http://pubsonline.informs.org/journal/ijoc

INFORMS JOURNAL ON COMPUTING
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0899-1499, EISSN 1526-5528

Submitted to *INFORMS Journal on Computing*

Truncated fusion learning on supervised clustering and its fast stagewise algorithm

Letian Li, Yang Li*, Jie Zhang, Zemin Zheng*
International Institute of Finance, The School of Management, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China

*Corresponding author

Contact: lilt@mail.ustc.edu.cn, tjly@ustc.edu.cn, zhangj94@ustc.edu.cn, zhengzm@ustc.edu.cn

Abstract. Supervised clustering has emerged to be a popular topic for studying heterogeneous effects in diverse areas such as risk management and medical science. In this paper, we introduce a general heterogeneity tracking model that accommodates various common paradigms for supervised clustering. Building on this new model, we propose a novel method called truncated fusion learning (TRUE) for conducting clustering, which involves a nonconvex fusion penalized optimization. For implementation, we innovate a thresholded fusion stagewise algorithm that allows for quickly tracing out the entire solution paths of TRUE. Under mild regularity conditions, we provide comprehensive theoretical guarantees including the convergence of the proposed algorithm, as well as consistency in parameter estimation, cluster recovery, and model selection. The superior performance of our method is demonstrated by several simulation examples and applications to two real datasets.

Key words: Fusion learning, Supervised clustering, Truncated L_1 penalty, Stagewise learning, Dual problem, Oracle properties

1. Introduction

Supervised clustering is currently a hot topic that has received wide concerns in diverse research areas, ranging from commercial marketing and risk management to medical science and bioinformatics (Chen et al. 2021, Wang et al. 2023). This increased interest is driven by the widespread heterogeneity found in contemporary datasets, especially with the emergence of big data. One natural application using supervised clustering to identify heterogeneity patterns is subgroup analysis, where individuals being studied would be grouped based on the explored heterogeneous effects. For instance, in insurance management, it is crucial to investigate the impacts of risk factors on insurance claim frequency, while these impacts often vary among customer groups within the large customer base. Consequently, different customers should be clustered based on heterogeneity information and provided with their own personalized insurance plans (Chen et al. 2019).

Moreover, supervised clustering has also been increasingly applied to various other popular big data problems that were previously considered primarily under homogeneous backgrounds but suffer from inherent heterogeneity. For example, in online learning, the streams of data are sometimes collected from multi-mode sources with different associations between state variable, action and outcomes (Chen et al. 2024). Results based on integrating over the entire population can be misleading, sometimes even dangerous in high-risk applications such as health care and autonomous driving. Similarly, in federated learning, data distributions among clients are also often distinct. As a result, many of recent studies have shifted towards clustered federated learning (Sattler et al. 2020, Ghosh et al. 2020, Liu et al. 2024) by establishing multiple personalized models instead of a single global shared model by clustering clients.

However, current approaches to supervised clustering typically encounter one or more of the following challenges. First, most existing work narrowly targets one specific model in a certain area, such as the heterogeneous Cox model in survival analysis (Hu et al. 2021). It's not straightforward or even difficult to extend these methods to other different models or application scenarios. Second, although many commonly used techniques for supervised clustering possess appealing theoretical properties, their nice performance typically relies on the unacquirable prior knowledge, such as the true number of clusters or parametric functions of the cluster boundaries (Chen et al. 2021). This poses challenges in practical implementation due to the inaccessibility of such prior information. Third, many existing methods involve complex nonconvex optimization problems, which could be computationally intensive or sometimes intractable, especially in the big data settings.

To tackle the above issues, in this paper we establish a broad fusion learning framework for effective supervised clustering. To be specific, we first introduce a general heterogeneity tracking model which accommodates various common models, including heterogeneous generalized linear models, heterogeneous survival analysis models, and so on. Building on this model, we propose a new method called truncated fusion learning (TRUE), which can automatically and accurately identify the latent clusters and estimate the heterogeneous coefficients without prior information on the true clusters number or clusters boundaries. Furthermore, to solve the associated truncated fusion regularization problem, we also tailor a thresholded fusion stagewise algorithm for quickly tracing out the entire solution paths of TURE. It inherits the merits of canonical stagewise algorithm and enjoys significant computational advantages by searching for the steepest coordinate descent direction in each iteration.

1.1. Main contributions

The main contributions of this article are threefold. First, our heterogeneity tracking model is set up with no specification of underlying distribution or data type allowing it to accommodate diverse application scenarios and thus flexibly contribute to many data science studies. Some application examples and managerial implications are given in Section A of the Supplementary Material. The heterogeneity in our proposed model is driven by individualized effects of predictors on responses, and thus represented by using subject-specific coefficients. Based on this model, we introduce a new method TRUE that subtly applies the truncated L_1 penalty to pairwise differences of subject-specific coefficients to achieve adaptive clustering without the need to know the true number of clusters. The proposed truncated L_1 fusion regularization represents a pioneering effort in the problems of supervised clustering, as its local convexity and truncation enable us to achieve both computational advantages and sharp statistical properties.

Second, to the best of our knowledge, the proposed thresholded fusion stagewise learning strategy is among the first works to develop a pathwise algorithm for supervised clustering. Existing work on stagewise learning focus primarily on standard regularization problems (Efron et al. 2004, Chen et al. 2022), which apparently differ from the fusion regularized regression. In addition, our stagewise algorithm is one of the few that enables stagewise learning for nonconvex optimization problems. It addresses the nonconvex problem through piecewisely optimizing convex subproblems with non-fully connected L_1 fusion penalties constructed by dropping redundant penalty terms, offering new insights for other similar nonconvex problems that contain locally convex domains. It should be emphasized that dropping redundant penalty terms is significantly helpful for reducing the computational complexity. Moreover, we also demonstrate the strict pathwise convergence of the proposed algorithm.

Third, we establish a comprehensive statistical theory for the TRUE method. Due to the complexity of nonconvex optimization problems, the theoretical framework commonly used in existing supervised clustering literature typically focuses on the statistical properties of a specific local minimizer, as detailed in Ma and Huang (2017), Ma et al. (2020), and Hu et al. (2021). However, this framework may lack practical significance, as the computable solution by any algorithm does not always converge to this specific local minimizer. In contrast, our TRUE method guarantees that any computable local minimizer that satisfies certain regularity conditions can enjoy appealing oracle properties, including consistency in parameter estimation, cluster recovery, and asymptotic normality, which demonstrates the completeness of our theoretical framework. Moreover, we design a GIC-type information criterion for selecting the optimal tuning parameters to choose the best candidate model in a heterogeneous setting. Although many existing criteria, such as AIC, BIC, and cross-validation, could potentially be used to select tuning parameters, no prior work has provided a theoretically guaranteed criterion for supervised clustering within a unified framework.

1.2. Literature review

Supervised clustering methodologies. In the literature, a popular class of methods for supervised clustering is to assume data as coming from a mixture of subgroups with their own sets of parameters, and then apply the finite mixture model analyses (Banfield and Raftery 1993, Hastie and Tibshirani 1996, McNicholas 2010, Andrews and McNicholas 2012, Wei and Kosorok 2013, Shen and He 2015). However, the mixture model approaches require specifying an underlying distribution for the data and the number of mixture components in advance, which is challenging in practice due to the inaccessibility of such prior information.

Many machine learning methods have also been introduced and applied for supervised clustering. For example, linear regression with two-way interactions becomes widely used. Yet such a method imposes strong parametric assumptions requiring that the underlying heterogeneity can be explained by those interactions (Greenland 2009). Nonparametric methods such as random forests are also popular, while their results remain less interpretable in practice (Wager and Athey 2018). Additionally, Wei and Kosorok (2013) introduced a new type of machine learning tool, named latent supervised learning, which combines the advantages of unsupervised learning by introducing latent labels. However, it relies on prior knowledge of the parametric functions defining cluster boundaries and is limited in the number of clusters.

To address these drawbacks, the fusion regularization approaches were proposed. This class of methods applies a shrinking penalty function to pairwise differences of subject-specific coefficients, thus can automatically identify the latent clusters in samples without knowledge of a prior classification or a natural basis for separating samples into subsets (Ma and Huang 2017, Ma et al. 2020, Chen et al. 2021, Wang and Su 2021, Zhang et al. 2024). Generally speaking, the concave or weighted L_1 penalization was widely used in existing fusion learning work. Nevertheless, the former is computationally unfriendly due to the nonconvex nature of optimization problem, while the later can be effected dramatically by the choice of weights and lacks theoretical guidelines. Besides, existing methods solving such fusion regularization problems necessitate repeating the optimization process for a sufficiently large grid of tuning parameters, bringing considerably expensive computational cost.

Stgewise learning. In parallel to the traditional learning scheme of “regularization + optimization” in solving the regularized problems, there has been a revival of interest into stagewise learning (Efron et al. 2004, Zhao and Yu 2007, Tibshirani 2015, He et al. 2018, Chen et al. 2022). Unlike traditional regularized optimization, a stagewise procedure builds a model from scratch, and gradually updates the model through releasing regularization constraints in a sequence of simple learning steps. For instance, stagewise lasso starts from the null model and updates coefficients by small increments, equilibrating the loss and L_1 penalty along the entire path (Efron et al. 2004). A comprehensive review of stagewise learning can be found in Tibshirani (2015), where its connections and differences with various optimization and machine learning approaches such as steepest descend, boosting, and path-following algorithms were discussed.

Stagewise learning possesses several notable features that make it particularly appealing for addressing penalized optimization problems. First, it allows for a flexible balance between statistical accuracy and computational efficiency through the choice of step size. Second, the learning process can halt early when a desirable model complexity is achieved, thereby avoiding excessive penalization. Lastly, it establishes a principled connection with regularized estimation. Specifically, for various problems, a stagewise procedure can be designed to approximate the solution paths of corresponding regularized estimation problems, with the convergence of these paths becoming exact as the step size approaches zero. However, despite its potential, stagewise learning is so far primarily confined to convex problems and standard regularization problems (Tibshirani 2015, Vaughan et al. 2017).

1.3. Organization

The rest of the paper is organized as follows. In Section 2, we demonstrate the problem setup and our method TRUE in detail. Section 3 develops a fast stagewsie algorithm for tracing out the solution paths of TRUE and shows its convergence analysis. Section 4 presents comprehensive theoretical properties of TRUE. The empirical performance of TRUE is illustrated through simulation studies in Section 5 and an empirical application to the motor insurance claim frequency data in Section 6. Section 7 concludes with some possible future works. Some supporting information is provided in the Supplementary Material, including the details of an ADMM algorithm for solving TURE, and an additional empirical application. Moreover, all the mathematical proofs are provided in an online document publicly available at `https://github.com/letianli059/REACH.`

Notations. For two positive sequences $\{u_n\}$ and $\{v_n\}$, we write $u_n \gg v_n$ or $v_n = o(u_n)$ if $v_n/u_n \to 0$. In addition, we write $u_n = O(v_n)$ if there exists a constant $c \geq 0$ such that $u_n/v_n \to c$. Similarly, we have $u_n = \Omega(v_n)$ if $v_n/u_n \to c$. Moreover, $v_n = o_P(u_n)$ means that $v_n/u_n \xrightarrow{P} 0$, and O_P is used similarly. For any vector $\boldsymbol{\xi} = (\xi_1, ..., \xi_d)^\top \in \mathbb{R}^d$, denote by $\|\boldsymbol{\xi}\|_q = (\sum_{i=1}^d |\xi_i|^q)^{1/q}$ for any $q \geq 0$. For any symmetric matrix $\mathbf{A}$, use $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ to denote the largest and smallest eigenvalue of $\mathbf{A}$, respectively.

2. Supervised clustering via the truncated fusion learning

2.1. Problem setup

We use $\{y_i, \mathbf{x}_i, \mathbf{z}_i\}_{i=1}^n$ to represent our mutually independent sample data, where y_i is the outcome, $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^\top$ is the p-dimensional vector of baseline covariates, and $\mathbf{z}_i = (1, z_{i1}, ..., z_{i(q-1)})^\top$ is the q-dimensional vector of heterogeneity covariates assumed to capture individualized effects on the outcome. Our goal is to uncover the association among the outcome and covariates, while simultaneously exploring heterogeneity/individualization within the data to facilitate personalized decision-making. To do this, we introduce a general heterogeneity tracking model grounded in a likelihood-based framework. Specifically, we consider that, conditional on $(\mathbf{x}_i, \mathbf{z}_i)$, the underlying likelihood of y_i has a general form of

$$f(y_i; \theta_i^*, \phi^*) \quad \text{with} \quad \theta_i^* = \mathbf{x}_i^\top \boldsymbol{\beta}^* + \mathbf{z}_i^\top \boldsymbol{\gamma}_i^*, \tag{1}$$

where $f(\cdot)$ is the likelihood function, θ_i^* is called the linear predictor, ϕ^* is the scalar parameter characterizing the distribution of y_i, $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is the population-shared coefficient vector, and $\boldsymbol{\gamma}_i^* \in \mathbb{R}^q$ is the subject-specific coefficient vector. Note that the first entry in each $\mathbf{z}_i$ is assumed to be 1. So $\boldsymbol{\gamma}_i^*$ includes the subject-specific intercept which integrates some unobserved individual effects. As we can see, model (1) is established within a general likelihood-based framework, where the response is connected to the explanatory variables through subject-specific linear predictors. The formulation $f(\cdot)$ offers clear interpretability without imposing additional assumptions on the function structure. Moreover, it is worth pointing that the representation of $\theta_i^* =$

$\mathbf{x}_i^\top \boldsymbol{\beta}^* + \mathbf{z}_i^\top \boldsymbol{\gamma}_i^*$ is a widely recognized form in the statistical learning literature for capturing data heterogeneity (Ma and Huang 2017, Ma et al. 2020, Chen et al. 2021). While most existing work focuses on linear regression, our study adopts this representation as a core to establish a unified likelihood-based framework capable of accommodating various data modalities, thereby facilitating a broad range of applications.

In the context of individualized modeling, we typically assume that y_i's are drawn from K^* different latent clusters. If y_i and y_j are in the same cluster, we have $\boldsymbol{\gamma}_i^* = \boldsymbol{\gamma}_j^*$. To characterize this point, let $\mathcal{G} = \{\mathcal{G}_1, ..., \mathcal{G}_{K^*}\}$ be the true cluster membership, which corresponds to the true cluster division of samples and is a mutually exclusive partition of $\{1, ..., n\}$. For $k = 1, ..., K^*$, let N_k be the size of the kth cluster $\mathcal{G}_k$. Denote by $\boldsymbol{\alpha}_k^*$ the common coefficient for the kth cluster. Accordingly, we have $\boldsymbol{\gamma}_i^* = \boldsymbol{\alpha}_k^*$ for all $i \in \mathcal{G}_k$. Throughout the paper, denote by $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}_1^{*\top}, ..., \boldsymbol{\alpha}_{K^*}^{*\top})^\top$ and $\boldsymbol{\gamma}^* = (\boldsymbol{\gamma}_1^{*\top}, ..., \boldsymbol{\gamma}_n^{*\top})^\top$.

We further demonstrate the generality of our model as follows. For instance, the heterogeneous generalized linear models (GLMs) (Hastie and Pregibon 2017) serve as popular examples of our setting. To be specific, the conditional distribution of y_i in GLMs belongs to the exponential family with the following density

$$f(y_i; \theta_i^*, \phi^*) = c(y_i) \exp\left\{ \frac{y_i \theta_i^* - b(\theta_i^*)}{\phi^*} \right\}, \tag{2}$$

where $\phi^* \in (0, \infty)$ is the known dispersion parameter, and $b(\cdot)$ and $c(\cdot)$ are known functions. Commonly used GLMs in practice include linear regression, logistic regression, and Poisson regression. Extensions of GLMs, such as the ordinal logistic model (McCullagh 1980) and the zero-inflated Poisson model (Lambert 1992), are also covered by (1) when considering the heterogeneous effects. Indeed, any response–predictor relationship with a known form of conditional distribution or likelihood can be modeled within the framework of (1). Moreover, it is worth noting that the partially observed data can also be modeled in the fashion of (1) by partial likelihood. The partially observed data, or specifically, the censored data, are usually found in survival analyses, where tools such as the Tobit model (Tobin 1958) and the accelerated failure time model (Wei 1992) are widely applied.

2.2. Truncated fusion learning

Without loss of generality, we consider the case where the parameter ϕ^* is known such that we can write $f(y_i; \theta_i^*, \phi^*) = f(y_i; \theta_i^*)$. We would like to emphasize that our method can be easily extended to the case where ϕ^* is unknown and need to be estimated. Clearly, the log-liklihood of the samples is given by

$$\mathcal{L}_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}, \boldsymbol{\gamma}_i) = \sum_{i=1}^{n} \log f(y_i; \theta_i),$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, ..., \boldsymbol{\gamma}_n^\top)^\top$ and $\theta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{\gamma}_i$. Here we assume that $\mathcal{L}_n(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is strictly concave and differentiable. To estimate the coefficients and identify the latent clusters simultaneously, we propose a new method called truncated fusion learning (TRUE) which minimizes

$$Q_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) = -\frac{1}{n}\mathcal{L}_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) + \lambda_n \sum_{i<j} p_{\kappa_n}(\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|_2), \tag{3}$$

where λ_n is the regularization parameter that controls the strength of penalization, and $p_{\kappa_n}(t) = |t|\mathbf{1}_{\{|t| \le \kappa_n\}} + \kappa_n \mathbf{1}_{\{|t| > \kappa_n\}}$ is the penalty function with $\kappa_n > 0$ being the thresholding parameter. Then we define our TRUE estimator as

$$(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\boldsymbol{\gamma}}(\lambda)) = \arg\min_{\boldsymbol{\beta}, \boldsymbol{\gamma}} Q_n(\boldsymbol{\beta}, \boldsymbol{\gamma})$$

with $\lambda = (\lambda_n, \kappa_n)$. As we can see, the penalty would shrink some of the fusion pairs $\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j$ to zero. Based on this, we can assign the samples into several clusters. Correspondingly, we denote by $\widehat{K}$ the estimated cluster number and $\widehat{\mathcal{G}} = \{\widehat{\mathcal{G}}_1, ..., \widehat{\mathcal{G}}_{\widehat{K}}\}$ the estimated cluster membership.

We now introduce the computational and theoretical merits of the proposed truncated fusion learning. Note that much of the existing related work typically utilizes concave penalties such as smoothly clipped absolute deviation (SCAD) (Fan and Li 2001) and minimax concave penalty (MCP) (Zhang 2010). While both concave penalties and the truncated L_1 penalty achieve asymptotic unbiasedness by moderating the shrinkage on large fusion terms, concave penalization methods are computationally unfriendly due to their concavity. Therefore, the truncated L_1 penalty can be regarded as a convex relaxation of concave penalties in order to capture computational benefits while retaining appealing statistical properties comparable to those of concave fusion methods. Such penalty has been also considered in other contexts such as the wavelet denoising (Antoniadis 1997), variable selection via integer programming (Liu and Wu 2007), and grouping pursuit (Shen and Huang 2010).

Some discussion about the two tuning parameters are as follows. The regularization parameter λ_n controls the strength of shrinkage, while the thresholding parameter κ_n determines which pairs' differences should be shrunken. It is intuitive that there should be a balance between these two parameters. We suggest to set $\kappa_n = a\lambda_n$ for some positive constant $a > 1$ (e.g. $a = 3$) by borrowing the idea from concave penalties. As illustrated in Figure 1, the truncated L_1, SCAD and MCP penalty functions exhibit similar growth patterns from the origin and reach plateaus at the same transition point under the suggested setting. In view of the appealing performance of concave penalties in balancing the penalization strength and the over-shrinkage prevention, we structure the truncated L_1 penalty by emulating concave penalties to achieve desirable estimation properties. Our arguments are supported by the theoretical and empirical results in Sections 4-6.

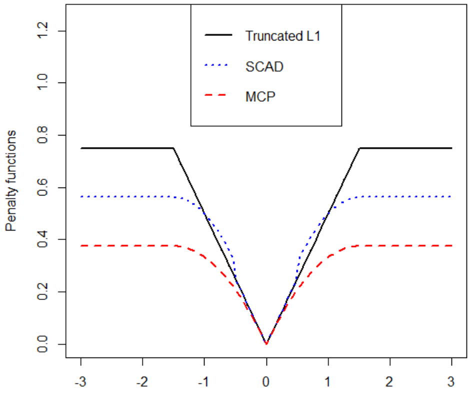

Figure 1 **The truncated L_1 penalty function with $\lambda_n = 0.5$ and $\kappa_n = 3\lambda_n$, and SCAD and MCP penalty functions with their concavity control parameters set to 3.**

3. Fast stagewise algorithm for TURE

3.1. Stagewise TURE

To solve the fusion penalized optimization in the classical linear model setting, the alternating direction method of multipliers (ADMM) developed by Ma and Huang (2017) is a commonly used choice. For further accommodating our optimization problem (3), a natural extension of the ADMM framework entails employing Newton-Raphson iterations for updating $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ numerically. We relegate the detailed procedures and the convergence guarantees of the corresponding ADMM algorithm for solving (3) to the Supplementary Material. However, it is worth noting that the number of fusion terms in the objective function (3) is of order $O(n^2)$, thus the ADMM algorithm suffers from a very high computational complexity. Additionally, with such a traditional optimization scheme, the computation process needs to be repeated for a sufficiently large grid of tuning parameters for locating the optimal model. Therefore, implementing TRUE via the ADMM algorithm would lead to skyrocketing computational costs, especially for large-scale problems.

To enhance the scalability, we innovate a thresholded stagewise fusion learning strategy to quickly trace out the whole solution paths of TRUE. Before delving into the specifics of our proposed algorithms, we would like to provide an overview of the main concepts and potential advantages of stagewise learning. Broadly speaking, stagewise learning is a gradual process for model construction by mapping out a spectrum of potential models through repetitive, straightforward calculations. In the context of fusion learning, a stagewise procedure begins with an unpenalized model, and progresses through simple, incremental steps to update the parameters and cluster the individuals simultaneously. At each step, the regularization parameters also increase, ensuring that the updated parameters achieve optimal values for the current objective function.

Although TRUE is a nonconvex problem with fusion penalization, we come to realize that efficient and principled stagewise learning remains possible. Our idea to address nonconvexity is simple yet elegant:

motivated by the local convexity and truncation of the truncated L_1 penalty, we trace out the entire solution paths in a piecewise manner; in each piece, we solve a subproblem with non-fully connected L_1 fusion penalties constructed by pre-thresholding fusion terms. We first make a simple observation. In certain regions where some paired differences $(\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j)$'s have the L_2-norms larger than the thresholding parameter κ_n, the optimization problem is equivalent to an L_1 fusion penalized problem with non-fully connected fusions. To be specific, by defining the active set as $S = \{(i,j) : \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|_2 < \kappa_n\}$, it can be seen that minimizing $Q_n(\boldsymbol{\beta}, \boldsymbol{\gamma})$ constrained on S is equivalent to solving

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \left\{ -\frac{1}{n}\mathcal{L}_n(\boldsymbol{\beta},\boldsymbol{\gamma}) + \lambda_n \sum_{(i,j)\in S} \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|_2 \right\}. \tag{4}$$

Therefore, we can approximate the solution paths by piecewisely solving the L_1 fusion penalized problem (4). Along the solution paths, the subproblem in each piece is strictly convex, and the corresponding minimizer is also a local minimizer of the objective function $Q_n(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Moreover, it is noteworthy that our strategy also significantly reduces the computational complexity by dropping redundant fusion terms, which shares similarities with the nearest-neighbor-based method (Chen et al. 2021) and the spanning-tree-based method (Li and Sang 2019, Zhang et al. 2024). In particular, a distinctive feature is that our approach introduces a specific thresholding level to filter fusion terms and progressively updates the penalization subset during the computation process.

On the other hand, to disentangle the nuisance fusion penalization in (4), we utilize its duality. Denote by L_n the loss function, i.e., $L_n(\boldsymbol{\beta},\boldsymbol{\gamma}) = -\frac{1}{n}\mathcal{L}_n(\boldsymbol{\beta},\boldsymbol{\gamma})$. Let $\boldsymbol{\Delta} = \widetilde{\boldsymbol{\Delta}} \otimes \mathbf{I}_q$, where $\widetilde{\boldsymbol{\Delta}} = \{(\mathbf{e}_i - \mathbf{e}_j), i<j\}^\top$, $\mathbf{e}_i$ is the $n \times 1$ unit vector with all 0's except for a 1 in its ith coordinate, and $\otimes$ denotes the Kronecker product. Specifically, we consider a dual formulation of (4):

$$\widehat{\boldsymbol{\eta}}(\lambda_n) = \arg\min_{\boldsymbol{\eta}\in\mathcal{S}(\boldsymbol{\eta})} L_n^*(\mathbf{0}, \boldsymbol{\Delta}^\top\boldsymbol{\eta}), \text{ subject to } \sup_{(i,j)\in S} \|\boldsymbol{\eta}_{ij}\|_2 \le \lambda_n, \tag{5}$$

where L_n^* is the convex conjugate of L_n, i.e., $L_n^*(\mathbf{u},\mathbf{v}) = \sup_{\boldsymbol{\beta},\boldsymbol{\gamma}}\{\mathbf{u}^\top\boldsymbol{\beta} + \mathbf{v}^\top\boldsymbol{\gamma} - L_n(\boldsymbol{\beta},\boldsymbol{\gamma})\}$, S is the active set, $\boldsymbol{\eta} = \{\boldsymbol{\eta}_{ij}^\top, i<j\}^\top$, and $\mathcal{S}(\boldsymbol{\eta}) = \{\boldsymbol{\eta} : \boldsymbol{\eta}_{ij} = \mathbf{0}, \forall (i,j) \in S^c\}$. According to the stationarity condition, the primal and dual solutions satisfy the following equations:

$$\begin{cases} \nabla_{\boldsymbol{\gamma}} L_n(\widehat{\boldsymbol{\beta}}(\lambda_n), \widehat{\boldsymbol{\gamma}}(\lambda_n)) - \boldsymbol{\Delta}^\top \widehat{\boldsymbol{\eta}}(\lambda_n) = \mathbf{0}, \\ \nabla_{\boldsymbol{\beta}} L_n(\widehat{\boldsymbol{\beta}}(\lambda_n), \widehat{\boldsymbol{\gamma}}(\lambda_n)) = \mathbf{0}. \end{cases} \tag{6}$$

We now present in detail our proposed stagewise algorithm for TRUE. Each iteration of our algorithm contains two stages: one is the pairs selection procedure, and the other is the parameter updating step. In the first stage, we construct the current active set S based on the previous estimate of $\boldsymbol{\gamma}$. Then, the second stage updates the parameters along the solution path of (4) conditional on S. It is important to point out that in our algorithm, we first computes the solution path of the dual problem (5), and then convert it into a primal solution path via (6).

Motivated by Tibshirani (2015), we apply a forward stagewise scheme that involves iteratively updating the current iterate in a direction that minimizes the inner product with the gradient of the loss function, and simultaneously restrict this direction to be small under the regularization constraint. Through repeating these updates, one can implicitly adjust the trade-off between minimizing the loss and the constraint, thereby approximating the solutions of the regularization problem pathwisely. In our optimization, the framework of the stagewise procedure is as follows: for a small step size $\epsilon > 0$, repeat for $t = 0, 1, 2, ...,$

$$\begin{aligned} &\boldsymbol{\eta}^{(t+1)} \leftarrow \alpha \boldsymbol{\eta}^{(t)} + \widehat{\boldsymbol{\delta}}, \\ \text{with } \widehat{\boldsymbol{\delta}} = \arg &\min_{\boldsymbol{\delta} \in \mathcal{S}(\boldsymbol{\delta})} \langle \nabla_{\boldsymbol{\eta}} L_n^*(\mathbf{0}, \boldsymbol{\Delta}^\top \boldsymbol{\eta}^{(t)}), \boldsymbol{\delta} \rangle, \text{subject to } \sup_{(i,j)\in S} \|\boldsymbol{\delta}_{ij}\|_2 \leq \epsilon, \end{aligned} \tag{7}$$

where $\boldsymbol{\delta} = \{\boldsymbol{\delta}_{ij}^\top, i < j\}^\top$, $\mathcal{S}(\boldsymbol{\delta}) = \{\boldsymbol{\delta} : \boldsymbol{\delta}_{ij} = \mathbf{0}, \forall (i,j) \in S^c\}$, and $\alpha \in (0,1)$ is the shrunken parameter which alternates the backward strategy to reduce bias. The above is just a profile of iterations, and the exact formula of the updating procedure is provided in the following proposition.

PROPOSITION 1. *Suppose L_n^* is differentiable. Then the increment $\widehat{\boldsymbol{\delta}}$ in (7) has the form:*

$$\widehat{\boldsymbol{\delta}}_{ij} = \begin{cases} -\frac{\gamma_i^{(t)} - \gamma_j^{(t)}}{\|\gamma_i^{(t)} - \gamma_j^{(t)}\|_2}\epsilon, & \gamma_i^{(t)} - \gamma_j^{(t)} \neq \mathbf{0}, \\ \mathbf{0}, & \gamma_i^{(t)} - \gamma_j^{(t)} = \mathbf{0}, \end{cases} \tag{8}$$

for all $(i,j) \in S$, and $\widehat{\boldsymbol{\delta}}_{i'j'} = \mathbf{0}$ for all $(i',j') \in S^c$.

The proof of Proposition 1, whose details are given in the Supplementary Material, follows from the Karush–Kuhn–Tucker (KKT) conditions and the primal-dual relationship. This proposition implies that the increament does not depend on L_n^* but only on $\boldsymbol{\gamma}^{(t)}$, so we do not need to know the specific expression of L_n^*. In fact, it is not easy to determine L_n^* for many commonly used likelihood functions.

Based on the discussion above, we summarize the pseudocode of the proposed stagewise learning algorithm in Algorithm 1. There remain some key points in Algorithm 1. First, there should be a reset of $\boldsymbol{\eta}$ in the pair selection stage if some pairs dropped by the active set. This optional step transforms the previous estimate into a reasonable start point for the new optimization problem. The detailed reasoning can be seen in the

Algorithm 1 Stagewise TRUE

```
Require: the scale parameter $a > 1$, the shrunken parameter $\alpha \in (0,1)$, a small step size $\epsilon > 0$, and the maximum iteration number $T$.
  Initialize $\lambda_n^{(0)} \leftarrow 0$, $\kappa_n^{(0)} \leftarrow a\lambda_n^{(0)}$, $\boldsymbol{\eta}^{(0)} \leftarrow \mathbf{0}$, $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$, $S^{(0)} \leftarrow \{(i,j) : \|\boldsymbol{\gamma}_i^{(0)} - \boldsymbol{\gamma}_j^{(0)}\|_2 < \kappa_n^{(0)}\}$.
  for $t = 0, 1, 2, \dots, T$ do
    Pairs selection:
    Update the active set $S^{(t+1)} \leftarrow \{(i,j) : \|\boldsymbol{\gamma}_i^{(t)} - \boldsymbol{\gamma}_j^{(t)}\|_2 < \kappa_n^{(t)}\}$.
    Set $\boldsymbol{\eta}_{i'j'}^{(t)} = \mathbf{0}$ if there are some pairs $(i', j')$ dropped by $S^{(t+1)}$.
    Forward updating:
    Compute the increament $\widehat{\boldsymbol{\delta}}$ by (8).
    Update the dual parameter $\boldsymbol{\eta}^{(t+1)} \leftarrow \alpha\boldsymbol{\eta}^{(t)} + \widehat{\boldsymbol{\delta}}$.
    Solve the primal-dual equations (6) to obtain the estimates $\boldsymbol{\beta}^{(t+1)}$ and $\boldsymbol{\gamma}^{(t+1)}$.
    Update the regularization parameter $\lambda_n^{(t+1)} \leftarrow \sup_{(i,j)\in S^{(t+1)}} \|\boldsymbol{\eta}_{ij}^{(t+1)}\|_2$.
    Update the thresholding parameter $\kappa_n^{(t+1)} \leftarrow a\lambda_n^{(t+1)}$.
  end for
  return the solution path of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$.
```

proof of the following Theorem 1. The second revolves around the solution of the primal-dual equations (6). For the classical linear regression model, the equations have a closed-form solution:

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}(\lambda_n) \\ \widehat{\boldsymbol{\gamma}}(\lambda_n) \end{pmatrix} = \begin{pmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\mathbf{Z} \\ \mathbf{Z}^\top\mathbf{X} & \mathbf{Z}^\top\mathbf{Z} \end{pmatrix}^+ \begin{pmatrix} \mathbf{X}^\top\mathbf{y} \\ n\boldsymbol{\Delta}^\top\widehat{\boldsymbol{\eta}}(\lambda_n) + \mathbf{Z}^\top\mathbf{y} \end{pmatrix},$$

where $^+$ stands for the Moore-Penrose inverse, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^\top$, $\mathbf{Z} = \mathrm{diag}(\mathbf{z}_1^\top, ..., \mathbf{z}_n^\top)$, and $\mathbf{y} = (y_1, ..., y_n)^\top$. While for other models, numerical approaches such as the Newton-Raphson method or gradient descent are applicable options if we can not obtain a closed-form solution. Since the step size ϵ is small, in order to obtain a solution in practice, a few iterations are often sufficient when using the last iteration estimate as the initial point. Third, Algorithm 1 tracks the solution path along λ_n while simultaneously adjusting $\kappa_n = a\lambda_n$. It facilitates the choice of κ_n and also ensures desirable estimates. Some discussion about the balance between the two tuning parameters has been given in Section 2.2.

Although we adopt the idea of the forward stagewise procedure proposed in Tibshirani (2015), our work is fundamentally different. The most important point is that their work targets the convex framework, whereas we deal with a nonconvex optimization problem and only draw on their idea to compute the path solutions of the subproblems. Moreover, our work is devoted to the problem of supervised clustering by the pairwise fusion regularization, which apparently differs from the standard regularization problem in their work. All of these are the new contributions of our study to stagewise learning.

The convergence of the proposed algorithm is guaranteed by the following theorem.

THEOREM 1. *(Convergence of stagewise TRUE) Let* $g_S(\boldsymbol{\eta}) = \sup_{(i,j)\in S} \|\boldsymbol{\eta}_{ij}\|_2$ *and* $g_S^*(\boldsymbol{\eta}) = \sum_{(i,j)\in S} \|\boldsymbol{\eta}_{ij}\|_2$ *for all possible* S*. Suppose that* L_n^* *is differentiable and convex, and* ∇L_n^* *is Lipschitz with respect to* g_S *and* g_S^* *for all* S*. If Algorithm 1 starts from* $(\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)})$*, a solution of* (3) *when* $\lambda_n = 0$*, and*

$$\frac{1-\alpha}{\epsilon} \to 0 \quad \text{and} \quad \frac{1-\alpha}{\epsilon^2} \to \infty$$

hold as $\epsilon \to 0$ *and* $\alpha \to 1$*, then for each* $t > 0$*,* $(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)})$ *converges to a local minimizer of* (3) *with respect to* $(\lambda_n^{(t)}, \kappa_n^{(t)})$*.*

Theorem 1 relies on some regularity conditions regarding the conjugate loss function L_n^*. Although L_n^* does not always have an explicit expression in certain applications, we would also like to emphasize that these regularity conditions are mild and can be guaranteed by the strong convexity and differentiability of the primal function L_n. First, with the aid of Danskin's theorem (Bertsekas 1997), if L_n is strongly convex and differentiable, then some simple algebra immediately yields that L_n^* is differentiable. Second, if L_n is strongly convex, it follows easily from Theorem 1 in Zhou (2018) that ∇L_n^* is Lipschitz with respect to g_S and g_S^* for all S.

3.2. Initialization and tuning

According to Theorem 1, the appealing performance of Algorithm 1 needs an initial point that minimizes the negative log-liklihood, i.e., $-\mathcal{L}_n(\boldsymbol{\beta}, \boldsymbol{\gamma})$. It is clear that the minimization of $-\mathcal{L}_n(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is challenging and its optimizer may be not unique due to the over-parameterization. To resolve this problem, we suggest a ridge fusion estimate. That is, we obtain $(\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)})$ by solving

$$\min_{\boldsymbol{\beta}, \boldsymbol{\gamma}} \left\{ -\frac{1}{n}\mathcal{L}_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) + \frac{\lambda^*}{2} \sum_{i<j} \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|_2^2 \right\},$$

where $\lambda^* \to 0$ is the tuning parameter. We would like to emphasize that $\lambda^* \to 0$ is assumed to guarantee the convergence property in Theorem 1. In practice, the tuning parameter can be set to be sufficiently small, e.g., $\lambda^* = 0.001$. The motivation is intuitive: the proposed initial estimate approximates the minimizer of $-\mathcal{L}_n(\boldsymbol{\beta}, \boldsymbol{\gamma})$ due to the sufficiently small penalty term, and it helps to address the unidentifiability of the optimization problem. Such strategy was also widely used in the ADMM algorithm; see, for example, Ma et al. (2020) and Hu et al. (2021).

On the other hand, the performance of the regularization optimization depends on the tuning parameters λ_n and κ_n. In practice, we can choose (λ_n, κ_n) by minimizing the following GIC-type information criterion

$$GIC(\lambda) = -\frac{1}{n}\mathcal{L}_n(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\boldsymbol{\gamma}}(\lambda)) + C_n \frac{\log\log n}{\sqrt{n}}(q\widehat{K}(\lambda) + p), \tag{9}$$

where $C_n \in [1, 5]$ provides a good choice. We shall establish the validity of this tuning parameter selector in Section 4.

3.3. Refinement

To overcome the overfitting issue, we suggest refining the local minimizer sequence $\{(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)})\}_{t=1}^{T}$ obtained by the stagewise TRUE algorithm through some unsupervised clustering techniques. Specifically, we can apply some state-of-the-art unsupervised clustering methods, such as K-means and hierarchical clustering, to perform secondary clustering on $\boldsymbol{\gamma}_i^{(t)}$ such that we can obtain the refined sequence $\{(\boldsymbol{\beta}^{(t)}, \widetilde{\boldsymbol{\gamma}}^{(t)})\}_{t=1}^{T}$. Based on this refined sequence, the final refined estimator $(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\boldsymbol{\gamma}}(\lambda))$ is tuned by the GIC criterion given in (9).

The main idea behind the suggested refinement through secondary clustering is motivated by a common practical scenario: although the local minimizer obtained by the stagewise TRUE algorithm yields small estimation error, it may result in redundant subclusters (i.e., clusters with small inter-class distances), leading to an overestimation of the number of clusters. The suggested refinement procedure effectively avoids these potential overfitted clusters, thereby improving clustering accuracy. It is worth pointing out that this refinement strategy also shares a similar spirit to the post thresholding for variable selection in the field of high-dimensional statistics (Fan and Lv 2013, Fan et al. 2019, Zheng et al. 2019).

4. Asymptotic Properties of TRUE

4.1. Technical assumptions

To begin with, we first introduce some necessary notation and definitions which will be used later on. Without loss of generality, we suppose the true number of clusters $K^* \geq 2$. Let

$$\mathcal{M}_{\mathcal{G}} = \left\{\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, ..., \boldsymbol{\gamma}_n^\top)^\top \in \mathbb{R}^{nq} : \boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j, \forall i, j \in \mathcal{G}_k, 1 \leq k \leq K^*\right\} \tag{10}$$

be a subset of $\mathbb{R}^{nq}$. Define $\mathbf{D}_0 = \{d_{ij}\}$ as the $n \times K^*$ cluster indicator matrix with $d_{ij} = 1$ if $j = k$ and otherwise $d_{ij} = 0$ given $i \in \mathcal{G}_k$. Denote by $\otimes$ the Kronecker product, and let $\mathbf{D} = \mathbf{D}_0 \otimes \mathbf{I}_m$. For each $\boldsymbol{\gamma} \in \mathcal{M}_{\mathcal{G}}$, it can be rewritten as $\boldsymbol{\gamma} = \mathbf{D}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, ..., \boldsymbol{\alpha}_{K^*}^\top)^\top \in \mathbb{R}^{qK^*}$, and $\boldsymbol{\alpha}_k$ is the kth cluster-specific parameter vector for $k = 1, ..., K^*$. Furthermore, we introduce the mapping $T_{\mathcal{G}} : \mathcal{M}_{\mathcal{G}} \to \mathbb{R}^{qK^*}$ such that $T_{\mathcal{G}}(\boldsymbol{\gamma}) = \boldsymbol{\alpha}$. Based on these, denote by $\widetilde{\mathbf{Z}} = \mathbf{D}\mathbf{Z}$. We define $N_{\max} = \max_k N_k$ and $N_{\min} = \min_k N_k$ as the maximal and minimal cluster sizes, respectively. Moreover, we let

$$d_n = \min_{i \in \mathcal{G}_k, j \in \mathcal{G}_{k'}, k \neq k'} \|\boldsymbol{\gamma}_i^* - \boldsymbol{\gamma}_j^*\|_2 = \min_{k \neq k'} \|\boldsymbol{\alpha}_k^* - \boldsymbol{\alpha}_{k'}^*\|_2 \tag{11}$$

be the minimal difference of the common values between two clusters. For $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\alpha}_k \in \mathbb{R}^q, k = 1, ..., K^*$, we write $\boldsymbol{\zeta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top$ and $\boldsymbol{\zeta}_k = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}_k^\top)^\top$. Correspondingly, $\boldsymbol{\zeta}^*$ and $\boldsymbol{\zeta}_k^*$ refer to the true value counterparts of $\boldsymbol{\zeta}$ and $\boldsymbol{\zeta}_k$, respectively.

ASSUMPTION 1. *Conditional on $(\mathbf{x}_i, \mathbf{z}_i)$ for $i = 1, ..., n$, the observations y_i are independent and have the identical form of likelihood $f(y_i; \theta_i^*)$ with a common support, and the model is identifiable. Furthermore, the first and second derivatives of the likelihood functions $\ell_i(\boldsymbol{\beta}, \boldsymbol{\alpha}_k) = \log f(y_i; \theta_i)$ satisfy the equations*

$$\mathbb{E}_{\boldsymbol{\zeta}_k^*}\left\{\frac{\partial \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_k^*)}{\partial \boldsymbol{\zeta}_k}\right\} = \mathbf{0}$$

and

$$\mathbb{E}_{\boldsymbol{\zeta}_k^*}\left[\left\{\frac{\partial \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_k^*)}{\partial \boldsymbol{\zeta}_k}\right\}\left\{\frac{\partial \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_k^*)}{\partial \boldsymbol{\zeta}_k}\right\}^\top\right] = -\mathbb{E}_{\boldsymbol{\zeta}_k^*}\left\{\frac{\partial^2 \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_k^*)}{\partial \boldsymbol{\zeta}_k^2}\right\}$$

for any $i \in \mathcal{G}_k, k = 1, ..., K^*$.

ASSUMPTION 2. *The Fisher information matrices*

$$I(\boldsymbol{\zeta}_k^*) = \mathbb{E}_{\boldsymbol{\zeta}_k^*}\left[\left\{\frac{\partial \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_k^*)}{\partial \boldsymbol{\zeta}_k}\right\}\left\{\frac{\partial \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_k^*)}{\partial \boldsymbol{\zeta}_k}\right\}^\top\right]$$

for all $i \in \mathcal{G}_k, k = 1, ..., K^*$ satisfy

$$0 < \max_k [\lambda_{\max}\{I(\boldsymbol{\zeta}_k^*)\}] < \infty.$$

Moreover, the augmented Fisher information matrix corresponding to $\mathcal{L}_n(\boldsymbol{\beta}, \mathbf{D}\boldsymbol{\alpha}) = \sum_{k=1}^{K^*}\sum_{i\in\mathcal{G}_k} \ell_i(\boldsymbol{\beta}, \boldsymbol{\alpha}_k)$, that is,

$$I_n(\boldsymbol{\zeta}^*) = \mathbb{E}_{\boldsymbol{\zeta}^*}\left[\left\{\frac{\partial \mathcal{L}_n(\boldsymbol{\beta}^*, \mathbf{D}\boldsymbol{\alpha}^*)}{\partial \boldsymbol{\zeta}}\right\}\left\{\frac{\partial \mathcal{L}_n(\boldsymbol{\beta}^*, \mathbf{D}\boldsymbol{\alpha}^*)}{\partial \boldsymbol{\zeta}}\right\}^\top\right]$$

satisfies

$$0 < N_{\min}^{-1}\lambda_{\min}\{I_n(\boldsymbol{\zeta}^*)\} < \infty.$$

ASSUMPTION 3. *The second moments for the second derivatives of the likelihood functions are bounded, that is,*

$$\mathbb{E}_{\boldsymbol{\zeta}_k^*}\left\{\frac{\partial^2 \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_k^*)}{\partial \zeta_{k,j}\partial \zeta_{k,l}}\right\}^2 < \infty$$

for $\boldsymbol{\zeta}_k = (\zeta_{k,1}, ..., \zeta_{k,p+q})^\top$, any $i \in \mathcal{G}_k, k = 1, ..., K^*$ and $j, l = 1, ..., p+q$. Moreover, there exists an open set $\boldsymbol{\Omega} \in \mathbb{R}^{qK^*+p}$ which contains the true parameter point $\boldsymbol{\zeta}^*$, such that for almost all $(y_i, \mathbf{x}_i, \mathbf{z}_i)$, the conditional likelihood admits all third derivatives for all $\boldsymbol{\zeta} \in \boldsymbol{\Omega}$. Moreover, we have

$$\mathbb{E}_{\boldsymbol{\zeta}_k}\left\{\left|\frac{\partial^3 \ell_i(\boldsymbol{\beta}, \boldsymbol{\alpha}_k)}{\partial \zeta_{k,j}\partial \zeta_{k,l}\partial \zeta_{k,m}}\right|\right\} < \infty$$

for all $\boldsymbol{\zeta} \in \boldsymbol{\Omega}$, $i \in \mathcal{G}_k, k = 1, ..., K^*$ and $j, l, m = 1, ..., p+q$.

Assumptions 1-3 are natural extensions of the assumptions that guarantee asymptotic normality of the ordinary maximum likelihood estimates for homogeneous modeling introduced in Fan and Li (2001) to our clustered setting. Similar conditions have also been used in other works; see, for example, Fan and Peng (2004), Jeon et al. (2017) and Su et al. (2018). Below, we provide further explanations of these assumptions.

Assumption 1 imposes conditions on the score function to regulate random perturbations in the data distribution, which is standard in the maximum likelihood framework (Lehmann and Casella 2006). To be specific, the first equation assumes that the score function is mean zero, which can be interpreted as a generalized zero-mean noise condition. The second equation establishes a relationship between the covariance of the score function and the second-order moment of the log-likelihood, that is, it supposes that the uncertainty of model is determined by the curvature (i.e., the second derivative) of the log-likelihood function.

Assumption 2 ensures model identifiability by constraining the collinearity among the predictors. In this context, the Fisher information matrices can be viewed as predictor covariance matrices, and constraining their eigenvalues serves to control the predictor collinearity. Notably, in the Gaussian model setting, the requirement on the augmented Fisher information matrix is equivalent to $\lambda_{\min}\{(\widetilde{\mathbf{Z}},\mathbf{X})^\top(\widetilde{\mathbf{Z}},\mathbf{X})\} > CN_{\min}$ for some constant $C > 0$. This aligns with the smallest eigenvalue assumption proposed by Ma et al. (2020), which is widely adopted in the current supervised clustering literature. Here, we extend this condition to a more general model setting.

Assumption 3 imposes mild constraints on the higher-order derivatives of the log-likelihood function to control the smoothness of the underlying distributions, which holds automatically in common heterogeneous generalized linear models, such as linear regression, logistic regression, and Poisson regression. In particular, the boundedness of the third-order derivative ensures that the log-likelihood function can be well-approximated by quadratic functions (Fan and Peng 2004), which further facilitates the application of Taylor's expansion in the derivation of the oracle properties of the TRUE estimator.

4.2. Theoretical results

In this section, we demonstrate the main theoretical results of TRUE.

THEOREM 2. *(Oracle properties of the TRUE estimator) Suppose Assumptions 1-3 hold. If* $N_{\min} \gg (p+q)n^{2/3}$, $d_n > a\lambda_n$, $\lambda_n \gg \tau_n$, *and* $\kappa_n = a\lambda_n$, *where* d_n *is defind in* (11), $\tau_n = CN_{\min}^{-1}\sqrt{(p+q)n}$, *and* a, C *are positive constants, then there exists a local minimizer* $(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\boldsymbol{\gamma}}(\lambda))$ *of the objective function* $Q_n(\boldsymbol{\beta},\boldsymbol{\gamma})$ *given in* (3) *satisfying*

$$\begin{aligned}
\|((\widehat{\boldsymbol{\alpha}}(\lambda)-\boldsymbol{\alpha}^*)^\top,(\widehat{\boldsymbol{\beta}}(\lambda)-\boldsymbol{\beta}^*)^\top)^\top\|_2 &= O\left(N_{\min}^{-1}\sqrt{(p+q)n}\right),\\
\|\widehat{\boldsymbol{\gamma}}(\lambda)-\boldsymbol{\gamma}^*\|_2 &= O\left(N_{\min}^{-1}\sqrt{N_{\max}(p+q)n}\right),\\
\sup_i \|\widehat{\boldsymbol{\gamma}}_i(\lambda)-\boldsymbol{\gamma}_i^*\|_2 &= O\left(N_{\min}^{-1}\sqrt{(p+q)n}\right).
\end{aligned} \tag{12}$$

with probability tending to 1 as $N_k \to \infty$ *for* $k = 1, \ldots, K^*$, *where* $\widehat{\boldsymbol{\alpha}}(\lambda) = T_{\mathcal{G}}(\widehat{\boldsymbol{\gamma}}(\lambda))$. *Moreover, it also enjoys the following results:*

(1) the estimated cluster membership $\widehat{\mathcal{G}}$ satisfies $\mathbb{P}(\widehat{\mathcal{G}} = \mathcal{G}) \to 1$;

(2) the estimated cluster number $\widehat{K}$ satisfies $\mathbb{P}(\widehat{K} = K^*) \to 1$;

(3) for any $m \times (qK^* + p)$ matrix $\mathbf{A}_n$ such that $\mathbf{A}_n \mathbf{A}_n^\top \to \mathbf{S}$, where $\mathbf{S}$ is a $m \times m$ symmetric positive definite matrix, the estimator has the limiting distribution

$$\sqrt{n}\mathbf{A}_n I_n^{1/2}(\boldsymbol{\zeta}^*)((\widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}^*)^\top, (\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*)^\top)^\top \xrightarrow{d} N(\mathbf{0}, \mathbf{S}), \tag{13}$$

or equivalently, it can be expressed as

$$\lim_{n\to\infty} \sqrt{n}\mathbf{A}_n I_n^{1/2}(\boldsymbol{\zeta}^*)((\widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}^*)^\top, (\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*)^\top)^\top \sim N(\mathbf{0}, \mathbf{S}).$$

Theorem 2 shows the asymptotic properties of a local minimizer of TRUE, including the oracle properties, model selection consistency and asymptotic normality. It is worth mentioning that the error bound in (12) coincides with the bound for the unpenalized estimator assuming that the true cluster structure is known, which implies the desirable accuracy of our estimator. Moreover, as we can see in Ma et al. (2020), the corresponding error bound for the Gaussian linear model is within the order of $O(N_{\min}^{-1}\sqrt{(qK^*+p)n})$. Notably, this order includes an extra factor K^*. Our bound releases the direct effect of the true cluster number K^* on the estimation error by imposing different conditions. To be specific, Assumption 2 restricts the largest eigenvalues of Fisher information matrices corresponding to each true cluster separately, whereas Ma et al. (2020) considered the design matrix assumption for the entire sample. Nevertheless, there remains some restrictions on K^* in our theoretical results. Given $N_{\min} \le n/K^*$, in view of the condition $N_{\min} \gg (p+q)n^{2/3}$, it must satisfy $K^* = o(\frac{n^{1/3}}{p+q})$.

With the asymptotic normality demonstrated in (13), we can conduct further statistical inference for the proposed estimator. In particular, the covariance matrix of $((\widehat{\boldsymbol{\alpha}}(\lambda))^\top, (\widehat{\boldsymbol{\beta}}(\lambda))^\top)^\top$ is $\frac{1}{n}I_n^{-1}(\boldsymbol{\zeta}^*)$. When the true cluster structure is recovered, we can use the following

$$\widehat{\boldsymbol{\Sigma}} = n\{\nabla^2\mathcal{L}_n(\widehat{\boldsymbol{\beta}}(\lambda), \mathbf{D}\widehat{\boldsymbol{\alpha}}(\lambda))\}^{-1}\widehat{\mathrm{cov}}\{\nabla\mathcal{L}_n(\widehat{\boldsymbol{\beta}}(\lambda), \mathbf{D}\widehat{\boldsymbol{\alpha}}(\lambda))\}\{\nabla^2\mathcal{L}_n(\widehat{\boldsymbol{\beta}}(\lambda), \mathbf{D}\widehat{\boldsymbol{\alpha}}(\lambda))\}^{-1},$$

as the estimated covariance matrix, where

$$\nabla^2\mathcal{L}_n(\widehat{\boldsymbol{\beta}}(\lambda), \mathbf{D}\widehat{\boldsymbol{\alpha}}(\lambda)) = \sum_{k=1}^{K^*}\sum_{i\in\mathcal{G}_k} \frac{\partial^2 \ell_i(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\boldsymbol{\alpha}}_k(\lambda))}{\partial\boldsymbol{\zeta}^2}$$

and

$$\widehat{\text{cov}}\{\nabla \mathcal{L}_n(\widehat{\boldsymbol{\beta}}(\lambda), \mathbf{D}\widehat{\boldsymbol{\alpha}}(\lambda))\} = \frac{1}{n}\sum_{k=1}^{K^*}\sum_{i\in\mathcal{G}_k}\left\{\frac{\partial \ell_i(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\boldsymbol{\alpha}}_k(\lambda))}{\partial \boldsymbol{\zeta}}\right\}\left\{\frac{\partial \ell_i(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\boldsymbol{\alpha}}_k(\lambda))}{\partial \boldsymbol{\zeta}}\right\}^{\top}$$
$$- \frac{1}{n^2}\left\{\sum_{k=1}^{K^*}\sum_{i\in\mathcal{G}_k}\frac{\partial \ell_i(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\boldsymbol{\alpha}}_k(\lambda))}{\partial \boldsymbol{\zeta}}\right\}\left\{\sum_{k=1}^{K^*}\sum_{i\in\mathcal{G}_k}\frac{\partial \ell_i(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\boldsymbol{\alpha}}_k(\lambda))}{\partial \boldsymbol{\zeta}}\right\}^{\top}.$$

The estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}$ can be proved to be consistent with $I_n^{-1}(\boldsymbol{\zeta}^*)$ according to Theorem 3 of Fan and Peng (2004). However, the associated formula above may look somewhat complex because we consider a general setting. To develop intuition for it, let us consider some concrete examples. For instance, when considering the Gaussian model, the estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}$ can be expressed as $(\mathbf{W}^{\top}\mathbf{W})^{-1}\mathbf{W}^{\top}\widehat{\mathbf{E}}\mathbf{W}(\mathbf{W}^{\top}\mathbf{W})^{-1}$, where $\mathbf{W} = (\mathbf{X}, \widetilde{\mathbf{Z}})$, $\widehat{\mathbf{E}} = \text{diag}(\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}^{\top}) - \frac{1}{n}\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}^{\top}$, and $\widehat{\boldsymbol{\varepsilon}} = (y_i - \widehat{\theta}_i, ..., y_n - \widehat{\theta}_n)^{\top}$. Here, $(\mathbf{W}^{\top}\mathbf{W})^{-1}$ evaluates the multicollinearity among predictors, and $\mathbf{W}^{\top}\widehat{\mathbf{E}}\mathbf{W}$ can be interpreted as a weight matrix governed by the correlation between predictors and residuals. Furthermore, in the context of GLMs, the estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}$ has the form of $(\mathbf{W}_g^{\top}\mathbf{W}_g)^{-1}\mathbf{W}^{\top}\widehat{\mathbf{E}}_g\mathbf{W}(\mathbf{W}_g^{\top}\mathbf{W}_g)^{-1}$ with $\mathbf{W}_g = \text{diag}^{1/2}(b''(\widehat{\boldsymbol{\theta}}))\mathbf{W}$, $\widehat{\mathbf{E}}_g = \text{diag}(\widehat{\boldsymbol{\varepsilon}}_g\widehat{\boldsymbol{\varepsilon}}_g^{\top}) - \frac{1}{n}\widehat{\boldsymbol{\varepsilon}}_g\widehat{\boldsymbol{\varepsilon}}_g^{\top}$ and $\widehat{\boldsymbol{\varepsilon}}_g = (y_1 - b'(\widehat{\theta}_1), ..., y_n - b'(\widehat{\theta}_n))^{\top}$, where $b(\cdot)$ is defined in (2).

Furthermore, let us interpret the technical assumptions as follows. The condition of $N_{\min} \gg (p+q)n^{2/3}$ demonstrates the need for sample size in each cluster. The conditions $d_n > a\lambda$ and $\lambda \gg \tau_n$ imply that we need $d_n \gg \tau_n$, where τ_n has the order of $O(N_{\min}^{-1}(p+q)n)$. This presents the requirement of the minimal difference of signals between clusters to consistently recover the true cluster structure. Moreover, $a\lambda = \kappa_n \gg \tau_n$ shows the regularization strength and thresholding level we need to identify the differences between cluster-specific coefficients from different clusters.

Importantly, the performance of TRUE hinges on the choice of the regularization parameter λ_n and the thresholding parameter κ_n. Although many existing criteria including AIC, BIC and cross-validation could be potentially employed to select the tuning parameters, there is no work provides a theoretically guaranteed criterion for supervised clustering in a general framework. Next, we will propose a GIC-type criterion yielding model selection consistency. Before outlining the model selection criterion, we first introduce some notation and definitions. We set a cluster size upper bound, denoted by K_U, with $K^* \leq K_U$. This means that we limit model search to submodels with cluster size no more than K_U. We define $\widetilde{N}_{\min}$ as the minimal cluster size for any candidate model. Moreover, we refer the model equipped with the local estimator given in Theorem 2 as the TRUE model.

THEOREM 3. *(Consistency of model selection) Suppose Assumptions 1-3, and Assumption 4 spelled in the supplement material hold, $\widetilde{N}_{\min} \gg (p+q)n^{1/2}$, and the cluster size upper bound K_U and the dimension*

q are finite. For any sequence ϕ_n satisfying $\phi_n \gg 1/\sqrt{n}$ and $n\phi_n = O(d_n^2 N_{\min})$, minimizing the following information criterion

$$GIC(\lambda) = -\frac{1}{n}\mathcal{L}_n(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\boldsymbol{\gamma}}(\lambda)) + (q\widehat{K}(\lambda) + p)\phi_n \tag{14}$$

will select the TRUE model with probability approaching to 1 as $N_k \to \infty$ for $k = 1, .., K^*$.

In practice, we often choose $\phi_n = C_n \frac{\log\log n}{\sqrt{n}}$ for some constant $C_n > 0$, as alluded to in Section 3.2. In fact, it can be shown that any wrong candidate model must belong to one of three classes: overfitted class, underfitted class, and wrongly-assigned class, as defined in the Supplementary Material. The proof of Theorem 3 demonstrates that the TRUE model achieves the value of (14) smaller than that of any candidate model from the aforementioned three classes, thus ensuring the consistency of model selection.

Theorems 2-3 establish that there exists a local minimizer of the TRUE problem enjoying appealing statistical properties. However, due to the nonconvexity of the TRUE problem, a natural question arises: whether these nice properties can be shared by the computable local minimizer by any algorithm, such as the proposed stagewise TRUE algorithm. The following theorem gives a positive answer.

THEOREM 4. *(Oracle properties of computable local minimizers) Let $\{(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)})\}_{t=1}^{T}$ be a sequence of computable local minimizers of (3), and $(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\boldsymbol{\gamma}}(\lambda))$ the corresponding refined estimator using hierarchical clustering as the refiner. Then under assumptions of Theorems 2-3, if $\widetilde{N}_{\min} = \Omega(N_{\min})$ and $\#\{\|\widehat{\boldsymbol{\gamma}}_i(\lambda) - \widehat{\boldsymbol{\gamma}}_j(\lambda)\|_2 < \kappa_n, i < j\} = o(\lambda_n^{-2}(p+q)n)$, the refined estimator $(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\boldsymbol{\gamma}}(\lambda))$ achieves the same oracle properties as those in Theorem 2, where $\#\{\cdot\}$ represents the cardinality of a set.*

Theorem 4 shows that the local minimizer obtained by the stagewise TRUE algorithm, after a simple refinement, can also enjoy the oracle properties outlined in Theorem 2. It is worth noting that considering hierarchical clustering for refinement in this theorem is only to facilitate our technical analysis. Moreover, the conditions $\widetilde{N}_{\min} = \Omega(N_{\min})$ and $\#\{\|\widehat{\boldsymbol{\gamma}}_i(\lambda) - \widehat{\boldsymbol{\gamma}}_j(\lambda)\|_2 < \kappa_n, i < j\} = o(\lambda_n^{-2}(p+q)n)$ represent requirements on the sparsity of the local minimizer, which are mild and sensible as the stagewise TRUE algorithm iteratively drops pairs $(\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_j)$ that are distant from each other during the computation process.

5. Simulation studies

In this section, we conduct numerical experiments to assess the finite-sample performance of our proposed method and algorithm. To be specific, we examine the path convergence of the stagewise algorithm, estimation accuracy, the impact of step size, and computational efficiency using simulated data. For comparisons, we introduce two sets of benchmarks: computational benchmarks and methodological benchmarks. The first set includes the ADMM and Trust Region Optimization (TRO) (Yuan 2000, 2015) algorithms, where

ADMM is guaranteed to converge (see the Supplementary Material for details). The second set consists of the following four approaches: the concave fusion approach (Concave) (Ma and Huang 2017, Ma et al. 2020), the weighted L_1 fusion approach (Weighted) (Chen et al. 2021), the Gaussian Mixture Model (GMM) (Reynolds 2009), and the Wasserstein Barycenter-enhanced Gaussian Mixture Model (Wasserstein-GMM) (Delon and Desolneux 2020, Lin et al. 2023). In particular, we utilize both the MCP and SCAD penalties for the concave fusion approach, but only report the results of the one that performs better in each example. In the weighted L_1 fusion approach, the penalty is

$$\lambda_n \sum_{i<j} p(\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|_2) = \lambda_n \sum_{i<j} \omega_{ij} \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|_2,$$

where ω_{ij} is adaptive weight required to be specified. Borrowing the idea in Chen et al. (2021), we set $\omega_{ij} = \min\{B_\omega, \frac{\iota_{ij}^m}{\|\widetilde{\boldsymbol{\gamma}}_i - \widetilde{\boldsymbol{\gamma}}_j\|_2}\}$, where ι_{ij}^m indicates whether the observation j is among i's m-nearest neighbors defined by the Euclidean distance, $\widetilde{\boldsymbol{\gamma}}$ is a preliminary estimate of $\boldsymbol{\gamma}^*$, and B_ω is a constant employed to control the value of ι_{ij}^m in the case the pair $(\widetilde{\boldsymbol{\gamma}}_i, \widetilde{\boldsymbol{\gamma}}_j)$ have values too close to each other. In our studies, we construct $\widetilde{\boldsymbol{\gamma}}$ by the local regression, and we use $B_\omega = 10$. These two methods can be both implemented by our proposed ADMM algorithm. Moreover, GMM and Wasserstein-GMM are both computed by the EM algorithm. Note that Wasserstein-GMM can be regarded as a modified version of GMM, which incorporates the Wasserstein barycenter to compute cluster centers to enhance robustness. Last but not least, for the selection of the optimal tuning parameter, we apply the GIC criterion in (9) for all the methods.

5.1. Simulations

We evaluate the proposed method using six different simulated examples. Specifically, Examples 1-3 are similar to the simulations in Chen et al. (2021). These examples are all linear models and have distinct settings according to different numbers of clusters or cluster boundaries. In particular, the different settings of the same linear model are designed to quantify the performance of the proposed method in solving problems of varying difficulty levels. Examples 4 and 5 are generated by two generalized linear models: the logistic model and the Poisson model, respectively. In addition, Example 6 is simulated by the Tobit model. It is worth pointing out that these simulation examples are not arbitrary; they each hold potential value for practical applications, as discussed in detail in Section A. In all these examples, the covariate vector that captures population-shared effects, $\mathbf{x}_i$, is generated by the multivaraite normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, in which $\boldsymbol{\Sigma} = \{\sigma_{jj'}\}$, $\sigma_{jj} = 1$ and $\sigma_{jj'} = 0.3$ for $j \neq j'$. However, the covariate vector with individual effects, $\mathbf{z}_i$, is generated in different ways. Moreover, all the simulations are realized for 100 times with the sample size of $n = 300$.

Example 1 ***(Linear model I)*** *Two clusters linear model with continuous* $\mathbf{z}_i$. *Suppose the underlying true model is linear with:*

$$y_i = \begin{cases} \mathbf{x}_i^\top \boldsymbol{\beta}^* + 1 + 2z_{i1} + \varepsilon_i, & z_{i2} \leq 0, \\ \mathbf{x}_i^\top \boldsymbol{\beta}^* - 1 - 2z_{i1} + \varepsilon_i, & z_{i2} > 0, \end{cases}$$

where $\boldsymbol{\beta}^* = (1,-2,3,-5,2)^\top$, z_{i1} is generated from the uniform distribution $U(0,2)$, z_{i2} is generated from the uniform distribution $U(-1,1)$, and the random noise $\varepsilon_i \sim N(0,0.5^2)$.

Example 2 ***(Linear model II)*** *Three clusters linear model with mixed-type $\mathbf{z}_i$. The true model has a higher dimension of $\mathbf{x}_i$ with noisy variables:*

$$y_i = \begin{cases} \mathbf{x}_i^\top \boldsymbol{\beta}^* + 2 + 3z_{i1} + \varepsilon_i, & z_{i3} \le 0 \,\text{and}\, z_{i2} = 1, \\ \mathbf{x}_i^\top \boldsymbol{\beta}^* + z_{i1} + \varepsilon_i, & z_{i3} > 0 \,\text{and}\, z_{i2} = 1, \\ \mathbf{x}_i^\top \boldsymbol{\beta}^* - 2 - 2z_{i1} + \varepsilon_i, & z_{i2} = 0, \end{cases}$$

where $\boldsymbol{\beta}^* = (1,-2,3,-5,2,0,0,0)^\top$, z_{i1} and z_{i2} are generated from the Bernoulli distribution with success probability 2/3, z_{i3} is generated from the uniform distribution $U(-1,1)$, and the random noise $\varepsilon_i \sim N(0,0.5^2)$.

Example 3 ***(Linear model III)*** *Three clusters linear model with nonlinear cluster boundaries. The true model has a higher dimension of $\mathbf{z}_i$:*

$$y_i = \begin{cases} \mathbf{x}_i^\top \boldsymbol{\beta}^* + 2 + 3z_{i1} + 3z_{i2} + \varepsilon_i, & z_{i1}^2 + \exp(z_{i2}) \le 1.5, \\ \mathbf{x}_i^\top \boldsymbol{\beta}^* + z_{i1} + z_{i2} + \varepsilon_i, & z_{i1}^2 + \exp(z_{i2}) > 3.5, \\ \mathbf{x}_i^\top \boldsymbol{\beta}^* - 2 - 2z_{i1} - z_{i2} + \varepsilon_i, & \text{otherwise}, \end{cases}$$

where $\boldsymbol{\beta}^* = (1,-5,3,-2,1)^\top$, all the z_{ij} are generated from the uniform distribution $U(-2,2)$, and the random noise $\varepsilon_i \sim N(0,0.5^2)$.

Example 4 ***(Logistic model)*** *Three clusters logistic model with mixed-type $\mathbf{z}_i$. The response y_i is generated from the Bernoulli distribution with conditional success probability $\frac{\exp(\theta_i^*)}{1+\exp(\theta_i^*)}$ for*

$$\theta_i^* = \begin{cases} \mathbf{x}_i^\top \boldsymbol{\beta}^* + 3 - 3z_{i1}, & z_{i1} \le 0 \,\text{and}\, z_{i2} = 1, \\ \mathbf{x}_i^\top \boldsymbol{\beta}^* - 1 - z_{i1}, & z_{i1} > 0 \,\text{and}\, z_{i2} = 1, \\ \mathbf{x}_i^\top \boldsymbol{\beta}^* - 3 + 3z_{i1}, & z_{i2} = 0, \end{cases}$$

where $\boldsymbol{\beta}^* = (1,-2,3,-5,2)^\top$, z_{i1} is generated from the uniform distribution $U(-1,1)$, and z_{i2} is generated from the Bernoulli distribution with success probability 2/3.

Example 5 ***(Poisson model)*** *Three clusters Poisson model with mixed-type $\mathbf{z}_i$. The response y_i is generated from the Poisson distribution with conditional mean $\exp(\theta_i^*)$:*

$$\theta_i^* = \begin{cases} \mathbf{x}_i^\top \boldsymbol{\beta}^* + 3 - 3z_{i1}, & z_{i1} \le 0 \,\text{and}\, z_{i2} = 0, \\ \mathbf{x}_i^\top \boldsymbol{\beta}^* - 1 - z_{i1}, & z_{i1} > 0 \,\text{and}\, z_{i2} = 0, \\ \mathbf{x}_i^\top \boldsymbol{\beta}^* - 3 + 3z_{i1}, & z_{i2} = 1, \end{cases}$$

where $\boldsymbol{\beta}^* = (2,-3,4,1,-3)^\top$, z_{i1} is generated from the uniform distribution $U(-1,1)$, and z_{i2} is generated from the Bernoulli distribution with success probability 2/3.

Example 6 ***(Tobit model)*** *Two clusters Tobit model with mixed-type $\mathbf{z}_i$. Suppose the underlying true model is:*

$$y_i^* = \begin{cases} \mathbf{x}_i^\top \boldsymbol{\beta}^* + 3 - 2z_{i1} + \varepsilon_i, & z_{i1} \le 0, \\ \mathbf{x}_i^\top \boldsymbol{\beta}^* + 2z_{i1} + \varepsilon_i, & z_{i1} > 0, \end{cases}$$

where $\boldsymbol{\beta}^* = (1, -2, 3, -5, 2)^\top$, z_{i1} is generated from the uniform distribution $U(-2,2)$, and the random noise $\varepsilon_i \sim N(0, 0.5^2)$. The observed responses are left-censored, i.e., $y_i = \max\{y_i^*, 0\}$. Moreover, the Tobit log-likelihood is given by

$$\mathcal{L}_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2) = \sum_{i=1}^n \left[d_i \left\{ -\log \sigma - \frac{1}{2\sigma^2}(y_i - \theta_i)^2 \right\} + (1 - d_i) \log \left\{ \Phi\left(\frac{-\theta_i}{\sigma} \right) \right\} \right],$$

where σ^2 is the variance parameter should be estimated, $d_i = \mathbf{1}_{\{y_i > 0\}}$ are the indicator variables, $\theta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{\gamma}_i$, and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

5.2. Path convergence of stagewise TRUE

We compare the stagewise path of TRUE with the solution path obtained by the ADMM to demonstrate the path convergence of our stagewise algorithm. We use the linear model I in Example 1 with the sample size $n = 100$. The results are illustrated in Figure 2. Note that the optimization of TRUE is nonconvex, so its solution path along λ_n is not unique and the paths by the two algorithms may be not identical. However, the stagewise paths of TRUE exhibit very similar merging trends and cluster outcomes as the ADMM solution paths. Therefore, we conclude that our proposed stagewise algorithm can produce nice solution paths of TRUE.

5.3. Estimation performance

For the accuracy comparison, we report the averages and standard deviations of mean squared errors of the estimates for $\boldsymbol{\beta}^*$ and $\boldsymbol{\gamma}^*$. In addition, we also compare the estimated number of clusters $\widehat{K}$. All the simulations are carried out for 100 times with the sample size of $n = 300$. We report the results of all the methods in Table 1. In general, the proposed method exhibits superior estimation and clustering performance compared to other competing methods. Among the TRUE estimates computed by different algorithms, the disparity between the stagewise and ADMM performance is not pronounced, whereas the performance of TRO lags significantly behind them. The unsatisfactory performance of the TRO algorithm arises from its poor adaptability to nonconvex fusion regularization and the complexity of hyperparameter tuning. In addition, the weighted L_1 approach performs the worst, as a suboptimal choice of the weights can dramatically influence the quality of the clustering solution. Unfortunately, there is no weight specification strategy that is adequately applicable to all possible model settings. Furthermore, since GMM can only deal with Gaussian data, we apply GMM and Wasserstein-GMM exclusively to Examples 1–3. These two methods perform worse than the proposed TRUE approach, as their stricter modeling assumptions prevent them from accurately capturing both population effects and heterogeneity effects.

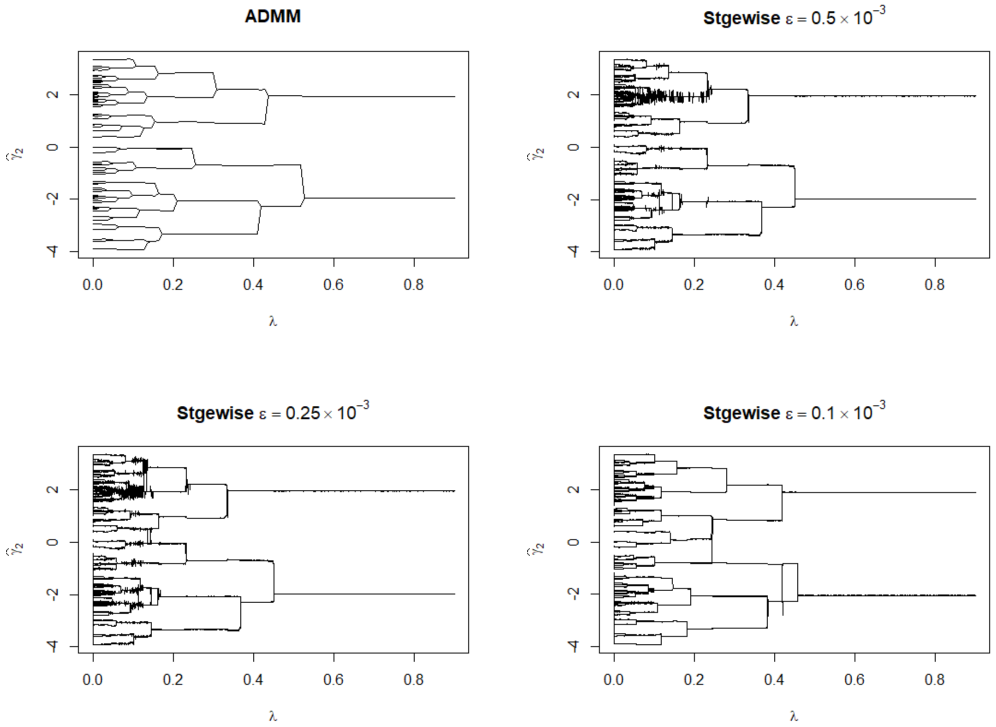

Figure 2 **Solution paths of $\widehat{\gamma}$ produced by the ADMM and satgewise algorithms.**

Moreover, we conduct statistical inference on the difference between clusters based on the asymptotic normality developed in Theorem 2. Following Ma and Huang (2017), we report the average p-values for testing coefficient differences between different clusters. In this experiment, we only consider our proposed TRUE method, as the other methods used for comparison do not have theoretical guarantees for statistical inference. As we can see in the following Table 2, the results show that the p-values are close to zero in all cases, thus further confirming the inter-class differences through the inference aspect.

5.4. Impact of step size

We investigate the impact of the step size ϵ in stagewise learning on the performance of the proposed methods. We use the linear model I in Example 1 for instance. We consider different step sizes, i.e., ϵ in $\{0.125, 0.25, 0.375, 0.5, 0.625\} \times 10^{-3}$. The experiment is replicated 100 times. Figure 3 shows the boxplots of the estimation errors for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\gamma}$, and computation time. The performance of the stagewise methods is stabilized when ϵ is small enough, i.e., $\epsilon \leq 0.25 \times 10^{-3}$ in this example. But the computational cost will increase if ϵ is too small and it is clear that there is a tradeoff between stepsize and accuracy. In practice, we suggest to conduct some pilot numerical analysis to identify a proper step size.

5.5. Computational efficiency

We report the computational time of our proposed method as the function of the sample size n, which is the main factor affecting the computational efficiency. The data are also simulated from the linear model I in

Table 1 Simulation results: The averages and standard deviations of mean squared errors with the best results in bold.

Case	TRUE-stagewise			TRUE-ADMM			TRUE-TRO			Concave		
	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\gamma}}$	$\widehat{K}$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\gamma}}$	$\widehat{K}$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\gamma}}$	$\widehat{K}$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\gamma}}$	$\widehat{K}$
1	0.162	**0.008**	**2.04**	**0.156**	0.011	2.08	0.185	0.057	2.57	0.170	0.015	2.06
	(0.023)	(0.001)	(0.26)	(0.018)	(0.002)	(0.32)	(0.068)	(0.014)	(0.77)	(0.037)	(0.002)	(0.30)
2	**0.221**	**0.012**	3.10	0.234	0.015	**3.08**	0.278	0.069	3.94	0.245	0.015	3.24
	(0.033)	(0.005)	(0.48)	(0.028)	(0.003)	(0.32)	(0.099)	(0.021)	(1.08)	(0.041)	(0.004)	(0.33)
3	**0.801**	0.134	**3.25**	0.814	**0.114**	3.33	0.864	0.165	4.14	0.855	0.165	3.30
	(0.139)	(0.022)	(0.36)	(0.127)	(0.035)	(0.41)	(0.191)	(0.184)	(1.25)	(0.142)	(0.019)	(0.28)
4	0.301	0.026	**3.14**	**0.299**	**0.020**	3.18	0.365	0.041	3.55	0.331	0.036	3.24
	(0.056)	(0.011)	(0.24)	(0.062)	(0.009)	(0.34)	(0.127)	(0.035)	(0.88)	(0.063)	(0.012)	(0.69)
5	0.269	0.021	**3.10**	**0.256**	**0.019**	3.11	0.321	0.033	3.34	0.277	0.031	3.15
	(0.034)	(0.008)	(0.36)	(0.031)	(0.007)	(0.24)	(0.094)	(0.029)	(0.71)	(0.028)	(0.014)	(0.32)
6	**0.398**	**0.041**	**2.20**	0.401	0.052	2.22	0.474	0.118	3.12	0.417	0.066	2.41
	(0.085)	(0.019)	(0.32)	(0.076)	(0.021)	(0.22)	(0.157)	(0.087)	(0.87)	(0.096)	(0.032)	(0.54)

Case	Weighted			GMM			Wasserstein-GMM		
	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\gamma}}$	$\widehat{K}$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\gamma}}$	$\widehat{K}$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\gamma}}$	$\widehat{K}$
1	0.195	0.024	2.24	0.182	0.023	2.34	0.179	0.022	2.41
	(0.011)	(0.010)	(0.55)	(0.026)	(0.003)	(0.24)	(0.029)	(0.002)	(0.31)
2	0.301	0.022	3.54	0.237	0.026	3.44	0.241	0.029	3.47
	(0.045)	(0.014)	(0.68)	(0.033)	(0.005)	(0.71)	(0.029)	(0.007)	(0.66)
3	1.023	0.345	3.62	0.911	0.154	3.87	0.907	0.142	3.73
	(0.215)	(0.103)	(0.62)	(0.174)	(0.041)	(0.88)	(0.166)	(0.038)	(0.82)
4	0.451	0.124	4.21	—	—	—	—	—	—
	(0.105)	(0.098)	(1.25)	—	—	—	—	—	—
5	0.411	0.134	4.13	—	—	—	—	—	—
	(0.112)	(0.112)	(1.01)	—	—	—	—	—	—
6	0.641	0.214	3.31	—	—	—	—	—	—
	(0.326)	(0.198	(0.99)	—	—	—	—	—	—

Table 2 Simulation results: The averages of p-values for pairwise comparison testing.

Case	TRUE-stagewise			TRUE-ADMM		
	1 vs. 2	1 vs. 3	2 vs. 3	1 vs. 2	1 vs. 3	2 vs. 3
1	<0.001	—	—	<0.001	—	—
2	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
3	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
4	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
5	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
6	<0.001	—	—	<0.001	—	—

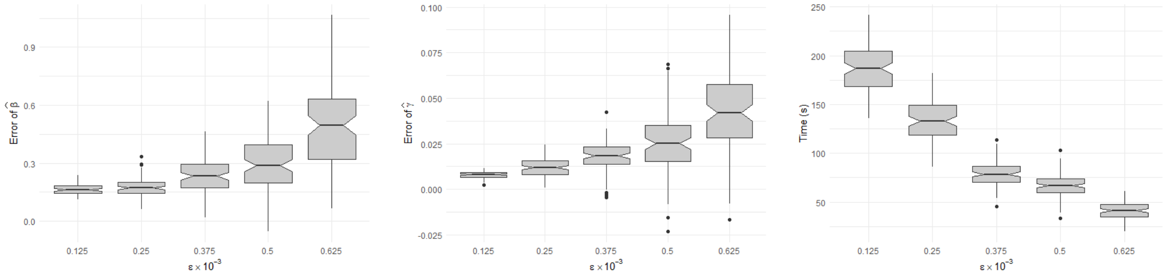

Figure 3 Impact of the step size ϵ on the performance of the stagewise learning method.

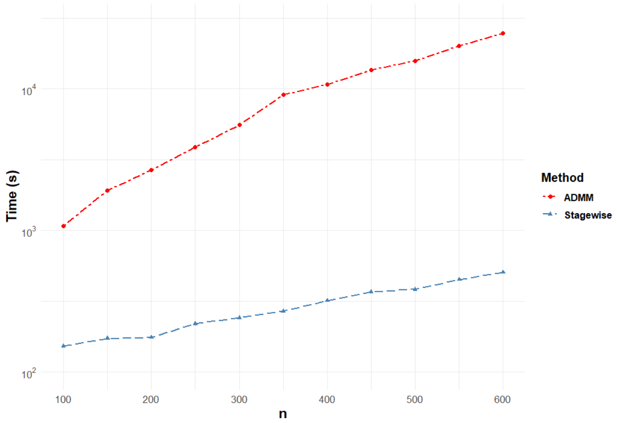

Figure 4 **Computation cost with the increasing sample size n.**

Example 1. We let n vary from 100 to 600 and each experiment is repeated 100 times. We consider both the stagewise algorithm and the ADMM algorithm. In particular, we align the grids of the tuning parameter λ_n for the ADMM with the paths traced by the stagewise solutions. The mean values of runtime for the two methods are illustrated in Figure 4. It is evident that the stagewise algorithm is scalable to large-sample problems, and the gain over the ADMM is dramatic.

6. Real data applications

Investigating the factors affecting insurance claim frequency is a crucial aspect of insurance management. Given the large customer base, heterogeneity often arises in distributions among different customer groups. This drives us to explore the potential latent heterogeneous factors affecting insurance claim frequency and the individual effects of some certain covariates, thereby facilitating the development of personalized insurance planning. In this section, we illustrate the usefulness of our proposed methodology by analyzing the motor insurance claim frequency data on 2812 individuals in one year period. This dataset is available from the SAS Enterprise Miner database and was analyzed in many researches such as Yip and Yau (2005), Tang et al. (2014), and Chen et al. (2019). The dataset contains information including claim profiles, policy details, driving records and personal particulars of policyholders. In the insurance industry, the no claim discount system, which is widely adopted by automobile insurers, leads to excessive zero claims because policyholders seldom make a claim if the loss is small. This phenomenon can be observed from this dataset: among the 2812 observations, 60.7% or 1706 individuals have no claim; 12.3% or 351 individuals have one claim; 14.5% or 408 individuals have two claims, and 12.3% or 347 individuals have three or more claims. To model the excessive zeros phenomenon in claim frequency distribution, we consider the heterogeneous zero-inflated Poisson (ZIP) model (Lambert 1992). Specifically, the response variable y_i, the insured claim frequency of the ith individual, follows

$$y_i \sim \begin{cases} 0, & \text{with probability}\, p, \\ \text{Poisson}(\theta_i^*), & \text{with probability}\, 1-p, \end{cases}$$

Table 3 Coding of risk factors.

Variables	Description
x_1	gender: 0 = female, 1 = male
x_2	marriage status: 0 = not married, 1 = married
x_3	driving area: 0 = suburban, 1 = urban
x_4	income of the policyholder
z_1	car use: 0 = private, 1 = business

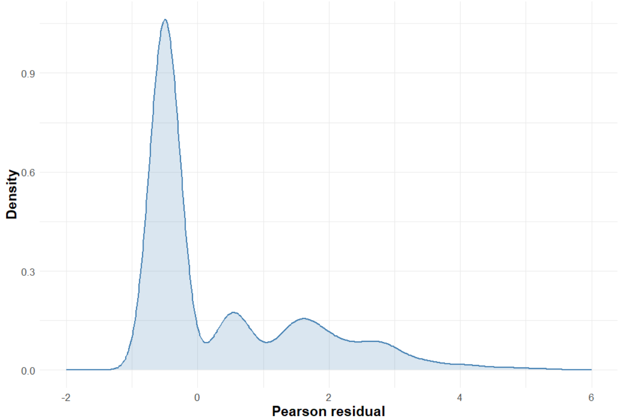

Figure 5 Kernel density plot of the Pearson residuals of the homogeneous ZIP model.

where p is a constant and $\log(\theta_i^*) = \mathbf{x}_i^\top \boldsymbol{\beta}^* + \mathbf{z}_i^\top \boldsymbol{\gamma}_i^*$. In terms of covariates, we adopt the five risk factors selected by Yip and Yau (2005), which are summarized in Table 3. In particular, we treat the binary of car use as a covariate with heterogeneous effects, and other factors as baseline covariates. That is, our goal is to identify if there are some individual effects of car use on motor insurance claims.

To see possible heterogeneous effects, we first fit the homogeneous ZIP model by the all the five factors, so that the effects of the baseline covariates are controlled. Then, we plot the kernel density estimates of the Pearson residuals of the fitted model in Figure 5. The Pearson residuals are defined as

$$r_i = \frac{y_i - \widehat{\mathbb{E}(y_i)}}{\sqrt{\widehat{\mathrm{Var}(y_i)}}},$$

where for the ZIP model, $\widehat{\mathbb{E}(y_i)} = (1-\widehat{p})\exp(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}} + \mathbf{z}_i^\top \widehat{\boldsymbol{\gamma}})$, $\widehat{\mathrm{Var}(y_i)} = \widehat{\mathbb{E}(y_i)}(1+\exp(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}} + \mathbf{z}_i^\top \widehat{\boldsymbol{\gamma}}) - \widehat{\mathbb{E}(y_i)})$, and $\widehat{p}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ are the maximum likelihood estimates (MLEs). It is clearly seen that after adjusting for the effects of covariates, the distribution of residuals still shows multiple modes. This indicates some potential heterogeneous, which motivates the use of the heterogeneous model.

In applying the proposed TRUE, the continuous covariates are standardized so that the coefficients are comparable. Also, the tuning parameter λ_n and κ_n are selected according to the GIC in (9). We consider the estimates produced by both the stagewise and ADMM algorithms. For comparison, the MLEs obtained by the homogeneous ZIP model are also involved. The main results are reported in Table 4, including the estimated coefficients (Coef), the approximate standard errors (s.e.), and the corresponding p-values,

Table 4 Main results of the fitted ZIP models by MLE and the proposed methods.

	MLE			TRUE-stagewise			TRUE-ADMM		
	Coef	s.e.	p-value	Coef	s.e.	p-value	Coef	s.e.	p-value
Intercept-1	-2.191	0.236	$<$0.001	-0.512	0.089	$<$0.001	-0.524	0.090	$<$0.001
Intercept-2	—	—	—	-3.365	0.245	$<$0.001	-3.358	0.239	$<$0.001
z_1-1	0.149	0.047	0.002	0.321	0.073	$<$0.001	0.318	0.071	$<$0.001
z_1-2	—	—	—	0.086	0.021	$<$0.001	0.091	0.026	$<$0.001
x_1	-0.051	0.022	0.020	0.074	0.018	$<$0.001	0.069	0.017	$<$0.001
x_2	-0.111	0.078	0.155	-0.082	0.037	0.027	-0.065	0.028	0.020
x_3	1.230	0.153	$<$0.001	1.231	0.095	$<$0.001	1.240	0.101	$<$0.001
x_4	-0.017	0.003	$<$0.001	-0.058	0.008	$<$0.001	-0.063	0.009	$<$0.001
p	0.447	—	—	≈ 0	—	—	≈ 0	—	—
GIC	3152.368			113.625			113.214		

estimated zero-inflation probability p, and the GIC values. It is worth noting that the standard errors and p-values of our proposed estimates are calculated according to Theorem 2. The results show that two major clusters are identified. To interpret the results, we categorize the insureds to clusters with high and low risk, according to the large and small value of the estimated intercepts, respectively. Moreover, car use has significantly heterogeneous effects on motor insurance claims. Another interesting finding is that after considering the heterogeneity in the data the estimated zero-inflation probability is close to 0. This indicates that the excessive zeros can be almost explained by clustering with individualized effects. On the other hand, our proposed method promotes interpretability of the variables which truly have affects on motor insurance claims. To be specific, using TRUE, the covariates x_1 and x_2 are more significant than that by the classical MLE. Last not but least, the GIC values also suggest that TRUE make a great improvement over the homogeneous ZIP model.

7. Discussion

In this paper, we targeted the problem of supervised clustering and established a general heterogeneity tracking model for studying this topic. Based on this new model, we proposed a new method via the truncated fusion learning and developed a novel fusion stagewise algorithm to compute the entire solution paths. Theoretical results and numerical studies demonstrate the statistical and computational accuracy of our proposed method. Through addressing the core issues of supervised clustering, TRUE can be flexibly extended to various modern scientific topics, including the identification of heterogeneous treatment effects, personalized federated learning, and heterogeneous environment online learning.

References

Andrews JL, McNicholas PD (2012) Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Stat. Comput.* 22(5):1021–1029.

Antoniadis A (1997) Wavelets in statistics: a review. *J. Ital. Statist. Soc.* 6:97–130.

Banfield JD, Raftery AE (1993) Model-based gaussian and non-gaussian clustering. *Biometrics* 49(3):803–821.

Bertsekas DP (1997) Nonlinear programming. *J. Oper. Res. Soc.* 48(3):334–334.

Chaudhuri A (2014) Modified fuzzy support vector machine for credit approval classification. *AI Commun.* 27(2):189–211.

Chen EY, Song R, Jordan MI (2024) Reinforcement learning in latent heterogeneous environments. *J. Am. Stat. Assoc.* 119(548):3113–3126.

Chen J, Tran-Dinh Q, Kosorok MR, Liu Y (2021) Identifying heterogeneous effect using latent supervised clustering with adaptive fusion. *J. Comput. Graph. Stat.* 30(1):43–54.

Chen K, Dong R, Xu W, Zheng Z (2022) Fast stagewise sparse factor regression. *J. Mach. Learn. Res.* 23(271):1–45.

Chen K, Huang R, Chan NH, Yau CY (2019) Subgroup analysis of zero-inflated poisson regression model with applications to insurance data. *Insur. Math. Econ.* 86:8–18.

Delon J, Desolneux A (2020) A wasserstein-type distance in the space of gaussian mixture models. *SIAM J. Imaging Sci.* 13(2):936–970.

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. *Ann. Statist.* 32(1):407–499.

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* 96(456):1348–1360.

Fan J, Peng H (2004) Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Stat.* 32(3):928–961.

Fan Y, Demirkaya E, Li G, Lv J (2019) Rank: Large-scale inference with graphical nonlinear knockoffs. *J. Am. Stat. Assoc.* 115(529):362–379.

Fan Y, Lv J (2013) Asymptotic equivalence of regularization methods in thresholded parameter space. *J. Am. Stat. Assoc.* 108(503):1044–1061.

Gandhi RT, Tashima KT, Smeaton LM, Vu V, Ritz J, Andrade A, Eron JJ, Hogg E, Fichtenbaum CJ (2020) Long-term outcomes in a large randomized trial of hiv-1 salvage therapy: 96-week results of aids clinical trials group a5241 (options). *J. Infect. Dis.* 221(9):1407–1415.

Ghosh A, Chung J, Yin D, Ramchandran K (2020) An efficient framework for clustered federated learning. *NeurIPS* 33:19586–19597.

Greenland S (2009) Interactions in epidemiology: relevance, identification, and estimation. *Epidemiology* 20(1):14–17.

Guotai C, Abedin MZ, Moula FE (2017) Modeling credit approval data with neural networks: an experimental investigation and optimization. *J. Bus. Econ. Manag.* 18(2):224–240.

Hastie T, Tibshirani R (1996) Discriminant analysis by gaussian mixtures. *J. R. Stat. Soc. Series B. Stat. Methodol.* 58(1):155–176.

Hastie TJ, Pregibon D (2017) Generalized linear models. *Statistical models in S* (Routledge).

He L, Chen K, Xu W, Zhou J, Wang F (2018) Boosted sparse and low-rank tensor regression. *NeurIPS* 31:1017–1026.

Hu X, Huang J, Liu L, Sun D, Zhao X (2021) Subgroup analysis in the heterogeneous cox model. *Stat. Med.* 40(3):739–757.

Jeon JJ, Kwon S, Choi H (2017) Homogeneity detection for the high-dimensional generalized linear model. *Comput. Stat. Data Anal.* 114:61–74.

Lambert D (1992) Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics* 34(1):1–14.

Lehmann EL, Casella G (2006) *Theory of point estimation* (Springer Science & Business Media).

Li F, Sang H (2019) Spatial homogeneity pursuit of regression coefficients for large datasets. *J. Am. Stat. Assoc.* 114(527):1050–1062.

Lin L, Shi W, Ye J, Li J (2023) Multisource single-cell data integration by maw barycenter for gaussian mixture models. *Biometrics* 79(2):866–877.

Liu W, Mao X, Zhang X, Zhang X (2024) Robust personalized federated learning with sparse penalization. *J. Am. Stat. Assoc.* 120(549):266–277.

Liu Y, Wu Y (2007) Variable selection via a combination of the l_0 and l_1 penalties. *J. Comput. Graph. Stat.* 16(4):782–798.

Ma S, Huang J (2017) A concave pairwise fusion approach to subgroup analysis. *J. Am. Stat. Assoc.* 112(517):410–423.

Ma S, Huang J, Zhang Z, Liu M (2020) Exploration of heterogeneous treatment effects via concave fusion. *Int. J. Biostat.* 16(1):1–26.

McCullagh P (1980) Regression models for ordinal data. *J. R. Stat. Soc. Series B. Stat. Methodol.* 42(2):109–127.

McNicholas PD (2010) Model-based classification using latent gaussian mixture models. *J. Stat. Plan. Inference.* 140(5):1175–1181.

Reynolds DA (2009) Gaussian mixture models. *Encyclopedia of biometrics* 741(659-663):3.

Runchi Z, Liguo X, Qin W (2023) An ensemble credit scoring model based on logistic regression with heterogeneous balancing and weighting effects. *Expert Syst. Appl.* 212:118732.

Sattler F, Müller KR, Samek W (2020) Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Trans. Neural Netw. Learn. Syst.* 32(8):3710–3722.

Shen J, He X (2015) Inference for subgroup analysis with a structured logistic-normal mixture model. *J. Am. Stat. Assoc.* 110(509):303–312.

Shen X, Huang HC (2010) Grouping pursuit through a regularization solution surface. *J. Am. Stat. Assoc.* 105(490):727–739.

Su X, Fan J, Levine RA, Nunn ME, Tsai CL (2018) Sparse estimation of generalized linear models (glm) via approximated information criteria. *Stat. Sin.* 28(3):1561–1581.

Tang Y, Xiang L, Zhu Z (2014) Risk factor selection in rate making: Em adaptive lasso for zero-inflated poisson regression models. *Risk Anal.* 34(6):1112–1127.

Tibshirani RJ (2015) A general framework for fast stagewise algorithms. *J. Mach. Learn. Res.* 16(1):2543–2588.

Tobin J (1958) Estimation of relationships for limited dependent variables. *Econometrica* 24–36.

Vaughan G, Aseltine R, Chen K, Yan J (2017) Stagewise generalized estimating equations with grouped variables. *Biometrics* 73(4):1332–1342.

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* 113(523):1228–1242.

Wang M, Yao T, Allen GI (2023) Supervised convex clustering. *Biometrics* 79(4):3846–3858.

Wang W, Su L (2021) Identifying latent group structures in nonlinear panels. *J. Econom.* 220(2):272–295.

Wei LJ (1992) The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Stat. Med.* 11(14-15):1871–1879.

Wei S, Kosorok MR (2013) Latent supervised learning. *J. Am. Stat. Assoc.* 108(503):957–970.

Yip KC, Yau KK (2005) On modeling claim frequency data in general insurance with extra zeros. *Insur. Math. Econ.* 2(36):153–163.

Yuan YX (2000) A review of trust region algorithms for optimization. *Iciam*, volume 99, 271–282.

Yuan YX (2015) Recent advances in trust region algorithms. *Math. Program.* 151:249–281.

Zhang C (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* 38(2):894–942.

Zhang X, Liu J, Zhu Z (2024) Learning coefficient heterogeneity over networks: A distributed spanning-tree-based fused-lasso regression. *J. Am. Stat. Assoc.* 119(545):485–497.

Zhao P, Yu B (2007) Stagewise lasso. *J. Mach. Learn. Res.* 8:2701–2726.

Zheng Z, Bahadori MT, Liu Y, Lv J (2019) Scalable interpretable multi-response regression via seed. *J. Mach. Learn. Res.* 20(107):1–34.

Zhou X (2018) On the fenchel duality between strong convexity and lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573* .

Supplementary Material to "Fusion learning on supervised clustering and its fast stagewise algorithm"

Letian Li, Yang Li*, Jie Zhang, Zemin Zheng*

International Institute of Finance, The School of Management

University of Science and Technology of China

Hefei, Anhui 230026, P. R. China

This Supplementary Material consists of three parts. Section B provides the details of the ADMM algorithm for solving TRUE. Section D lists the key lemmas and presents the proofs for main results. Additional technical proofs for the lemmas are provided in Section E. All the notations are the same as defined in the main body of the paper.

A. Application examples and managerial implications

The proposed model can be applied in a wide range of applications. In what follows, we list several common application examples regarding different data types of the response and discuss the corresponding managerial implications.

EXAMPLE 1. **(Continuous data in healthcare)** Exploring heterogeneous treatment effects in clinical trials is a key objective of precision medicine. It is generally understood that patients' medical indicators are influenced by their biological traits and the medical treatments applied, while the same treatments may yield distinct effects across different patient groups. For instance, in the AIDS Clinical Trials Group Study Ma et al. (2020), the problem can be modeled as $y_i = \theta_i^* + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta}^* + z_i\gamma_i^* + \varepsilon_i$, where the response y_i is the log-transformed value of the CD4 counts of the ith patient, the covariate vector $\mathbf{x}_i$ contains the ith patient's baseline characteristics such as age, weight and Karnofsky score, the binary variable z_i denotes the treatment for the ith patient, and ε_i is the random noise. If the noise distribution is Gaussian, the conditional likelihood can be expressed as $f(y_i;\theta_i^*,\phi^*) = \frac{1}{\sqrt{2\pi\phi^*}}\exp\{-\frac{1}{2\phi^*}(y_i-\theta_i^*)^2\}$. In some complex cases where the association can not be adequately captured by the linear relationship, a nonlinear form $y_i = g(\theta_i^*) + \varepsilon_i$ can be applied, where $g(\cdot)$ could be, for example, a polynomial function.

In particular, due to the technical limitations of the assays employed for certain clinical measurements, the observed data occasionally are censored. For example, in the OPTIONS trial Gandhi et al. (2020), the aim is to investigate the association between HIV viral load and mutations in the virus's genome to identify drug-resistant mutations and quantify the degree of resistance that these mutations confer to different antiretroviral therapy treatments. Given that the HIV viral load is scarcely measurable when it is below the threshold of 50 copies/ml, the response is left-censored. Borrowing the idea of the Tobit model, this problem can be modeled by a partial likelihood $f(y_i;\theta_i^*,\phi^*) = [\frac{1}{\sqrt{2\pi\phi^*}}\exp\{-\frac{1}{2\phi^*}(y_i-\theta_i^*)^2\}]^{d_i}[\Phi(-\theta_i^*/\sqrt{\phi^*})]^{1-d_i}$, where $d_i = \mathbf{1}_{\{y_i\geq 50\}}$ is the indicator variable and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. In these medical applications, by recovering the coefficients $\boldsymbol{\beta}^*$ and γ_i^*, experts can personalize treatments for each patient based on their expected benefits from different treatments, thereby enhancing therapeutic efficacy.

EXAMPLE 2. **(Categorical data in financial risk management)** In the field of financial risk management, data science methodologies play a pivotal role in enabling intelligent decision-making. A key challenge in this field involves handling heterogeneous/individualized datasets. For instance, in customer credit scoring, an individual's credit risk level is influenced by factors such as financial status, credit history, and demographic characteristics. These influences can vary significantly across individuals, reflecting personalized effects Runchi et al. (2023). Let y_i be the binary variable indicating whether the customer has a high credit risk, $\mathbf{x}_i$ represent the attribute covariates, and $\mathbf{z}_i$ include some factors carrying individualized effects, such as tendency to prepay, or unobserved individual-specific intercepts. Then this problem can be modeled by a logistic likelihood $f(y_i;\theta_i^*) = (\frac{e^{\theta_i^*}}{1+e^{\theta_i^*}})^{y_i}(\frac{1}{1+e^{\theta_i^*}})^{1-y_i}$, where $\theta_i^* = \mathbf{x}_i^\top\boldsymbol{\beta}^* + \mathbf{z}_i^\top\boldsymbol{\gamma}_i^*$. When y_i takes on more than two levels or exhibits an ordinal structure, the model can be naturally extended based on a multinomial logistic likelihood or an ordinal logistic likelihood. Estimating the coefficients $\boldsymbol{\beta}^*$ and $\boldsymbol{\gamma}_i^*$ provides actionable insights for financial institutions to implement personalized lending and investment strategies. Similarly, enterprises can also leverage these insights to design tailored product recommendation systems.

EXAMPLE 3. **(Count data in commercial marketing)** The increasing demand for personalized rate offerings has made handling heterogeneous count data more crucial than ever across various commercial marketing fields. For instance, to analyze the heterogeneous claim frequency data in the insurance industry, one can employ a Poisson likelihood $f(y_i;\theta_i^*) = \frac{e^{\theta_i^*}(\theta_i^*)^{y_i}}{y_i!}$ with $\theta_i^* = \mathbf{x}_i^\top\boldsymbol{\beta}^* + \mathbf{z}_i^\top\boldsymbol{\gamma}_i^*$, where y_i counts the claim frequency, $\mathbf{x}_i$ denotes the attribute vector, and $\mathbf{z}_i$ captures factors with individualized effects, such as the frequency of night-time driving, or unobserved individual-specific intercepts. In particular, a common issue in automobile insurance is the presence of excessive zero claims, as policyholders often refrain from filing claims for minor losses. In such cases, a zero-inflated Poisson likelihood Chen et al. (2019) offers a more appropriate modeling framework, which also falls into the scope of model (1). By leveraging the proposed model, these practical problems can be effectively addressed, enabling the design of more personalized insurance solutions.

B. ADMM algorithm

In this section, we show the details of the ADMM algorithm for solving TRUE. We establish the our ADMM algorithm following the similar lines of that proposed in Ma and Huang (2017). First of all, we reparameterize the criterion in (3) by introducing a new set of parameters $\boldsymbol{\eta}_{ij} = \boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j$. Then, the minimization of (3) is equivalent to the constraint optimization problem

$$S(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\eta}) = -\frac{1}{n}\mathcal{L}_n(\boldsymbol{\beta},\boldsymbol{\gamma}) + \lambda_n \sum_{i<j} p_{\kappa_n}(\|\boldsymbol{\eta}_{ij}\|_2),$$

$$\text{subject to}\, \boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j - \boldsymbol{\eta}_{ij} = \mathbf{0},$$

where $\boldsymbol{\eta} = \{\boldsymbol{\eta}_{ij}^\top, i < j\}^\top$. By the augmented Lagrangian method, the estimates of the parameters can be obtained by minimizing

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\upsilon}) =& S(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}) + \sum_{i<j} \boldsymbol{\upsilon}_{ij}^\top (\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j - \boldsymbol{\eta}_{ij}) \\ & + \frac{\rho}{2} \sum_{i<j} \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j - \boldsymbol{\eta}_{ij}\|_2^2, \end{aligned}$$

where the dual variables $\boldsymbol{\upsilon} = \{\boldsymbol{\upsilon}_{ij}^\top, i < j\}^\top$ are Lagrange multipliers and $\rho > 0$ is the penalty parameter. In the ADMM algorithm, we compute the estimates of $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\upsilon})$. The ADMM iterations have the following three steps:

$$\begin{aligned} (\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)}) \leftarrow& \arg\min_{\boldsymbol{\beta}, \boldsymbol{\gamma}} \left\{ -\frac{1}{n} \mathcal{L}_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) + \frac{\rho}{2} \|\boldsymbol{\Delta}\boldsymbol{\gamma} - \boldsymbol{\eta}^{(t)} + \boldsymbol{\upsilon}^{(t)}/\rho\|_2^2 \right\}, && (15) \\ \boldsymbol{\eta}^{(t+1)} \leftarrow& \arg\min_{\boldsymbol{\eta}} \left\{ \frac{\rho}{2} \|\boldsymbol{\Delta}\boldsymbol{\gamma}^{(t+1)} - \boldsymbol{\eta} + \boldsymbol{\upsilon}^{(t)}/\rho\|_2^2 + \lambda_n \sum_{i<j} p_{\kappa_n}(\|\boldsymbol{\eta}_{ij}\|_2) \right\}, && (16) \\ \boldsymbol{\upsilon}^{(t+1)} \leftarrow& \boldsymbol{\upsilon}^{(t)} + \rho(\boldsymbol{\Delta}\boldsymbol{\gamma} - \boldsymbol{\eta}^{(t+1)}). \end{aligned}$$

The main ingredients of the algorithm are as follows. The $(\boldsymbol{\beta}, \boldsymbol{\gamma})$-step in (15) can be seen as a generalized ridge penalized maximum likelihood estimation problem. When there is no close-form solution for (15), it can be solved by some numerical computation methods, such as the Newton-Raphson method. The $\boldsymbol{\eta}$-step in (16) is a truncated L_1 penalized least square regression. Its has a close-form expression:

$$\boldsymbol{\eta}_{ij}^{(t+1)} \leftarrow \begin{cases} \mathrm{ST}(\widetilde{\boldsymbol{\eta}}_{ij}^{(t)}, \lambda_n/\rho), & \text{if } \|\widetilde{\boldsymbol{\eta}}_{ij}^{(t)}\|_2 \le \kappa_n \\ \widetilde{\boldsymbol{\eta}}_{ij}^{(t)}, & \text{if } \|\widetilde{\boldsymbol{\eta}}_{ij}^{(t)}\|_2 > \kappa_n, \end{cases}$$

for all $i < j$, where $\widetilde{\boldsymbol{\eta}}_{ij}^{(t)} = \boldsymbol{\gamma}_i^{(t)} - \boldsymbol{\gamma}_j^{(t)}$, $\mathrm{ST}(\mathbf{a}, \lambda) = (1 - \lambda/\|\mathbf{a}\|_2)_+ \mathbf{a}$ is the groupwise soft thresholding operator, and $(\cdot)_+$ denotes the positive part.

Given the initial estimates $(\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)})$, $\boldsymbol{\eta}^{(0)} = \boldsymbol{\Delta}\boldsymbol{\gamma}^{(0)}$, and $\boldsymbol{\upsilon} = \mathbf{0}$, we can obtain the final estimator $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\eta}})$ with enough iterations. This algorithm enables some $\widehat{\boldsymbol{\eta}}_{ij} = \mathbf{0}$, and we assign individuals i and j to one cluster. As a result, we have $\widehat{K}$ estimated clusters $\widehat{\mathcal{G}}_1, ..., \widehat{\mathcal{G}}_{\widehat{K}}$. We set $\widehat{\boldsymbol{\alpha}}_k = |\widehat{\mathcal{G}}_k|^{-1} \sum_{i \in \widehat{\mathcal{G}}_k} \widehat{\boldsymbol{\gamma}}_i$ as the estimated common value for the kth cluster, where $|\widehat{\mathcal{G}}_k|$ is the cardinality of $\widehat{\mathcal{G}}_k$. We next show the convergence properties of the ADMM algorithm.

THEOREM 5. *(Convergence of the ADMM) Let* $\mathbf{r}^{(t)} = \boldsymbol{\Delta}\boldsymbol{\gamma}^{(t)} - \boldsymbol{\eta}^{(t)}$ *and* $\mathbf{s}^{(t+1)} = \rho\boldsymbol{\Delta}^\top(\boldsymbol{\eta}^{(t+1)} - \boldsymbol{\eta}^{(t)})$ *be the primal residual and the dual residual, respectively. It holds that* $\lim_{t\to\infty} \|\mathbf{r}^{(t)}\|_2^2 = 0$ *and* $\lim_{t\to\infty} \|\mathbf{s}^{(t)}\|_2^2 = 0$.

This theorem shows that the primal feasibility and the dual feasibility are achieved by the algorithm. Therefore, it converges to an optimal point. This optimal point may be a local minimum of the objective function due to its nonconvexity.

C. Additional empirical application to Australian credit approval data

Examining the heterogeneity in credit rating evaluations plays a vital role in risk management. Due to the extensive customer base, variations naturally emerge across different consumer segments. This underscores the need to identify latent heterogeneity within the population seeking credit cards, which in turn supports the development of more tailored credit approval policies. To illustrate the applicability of our proposed approach, we analyze the Australian credit approval dataset, which comprises 690 individuals. This dataset, sourced from the UCI Machine Learning Repository, has been extensively examined in prior studies, including Chaudhuri (2014), Guotai et al. (2017), and Runchi et al. (2023). It is particularly noteworthy for its diverse mix of variables, including six numerical and eight categorical attributes, with categorical features varying in the number of unique values. To safeguard data confidentiality, all variable names and values have been anonymized. The target variable is a binary indicator that denotes whether an individual's credit status is classified as good. The target variable exhibits a balanced distribution, with 307 positive cases (44.5%) and 383 negative cases (55.5%). To jointly capture the association between predictor variables and the response while accounting for individual-level heterogeneity, we adopt a heterogeneous logistic model. Specifically, the response variable y_i follows:

$$y_i \sim \text{Bernoulli}\left(\frac{\exp(\theta_i^*)}{1+\exp(\theta_i^*)}\right),$$

where $\theta_i^* = \mathbf{x}_i^\top \boldsymbol{\beta}^* + \gamma_i^*$, with $\mathbf{x}_i$ representing the predictor vector and γ_i^* serving as the individual-specific intercept. Regarding the covariates, we first transform categorical variables into dummy variables. Next, we refine the predictor set through a backward stepwise selection procedure. Ultimately, we retain $p = 12$ variables that exhibit statistical significance for our analysis.

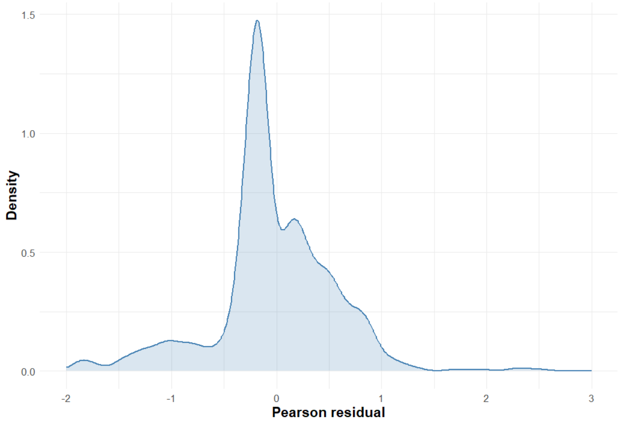

Figure 6 Kernel density plot of the Pearson residuals of the homogeneous logistic model.

Table 5 Classification metrics for TRUE and MLE.

Method	Accuracy	Precision	Recall	F1-Score
TRUE	0.886	0.939	0.849	0.892
MLE	0.877	0.916	0.856	0.885

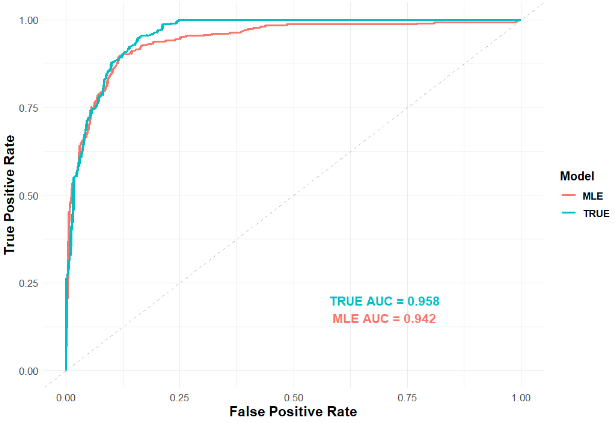

Figure 7 ROC curves for TRUE and MLE.

To assess potential heterogeneous effects, we first fit a homogeneous logistic model using all 12 predictors along with a constant intercept, ensuring that the baseline covariates are appropriately controlled. We then plot the kernel density estimates of the Pearson residuals from the fitted model in Figure 6. For the logistic model, the Pearson residuals are defined as

$$r_i = \frac{y_i - \widehat{p}_i}{\sqrt{\widehat{p}_i(1-\widehat{p}_i)}},$$

where $\widehat{p}_i = \frac{\exp(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}} + \widehat{\gamma})}{1+\exp(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}} + \widehat{\gamma})}$ is the predicted probability with $\widehat{\boldsymbol{\beta}}$ and $\widehat{\gamma}$ representing the maximum likelihood estimates (MLEs) of the model parameters. As shown in the figure, even after adjusting for covariate effects, the residual distribution exhibits multiple modes. This suggests the presence of latent heterogeneity, further motivating the use of a heterogeneous model.

We employ our TRUE method to demonstrate the significance of supervised clustering in this application. The proposed stagewise algorithm and the ADMM algorithm are both used to compute the TRUE estimators; however, we report only the results from the stagewise algorithm, as the two methods yield highly similar estimates. The tuning parameters λ_n and κ_n are selected based on the GIC criterion in (9). For comparison, we also include results from the homogeneous logistic model estimated via MLE. Through the supervised clustering of TRUE, we identify $\widehat{K} = 2$ latent clusters. Based on this estimated clustering structure, we perform personalized credit classification. The classification results are presented below. Table 5 reports various commonly used classification metrics for both methods, while Figure 7 displays the ROC curves of the two classifiers. The results indicate that after identifying latent clusters using the TRUE method, the classifier achieves more accurate classification performance. Notably, its higher precision value suggests that this personalized decision-making strategy is more effective in reducing false positives for creditworthy customers—an aspect of great importance in commercial operations, as it enhances a company's reputation.

Mathematical Proofs for "Fusion learning on supervised clustering and its fast stagewise algorithm"

Letian Li, Yang Li*, Jie Zhang, Zemin Zheng*
International Institute of Finance, The School of Management
University of Science and Technology of China
Hefei, Anhui 230026, P. R. China

This Supplementary Material consists of three parts. Section B provides the details of the ADMM algorithm for solving TRUE and proves its convergence theoretically. Section D lists the key lemmas and presents the proofs for main results. Additional technical proofs for the lemmas are provided in Section E. All the notations are the same as defined in the main body of the paper.

D. Proofs of main results

Before introducing the following two lemmas, we define the oracle estimators which contain the underlying cluster information. The oracle MLE for $(\boldsymbol{\beta},\boldsymbol{\gamma})$ is

$$(\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\gamma}}^{or}) = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\gamma}\in\mathcal{M}_{\mathcal{G}}} \mathcal{L}_n(\boldsymbol{\beta},\boldsymbol{\gamma}). \tag{17}$$

For the common coefficients we have $\boldsymbol{\gamma}=\mathbf{D}\boldsymbol{\alpha}$, thus the oracle MLE for $(\boldsymbol{\beta},\boldsymbol{\alpha})$ is given by

$$(\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\alpha}}^{or}) = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\alpha}\in\mathbb{R}^{qK^*}} \mathcal{L}_n(\boldsymbol{\beta},\mathbf{D}\boldsymbol{\alpha}). \tag{18}$$

Equivalently, $\widehat{\boldsymbol{\alpha}}^{or}$ can be expressed as $T_{\mathcal{G}}(\widehat{\boldsymbol{\gamma}}^{or})$. These oracle estimators are not real estimators but used as a bridge to establish the properties of our method. The following lemmas will be used in the proofs of the main results.

LEMMA 1. *(Consistency of the oracle estimators) Suppose Assumptions 1-3 and $N_{\min}\gg(p+q)n^{1/2}$ hold, the oracle estimators obtained by (17) and (18) satisfy*

$$\|((\widehat{\boldsymbol{\alpha}}^{or}-\boldsymbol{\alpha}^*)^\top,(\widehat{\boldsymbol{\beta}}^{or}-\boldsymbol{\beta}^*)^\top)^\top\|_2 = O_P\left(N_{\min}^{-1}\sqrt{(p+q)n}\right), \tag{19}$$

$$\|\widehat{\boldsymbol{\gamma}}^{or}-\boldsymbol{\gamma}^*\|_2 = O_P\left(N_{\min}^{-1}\sqrt{N_{\max}(p+q)n}\right), \tag{20}$$

$$\sup_i \|\widehat{\boldsymbol{\gamma}}_i^{or}-\boldsymbol{\gamma}_i^*\|_2 = O_P\left(N_{\min}^{-1}\sqrt{(p+q)n}\right). \tag{21}$$

with probability tending to 1 as $N_k\to\infty$ for $k=1,\ldots,K^*$.

LEMMA 2. *(Asymptotic normality of the oracle estimator) Suppose Assumptions 1-3 hold and* $N_{\min} \gg (p+q)n^{2/3}$*. For any* $m \times (qK^* + p)$ *matrix* $\mathbf{A}_n$ *such that* $\mathbf{A}_n\mathbf{A}_n^\top \to \mathbf{S}$*, where* $\mathbf{S}$ *is a* $m \times m$ *symmetric positive definite matrix, we have*

$$\mathbf{A}_n I_n^{1/2}(\boldsymbol{\zeta}^*)((\widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^*)^\top, (\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^*)^\top)^\top \xrightarrow{d} N(\mathbf{0}, \mathbf{S}) \tag{22}$$

as $N_k \to \infty$ *for* $k = 1, ..., K^*$.

We will provide the proofs of the above lemmas in Section E.

D.1. Proof of Proposition 1

The proof of Proposition 1 follows the Karush–Kuhn–Tucker (KKT) conditions and the primal-dual relationship. We write $L^{(t)} = L_n^*(\mathbf{0}, \mathbf{\Delta}^\top \boldsymbol{\eta}^{(t)})$ for abbreviation. Consider the Lagrangian of (8)

$$L(\boldsymbol{\delta}, \boldsymbol{\mu}) = \nabla^\top L^{(t)} \mathbf{\Delta}^\top \boldsymbol{\delta} + \sum_{(i,j)\in S} \mu_{ij}(\|\boldsymbol{\delta}_{ij}\|_2^2 - \epsilon^2) = \sum_{(i,j)\in S} \nabla^\top L^{(t)} \mathbf{\Delta}_{ij} \boldsymbol{\delta}_{ij} + \sum_{(i,j)\in S} \mu_{ij}(\|\boldsymbol{\delta}_{ij}\|_2^2 - \epsilon^2)$$

where $\mathbf{\Delta} = \{\mathbf{\Delta}_{ij}, (i,j) \in S\}^\top$, and μ_{ij} are Lagrange multipliers. By the KKT conditions and the convexity of (8), the solution of (8) can be obtained by solving

$$\begin{cases} \frac{\partial L(\boldsymbol{\delta}, \boldsymbol{\mu})}{\partial \boldsymbol{\delta}_{ij}} = \mathbf{\Delta}_{ij}^\top \nabla L^{(t)} + 2\mu_{ij}\boldsymbol{\delta}_{ij} = \mathbf{0}, \\ \|\boldsymbol{\delta}_{ij}\|_2^2 \le \epsilon^2, \\ \mu_{ij} \ge 0, \\ \mu_{ij}(\|\boldsymbol{\delta}_{ij}\|_2^2 - \epsilon^2) \end{cases}$$

for all $(i,j) \in S$, which yields

$$\widehat{\boldsymbol{\delta}}_{ij} = \begin{cases} -\frac{\mathbf{\Delta}_{ij}^\top \nabla L^{(t)}}{\|\mathbf{\Delta}_{ij}^\top \nabla L^{(t)}\|_2}\epsilon, & \mathbf{\Delta}_{ij}^\top \nabla L^{(t)} \ne \mathbf{0} \\ \mathbf{0}, & \mathbf{\Delta}_{ij}^\top \nabla L^{(t)} = \mathbf{0}. \end{cases} \tag{23}$$

Note that calculating $\nabla L^{(t)}$ can be challenging, so an equivalent expression is absolutely necessary. There is a key relationship between L_n and its conjugate L_n^* that simplifies the update direction in (23) considerably. At the $(t+1)$th iteration, observe that

$$\nabla L_n^*(\mathbf{0}, \mathbf{\Delta}^\top \boldsymbol{\eta}^{(t)}) = \nabla L_n^*(\mathbf{0}, \nabla_{\boldsymbol{\gamma}} L_n(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)})) = \boldsymbol{\gamma}^{(t)}.$$

The first equality comes from the primal-dual relationship (6) at the tth iteration, and the second is due to the fact that

$$\nabla L_n(\boldsymbol{\beta}, \gamma) = \begin{pmatrix} \mathbf{0} \\ \mathbf{\Delta}\boldsymbol{\eta} \end{pmatrix} \iff \nabla L_n^*(\mathbf{0}, \mathbf{\Delta}\boldsymbol{\eta}) = \begin{pmatrix} \boldsymbol{\beta} \\ \gamma \end{pmatrix}.$$

Thus, the conclusion in (8) follows.

D.2. Proof of Theorem 1

The proof of this theorem is established based on Theorem 2 in Tibshirani (2015), which develops the convergence properties of a general convex stagewise learning framework. Note that Algorithm 1 divides the optimization problem into several subproblems piecewisely through the solution path due to the pairs selection step, and all the subproblems are convex. Thus, the convergence for each piece of the solution path can be immediately obtained by Theorem 2 in Tibshirani (2015) if the start point is strictly minimizing the subproblem. To prove Theorem 1, now it suffices to show: (i) the estimate at the each transition point (where the active set changes) is a reasonable start point for the next piece; (ii) among each piece, the minimizer of the corresponding subproblem is also a local minimizer of the objective function (3). In the remainder of this part, we will prove the aforementioned points (i) and (ii) separately.

To facilitate theoretical analysis, we first introduce some necessary notations. Suppose there are $s \geq 0$ transition points, so there are $s+1$ pieces along the solution path. We denote by $t_0 = 0$ and by $t_1, ..., t_s$ the transition points that the active set changes. Since we study the properties in the limiting scenario (i.e., as $\epsilon \to 0$), we assume that at each transition point, there is only one pair is merged or dropped in the active set, without loss of generality.

Proof of (i): We consider the dual problem here. Note that the problem in (5) is equivalent to

$$\min_{\boldsymbol{\eta}} L_n^*(\mathbf{0}, \boldsymbol{\Delta}^\top \boldsymbol{\eta}), \text{ subject to } \sup_{(i,j)\in S} \|\boldsymbol{\eta}_{ij}\|_2 \leq \lambda_n, \sup_{(i,j)\in S^c} \|\boldsymbol{\eta}_{ij}\|_2 = 0. \tag{24}$$

By the KKT conditions, the solution of (24) must satisfy

$$\begin{cases} \boldsymbol{\Delta}_{ij}^\top \nabla L_n^*(\mathbf{0}, \boldsymbol{\Delta}^\top \boldsymbol{\eta}) + 2\mu_{ij}\boldsymbol{\eta}_{ij} = \mathbf{0}, & \forall\, i, j, \\ \|\boldsymbol{\eta}_{ij}\|_2^2 \leq \lambda_n, & (i,j) \in S, \\ \mu_{ij}(\|\boldsymbol{\eta}_{ij}\|_2^2 - \lambda_n^2), & (i,j) \in S, \\ \boldsymbol{\eta}_{ij} = \mathbf{0}, & (i,j) \in S^c, \end{cases} \tag{25}$$

where $\mu_{ij} > 0$ are Lagrange multipliers.

We should consider two classes of transition points: $\mathcal{C}_1$, which comprises the transition points where the active set merge a new element, and $\mathcal{C}_2$, consists of the transition points where the active set drop a contained element. We first consider the class $\mathcal{C}_1$. For any $t \in \mathcal{C}_1$, the estimate $\boldsymbol{\eta}^{(t-1)}$ satisfies (25) for $S = S^{(t-1)}$ given $\lambda_n^{(t-1)}$. Denote by (i', j') the pair merged by S at t, that is, $S^{(t)} = S^{(t-1)} \cup \{(i', j')\}$. Note that $\boldsymbol{\eta}^{(t-1)}$ also satisfies (25) for $S = S^{(t)}$ when giving $\mu_{i'j'} = 0$, thus it minimizes the subproblem for the next piece, which is reasonable to be a start point.

Next, we consider the class $\mathcal{C}_2$. Let (i',j') be the pair dropped at t such that $S^{(t)} = S^{(t-1)} - \{(i',j')\}$. We set $\boldsymbol{\eta}'_{i'j'} = \mathbf{0}$ and $\boldsymbol{\eta}'_{ij} = \boldsymbol{\eta}^{(t-1)}_{ij}$ for all $(i,j) \neq (i',j')$. By the definition of L^*_n, we have

$$L^*_n(\mathbf{0}, \boldsymbol{\Delta}^\top \boldsymbol{\eta}) = \sup_{\boldsymbol{\gamma}} \boldsymbol{\gamma}^\top \boldsymbol{\Delta}^\top \boldsymbol{\eta} - \sup_{\boldsymbol{\gamma}} L_n(\mathbf{0}, \boldsymbol{\gamma}) = \boldsymbol{\gamma}_*^\top \boldsymbol{\Delta}^\top \boldsymbol{\eta} + C,$$

where $\boldsymbol{\gamma}_*$ maximizes $\boldsymbol{\gamma}^\top \boldsymbol{\Delta}^\top \boldsymbol{\eta}$ and C is some constant independent of $\boldsymbol{\eta}$. It can be observed that $\boldsymbol{\Delta}_{ij}^\top \nabla L^*_n(\mathbf{0}, \boldsymbol{\Delta}^\top \boldsymbol{\eta})$ only depends on $\boldsymbol{\eta}_{ij}$. Thus, $\boldsymbol{\eta}'$ satisfies (25) for $S = S^{(t)}$ given $\lambda_n^{(t-1)}$. This implies that $\boldsymbol{\eta}'$ is the minimizer of the subproblem for the next piece, so $\boldsymbol{\eta}^{(t+1)} = \boldsymbol{\eta}'$ is reasonable to be a start point in this case.

Proof of (ii): We consider the primal problem here. For the piece of the solution path with the active set S, the optimization problem is

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} Q_S(\boldsymbol{\beta},\boldsymbol{\gamma}) = \min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \left\{ L_n(\boldsymbol{\beta},\boldsymbol{\gamma}) + \lambda_n \sum_{(i,j)\in S} \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|_2 \right\}.$$

We define

$$\boldsymbol{\Gamma}_S = \left\{(\boldsymbol{\beta},\boldsymbol{\gamma}) \in \mathbb{R}^p \times \mathbb{R}^{nq} : \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|_2 < \kappa_n, \forall (i,j) \in S, \|\boldsymbol{\gamma}_{i'} - \boldsymbol{\gamma}_{j'}\|_2 \geq \kappa_n, \forall (i',j') \in S^c \right\}.$$

One can note that, conditional on $\boldsymbol{\Gamma}_S$, the objective function in (3) is

$$Q_n(\boldsymbol{\beta},\boldsymbol{\gamma}) = L_n(\boldsymbol{\beta},\boldsymbol{\gamma}) + \lambda_n \sum_{(i,j)\in S} \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|_2 + (N - |S|)\lambda_n \kappa_n,$$

where $N = n \times (n-1)/2$ and $|S|$ is the cardinality of S. Thus, conditional on $\boldsymbol{\Gamma}_S$, minimizing $Q_n(\boldsymbol{\beta},\boldsymbol{\gamma})$ is equivalent to minimizing $Q_S(\boldsymbol{\beta},\boldsymbol{\gamma})$. Note that each update along the path for the piece corresponding to the active set S minimizes $Q_S(\boldsymbol{\beta},\boldsymbol{\gamma})$ in $\mathbb{R}^p \times \mathbb{R}^{nq}$, it is also the minimizer within the boundary of $\boldsymbol{\Gamma}_S$. Thus, one can see that each update minimizing $Q_S(\boldsymbol{\beta},\boldsymbol{\gamma})$ is a local minimizer of $Q_n(\boldsymbol{\beta},\boldsymbol{\gamma})$.

D.3. Proof of Theorem 2

In this section, we will show that the oracle estimator is a local minimizer of the objective function (3). Theorem 2 immediately follows from this result and Lemmas 1-2. We follow the similar steps of the proof for Theorem 2 in Ma and Huang (2017). Different from their work, we should tackle the truncated L_1 penalty instead of concave penalties and study in general models rather than only considering the linear model.

With the underlying division $\mathcal{G}_1, ..., \mathcal{G}_{K^*}$, define

$$L_n(\boldsymbol{\beta},\boldsymbol{\gamma}) = -\frac{1}{n}\mathcal{L}_n(\boldsymbol{\beta},\boldsymbol{\gamma}), P_n(\boldsymbol{\gamma}) = \lambda_n \sum_{i<j} p_{\kappa_n}(\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|_2),$$

$$L^{\mathcal{G}}_n(\boldsymbol{\beta},\boldsymbol{\alpha}) = -\frac{1}{n}\mathcal{L}_n(\boldsymbol{\beta},\mathbf{D}\boldsymbol{\alpha}), P^{\mathcal{G}}_n(\boldsymbol{\alpha}) = \lambda_n \sum_{k<k'} N_k N_{k'} p_{\kappa_n}(\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_{k'}\|_2),$$

and denote

$$Q_n(\boldsymbol{\beta},\boldsymbol{\gamma}) = L_n(\boldsymbol{\beta},\boldsymbol{\gamma}) + P_n(\boldsymbol{\gamma}),$$

$$Q_n^{\mathcal{G}}(\boldsymbol{\beta},\boldsymbol{\alpha}) = L_n^{\mathcal{G}}(\boldsymbol{\beta},\boldsymbol{\alpha}) + P_n^{\mathcal{G}}(\boldsymbol{\alpha}).$$

Recall that $T_{\mathcal{G}} : \mathcal{M}_{\mathcal{G}} \to \mathbb{R}^{qK^*}$ is the mapping such that $T_{\mathcal{G}}(\boldsymbol{\gamma})$ is the qK^*-dimensional vector whose kth component equals to the common value of $\boldsymbol{\gamma}_i$ for $i \in \mathcal{G}_k$. Let $T_a : \mathbb{R}^{nq} \to \mathbb{R}^{qK^*}$ be the mapping such that $T_a(\boldsymbol{\gamma}) = \{\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, ..., \boldsymbol{\alpha}_{K^*}^\top)^\top : \boldsymbol{\alpha}_k = N_k^{-1}\sum_{i\in\mathcal{G}_k}\boldsymbol{\gamma}_i, k = 1, ..., K^*\}$. Obviously, when $\boldsymbol{\gamma} \in \mathcal{M}_{\mathcal{G}}$, we have $T_{\mathcal{G}}(\boldsymbol{\gamma}) = T_a(\boldsymbol{\gamma})$. Moreover, for every $\boldsymbol{\gamma} \in \mathcal{M}_{\mathcal{G}}$, we have $P_n(\boldsymbol{\gamma}) = P_n^{\mathcal{G}}(T(\boldsymbol{\gamma}))$, and for every $\boldsymbol{\alpha} \in \mathbb{R}^{qK^*}$ we have $P_n(T_{\mathcal{G}}^{-1}(\boldsymbol{\alpha})) = P_n^{\mathcal{G}}(\boldsymbol{\alpha})$. Hence,

$$Q_n(\boldsymbol{\beta},\boldsymbol{\gamma}) = Q_n^{\mathcal{G}}(\boldsymbol{\beta}, T(\boldsymbol{\gamma})),\ Q_n^{\mathcal{G}}(\boldsymbol{\beta},\boldsymbol{\alpha}) = Q_n(\boldsymbol{\beta}, T_{\mathcal{G}}^{-1}(\boldsymbol{\alpha})).$$

Consider the neighborhood of $(\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*)$:

$$\Omega = \left\{(\boldsymbol{\beta},\boldsymbol{\gamma}) \in \mathbb{R}^p \times \mathbb{R}^{nq} : \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \le \tau_n,\ \sup_i \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^*\|_2 \le \tau_n\right\}.$$

Define the event $\mathcal{F}_1 = \{(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or}) \in \Omega\}$. By Lemma 1, we have $\mathbb{P}(\mathcal{F}_1^c) \to 1$ as $N_k \to \infty$ for $k = 1, ..., K^*$. For any $\boldsymbol{\gamma} \in \mathbb{R}^{nq}$, let $\boldsymbol{\gamma}^a = T_{\mathcal{G}}^{-1}(T_a(\boldsymbol{\gamma}))$. We will prove that $(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or})$ is a local minimizer of the objective function $Q_n(\boldsymbol{\beta},\boldsymbol{\gamma})$ with probability approaching to one for sufficiently large N_k for $k = 1, ..., K^*$ through the following two steps:

(i) On the event $\mathcal{F}_1$, we have $Q_n(\boldsymbol{\beta},\boldsymbol{\gamma}^a) > Q_n(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\gamma}}^{or})$ for any $(\boldsymbol{\beta},\boldsymbol{\gamma}^a) \in \Omega$ and $(\boldsymbol{\beta},\boldsymbol{\gamma}^a) \ne (\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\gamma}}^{or})$.

(ii) There is an event $\mathcal{F}_2 = \left\{\sup_{i\in\mathcal{G}_k, 1\le k\le K^*} \left\|\frac{\partial^2 \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_k^*)}{\partial\boldsymbol{\alpha}_k \partial\boldsymbol{\zeta}_k}\right\|_2 \le \sqrt{\tau q(p+q)n}\right\}$ for a large τ such that $P(\mathcal{F}_2^c) = O(\tau^{-1})$. On $\mathcal{F}_1 \cap \mathcal{F}_2$ there is a neighborhood of $(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\gamma}}^{or})$, denoted by $\Theta_n = \{(\boldsymbol{\beta},\boldsymbol{\gamma}) : \sup_i \|\boldsymbol{\gamma}_i - \widehat{\boldsymbol{\gamma}}_i^{or}\|_2 \le t_n\}$, such that $Q_n(\boldsymbol{\beta},\boldsymbol{\gamma}) > Q_n(\boldsymbol{\beta},\boldsymbol{\gamma}^a)$ for any $(\boldsymbol{\beta},\boldsymbol{\gamma}) \in \Theta_n \cap \Omega$ for sufficiently large N_k for $k = 1, ..., K^*$.

Therefore, by the result of (i) and (ii), we have $Q_n(\boldsymbol{\beta},\boldsymbol{\gamma}) > Q_n(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\gamma}}^{or})$ for any $(\boldsymbol{\beta},\boldsymbol{\gamma}) \in \Theta_n \cap \Omega$, so that $(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\gamma}}^{or})$ is a strictly local minimizer of the objective function $Q_n(\boldsymbol{\beta},\boldsymbol{\gamma})$ on the event $\mathcal{F}_1 \cap \mathcal{F}_2$ which satisfies $\mathbb{P}(\mathcal{F}_1 \cap \mathcal{F}_2) \to 1$ as $N_k \to \infty$ for $k = 1, ..., K^*$.

Step (i): Since $(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or})$ is the global minimizer of $L_n^{\mathcal{G}}(\boldsymbol{\beta},\boldsymbol{\alpha})$, $L_n^{\mathcal{G}}(\boldsymbol{\beta}, T_a(\boldsymbol{\gamma})) > L_n^{\mathcal{G}}(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or})$ for all $(\boldsymbol{\beta}, T_a(\boldsymbol{\gamma})) \in \Omega$. And then we will derive that $P_n^{\mathcal{G}}(T_a(\boldsymbol{\gamma}))$ is a constant which is independent on $\boldsymbol{\gamma}$ for $\boldsymbol{\gamma} \in \Omega$. Let $T_a(\boldsymbol{\gamma}) = \boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_{K^*})^\top$. By the sub-additivity of the L_2-norm, for $k \ne k'$,

$$\begin{aligned}
\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_{k'}\|_2 &\ge \|\boldsymbol{\alpha}_k^* - \boldsymbol{\alpha}_{k'}^*\|_2 - \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^*\|_2 - \|\boldsymbol{\alpha}_{k'} - \boldsymbol{\alpha}_{k'}^*\|_2 \\
&\ge \|\boldsymbol{\alpha}_k^* - \boldsymbol{\alpha}_{k'}^*\|_2 - 2\sup_k \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^*\|_2,
\end{aligned}$$

and

$$\begin{aligned}\sup_k \|\boldsymbol{\alpha}_k-\boldsymbol{\alpha}_k^*\|_2^2 &= \sup_k \left\| N_k^{-1}\sum_{i\in\mathcal{G}_k}\boldsymbol{\gamma}_i-\boldsymbol{\alpha}_k^*\right\|_2^2 = \sup_k \left\| N_k^{-1}\sum_{i\in\mathcal{G}_k}(\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^*)\right\|_2^2 \\ &= \sup_k N_k^{-2}\left\|\sum_{i\in\mathcal{G}_k}(\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^*)\right\|_2^2 \le \sup_k N_k^{-1}\sum_{i\in\mathcal{G}_k}\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^*\|_2^2 \\ &\le \sup_i \|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^*\|_2^2 \le \tau_n^2.\end{aligned} \tag{26}$$

As such, we obtain

$$\|\boldsymbol{\alpha}_k-\boldsymbol{\alpha}_{k'}\|_2 \ge d_n - 2\tau_n.$$

Since $d_n > a\lambda_n = \kappa_n$ and $\kappa_n = a\lambda_n \gg \tau_n$, we have

$$\|\boldsymbol{\alpha}_k-\boldsymbol{\alpha}_{k'}\|_2 > \kappa_n,$$

which leads to $p_{\kappa_n}(\|\boldsymbol{\alpha}_k-\boldsymbol{\alpha}_{k'}\|_2)$ be a constant independent with $\boldsymbol{\gamma}$, and thus $P_n^{\mathcal{G}}(T_a(\boldsymbol{\gamma}))$ is also a constant for any $\boldsymbol{\gamma}\in\Omega$. On conclusion, we have $Q_n^{\mathcal{G}}(\boldsymbol{\beta},T_a(\boldsymbol{\gamma})) > Q_n^{\mathcal{G}}(\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\alpha}}^{or})$ for all $(\boldsymbol{\beta},T_a(\boldsymbol{\gamma}))\in\Omega$. In addition, $Q_n^{\mathcal{G}}(\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\alpha}}^{or}) = Q_n(\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\alpha}}^{or})$ and $Q_n^{\mathcal{G}}(\boldsymbol{\beta},T_a(\boldsymbol{\gamma})) = Q_n(\boldsymbol{\beta},T_{\mathcal{G}}^{-1}(T_a(\boldsymbol{\gamma}))) = Q_n(\boldsymbol{\beta},\boldsymbol{\gamma}^a)$. Hence, $Q_n(\boldsymbol{\beta},\boldsymbol{\gamma}^a) > Q_n(\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\alpha}}^{or})$ for all $(\boldsymbol{\beta},T_a(\boldsymbol{\gamma}))\in\Omega$.

Step (ii): First, we introduce a neighborhood $\Theta_n = \{(\boldsymbol{\beta},\boldsymbol{\gamma}) : \sup_i \|\boldsymbol{\gamma}_i-\widehat{\boldsymbol{\gamma}}_i^{or}\|_2 \le t_n\}$ for a positive sequence t_n. For any $(\boldsymbol{\beta},\boldsymbol{\gamma})\in\Theta_n\cap\Omega$, by the Taylor's expansion, we have

$$\begin{aligned}Q_n(\boldsymbol{\beta},\boldsymbol{\gamma}) - Q_n(\boldsymbol{\beta},\boldsymbol{\gamma}^a) &= -\frac{1}{n}\sum_{i=1}^n\left\{\frac{\partial \ell_i(\boldsymbol{\beta},\boldsymbol{\gamma}^t)}{\partial\boldsymbol{\gamma}_i}\right\}^\top(\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^a) + \sum_{i=1}^n\left\{\frac{\partial P_n(\boldsymbol{\gamma}^t)}{\partial\boldsymbol{\gamma}_i}\right\}^\top(\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^a) \\ &\equiv \Lambda_1+\Lambda_2,\end{aligned}$$

where $\boldsymbol{\gamma}^t = t\boldsymbol{\gamma}+(1-t)\boldsymbol{\gamma}^a$ for some constant $t\in(0,1)$. We will tackle Λ_1 and Λ_2 separately.

First, we deal with Λ_1. By denoting $\boldsymbol{\psi}_i = \frac{\partial \ell_i(\boldsymbol{\beta}, \boldsymbol{\gamma}_i^t)}{\partial \boldsymbol{\gamma}_i}$ for $i = 1, ..., n$, we have

$$\begin{aligned}
\Lambda_1 &= -\frac{1}{n} \sum_{k=1}^{K^*} \sum_{i \in \mathcal{G}_k} \langle \boldsymbol{\psi}_i, \boldsymbol{\gamma}_i - \boldsymbol{\gamma}^a \rangle = -\frac{1}{n} \sum_{k=1}^{K^*} \sum_{i \in \mathcal{G}_k} \left\langle \boldsymbol{\psi}_i, \boldsymbol{\gamma}_i - \sum_{j \in \mathcal{G}_k} \frac{\boldsymbol{\gamma}_j}{N_k} \right\rangle \\
&= -\sum_{k=1}^{K^*} \sum_{i,j \in \mathcal{G}_k} \frac{\langle \boldsymbol{\psi}_i, \boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j \rangle}{n N_k} = -\sum_{k=1}^{K^*} \sum_{i,j \in \mathcal{G}_k} \frac{\langle \boldsymbol{\psi}_j - \boldsymbol{\psi}_i, \boldsymbol{\gamma}_j - \boldsymbol{\gamma}_i \rangle}{2 n N_k} \\
&= -\sum_{k=1}^{K^*} \sum_{i,j \in \mathcal{G}_k, i<j} \frac{\langle \boldsymbol{\psi}_j - \boldsymbol{\psi}_i, \boldsymbol{\gamma}_j - \boldsymbol{\gamma}_i \rangle}{n N_k} \\
&\geq -n^{-1} N_{\min}^{-1} \sum_{k=1}^{K^*} \sum_{i,j \in \mathcal{G}_k, i<j} \max_{i,j} \|\boldsymbol{\psi}_j - \boldsymbol{\psi}_i\|_2 \|\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_i\|_2,
\end{aligned} \tag{27}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, and the fourth equation stems from interchanging the subscripts of i and j. We will only bound $\max_{i,j} \|\boldsymbol{\psi}_j - \boldsymbol{\psi}_i\|_2$. One can see that

$$\begin{aligned}
\max_{i,j} \|\boldsymbol{\psi}_j - \boldsymbol{\psi}_i\|_2 = \max_{i,j} &\left\| \left\{ \frac{\partial \ell_i(\boldsymbol{\beta}, \boldsymbol{\gamma}_i^t)}{\partial \boldsymbol{\gamma}_i} - \frac{\partial \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\gamma}_i^*)}{\partial \boldsymbol{\gamma}_i} \right\} - \left\{ \frac{\partial \ell_i(\boldsymbol{\beta}, \boldsymbol{\gamma}_i^t)}{\partial \boldsymbol{\gamma}_i} - \frac{\partial \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\gamma}_i^*)}{\partial \boldsymbol{\gamma}_i} \right\} \right\|_2 \\
&\leq 2 \sup_i \left\| \frac{\partial \ell_i(\boldsymbol{\beta}, \boldsymbol{\gamma}_i^t)}{\partial \boldsymbol{\gamma}_i} - \frac{\partial \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\gamma}_i^*)}{\partial \boldsymbol{\gamma}_i} \right\|_2 .
\end{aligned}$$

On the other hand, we have under the L_2-norm

$$\frac{\partial \ell_i(\boldsymbol{\beta}, \boldsymbol{\gamma}_i^t)}{\partial \boldsymbol{\gamma}_i} - \frac{\partial \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\gamma}_i^*)}{\partial \boldsymbol{\gamma}_i} = \{1 + o(1)\} \frac{\partial^2 \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_k^*)}{\partial \boldsymbol{\alpha}_k \partial \boldsymbol{\zeta}_k} (\boldsymbol{\xi}_i^t - \boldsymbol{\zeta}_k^*),$$

where $\boldsymbol{\xi}_i^t = (\boldsymbol{\beta}^\top, (\boldsymbol{\gamma}_i^t)^\top)^\top$, $\boldsymbol{\zeta}_k = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}_k^\top)^\top$, and $\boldsymbol{\zeta}_k^* = (\boldsymbol{\beta}^{*\top}, \boldsymbol{\alpha}_k^{*\top})^\top$ for all $i \in \mathcal{G}_k$. Define the event

$$\mathcal{F}_2 = \left\{ \sup_{i \in \mathcal{G}_k, 1 \leq k \leq K^*} \left\| \frac{\partial^2 \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_k^*)}{\partial \boldsymbol{\alpha}_k \partial \boldsymbol{\zeta}_k} \right\|_2 \leq \sqrt{\tau q(p+q)n} \right\}$$

for $\tau \in (0, \infty)$. By the Markov's inequality, we obtain

$$\begin{aligned}
\mathbb{P}(\mathcal{F}_2^c) &\leq \sum_{k=1}^{K^*} \sum_{i \in \mathcal{G}_k} \mathbb{P} \left\{ \left\| \frac{\partial^2 \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_k^*)}{\partial \boldsymbol{\alpha}_k \partial \boldsymbol{\zeta}_k} \right\|_2^2 > \tau q(p+q)n \right\} \\
&\leq \frac{\tau^{-1}}{q(p+q)n} \sum_{k=1}^{K^*} \sum_{i \in \mathcal{G}_k} \mathbb{E} \left\{ \left\| \frac{\partial^2 \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_k^*)}{\partial \boldsymbol{\alpha}_k \partial \boldsymbol{\zeta}_k} \right\|_2^2 \right\} \\
&\leq \frac{\tau^{-1}}{q(p+q)n} \sum_{j=1}^{q} \sum_{l=1}^{p+q} \sum_{k=1}^{K^*} \sum_{i \in \mathcal{G}_k} \mathbb{E} \left\{ \frac{\partial^2 \ell_i(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_k^*)}{\partial \boldsymbol{\zeta}_{k,j} \partial \boldsymbol{\zeta}_{k,l}} \right\}^2 \\
&= O(\tau^{-1}),
\end{aligned}$$

where the last equality holds due to Assumption 3. When τ is sufficiently large, $\mathbb{P}(\mathcal{F}_2) \to 1$, and conditional on this event we have

$$\max_{i,j} \|\boldsymbol{\psi}_j - \boldsymbol{\psi}_i\|_2 \le O(\sqrt{q(p+q)n}) \sup_{i\in\mathcal{G}_k, 1\le k\le K^*} \|\boldsymbol{\xi}_i^t - \boldsymbol{\zeta}_k^*\|_2.$$

As shown in (26), we obtain that

$$\sup_i \|\boldsymbol{\gamma}_i^a - \boldsymbol{\gamma}_i^*\|_2 = \sup_k \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^*\| \le \sup_i \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^*\|_2.$$

Then, we have

$$\begin{aligned}
\sup_i \|\boldsymbol{\gamma}_i^* - \boldsymbol{\gamma}_i^t\|_2 &= \sup_i \|\boldsymbol{\gamma}_i^* - t\boldsymbol{\gamma}_i - (1-t)\boldsymbol{\gamma}^a\|_2 \\
&= t \sup_i \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^*\|_2 + (1-t) \sup_i \|\boldsymbol{\gamma}^a - \boldsymbol{\gamma}_i^*\|_2 \\
&\le \sup_i \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^*\|_2 \le \tau_n.
\end{aligned} \tag{28}$$

Thus,

$$\sup_{i\in\mathcal{G}_k, 1\le k\le K^*} \|\boldsymbol{\xi}_i^t - \boldsymbol{\zeta}_k^*\|_2^2 = \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + \sup_i \|\boldsymbol{\gamma}_i^t - \boldsymbol{\gamma}_i^*\|_2^2 \le 2\tau_n^2.$$

As such, we can see that

$$\max_{i,j} \|\boldsymbol{\psi}_j - \boldsymbol{\psi}_i\|_2 = O(\sqrt{q(p+q)n}\tau_n). \tag{29}$$

Combining (27) and (29), we obtain

$$\Lambda_1 \ge -\sum_{k=1}^{K^*} \sum_{i,j\in\mathcal{G}_k, i<j} O(N_{\min}^{-1}\sqrt{q(p+q)/n})\tau_n \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|_2. \tag{30}$$

Now we address Λ_2. By simple algebraic operation, one can see that

$$\begin{aligned}
\Lambda_2 =& \lambda_n \sum_{i=1}^n \sum_{j\ne i} p'_{\kappa_n}(\|\boldsymbol{\gamma}_i^t - \boldsymbol{\gamma}_j^t\|_2) \|\boldsymbol{\gamma}_i^t - \boldsymbol{\gamma}_j^t\|_2^{-1} (\boldsymbol{\gamma}_i^t - \boldsymbol{\gamma}_j^t)^\top (\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^a) \\
=& \lambda_n \sum_{j<i} p'_{\kappa_n}(\|\boldsymbol{\gamma}_i^t - \boldsymbol{\gamma}_j^t\|_2) \|\boldsymbol{\gamma}_i^t - \boldsymbol{\gamma}_j^t\|_2^{-1} (\boldsymbol{\gamma}_i^t - \boldsymbol{\gamma}_j^t)^\top (\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^a) \\
&+ \lambda_n \sum_{i<j} p'_{\kappa_n}(\|\boldsymbol{\gamma}_i^t - \boldsymbol{\gamma}_j^t\|_2) \|\boldsymbol{\gamma}_i^t - \boldsymbol{\gamma}_j^t\|_2^{-1} (\boldsymbol{\gamma}_i^t - \boldsymbol{\gamma}_j^t)^\top (\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^a).
\end{aligned}$$

Swap i and j in the first term of the second equation,

$$\begin{aligned}\Lambda_2 =&\lambda_n \sum_{i<j} p'_{\kappa_n}(\|\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t\|_2)\|\boldsymbol{\gamma}_j^t-\boldsymbol{\gamma}_i^t\|_2^{-1}(\boldsymbol{\gamma}_j^t-\boldsymbol{\gamma}_i^t)^\top(\boldsymbol{\gamma}_j-\boldsymbol{\gamma}_j^a)\\ &+\lambda_n \sum_{i<j} p'_{\kappa_n}(\|\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t\|_2)\|\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t\|_2^{-1}(\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t)^\top(\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^a)\\ =&\lambda_n \sum_{i<j} p'_{\kappa_n}(\|\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t\|_2)\|\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t\|_2^{-1}(\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t)^\top\{(\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^a)-(\boldsymbol{\gamma}_j-\boldsymbol{\gamma}_j^a)\}.\end{aligned}$$

When $i,j\in\mathcal{G}_k$, $\boldsymbol{\gamma}_i^a=\boldsymbol{\gamma}_j^a$ and $\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t=t(\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_j)$. Hence,

$$\begin{aligned}\Lambda_2 =& \lambda_n \sum_{k=1}^{K^*}\sum_{i,j\in\mathcal{G}_k,i<j} p'_{\kappa_n}(\|\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t\|_2)\|\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t\|_2^{-1}(\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t)^\top(\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_j)\\ &+\lambda_n \sum_{k<k'}\sum_{i\in\mathcal{G}_k,j\in\mathcal{G}_{k'}} p'_{\kappa_n}(\|\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t\|_2)\|\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t\|_2^{-1}(\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t)^\top\{(\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^a)-(\boldsymbol{\gamma}_j-\boldsymbol{\gamma}_j^a)\}.\end{aligned}$$

Next, we will show that the second summation of the above equation equals to zero. For $k\neq k'$, $i\in\mathcal{G}_k$, $j\in\mathcal{G}_{k'}$, by the triangle inequality and (28), it shows that

$$\begin{aligned}\|\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t\|_2 &\geq \|\boldsymbol{\gamma}_i^*-\boldsymbol{\gamma}_j^*\|_2-\|\boldsymbol{\gamma}_i^*-\boldsymbol{\gamma}_i^t\|_2-\|\boldsymbol{\gamma}_j^*-\boldsymbol{\gamma}_j^t\|_2\\ &\geq \min_{i\in\mathcal{G}_k,j\in\mathcal{G}_{k'}}\|\boldsymbol{\gamma}_i^*-\boldsymbol{\gamma}_j^*\|_2-2\sup_i\|\boldsymbol{\gamma}_i^*-\boldsymbol{\gamma}_i^t\|_2\\ &\geq d_n-2\tau_n>\kappa_n.\end{aligned}$$

Thus, we obtain that $p'_{\kappa_n}(\|\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t\|_2)=0$. Then, it shows that

$$\begin{aligned}\Lambda_2 &= \lambda_n \sum_{k=1}^{K^*}\sum_{i,j\in\mathcal{G}_k,i<j} p'_{\kappa_n}(\|\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t\|_2)\|\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t\|_2^{-1}(\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t)^\top(\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_j)\\ &= \lambda_n \sum_{k=1}^{K^*}\sum_{i,j\in\mathcal{G}_k,i<j} p'_{\kappa_n}(\|\boldsymbol{\gamma}_i^t-\boldsymbol{\gamma}_j^t\|_2)\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_j\|_2.\end{aligned}$$

Moreover, by the same reasoning as (26), we have

$$\begin{aligned}\|\boldsymbol{\gamma}_i^t - \boldsymbol{\gamma}_j^t\|_2 &\le 2\sup_i \|\boldsymbol{\gamma}_i^t - \boldsymbol{\gamma}_i^a\|_2 \le 2\sup_i \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^a\|_2 = 2\sup_i \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^a\|_2 \\ &\le 2(\sup_i \|\boldsymbol{\gamma}_i - \widehat{\boldsymbol{\gamma}}_i^{or}\|_2 + \sup_i \|\boldsymbol{\gamma}_i^a - \widehat{\boldsymbol{\gamma}}_i^{or}\|_2) \\ &\le 4\sup_i \|\boldsymbol{\gamma}_i - \widehat{\boldsymbol{\gamma}}_i^{or}\|_2 \le 4t_n.\end{aligned}$$

Let $t_n = o(1)$, we have $p'_{\kappa_n}(4t_n) \to 1$. Thus,

$$\Lambda_2 \ge \sum_{k=1}^{K^*} \sum_{i,j\in\mathcal{G}_k, i<j} \lambda_n \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|_2 \tag{31}$$

for sufficiently large N_k for $k = 1, ..., K^*$.

Since $\tau_n = o(1)$, $N_{\min} \gg \sqrt{(p+q)n}$. Together with $q = o(n)$, it shows that $N_{\min}^{-1}\sqrt{q(p+q)/n}\tau_n = o(1)$. Therefore, combining (30) and (31) we obtain

$$Q_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) - Q_n(\boldsymbol{\beta}, \boldsymbol{\gamma}^a) = \Lambda_1 + \Lambda_2 \ge \sum_{k=1}^{K^*} \sum_{i,j\in\mathcal{G}_k, i<j} \left\{ -O(N_{\min}^{-1}\sqrt{q(p+q)/n})\tau_n + \lambda_n \right\} \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|_2 > 0$$

when $\lambda_n \gg \tau_n$ for some large constant $A > 0$, so that Step (ii) is completed.

D.4. Proof of Theorem 3

We first introduce four classes of model. We call $\mathcal{M}_{\mathcal{G}}$, which is defined in (10), the oracle class. Moreover, we consider three classes of wrong model with K number of clusters: (1) overfitted class ($\mathcal{M}_O$) for which $K > K^*$ and each cluster contains only units from the same cluster; (2) underfitted class ($\mathcal{M}_U$) for which $K < K^*$ and at least one cluster contains all units from more than one cluster; (3) wrongly-assigned class ($\mathcal{M}_W$) if the model is neither $\mathcal{M}_O$ nor $\mathcal{M}_U$. Any candidate wrong model must belong to one of the three classes. We give the mathmetical definitions:

$$\mathcal{M}_O = \left\{ \boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, ..., \boldsymbol{\gamma}_n^\top)^\top \in \mathbb{R}^{nq} : \boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j, \forall i,j \in \mathcal{G}_{kk'}, 1 \le k \le K^* < K \right\},$$

where $\mathcal{G}_{kk'}$ are subclusters of the k-th true cluster $\mathcal{G}_k$ such that $\mathcal{G}_k = \cup_{k'}\mathcal{G}_{kk'}$ and $\sum_{k=1}^{K^*}\sum_{k'} 1 = K$,

$$\mathcal{M}_U = \left\{ \boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, ..., \boldsymbol{\gamma}_n^\top)^\top \in \mathbb{R}^{nq} : \boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j, \forall i,j \in \bar{\mathcal{G}}_k, 1 \le k \le K < K^* \right\},$$

where $\bar{\mathcal{G}}_k = \cup_{k'}\mathcal{G}_{k'}$ are mutually exclusive mergings of the true cluster membership, and

$$\mathcal{M}_W = \{\boldsymbol{\gamma} \in \mathbb{R}^{nq}\} - \mathcal{M}_{\mathcal{G}} - \mathcal{M}_O - \mathcal{M}_U.$$

Moreover, we rewrite the GIC criterion in (14) as another version, which varies with the estimates of $(\boldsymbol{\beta},\boldsymbol{\gamma})$ instead of the regularization parameter λ_n:

$$GIC(\boldsymbol{\beta},\boldsymbol{\gamma}) = -\frac{1}{n}\mathcal{L}_n(\boldsymbol{\beta},\boldsymbol{\gamma}) + (qK+p)\phi_n. \tag{32}$$

Our proof contains three parts, where we compare the GIC values of the oracle model and wrong models for the three class. We only consider the unpenalized model of each class, which maximizes the log-likelihood function $\mathcal{L}_n(\boldsymbol{\beta},\boldsymbol{\gamma})$ restricted on the corresponding class.

Part 1: overfitted model. We consider an arbitrary model belonging to the class $\mathcal{M}_O$ with the cluster number of K. We let $\widetilde{\boldsymbol{\zeta}}^* = (\boldsymbol{\beta}^{*\top},\widetilde{\boldsymbol{\alpha}}^{*\top})^\top$, where $\widetilde{\boldsymbol{\alpha}}^* = (\widetilde{\boldsymbol{\alpha}}_1^{*\top},...,\widetilde{\boldsymbol{\alpha}}_K^{*\top})^\top$ consists of the true common values of all the subclusters $\mathcal{G}_{kk'}$. Then there is some $n\times K$ cluster indicator matrix $\widetilde{\mathbf{D}}$, which has the structure like $\mathbf{D}$, such that $\widetilde{\mathbf{D}}\boldsymbol{\alpha}\in\mathcal{M}_O$ for any $\boldsymbol{\alpha}\in\mathbb{R}^{qK}$. In addition, we denote by $\widehat{\boldsymbol{\zeta}}^{ov} = ((\widehat{\boldsymbol{\beta}}^{ov})^\top,(\widehat{\boldsymbol{\alpha}}^{ov})^\top)^\top$ the unpenalized estimator. That is,

$$(\widehat{\boldsymbol{\beta}}^{ov},\widehat{\boldsymbol{\alpha}}^{ov}) = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\alpha}\in\mathbb{R}^{qK}}\mathcal{L}_n(\boldsymbol{\beta},\widetilde{\mathbf{D}}\boldsymbol{\alpha}).$$

Correspondingly, we denote by $\widehat{\boldsymbol{\gamma}}^{ov}$ the unpenalized estimator for $\boldsymbol{\gamma}$. We also write $\mathcal{L}_n^{ov}(\boldsymbol{\zeta}) = \mathcal{L}_n(\boldsymbol{\beta},\widetilde{\mathbf{D}}\boldsymbol{\alpha})$ with $\boldsymbol{\zeta} = (\boldsymbol{\beta}^\top,\boldsymbol{\alpha}^\top)^\top$. Now we introduce an additional assumption for the overfitted class.

ASSUMPTION 4. *The augmented Fisher information matrices for the overfitted class $\mathcal{M}_O$ satisfies*

$$0 < \min_{\boldsymbol{\gamma}\in\mathcal{M}_O,K^*<K\le K_U}[\widetilde{N}_{\min}^{-1}\lambda_{\min}\{\widetilde{I}_n(\widetilde{\boldsymbol{\zeta}}^*)\}] < \infty,$$

where

$$\widetilde{I}_n(\widetilde{\boldsymbol{\zeta}}^*) = \mathbb{E}_{\widetilde{\boldsymbol{\zeta}}^*}\left[\left\{\frac{\partial\mathcal{L}_n(\boldsymbol{\beta}^*,\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\alpha}}^*)}{\partial\widetilde{\boldsymbol{\zeta}}}\right\}\left\{\frac{\partial\mathcal{L}_n(\boldsymbol{\beta}^*,\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\alpha}}^*)}{\partial\widetilde{\boldsymbol{\zeta}}}\right\}^\top\right],$$

$\widetilde{N}_{\min}$ *is the minimal cluster size, and $\widetilde{\boldsymbol{\zeta}}$ is the vector consisting of $\boldsymbol{\beta}$ and the common values of $\boldsymbol{\gamma}\in\mathcal{M}_O$.*

Assumption 4 is reasonable, since we only consider the identifiable models. Under this condition, we will follow the similar lines of Proof of Lemma 1 to proceed the proof.

When $K = O(K^*)$, $\|\nabla\mathcal{L}_n^{ov}(\widetilde{\boldsymbol{\zeta}}^*)\|_2 = O(\|\nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\|_2)$ and $\|\nabla^2\mathcal{L}_n^{ov}(\widetilde{\boldsymbol{\zeta}}^*) - \mathbb{E}\{\nabla^2\mathcal{L}_n^{ov}(\widetilde{\boldsymbol{\zeta}}^*)\}\|_2 = O(\|\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) - \mathbb{E}\{\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\}\|_2)$. Thus, under the event $\mathcal{E}_1\cap\mathcal{E}_2$ and the condition $\widetilde{N}_{\min}\gg\sqrt{(p+q)^2n}$, we can obtain that $\|\widehat{\boldsymbol{\zeta}}^{ov}-\widetilde{\boldsymbol{\zeta}}^*\|_2 = O_P(\widetilde{N}_{\min}^{-1}\sqrt{(p+q)n})$ and

$$\begin{aligned}
&\mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{ov},\widetilde{\mathbf{D}}\widehat{\boldsymbol{\alpha}}^{ov}) - \mathcal{L}_n(\boldsymbol{\beta}^*,\mathbf{D}\boldsymbol{\alpha}^*) = \mathcal{L}_n^{ov}(\widehat{\boldsymbol{\zeta}}^{ov}) - \mathcal{L}_n^{ov}(\widetilde{\boldsymbol{\zeta}}^*)\\
\le& O(\sqrt{(p+q)n})\|\widehat{\boldsymbol{\zeta}}^{ov}-\widetilde{\boldsymbol{\zeta}}^*\|_2 + \{\lambda_{\min}\{\widetilde{I}_n(\boldsymbol{\zeta}^*)\} - O(\sqrt{(p+q)^2n})\}\|\widehat{\boldsymbol{\zeta}}^{ov}-\widetilde{\boldsymbol{\zeta}}^*\|_2^2\\
\le& O_P(\widetilde{N}_{\min}^{-1}(p+q)n) = o_P(\sqrt{n}),
\end{aligned}$$

where $\widetilde{N}_{\min}$ is the minimal cluster size of the model. The details are omitted to reduce space. Now, we can compare the GIC values of the oracle model and the overfitted model. One can obtain that

$$\begin{aligned}GIC(\widehat{\boldsymbol{\beta}}^{ov},\widehat{\boldsymbol{\gamma}}^{ov})-GIC(\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\gamma}}^{or})&=-\frac{1}{n}\left\{\mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{ov},\widehat{\boldsymbol{\gamma}}^{ov})-\mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\gamma}}^{or})\right\}+q(K-K^*)\phi_n\\&>-\frac{1}{n}\left\{\mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{ov},\widehat{\boldsymbol{\gamma}}^{ov})-\mathcal{L}_n(\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*)\right\}+q(K-K^*)\phi_n\\&>0\end{aligned}$$

when $\phi_n \gg 1/\sqrt{n}$. Due to the arbitrariness of the model we discuss, we can conclude that

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\gamma}\in\mathcal{M}_O,K^*<K\le K_U} GIC(\boldsymbol{\beta},\boldsymbol{\gamma})>GIC(\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\gamma}}^{or})$$

with probability approaching to 1.

Part 2: underfitted model. Here we also consider an arbitrary model belonging to the class $\mathcal{M}_U$ with the cluster number of K. We let $\bar{\boldsymbol{\zeta}}^*=(\boldsymbol{\beta}^{*\top},\bar{\boldsymbol{\alpha}}^{*\top})^\top$, where $\bar{\boldsymbol{\alpha}}^*=(\bar{\boldsymbol{\alpha}}_1^{*\top},...,\bar{\boldsymbol{\alpha}}_{K^*}^{*\top})^\top$ and $\bar{\boldsymbol{\alpha}}_k^*$ is the mean vector of true common values within the merged cluster containing $\mathcal{G}_k$. That is, $\bar{\boldsymbol{\alpha}}_k^*=\frac{1}{|\bar{\mathcal{G}}_l|}\sum_{k'=1}^{K^*}\mathbf{1}_{\{\mathcal{G}_{k'}\subset\bar{\mathcal{G}}_l\}}|\mathcal{G}_{k'}|\boldsymbol{\alpha}_{k'}^*$ for $\mathcal{G}_k\subset\bar{\mathcal{G}}_l$. Moreover, we define the unpenalized estimator

$$(\widehat{\boldsymbol{\beta}}^{un},\widehat{\boldsymbol{\gamma}}^{un})=\arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\gamma}\in\mathcal{M}_U}\mathcal{L}_n(\boldsymbol{\beta},\boldsymbol{\gamma}).$$

Then we denote $\widehat{\boldsymbol{\zeta}}^{un}=((\widehat{\boldsymbol{\beta}}^{un})^\top,(\widehat{\boldsymbol{\alpha}}^{un})^\top)^\top$, where $\widehat{\boldsymbol{\alpha}}^{un}$ is a qK^*-dimensional vector such that $\widehat{\boldsymbol{\gamma}}^{un}=\mathbf{D}\widehat{\boldsymbol{\alpha}}^{un}$.

We first evaluate $\|\bar{\boldsymbol{\zeta}}^*-\boldsymbol{\zeta}^*\|_2$ which is important for our subsequent derivations. Note that

$$\begin{aligned}\|\bar{\boldsymbol{\zeta}}^*-\boldsymbol{\zeta}^*\|_2^2&=\sum_{k=1}^{K^*}\|\bar{\boldsymbol{\alpha}}_k^*-\boldsymbol{\alpha}_k^*\|_2^2=\sum_{l=1}^{K}\sum_{k=1}^{K^*}\mathbf{1}_{\{\mathcal{G}_k\subset\bar{\mathcal{G}}_l\}}\left\|\frac{1}{|\bar{\mathcal{G}}_l|}\sum_{k'=1}^{K^*}\mathbf{1}_{\{\mathcal{G}_{k'}\subset\bar{\mathcal{G}}_l\}}|\mathcal{G}_{k'}|\boldsymbol{\alpha}_{k'}^*-\boldsymbol{\alpha}_k^*\right\|_2^2\\&=\sum_{l=1}^{K}\sum_{k=1}^{K^*}\mathbf{1}_{\{\mathcal{G}_k\subset\bar{\mathcal{G}}_l\}}|\bar{\mathcal{G}}_l|^{-2}\left\|\sum_{k'=1}^{K^*}\mathbf{1}_{\{\mathcal{G}_{k'}\subset\bar{\mathcal{G}}_l\}}|\mathcal{G}_{k'}|(\boldsymbol{\alpha}_{k'}^*-\boldsymbol{\alpha}_k^*)\right\|_2^2.\end{aligned}$$

We consider a merged cluster $\bar{\mathcal{G}}_l$ that contains at least two true clusters $\mathcal{G}_k$ and $\mathcal{G}_j$. One can see that

$$\begin{aligned}&\sum_{k=1}^{K^*}\mathbf{1}_{\{\mathcal{G}_k\subset\bar{\mathcal{G}}_l\}}|\bar{\mathcal{G}}_l|^{-2}\left\|\sum_{k'=1}^{K^*}\mathbf{1}_{\{\mathcal{G}_{k'}\subset\bar{\mathcal{G}}_l\}}|\mathcal{G}_{k'}|(\boldsymbol{\alpha}_{k'}^*-\boldsymbol{\alpha}_k^*)\right\|_2^2\\\ge&|\bar{\mathcal{G}}_l|^{-2}\left\|\sum_{k'=1}^{K^*}\mathbf{1}_{\{\mathcal{G}_{k'}\subset\bar{\mathcal{G}}_l\}}|\mathcal{G}_{k'}|(\boldsymbol{\alpha}_{k'}^*-\boldsymbol{\alpha}_k^*)+\sum_{k'=1}^{K^*}\mathbf{1}_{\{\mathcal{G}_{k'}\subset\bar{\mathcal{G}}_l\}}|\mathcal{G}_{k'}|(\boldsymbol{\alpha}_j^*-\boldsymbol{\alpha}_{k'}^*)\right\|_2^2\\=&|\bar{\mathcal{G}}_l|^{-2}\left\|\sum_{k'=1}^{K^*}\mathbf{1}_{\{\mathcal{G}_{k'}\subset\bar{\mathcal{G}}_l\}}|\mathcal{G}_{k'}|(\boldsymbol{\alpha}_j^*-\boldsymbol{\alpha}_k^*)\right\|_2^2\\=&\|\boldsymbol{\alpha}_j^*-\boldsymbol{\alpha}_k^*\|_2^2.\end{aligned}$$

Thus, we have

$$\|\bar{\boldsymbol{\zeta}}^* - \boldsymbol{\zeta}^*\|_2 \geq \min_{k \neq k'} \|\boldsymbol{\alpha}_k^* - \boldsymbol{\alpha}_{k'}^*\|_2 = d_n.$$

On the other hand, using a similar argument invoking the Taylor's expansion in Lemma 1, we can deduce that

$$\begin{aligned} &\mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{un}, \mathbf{D}\widehat{\boldsymbol{\alpha}}^{un}) - \mathcal{L}_n(\boldsymbol{\beta}^*, \mathbf{D}\boldsymbol{\alpha}^*) = \mathcal{L}_n^{or}(\widehat{\boldsymbol{\zeta}}^{un}) - \mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) \\ \leq & O(\sqrt{(p+q)n})\|\widehat{\boldsymbol{\zeta}}^{un} - \boldsymbol{\zeta}^*\|_2 + \{O(\sqrt{(p+q)^2 n}) - \lambda_{\min}\{I_n(\boldsymbol{\zeta}^*)\}\}\|\widehat{\boldsymbol{\zeta}}^{un} - \boldsymbol{\zeta}^*\|_2^2 \end{aligned}$$

holds under the event $\mathcal{E}_1 \cap \mathcal{E}_2$. Together with the fact

$$\|\widehat{\boldsymbol{\zeta}}^{un} - \boldsymbol{\zeta}^*\|_2 \geq \inf_{\boldsymbol{\zeta}} \|\boldsymbol{\zeta} - \boldsymbol{\zeta}^*\|_2 = \|\bar{\boldsymbol{\zeta}}^* - \boldsymbol{\zeta}^*\|_2,$$

we have

$$\mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{un}, \mathbf{D}\widehat{\boldsymbol{\alpha}}^{un}) - \mathcal{L}_n(\boldsymbol{\beta}^*, \mathbf{D}\boldsymbol{\alpha}^*) \leq -\Omega(N_{\min})\|\widehat{\boldsymbol{\zeta}}^{un} - \boldsymbol{\zeta}^*\|_2^2$$

when $d_n \gg N_{\min}^{-1}\sqrt{(p+q)n}$. Thus, the difference between GIC values of the underfitted model and the oracle model satisfies

$$\begin{aligned} GIC(\widehat{\boldsymbol{\beta}}^{un}, \widehat{\boldsymbol{\gamma}}^{un}) - GIC(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\gamma}}^{or}) &= -\frac{1}{n}\left\{\mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{un}, \widehat{\boldsymbol{\gamma}}^{un}) - \mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\gamma}}^{or})\right\} + q(K - K^*)\phi_n \\ &> -\frac{1}{n}\left\{\mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{un}, \widehat{\boldsymbol{\gamma}}^{un}) - \mathcal{L}_n(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)\right\} - q(K^* - K)\phi_n \\ &> 0 \end{aligned}$$

when $n\phi_n = O(d_n^2 N_{\min})$. Due to the arbitrariness of the model we discuss, we can conclude that

$$\inf_{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\gamma} \in \mathcal{M}_U, K < K^*} GIC(\boldsymbol{\beta}, \boldsymbol{\gamma}) > GIC(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\gamma}}^{or})$$

with probability approaching to 1.

Part 3: wrongly-assigned model. For any wrongly-assigned model, we can construct an intermediate model such that it is overfitted for the oracle model and the wrongly-assigned model is underfitted for it. For clarity and comprehension purposes, we present a simple example to elucidate this fact. We assume the true cluster membership is $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\}$ with $K^* = 3$. In the candidate model which is wrongly-assigned, $\mathcal{G}_3$ is divided into $\mathcal{G}_{31}$ and $\mathcal{G}_{32}$, $\mathcal{G}_{31}$ is merged with $\mathcal{G}_1$, and $\mathcal{G}_{32}$ is merged with $\mathcal{G}_2$, so this model is under the cluster structure of $\{\mathcal{G}_1 \cup \mathcal{G}_{31}, \mathcal{G}_2 \cup \mathcal{G}_{32}\}$ with $K = 2$. In this situation, we can introduce an intermediate model which has the cluster structure of $\{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_{31}, \mathcal{G}_{32}\}$. Denoted by $(\widehat{\boldsymbol{\beta}}^{im}, \widehat{\boldsymbol{\gamma}}^{im})$ the unpenalized estimator for any intermediate model, we can obtain that

$$
\begin{aligned}
&\inf_{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\gamma}\in\mathcal{M}_W,K<K_U} GIC(\boldsymbol{\beta},\boldsymbol{\gamma}) - GIC(\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\gamma}}^{or})\\
=&\inf_{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\gamma}\in\mathcal{M}_W,K<K_U} -\frac{1}{n}\{\mathcal{L}_n(\boldsymbol{\beta},\boldsymbol{\gamma}) - \mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{or},\boldsymbol{\gamma}^{or})\} + q(K-K^*)\phi_n\\
\geq&\inf_{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\gamma}\in\mathcal{M}_W,K<K_U} -\frac{1}{n}\{\mathcal{L}_n(\boldsymbol{\beta},\boldsymbol{\gamma}) - \mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{im},\widehat{\boldsymbol{\gamma}}^{im})\}\\
&+\inf_{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\gamma}\in\mathcal{M}_O,K<K_U K^*} -\frac{1}{n}\{\mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{im},\widehat{\boldsymbol{\gamma}}^{im}) - \mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{or},\boldsymbol{\gamma}^{or})\} + q(K-K^*)\phi_n,\\
>&0
\end{aligned}
$$

where the first inequality holds due to the same arguments as in **Part 1** and **Part 2**.

D.5. Proof of Theorem 4

We continue to employ the four classes of models and all technical notations defined in Section D.4. This proof proceeds in three steps: (i) we first demonstrate that oracle models and overfitted models satisfying the conditions of Theorem 4 exhibit favorable error bounds; (ii) we then show that the refined estimators of these models, obtained through hierarchical clustering, yield the oracle properties outlined in Theorem 2; and (iii) we finally prove that the GIC values of these refined models are close to that of the TRUE model, so that underfitted and wrongly-assigned models would not be selected by our proposed GIC criterion.

Step (i): In this part, we only consider overfitted models, as the cases of oracle models are analogous in this context. By the KKT conditions with respect to the optimization problem (3), we have for any local minimizer $(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}})$ that

$$
\begin{cases}
\frac{\partial \ell_i(\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\gamma}}_i)}{\partial \boldsymbol{\gamma}_i} + \lambda_n \sum_{j\neq i} p'_{\kappa_n}(\|\widetilde{\boldsymbol{\gamma}}_i - \widetilde{\boldsymbol{\gamma}}_j\|_2)\frac{\widetilde{\boldsymbol{\gamma}}_i-\widetilde{\boldsymbol{\gamma}}_j}{\|\widetilde{\boldsymbol{\gamma}}_i-\widetilde{\boldsymbol{\gamma}}_j\|_2} = \mathbf{0},\ i = 1,\dots,n,\\
\sum_{i=1}^n \frac{\partial \ell_i(\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\gamma}}_i)}{\partial \boldsymbol{\beta}} = \mathbf{0}.
\end{cases}
$$

Recall that $\widetilde{\mathbf{D}}$ is the cluster indicator matrix for overfitted models. We write $\widetilde{\boldsymbol{\gamma}} = \widetilde{\mathbf{D}}\widetilde{\boldsymbol{\alpha}}$, where $\widetilde{\boldsymbol{\alpha}} = \{\widetilde{\boldsymbol{\alpha}}_{kk'}^\top\}^\top$ is the common value vector for the estimator $\widetilde{\boldsymbol{\gamma}}$. Summing up the above equations according to the cluster index, we can obtain

$$
\begin{cases}
\sum_{i\in\mathcal{G}_{kk'}} \frac{\partial \ell_i(\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\alpha}}_{kk'})}{\partial \boldsymbol{\alpha}_{kk'}} + \lambda_n \sum_{i\in\mathcal{G}_{kk'}} \sum_{j\neq i} p'_{\kappa_n}(\|\widetilde{\boldsymbol{\gamma}}_i - \widetilde{\boldsymbol{\gamma}}_j\|_2)\frac{\widetilde{\boldsymbol{\gamma}}_i-\widetilde{\boldsymbol{\gamma}}_j}{\|\widetilde{\boldsymbol{\gamma}}_i-\widetilde{\boldsymbol{\gamma}}_j\|_2} = \mathbf{0},\ \forall k, k',\\
\sum_{i=1}^n \frac{\partial \ell_i(\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\gamma}}_i)}{\partial \boldsymbol{\beta}} = \mathbf{0}.
\end{cases}
$$

The above can be expressed as a matrix form of

$$
\begin{cases}
\nabla_{\boldsymbol{\alpha}}\mathcal{L}_n(\widetilde{\boldsymbol{\beta}}, \widetilde{\mathbf{D}}\widetilde{\boldsymbol{\alpha}}) + \lambda_n \boldsymbol{\xi}_n = \mathbf{0},\\
\nabla_{\boldsymbol{\beta}}\mathcal{L}_n(\widetilde{\boldsymbol{\beta}}, \widetilde{\mathbf{D}}\widetilde{\boldsymbol{\alpha}}) = \mathbf{0},
\end{cases}
$$

where $\boldsymbol{\xi}_n$ is stacked by $\sum_{i\in\mathcal{G}_{kk'}}\sum_{j\neq i} p'_{\kappa_n}(\|\widetilde{\boldsymbol{\gamma}}_i-\widetilde{\boldsymbol{\gamma}}_j\|_2)\frac{\widetilde{\boldsymbol{\gamma}}_i-\widetilde{\boldsymbol{\gamma}}_j}{\|\widetilde{\boldsymbol{\gamma}}_i-\widetilde{\boldsymbol{\gamma}}_j\|_2}, \forall k,k'$.

By definition of the truncated L_1 penalty function, we see that $\|p'_{\kappa_n}(\|\widetilde{\boldsymbol{\gamma}}_i-\widetilde{\boldsymbol{\gamma}}_j\|_2)\frac{\widetilde{\boldsymbol{\gamma}}_i-\widetilde{\boldsymbol{\gamma}}_j}{\|\widetilde{\boldsymbol{\gamma}}_i-\widetilde{\boldsymbol{\gamma}}_j\|_2}\|_2\le 1$ for any $\|\widetilde{\boldsymbol{\gamma}}_i-\widetilde{\boldsymbol{\gamma}}_j\|_2\le\kappa_n$ and $p'_{\kappa_n}(\|\widetilde{\boldsymbol{\gamma}}_i-\widetilde{\boldsymbol{\gamma}}_j\|_2)=0$ for any $\|\widetilde{\boldsymbol{\gamma}}_i-\widetilde{\boldsymbol{\gamma}}_j\|_2>\kappa_n$. Due to the condition that $\#\{\|\widetilde{\boldsymbol{\gamma}}_i-\widetilde{\boldsymbol{\gamma}}_j\|_2<\kappa_n, i<j\}=o(\lambda_n^{-2}(p+q)n)$, we have $\|\lambda_n\boldsymbol{\xi}_n\|_2=o(\sqrt{(p+q)n})$. Following the similar lines of **Part 1** in Section E.1, we can deduce that $\|\lambda_n\boldsymbol{\xi}_n\|_2=o_P(\|\nabla_{\boldsymbol{\alpha}}\mathcal{L}_n(\widetilde{\boldsymbol{\beta}},\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\alpha}})\|_2)$ under the event $\mathcal{E}_1\cap\mathcal{E}_2$, implying that $\nabla_{\boldsymbol{\alpha}}\mathcal{L}_n(\widetilde{\boldsymbol{\beta}},\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\alpha}})=\mathbf{0}$ under the L_2-norm. Thus, we obtain $\nabla\mathcal{L}_n^{ov}(\widetilde{\boldsymbol{\zeta}})=\mathbf{0}$ for $\widetilde{\boldsymbol{\zeta}}=(\widetilde{\boldsymbol{\beta}}^\top,\widetilde{\boldsymbol{\alpha}}^\top)^\top$. Furthermore, we have $\|\nabla\mathcal{L}_n^{ov}(\widetilde{\boldsymbol{\zeta}}^*)\|_2=O(\|\nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\|_2)$ and $\|\nabla^2\mathcal{L}_n^{ov}(\widetilde{\boldsymbol{\zeta}}^*)-\mathbb{E}\{\nabla^2\mathcal{L}_n^{ov}(\widetilde{\boldsymbol{\zeta}}^*)\}\|_2=O(\|\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)-\mathbb{E}\{\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\}\|_2)$ when $K=O(K^*)$. Since $\nabla\mathcal{L}_n^{ov}(\widetilde{\boldsymbol{\zeta}})=\mathbf{0}$, under the event $\mathcal{E}_1\cap\mathcal{E}_2$ and the condition $\widetilde{N}_{\min}\gg\sqrt{(p+q)^2n}$, we obtain that $\|\widetilde{\boldsymbol{\zeta}}-\widetilde{\boldsymbol{\zeta}}^*\|_2=O_P(\widetilde{N}_{\min}^{-1}\sqrt{(p+q)n})$.

Step (ii): First, we show that the true cluster structure can be recovered by refining overfitted models. When $\widetilde{N}_{\min}=\Omega(N_{\min})$, it shows that $\|\widetilde{\boldsymbol{\zeta}}-\widetilde{\boldsymbol{\zeta}}^*\|_2=O_P(\tau_n)$. Hence, we have

$$\|\widetilde{\boldsymbol{\gamma}}_i-\widetilde{\boldsymbol{\gamma}}_j\|_2=\|\widetilde{\boldsymbol{\alpha}}_{kk'}-\widetilde{\boldsymbol{\alpha}}_{kk''}\|_2\le\|\widetilde{\boldsymbol{\alpha}}_{kk'}-\boldsymbol{\alpha}_k^*\|_2+\|\widetilde{\boldsymbol{\alpha}}_{kk''}-\boldsymbol{\alpha}_k^*\|_2=O_P(\tau_n)$$

for any $i,j\in\mathcal{G}_k$ and k',k'', and

$$\|\widetilde{\boldsymbol{\gamma}}_i-\widetilde{\boldsymbol{\gamma}}_j\|_2=\|\widetilde{\boldsymbol{\alpha}}_{k_1k'}-\widetilde{\boldsymbol{\alpha}}_{k_2k''}\|_2\ge\|\boldsymbol{\alpha}_{k_1}^*-\boldsymbol{\alpha}_{k_2}^*\|_2-\|\widetilde{\boldsymbol{\alpha}}_{k_1k'}-\boldsymbol{\alpha}_{k_1}^*\|_2-\|\widetilde{\boldsymbol{\alpha}}_{k_2k''}-\boldsymbol{\alpha}_{k_2}^*\|_2\ge d_n-O_P(\tau_n)$$

when $d_n\gg\tau$ for any $i\in\mathcal{G}_{k_1}, j\in\mathcal{G}_{k_2}, k_1\neq k_2$ and k',k''. This means that the pairs $(\widetilde{\boldsymbol{\gamma}}_i,\widetilde{\boldsymbol{\gamma}}_j)$ can be separated into the true clusters by the distance d_n. Then the true cluster structure can be recovered if we refine the model using the hierarchical clustering when the true cluster number K^* is known.

Recall that the estimator of the refined model $\widehat{\boldsymbol{\alpha}}=(\widehat{\boldsymbol{\alpha}}_1^\top,\dots,\widehat{\boldsymbol{\alpha}}_{K^*}^\top)^\top$ yields $\widehat{\boldsymbol{\alpha}}_k=\sum_{k'}\frac{|\mathcal{G}_{kk'}|}{|\mathcal{G}_k|}\widetilde{\boldsymbol{\alpha}}_{kk'}$ for all $k=1,\dots,K^*$. Then we have

$$\begin{aligned}\|\widehat{\boldsymbol{\alpha}}_k-\boldsymbol{\alpha}_k^*\|_2&=\left\|\sum_{k'}\frac{|\mathcal{G}_{kk'}|}{|\mathcal{G}_k|}\widetilde{\boldsymbol{\alpha}}_{kk'}-\boldsymbol{\alpha}_k^*\right\|_2=\left\|\sum_{k'}\frac{|\mathcal{G}_{kk'}|}{|\mathcal{G}_k|}(\widetilde{\boldsymbol{\alpha}}_{kk'}-\boldsymbol{\alpha}_k^*)\right\|_2\\&\le\sum_{k'}\frac{|\mathcal{G}_{kk'}|}{|\mathcal{G}_k|}\|\widetilde{\boldsymbol{\alpha}}_{kk'}-\boldsymbol{\alpha}_k^*\|_2=\|\widetilde{\boldsymbol{\alpha}}_{kk'}-\boldsymbol{\alpha}_k^*\|_2.\end{aligned}$$

Thus, we obtain that the refined estimator $\widehat{\boldsymbol{\zeta}}=(\widehat{\boldsymbol{\beta}}^\top,\widehat{\boldsymbol{\alpha}}^\top)^\top$ satisfies $\|\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}^*\|_2=O_P(\tau_n)$. Combing the true cluster recovery, we conclude that the refined model enjoys the oracle properties in Theorem 2.

Step(iii): First, we derive the difference of loss functions between the refined model and the TRUE model. It shows that

$$|\mathcal{L}_n(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})-\mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\gamma}}^{or})|=|\mathcal{L}_n^{or}(\widehat{\boldsymbol{\zeta}})-\mathcal{L}_n^{or}(\widehat{\boldsymbol{\zeta}}^{or})|\le|\mathcal{L}_n^{or}(\widehat{\boldsymbol{\zeta}})-\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)|+|\mathcal{L}_n^{or}(\widehat{\boldsymbol{\zeta}}^{or})-\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)|.$$

We now evaluate $|\mathcal{L}_n^{or}(\widehat{\boldsymbol{\zeta}})-\mathcal{L}_n^{or}(\widehat{\boldsymbol{\zeta}}^{or})|$. By applying the Taylor's expansion, we have

$$\mathcal{L}_n^{or}(\widehat{\boldsymbol{\zeta}})-\mathcal{L}_n^{or}(\widehat{\boldsymbol{\zeta}}^{or})=\nabla^\top\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)(\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}^*)+\frac{1}{2}(\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}^*)^\top\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)(\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}^*)\{1+o(1)\}.$$

Based on the results given in Section E.1, we have $\|\nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\|_2 = O_P(\sqrt{(p+q)n})$ and $\|\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) - \mathbb{E}\left(\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\right)\|_2 = O_P(\sqrt{\tau(p+q)^2 n})$ under the event $\mathcal{E}_1 \cap \mathcal{E}_2$. Then we deduce that

$$\begin{aligned}|\mathcal{L}_n^{or}(\widehat{\boldsymbol{\zeta}}) - \mathcal{L}_n^{or}(\widehat{\boldsymbol{\zeta}}^{or})| \leq & \|\nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\|_2\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}^*\|_2 + \frac{1}{2}\lambda_{\min}\{I_n(\boldsymbol{\zeta}^*)\}\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}^*\|_2^2 \\ & - \frac{1}{2}\|\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) - \mathbb{E}\left(\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\right)\|_2\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}^*\|_2^2 \\ = & O_P(N_{\min}^{-1}(p+q)n) = o_P(n^{1/3}).\end{aligned}$$

Following the similar way, we can also get $|\mathcal{L}_n^{or}(\widehat{\boldsymbol{\zeta}}^{or}) - \mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)| = o_P(n^{1/3})$. Thus, it shows $|\mathcal{L}_n^{or}(\widehat{\boldsymbol{\zeta}}) - \mathcal{L}_n^{or}(\widehat{\boldsymbol{\zeta}}^{or})| = o_P(n^{1/3})$.

Recall that the estimates version of the GIC criterion has the form of

$$GIC(\boldsymbol{\beta},\boldsymbol{\gamma}) = -\frac{1}{n}\mathcal{L}_n(\boldsymbol{\beta},\boldsymbol{\gamma}) + (qK+p)\phi_n.$$

When $\phi_n \gg 1/\sqrt{n}$, $-\frac{1}{n}\{\mathcal{L}_n(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}}) - \mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\gamma}}^{or})\} = o_P((qK+p)\phi_n)$. This implies that the difference between $\mathcal{L}_n(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})$ and $\mathcal{L}_n(\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\gamma}}^{or})$ can be neglected in the context of the GIC criterion. According to the results proved in **Parts 1-2** of Section D.4, the GIC values of refined models are lower than those of underfitted and wrongly-assigned models.

D.6. Poof of Theorem 5

By the definition of $\boldsymbol{\eta}^{(t+1)}$, we have

$$L(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\gamma}^{(t+1)},\boldsymbol{\eta}^{(t+1)},\boldsymbol{\upsilon}^{(t)}) \leq L(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\gamma}^{(t+1)},\boldsymbol{\eta},\boldsymbol{\upsilon}^{(t)})$$

for any $\boldsymbol{\eta}$. We define

$$\begin{aligned}f^{(t+1)} &= \inf_{\boldsymbol{\Delta\gamma}^{(t+1)}-\boldsymbol{\eta}=\mathbf{0}}\left\{-\frac{1}{n}\mathcal{L}_n(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\gamma}^{(t+1)}) + \lambda_n\sum_{i<j}p_{\kappa_n}(\|\boldsymbol{\eta}_{ij}\|_2)\right\} \\ &= \inf_{\boldsymbol{\Delta\gamma}^{(t+1)}-\boldsymbol{\eta}=\mathbf{0}} L(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\gamma}^{(t+1)},\boldsymbol{\eta},\boldsymbol{\upsilon}^{(t)}).\end{aligned}$$

Then we can see that

$$L(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\gamma}^{(t+1)},\boldsymbol{\eta}^{(t+1)},\boldsymbol{\upsilon}^{(t)}) \leq f^{(t+1)}.$$

Let k be an arbitrary integer. Since $\boldsymbol{v}^{(t+1)} = \boldsymbol{v}^{(t)} + \rho \sum_{j=1}^{k-1}(\boldsymbol{\Delta\gamma}^{t+j} - \boldsymbol{\eta}^{t+j})$, we can obtain

$$\begin{aligned}
L(\boldsymbol{\beta}^{(t+k)}, \boldsymbol{\gamma}^{(t+k)}, \boldsymbol{\eta}^{(t+k)}, \boldsymbol{v}^{(t+k-1)}) =& -\frac{1}{n}\mathcal{L}_n(\boldsymbol{\beta}^{(t+k)}, \boldsymbol{\gamma}^{(t+k)}) + (\boldsymbol{v}^{(t+k-1)})^\top(\boldsymbol{\Delta\gamma}^{(t+k)} - \boldsymbol{\eta}^{(t+k)}) \\
&+ \frac{\rho}{2}\|\boldsymbol{\Delta\gamma}^{(t+k)} - \boldsymbol{\eta}^{(t+k)}\|_2^2 + \lambda_n \sum_{i<j} p_{\kappa_n}(\|\boldsymbol{\eta}_{ij}\|_2) \\
=& -\frac{1}{n}\mathcal{L}_n(\boldsymbol{\beta}^{(t+k)}, \boldsymbol{\gamma}^{(t+k)}) + (\boldsymbol{v}^{(t)})^\top(\boldsymbol{\Delta\gamma}^{(t+k)} - \boldsymbol{\eta}^{(t+k)}) \\
&+ \rho\sum_{j=1}^{k-1}(\boldsymbol{\Delta\gamma}^{t+j} - \boldsymbol{\eta}^{t+j})^\top(\boldsymbol{\Delta\gamma}^{(t+k)} - \boldsymbol{\eta}^{(t+k)}) \\
&+ \frac{\rho}{2}\|\boldsymbol{\Delta\gamma}^{(t+k)} - \boldsymbol{\eta}^{(t+k)}\|_2^2 + \lambda_n \sum_{i<j} p_{\kappa_n}(\|\boldsymbol{\eta}_{ij}\|_2) \\
\le& f^{(t+k)}.
\end{aligned}$$

Since the objective function $L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{v})$ is differentiable with respect to $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and is convex with respect to $\boldsymbol{\eta}$, by applying the results in Theorem 4.1 of Tseng (2001), the sequence $(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\eta}^{(t)})$ has a limit point, denoted by $(\boldsymbol{\beta}^{(*)}, \boldsymbol{\gamma}^{(*)}, \boldsymbol{\eta}^{(*)})$. Then we obtain

$$f^{(*)} = \lim_{t\to\infty} f^{(t+1)} = \lim_{t\to\infty} f^{(t+k)} = \inf_{\boldsymbol{\Delta\gamma}^{(*)} - \boldsymbol{\eta} = \mathbf{0}} \left\{ -\frac{1}{n}\mathcal{L}_n(\boldsymbol{\beta}^{(*)}, \boldsymbol{\gamma}^{(*)}) + \lambda_n \sum_{i<j} p_{\kappa_n}(\|\boldsymbol{\eta}_{ij}\|_2) \right\},$$

and for all $k \ge 0$,

$$\begin{aligned}
&\lim_{t\to\infty} L(\boldsymbol{\beta}^{(t+k)}, \boldsymbol{\gamma}^{(t+k)}, \boldsymbol{\eta}^{(t+k)}, \boldsymbol{v}^{(t+k-1)}) \\
=& -\frac{1}{n}\mathcal{L}_n(\boldsymbol{\beta}^{(*)}, \boldsymbol{\gamma}^{(*)}) + \lambda_n \sum_{i<j} p_{\kappa_n}(\|\boldsymbol{\eta}_{ij}\|_2) + \lim_{t\to\infty}(\boldsymbol{v}^{(t)})^\top(\boldsymbol{\Delta\gamma}^{(*)} - \boldsymbol{\eta}^{(*)}) + (k - \frac{1}{2})\rho\|\boldsymbol{\Delta\gamma}^{(*)} - \boldsymbol{\eta}^{(*)}\|_2^2 \\
\le& f^{(*)}.
\end{aligned}$$

Hence, we have $\lim_{t\to\infty}\|\mathbf{r}^{(t)}\|_2^2 = \|\boldsymbol{\Delta\gamma}^{(*)} - \boldsymbol{\eta}^{(*)}\|_2^2 = 0$.

On the other hand, since $(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)})$ minimizes $L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}^{(t)}, \boldsymbol{v}^{(t)})$ by definition, we see that $\partial L(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)}, \boldsymbol{\eta}^{(t)}, \boldsymbol{v}^{(t)})/\partial\boldsymbol{\gamma} = \mathbf{0}$, and moreover,

$$\begin{aligned}
&\partial L(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)}, \boldsymbol{\eta}^{(t)}, \boldsymbol{v}^{(t)})/\partial\boldsymbol{\gamma} \\
=& -n^{-1}\partial\mathcal{L}_n(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)})/\partial\boldsymbol{\gamma} + \boldsymbol{\Delta}^\top\boldsymbol{v}^{(t)} + \rho\boldsymbol{\Delta}^\top(\boldsymbol{\Delta\gamma}^{(t+1)} - \boldsymbol{\eta}^{(t)}) \\
=& -n^{-1}\partial\mathcal{L}_n(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)})/\partial\boldsymbol{\gamma} + \boldsymbol{\Delta}^\top\{\boldsymbol{v}^{(t)} + \rho(\boldsymbol{\Delta\gamma}^{(t+1)} - \boldsymbol{\eta}^{(t)})\} \\
=& -n^{-1}\partial\mathcal{L}_n(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)})/\partial\boldsymbol{\gamma} + \boldsymbol{\Delta}^\top\{\boldsymbol{v}^{(t+1)} - \rho(\boldsymbol{\Delta\gamma}^{(t+1)} - \boldsymbol{\eta}^{(t+1)}) + \rho(\boldsymbol{\Delta\gamma}^{(t+1)} - \boldsymbol{\eta}^{(t)})\} \\
=& -n^{-1}\partial\mathcal{L}_n(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)})/\partial\boldsymbol{\gamma} + \boldsymbol{\Delta}^\top\boldsymbol{v}^{(t+1)} + \rho\boldsymbol{\Delta}^\top(\boldsymbol{\eta}^{(t+1)} - \boldsymbol{\eta}^{(t)}).
\end{aligned}$$

Thus, we have

$$\mathbf{s}^{(t+1)} = \rho\boldsymbol{\Delta}^\top(\boldsymbol{\eta}^{(t+1)} - \boldsymbol{\eta}^{(t)}) = n^{-1}\partial\mathcal{L}_n(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)})/\partial\boldsymbol{\gamma} - \boldsymbol{\Delta}^\top\boldsymbol{v}^{(t+1)}.$$

Moreover, since $\|\boldsymbol{\Delta}\boldsymbol{\gamma}^{(*)} - \boldsymbol{\eta}^{(*)}\|_2^2 = 0$,

$$\begin{aligned}
&\lim_{t\to\infty} L(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)}, \boldsymbol{\eta}^{(t)}, \boldsymbol{v}^{(t)})/\partial\boldsymbol{\gamma}\\
=&\lim_{t\to\infty}\{-n^{-1}\partial\mathcal{L}_n(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)})/\partial\boldsymbol{\gamma} + \boldsymbol{\Delta}^\top\boldsymbol{v}^{(t+1)}\}\\
=&-n^{-1}\partial\mathcal{L}_n(\boldsymbol{\beta}^{(*)}, \boldsymbol{\gamma}^{(*)})/\partial\boldsymbol{\gamma} + \boldsymbol{\Delta}^\top\boldsymbol{v}^{(*)} = \mathbf{0}.
\end{aligned}$$

Therefore, $\lim_{t\to\infty}\|\mathbf{s}^{(t)}\|_2^2 = 0$.

E. Additional technical details

This part presents the proofs of the lemmas that are used in the proofs of the main theoretical results.

E.1. Proof of Lemma 1

The proof of this lemma builds mainly on the definition of local optimality and the Taylor's expansion. In this part, we use $\mathcal{L}_n^{or}(\boldsymbol{\zeta})$ as the abbreviation for $\mathcal{L}_n(\boldsymbol{\beta}, \mathbf{D}\boldsymbol{\alpha})$ with $\boldsymbol{\zeta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top$. Let $\mathcal{N}_\tau = \{\boldsymbol{\zeta} \in \mathbb{R}^{qK^*+p} : \|\boldsymbol{\zeta} - \boldsymbol{\zeta}^*\|_2 \le N_{\min}^{-1}\sqrt{(p+q)n\tau}\}$ be a closed set with $\tau \in (0, \infty)$, and denote by $\partial\mathcal{N}_\tau$ the boundary of $\mathcal{N}_\tau$. To prove the error bounds in Lemma 1, we will first show that

$$\mathbb{P}\left\{\max_{\boldsymbol{\zeta}\in\partial\mathcal{N}_\tau} \mathcal{L}_n^{or}(\boldsymbol{\zeta}) < \mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\right\} \to 1$$

as $N_k \to \infty, k = 1, ..., K^*$ with large τ. This implies that with probability approaching to 1, there is a local maximizer $\widehat{\boldsymbol{\zeta}}^{or}$ in the set $\mathcal{N}_\tau$ such that $\|\widehat{\boldsymbol{\zeta}}^{or} - \boldsymbol{\zeta}^*\|_2 = O_P(N_{\min}^{-1}\sqrt{(p+q)n})$. Note that by the convexity of $\mathcal{L}_n^{or}(\boldsymbol{\zeta})$, the local maximizer $\widehat{\boldsymbol{\zeta}}^{or}$ consists of the oracle MLEs we defined in Section 4. By applying the Taylor's expansion, we obtain

$$\begin{aligned}
\mathcal{L}_n^{or}(\boldsymbol{\zeta}) - \mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) &= \nabla^\top\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)(\boldsymbol{\zeta} - \boldsymbol{\zeta}^*) + \frac{1}{2}(\boldsymbol{\zeta} - \boldsymbol{\zeta}^*)^\top\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)(\boldsymbol{\zeta} - \boldsymbol{\zeta}^*)\{1 + o(1)\}\\
&\equiv I_1 + I_2.
\end{aligned}$$

We first bound I_1. Consider the event

$$\mathcal{E}_1 = \left\{\|\nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\|_2 \le \sqrt{\tau(p+q)n}\right\}.$$

Since $\max_k[\lambda_{\max}\{I(\boldsymbol{\zeta}_k^*)\}] = O(1)$ due to Assumption 2,

$$\mathbb{E}\left\{\|\nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\|_2^2\right\} = \sum_{j=1}^{p+q}\sum_{k=1}^{K^*}\sum_{i\in\mathcal{G}_k}\mathbb{E}\left\{\frac{\partial\ell_i(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_k^*)}{\partial\zeta_{k,j}}\right\}^2 = \sum_{k=1}^{K^*}\sum_{i\in\mathcal{G}_k}\mathrm{tr}\{I(\boldsymbol{\zeta}_k^*)\} = O((p+q)n).$$

Then by the Markov's inequality, we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}_1^c) &= \mathbb{P}\left\{\|\nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\|_2 > \sqrt{\tau(p+q)n}\right\}\\ &\le \frac{\tau^{-1}}{(p+q)n}\mathbb{E}\{\|\nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\|_2\}\\ &= \frac{\tau^{-1}}{(p+q)n}\left[\mathbb{E}\left\{\|\nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\|_2^2\right\}\right]^{1/2}\\ &= O(\tau^{-1}).\end{aligned} \tag{33}$$

Thus, under the event $\mathcal{E}_1$ we have

$$|I_1| \le \|\nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\|_2\|\boldsymbol{\zeta}-\boldsymbol{\zeta}^*\|_2 = \sqrt{\tau(p+q)n}\|\boldsymbol{\zeta}-\boldsymbol{\zeta}^*\|_2. \tag{34}$$

Then we deal with I_2. By Assumption 1 and ignoring a small order term, one can see that

$$I_2 = \frac{1}{2}(\boldsymbol{\zeta}-\boldsymbol{\zeta}^*)^\top\left[\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) - \mathbb{E}\left\{\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\right\}\right](\boldsymbol{\zeta}-\boldsymbol{\zeta}^*) - \frac{1}{2}(\boldsymbol{\zeta}-\boldsymbol{\zeta}^*)I_n(\boldsymbol{\zeta}^*)^\top(\boldsymbol{\zeta}-\boldsymbol{\zeta}^*).$$

Define the event

$$\mathcal{E}_2 = \left\{\|\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) - \mathbb{E}\left(\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\right)\|_2 \le \sqrt{\tau(p+q)^2 n}\right\}.$$

Under Assumption 3, one can see that

$$\begin{aligned}&\mathbb{E}\left\{\|\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) - \mathbb{E}\left(\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\right)\|_2^2\right\}\\ \le&\sum_{j,l=1}^{p+q}\sum_{k=1}^{K^*}\sum_{i\in\mathcal{G}_k}\mathbb{E}\left[\frac{\partial\ell_i^2(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_k^*)}{\partial\zeta_{k,j}\partial\zeta_{k,l}} - \mathbb{E}\left\{\frac{\partial\ell_i^2(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_k^*)}{\partial\zeta_{k,j}\partial\zeta_{k,l}}\right\}\right]^2\\ \le&\sum_{j,l=1}^{p+q}\sum_{k=1}^{K^*}\sum_{i\in\mathcal{G}_k}\mathbb{E}\left\{\frac{\partial\ell_i^2(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_k^*)}{\partial\zeta_{k,j}\partial\zeta_{k,l}}\right\}^2\\ =&O(n(p+q)^2).\end{aligned}$$

Further, by the Markov's inequality, we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}_2^c) &= \mathbb{P}\left\{\|\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) - \mathbb{E}\left(\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\right)\|_2 > \sqrt{\tau(p+q)^2n}\right\} \\ &\leq \frac{\tau^{-1}}{(p+q)^2n}\mathbb{E}\left\{\|\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) - \mathbb{E}\left(\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\right)\|_2\right\} \\ &= \frac{\tau^{-1}}{(p+q)^2n}\left[\mathbb{E}\left\{\|\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) - \mathbb{E}\left(\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\right)\|_2^2\right\}\right]^{1/2} \\ &= O(\tau^{-1}).\end{aligned} \tag{35}$$

Thus, conditional on the event $\mathcal{E}_2$, we have

$$\begin{aligned}I_2 &\leq \frac{1}{2}\|\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) - \mathbb{E}\left(\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\right)\|_2\|\boldsymbol{\zeta}-\boldsymbol{\zeta}^*\|_2^2 - \frac{1}{2}\lambda_{\min}\{I_n(\boldsymbol{\zeta}^*)\}\|\boldsymbol{\zeta}-\boldsymbol{\zeta}^*\|_2^2 \\ &\leq \frac{1}{2}\{\sqrt{\tau(p+q)^2n} - \Omega(N_{\min})\}\|\boldsymbol{\zeta}-\boldsymbol{\zeta}^*\|_2^2\end{aligned} \tag{36}$$

by Assumption 2.

From the results of (33)-(36), one can see that, conditional on $\mathcal{E}_1 \cap \mathcal{E}_1$, I_1 is dominated by I_2 and it is negative when τ is large enough and $N_{\min} \gg \sqrt{(p+q)^2n}$. Therefore, we have

$$\mathbb{P}\left\{\max_{\boldsymbol{\zeta}\in\partial\mathcal{N}_\tau}\mathcal{L}_n^{or}(\boldsymbol{\zeta}) < \mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\right\} \geq 1 - O(\tau^{-1}),$$

which proves $\|\widehat{\boldsymbol{\zeta}}^{or} - \boldsymbol{\zeta}^*\|_2 = O_P(N_{\min}^{-1}\sqrt{(p+q)n})$, that is, the error bound (19). Moreover, (20)-(21) follow from

$$\begin{aligned}\|\widehat{\boldsymbol{\gamma}}^{or} - \boldsymbol{\gamma}^*\|_2^2 &= \sum_{k=1}^{K^*}\sum_{i\in\mathcal{G}_k}\|\widehat{\boldsymbol{\alpha}}_k^{or} - \boldsymbol{\alpha}_k^*\|_2^2 \leq N_{\max}\sum_{k=1}^{K^*}\|\widehat{\boldsymbol{\alpha}}_k^{or} - \boldsymbol{\alpha}_k^*\|_2^2 \\ &= N_{\max}\|\widehat{\boldsymbol{\zeta}}^{or} - \boldsymbol{\zeta}^*\|_2^2 = O_P(N_{\min}^{-2}N_{\max}(p+q)n)\end{aligned}$$

and

$$\sup_i\|\widehat{\boldsymbol{\gamma}}_i^{or} - \boldsymbol{\gamma}_i^*\|_2 = \sup_k\|\widehat{\boldsymbol{\alpha}}_k^{or} - \boldsymbol{\alpha}_k^*\|_2 \leq \|\widehat{\boldsymbol{\zeta}}^{or} - \boldsymbol{\zeta}^*\|_2 = O_P(N_{\min}^{-1}\sqrt{(p+q)n}).$$

E.2. Proof of Lemma 2

The proof of this lemma is mainly established based on the consistency of the oracle estimator derived in Lemma 1 and the Lindeberg-Feller central limit theorem (Van der Vaart 2000). We continue to adopt the notation in the proof of Lemma 2. In the rest of this part, we consider the properties of the oracle estimators conditional on the event $\mathcal{E}_1 \cap \mathcal{E}_2$ with $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \to 1$ as $N_k \to \infty$ for $k = 1, ..., K^*$.

Recall that $\widehat{\boldsymbol{\zeta}}^{or}$ is the maximizer of $\mathcal{L}_n^{or}(\boldsymbol{\zeta})$, we have $\nabla\mathcal{L}_n^{or}(\widehat{\boldsymbol{\zeta}}^{or}) = \mathbf{0}$. We expand $\nabla\mathcal{L}_n^{or}(\widehat{\boldsymbol{\zeta}}^{or})$ around $\boldsymbol{\zeta}^*$ to the first order componentwise. According to the proof of Lemma 1, under the event $\mathcal{E}_1 \cap \mathcal{E}_2$ we have $\|\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) - \mathbb{E}\{\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\}\|_2 = O_P(\sqrt{(p+q)^2 n})$ and $\|\widehat{\boldsymbol{\zeta}}^{or} - \boldsymbol{\zeta}^*\|_2 = O_P(N_{\min}^{-1}\sqrt{(p+q)n})$. One can obtain under the L_2-norm

$$
\begin{aligned}
I_n(\boldsymbol{\zeta}^*)(\widehat{\boldsymbol{\zeta}}^{or} - \boldsymbol{\zeta}^*) &= \nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) + [\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) - \mathbb{E}\left\{\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)\right\}](\widehat{\boldsymbol{\zeta}}^{or} - \boldsymbol{\zeta}^*) + o(1)\nabla^2\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)(\widehat{\boldsymbol{\zeta}}^{or} - \boldsymbol{\zeta}^*) \\
&= \nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) + O_P(N_{\min}^{-1}(p+q)^{3/2}n) \\
&= \nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) + o_P(N_{\min}^{1/2})
\end{aligned}
$$

when $N_{\min} \gg (p+q)n^{2/3}$. Thus, we have

$$
\mathbf{A}_n I_n^{1/2}(\boldsymbol{\zeta}^*)(\widehat{\boldsymbol{\zeta}}^{or} - \boldsymbol{\zeta}^*) = \mathbf{A}_n I_n^{-1/2}(\boldsymbol{\zeta}^*)\nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) + o_P(N_{\min}^{1/2}\mathbf{A}_n I_n^{-1/2}(\boldsymbol{\zeta}^*)).
$$

Since $\|I_n(\boldsymbol{\zeta}^*)^{-1/2}\|_2 = O(N_{\min}^{-1/2})$ by Assumption 2, we have the last term of $o_P(1)$. By Slutsky's lemma, to show that

$$
\mathbf{A}_n I_n^{1/2}(\boldsymbol{\zeta}^*)(\widehat{\boldsymbol{\zeta}}^{or} - \boldsymbol{\zeta}^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{S}),
$$

it suffices to prove $\mathbf{A}_n I_n^{-1/2}(\boldsymbol{\zeta}^*)\nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{S})$.

To this end, we denote

$$
\mathbf{A}_n I_n^{-1/2}(\boldsymbol{\zeta}^*)\nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*) = \sum_{k=1}^{K^*}\sum_{i\in\mathcal{G}_k}\boldsymbol{\xi}_i,
$$

where $\boldsymbol{\xi}_i = \mathbf{A}_n I_n^{-1/2}(\boldsymbol{\zeta}^*)\left\{\frac{\ell_i(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_k^*)}{\partial\boldsymbol{\zeta}}\right\}$ for $i \in \mathcal{G}_k, k = 1, ..., K^*$. By the Cauchy-Schwarz inequality, it follows that

$$
\sum_{k=1}^{K^*}\sum_{i\in\mathcal{G}_k}\mathbb{E}[\|\boldsymbol{\xi}_i\|_2^2\mathbf{1}\{\|\boldsymbol{\xi}_i\|_2 > \epsilon\}] \le \sum_{k=1}^{K^*}\sum_{i\in\mathcal{G}_k}\{\mathbb{E}(\|\boldsymbol{\xi}_i\|_2^4)\}^{1/2}\{\mathbb{P}(\|\boldsymbol{\xi}_i\|_2 > \epsilon)\}^{1/2}
$$

for any $\epsilon > 0$. By the Markov's inequality and Assumptions 2-3, we have

$$
\begin{aligned}
\mathbb{P}(\|\boldsymbol{\xi}_i\|_2 > \epsilon) &\le \frac{1}{\epsilon^2}\mathbb{E}\left\{\left\|\mathbf{A}_n I_n^{-1/2}(\boldsymbol{\zeta}^*)\left(\frac{\ell_i(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_k^*)}{\partial\boldsymbol{\zeta}}\right)\right\|_2^2\right\} \le \frac{1}{\epsilon^2}\|I_n^{-1/2}(\boldsymbol{\zeta}^*)\|_2^2\sum_{j=1}^{p+q}\mathbb{E}\left\{\frac{\ell_i(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_k^*)}{\partial\zeta_{k,j}}\right\}^2 \\
&= O(N_{\min}^{-1})\sum_{j=1}^{p+q}\mathbb{E}\left\{\frac{\ell_i(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_k^*)}{\partial\zeta_{k,j}}\right\}^2 = O(N_{\min}^{-1})\mathrm{tr}\{I(\boldsymbol{\zeta}_k^*)\} \\
&= O(N_{\min}^{-1}(p+q))
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}(\|\boldsymbol{\xi}_i\|_2^4) &= \mathbb{E}\left\{\left\|\mathbf{A}_n I_n^{-1/2}(\boldsymbol{\zeta}^*)\left(\frac{\ell_i(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_k^*)}{\partial\boldsymbol{\zeta}}\right)\right\|_2^4\right\} = O(N_{\min}^{-2})\left[\sum_{j=1}^{p+q}\mathbb{E}\left\{\frac{\ell_i(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_k^*)}{\partial\zeta_{k,j}}\right\}^2\right]^2 \\
&= O(N_{\min}^{-2})[\mathrm{tr}\{I(\boldsymbol{\zeta}_k^*)\}]^2 = O(N_{\min}^{-2}(p+q)^2).
\end{aligned}
$$

Combining the above, it can be seen that

$$\sum_{k=1}^{K^*}\sum_{i\in\mathcal{G}_k}\mathbb{E}[\|\boldsymbol{\xi}_i\|_2^2\mathbf{1}\{\|\boldsymbol{\xi}_i\|_2>\epsilon\}]=O(N_{\min}^{-3/2}(p+q)^{3/2}n)=o(1).$$

On the other hand, as $\mathbf{A}_n\mathbf{A}_n^\top\to\mathbf{S}$, we have

$$\sum_{k=1}^{K^*}\sum_{i\in\mathcal{G}_k}\mathrm{Var}(\boldsymbol{\xi}_i)=\mathbf{A}_nI_n^{-1/2}(\boldsymbol{\zeta}^*)I_n(\boldsymbol{\zeta}^*)I_n^{-1/2}(\boldsymbol{\zeta}^*)\mathbf{A}_n^\top=\mathbf{A}_n\mathbf{A}_n^\top\to\mathbf{S}.$$

Thus, $\boldsymbol{\xi}_i$ satisfy the conditions of the Lindeberg-Feller central limit theorem, which entails

$$\mathbf{A}_nI_n^{-1/2}(\boldsymbol{\zeta}^*)\nabla\mathcal{L}_n^{or}(\boldsymbol{\zeta}^*)=\sum_{k=1}^{K^*}\sum_{i\in\mathcal{G}_k}^{n}\boldsymbol{\xi}_i\xrightarrow{d}N(\mathbf{0},\mathbf{S}).$$

The conclusion follows.

References

Ma S, Huang J (2017) A concave pairwise fusion approach to subgroup analysis. *J. Am. Stat. Assoc.* 112(517):410–423.

Tibshirani RJ (2015) A general framework for fast stagewise algorithms. *J. Mach. Learn. Res.* 16(1):2543–2588.

Tseng, P (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.* 109:109.

Van der Vaart AW (2015) *Asymptotic statistics*, volume 3 (Cambridge university press).