kk'}^{t-1} \| \theta^{(k)} - \theta^{(k')} \|_2 + \sum_{j=2}^{p} \omega_j \| \theta_j \|_2, \\
\text{subject to} \ & \theta^{(k)} - \theta^{(k')} = \alpha_{kk'}, \ 1 \leq k < k' \leq K,
\end{align*}
\]

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

\[
\begin{align*}
\arg\min_{\theta, \alpha} \left\{ \ell(\theta) := g(\theta) + h_1(\alpha) + h_2(\theta) \right\} \\
&= \frac{1}{N} \sum_{k=1}^{K} n_k \left( \theta^{(k)\top} \tilde{V}^{(k)} \theta^{(k)} - 2 \theta^{(k)\top} \tilde{\zeta}^{(k)} \right) + \sum_{k<k'} \omega_{kk'}^{t-1} \| \alpha_{kk'} \|_2 + \sum_{j=2}^{p} \omega_j \| \theta_j \|_2, \\
\text{subject to} \ & \theta^{(k)} - \theta^{(k')} = \alpha_{kk'}, \ 1 \leq k < k' \leq K,
\end{align*}
\]

where \( \alpha = (\alpha_{12}, \ldots, \alpha_{(K-1)K}) \) 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

\[
\ell_{\nu}(\theta, \alpha, \xi) = \ell(\theta) + \sum_{k<k'} \xi_{kk'}^{(k)\top} (\theta^{(k)} - \theta^{(k')} - \alpha_{kk'}) + \frac{\nu}{2} \sum_{k<k'} \| \theta^{(k)} - \theta^{(k')} - \alpha_{kk'} \|_2^2, \quad (4)
\]

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

\[
(\theta^t, \alpha^t) = \arg\min_{\theta, \alpha} \ell_{\nu}(\theta, \alpha, \xi^{t-1}),
\]

\[
\xi_{kk'}^t = \xi_{kk'}^{t-1} + \nu (\theta^{(k),t} - \theta^{(k'),t} - \alpha_{kk'}^t), \ 1 \leq k < k' \leq K.
\]
To update \((\theta, \alpha)\) via (4), we minimize \(\eta(\theta)\) defined by

\[
\eta(\theta) := \arg \min_{\alpha} \ell_\nu(\theta, \alpha, \xi^{t-1}) = \arg \min_{\alpha} \left\{ \sum_{k<k'} [\omega_{kk'}^{t-1} \|\alpha_{kk'}\|_2 + \xi_{kk'}^{t-1 \top} (\theta^{(k)} - \theta^{(k')}) - \alpha_{kk'}] + \frac{\nu}{2} \|\theta^{(k)} - \theta^{(k')} - \alpha_{kk'}\|_2^2 \right\}
+ g(\theta) + h_2(\theta)
\]

\[
:= \arg \min_{\alpha} \eta_1(\theta, \alpha) + g(\theta) + h_2(\theta) := \eta_2(\theta) + h_2(\theta).
\]

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

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

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

\[
\alpha(\theta) = \text{prox}_{\nu^{-1}h_1}(\theta A + \nu^{-1}\xi^{t-1}),
\]

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

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

where \(\theta_g = 2/N[n_1(\tilde{V}^{(1)}\theta^{(1)} - \tilde{\zeta}^{(1)}), \ldots, n_K(\tilde{V}^{(K)}\theta^{(K)} - \tilde{\zeta}^{(K)})]\) is a \(p \times K\) matrix and \(h_1^*\) is the conjugate function of \(h_1\). Then, we can adopt proximal gradient techniques, such
as FISTA (Parikh et al., 2014), to obtain the solution of the first minimization problem in (5). Further, the second iterative step in (5) can be reformulated as

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

The proposed proximal ADMM algorithm is summarized as follows.

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

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

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

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

   $$\omega_{kk'}^{t-1,s} \leftarrow p'_\tau(||u^{(k),t-1,s-1} - u^{(k'),t-1,s-1}||_2, \lambda_2), \quad 1 \leq k < k' \leq K,$$
   $$\theta^{t-1,s} \leftarrow \text{prox}_{\chi_2} \left[ u^{t-1,s-1} - \zeta \frac{\partial \eta_2(u^{t-1,s-1})}{\partial \theta} \right],$$
   $$\rho_s \leftarrow \frac{1 + \sqrt{1 + 4\rho_{s-1}^2}}{2}, \quad u^{t-1,s} \leftarrow \theta^{t-1,s} + \frac{\rho_{s-1} - 1}{\rho_s} (\theta^{t-1,s} - \theta^{t-1,s-1}).$$

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

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

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

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

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

**Remark 1.** There exist closed-form solutions for the proximal maps of $\nu h_1^*$, $\nu^{-1}h_1$, and $\zeta h_2$. Specifically, the proximal map of $\nu h_1^*$ is a projection function. And the proximal maps
of $\nu^{-1}h_1$ and $\varsigma h_2$ can be easily derived as in Ma and Huang (2017) with $\tau > \varsigma$. In Step 2.2, $\varsigma$ denotes the step size. As in Shimmura and Suzuki (2022), we can derive the Lipschitz constant of $\eta_2(\theta)$, denoted by $L_\eta = 1 + 2\nu \max_{k \in [K]} (AA^\top)_{k,k}$, and then set $\varsigma = L_\eta^{-1}$. With $\nu = 1$ and $\tau = 3$, $\tau > \varsigma$ since $\varsigma \leq 1/3$.

**Tuning parameter selection** Following the literature, we set $\nu = 1$ and the concavity related parameter $\tau = 3$. Following Yang et al. (2019), we select $\lambda_1$ and $\lambda_2$ with the modified BIC defined as

$$mBIC(\lambda_1, \lambda_2) = \frac{1}{N} \sum_{k=1}^{K} n_k (\theta^{(k)^\top} \bar{V}^{(k)} \theta^{(k)} - 2 \theta^{(k)^\top} \bar{\zeta}^{(k)}) + C_N \frac{\log N}{N} \tilde{q},$$

where $\tilde{q}$ 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 unique value.

3.1 **Notations and Definitions**

For a vector $z = (z_1, \ldots, z_p) \in \mathbb{R}^p$, and $1 \leq l < \infty$, define $||z||_l = (\sum_{j=1}^{p} |z_j|^l)^{1/l}$ and $||z||_\infty = \max_{j \in [p]} |z_j|$. Given an index set $S$, let $z_S$ denote the subvector of $z$ corresponding to the elements of $S$. For a matrix $Z_{s \times p}$, let $||Z||_2 = \sup_{v \in \mathbb{R}^p, ||v||_2 = 1} ||Zv||_2$, $||Z||_\infty = \max_{j \in [p]} |Z_{j,:}|$.
\[\max_{1 \leq i \leq s} \sum_{j=1}^{p} |Z_{ij}|, \quad \|Z\|_{\text{max}} = \max_{1 \leq i \leq s, 1 \leq j \leq p} |Z_{ij}|, \quad \text{and} \quad \|Z\|_{F} = \sqrt{\sum_{i=1}^{s} \sum_{j=1}^{p} Z_{ij}^2}.\]

For two index sets \(S_1\) and \(S_2\), let \(Z_{S_1,S_2}\) denote the submatrix of \(Z\) corresponding to the rows in \(S_1\) and columns in \(S_2\), and \(Z_{S^c}\) denote the submatrix of \(Z\) corresponding to the rows in \(S_1\). For a vector \(v_0 \in \mathbb{R}^p\), let \(B_r(v_0) = \{v \in \mathbb{R}^p : \|v - v_0\|_2 \leq r\}\) be the \(\ell_2\)-ball around \(v_0\) with radius \(r > 0\). For a random variable \(X\), its sub-Gaussian norm is defined by \(\|X\|_{\psi_2} = \sup_{s \geq 1} s^{-1/2}(\mathbb{E}|X|^s)^{1/s}\). For a random vector \(z \in \mathbb{R}^p\), its sub-Gaussian norm is defined by \(\|z\|_{\psi_2} = \sup_{v \in B_1(0)} ||v^\top z||_{\psi_2}\). For a symmetric matrix \(H\), its maximum and minimum eigenvalues are denoted by \(\Lambda_{\text{max}}(H)\) and \(\Lambda_{\text{min}}(H)\), respectively. For two sequences of real numbers \(\{a_n\} \geq 1\) and \(\{b_n\} \geq 1\), let \(a_n \gg b_n\) (or \(b_n \ll a_n\)) denote \(b_n/a_n = o(1)\), and \(a_n \asymp b_n\) denote \(a_n\) is of the same order as \(b_n\). 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 \(V^{(k)}(\theta^{(k)}) = \partial^2 \tilde{L}_k(\theta^{(k)})/\partial \theta^{(k)}\partial \theta^{(k)\top}\) and \(g^{(k)}(\theta^{(k)}) = \partial \tilde{L}_k(\theta^{(k)})/\partial \theta^{(k)}\). We further denote \(\tilde{V}^{(k)} = V^{(k)}(\theta^{(k)})\), \(V^{*^{(k)}}(\theta^{*^{(k)}}) = V^{(k)}(\theta^{*^{(k)}}),\) \(\tilde{g}^{(k)} = g^{(k)}(\theta^{(k)})\) and \(g^{*^{(k)}} = g^{(k)}(\theta^{*^{(k)}})\) for simplicity. Let \(\varphi^{(k)} = \|E(V_{A^c,A}^{(k)})[E(V_{A,A}^{(k)})]^{-1}\|_{\infty}\) for any \(k \in [K]\) and \(\varphi_{\text{max}} = \max_{k \in [K]} \varphi^{(k)}\).

Let \(N_m = \sum_{k \in G^{(m)}} n_k, \quad N_{\text{max}} = \max_{m \in [M]} N_m,\) and \(N_{\text{min}} = \min_{m \in [M]} N_m.\) Let \(|G^{(m)}|\) be the cardinality of index set \(G^{(m)}\) with \(m \in [M]\), and denote \(|G_{\text{max}}| = \max_{m \in [M]} |G^{(m)}|\) and \(|G_{\text{min}}| = \min_{m \in [M]} |G^{(m)}|\).

Let \(M_G\) be a subspace of \(\mathbb{R}^{p \times K}\) defined as \(M_G = \{\theta \in \mathbb{R}^{p \times K} : \theta^{(k)} = \theta^{(k')}, \text{for any } k, k' \in G^{(m)}, 1 \leq m \leq M\}\),

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

If the underlying true cluster memberships $G^{(1)}, \ldots, G^{(M)}$ are known, we can define the cluster-oracle objective function for $\theta$ by

$$\arg\min_{\theta \in M} \left\{ L(\theta) = \frac{1}{N} \sum_{k=1}^{K} n_k \left( (\theta^{(k)})^T \tilde{V}^{(k)} \theta^{(k)} - 2 \theta^{(k)}^T \tilde{\zeta}^{(k)} \right) + \sum_{j=2}^{p} p_r(||\theta_j||_2, \lambda_1) \right\}. \quad (8)$$

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

$$L^G(\psi) = \frac{1}{N} \sum_{m=1}^{M} \left[ (\psi^{(m)})^T \left( \sum_{k \in G^{(m)}} n_k \tilde{V}^{(k)} \right) \psi^{(m)} - 2 \psi^{(m)}^T \left( \sum_{k \in G^{(m)}} n_k \tilde{\zeta}^{(k)} \right) \right]$$

$$+ \sum_{j=2}^{p} p_r \left( \sqrt{\sum_{m=1}^{M} \left( ||G^{(m)}||_1/2 \psi^{(m)}_j \right)^2}, \lambda_1 \right). \quad (9)$$

### 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 $i \in [n_k]$, $\{x_i^{(k)}, y_i^{(k)}\}$’s are independent and identically distributed. There exists a constant $C_x > 0$ such that $\max_{k \in [K], i \in [n_k]} ||x_i^{(k)}||_\infty \leq C_x$ and $\max_{x \in B_1(0)} \mathbb{E}(x^T x_i^{(k)})^2 \leq C_x$.

(C2) For each $k \in [K]$ and $i \in [n_k]$, $f'(\theta^{*T} x_i^{(k)}, y_i^{(k)})$’s are sub-Gaussian. That is, there exists a constant $\kappa_x > 0$ such that $||f'(\theta^{*T} x_i^{(k)}, y_i^{(k)})||_{\psi_2} \leq \kappa_x$. 

14
For each $k \in [K]$, there exist two constants $C_{\text{min}}$ and $C_{\text{max}}$ such that $0 < C_{\text{min}} \leq \Lambda_{\text{min}}(\mathbb{E}(V^{*\text{(k)}}_{\mathcal{A}})) \leq \Lambda_{\text{max}}(\mathbb{E}(V^{*\text{(k)}}_{\mathcal{A}})) \leq C_{\text{max}}$.

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

$$|f''(\theta^{(k)}_{i} x^{(k)}_{ij}, y^{(k)}_{ij})| \leq C_L, \quad \text{for all } \theta^{(k)} \in B_{\delta}(\theta^{*\text{(k)}}) .$$

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

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

The local estimators satisfy

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

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 $\lambda^{-1} p_{\tau}'(0+, \lambda) = 1$.

Condition (C1) assumes that all covariates are uniformly bounded. Similar conditions have been commonly assumed in the literature, especially including Cai et al. (2021). It is satisfied under many practical scenarios. Condition (C2) controls the tail behavior of $X^{(k)} f'(a, y)$ and bounds the random error $g^{*\text{(k)}}$. Condition (C3) has been commonly assumed to ensure that the eigenvalues of $\mathbb{E}(V^{*\text{(k)}}_{\mathcal{A}})$ 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. (2021) 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.

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

$$
||\hat{\psi}^{or}_A - \psi^*_A||_F = O_p(r_{1N}), \quad \hat{\psi}^{or}_{A^c} = 0,
$$

with probability approaching 1 as $N \to \infty$, where

$$
r_{1N} = \sqrt{\left(\frac{K}{|G_{\min}|}\right)q} \left(\frac{|G_{\max}|^2q^3/2 \log p}{N_{\min}}\right),$$

$$
r_{2N} = \sqrt{\left(\frac{|G_{\max}|/|G_{\min}|^2}{KN}\right)M \log p + \left(\frac{|G_{\max}|/|G_{\min}|^1/2}{N}\right)M^{1/2}q \log p}.
$$

Theorem 1 establishes the estimation consistency and model selection consistency of the oracle estimator $\hat{\psi}^{or}$. 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 $M|G_{\max}| = o(\sqrt{N/q^2(\log p)^2})$, 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 $|G_m| \asymp K/M$ 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 $K = o(\sqrt{N/q^2(\log p)^2})$, and this condition is similar to that for the distributed sparse high-dimensional models (Battey et al., 2018).

**Corollary 1.** Suppose that the conditions of Theorem 1 hold. Then there exists a strictly...
local minimizer \( \hat{\theta}^{or} \) of \( \mathcal{L}(\theta) \) such that

$$||\hat{\theta}^{or}_{A} - \theta^{*}_{A}||_{F} = O_{p}(|\mathcal{G}_{\max}|^{1/2}r_{1N}), \quad \hat{\theta}^{or}_{A^{c}} = 0,$$

with probability approaching 1 as \( N \to \infty \).

Corollary 1 directly follows from Theorem 1 and can be used to derive the following theorem.

**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 \( \hat{\theta} \) of \( \hat{\mathcal{Q}}_{ICR}(\theta) \)
that satisfies

$$P(\hat{\theta} = \hat{\theta}^{or}) \to 1.$$  

Theorem 2 shows that the oracle estimator \( \hat{\theta}^{or} \) is a local minimizer with a high probability. As clustering is based on estimation, the clustering consistency follows from Theorems 1 and 2.

**4 Simulation Study**

We conduct simulation to gauge performance of the proposed approach. For benchmarking, we consider: (a) the Local estimator obtained by minimizing a local loss function for each client separately; (b) the SK estimators obtained by applying the sparse K-means approach \( \text{[Witten and Tibshirani 2010]} \) to the local estimators. This is achieved using R package \textit{sparcl}. We adopt two criteria, namely the Hartigan statistic \( \text{[Hartigan 1975]} \) and gap statistic \( \text{[Tibshirani et al. 2001]} \), to choose the number of clusters – this is realized using R package \textit{NbClust}. The two approaches are accordingly referred to as SK(har)
and SK(gap), respectively; (c) the 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); and (d) the Oracle estimator obtained by minimizing objective function $L_{\text{or},G}(\psi)$ in (A.3). Note that this alternative is not realistic in practice. Here, it serves as the ideal target. For the alternatives, tuning parameters are chosen in a way compatible with the proposed.

Four examples are designed. Examples 1-2 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 and 4. For all examples, $n_1 = \cdots = n_K = n$. More specific settings are as follows.

**Example 1.** $\psi^{(1)} = (0.4 \times 1_8^\top, 0_{p-8}^\top)^\top$ and $\psi^{(2)} = (-0.4 \times 1_8^\top, 0_{p-8}^\top)^\top$. We generate $x_{i_{-1}, k}^{(k)}, i \in [n_k], k \in [K]$, where $x_{i_{-1}, k}^{(k)} = (x_{i_2}^{(k)}, \ldots, x_{ip}^{(k)})^\top$, from a multivariate normal distribution with mean 0 and cov$(X_w, X_t) = \rho^{|w-t|}$ for $w, t \in \{2, \cdots, p\}$ and $\rho = 0.5$. Given $X^{(k)}$, we generate $Y^{(k)}$ from a logistic model. We set the number of datasets in each cluster as $|G_1| = |G_2| = K/2$. We further set $n = 200$ and consider $K \in \{16, 32, 64\}$ and $p \in \{100, 500\}$.  

**Example 2.** $\psi^{(1)} = (0.6 \times 1_4^\top, -0.6 \times 1_4^\top, 0_{p-8}^\top)^\top$, $\psi^{(2)} = (0.6 \times 1_2^\top, -0.6 \times 1_2^\top, 0_6 \times 1_2^\top, -0.6 \times 1_2^\top, 0_6 \times 1_2^\top, 0_{p-8}^\top)^\top$, $\psi^{(3)} = (-0.6 \times 1_2^\top, 0_6 \times 1_2^\top, -0.6 \times 1_2^\top, 0_6 \times 1_2^\top, 0_{p-8}^\top)^\top$, and $\psi^{(4)} = (-0.6 \times 1_4^\top, 0_6 \times 1_4^\top, 0_{p-8}^\top)^\top$. We set the number of datasets in each cluster as $|G_1| = |G_2| = |G_3| = |G_4| = K/4$. $X^{(k)}$ and $Y^{(k)}$ 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 $N(0, \sigma^2)$. We consider $K \in \{16, 32, 64\}$ and $\sigma \in \{1, 2\}$ and set $p = 100$ and $n = 100$. 

18
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 $K = 64$.

For each example, we generate 100 replicates. We first observe that the proposed computational algorithm has satisfactory convergence properties. With all of our simulated datasets, convergence is achieved within 100 iterations. Additionally, the proposed approach is computationally affordable. For example, the analysis of one simulated dataset 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 $\hat{A} = \{j : \hat{\theta}_j \neq 0\}$. 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 $MS = \sum_{j=1}^{p} \hat{q}_j$, where $\hat{q}_j$ is the number of distinct nonzero coefficients of the $j$th variable. For evaluating clustering accuracy, we consider: (4) $\hat{M}$, the number of identified clusters; (5) Per, the percentage of fully accurate identification; (6) RI, the Rand Index defined as $RI = (TP + TN)/(TP + FP + FN + TN)$, where TP (true positive) is the number of pairs of datasets 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 $ARI = (RI - \mathbb{E}(RI))/\max(\text{RI})$, where $\mathbb{E}(RI)$ and $\max(\text{RI})$ 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 evaluating estimation, we consider: (8) RMSE, the root mean squared error of $\hat{\theta}$ defined as $\text{RMSE} = \sqrt{\frac{\sum_{j \in A} ||\hat{\theta}_j - \theta^*_j||^2}{2}}/K$.

Results for Example 1 are summarized in Table 1 Table 2 and Figure 2. It is observed that the proposed approach has larger TPR and smaller FPR values, and performance improves as $K$ increases. It has the average MS values around 16, which is the true model size, while the alternatives generate much larger models. When $K$ is sufficiently large, it outperforms the alternatives in variable selection. Compared to the integrative analysis alternatives SHIR and SMA, the Local approach has worse variable selection performance, which may be caused by the limited sample sizes. Besides, compared to the two-step alternatives SK(har) and SK(gap), it has better variable selection performance. The proposed method and SK(gap) have higher clustering accuracy, while SK(har) often overestimates the number of clusters. Figure 2 suggests that estimation accuracy of the proposed approach is very close to that of Oracle, especially when $K$ is large.

Results for Example 2 are summarized in Table 4 Table 5 and Figure 3 (Appendix B). It is observed that the proposed approach has higher variable selection and estimation accuracy, especially when $n$ is large. Its performance is much closer to Oracle. Compared to the Local approach, SHIR and SMA has worse estimation performance, which suggests that inappropriate data integration may not help. Table 5 shows that, with the proposed approach 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. A larger $n$ leads to better performance. In comparison, SK(gap) usually underestimates the number of clusters and has much lower clustering accuracy.
Results for Examples 3 and 4 are summarized in Tables 5-9 and Figures 4-5 (Appendix B). The overall findings are very similar to Examples 1-2.

Table 1: Simulation Example 1, variable selection accuracy: mean (sd) based on 100 replicates.

| Method   | TPR  | FPR  | MS    | TPR  | FPR  | MS    | TPR  | FPR  | MS    |
|----------|------|------|-------|------|------|-------|------|------|-------|
|          | K = 16 |       |       | K = 32 |       |       | K = 64 |       |       |
| p = 100  |       |       |       |       |       |       |       |       |       |
| ICR      | 0.990 (0.058) | 0.000 (0.000) | 15.840 (0.929) | 1.000 (0.000) | 0.000 (0.000) | 16.000 (0.000) | 1.000 (0.000) | 0.000 (0.000) | 16.000 (0.000) |
| SHIR     | 1.000 (0.000) | 0.010 (0.012) | 128.900 (1.087) | 1.000 (0.000) | 0.008 (0.009) | 256.730 (0.863) | 1.000 (0.000) | 0.009 (0.010) | 512.670 (0.922) |
| SMA      | 1.000 (0.000) | 0.008 (0.011) | 128.780 (0.991) | 1.000 (0.000) | 0.007 (0.008) | 256.660 (0.742) | 1.000 (0.000) | 0.007 (0.007) | 512.920 (6.402) |
| Local    | 0.889 (0.026) | 0.102 (0.020) | 264.180 (29.586) | 0.893 (0.018) | 0.100 (0.013) | 524.470 (37.805) | 0.891 (0.013) | 0.102 (0.010) | 1056.990 (104.202) |
| SK(har)  | 0.999 (0.051) | 0.004 (0.161) | 173.690 (57.414) | 1.000 (0.000) | 0.004 (0.009) | 257.730 (3.681) | 1.000 (0.000) | 0.004 (0.010) | 513.080 (14.193) |
| SK(gap)  | 1.000 (0.000) | 0.003 (0.010) | 120.000 (20.013) | 1.000 (0.000) | 0.003 (0.007) | 257.230 (13.775) | 1.000 (0.000) | 0.003 (0.008) | 512.670 (4.613) |

Table 2: Simulation Example 1, clustering accuracy: mean (sd) based on 100 replicates.

| Method   | $\hat{M}$ | Per | RI   | ARI  | $\hat{M}$ | Per | RI   | ARI  | $\hat{M}$ | Per | RI   | ARI  |
|----------|------------|-----|------|------|------------|-----|------|------|------------|-----|------|------|
| p = 100  |            |     |      |      |            |     |      |      |            |     |      |      |
| ICR      | 2.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | 2.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | 2.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) |
| SK(har)  | 4.640 (1.508) | 0.000 (0.000) | 0.779 (0.097) | 0.539 (0.208) | 5.230 (1.847) | 0.000 (0.098) | 0.759 (0.202) | 0.508 (1.495) | 5.220 (1.847) | 0.000 (0.098) | 0.734 (0.202) | 0.462 (1.313) |
| SK(gap)  | 2.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | 2.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | 2.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) |

Table 2: Simulation Example 1, clustering accuracy: mean (sd) based on 100 replicates.

| Method   | $\hat{M}$ | Per | RI   | ARI  | $\hat{M}$ | Per | RI   | ARI  | $\hat{M}$ | Per | RI   | ARI  |
|----------|------------|-----|------|------|------------|-----|------|------|------------|-----|------|------|
| p = 500  |            |     |      |      |            |     |      |      |            |     |      |      |
| ICR      | 2.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | 2.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | 2.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) |
| SK(har)  | 4.580 (1.505) | 0.000 (0.101) | 0.793 (0.216) | 0.570 (1.684) | 5.440 (1.684) | 0.000 (0.097) | 0.754 (0.199) | 0.499 (1.720) | 5.150 (1.720) | 0.000 (0.098) | 0.766 (0.199) | 0.528 (0.199) |
| SK(gap)  | 2.010 (0.100) | 0.999 (0.006) | 0.999 (0.012) | 2.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | 2.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) |
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, 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 request 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 6 (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 10, 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 GET and POST parameter, 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 \( G_{lx} \) and \( P_{lx} \), 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
\( G_{tx} \) and \( P_{tx} \), 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 10
(Appendix B) as an example. We can obtain the following feature values: \( G_{num} = 1 \), \( G_{len} = 9 \), \( P_{num} = 4 \), \( P_{len} = 72 \), \( Gl0 = 9 \), \( Gl1 = \cdots = Gl19 = 0 \), \( Pl0 = 19 \), \( Pl1 = 16 \), \( Pl2 = 10 \),
\( Pl3 = 24 \), \( Pl4 = \cdots = Pl19 = 0 \), \( Gt0 = "str" \), \( Gt1 = \cdots = Gt19 = "na" \), \( Pt0 = \cdots = Pt2 = "str" \), \( Pt3 = "num" \), \( Pt4 = \cdots = Pt19 = "na" \). Since the values of \( Gl1, \ldots, Gl19 \)
are 0 for all requests, we delete these features. To further utilize the character strings, we
concatenate \( Gt0 - Gt19 \) sequentially into a sequence of strings and train the Skip-gram
model (which is a popular model of word2vec) to obtain a 40-dimensional continuous word
vector features, denoted by \( Gw1, \ldots, Gw40 \). Similarly, we obtain another 40-dimensional
continuous word vector features, denoted by \( Pw1, \ldots, Pw40 \), from \( Pt0 - Pt19 \). 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 seven nontrivial clusters (with sizes larger
than one and denoted as \( \text{ICR}^{(1)}, \ldots, \text{ICR}^{(7)} \)), which have sizes 35, 34, 8, 7, 3, 2 and 2,
respectively. Additionally, there are 32 interfaces forming their own individual clusters.
In Table 3, we present important variables identified by the proposed method and/or the
two integrative analysis alternatives SHIR and SMA. 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. From Table 3, we can see that, with the proposed method, Gnum has a negative effect on the probability of abnormal request, while Gl0, Pl0, Pl1 and Pl2 have positive effects. This suggests that requests with less GET key-value pairs, 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 security rule for the initial screening of abnormal requests. Here we note that the traditional anomaly detection for logs is to extract security rules from samples (El Hadj et al., 2018). Additionally, the proposed analysis has reduced the number of models to 39 (7 clustered ones and 32 individual ones), which corresponds to a much lower cost of maintaining models than the alternative client-specific modeling methods (with 123 models).

With practical data, it is difficult to objectively evaluate identification and estimation results. To support our finding, we conduct a prediction evaluation. Specifically, we randomly select 3/4 of the samples and form the training data. In this selection, the normal:abnormal ratio is retained. The remaining samples forms the testing data. Estimation is conducted using the training data, and we evaluate prediction performance on the testing data using AUC and brier score. This process is repeated 100 times, and the results are summarized in Figure 7 (Appendix B). It is observed that the proposed method has better prediction performance than the alternatives. Additionally, the Local method also has competitive performance, which suggests that inappropriate integration may lead to inferior prediction. Compared to Local, the proposed method can have better interpretability.
Table 3: Data analysis: identified important variables and their estimates using the three integrative analysis methods. For the proposed method, only estimates for the nontrivial clusters are shown.

| Variable | ICR$^{(1)}$ | ICR$^{(2)}$ | ICR$^{(3)}$ | ICR$^{(4)}$ | ICR$^{(5)}$ | ICR$^{(6)}$ | ICR$^{(7)}$ | SHIR | SMA |
|----------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|------|-----|
| Intercept | -0.168      | 3.967       | -1.207      | 9.173       | 4.863       | 4.330       | 1.150       | -1.757 | -1.754 |
| Gnum     | -0.027      | -0.522      | -0.002      | -0.430      | -0.054      | -0.116      | -0.077      | –     | –   |
| Glen     | 0.011       | -0.278      | 0.096       | -0.709      | 0.126       | 0.292       | -0.049      | -0.054 | -0.055 |
| Pnum     | 0.062       | 0.036       | -0.052      | 0.036       | -0.079      | 0.736       | -0.228      | 0.004  | –   |
| Plen     | 0.164       | -0.066      | -0.740      | -0.754      | 0.465       | -1.374      | -0.109      | –     | –   |
| Gl0      | 0.142       | 2.047       | 0.603       | 2.377       | 0.005       | 1.564       | 2.559       | 0.152  | 0.152 |
| P10      | 0.197       | 1.841       | 1.463       | 2.384       | 0.857       | 2.170       | 0.930       | 0.206  | 0.212 |
| P11      | 0.523       | 5.356       | 4.939       | 8.852       | 3.925       | 8.744       | 1.271       | 0.533  | 0.542 |
| P12      | 0.396       | 3.984       | 5.607       | 9.839       | 8.197       | 8.885       | 3.925       | 0.437  | 0.449 |
| P14      | 0.034       | 0.011       | 0.061       | 0.017       | -0.016      | -0.005      | -0.002      | 0.068  | 0.068 |
| P15      | 0.006       | 0.025       | 0.133       | 0.011       | -0.024      | 0.201       | -0.016      | 0.026  | 0.024 |
| P16      | 0.004       | -0.002      | 0.009       | 0.010       | -0.021      | 0.002       | -0.018      | –     | –   |
| P19      | –           | –           | –           | –           | –           | –           | 0.024       | 0.025  | –   |
| P110     | –           | –           | –           | –           | –           | –           | 0.010       | 0.010  | –   |
| P112     | 0.005       | 0.005       | 0.007       | 0.032       | -0.007      | -0.530      | 0.005       | –     | –   |
| P117     | -0.003      | 0.008       | 0.014       | -0.039      | -0.005      | -0.001      | 0.001       | –     | –   |
| Gw11     | –           | –           | –           | –           | –           | –           | 0.026       | 0.026  | –   |
| Gw31     | –           | –           | –           | –           | –           | –           | –           | -0.007 | –   |
| Gw32     | –           | –           | –           | –           | –           | –           | -0.005      | –     | –   |
| Gw34     | –           | –           | –           | –           | –           | –           | -0.006      | -0.006 | –   |
| Gw36     | –           | –           | –           | –           | –           | –           | 0.003       | 0.004  | –   |
| Pw1      | 0.012       | -0.062      | 0.116       | -0.043      | -0.104      | -1.165      | -0.266      | –     | –   |
| Pw27     | –           | –           | –           | –           | –           | –           | 0.013       | 0.018  | –   |
| Pw37     | –           | –           | –           | –           | –           | –           | 0.018       | 0.015  | –   |

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 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 between the local clients and 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 communications (which may lead to higher computational cost), 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. The proposed strategy can be potentially applied to much broader models/loss functions. Another possible extension, as previously mentioned, is to apply two-level penalized selection and allow different sparsity structures for multiple clients. These aforementioned extensions will be postponed to future research.

**Acknowledgements**

This study has been partly supported by NSF 1916251 and 2209685.

**References**

Abbe, E. A., Khandani, A. E., and Lo, A. W. (2012), “Privacy-preserving methods for sharing financial risk exposures,” *American Economic Review*, 102(3), 65–70.

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

Bhowmick, A., Duchi, J., Freudiger, J., Kapoor, G., and Rogers, R. (2018), “Protection
against reconstruction and its applications in private federated learning,” *arXiv preprint arXiv:1812.00984*, .

Bühlmann, P., and van de Geer, S. (2011), *Statistics for high-dimensional data: methods, theory and applications*, Heidelberg: Springer.

Cai, T., Liu, M., and Xia, Y. (2021), “Individual data protected integrative regression analysis of high-dimensional heterogeneous data,” *Journal of the American Statistical Association*, pp. 1–15.

Chen, J., Tran-Dinh, Q., Kosorok, M. R., and Liu, Y. (2021), “Identifying heterogeneous effect using latent supervised clustering with adaptive fusion,” *Journal of Computational and Graphical Statistics*, 30(1), 43–54.

Chi, E. C., and Lange, K. (2015), “Splitting methods for convex clustering,” *Journal of Computational and Graphical Statistics*, 24(4), 994–1013.

Dobriban, E., and Sheng, Y. (2020), “WONDER: Weighted One-shot Distributed Ridge Regression in High Dimensions,” *Journal of Machine Learning Research*, 21(66), 1–52.

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

El Hadj, M. A., Erradi, M., Khoumsi, A., and Benkaouz, Y. (2018), “Validation and correction of large security policies: a clustering and access log based approach,” *2018 IEEE international conference on big Data (Big Data)*, pp. 5330–5332.

Fan, J., and Li, R. (2001), “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, 96(456), 1348–1360.
Fan, J., and Lv, J. (2008), “Sure independence screening for ultrahigh dimensional feature space,” *Journal of the Royal Statistical Society, Series B*, 70, 849–911.

Ghosh, A., Chung, J., Yin, D., and Ramchandran, K. (2020), “An efficient framework for clustered federated learning,” *Advances in Neural Information Processing Systems*, 33, 19586–19597.

Guo, Y., Wu, Y., Zhu, Y., Yang, B., and Han, C. (2021), “Anomaly detection using distributed log data: A lightweight federated learning approach,” *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Hartigan, J. A. (1975), *Clustering algorithms*, New York: Wiley.

He, Q., Zhang, H. H., Avery, C. L., and Lin, D. Y. (2016), “Sparse meta-analysis with high-dimensional data,” *Biostatistics*, 17(2), 205–220.

Hu, Q., Tang, B., and Lin, D. (2017), “Anomalous user activity detection in enterprise multi-source logs,” *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 797–803.

Huang, Y., Zhang, Q., Zhang, S., Huang, J., and Ma, S. (2017), “Promoting similarity of sparsity structures in integrative analysis with penalization,” *Journal of the American Statistical Association*, 112(517), 342–350.

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

Lee, J. D., Liu, Q., Sun, Y., and Taylor, J. E. (2017), “Communication-efficient sparse regression,” *The Journal of Machine Learning Research*, 18(1), 115–144.
Liu, J., Huang, J., Zhang, Y., Lan, Q., Rothman, N., Zheng, T., and Ma, S. (2014), “Integrative analysis of prognosis data on multiple cancer subtypes,” *Biometrics*, 70(3), 480–488.

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

Marfoq, O., Neglia, G., Bellet, A., Kameni, L., and Vidal, R. (2021), “Federated multi-task learning under a mixture of distributions,” *Advances in Neural Information Processing Systems*, 34, 15434–15447.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), “A unified framework for high-dimensional analysis of estimators with decomposable regularizers,” *Statistical Science*, 27(4), 538–557.

Parikh, N., Boyd, S. et al. (2014), “Proximal algorithms,” *Foundations and trends® in Optimization*, 1(3), 127–239.

Shimmura, R., and Suzuki, J. (2022), “Converting ADMM to a proximal gradient for efficient sparse estimation,” *Japanese Journal of Statistics and Data Science*, pp. 1–21.

Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. S. (2017), “Federated multi-task learning,” *Advances in Neural Information Processing Systems*, 30.

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

Tang, X., Xue, F., and Qu, A. (2021), “Individualized multidirectional variable selection,” *Journal of the American Statistical Association*, 116(535), 1280–1296.
Tibshirani, R., Walther, G., and Hastie, T. (2001), “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.

Ünal, U., and Dağ, H. (2022), “Anomalyadapters: Parameter-efficient multi-anomaly task detection,” *IEEE Access*, 10, 5635–5646.

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

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

Xue, L., Zou, H., and Cai, T. (2012), “Nonconcave penalized composite conditional likelihood estimation of sparse Ising models,” *The Annals of Statistics*, 40(3), 1403–1429.

Yang, X., Yan, X., and Huang, J. (2019), “High-dimensional integrative analysis with homogeneity and sparsity recovery,” *Journal of Multivariate Analysis*, 174, 104529.

Zhang, C.-H. (2010), “Nearly unbiased variable selection under minimax concave penalty,” *The Annals of Statistics*, 38(2), 894–942.

Zhao, T., Cheng, G., and Liu, H. (2016), “A partially linear framework for massive heterogeneous data,” *Annals of statistics*, 44(4), 1400.

Zhu, X., Li, F., and Wang, H. (2021), “Least-square approximation for a distributed system,” *Journal of Computational and Graphical Statistics*, 30(4), 1004–1018.

Zou, H., and Li, R. (2008), “One-step sparse estimates in nonconcave penalized likelihood models,” *Annals of statistics*, 36(4), 1509.
Appendix A

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

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

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

**Proof.** Lemma 1 follows directly from Lemma 14.3, Chapter 14.2.2 of Bühlmann and van de Geer (2011).

**Lemma 2.** Assume that Conditions (C1), (C4), and (C5) hold. If $K \ll p$ and $q \log p \ll n^*$, then with probability approaching 1,

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

**Proof.** Note that,

$$\max_{k \in [K]} \left\| \widetilde{V}^{(k)} - \mathbb{E}(V^{*(k)}) \right\|_{\text{max}} \leq \max_{k \in [K]} \left\| \widetilde{V}^{(k)} - \mathbb{E}(\widetilde{V}^{(k)}) \right\|_{\text{max}} + \max_{k \in [K]} \left\| \mathbb{E}(\widetilde{V}^{(k)}) - \mathbb{E}(V^{*(k)}) \right\|_{\text{max}} \triangleq I_1 + I_2. \tag{A.1}$$

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

$$\left| \tilde{v}_{j_1j_2}^{(k)} \right| = \left| n_k^{-1} \sum_{i=1}^{n_k} f''(\tilde{\theta}^{(k)^\top} \mathbf{x}_i^{(k)}, y_i^{(k)}) x_{ij_1} x_{ij_2} \right| \leq C_L C_x^2.$$

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

$$P(I_1 > t) \leq 2Kp^2 \exp \left\{ - \frac{\min_{k \in [K]} n_k t^2}{2C'_x C_L^2} \right\},$$

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),

$$
\left| \mathbb{E}(\tilde{V}_{j_1 j_2}^{(k)}) - \mathbb{E}(V^{*(k)}_{j_1 j_2}) \right| \leq \mathbb{E}\left( \left| f''(\tilde{\theta}^{(k)}_i x^{(k)}_i, y^{(k)}_i) x_{i j_1} x_{i j_2} - f''(\theta^{*(k)}_i x^{(k)}_i, y^{(k)}_i) x_{i j_1} x_{i j_2} \right| \right)
$$

$$
\leq C_L C_x^2 \mathbb{E}\left[ \left| x^{(k)}_i (\tilde{\theta}^{(k)} - \theta^{*(k)}) \right| \right]
$$

$$
\leq C_L C_x^2 \max_{v \in \mathbb{R}^{(p)}} \mathbb{E}\left[ (v^\top x_i)^2 \right] \left\| \tilde{\theta}^{(k)} - \theta^{*(k)} \right\|_2^{1/2}
$$

$$
\leq C_L C_x^{5/2} \left\| \tilde{\theta}^{(k)} - \theta^{*(k)} \right\|_2.
$$

Thus,

$$
I_2 \leq C_L C_x^{5/2} \max_{k \in [K]} \left\| \tilde{\theta}^{(k)} - \theta^{*(k)} \right\|_2 = O_p(\sqrt{q \log p/n^*}).
$$

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

**Proof of Theorem 1**

Recall the definitions of $L(\theta)$ in (8) and $L^G(\psi)$ in (9). We further define two objective functions. If the true clustering structure and important covariate set are known, we can define the oracle estimator $\hat{\theta}^{or}$ as

$$
\arg\min_{\theta \in \mathcal{M}_G, \theta_{Ac} = 0} L^{or}(\theta) = \frac{1}{N} \sum_{k=1}^{K} n_k \left( \theta^{(k)}_i \tilde{V}^{(k)} \theta^{(k)} - 2 \theta^{(k)}_i \tilde{\zeta}^{(k)} \right).
$$

(A.2)

Accordingly, the oracle estimator $\hat{\psi}^{or} = (\hat{\psi}_A^{or\top}, 0^{(p-q)\times K})^\top$ of $\psi$ can be defined as

$$
\arg\min_{\psi \in \mathbb{R}^{p \times M}, \psi_{Ac} = 0} L^{or,G}(\psi) = \frac{1}{N} \sum_{m=1}^{M} \left[ \psi^{(m)^\top} \left( \sum_{k \in \mathcal{G}(m)} n_k \tilde{V}^{(k)} \right) \psi^{(m)} - 2 \psi^{(m)^\top} \left( \sum_{k \in \mathcal{G}(m)} n_k \tilde{\zeta}^{(k)} \right) \right].
$$

(A.3)

This theorem can be proved via two steps.

**Step 1:** Let $\psi_A = (\psi_A^{(1)}, \ldots, \psi_A^{(M)})$ and $\psi_A^{(m)} = (\psi_A^{(m)}_1, \ldots, \psi_A^{(m)}_q)^\top$. Based on the defini-
tion of $L_{\mathcal{A}}^{or, G}(\psi)$, we can further define

$$L_{\mathcal{A}}^{or, G}(\psi_{\mathcal{A}}) = \frac{1}{N} \sum_{m=1}^{M} \left[ \psi^{(m)}_{\mathcal{A}}^T \left( \sum_{k \in G^{(m)}} n_k \tilde{V}^{(k)}_{\mathcal{A}} \right) \psi^{(m)}_{\mathcal{A}} - 2 \psi^{(m)}_{\mathcal{A}}^T \left( \sum_{k \in G^{(m)}} n_k \tilde{\zeta}^{(k)}_{\mathcal{A}} \right) \right]. \quad (A.4)$$

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

$$\sum_{m=1}^{M} \left( \tilde{\psi}_{\mathcal{A}}^{or(m)}^T \tilde{V}^{G(m)}_{\mathcal{A}} \tilde{\psi}_{\mathcal{A}}^{or(m)} - 2 \tilde{\psi}_{\mathcal{A}}^{or(m)}^T \tilde{\zeta}^{G(m)}_{\mathcal{A}} \right) \leq \sum_{m=1}^{M} \left( \psi_{\mathcal{A}}^{*(m)}^T \tilde{V}^{G(m)}_{\mathcal{A}} \psi_{\mathcal{A}}^{*(m)} - 2 \psi_{\mathcal{A}}^{*(m)}^T \tilde{\zeta}^{G(m)}_{\mathcal{A}} \right), \quad (A.5)$$

where

$$\tilde{V}^{G(m)} = N^{-1} \sum_{k \in G^{(m)}} n_k \tilde{V}^{(k)}_G, \quad \tilde{\zeta}^{G(m)} = N^{-1} \sum_{k \in G^{(m)}} n_k \tilde{\zeta}^{(k)}_G.$$ 

Motivated by Cai et al. (2021), 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 $-\tilde{\zeta}^{(k)}$ in (A.5) into

$$\tilde{g}^{(k)} - \tilde{V}^{(k)} \theta^{(k)} = \tilde{g}^{*^{(k)}} - \tilde{v}^{(k)} \theta^{*^{(k)}} + \int_0^1 \left\{ v^{(k)} \left[ \theta^{*^{(k)}} + t(\tilde{\theta}^{(k)} - \theta^{*^{(k)})} \right] - \tilde{v}^{(k)} \right\} \left( \tilde{\theta}^{(k)} - \theta^{*^{(k)}} \right) dt. \quad (A.6)$$

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

$$\sum_{m=1}^{M} \left[ \left( \tilde{\psi}_{\mathcal{A}}^{or(m)} - \psi_{\mathcal{A}}^{*(m)} \right)^T \tilde{V}^{G(m)}_{\mathcal{A}} \left( \tilde{\psi}_{\mathcal{A}}^{or(m)} - \psi_{\mathcal{A}}^{*(m)} \right) \right] \leq - \sum_{m=1}^{M} \left( \tilde{\psi}_{\mathcal{A}}^{or(m)} - \psi_{\mathcal{A}}^{*(m)} \right)^T \left[ N^{-1} \sum_{k \in G^{(m)}} n_k \xi^{*(k)} \right]_{\mathcal{A}} - \sum_{m=1}^{M} \left( \tilde{\psi}_{\mathcal{A}}^{or(m)} - \psi_{\mathcal{A}}^{*(m)} \right)^T$$

$$\times \left[ N^{-1} \sum_{k \in G^{(m)}} n_k \int_0^1 \left\{ v^{(k)} \left[ \theta^{*^{(k)}} + t(\tilde{\theta}^{(k)} - \theta^{*^{(k)})} \right] - \tilde{v}^{(k)} \right\} \left( \tilde{\theta}^{(k)} - \theta^{*^{(k)}} \right) dt \right]_{\mathcal{A}}. \quad (A.7)$$

Let $\alpha = [ \text{vec}(\tilde{\psi}_{\mathcal{A}}^{or}) - \text{vec}(\psi_{\mathcal{A}}^*) ]$, $\tilde{V}^{G,A} = \text{bdiag}(\tilde{V}^{G(1)}_{\mathcal{A}}, \ldots, \tilde{V}^{G(M)}_{\mathcal{A}})$, $\xi = (\xi^{(1)}, \ldots, \xi^{(M)})$,
\( \eta = \left( \eta^{(1)}, \ldots, \eta^{(M)} \right) \), where

\[
\xi^{(m)} = \left[ N^{-1} \sum_{k \in G^{(m)}} n_k g^{(k)} \right],
\]

\[
\eta^{(m)} = \left[ N^{-1} \sum_{k \in G^{(m)}} n_k \int_0^1 \left\{ V^{(k)} \left( \left[ \theta^{(k)} + t(\tilde{\theta}^{(k)} - \theta^{(k)}) \right] - \tilde{V}^{(k)} \right) \right\} (\tilde{\theta}^{(k)} - \theta^{(k)}) \right] dt \right].
\]

Then by (A.7), we have

\[
\alpha^\top \tilde{V}^{(G, A)} \alpha \leq ||\alpha^\top \text{vec}(\xi_A)||_2 + ||\alpha^\top \text{vec}(\eta_A)||_2. \tag{A.8}
\]

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

\[
||\alpha||_2^2 \Lambda_{\min}(\tilde{V}^{(G, A)}) \leq ||\alpha^\top \text{vec}(\xi_A)||_2 + ||\alpha^\top \text{vec}(\eta_A)||_2. \tag{A.9}
\]

To bound \( \Lambda_{\min}(\tilde{V}^{(k)}_{AA}) \), noting that for each \( k \in [K] \), we have

\[
\Lambda_{\min}(\tilde{V}^{(k)}_{AA}) \geq \Lambda_{\min} \left[ \tilde{V}^{(k)}_{AA} - \mathbb{E}(V^{(k)}_{AA}) \right] + \Lambda_{\min} \left[ \mathbb{E}(V^{(k)}_{AA}) \right] \geq - ||\tilde{V}^{(k)}_{AA} - \mathbb{E}(V^{(k)}_{AA})||_F + \Lambda_{\min} \left[ \mathbb{E}(V^{(k)}_{AA}) \right]. \tag{A.10}
\]

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

\[
||\tilde{V}^{(k)}_{AA} - \mathbb{E}(V^{(k)}_{AA})||_F \leq \left( q^2 \max_{k \in [K]} ||\tilde{V}^{(k)} - \mathbb{E}(V^{(k)})||_2^2 \right)^{1/2} = \mathcal{O}_p(\sqrt{q^3 \log p/n^*}) = o_p(1). \tag{A.11}
\]

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

\[
\frac{C_{\min} N_{\min}}{2N} ||\alpha||_2^2 \leq ||\alpha||_2 \left| \left| \text{vec}(\xi_A) \right| \right|_2 + ||\alpha||_2 \left| \left| \text{vec}(\eta_A) \right| \right|_2. \tag{A.12}
\]

Now we bound \( \left| \left| \text{vec}(\xi_A) \right| \right|_2 \) and \( \left| \left| \text{vec}(\eta_A) \right| \right|_2 \). For \( \left| \left| \text{vec}(\xi_A) \right| \right|_2 \), since the product of a sub-Gaussian random variable and a bounded random variable is also sub-Gaussian and
\[ \mathbb{E}(n_k g^{(k)}) = \mathbb{E}(X^{(k) \top} \Phi^{(k)}) \text{ with } \Phi^{(k)} = (f'(x_1^{(k) \top} \theta^{*(k)}), \ldots, f'(x_{n_k}^{(k) \top} \theta^{*(k)}, y_{n_k}^{(k)}))^{\top}, \]
by Conditions (C1), (C2), and Lemma [1] for all \( m \in [M], j \in [p] \) and any \( t > 0 \),
\[
P\left( \frac{1}{\sqrt{N_m}} \sum_{k \in G^{(m)}} n_k g_j^{*(k)} \geq t \right) \leq 2 \exp\left( -\frac{C_1 t^2}{C_2^2 N_x^2} \right).
\]
Accordingly, there exists a constant \( C_2 > 0 \) such that \( \mathbb{E}[(\sum_{k \in G^{(m)}} n_k g_j^{*(k)})^2] \leq C_2 N_m \). Then,
\[
P\left( \|\text{vec}(\xi_A)\|_2 \geq t \right) \leq \sum_{m=1}^m \mathbb{E}[\|\xi_A^{(m)}\|_2^2] \leq \sum_{m=1}^M \sum_{j \in A} (\sum_{k \in G^{(m)}} n_k g_j^{*(k)})^2 \leq \frac{C_2 q}{N t^2},
\]
which leads to \( \|\text{vec}(\xi_A)\|_2 = O_p(\sqrt{q/N}) \).

For \( \|\text{vec}(\eta_A)\|_2 \), note that, for all \( k \in [K] \) and any \( t \in [0, 1] \), by Conditions (C1), (C4) and (C5), we have
\[
\left\| \left\{ \mathbf{V}^{(k)} \left( [\theta^{*(k)} + t(\tilde{\theta}^{(k)} - \theta^{*(k)})] - \tilde{\mathbf{V}}^{(k)} \right) \right\} (\tilde{\theta}^{(k)} - \theta^{*(k)}) \right\|_\infty
\leq \frac{1}{n_k} \sum_{i=1}^{n_k} |x_i^{(k) \top} \mathbf{V}^{(k)} \left( [\theta^{*(k)} + t(\tilde{\theta}^{(k)} - \theta^{*(k)})] - \tilde{\mathbf{V}}^{(k)} \right) x_i^{(k)}| \cdot C_L \left( 1 - t \right) |(\tilde{\theta}^{(k)} - \theta^{*(k)} \top) x_i^{(k)}| \leq \frac{C_L C_x}{n_k} \|X^{(k)} (\tilde{\theta}^{(k)} - \theta^{*(k)})\|_2^2 = O_p(q \log p/n^*).
\]
Thus, we have
\[
\|\text{vec}(\eta_A)\|_2 = \sqrt{\sum_{m=1}^M \|\eta_A^{(m)}\|_2^2} \leq \sqrt{\sum_{m=1}^M (\sqrt{q} \|\eta^{(m)}\|_\infty)^2} = O_p\left( \sqrt{\sum_{m=1}^M (G_{\max} q^{3/2} \log p)^2} N^2 \right) = O_p\left( M^{1/2} G_{\max} q^{3/2} \log p \right).
\]
Combining (A.12)–(A.14), with probability approaching 1, we have
\[
\|\alpha\|_2 \leq O_p(r_{1N}),
\]
(A.15)
where

\[ r_{1N} = \sqrt{\frac{(K/|G_{\min}|)q}{N_{\min}}} + \frac{|G_{\max}|M^{1/2}q^{3/2}\log p}{N_{\min}}. \]

Step 2: Let \( \hat{\psi}^{or} = (\hat{\psi}_{A}^{\top}, 0_{(p-q)\times M})^{\top} \). We show that \( \hat{\psi}^{or} \) is a local minimizer of \( L^G(\psi) \) in \( \Theta \) through verifying the KKT conditions,

\[
\sqrt{\sum_{m=1}^{M} \left( |G^{(m)}|^{1/2} \hat{\psi}^{or(m)} \right)^2} > \tau \lambda_1, \quad \text{in } A,
\]

\[
\frac{\partial L^G(\hat{\psi}^{or})}{\partial \psi_j} + \partial p_\tau \left( \sqrt{\sum_{m=1}^{M} \left( |G^{(m)}|^{1/2} \hat{\psi}^{or(m)} \right)^2}, \lambda_1 \right) / \partial \psi_j = 0, \quad \text{in } A^c,
\]

which are satisfied if the following conditions hold

\[
|G_{\min}|^{1/2} \left| \hat{\psi}^{or} \right|_2 > \tau \lambda_1, \quad \text{in } A, \quad \text{(A.16)}
\]

\[
\left| \frac{\partial L^G(\hat{\psi}^{or})}{\partial \psi_j} \right|_2 \leq \lambda_1 |G_{\min}|^{1/2}, \quad \text{in } A^c. \quad \text{(A.17)}
\]

First, we show that Condition (A.16) is satisfied with probability approaching 1. By the triangle inequality and (A.15), when \( N \) is sufficiently large,

\[
|G_{\min}|^{1/2} \min_{j \in A} ||\hat{\psi}^{or}\|_2 \geq |G_{\min}|^{1/2} \left( \min_{j \in A} ||\psi_j^*\|_2 - \max_{j \in A} ||\hat{\psi}^{or} - \psi_j^*\|_2 \right)
\]

\[
\geq |G_{\min}|^{1/2} \left( \min_{j \in A} ||\psi_j^*\|_2 - ||\alpha\|_2 \right) \geq |G_{\min}|^{1/2}(d_1 - Cr_{1N}) > \tau \lambda_1,
\]

where \( C \) is a constant. The last inequality is satisfied since \( |G_{\min}|^{1/2}d_1 > \tau \lambda_1 \) and \( \lambda_1 \gg |G_{\min}|^{1/2}r_{1N} \). Accordingly, (A.16) is satisfied with probability approaching 1 when \( N \to \infty \).

Second, we show that Condition (A.17) is satisfied with probability approaching 1. Note that \( ||\partial L^G(\hat{\psi}^{or})/\partial \psi_j||_2 = \sqrt{\sum_{m=1}^{M} (\partial L^G(\hat{\psi}^{or})/\partial \psi_j^{(m)})^2} \). Then Condition (A.17) holds if

\[
\left| \frac{\partial L^G(\hat{\psi}^{or})}{\partial \psi_j^{(m)}} \right|_{\infty} \leq \lambda_1 (|G_{\min}|/M)^{1/2}, \quad m \in [M]. \quad \text{(A.18)}
\]
Since for each \( m \in [M] \), \( \partial L^{or,G}(\tilde{\psi}^{or}_A)/\partial \psi^{(m)}_A = 0 \), we have
\[
\tilde{\psi}^{or(m)}_A - \psi^{* (m)}_A = \left( \tilde{V}^{G(m)}_{A,A} \right)^{-1} (\xi^{(m)}_A + \eta^{(m)}_A), \quad m \in [M]. \tag{A.19}
\]

Thus, combining (A.18) and (A.19) leads to
\[
\frac{1}{2} \left\| \frac{\partial L^{G}(\tilde{\psi}^{or})}{\partial \psi^{(m)}_A} \right\|_{\infty} = \left\| \tilde{V}^{G(m)}_{A',A} \left( \tilde{\psi}^{or(m)}_A - \psi^{* (m)}_A \right) + (\xi^{(m)}_A + \eta^{(m)}_A) \right\|_{\infty}
\]
\[
\leq \left\| \tilde{V}^{G(m)}_{A',A} \left( \tilde{V}^{G(m)}_{A,A} \right)^{-1} (\xi^{(m)}_A + \eta^{(m)}_A) \right\|_{\infty} + \left\| (\xi^{(m)}_A + \eta^{(m)}_A) \right\|_{\infty}
\]
\[
\leq \left( \left\| \tilde{V}^{G(m)}_{A',A} \left( \tilde{V}^{G(m)}_{A,A} \right)^{-1} \right\|_{\infty} + 1 \right) \left( \left\| (\xi^{(m)}_A + \eta^{(m)}_A) \right\|_{\infty} \right). \tag{A.20}
\]

Following Xue et al. (2012), we can derive the upper bound of \( \left\| \tilde{V}^{G(m)}_{A',A} \left( \tilde{V}^{G(m)}_{A,A} \right)^{-1} \right\|_{\infty}, m \in [M] \). Following the definition of \( \tilde{V}^{G(m)}(m = 1, \ldots, M) \), we can also define \( V^{*G(m)}(m = 1, \ldots, M) \) accordingly. By Condition (C3), we define
\[
c_m := \left\| \left[ E \left( V^{*G(m)}_A \right) \right]^{-1} \right\|_{\infty} \leq \sqrt{q} \left\| \left[ E \left( V^{*G(m)}_A \right) \right]^{-1} \right\|_2 \leq \frac{\sqrt{q}K}{|G_{\min}|C_{\min}}. \tag{A.21}
\]

Furthermore, we define
\[
\phi_m = \left\| \tilde{V}^{G(m)}_{A',A} \left( \tilde{V}^{G(m)}_{A,A} \right)^{-1} - E \left( V^{*G(m)}_{A',A} \right) \left[ E \left( V^{*G(m)}_A \right) \right]^{-1} \right\|_{\infty},
\]
\[
\phi_{1m} = \left\| \left( \tilde{V}^{G(m)}_{A,A} \right)^{-1} - \left[ E \left( V^{*G(m)}_A \right) \right]^{-1} \right\|_{\infty},
\]
\[
\phi_{2m} = \left\| \tilde{V}^{G(m)}_{A,A} - E \left( V^{*G(m)}_A \right) \right\|_{\infty}, \quad \phi_{3m} = \left\| \tilde{V}^{G(m)}_{A',A} - E \left( V^{*G(m)}_{A',A} \right) \right\|_{\infty}.
\]
Then by definition,

\[
\phi_m = \left\| \left[ \hat{V}^{(m)}_{A,A} - \mathbb{E}(V^{*G(m)}_{A,A}) \right] \left[ \hat{V}^{(m)}_{A,A} - \mathbb{E}(V^{*G(m)}_{A,A}) \right]^{-1} \right\|_{\infty} + \mathbb{E}(V^{*G(m)}_{A,A}) \left[ \hat{V}^{(m)}_{A,A} - \mathbb{E}(V^{*G(m)}_{A,A}) \right]^{-1} \left( \mathbb{E}(V^{*G(m)}_{A,A}) \right)^{-1} \\
+ \mathbb{E}(V^{*G(m)}_{A,A}) \left[ \mathbb{E}(V^{*G(m)}_{A,A}) \right]^{-1} \left[ - \hat{V}^{(m)}_{A,A} + \mathbb{E}(V^{*G(m)}_{A,A}) \right] \left( \hat{V}^{(m)}_{A,A} \right)^{-1} \\
+ \left[ \hat{V}^{(m)}_{A,A} - \mathbb{E}(V^{*G(m)}_{A,A}) \right] \left[ \mathbb{E}(V^{*G(m)}_{A,A}) \right]^{-1} \left\|_{\infty}
\]

\[
\leq \phi_{3m} \phi_{1m} + \varphi^{G(m)} \phi_{2m} \left\| \left( \hat{V}^{(m)}_{A,A} \right)^{-1} \right\|_{\infty} + \phi_{3m} c_m
\]

\[
\leq \phi_{3m} \phi_{1m} + \varphi^{G(m)} \phi_{2m} (c_m + \phi_{1m}) + \phi_{3m} c_m.
\]

Besides, \( \phi_{1m} \) can be reformulated as

\[
\phi_{1m} = \left\| \left( \hat{V}^{(m)}_{A,A} \right)^{-1} \left[ \mathbb{E}(V^{*G(m)}_{A,A}) - \hat{V}^{(m)}_{A,A} \right] \left[ \mathbb{E}(V^{*G(m)}_{A,A}) \right]^{-1} \right\|_{\infty}
\]

\[
\leq \left\| \left( \hat{V}^{(m)}_{A,A} \right)^{-1} \right\|_{\infty} \cdot \left\| \mathbb{E}(V^{*G(m)}_{A,A}) - \hat{V}^{(m)}_{A,A} \right\|_{\infty} \cdot \left\| \left[ \mathbb{E}(V^{*G(m)}_{A,A}) \right]^{-1} \right\|_{\infty}
\]

\[
\leq (c_m + \phi_{1m}) \phi_{2m} c_m.
\]

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

\[
\phi_{m} \leq (\phi_{3m} + \varphi^{G(m)} \phi_{2m}) \frac{c_m}{1 - \phi_{2m} c_m}.
\]

(A.22)

Then, by Lemma \[2\]

\[
\phi_{2m} = \left\| \hat{V}^{(m)}_{A,A} - \mathbb{E}(V^{*G(m)}_{A,A}) \right\|_{\infty} \leq \frac{q \sum_{k \in G^{(m)}} n_k \left| \hat{V}^{(k)}_{A,A} - \mathbb{E}(V^{*G(m)}_{A,A}) \right|_{\max}}{N}
\]

\[
= O_p \left( \frac{|G_{\text{max}}|}{K} \cdot \sqrt{\frac{q^4 \log p}{n^*}} \right).
\]

(A.23)

Combining (A.21), (A.23), and Condition (C7), we have

\[
\phi_{2m} c_m = O_p \left( \frac{|G_{\text{max}}|}{|G_{\text{min}}|} \cdot \sqrt{\frac{q^4 \log p}{n^*}} \right) = o_p(1).
\]

(A.24)

Following the proof of (A.24), we also have

\[
\phi_{3m} c_m = O_p \left( \frac{|G_{\text{max}}|}{|G_{\text{min}}|} \cdot \sqrt{\frac{q^4 \log p}{n^*}} \right) = o_p(1).
\]

(A.25)
Combining (A.22), (A.24) and (A.25), with probability approaching 1, for all \( m \in [M] \)
\[
\phi_m = o_p(\varphi_{\text{max}}),
\]
\[
\max_{m \in [M]} \left\| \nabla G^{(m)} \left( \nabla \hat{g}_{\hat{A}} \right)^{-1} \right\|_\infty \leq \max_{m \in [M]} \phi_m + \max_{m \in [M]} \mathbb{E} \left( \nabla^{*} G^{(m)} \left[ \mathbb{E} \left( \nabla^{*} \hat{g}_{\hat{A}} \right) \right]^{-1} \right) \leq 2\varphi_{\text{max}}.
\]  
(A.26)

Now we focus on the term \( ||\xi^{(m)}||_\infty, m \in [M] \). By Conditions (C1), (C2), Lemma 1, and the union bound
\[
P \left( \max_{m \in [M]} \max_{j \in [p]} \left| \sum_{k \in G^{(m)}} n_k g^{(k)}_j \right| \geq t\sqrt{N_{\text{max}}} \log p \right) \leq 2pM \exp \left( -\frac{C_1 t^2 \log p}{C_2^2 \kappa_x^2} \right) \leq 2p^{2-C_3},
\]
where \( C_3 = C_1 t^2 / (C_2^2 \kappa_x^2) \). When \( t \) is sufficiently large and \( 2 - C_3 < 0, 2p^{2-C_3} \to 0 \) when \( p \to \infty \). Hence,
\[
\max_{m \in [M]} ||\xi^{(m)}||_\infty = O_p \left( \sqrt{\frac{|G_{\text{max}}| \log p}{KN}} \right).
\]  
(A.27)

Following the proof of (A.14), we have
\[
\max_{m \in [M]} ||\eta^{(m)}||_\infty = O_p \left( \frac{|G_{\text{max}}| q \log p}{N} \right).
\]  
(A.28)

Combining (A.20), (A.26), (A.27), (A.28), since \( \varphi_{\text{max}} r_{2N} \ll \lambda_1 \), where
\[
r_{2N} = \sqrt{\left( \frac{|G_{\text{max}}| / |G_{\text{min}}|}{KN} \right) M \log p + \left( \frac{|G_{\text{max}}| / |G_{\min}|^{1/2}}{N} \right) M^{1/2} q \log p},
\]
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.

**Proof of Theorem 2**

Define
\[
Q(\theta) = \mathcal{L}(\theta) + \mathcal{P}_1(\theta) + \mathcal{P}_2(\theta), \quad Q^G(\psi) = \mathcal{L}^G(\psi) + \mathcal{P}^G_1(\psi) + \mathcal{P}^G_2(\psi),
\]

40
where

\[
\mathcal{L}(\theta) = \frac{1}{N} \sum_{k=1}^{K} n_k \left( (\theta^{(k)} \top \bar{V}^{(k)} \theta^{(k)}) - 2 \theta^{(k)} \top \bar{Z}^{(k)} \right) + \sum_{j=2}^{p} p_j \left( ||\theta_j||_2, \lambda_1 \right),
\]

\[
\mathcal{P}_2(\theta) = \sum_{k<k'} p_j \left( ||\theta^{(k)} - \theta^{(k')}||_2, \lambda_2 \right),
\]

\[
\mathcal{L}^G(\psi) = \frac{1}{N} \sum_{m=1}^{M} \left[ \psi^{(m)} \top \left( \sum_{k \in G^{(m)}} n_k \bar{V}^{(k)} \right) \psi^{(m)} - 2 \psi^{(m)} \top \left( \sum_{k \in G^{(m)}} n_k \bar{Z}^{(k)} \right) \right] + \sum_{j=2}^{p} p_j \left( \left[ \sum_{m=1}^{M} |G^{(m)}| \psi_j^{(m)} |^2 \right]^{1/2}, \lambda_1 \right),
\]

\[
\mathcal{P}_2^G(\psi) = \sum_{m<m'} |G^{(m)}| \left| G^{(m')} \right| \left[ p_j \left( \left| \psi^{(m)} - \psi^{(m')} \right| ||_2, \lambda_2 \right) \right].
\]

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

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

\[
Q(\theta) = Q^G(T^G(\theta)), \quad Q^G(\psi) = Q(T^{-1}(\psi)). \quad \text{(A.29)}
\]

Consider the neighborhood of \( \theta^* \), denoted by \( \Theta \), which is defined as

\[
\Theta_1 = \left\{ \theta \in \mathbb{R}^{p \times K} : ||\theta - \theta^*||_F \leq C|G_{\max}|^{1/2}r_{1N} \right\}.
\]

Define the event

\[
E_1 = \left\{ ||\hat{\theta}^{or} - \theta^*||_F \leq C|G_{\max}|^{1/2}r_{1N} \right\}.
\]

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

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

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

(a) On event \( E_1 \), \( Q(\tilde{\theta}^G) > Q(\tilde{\theta}^{or}) \) for any \( \theta \in \Theta_1 \) and \( \theta^G \neq \tilde{\theta}^{or} \);

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

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

$$
\mathcal{P}_{T}(\mathcal{T}(\tilde{\theta})) = \sum_{m < m'} |G^{(m)}||G^{(m')}|p_T \left( \left\| \tilde{\psi}^{(m)} - \tilde{\psi}^{(m')} \right\|_2, \lambda_2 \right),
$$

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

$$
\left\| \tilde{\psi}^{(m)} - \tilde{\psi}^{(m')} \right\|_2 \geq \min_{m, m' \in [M], m \neq m'} \left\| \psi^{*}(m) - \psi^{*}(m') \right\|_2 - 2 \cdot \max_{m \in [M]} \left\| \tilde{\psi}^{(m)} - \psi^{*}(m) \right\|_2
\geq d_2 - 2 \cdot \max_{m \in [M]} \left\| G^{(m)} \right\|^{-1} \sum_{k \in G^{(m)}} (\tilde{\theta}^{(k)} - \theta^{*}(k)) \right\|_2
\geq d_2 - 2 \cdot \max_{k \in [K]} \left\| \tilde{\theta}^{(k)} - \theta^{*}(k) \right\|_2 \geq d_2 - 2C |G_{max}|^{1/2} r_{1N} > \tau \lambda_2,
$$

where \( C \) is a constant and the last inequality follows from \( d_2 > \tau \lambda_2 \) and \( \lambda_2 \gg |G_{max}|^{1/2} r_{1N} \).

Consequently, for any \( \theta \in \Theta_1 \), \( \mathcal{P}_{T}(\mathcal{T}(\tilde{\theta})) = C_N \), and \( C_N > 0 \) is a constant. By the result of Theorem 1, \( \tilde{\psi}^{or} \) is the local minimizer of objective function \( \mathcal{L}^G(\psi) \) with probability approaching 1. Thus we have \( \mathcal{L}^G(\tilde{\psi}^{or}) < \mathcal{L}^G(\mathcal{T}(\tilde{\theta})) \) for any \( \theta \in \Theta_1 \) and \( \mathcal{T}(\tilde{\theta}) \neq \tilde{\psi}^{or} \). Combining this with \( \mathcal{P}_{T}(\mathcal{T}(\tilde{\theta})) = C_N \) for \( \theta \in \Theta_1 \), we have \( \mathcal{Q}^G(\tilde{\psi}^{or}) < \mathcal{Q}^G(\mathcal{T}(\tilde{\theta})) \). By (A.29), we have

$$
\mathcal{Q}^G(\tilde{\psi}^{or}) = \mathcal{Q}(\mathcal{T}^{-1}(\tilde{\psi}^{or})) = \mathcal{Q}(\tilde{\theta}^{or}), \quad \mathcal{Q}^G(\mathcal{T}(\tilde{\theta})) = \mathcal{Q}(\mathcal{T}^{-1}(\mathcal{T}(\tilde{\theta}))) = \mathcal{Q}(\tilde{\theta}^G).
$$
Accordingly, we have $Q(\hat{\theta}^{or}) < Q(\tilde{\theta}^v)$ for any $\theta \in \Theta_1$ and $\tilde{\theta}^v \neq \hat{\theta}^{or}$. 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 $E_1$, $Q(\theta) \geq Q(\tilde{\theta})$ for any $\theta \in \Theta_1 \cap \Theta_2$. By the Taylor series expansion, we have

$$Q(\theta) - Q(\tilde{\theta}) = \Omega_{11} + \Omega_{12} + \Omega_{13}, \quad (A.31)$$

where

$$\Omega_{11} = \sum_{j=1}^p \frac{\partial L(\bar{\theta})}{\partial \theta_j} (\theta_j - \tilde{\theta}_j), \quad \Omega_{12} = \sum_{j=2}^p \frac{\partial P_1(\bar{\theta})}{\partial \theta_j} (\theta_j - \tilde{\theta}_j),$$

$$\Omega_{13} = \sum_{k<k'} \left[ p_r(||\theta^{(k)} - \theta^{(k')}||_2, \lambda_2) - p_r(||\tilde{\theta}^{(k)} - \tilde{\theta}^{(k')}||_2, \lambda_2) \right],$$

in which $\bar{\theta} = \delta_1 \theta + (1 - \delta_1) \tilde{\theta}$ for some $\delta_1 \in (0, 1)$. Note that, for any $j \in A$, $\theta_j = \tilde{\theta}_j$, and for any $j \in A^c$, $\tilde{\theta}_j = 0$ and $\bar{\theta}_j = \delta_1 \theta_j$. Since $p_r(t, \lambda_2)$ is a nondecreasing function of $t$ with $t \in [0, \infty)$, $\Omega_{13} \geq 0$. Besides,

$$\Omega_{12} = \sum_{j \in A^c} p'_r(||\bar{\theta}_j||_2, \lambda_1) \frac{\tilde{\theta}_j^T (\theta_j - \tilde{\theta}_j)}{||\bar{\theta}_j||_2} = \sum_{j \in A^c} p'_r(\delta_1 ||\theta_j||_2, \lambda_1) ||\theta_j||_2 \geq \sum_{j \in A^c} p'_r(t_N, \lambda_1) ||\theta_j||_2, \quad (A.32)$$

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

$$|\Omega_{11}| = \left| \sum_{j \in A^c} \frac{\partial L(\bar{\theta})}{\partial \theta_j} (\theta_j - \tilde{\theta}_j) \right| \leq \sum_{j \in A^c} \left| \frac{\partial L(\bar{\theta})}{\partial \theta_j} \right|_2 ||\theta_j||_2 \leq \sqrt{K} \max_{k \in [K]} \left| \frac{\partial L(\bar{\theta})}{\partial \theta_k} \right|_\infty \sum_{j \in A^c} ||\theta_j||_2, \quad (A.33)$$

Combining $(A.31)$, $(A.32)$, and $(A.33)$, we have

$$Q(\theta) - Q(\tilde{\theta}) = \Omega_{11} + \Omega_{12} + \Omega_{13} \geq \sum_{j \in A^c} \left[ p'_r(t_N, \lambda_1) - \sqrt{K} \max_{k \in [K]} \left| \frac{\partial L(\bar{\theta})}{\partial \theta_k} \right|_\infty \right] ||\theta_j||_2. \quad (A.34)$$
Following the proof from (A.19) to (A.28), we have
\[
\sqrt{K} \max_{k \in K} \left| \frac{\partial \mathcal{L}(\bar{\theta})}{\partial \theta^{(k)}_A} \right|_\infty = O_p \left( \varphi_{\max} \left[ \sqrt{\log p/N} + K^{1/2} q \log p/N \right] + (Kq)^{1/2} t_N \right).
\]
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
\[
\lambda_1 \gg \varphi_{\max} \left[ \sqrt{\log p/N} + K^{1/2} q \log p/N \right] + (Kq)^{1/2} t_N.
\]
(A.35)

Then, by (A.34) and (A.35), when \( N \) is sufficiently large, with probability approaching 1,
\[
\mathcal{Q}(\theta) - \mathcal{Q}(\bar{\theta}) \geq 0.
\]

Next, we show that, on event \( E_1 \), \( \mathcal{Q}(\bar{\theta}) \geq \mathcal{Q}(\bar{\theta}^\mathcal{G}) \) for any \( \theta \in \Theta_1 \cap \Theta_2 \). By the Taylor series expansion, we have
\[
\mathcal{Q}(\bar{\theta}) - \mathcal{Q}(\bar{\theta}^\mathcal{G}) = \Omega_{21} + \Omega_{22} + \Omega_{23},
\]
where
\[
\Omega_{21} = \sum_{k=1}^K \frac{\partial \mathcal{L}(\bar{\theta}^h)}{\partial \theta^{(k)}_A} (\bar{\theta}^{(k)} - \bar{\theta}^\mathcal{G}(k)), \quad \Omega_{22} = \sum_{j=2}^p \frac{\partial \mathcal{P}_1(\bar{\theta}^h)}{\partial \theta^j} (\bar{\theta}^j - \bar{\theta}^\mathcal{G}_j),
\]
\[
\Omega_{23} = \sum_{k=1}^K \frac{\partial \mathcal{P}_2(\bar{\theta}^h)}{\partial \theta^{(k)}_A} (\bar{\theta}^{(k)} - \bar{\theta}^\mathcal{G}(k)),
\]
in which \( \bar{\theta}^h = \delta_2 \bar{\theta} + (1 - \delta_2) \bar{\theta}^\mathcal{G} \) for some \( \delta_2 \in (0, 1) \). Next, we bound \( \Omega_{21}, \Omega_{22}, \) and \( \Omega_{23}. \)

Recall that, for any \( j \in \mathcal{A}^c, \bar{\theta}^{h}_{A^c} = \bar{\theta}^\mathcal{G}_{A^c} = 0. \) Then,
\[
\Omega_{22} = \sum_{j=2}^q \frac{\partial \mathcal{P}_1(\bar{\theta}^h)}{\partial \theta^j} (\bar{\theta}^j - \bar{\theta}^\mathcal{G}_j) = \sum_{j=2}^q p'_\tau(\|\bar{\theta}^h_j\|_2, \lambda_1) \frac{\bar{\theta}^{h,\top}_j (\bar{\theta}^j - \bar{\theta}^\mathcal{G}_j)}{\|\bar{\theta}^h_j\|_2}.
\]
Note that,

\[ ||\hat{\theta}_j^g - \theta_j^*||_2 = \sqrt{\sum_{m=1}^{M} |G^{(m)}| \left( \frac{\sum_{k \in G^{(m)}} (\hat{\theta}_j^{(k)} - \psi_j^{*(m)})}{|G^{(m)}|} \right)^2} \]

\[ \leq \sqrt{\sum_{m=1}^{M} |G^{(m)}| \cdot \frac{|G^{(m)}| \left( \sum_{k \in G^{(m)}} (\hat{\theta}_j^{(k)} - \psi_j^{*(m)}) \right)^2}{|G^{(m)}|^2}} = ||\hat{\theta}_j - \theta_j^*||_2. \]

Besides, since \( \hat{\theta}_j^h = \delta_2 \hat{\theta}_j + (1 - \delta_2) \hat{\theta}_j^g \),

\[ ||\hat{\theta}_j^h - \theta_j^*||_2 = \delta_2 ||\hat{\theta}_j - \theta_j^*||_2 + (1 - \delta_2)||\hat{\theta}_j^g - \theta_j^*||_2 \leq ||\hat{\theta}_j - \theta_j^*||_2. \]

Hence, for any \( j \in A \), by the triangle inequality,

\[ ||\hat{\theta}_j^h||_2 \geq ||\theta_j^*||_2 - ||\hat{\theta}_j - \theta_j^*||_2 \geq \min_{j \in A} ||\theta_j^*||_2 - \max_{j \in A} ||\hat{\theta}_j - \theta_j^*||_2 \]

\[ \geq |G_{\min}|^{1/2} d_1 - C |G_{\max}|^{1/2} r_1 N > \tau \lambda_1, \tag{A.37} \]

where the last inequality follows from \( |G_{\min}|^{1/2} d_1 > \tau \lambda_1 \gg |G_{\max}|^{1/2} r_1 N \). Combining (A.36) and (A.37), since \( p_r(t, \lambda_1) \) is a constant for \( t \geq \tau \lambda_1 \), we have \( \Omega_{22} = 0 \).

Now consider \( \Omega_{23} \). Recall that, for any \( j \in A^c \), \( \hat{\theta}_j^{A^c} = \hat{\theta}_j^c = \hat{\theta}_j^g = 0 \). Then,

\[ \Omega_{23} = \sum_{k < k'} \left\{ p_r^{(1)} \left( ||\hat{\theta}_j^{A^c} - \theta_j^{A^c}||_2, \lambda_2 \right) ||\hat{\theta}_j^{A^c} - \theta_j^{A^c}||_2^{-1} \right\} \]

\[ \times \left( \hat{\theta}_j^{A^c} - \theta_j^{A^c} \right)^\top \left[ \left( \hat{\theta}_j^{A^c} - \theta_j^{A^c} \right) - \left( \hat{\theta}_j^{A^c} - \theta_j^{A^c} \right) \right] \}

\[ = \sum_{m=1}^{M} \sum_{k, k' \in G^{(m)}, k < k'} \left\{ p_r^{(2)} \left( ||\hat{\theta}_j^{A^c} - \theta_j^{A^c}||_2, \lambda_2 \right) ||\hat{\theta}_j^{A^c} - \theta_j^{A^c}||_2^{-1} \right\} \]

\[ \times \left( \hat{\theta}_j^{A^c} - \theta_j^{A^c} \right)^\top \left[ \left( \hat{\theta}_j^{A^c} - \theta_j^{A^c} \right) - \left( \hat{\theta}_j^{A^c} - \theta_j^{A^c} \right) \right] \tag{A.38} \]

where the first term of the second equality follows from the fact that, when \( k, k' \in G^{(m)} \),
\[ \theta^G_A = \theta^G_{A'} \] and \[ \theta^h_A - \theta^h_{A'} = \delta_2(\theta^h_A - \theta^h_{A'}) \]. Note that, for any \( m \in [M], k \in G(m) \),

\[ ||\theta^G_A - \theta^g_A||_2 = \left| \left| \sum_{k \in G(m)} \frac{\theta^h_A}{|G(m)|} - \psi^g_A \right| \right| \leq \max_{k \in [K]} ||\theta^h_A - \theta^g_A||_2. \]

And then for any \( k \in [K] \), we have \[ ||\theta^h_A - \theta^g_A||_2 \leq \max_{k \in [K]} ||\theta^h_A - \theta^g_A||_2. \] Hence, similar to (A.30), for any \( m < m', k \in G(m), k' \in G(m') \),

\[ ||\theta^h_A - \theta^h_{A'}||_2 \geq \min_{k \in G(m), k' \in G(m')} ||\theta^g_A - \theta^g_{A'}||_2 - 2 \max_{k \in [K]} ||\theta^h_A - \theta^g_A||_2 \]

\[ \geq \min_{k \in G(m), k' \in G(m')} ||\theta^g_A - \theta^g_{A'}||_2 - 2 \max_{k \in [K]} ||\theta^h_A - \theta^g_A||_2 \quad (A.39) \]

\[ \geq d_2 - 2C|G_{max}|^{1/2}r_{1N} > \tau \lambda_2. \]

Combining (A.38) and (A.39), since \[ ||\theta^h_A - \theta^h_{A'}||_2 \leq 2 \max_{k \in [K]} ||\theta^h_A - \theta^g_A||_2 \leq \]

\[ 2C|G_{max}|^{1/2}r_{1N}, \] we have

\[ \Omega_{23} = \sum_{m=1}^{M} \sum_{k,k' \in G(m), k < k'} p^f \left( ||\theta^h_A - \theta^h_{A'}||_2, \lambda_2 \right) \left| \theta^h_A - \theta^h_{A'} \right|_2 \]

\[ \geq \sum_{m=1}^{M} \sum_{k,k' \in G(m), k < k'} p^f \left( 2C|G_{max}|^{1/2}r_{1N}, \lambda_2 \right) \left| \theta^h_A - \theta^h_{A'} \right|_2 \quad (A.40) \]

\[ \geq \sum_{m=1}^{M} \sum_{k,k' \in G(m), k < k'} \frac{\lambda_2}{2} \left| \theta^h_A - \theta^h_{A'} \right|_2, \]

where the last inequality follows from \( 2C|G_{max}|^{1/2}r_{1N} \rightarrow 0 \) when \( N \rightarrow \infty \).

We now consider the bound of \( \Omega_{21} \). For \( k \in [K] \), we define

\[ \omega^{(k)} := \frac{\partial L(\theta^h_A)}{\partial \theta^h_A} = 2\left\{ \left[(n_k/N)^2 V_A^{(k)} \right] \theta^h_A - (n_k/N) \zeta^{(k)}_A \right\} \]

\[ = 2\left\{ \left[(n_k/N)^2 V_A^{(k)} \right] \theta^h_A - \theta^g_A \right\} + (n_k/N) \gamma^{(k)}_A \]

\[ + (n_k/N) \int_0^1 \left\{ V^{(k)} \left[ \theta^g_A + t(\theta^h_A - \theta^g_A) \right] - V^{(k)} \right\} \left( \theta^h_A - \theta^g_A \right) \, dt \]

\[ := \omega^{(k)}_1 + \omega^{(k)}_2 + \omega^{(k)}_3. \]
Then,

\[
\Omega_{21} = \sum_{k=1}^{K} \omega^{(k)} \mathsf{T} (\hat{\theta}_A - \hat{\theta}_{G(k)}^{(k)}) = \sum_{m=1}^{M} \sum_{k,k' \in \mathcal{G}(m)} \omega^{(k)} \mathsf{T} \frac{(\hat{\theta}_A - \hat{\theta}_{G(k)}^{(k)})}{|\mathcal{G}(m)|}
\]

\[
= \sum_{m=1}^{M} \sum_{k,k' \in \mathcal{G}(m)} \omega^{(k)} \mathsf{T} \frac{(\hat{\theta}_A - \hat{\theta}_{G(k)}^{(k)})}{2|\mathcal{G}(m)|} + \sum_{m=1}^{M} \sum_{k,k' \in \mathcal{G}(m)} \omega^{(k)} \mathsf{T} \frac{(\hat{\theta}_A - \hat{\theta}_{G(k)}^{(k')})}{2|\mathcal{G}(m)|}
\]

\[
= \sum_{m=1}^{M} \sum_{k,k' \in \mathcal{G}(m)} \omega^{(k)} \mathsf{T} \frac{(\hat{\theta}_A - \hat{\theta}_{G(k)}^{(k')})}{2|\mathcal{G}(m)|}.
\]

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

\[
\max_{k \in [K]} \|\omega^{(k)}\|_2 = O_p \left( \frac{|\mathcal{G}_{\max}|^{1/2} r_1 N}{K} \right), \quad \max_{k \in [K]} \|\omega^{(k)}_2\|_2 = O_p \left( \sqrt{q/(KN)} \right),
\]

\[
\max_{k \in [K]} \|\omega^{(k)}_1\|_2 = O_p \left( q^{3/2} \log p / N \right).
\]

Then,

\[
|\Omega_{21}| \leq \sum_{m=1}^{M} \sum_{k,k' \in \mathcal{G}(m), k < k'} \frac{2 \max_{k \in [K]} \|\omega^{(k)}\|_2}{|\mathcal{G}_{\min}|} \|\hat{\theta}_A^{(k)} - \hat{\theta}_A^{(k')}\|_2.
\]

Combining (A.40) and (A.41), since \(\lambda_2 \gg |\mathcal{G}_{\max}|^{1/2} r_1 N\), we have

\[
\lambda_2 \gg \frac{|\mathcal{G}_{\max}|^{1/2} r_1 N}{K |\mathcal{G}_{\min}|} + \sqrt{\frac{q}{KN |\mathcal{G}_{\min}|^2}} + \frac{q^{3/2} \log p}{N |\mathcal{G}_{\min}|},
\]

which leads to

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

for a sufficiently large \(N\) with probability approaching 1. Thus, we have proved the result in (b). This finishes all the proof of Theorem 2.
## Appendix B

This section contains additional simulation results and data application results.

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

| Method  | n = 200 | n = 400 | n = 800 |
|---------|---------|---------|---------|
|         | 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) | (0.003) | (6.400) | (0.000) | (0.005) | (30.463) |
| SHIR    | 0.760   | 0.001   | 389.220 | 1.000   | 0.001   | 513.330 | 1.000   | 0.003   | 528.000 |
|         | (0.042) | (0.004) | (21.718) | (0.000) | (0.003) | (9.002) | (0.000) | (0.005) | (27.852) |
| 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| 1.000   | 0.198   | 1673.740| 1.000   | 0.218   | 1794.750|
|         | (0.020) | (0.010) | (63.847) | (0.005) | (0.010) | (9.002) | (0.000) | (0.011) | (63.438) |
| SK(har) | 0.945   | 0.526   | 222.880 | 1.000   | 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   | 1.000   | 200.000 |
|         | (0.000) | (0.000) | (1.518) | (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) |
| 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) | (7.335)  |
| 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) | (7.335)  |
| Local   | 0.851   | 0.147   | 2602.260| 1.000   | 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   | 0.000   | 216.400 | 1.000   | 0.004   | 1067.530| 1.000   | 0.011   | 1152.000|
|         | (0.121) | (0.000) | (189.078)| (0.000) | (0.000) | (38.757)| (0.000) | (0.000) | (85.526) |
| SK(gap) | 2.000   | 0.000   | 0.746   | 0.487   | 2.000   | 0.000   | 0.748   | 0.494   | 3.030   |
|         | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (1.000) | (0.000) | (1.000) |

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

| Method  | n = 200 | n = 400 | n = 800 |
|---------|---------|---------|---------|
|         |  |         |         |         |         |         |         |         |
|         |  |         |         |         |         |         |         |         |
|         |  |         |         |         |         |         |         |         |

48
Figure 3: Simulation Example 2: boxplot of RMSE.

Table 6: Simulation Example 3, variable selection accuracy: mean (sd) based on 100 replicates.
Table 7: Simulation Example 3, clustering accuracy: mean (sd) based on 100 replicates.

| Method | $\hat{M}$  | Per | RI | ARI | $\hat{M}$  | Per | RI | ARI | $\hat{M}$  | Per | RI | ARI |
|--------|------------|-----|----|-----|------------|-----|----|-----|------------|-----|----|-----|
| $\sigma = 1$ | ICR | 2.000 | 1.000 | 1.000 | 2.000 | 1.000 | 1.000 | 2.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (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) | (0.000) | (0.088) | (0.182) | (0.000) | (0.088) | (0.182) | (0.000) | (0.088) | (0.182) | (0.000) |
| $\sigma = 2$ | ICR | 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) | (0.007) | (0.008) | (0.017) | (0.000) | (0.007) | (0.008) | (0.017) | (0.000) | (0.007) | (0.008) | (0.017) | (0.000) |
| | 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) | (0.007) | (0.007) | (0.017) | (0.000) | (0.007) | (0.007) | (0.017) | (0.000) | (0.007) | (0.007) | (0.017) | (0.000) |

Figure 4: Simulation Example 3: boxplot of RMSE.
Table 8: Simulation Example 4, variable selection accuracy: mean (sd) based on 100 replicates.

| Method  | n = 100 | n = 200 | n = 400 |
|---------|---------|---------|---------|
| $\sigma = 1$ | | | |
| ICR | 1.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) |
| SHIR | 1.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) |
| SMA | 1.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) |
| Local | 0.989 (0.005) | 1.000 (0.000) | 1.000 (0.000) |
| SK(har) | 0.999 (0.013) | 1.000 (0.000) | 1.000 (0.000) |
| SK(gap) | 1.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) |
| $\sigma = 2$ | | | |
| ICR | 1.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) |
| SHIR | 1.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) |
| SMA | 1.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) |
| Local | 0.719 (0.030) | 0.944 (0.014) | 1.000 (0.000) |
| SK(har) | 0.908 (0.106) | 0.998 (0.000) | 1.000 (0.000) |
| SK(gap) | 2.040 (0.281) | 3.140 (0.140) | 3.890 (0.170) |

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

| Method  | $\hat{M}$ | Per | RI | ARI | $\hat{M}$ | Per | RI | ARI | $\hat{M}$ | Per | RI | ARI |
|---------|-----------|-----|----|----|-----------|-----|----|----|-----------|-----|----|----|
| $\sigma = 1$ | | | | | | | | | | | | | |
| ICR | 4.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | 4.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | 4.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | |
| SK(har) | 4.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | 4.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | 4.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | |
| SK(gap) | 2.040 (0.281) | 0.751 (0.036) | 0.948 (0.072) | 3.140 (0.140) | 0.891 (0.126) | 0.780 (0.255) | 3.350 (0.170) | 0.917 (0.119) | 0.833 (0.240) | |
| $\sigma = 2$ | | | | | | | | | | | | | |
| ICR | 4.050 (0.219) | 1.000 (0.002) | 0.999 (0.005) | 4.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | 4.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | |
| SK(har) | 4.150 (0.386) | 0.977 (0.021) | 0.935 (0.058) | 4.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | 4.000 (0.000) | 1.000 (0.000) | 1.000 (0.000) | |
| SK(gap) | 2.040 (0.000) | 0.745 (0.010) | 0.485 (0.000) | 2.000 (0.000) | 0.746 (0.000) | 0.488 (0.000) | 2.340 (0.075) | 0.789 (0.096) | 0.575 (0.193) | |

Table 10: Data analysis: one record of the initial request logs.

| URL interface | GET Parameter | POST Parameter |
|--------------|---------------|----------------|
| ajaxNoSessionGetSmsAction | s=captcha | _method=_construct&filter[]=phpinfo&method=get&server[REQUEST_METHOD]=1 |

51
Figure 5: Simulation Example 4: boxplot of RMSE.

Figure 6: Data analysis: abnormal proportions among the 47 URL interfaces.
Figure 7: Data analysis: Boxplots of (a) AUC and (b) Brier Score based on 100 random splits.