Biometrics 000, 000–000 DOI: 000
000 0000

# Supplementary Materials for "Subgroup identification and membership prediction" by Lu Chen, Xuerong Chen, Xinzhou, Guo and Yi Li

**Lu Chen,[1], Xuerong Chen,[1,*], Xinzhou, Guo[2] and Yi Li[3]**

[1]Center of Statistical Research, Southwestern University of Finance and Economics

[2]Department of Mathematics, The Hong Kong University of Science and Technology

[3]Department of Biostatistics, University of Michigan

**email:* chenxuerong@swufe.edu.cn

This paper has been submitted for consideration for publication in *Biometrics*

The Appendix provides notation details, formula details for section of methodology and theorectical property section. It gives computational details and discusses the convergence properties of the block-wise alternating directions method of multipliers algorithm. It also includes additional simulation studies, additional real data analysis and technical proofs for the theorem presented in Section 4.

## A1. Methodology

### Two-group partition recovery and group membership prediction

Let $\mathcal{H}_{\mathcal{K}}$ is the reproducing kernel Hilbert space (RKHS) associate with kernel function $\mathcal{K} : \mathcal{U} \times \mathcal{U} \to R$, which is the completion of the linear span of all functions $\{\mathcal{K}(\cdot, \boldsymbol{u}), \boldsymbol{u} \in \mathcal{U}\}$. The norm in $\mathcal{H}_{\mathcal{K}}$, denoted by $\|\cdot\|_{\mathcal{K}}$, is induced by the inner product, $\langle f, g\rangle_{\mathcal{K}} = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \mathcal{K}(\boldsymbol{u}_i, \boldsymbol{u}_j)$ for $f(\cdot) = \sum_{i=1}^{n} a_i \mathcal{K}(\cdot, \boldsymbol{u}_i)$ , $g(\cdot) = \sum_{j=1}^{m} b_j \mathcal{K}(\cdot, \boldsymbol{u}_j)$ and $\|f_g^*(\boldsymbol{U}_i)\|_{\mathcal{K}}^2 = \left\langle f_g^*(\boldsymbol{U}_i), f_g^*(\boldsymbol{U}_i)\right\rangle_{\mathcal{K}}$.

After obtaining $\widehat{f_g}(\boldsymbol{U})$, define $C_{gm} = \#\left\{i \in \{1, \cdots, n\} : \widehat{\alpha}_{ig} = \widehat{\theta}_{gm}, \widehat{f_g}(\boldsymbol{U}_i) \leqslant 0\right\}, m = 1, 2$. After obtaining $\widehat{f_g}(\boldsymbol{U})$, define $C_{gm} = \#\left\{i \in \{1, \cdots, n\} : \widehat{\alpha}_{ig} = \widehat{\theta}_{gm}, \widehat{f_g}(\boldsymbol{U}_i) \leqslant 0\right\}, m = 1, 2$. The partition estimation $\widehat{\mathcal{U}}_{g1} = \{\boldsymbol{U} : \widehat{f_g}(\boldsymbol{U}) \leqslant 0\}$, $\widehat{\mathcal{U}}_{g2} = \{\boldsymbol{U} : \widehat{f_g}(\boldsymbol{U}) > 0\}$ for $C_{g1} > C_{g2}$. Otherwise, $\widehat{\mathcal{U}}_{g1} = \{\boldsymbol{U} : \widehat{f_g}(\boldsymbol{U}) > 0\}$, $\widehat{\mathcal{U}}_{g2} = \{\boldsymbol{U} : \widehat{f_g}(\boldsymbol{U}) \leqslant 0\}$ for $C_{g1} < C_{g2}$. When a new subject arrives, we can predict its group membership by determining which partition block its $\boldsymbol{U}_i$ falls into.

### Multi-group partitions recovery and group membership prediction

If for some $g \in \{1, \ldots, p\}$, $\widehat{M}_g > 2$, then the $g$-th partition estimation could be a multicategory classification problem. In following part, we will use the support vector machine method for multiclass to obtain the partition estimates, which will facilitate subgroup membership prediction.

In the multi-classification problem, the task is to learn a classification rule $\widehat{\phi}_g : \boldsymbol{U} \to \{1, \ldots, \widehat{M}_g\}$. Similarly, we assign artificial labels to $\widehat{M}_g$ groups firstly, let $\widetilde{Y}_{gi}^* = m$ when $i$-th subjects' belongs to $\widehat{G}_{gm}$. Motivated by Lee et al. (2004), we extend artificial label $\widetilde{Y}_{gi}^*$ into the form of $\widehat{M}_g$-dimensional vector $\widetilde{\boldsymbol{Y}}_{gi}^*$. For instance, if $i$-th subjects' belong to $\widehat{G}_{gm}$, that is $\widetilde{Y}_{gi}^* = m$, define $\widetilde{\boldsymbol{Y}}_{gi}^*$ as a $\widehat{M}_g$-dimensional vector with 1 in the $m$th coordinate and $-1/(\widehat{M}_g - 1)$ elsewhere. Accordingly, we

define a $\widehat{M}_g$-tuple of separating functions $f_g(\boldsymbol{U}) = (f_{g1}(\boldsymbol{U}), \ldots, f_{g\widehat{M}_g}(\boldsymbol{U}))^T$ with the sum-to-zero constraint, $\sum_{m=1}^{\widehat{M}_g} f_{gm}(\boldsymbol{U}) = 0$ for any $\boldsymbol{U} \in \mathbb{R}^r$. Then the $f_g(\boldsymbol{U}) = (f_{g1}(\boldsymbol{U}), \ldots, f_{g\widehat{M}_g}(\boldsymbol{U}))^T$ could be obtained by solving following optimization problem:

$$\arg \min_{f_g \in \prod_{m=1}^{\widehat{M}_g}(\overline{\mathcal{H}}_\mathcal{K}), \sum_{m=1}^{\widehat{M}_g} f_{gm}(\boldsymbol{U})=0} \left\{ \frac{1}{n} \sum_{i=1}^{n} C(\widetilde{\boldsymbol{Y}}^*_{gi}) \cdot \left( f_g(\boldsymbol{U}_i) - \widetilde{\boldsymbol{Y}}^*_{gi} \right)_+ + \frac{\lambda_n}{2} \sum_{m=1}^{\widehat{M}_g} \| f^*_{gm}(\boldsymbol{U}_i) \|^2_\mathcal{K} \right\}, \quad \text{(A.1)}$$

in which the notation details please refer to supplementary material. And the multi-classifier $\widehat{\phi}_g : \boldsymbol{U} \to \{1, \ldots, \widehat{M}_g\}$ induced by $f_g(\boldsymbol{U})$ is naturally $\widehat{\phi}_g(\boldsymbol{U}) = \arg\max_m \widehat{f}_{gm}(\boldsymbol{U})$. Based on $\widehat{f}_g(\boldsymbol{U})$, the latent partition can be recovered. When a new subject come, we also can predict the group membership by fitting it's $\boldsymbol{U}_i$ into the classifier $\widehat{\phi}_g(\boldsymbol{U})$.

These constraints reflect the implicit nature of the subject in classification problems takes one and only one group from $\{1, \cdots, \widehat{M}_g\}$. Analogous to the two category case, we consider $f_g(\boldsymbol{U}) = (f_{g1}(\boldsymbol{U}), \cdots, f_{g\widehat{M}_g}(\boldsymbol{U}))^T \in \prod_{m=1}^{\widehat{M}_g}(\mathcal{H}_\mathcal{K} + R)$, $f_{gm}(\boldsymbol{U}_i) = f^*_{gm}(\boldsymbol{U}_i) - \gamma^\star_{gm}$ with $f^*_{gm}(\boldsymbol{U}) \in \mathcal{H}_\mathcal{K}$ and $\gamma^\star_{gm} \in \mathbb{R}$. In (A.1), $\left( f_g(\boldsymbol{U}_i) - \widetilde{\boldsymbol{Y}}^*_{gi} \right)_+$ means by taking the truncate function $(\cdot)_+$ componentwise, and the "$\cdot$" operation in the data fit functional indicates the Euclidean inner product, $C(\widetilde{\boldsymbol{Y}}^*_{gi})$ is a $\widehat{M}_g$ dimensional vector with 0 in the $m$th coordinate if $i$-th subjects' belong to $\widehat{G}_{gm}$, and 1 elsewhere.

In the multi-group case, partition blocks are often formed by the intersection of two or more boundary curves (e.g., Scenarios III and IV in Figure 1), making it difficult to obtain a general explicit expression for the estimated partitions. However, this does not affect group membership predictions.

## A2. Theoretical properties

### A2.1 *Notation and conditions*

For $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_n) = ([\boldsymbol{\alpha}]_1, \cdots, [\boldsymbol{\alpha}]_p)$, take $[\boldsymbol{\alpha}]_g$ as an example to illustrate the symbols'. Let $\mathcal{M}^g_\mathcal{G}$ be the subspace of $\mathbb{R}^n$, $\mathcal{M}^g_\mathcal{G} = \{[\boldsymbol{\alpha}]_g \in \mathbb{R}^n : \alpha_{ig} = \alpha_{jg}, \text{for any } i, j \in G^0_{gm}, 1 \leqslant m \leqslant M^0_g\}$. For each $[\boldsymbol{\alpha}]_g \in \mathcal{M}^g_\mathcal{G}$, it can be written as $[\boldsymbol{\alpha}]_g = \boldsymbol{W}_g \boldsymbol{\theta}_g$, where $\boldsymbol{W}_g = (\boldsymbol{W}_{1,g}, \cdots, \boldsymbol{W}_{n,g})^T$. By matrix calculation, we have $\boldsymbol{D}_g = \boldsymbol{W}_g^T \boldsymbol{W}_g = \text{diag}(|G^0_{g1}|, \cdots, |G^0_{gM^0_g}|)$, where $|G^0_{gm}|$ denotes the group

size of $G^0_{gm}$. Define $|G^0_{\min}| = \min_{1\leqslant g\leqslant p, 1\leqslant m\leqslant M^0_g} |G^0_{gm}|$, $|G^0_{\max}| = \max_{1\leqslant g\leqslant p, 1\leqslant m\leqslant M^0_g} |G^0_{gm}|$. For any positive numbers $a_n$ and $b_n$, let $a_n \gg b_n$ denote $a_n^{-1} b_n = o(1)$. For any vector $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_s)^T \in \mathbb{R}^s$, let $\|\boldsymbol{\zeta}\|_\infty = \max_{1\leqslant l\leqslant s} |\zeta_l|$. For any symmetric matrix $A_{s\times s}$, denote its $L_2$ norm by $\|A\| = \max_{\zeta\in\mathbb{R}^s, \|\zeta\|=1} \|A\zeta\|$, and let $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ be the smallest and largest eigenvalues of $A$, respectively. For any matrix $\boldsymbol{A} = (A_{ij})_{i=1,j=1}^{s,t}$, denote $\|\boldsymbol{A}\|_\infty = \max_{1\leqslant i\leqslant s} \sum_{j=1}^t |A_{ij}|$. Let $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$, $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^T$, and $\boldsymbol{Z}^g = diag(Z_{g1}, \cdots, Z_{gn})$. Denote $\widetilde{\boldsymbol{Z}}_g = \boldsymbol{Z}^g \boldsymbol{W}_g$ and $\boldsymbol{E} = (\boldsymbol{X}, \boldsymbol{Z}^1\boldsymbol{W}_1, \cdots, \boldsymbol{Z}^p\boldsymbol{W}_p)$. Finally, denoted the scaled penalty function by $\rho(t) = \lambda^{-1} p_\gamma(t, \lambda)$ and $\overline{\rho}(t) = \rho'(|t|) sgn(t)$.

In addition, we assume the threshold covariates satisfy $\boldsymbol{U}_i \overset{iid}{\sim} p(\boldsymbol{u})$, with respect to the measure space $(\boldsymbol{\mathcal{U}}, \mathcal{B}(\boldsymbol{\mathcal{U}}), v)$, where $\boldsymbol{\mathcal{U}}$ represents covariate space, $\mathcal{B}(\boldsymbol{\mathcal{U}})$ are Borel sigma algebra, induced by the metric $d_{\boldsymbol{\mathcal{U}}}$, and $v$ be a measure on $\mathcal{B}(\boldsymbol{\mathcal{U}})$. Here, we introduce some regular conditions.

(*C*1) Assume $\sum_{i=1}^n X_{ig}^2 = n$ for $1 \leqslant g \leqslant q$, and $\sum_{i=1}^n Z_{ig}^2 1\{i \in G^0_{gm}\} = |G^0_{gm}|$ for $1 \leqslant g \leqslant p$, $\lambda_{\min}(\boldsymbol{E}^T\boldsymbol{E}) \geqslant C_1 |G^0_{\min}|$, $\sup_i \|\boldsymbol{Z}_i\| \leqslant C_2\sqrt{p}$ and $\sup_i \|\boldsymbol{X}_i\| \leqslant C_3\sqrt{q}$ for some constants $0 < C_1 < \infty, 0 < C_2 < \infty$ and $0 < C_3 < \infty$.

(*C*2) The function $\rho_\gamma(t, \lambda)$ is a symmetric function of $t$, and it is non-decreasing and concave in $t$ for $t \in [0, \infty)$. $\rho(t)$ is a constant for all $t \geqslant a\lambda$ for some constant $a > 0$, and $\rho(0) = 0, \rho'(t)$ exists and is continuous except for a finite number values of $t$ and $\rho'(0+) = 1$.

(*C*3) The noise vector $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^\top$ has sub-Gaussian tails such that $pr(|\boldsymbol{A}^T\boldsymbol{\epsilon}| > \|\boldsymbol{A}\|x) \leqslant 2\exp(-c_1x^2)$ for any vector $\boldsymbol{A} \in \mathbb{R}^n$ and $x > 0$, where $0 < c_1 < \infty$.

(*C*4) The threshold covariates density $p$ satisfies $0 < p_{min} < p(\boldsymbol{u}) < p_{max}$ for all $\boldsymbol{u}$. where $p_{min}, p_{max} \in \mathbb{R}$.

(*C*5) The base measure $v$ in $\boldsymbol{\mathcal{U}}$ satisfies $c_{1,r} d^r \leqslant v\{B_d(\boldsymbol{U}) \cap \boldsymbol{\mathcal{U}}^0_{gm}\} \leqslant v\{B_d(\boldsymbol{U})\} \leqslant c_{2,r} d^r$ for all $\boldsymbol{U} \in \boldsymbol{\mathcal{U}}^0_{gm}$, $g = 1, \cdots, p, m = 1, \cdots, M^0_g$, $0 < d < d_0$. where $c_{1,r}, c_{2,r}, d_0$ are positive constants, $B_d(\boldsymbol{U}) = \{\boldsymbol{U}' : \boldsymbol{U}' \in \boldsymbol{\mathcal{U}} \text{ with } d_{\boldsymbol{\mathcal{U}}}(\boldsymbol{U}, \boldsymbol{U}') \leqslant d\}$.

(*C*6) Let $c_{n1} = \max_{1\leqslant j\leqslant q_s}\{p'(|\beta_j|, \lambda_{s2})\}$ and $c_{n2} = \max_{1\leqslant j\leqslant q_s}\{p''(|\beta_j|, \lambda_{s2})\}$. Assume $c_{n1} =$

$0(1/\sqrt{nq})$, $\liminf_{n\to\infty}\liminf_{\theta\to 0+} p'(\theta,\lambda_{s2})/\lambda_s\} > 0$, $c_{n2}\to 0$, as $n\to\infty$, $c_{n2} = o_p(1/\sqrt{q})$, and there are constants $c_2, c_3$ such that, when $\theta_1,\theta_2 > c_1\lambda_{s2}$, $|p''(\theta_1,\lambda_{s2}) - p''(\theta_2,\lambda_{s2})| \leqslant c_3|\theta_1-\theta_2|$.

(*C*7) Assume $\min_{1\leqslant j\leqslant q_s}|\beta^0_{s,j}|/\lambda_{s2}\to\infty$, as $n\to\infty$.

It is commonly assumed that the smallest eigenvalue of the transpose of the design matrix multiplied by the design matrix is bounded by $C_1 n$, which may not hold for $\boldsymbol{E}^T\boldsymbol{E}$. By some calculation and $\widetilde{\boldsymbol{Z}}_g = \boldsymbol{Z}^g\boldsymbol{W}_g, g = 1,\cdots p$, we have $\widetilde{\boldsymbol{Z}}_g^T\widetilde{\boldsymbol{Z}}_g = \sum_{i\in G_{gm}} Z_{ig}^2 = |G^0_{gm}|, m = 1,\cdots, M^0_g, g = 1,\cdots,p$. By assuming that $\lambda_{\min}(\widetilde{\boldsymbol{Z}}_g^T\widetilde{\boldsymbol{Z}}_g) \geqslant c|G^0_{gm}|$ for some constant $0 < c < \infty$. If $\boldsymbol{X}^T\widetilde{\boldsymbol{Z}}_g = 0(g = 1,\cdots p)$ and $\lambda_{\min}\ (\boldsymbol{X}^T\boldsymbol{X}) \geqslant Cn$, we have $\lambda_{\min}(\boldsymbol{E}^T\boldsymbol{E}) \geqslant \min\{\lambda_{\min}(\boldsymbol{X}^T\boldsymbol{X}), \lambda_{\min}(\widetilde{\boldsymbol{Z}}_1^T\widetilde{\boldsymbol{Z}}_1),\cdots,\lambda_{\min}(\widetilde{\boldsymbol{Z}}_p^T\widetilde{\boldsymbol{Z}}_p)\} = \min(|G^0_{\min}|, Cn)$, and $|G^0_{\min}| \leqslant n/(\max_{1\leqslant g\leqslant p} M^0_g)$. Therefore, we let the smallest eigenvalue in Condition (*C*1) be bounded below by $C_1|G^0_{\min}|$. Conditions (*C*2)–(*C*3) are common assumptions in penalized regression in high-dimensional settings, with concave penalties such as minimax concave penalty and smoothly clipped absolute deviation penalty satisfying Condition (*C*2). Conditions (*C*4)–(*C*5) are similar to those in Madrid Padilla et al. (2020), ensuring that $\nabla G$ can effectively capture the "similarity" among threshold variables $Ui$, and each vertex in $\nabla_G$ is connected to some vertex within the same group. Conditions (*C*6)–(*C*7) are conventional conditions used to ensure the validity of variable selection (Fan and Peng, 2004)

Remark 1: Since $\left|G^0_{min}\right| \leqslant n/(\max_{1\leqslant g\leqslant p} M^0_g)$, by the condition $\left|G^0_{min}\right| \gg \sqrt{(q+\sum_{g=1}^p M^0_g)n\log n}$, then $q, p, M^0_g(g = 1\ldots, p)$ must satisfy $(\max_{1\leqslant g\leqslant p} M^0_g)\sqrt{(q+\sum_{g=1}^p M^0_g)} = o\{\sqrt{n(\log n)^{-1}}\}$.

A2.2 *Error analysis of multi-partition recovery*

We establish the Fisher consistency of the proposed partition recovery method based on the support vector machine 1-norm soft margin classifier within the regularization framework, as outlined in (9). To do so, we begin with some notation for the classification rule. Define $P_g(\cdot \mid \boldsymbol{u})$ to be the conditional probability measure of $\widetilde{Y}^*_g$ given $\boldsymbol{u}$, consider $\widetilde{Y}^*_g \in \{1,\ldots,\widehat{M}_g\}$. Thus, the misclassification error for a classifier $f_g : \boldsymbol{U}\to\widetilde{Y}^*_g$ is defined to be the probability of the event

$\{f_g(\boldsymbol{U}) \neq \widetilde{Y}_g^*\}$, i.e $\mathcal{R}(f_g(\boldsymbol{U})) = pr(f_g(\boldsymbol{U}) \neq \widetilde{Y}_g^*)$. The classifier minimizing the misclassification error is called the Bayes rule $\widehat{f}_{gb}(\boldsymbol{U})$.

Similarly, when $M_g^0 > 2$, if $i$th subjects' belong to $G_{gm}^0$, let $\widetilde{Y}_{gi}^{*0} = m$, $\widetilde{\boldsymbol{Y}}_{gi}^{*0}$ is a $M_g^0$-dimensional vector with 1 in the $m$th coordinate and $-1/(M_g^0 - 1)$ elsewhere and $\boldsymbol{C}(\widetilde{\boldsymbol{Y}}_{gi}^{*0})$ is a $M_g^0$-dimensional vector with 0 in the $m$th coordinate and 1 elsewhere . Motivated by Lee et al. (2004), define the loss function of classification $V(\widetilde{\boldsymbol{Y}}_g^{*0}, f_g(\boldsymbol{U})) = (\widetilde{\boldsymbol{Y}}_g^{*0}) \cdot \left(f_g(\boldsymbol{U}) - \widetilde{\boldsymbol{Y}}_g^{*0}\right)_+$, the generalization error $\mathcal{E}(f_g(\boldsymbol{U})) = EV(\widetilde{\boldsymbol{Y}}_g^{*0}, f_g(\boldsymbol{U}))$ and the empirical error is $\mathcal{E}_z(f_g(\boldsymbol{U})) = \frac{1}{n}\sum_{i=1}^n V(\widetilde{\boldsymbol{Y}}_{gi}^{*0}, f_g(\boldsymbol{U}_i))$. If the underlying group memberships $G_{g1}^0, \cdots, G_{gM_g^0}^0$ for $g = 1, \cdots, p$ were konwn, let $f_{gb}^{or}(\boldsymbol{U})$ be the minimizer of $\mathcal{E}(f_g(\boldsymbol{U}))$, similar to Lemma 4.1 of Lee et al. (2004), we have $f_{gb}^{or}(\boldsymbol{U}) = ([f_{gb}^{or}(\boldsymbol{U})]_1, \cdots, [f_{gb}^{or}(\boldsymbol{U})]_{M_g^0})$ with

$$[f_{gb}^{or}(\boldsymbol{U})]_m = \begin{cases} 1, & if\ m = \arg\max_m p_{gm}(\boldsymbol{U}), \\ -\frac{1}{\widehat{M}_g - 1}, & \text{otherwise}, \end{cases}$$

where $[\cdot]_m$ represents the $m$th component of the vector, $p_{gm}(\boldsymbol{U}) = P_g(\widetilde{Y}_{gi}^{*0} = m \mid \boldsymbol{U} = \boldsymbol{u})$ for $g = 1, \cdots, p, m = 1, \cdots, M_g^0$. The true oracle Bayes rule $f_{gb}^{or}(\boldsymbol{U}) = \arg\max_{1 \leqslant m \leqslant M_g^0} [f_{gb}^{or}(\boldsymbol{U})]_m = \arg\max_{1 \leqslant m \leqslant M_g^0} p_{gm}(\boldsymbol{U})$.

The following theorem establishes that the estimated classifier rule $\widehat{\phi}_g(\boldsymbol{U}) = \arg\max_m \widehat{f}_{gm}(\boldsymbol{U})$, is Fisher consistent and asymptotically equivalent to the true oracle Bayes rule as the sample size tends to infinity, provided the number of subgroups is greater than 2.

THEOREM A.0: *Consider $M_0^g > 2$, under the conditions in Theorem 2, for every $\lambda_n > 0$ satisfy $\lambda_n \to 0$ and $\frac{1}{n\widehat{M}_g\sqrt{\lambda_n}} \log\left(\mathcal{N}\left(\sqrt{\lambda_n}\right)\right) \to 0$, where $\mathcal{N}(\epsilon)$ is the covering number. Besides, define the regularization error $\mathcal{D}(\lambda_n) = \inf_{f_g \in \prod_{m=1}^{M_g^0} \overline{\mathcal{H}}_{\mathcal{K}}} \left\{\mathcal{E}(f_g(\boldsymbol{U})) - \mathcal{E}(f_{gb}^{or}(\boldsymbol{U})) + \frac{1}{2}\lambda_n \sum_{m=1}^{\widehat{M}_g} \|f_{gm}^*\|_{\mathcal{K}}^2\right\}$, where $f_{gb}^{or}(\boldsymbol{U})$ is the minimizer of $\mathcal{E}(f_g(\boldsymbol{U}))$, $\sum_{m=1}^{M_g^0} f_{gM_g^0}(\boldsymbol{U}) = 0$. As $\lim_{\lambda_n \to 0} \mathcal{D}(\lambda_n) = 0$, we have $pr\left\{\mathcal{R}(\widehat{\phi}_g(\boldsymbol{U})) = \mathcal{R}(f_{gb}^{or}(\boldsymbol{U}))\right\} \to 1$ for $g = 1, \cdots, p$.*

We do not present the property $pr(\widehat{\mathcal{U}}_{g1} = \mathcal{U}_{g1}^0, \cdots, \widehat{\mathcal{U}}_{g\widehat{M}_g} = \mathcal{U}_{gM_g^0}^0) \to 1$ in this theorem, as it is

challenging to obtain a general explicit expression for the estimated partition blocks, as discussed in Section 3.2. However, this property also holds for the estimated partition blocks of the underlying potential groups.

## A3. Computation

A3.1 *Coordinate alternating directions method of multipliers algorithm*

Firstly, the problem of minimizing the objective (8) can be reformulated as

$$\arg\min_{\boldsymbol{\beta},\boldsymbol{\alpha}_i,\boldsymbol{\delta}} \sum_{i=1}^{n} \frac{1}{2}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} - \boldsymbol{Z}_i^T\boldsymbol{\alpha}_i)^2 + \sum_{i=1}^{n}\sum_{j\in\mathcal{N}_K^i}\sum_{g=1}^{p} p(|[\delta_{ij}]_g|, \lambda_g),$$

$$s.t.\ \nabla_G\ [\boldsymbol{\alpha}]_g = \boldsymbol{\delta}_g, \tag{A.2}$$

where $\boldsymbol{\delta}_g = ([\boldsymbol{\delta}_{1j_1}]_g, \cdots, [\boldsymbol{\delta}_{nj_n}]_g, j_l \in \mathcal{N}_K^l)^T$, $[\boldsymbol{\delta}_{ij}]_g = \alpha_{ig} - \alpha_{jg}$, and $\nabla_G$ is an adjacency matrix of the graph $G$. And graph $G$ is constructed based on the second neighbor definition, $\mathcal{N}_K^i = \{j : \boldsymbol{U}_j$ is one of the $K$ nearest to $\boldsymbol{U}_i$ among all the individuals $\}$. We define $\nabla_G$ as follows: each row of the matrix corresponds to one edge in $G$; for instance, if the $l$-th edge in $G$ connects the $\boldsymbol{U}_i$ and $\boldsymbol{U}_j$, then

$$(\nabla_G)_{l,l'} = \begin{cases} 1, \text{ if } l' = i, \\ -1, \text{ if } l' = j, \\ 0, \ otherwise. \end{cases}$$

Model (A.2) can be reformulated as

$$L_0(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\delta}) = \frac{1}{2}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} - \boldsymbol{Z}_i^T\boldsymbol{\alpha}_i)^2 + \sum_{i=1}^{n}\sum_{j\in\mathcal{N}_K^i}\sum_{g=1}^{p} p(|[\delta_{ij}]_g|, \lambda_g),$$

$$s.t.\ \nabla_G\ [\boldsymbol{\alpha}]_g - \boldsymbol{\delta}_g = \mathbf{0}, g = 1, \cdots p.$$

where $\boldsymbol{\delta}_g = ([\boldsymbol{\delta}_{1j_1}]_g, \cdots, [\boldsymbol{\delta}_{1j_K}]_g, \cdots, [\boldsymbol{\delta}_{nj_1}], \cdots, [\boldsymbol{\delta}_{nj_K}]_g, j_l \in \mathcal{N}_K^l)'$, $[\boldsymbol{\delta}_{ij}]_g = \alpha_{ig} - \alpha_{jg}$. The augmented Larangian is

$$L(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\nu}) = L_0(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\delta}) + \sum_{g=1}^{p}\langle \boldsymbol{\nu}_g, \nabla_G[\boldsymbol{\alpha}]_g - \boldsymbol{\delta}_g\rangle + \frac{\upsilon}{2}\sum_{g=1}^{p} \| \nabla_G\ [\boldsymbol{\alpha}]_g - \boldsymbol{\delta}_g\|^2,$$

where the dual variables $\boldsymbol{\nu}_g = \{\nu_{g1}, \cdots, \nu_{g(nK)}\}, g = 1, \cdots p$ are Lagrange multipliers and $\upsilon$ is a penalty parameter. For a given value of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ at step $t$, we can update the estimator according

following steps:

$$
\begin{aligned}
(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\alpha}^{(t+1)}) &= \arg\min_{\boldsymbol{\beta},\boldsymbol{\alpha}} L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\delta}^{(t)}, \boldsymbol{\nu}^{(t)}), && \text{(A.3)}\\
\boldsymbol{\delta}^{(t+1)} &= \arg\min_{\boldsymbol{\delta}} L(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}, \boldsymbol{\nu}^{(t)}), && \text{(A.4)}\\
\boldsymbol{\nu}_g^{(t+1)} &= \boldsymbol{\nu}_g^{(t)} + \upsilon(\triangledown_G [\boldsymbol{\alpha}]_g^{(t+1)} - \boldsymbol{\delta}_g^{(t+1)}). && \text{(A.5)}
\end{aligned}
$$

Minimizing $L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\delta}^{(t)}, \boldsymbol{\nu}^{(t)})$ in (A.3) is equivalent to minimizing following function

$$
f(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^T\boldsymbol{\beta} - \mathbf{Z}_i^T\boldsymbol{\alpha}_i)^2 + \frac{\upsilon}{2}\sum_{g=1}^{p} \| \triangledown_G\, [\boldsymbol{\alpha}]_g - \boldsymbol{\delta}_g^{(t)} + \upsilon^{-1}\boldsymbol{\nu}_g^{(t)}\|^2 + C,
$$

where $C$ is a constant independent of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$. It easy to see that we can write $f(\boldsymbol{\beta}, \boldsymbol{\alpha})$ as

$$
f(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{2}\|(\mathbf{X}\boldsymbol{\beta} + \sum_{g=1}^{p}\mathbf{Z}^g[\boldsymbol{\alpha}]_g - \boldsymbol{Y})\|^2 + \frac{\upsilon}{2}\sum_{g=1}^{p} \| \triangledown_G\, [\boldsymbol{\alpha}]_g - \boldsymbol{\delta}_g^{(t)} + \upsilon^{-1}\boldsymbol{\nu}_g^{(t)}\|^2 + C, \qquad \text{(A.6)}
$$

where $\boldsymbol{X}$ is a $n \times q$ dimension matrix, $\mathbf{Z}^g = diag(Z_{g1}, \cdots, Z_{gn})$. Thus for given $\boldsymbol{\nu}^{(t)}$ and $\boldsymbol{\delta}^{(t)}$ at the $t$-th step, we can update $[\boldsymbol{\alpha}]_g^{(t+1)}$ and $\boldsymbol{\beta}^{(t+1)}$ via following iteration steps

$$
\begin{aligned}
&[\boldsymbol{\alpha}]_1^{(t+1)} = ((\boldsymbol{Z}^1)^T\boldsymbol{Q}_x\boldsymbol{Z}^1 + \upsilon\, \triangledown_G^T\, \triangledown_G)^{-1}[(\boldsymbol{Z}^1)^T\boldsymbol{Q}_x(\boldsymbol{Y} - \sum_{g=2}^{p}\boldsymbol{Z}^g[\boldsymbol{\alpha}]_g^{(t)}) + \upsilon\, \triangledown_G^T\, (\boldsymbol{\delta}_1^{(t)} - \upsilon^{-1}\boldsymbol{\nu}_1^{(t)})],\\
&\vdots\\
&[\boldsymbol{\alpha}]_j^{(t+1)} = ((\boldsymbol{Z}^j)^T\boldsymbol{Q}_x\boldsymbol{Z}^j + \upsilon\, \triangledown_G^T\, \triangledown_G)^{-1}[\boldsymbol{Z}^j\mathbf{Q}_x(Y - \sum_{g=1}^{j-1}\mathbf{Z}^g[\boldsymbol{\alpha}]_g^{(t+1)} - \sum_{g=j+1}^{p}\mathbf{Z}^g[\boldsymbol{\alpha}]_g^{(t)})\\
&\qquad\qquad +\upsilon\, \triangledown_G^T\, (\boldsymbol{\delta}_j^{(t)} - \upsilon^{-1}\boldsymbol{\nu}_j^{(t)})],\\
&\vdots\\
&[\boldsymbol{\alpha}]_p^{(t+1)} = ((\mathbf{Z}^p)^T\mathbf{Q}_x\mathbf{Z}^p + \upsilon\, \triangledown_G^T\, \triangledown_G)^{-1}[(\mathbf{Z}^p)^T\mathbf{Q}_x(\boldsymbol{Y} - \sum_{g=1}^{p-1}\mathbf{Z}^g[\boldsymbol{\alpha}]_g^{(t+1)}) + \upsilon\, \triangledown_G^T\, (\boldsymbol{\delta}_p^{(t)} - \upsilon^{-1}\boldsymbol{\nu}_p^{(t)})],\\
&\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\boldsymbol{Y} - \sum_{g=1}^{p}\mathbf{Z}^g[\boldsymbol{\alpha}]_g^{(t+1)}), \qquad \text{(A.7)}
\end{aligned}
$$

where $\mathbf{Q}_x = \boldsymbol{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Note that the dimension of the entire vector $\boldsymbol{\alpha} = ([\boldsymbol{\alpha}]_1, \cdots, [\boldsymbol{\alpha}]_p)^T$ is $np$, which increases linearly with the sample size $n$. Therefore, we implement a coordinate (blockwise) alternating directions method of multipliers algorithm to manage the computational

burden efficiently. In (A.4), after discarding the terms independent of $\boldsymbol{\delta}$, we need to minimize

$$\begin{cases} \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_K^i} p(|[\boldsymbol{\delta}_{ij}]_1|, \lambda) + \frac{\upsilon}{2} \| \nabla_G [\boldsymbol{\alpha}]_1^{(t)} + \upsilon^{-1} \boldsymbol{\nu}_1^{(t)} - \boldsymbol{\delta}_1 \|^2, \\ \vdots \\ \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_K^i} p(|[\boldsymbol{\delta}_{ij}]_k|, \lambda) + \frac{\upsilon}{2} \| \nabla_G [\boldsymbol{\alpha}]_k^{(t)} + \upsilon^{-1} \boldsymbol{\nu}_k^{(t)} - \boldsymbol{\delta}_k \|^2, \\ \vdots \\ \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_K^i} p(|[\boldsymbol{\delta}_{ij}]_p|, \lambda) + \frac{\upsilon}{2} \| \nabla_G [\boldsymbol{\alpha}]_p^{(t)} + \upsilon^{-1} \boldsymbol{\nu}_p^{(t)} - \boldsymbol{\delta}_p \|^2, \end{cases} \tag{A.8}$$

with respect to $\boldsymbol{\delta}_g$, where $\boldsymbol{\zeta}_g^{(t)} = ([\boldsymbol{\zeta}_{1j_1}^{(t)}]_g, \cdots, [\boldsymbol{\zeta}_{1j_K}^{(t)}]_g, \cdots, [\boldsymbol{\zeta}_{nj_1}^{(t)}]_g, \cdots, [\boldsymbol{\zeta}_{nj_K}^{(t)}]_g)^T = \nabla_G [\boldsymbol{\alpha}]_g^{(t)} + \upsilon^{-1} \boldsymbol{\nu}_g^{(t)}$. This is a groupwise thresholding operator corresponding to $\rho_\lambda$. For the lasso penalty, the solution is

$$[\boldsymbol{\delta}_{ij}^{(t+1)}]_g = \boldsymbol{S}([\boldsymbol{\zeta}_{ij}^{(t)}]_g, \lambda/\upsilon), \tag{A.9}$$

where $S(z, t) = (1 - t/|z|)_+ z$ is the groupwise soft thresholding operator. Here $(x)_+ = x$ if $x > 0$ and $= 0$, otherwise. For the the minimax concave penalty with $\gamma > 1/\upsilon$, the solution is

$$[\boldsymbol{\delta}_{ij}^{(t+1)}]_g = \begin{cases} \frac{S([\boldsymbol{\zeta}_{ij}^{(t)}]_g, \lambda/\upsilon)}{1 - 1/(\gamma\upsilon)}, \text{if } |[\boldsymbol{\zeta}_{ij}^{(t)}]_g| \leqslant \gamma\lambda, \\ [\boldsymbol{\zeta}_{ij}^{(t)}]_g, \text{if } |[\boldsymbol{\zeta}_{ij}^{(t)}]_g| > \gamma\lambda. \end{cases} \tag{A.10}$$

For the smoothly clipped absolute deviation penalty with $\gamma > 1/\upsilon + 1$, the solution is

$$[\boldsymbol{\delta}_{ij}^{(t+1)}]_g = \begin{cases} S([\boldsymbol{\zeta}_{ij}^{(t)}]_g, \lambda/\upsilon), \text{if } |[\boldsymbol{\zeta}_{ij}^{(t)}]_g| \leqslant \lambda + \lambda/\upsilon, \\ \frac{S([\boldsymbol{\zeta}_{ij}^{(t)}]_g, \lambda\gamma/((\gamma-1)\upsilon))}{1 - 1/((\gamma-1)\upsilon)}, \text{if } \lambda + \lambda/\upsilon < |[\boldsymbol{\zeta}_{ij}^{(t)}]_g| \leqslant \gamma\lambda, \\ [\boldsymbol{\zeta}_{ij}^{(t)}]_g, \text{if } |[\boldsymbol{\zeta}_{ij}^{(t)}]_g| > \gamma\lambda. \end{cases} \tag{A.11}$$

Finally, the update of $\boldsymbol{\nu}_g$ is given in (A.5).

Instead of minimizing the objective function with respect to the entire vector $\boldsymbol{\alpha}$, the proposed coordinate alternating directions method of multipliers minimizes it with respect to the subvector $[\boldsymbol{\alpha}]_g$ at a time. Compared to the standard alternating directions method of multipliers, the coordinate alternating directions method of multipliers effectively reduces the computational burden.

Similarly, the problem of minimizing the objective (11) can be reformulated as

$$\arg\min_{\boldsymbol{\beta},\boldsymbol{\alpha}_i,\boldsymbol{\delta}} \sum_{i=1}^{n} \frac{1}{2}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} - \boldsymbol{Z}_i^T\boldsymbol{\alpha}_i)^2 + \sum_{i=1}^{n}\sum_{j\in\mathcal{N}_K^i}\sum_{g=1}^{p} p(|[\delta_{ij}]_g|, \lambda_g) + \sum_{g=1}^{q} p(|\beta_g|, \lambda_s),$$
$$s.t.\ \nabla_G\,[\boldsymbol{\alpha}]_g = \boldsymbol{\delta}_g, \tag{A.12}$$

where $\boldsymbol{\delta}$ and adjacency matrix is same as above. The implementation can be conducted using the proposed algorithm with the updates $(\boldsymbol{\beta},\boldsymbol{\alpha})$ at step $t+1$ replaced by the following procedure:

$$\begin{aligned}
\boldsymbol{\beta}^{(t+1)} &= \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|(\boldsymbol{X}\boldsymbol{\beta} + \sum_{g=1}^{p}\boldsymbol{Z}^g[\boldsymbol{\alpha}]_g^{(t)} - \boldsymbol{Y})\|^2 + \sum_{g=1}^{q} p(|\boldsymbol{\beta}_g|, \lambda_{s2}),\\
\boldsymbol{\alpha}^{(t+1)} &= \arg\min_{\boldsymbol{\alpha}} \frac{1}{2}\|(\boldsymbol{X}\boldsymbol{\beta}^{(t+1)} + \sum_{g=1}^{p}\boldsymbol{Z}^g[\boldsymbol{\alpha}]_g - \boldsymbol{Y})\|^2 + \frac{\upsilon}{2}\sum_{g=1}^{p}\|\nabla_G\,[\boldsymbol{\alpha}]_g - \boldsymbol{\delta}_g^{(t)} + \upsilon^{-1}\boldsymbol{\nu}_g^{(t)}\|^2
\end{aligned}$$

The update of $\boldsymbol{\delta}$ and $\boldsymbol{\nu}$ is same as $(A.10)$ and $(A.5)$.

### A3.2 *Convergence of the algorithm*

In this section, we derive the convergence and accuracy properties of the proposed algorithm.

Proposition 1: Let $\boldsymbol{r}_g^{(t)} = \nabla_G[\boldsymbol{\alpha}]_g^{(t)} - \boldsymbol{\delta}_g^{(t)}$ and $\boldsymbol{s}_g^{(t+1)} = \upsilon\,\nabla_G^T\,(\boldsymbol{\delta}_g^{(t+1)} - \boldsymbol{\delta}_g^{(t)}), g = 1\cdots, p$ be the primal residual and the dual residual in the block-wise alternating directions method of multipliers described above, respectively. It holds that $\lim_{t\to\infty}\|\boldsymbol{r}^{(t)}\|^2 = 0$ and $\lim_{t\to\infty}\|\boldsymbol{s}^{(t)}\|^2 = 0$ for the minimax concave penalty penalties. where $\boldsymbol{r} = (\boldsymbol{r}_1^T, \cdots, \boldsymbol{r}_p^T), \boldsymbol{s} = (\boldsymbol{s}_1^T, \cdots, \boldsymbol{s}_p^T)$.

This proposition demonstrates that both primal and dual feasibility are achieved by the alternating directions method of multipliers algorithm. The proof can be readily proved by following the proof of the corresponding results in Ma et al. (2020).

### A3.3 *Initial values choose*

To use the alternating directions method of multipliers algorithm described above, it is important to choose a reasonable initial value firstly. For this purpose, we consider the ridge fusion criterion given by

$$L_R(\boldsymbol{\beta},\boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} - \boldsymbol{Y}\|^2 + \frac{\lambda^*}{2}\sum_{g=1}^{p}\|\nabla_G\,[\boldsymbol{\alpha}]_g\|^2,$$

where $\lambda^*$ is the tuning parameter having a small value and $\mathbf{Z} = diag(\mathbf{Z}_1^T, \cdots, \mathbf{Z}_n^T), \boldsymbol{\alpha} \in \mathbb{R}^{np}$. Note that $L_R(\boldsymbol{\beta}, \boldsymbol{\alpha})$ can be written as

$$L_R(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} - \boldsymbol{Y}\|^2 + \frac{\lambda^*}{2}\|\mathbf{A}\boldsymbol{\alpha}\|^2,$$

where $\mathbf{A}$ is defined $\mathbf{A} = \nabla_G \otimes I_p$. The solutions are

$$\begin{aligned}
\boldsymbol{\alpha}_R(\lambda^*) &= (\boldsymbol{\alpha}_{R,1}^T(\lambda^*), \cdots, \boldsymbol{\alpha}_{R,n}^T(\lambda^*))^T = (\mathbf{Z}^T\mathbf{Q}_x\mathbf{Z} + \lambda^*\mathbf{A}^T\mathbf{A})^{-1}\mathbf{Z}^T\mathbf{Q}_x\boldsymbol{Y}, \\
\boldsymbol{\beta}_R(\lambda^*) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\boldsymbol{Y} - \mathbf{Z}\boldsymbol{\alpha}_R(\lambda^*)).
\end{aligned} \tag{A.13}$$

where $\mathbf{Q}_x = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Next, we can rewrite $\boldsymbol{\alpha}_R(\lambda^*)$ as matrix $(\boldsymbol{\alpha}_{R1}(\lambda^*), \cdots, \boldsymbol{\alpha}_{Rp}(\lambda^*))$. For the $g$-th component, we assign the subjects to $M_g^*$ groups by ranking the medians values of $\boldsymbol{\alpha}_{Rg,i}(\lambda^*)$. Let $M_g^* = \lfloor n^{1/2} \rfloor$ to ensure that it is sufficiently large, where $\lfloor a \rfloor$ denotes the largest integer no greater than $a$.

After that, let $M_g^*$ medians as the initial values $([\boldsymbol{\alpha}]_g)^*$, $\boldsymbol{\beta}^* = \boldsymbol{\beta}_R(\lambda^*)$. And $\boldsymbol{\delta}_g^*$ component $\delta_{g,ij}^* = \alpha_{ig}^* - \alpha_{jg}^*$ for $i = 1, \cdots, n$, $j = 1, \cdots, K$ which $j \in \mathcal{N}_K^i$. The specific process of initial value selection is shown in algorithm 1.

### A3.4 *Additional simulation study*

We consider the following cases and provide the performance results for group identification and coefficient estimation. Case 6 introduces a general threshold change plane model with varying numbers of subgroups and partition boundaries for different covariates, where boundaries within the same coordinate remain parallel. Case 7 is an extension of Case 1, where the number of coefficient groups is 3. The corresponding partition estimates differ for the different components of $\mathbf{Z}_i$ , while the univariate outcome $\mathbf{U}_i$ remains the same. Case 8 is an extension of Case 2, where the partition estimates again differ for the components of $\mathbf{Z}_i$. Unlike Case 4, Case 9 assumes that the number of subgroups is the same across all components of $\mathbf{Z}_i$. Case 10 considers the scenario where $\mathbf{U}_i$ includes an interaction term. The specific setting is given as follows.

[Figure 1 about here.]

**Algorithm 1** The initial value of the alternating directions method of multipliers

**Input:** Number of nearest neighbor: $K$; tuning parameter: $\lambda^*$; Number of groups: $M_g^* = \lfloor n^{1/2} \rfloor$ for $g = 1, \cdots, p$. Size of every groups: $\boldsymbol{d} = (d_1, \cdots, d_{\lfloor n^{1/2} \rfloor})$ in which $\max_i\{d_i\} - \min_j\{d_j\} \leqslant 1$ and $\sum_{i=1}^{\lfloor n^{1/2} \rfloor} d_i = n$

**Data:** $\mathbf{X} \in \mathbb{R}^{n\times q}$,$\mathbf{Z} \in \mathbb{R}^{n\times p}$, $\boldsymbol{Y} \in \mathbb{R}^n$, $\mathbf{U} \in \mathbb{R}^{n\times r}$

**Output:** $\boldsymbol{\beta}^* \in \mathbb{R}^q$, $[\boldsymbol{\alpha}]_g^* \in \mathbb{R}^n$, $\boldsymbol{\delta}_g^* \in \mathbb{R}^n$, $(g = 1, \cdots, p)$

**1.** Compute incidence matrix $\nabla_G$ from $\mathbf{U}$.

**2.** Based (A.13) calculation $\boldsymbol{\alpha}_R(\lambda^*) = (\boldsymbol{\alpha}_{R,1}^T(\lambda^*), \cdots, \boldsymbol{\alpha}_{R,n}^\top(\lambda^*))^T$ and $\boldsymbol{\beta}_R(\lambda^*)$, for $g = 1, \cdots, p$

(a) let $\boldsymbol{\alpha}_{Rg}(\lambda^*) = \{$ the $g$-th component of $\boldsymbol{\alpha}_{R,i}(\lambda^*)$ for $i = 1, \cdots, n\}$, ranking the elements in $\boldsymbol{\alpha}_{Rg}(\lambda^*)$ and divide them into $M_g^*$ groups with $d$.

(b) The median for each group was calculated separately, as the initial value for this group, merge this medians and recorded as $[\boldsymbol{\alpha}^{med}]_g$

**3.** Let $[\boldsymbol{\alpha}^*]_g = [\boldsymbol{\alpha}^{med}]_g$ for $g = 1, \cdots, p$ and $\boldsymbol{\beta}_{Rg}(\lambda^*) = \boldsymbol{\beta}^*$

In Cases 6-10, the random error values of $\epsilon_i$ are generated from $N(0, 0.5^2)$, $\mathbf{X}_i = (1, X_{1i}, X_{2i})^T$, $\mathbf{Z}_i = (Z_{1i}, Z_{2i})^T$, the population parameters $\boldsymbol{\beta}^0 = (1, 1, 1)^T$. The latent partition of $\mathbf{U}$, i.e., the subgroup structures, are depicted in Figure A.1 (Scenario $V$ - Scenario $VIII$), corresponding to Cases 6-9. Under Case 2, we consider Case 6 with shown in Figure A.1 Scenario $V$, $\alpha_{1i}^0 = \theta_{11}^0$ for $i \in \mathcal{U}_{11}^0$, $\alpha_{1i}^0 = \theta_{12}^0$ for $i \in \mathcal{U}_{12}^0$, $\alpha_{1i}^0 = \theta_{13}^0$ for $i \in \mathcal{U}_{13}^0$, and $\alpha_{2i}^0 = \theta_{21}^0$ for $i \in \mathcal{U}_{21}^0$, $\alpha_{2i}^0 = \theta_{22}^0$ for $i \in \mathcal{U}_{22}^0$. Where $(\theta_{11}^0, \theta_{12}^0, \theta_{13}^0) = (-2, 0, 2)$, $(\theta_{21}^0, \theta_{22}^0) = (-1.5, 0.5)$. The remaining settings $\mathbf{X}_i, \mathbf{Z}_i, \mathbf{U}_i$ are same as Case 2. In Case 7, we assume $X_{1i}$ generated from $U(0, 1)$, $X_{2i}$ generated from $N(0, 1)$ and $U_{1i} = X_{1i}$, the group structure as shown in Figure A.1 Scenario $VI$. $\alpha_{1i}^0 = \theta_{11}^0$ for $i \in \mathcal{U}_{11}^0$, $\alpha_{1i}^0 = \theta_{12}^0$ for $i \in \mathcal{U}_{12}^0$, $\alpha_{1i}^0 = \theta_{13}^0$ for $i \in \mathcal{U}_{13}^0$, and $\alpha_{2i}^0 = \theta_{21}^0$ for $i \in \mathcal{U}_{21}^0$, $\alpha_{2i}^0 = \theta_{22}^0$ for $i \in \mathcal{U}_{22}^0$, $\alpha_{2i}^0 = \theta_{23}^0$ for $i \in \mathcal{U}_{23}^0$, where $(\theta_{11}^0, \theta_{12}^0, \theta_{13}^0) = (-2, 0, 2)$, $(\theta_{21}^0, \theta_{22}^0, \theta_{23}^0) = (-1.5, 0.5, 2.5)$. Two settings for the covariate $(Z_{1i}, Z_{2i})$ are similar to Case 1. In Case 8, $(Z_{1i}, Z_{2i})$ are generated from $N((1, 2)^T, \boldsymbol{\Sigma}_1)$ where $\boldsymbol{\Sigma}_1 = \{\sigma_{jj'}\}, \sigma_{jj} = 1$ and $\sigma_{jj'} = 0.3$ for $j \neq j'$. The settings for

the covariate $(X_{1i}, X_{2i})$ and $(U_{1i}, U_{2i})$ are similar to Case 2. The group structure as shown in Figure A.1 Scenario $VII$, where $(\theta^0_{11}, \theta^0_{12}, \theta^0_{13}) = (-2, 0, 2), (\theta^0_{21}, \theta^0_{22}, \theta^0_{23}) = (-1.5, 0.5, 2.5)$. Under Case 3 covariate setting, we consider Case 9 with shown in Figure A.1 Scenario $VIII$. In Case 10, $X_{1i}, Z_{1i}$ independent generated from $U(0, 1)$, $X_{2i}, Z_{2i}$ independent generated from $N(0, 1)$, the random error values of $\epsilon_i$ are generated from $N(0, 0.5^2)$. The heterogeneity parameters $\alpha^0_{1i} = \theta^0_{11}$ for $i \in \mathcal{U}^0_{11}$, $\alpha^0_{1i} = \theta^0_{12}$ for $i \in \mathcal{U}^0_{12}$, and $\alpha^0_{2i} = \theta^0_{21}$ for $i \in \mathcal{U}^0_{21}$, $\alpha^0_{2i} = \theta^0_{22}$ for $i \in \mathcal{U}^0_{22}$. Where $(\theta^0_{11}, \theta^0_{12}) = (-1, 1), (\theta^0_{21}, \theta^0_{22}) = (-1.5, 0.5)$, $\mathcal{U}^0_{11} = \mathcal{U}^0_{21} = \{U_i \leqslant 0.2\}$ and $\mathcal{U}^0_{11} = \mathcal{U}^0_{21} = \{U_i > 0.2\}$. Simulation experiment are based on sample size $n = 400$ or $n = 800$ and $B = 200$ replicates.

To evaluate the performance of the proposed subgroup identification, Table A.1 reports the median, bias and standard (s.d.) of the estimated number of groups, the average value of Rand Index (RI) for measuring clustering accuracy, and the percentage (per) of $\widehat{M}_g$ equaling to the true number of subgroups. From Table A.1, we observe that the medians of $\widehat{M}_1, \widehat{M}_2$ match the true number of subgroups across all cases. As the sample size $n$ increases, both the RI values and the proportion of correctly selecting the number of subgroups approach 1, indicating enhanced clustering performance.

[Table 1 about here.]

To evaluate the performance of partition recovery, Table A.1 presents the average accuracy (ACC) of the partition recovery. The value of $\text{ACC}_g, g = 1, 2$ and ACC defined by $\text{ACC}_g = \frac{1}{n}\sum_{m=1}^{M^0_g}\sum_{i\in G^0_{gm}} I\{w^g_{im} = \widehat{w}^g_{im}\}$ and $\text{ACC} = \frac{1}{n}\sum_{m=1}^{M^0_1}\sum_{m'=1}^{M^0_2}\sum_{i\in G^0_{1m}}\sum_{i'\in G^0_{2m'}} I\{w^1_{im} = \widehat{w}^1_{im}, w^2_{i'm'} = \widehat{w}^2_{i'm'}\}$ , in which $w^g_{im} = 1$ for $i \in G^0_{gm}$, $w^g_{im} = 0$ otherwise, and $\widehat{w}^g_{im} = 1$ for $U_i \in \widehat{\mathcal{U}}_{gm}$ and $\widehat{w}^g_{im} = 0$ otherwise. As $n$ increases, the accuracy gets close to 1. To evaluate the accuracy of the group membership predictions, we tested the proposed method by predicting the group membership of 100 test subjects using the recovered partitions. $\text{ACC}^p_1$, $\text{ACC}^p_2$ and $\text{ACC}^p$ shows the average prediction accuracy for the 100 testing sample. From the Table A.1, it is clear that all ACC-related indices are very close to 1, indicating that both the proposed partition recovery method and the

membership prediction based on it perform exceptionally well. Additionally, Figure A.2 and A.3 visualizes the recovered partitions for cases 6-10 when $n = 400$ and $n = 800$, respectively. In this figure, the dotted curves represent the true boundary curves, while the differently colored areas depict the estimated partition blocks. The boundaries formed by adjacent estimated blocks represent the estimated boundaries. From Figure A.2 and A.3, we can observe that the estimated partition blocks closely align with the true partitions in most cases.

The existing methods cannot handle general cases like case 3-9. We compare our method with Li et al. (2021) in the simpler situations of case 1 and case 2. We use the three metrics, per, ACC and $\text{ACC}^p$, to compare the performance of subgroup identification and prediction between the two methods, as shown in Table A.3. The performance of both methods is good, as they are able to correctly identify group structures and effectively predict group members with a probability of 1. Figure 3 are representative to all parameter estimation error (estimator value - the true value) under 200 replications. Our method is as effective as the Li et al. (2021). To demonstrate the superiority of our method, we use Li et al. (2021) method to handle the situation under case 4 (Any selection from Case 3 - 9 is acceptable). Under 200 replications, the number of identified subgroups was as follows: 94 times for 1 subgroup, 91 times for 2 subgroups, 14 times for 3 subgroups, and 1 time for 4 subgroups. It is clear that their method does not work in remaining cases.

[Figure 2 about here.]

[Figure 3 about here.]

[Table 2 about here.]

[Figure 4 about here.]

[Table 3 about here.]

Our simulation study was conducted on a high-performance server equipped with 4 Intel Xeon Gold 6238 CPUs (base frequency 2.10 GHz, max turbo frequency 3.70 GHz, 176 threads) and 1 TB

of main memory (RAM). Code was executed in a Python 3.10 environment. The computation time primarily depends on the sample size $n$ and the number of variables. Figure A.5 shows the average computation time of our algorithm across 50 Monte Carlo repetitions under different sample sizes for Cases 1 to 5. The computational cost primarily stems from two aspects: the ADMM algorithm for solving the penalized optimization problem to detect the underlying group structure, and the support vector machine procedure for recovering the group partitions. The computation times for Cases 1–3 are similar across different sample sizes. Due to the use of nonlinear support vector machines for recovering group partitions, the computation time for Case 4 is slightly longer than that of Cases 1–3. Case 5 represents a high-dimensional setting where the number of variables increases with the sample size, resulting in significantly longer computation times compared to Cases 1–4. Although our method is more computationally intensive than standard linear regression, it is important to note that it requires less computation compared to many other algorithms handling heterogeneity, such as pairwise fusion penalties (Ma and Huang, 2017).

[Figure 5 about here.]

### A3.5 *Analysis of Panitumumab trial data*

The Panitumumab trial data is a right censored data set. To make our method applicable for analyzing these data, we removed the subjects with missing covariates and the right-censored subjects, then took the logarithm of progression-free survival of the uncensored subjects as $Y$, using ordinary least squares for fitting, namely, $\log T_i = \mathbf{X}_i^T\beta + Z_i\alpha_i(\mathbf{U}_i) + \epsilon_i, i = 1, \cdots, 804$, where $T_i$ is the progression-free survival day for subject $i$. Hence, our analysis method is essentially an Accelerated Failure Time (AFT) model applied to the specific case of complete (non-censored) failure time data.

In addition, for comparison, we also conducted an alternative analysis by directly treating the censoring time of patients without an observed disease progression as the final PFS value. The details presented as follows:

After removing 95 records with missing data, a total of 851 subjects were included in the

analysis. Similarly, we first fit a homogeneous linear regression model using $Y_i$ as response, the 8 baseline covariates, and the intercept, along with the treatment variable $Z_i$ as predictors. The estimated coefficient under the homogeneous model is 0.071 ($p < 0.05$), corresponding to a hazard ratio of approximately 1.07. While the overall effect is modest in magnitude, suggesting limited average clinical benefit at the population level, the statistically significant signal may still reflect underlying biological heterogeneity and supports further investigation of subgroup-specific treatment responses. This is consistent with the limited overall efficacy that initially led the European Medicines Agency to decline panitumumab's application for mCRC. We also fit the proposed heterogeneous linear model $Y_i = \mathbf{X}_i^T\beta + Z_i\alpha_i(\mathbf{U}_i) + \epsilon_i, i = 1, \cdots, n$, where $\mathbf{X}_i$, $Z_i$ and $\mathbf{U}_i$ are defined consistently with the notations in the Section 6. All the predictors are centered and standardized before applying the regularization methods.

[Figure 6 about here.]

Patients were assigned to three subgroups ($\widehat{G}_1$, $\widehat{G}_2$, $\widehat{G}_3$) with sizes of 159, 334, and 358, respectively. Performing post-grouping estimation on the identified subgroup structures, we obtain the following treatment effect estimates: $\widehat{\theta}_1 = -0.494$ (statistically significant, $p = 0.000$), $\widehat{\theta}_2 = 0.043$ (not significant, $p = 0.228$), and $\widehat{\theta}_3 = 0.366$ (statistically significant, $p = 0.000$). This suggests that for patients in subgroup 3, panitumumab plus FOLFIRI demonstrates superior efficacy compared to FOLFIRI alone. However, no significant difference was observed between the two treatment regimens for patients in subgroup 2, whereas for those in subgroup 1, panitumumab was associated with adverse effects. These results diverge from the established view that panitumumab provides clinical benefit in wild-type KRAS patients. This conclusion is largely consistent with the analysis that directly treats censored data as missing data. Additionally, we attempt to recover the partition using a Gaussian kernel-based support vector machine, the prediction accuracy of the support vector machine in recovering the partition membership is 92.48%, which demonstrates that the proposed method works well. We further analyze the distribution of threshold variables across the different

subgroups. Figures A.6(a)–(d) display the counts for each category of the variables gender, KRAS gene mutation status, ECOG performance status, and baseline metastatic site number within the group structures $\widehat{G}_1$, $\widehat{G}_2$ and $\widehat{G}_3$. The boxplots of the threshold variables age at baseline in different subgroups are presented in Figure A.6(e). Based on Figures A.6(a)–(e), the distributions of the three discrete threshold variables (gender, KRAS mutation status,ECOG performance status) in subgroup $\widehat{G}_1$ exhibit a distinct pattern compared to subgroups $\widehat{G}_2$ and $\widehat{G}_3$, while the latter two subgroups show generally similar distributions for these variables. In contrast, the number of baseline metastatic sites displays markedly different distribution patterns between $\widehat{G}_2$ and $\widehat{G}_3$. Figure A.6(e) reveals that both the mean and median age in subgroup 2 are higher than those in the other two subgroups.

We fully acknowledge that the first approach—treating censored observations as missing and fitting the model using only complete data, may lead to information loss and estimation bias if the missingness is nonignorable. Similarly, directly imputing censored times as the observed progression-free survival (PFS) values may underestimate the regression coefficients, since the true survival time for censored subjects should exceed the recorded censoring time. Nevertheless, a comparison of the results obtained from these two methods, directly excluding all censored subjects and treating the censored time directly as the final PFS value, shows that the estimates and conclusions are very similar. This suggests that, in this particular dataset, the handling of censored observations (which constitute only 5.5% of the sample) may not substantially influence the overall results.

Both approaches, however, have methodological limitations. Therefore, an important direction for future research is to extend the proposed method to established time-to-event models, such as the Cox proportional hazards model or accelerated failure time (AFT) models, while appropriately accounting for right-censored outcomes. In particular, for AFT-type extensions, one could incorporate classical imputation-based strategies for censoring, such as the Buckley–James imputation method, within our subgroup-identification and post-recovery estimation framework. We have

acknowledged this limitation in the revised manuscript and have highlighted formal time-to-event analysis, including principled handling of censoring, as a key direction for future research.

### A3.6 *Proof of Theorem 1*

Take the $g$-th component as an example, for every $[\boldsymbol{\alpha}]_g \in \mathcal{M}^g_{\mathcal{G}}$, it can be written as $[\boldsymbol{\alpha}]_g = \mathbf{W}_g\boldsymbol{\theta}_g$. Recall $\boldsymbol{E} = (\mathbf{X}, \mathbf{Z}_{(1)}\mathbf{W}_1, \cdots, \mathbf{Z}_{(p)}\mathbf{W}_p)$. We have

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}^{or} \\ \widehat{\boldsymbol{\theta}}_1^{or} \\ \cdots \\ \widehat{\boldsymbol{\theta}}_p^{or} \end{pmatrix} = \arg\min_{\boldsymbol{\beta}\in R^q, \boldsymbol{\theta}_g \in M_g^0} \frac{1}{2}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta} - \sum_{g=1}^{p} \mathbf{Z}^g [\boldsymbol{\alpha}]_g\|^2 = \arg\min_{\boldsymbol{\beta}\in R^q, \boldsymbol{\theta}_g \in M_g^0} \frac{1}{2}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta} - \sum_{g=1}^{p} \mathbf{Z}^g \mathbf{W}_g\boldsymbol{\theta}_g\|^2.$$

Then, we have

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}^{or} \\ \widehat{\boldsymbol{\theta}}_1^{or} \\ \cdots \\ \widehat{\boldsymbol{\theta}}_p^{or} \end{pmatrix} = (\boldsymbol{E}^T\boldsymbol{E})^{-1}\boldsymbol{E}^T\boldsymbol{Y},$$

and

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0 \\ \widehat{\boldsymbol{\theta}}_1^{or} - \boldsymbol{\theta}_1^0 \\ \cdots \\ \widehat{\boldsymbol{\theta}}_p^{or} - \boldsymbol{\theta}_p^0 \end{pmatrix} = (\boldsymbol{E}^T\boldsymbol{E})^{-1}\boldsymbol{E}^T\boldsymbol{\epsilon}.$$

Hence

$$\left\| \begin{pmatrix} \widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0 \\ \widehat{\boldsymbol{\theta}}_1^{or} - \boldsymbol{\theta}_1^0 \\ \cdots \\ \widehat{\boldsymbol{\theta}}_p^{or} - \boldsymbol{\theta}_p^0 \end{pmatrix} \right\| \leqslant \|(\boldsymbol{E}^T\boldsymbol{E})^{-1}\|\|\boldsymbol{E}^T\boldsymbol{\epsilon}\|. \tag{A.14}$$

By condition $(C1)$, we get $\|(\boldsymbol{E}^T\boldsymbol{E})^{-1}\| \leqslant C_1^{-1}|\boldsymbol{G}_{\min}|^{-1}$ and thus

$$\|(\boldsymbol{E}^T\boldsymbol{E})^{-1}\|_\infty \leqslant \sqrt{q + \sum_{g=1}^{p} M_g^0} C_1^{-1}|\boldsymbol{G}_{\min}|^{-1}. \tag{A.15}$$

Moreover

$$P(\|\boldsymbol{E}^T\boldsymbol{\epsilon}\|_\infty > C\sqrt{n\log n}) \leqslant P(\|\mathbf{X}^T\boldsymbol{\epsilon}\|_\infty > C\sqrt{n\log n}) + \sum_{g=1}^{p} P(\|(\mathbf{Z}^g\mathbf{W}_g)^T\boldsymbol{\epsilon}\|_\infty > C\sqrt{n\log n}),$$

for some constant $0 < C < \infty$. And $\mathbf{Z}^g\mathbf{W}_g = [Z_{ig}^T 1\{i \in \boldsymbol{G}_{gm}\}]_{i=1,m=1}^{n,M_g^0}$, we have $\|(\mathbf{Z}^g\mathbf{W}_g)^T\boldsymbol{\epsilon}\|_\infty = \sup_{g,M_g^0} |\sum_{i=1}^n Z_{ig}\epsilon_i 1\{i \in \boldsymbol{G}_{gm}\}|$ and by union bound, Condition $(C1)$ that $\sum_{i=1}^n Z_{ig}^2 1\{i \in \boldsymbol{G}_{gm}\} = |\boldsymbol{G}_{gm}|$ and Condition $(C3)$,

$$\begin{aligned}
&P(\|(\mathbf{Z}^g\mathbf{W}_g)^T\boldsymbol{\epsilon}\|_\infty > C\sqrt{n\log n})\\
&\leqslant \sum_{m=1}^{M_g^0} P(|\sum_{i\in\boldsymbol{G}_{gm}} Z_{ig}\epsilon_i| > C\sqrt{n\log n}),\\
&\leqslant \sum_{k=1}^{M_g^0} P(|\sum_{i\in\boldsymbol{G}_{gm}} Z_{ig}\epsilon_i| > \sqrt{|\boldsymbol{G}_{gm}|}C\sqrt{\log n}),\\
&\leqslant 2M_g^0\exp(-c_1C^2\log n) = 2M_g^0 n^{-c_1C^2}.
\end{aligned}$$

By union bound, Condition $(C1)$ that $\|\mathbf{X}_m\| = \sqrt{n}$ and $(C3)$,

$$\begin{aligned}
&P(\|\mathbf{X}^T\boldsymbol{\epsilon}\|_\infty > C\sqrt{n\log n})\\
&\leqslant \sum_{g=1}^{q} P(|\mathbf{X}_g^T\boldsymbol{\epsilon}| > C\sqrt{n\log n}),\\
&\leqslant 2q\exp(-c_1C^2\log n) = 2qn^{-c_1C^2}.
\end{aligned}$$

By the above results, we have

$$P(\|\boldsymbol{E}^T\boldsymbol{\epsilon}\|_\infty > C\sqrt{n\log n}) \leqslant 2(q + \sum_{g=1}^{p} M_g^0)n^{-c_1C^2}.$$

Since $\|\boldsymbol{E}^T\boldsymbol{\epsilon}\| \leqslant \sqrt{q + \sum_{g=1}^p M_g^0}\|\boldsymbol{E}^T\epsilon\|_\infty$, then

$$P(\|\boldsymbol{E}^T\boldsymbol{\epsilon}\| > C\sqrt{q + \sum_{g=1}^{p} M_g^0}\sqrt{n\log n}) \leqslant 2(q + \sum_{g=1}^{p} M_g^0)n^{-c_1C^2}. \tag{A.16}$$

Therefore, by (A.14), (A.15), (A.16), we have with probability at least $1 - 2(q + \sum_{g=1}^{p} M_g^0)n^{-c_1 C^2}$,

$$\left\| \begin{pmatrix} \widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0 \\ \widehat{\boldsymbol{\theta}}_1^{or} - \boldsymbol{\theta}_1^0 \\ \cdots \\ \widehat{\boldsymbol{\theta}}_p^{or} - \boldsymbol{\theta}_p^0 \end{pmatrix} \right\| \leqslant C_1^{-1} |\boldsymbol{G}_{\min}|^{-1} C \sqrt{n \log n} \sqrt{q + \sum_{g=1}^{p} M_g^0} \tag{A.17}$$

The result (1) in Theorem 1 is proved by letting $C = c_1^{-1/2}$. Moreover

$$\left\| \begin{pmatrix} [\widehat{\boldsymbol{\alpha}}^{or}]_1 - [\boldsymbol{\alpha}^0]_1 \\ \cdots \\ [\widehat{\boldsymbol{\alpha}}^{or}]_p - [\boldsymbol{\alpha}^0]_p \end{pmatrix} \right\|^2 = \sum_{g=1}^{p} \sum_{m=1}^{M_g^0} \sum_{j \in G_{gm}^0} |\widehat{\theta}_{gm}^{or} - \widehat{\theta}_{gm}^0|^2 \leqslant |\max_{g',m'}(G_{g'm'}^0)| \sum_{g=1}^{p} \sum_{m=1}^{M_g^0} |\widehat{\theta}_{gm}^{or} - \widehat{\theta}_{gm}^0|^2$$

$$= |\max_{g',m'}(G_{g'm'}^0)| \left\| \begin{pmatrix} \widehat{\theta}_1^{or} - \theta_1^0 \\ \cdots \\ \widehat{\theta}_p^{or} - \theta_p^0 \end{pmatrix} \right\|^2 \leqslant |G_{\max}^0| \phi_n^2,$$

where $|\boldsymbol{G}_{\max}^0| = |\max_{g',m'}(\boldsymbol{G}_{g'm'}^0)|$ and

$$\sup_i \|\widehat{\boldsymbol{\alpha}}_i^{or} - \boldsymbol{\alpha}_i^0\| \leqslant \left( \sum_{g=1}^{p} \sup_m |\widehat{\theta}_{gm}^{or} - \theta_{gm}^0|^2 \right)^{1/2} \leqslant \|\widehat{\boldsymbol{\theta}}^{or} - \boldsymbol{\theta}^0\| \leqslant \phi_n.$$

Note $\boldsymbol{E}$ is a $n \times (q + \sum_{g=1}^{p} M_g^0)$ matrix, let $\boldsymbol{E} = (\boldsymbol{E}_1, \cdots, \boldsymbol{E}_n)^T$, and $\Xi_n = \boldsymbol{E}^T \boldsymbol{E}$, for any vector $\mathbf{a}_n \in R^{q + \sum_{g=1}^{p} M_g^0}$ with $\|\mathbf{a}_n\| = 1$. Then

$$\mathbf{a}_n^T((\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0)^T, (\widehat{\boldsymbol{\theta}}_1^{or} - \boldsymbol{\theta}_1^0)^T, \cdots, (\widehat{\boldsymbol{\theta}}_p^{or} - \boldsymbol{\theta}_p^0)^T)^T = \sum_{i=1}^{n} \mathbf{a}_n^T \Xi_n^{-1} \boldsymbol{E}_i \epsilon_i.$$

and

$$E\{\mathbf{a}_n^T((\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0)^T, (\widehat{\boldsymbol{\theta}}_1^{or} - \boldsymbol{\theta}_1^0)^T, \cdots, (\widehat{\boldsymbol{\theta}}_p^{or} - \boldsymbol{\theta}_p^0)^T)^T\} = 0,$$

$$var\{\mathbf{a}_n^T((\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0)^T, (\widehat{\boldsymbol{\theta}}_1^{or} - \boldsymbol{\theta}_1^0)^T, \cdots, (\widehat{\boldsymbol{\theta}}_p^{or} - \boldsymbol{\theta}_p^0)^T)^T\} = \sigma^2[\mathbf{a}_n^T(\boldsymbol{E}^T\boldsymbol{E})^{-1}\mathbf{a}_n] = \sigma^2 \mathbf{a}_n^T \Xi_n^{-1} \mathbf{a}_n.$$

Moreover, for any $\epsilon > 0$,

$$\begin{aligned} & \sum_{i=1}^{n} E[(\mathbf{a}_n^T \Xi^{-1} \boldsymbol{E}_i \epsilon_i)^2 \cdot 1_{\{|\mathbf{a}_n^T \Xi^{-1} \boldsymbol{E}_i \epsilon_i| > \epsilon \sigma_n(\mathbf{a}_n)\}}] \\ \leqslant \quad & \sum_{i=1}^{n} \{E(\mathbf{a}_n^T \Xi^{-1} \boldsymbol{E}_i \epsilon_i)^4\}^{1/2} [P\{|\mathbf{a}_n^T \Xi^{-1} \boldsymbol{E}_i \epsilon_i| > \epsilon \sigma_n(\mathbf{a}_n)\}]^{1/2}. \end{aligned}$$

By condition $(C3)$ $E(\epsilon_i^4) \leqslant c$ for some constant $c \in (0, \infty)$,then

$$\sum_{i=1}^{n} \{E(\mathbf{a}_n^T \Xi^{-1} \boldsymbol{E}_i \epsilon_i)^4\}^{1/2} \leqslant c^{1/2} \mathbf{a}_n^T \Xi_n^{-1} \mathbf{a}_n.$$

By Markov's inequality, we can get

$$\begin{aligned} & \max_i pr\{|\mathbf{a}_n^T \Xi^{-1} \boldsymbol{E}_i \epsilon_i| > \epsilon \sigma_n(\mathbf{a}_n)\} \\ \leqslant \quad & \max_i E(\mathbf{a}_n^T \Xi^{-1} \boldsymbol{E}_i \epsilon_i)^2 / \{\epsilon^2 \sigma_n^2(\mathbf{a}_n)\} \\ \leqslant \quad & c' \epsilon^{-2} (q + \sum_{g=1}^{p} M_g^0) \mathbf{a}_n^T \Xi_n^{-1} \Xi_n^{-1} \mathbf{a}_n / (\mathbf{a}_n^T \Xi_n^{-1} \mathbf{a}_n), \end{aligned}$$

for some constant $c' \in (0, \infty)$. Therefore,

$$\begin{aligned} & \sigma_n^{-2}(\mathbf{a}_n) \sum_{i=1}^{n} E[(\mathbf{a}_n^T \Xi^{-1} \boldsymbol{E}_i \epsilon_i)^2 \cdot 1_{\{|\mathbf{a}_n^T \Xi^{-1} \boldsymbol{E}_i \epsilon_i| > \epsilon \sigma_n(\mathbf{a}_n)\}}] \\ \leqslant \quad & \{\sigma^2 \mathbf{a}_n^T \Xi_n^{-1} \mathbf{a}_n\}^{-1} c^{-1/2} \mathbf{a}_n^T \Xi_n^{-1} \mathbf{a}_n \{c' \epsilon^{-2} (q + \sum_{g=1}^{p} M_g^0) \mathbf{a}_n^T \Xi_n^{-1} \Xi_n^{-1} \mathbf{a}_n / (\mathbf{a}_n^T \Xi_n^{-1} \mathbf{a}_n)\}^{1/2} \\ \leqslant \quad & c^{1/2} c'^{1/2} C_1^{-1/2} \epsilon^{-1} \sigma^{-1} (q + \sum_{g=1}^{p} M_g^0)^{1/2} |\boldsymbol{G}_{\min}|^{-1/2} = o(1). \end{aligned}$$

From Linderberg Central Limit Theorem, the result Theorem 1 (2) proved. □

## A3.7 *Proof of Theorem 2*

In this section, we show the results in Theorem 2. Note that $\mathbf{U}_i \overset{iid}{\sim} p(\mathbf{U}), i = 1, \cdots, n$. First, we introduce the following lemma.

LEMMA 1: *Denote by* $R_K(\mathbf{U})$ *the distance from* $\mathbf{U} \in \boldsymbol{\mathcal{U}}$ *to its* $K$*th nearest neighbor in the set* $\{\mathbf{U}_1, \cdots, \mathbf{U}_n\}$. *Setting*

$$R_{K,\max} = \max_{1 \leqslant i \leqslant n} R_K(\mathbf{U}_i),$$

$$R_{K,\min} = \min_{1 \leqslant i \leqslant n} R_K(\mathbf{U}_i),$$

*we have that*

$$pr\{a(\frac{K}{n})^{1/r} \leqslant R_{K,\min} \leqslant R_{K,\max} \leqslant \widetilde{a}(\frac{K}{n})^{1/r}\} \geqslant 1 - n\exp(-K/3) - n\exp(-K/12).$$

*under condition* (4)-(5)*, where* $a = 1/(2c_{2,r}p_{\max})^{1/r}$ *and* $\widetilde{a} = 2^{1/r}/(c_{1,r}p_{\min})^{1/r}$.

**proof**. see the proof of Lemma S3. in (Madrid Padilla et al., 2020).

Next, we proceed to find a lower bound on the degree of the $KNN$ graph. Defining the sets

$$B_{gi}(\mathbf{U}) = \{j \in [n]\backslash\{i\} : \mathbf{U}_{gj} \in B_{a(K/n)^{1/r}}(\mathbf{U}) \cap \mathcal{U}^0_{g\cdot}(\mathbf{U})\},$$

$$B_{gi} = \{j \in [n]\backslash\{i\} : \mathbf{U}_{gj} \in B_{a(K/n)^{1/r}}(\mathbf{U}_i) \cap \mathcal{U}^0_{g\cdot}(\mathbf{U}_i)\},$$

for $i \in [n]$, and $\mathbf{U} \in \mathcal{U}, \mathcal{U}^0_{g\cdot}(\mathbf{U})$ reprents the partition of $\mathbf{U}$ belongs in $\{\mathcal{U}^0_{gm}, m = 1, \cdots, M^0_g\}$ , and where $a$ is given as in Lemma 1. Then

$$|B_{gi}(\mathbf{U})| \sim Binomial[n-1, \mathrm{pr}\{\mathbf{U}_1 \in B_{a(K/n)^{1/r}}(\mathbf{U}) \cap \mathcal{U}^0_{g\cdot}(\mathbf{U})\}],$$

where $p_{\min}c_{1,r}a^r\frac{K}{n} \leqslant \mathrm{pr}\{\mathbf{U}_1 \in B_{a(K/n)^{1/r}}(\mathbf{U}) \cap \mathcal{U}^0_{g\cdot}(\mathbf{U})\} \leqslant p_{\max}c_{2,r}a^r\frac{K}{n}$ by condition (C4). According to the concentration inequality for binomial random variables, we can get

$$\begin{aligned}
&\mathrm{pr}\left\{|B_{gi}(\mathbf{U})| \leqslant \frac{n-1}{2}\mathrm{pr}\{\mathbf{U}_1 \in B_{a(K/n)^{1/r}}(\mathbf{U}) \cap \mathcal{U}^0_{g\cdot}(\mathbf{U})\}\right\} \\
&\leqslant \exp\left(-\frac{(n-1)\mathrm{pr}\{\mathbf{U}_1 \in B_{a(K/n)^{1/r}}(\mathbf{U}) \cap \mathcal{U}^0_{g\cdot}(\mathbf{U})\}}{12}\right) \leqslant \exp\left(-\frac{(n-1)p_{\min}c_{1,r}a^rK}{12n}\right), \\
&\leqslant \exp\left(-\frac{p_{\min}c_{1,r}K}{48c_{2,r}p_{\max}}\right), \quad \text{for } n \geqslant 2.
\end{aligned}$$

Hence, for $n \geqslant 2$

$$\begin{aligned}
&\mathrm{pr}\left\{|B_{gi}| \leqslant \frac{n-1}{2}\mathrm{pr}\{\mathbf{U}_1 \in B_{a(K/n)^{1/r}}(\mathbf{U}) \cap \mathcal{U}^0_{g\cdot}(\mathbf{U})\}\right\} \\
&= \int_{\mathcal{U}} \mathrm{pr}\left\{|B_{gi}(\mathbf{U})| \leqslant \frac{n-1}{2}\mathrm{pr}\{\mathbf{U}_1 \in B_{a(K/n)^{1/r}}(\mathbf{U}) \cap \mathcal{U}^0_{g\cdot}(\mathbf{U})\}\right\} p(\mathbf{U})\nu(d\mathbf{U}) \leqslant \exp\left(-\frac{p_{\min}c_{1,r}K}{48c_{2,r}p_{\max}}\right).
\end{aligned}$$

Therefore, if $d_i$ is the degree associated with $\mathbf{U}_i$, then $d_i = |\{j \in \{1, \cdots, n\}\backslash\{i\} : d_{\mathcal{U}}(\mathbf{U}_i, \mathbf{U}_j) \leqslant R_K(\mathbf{U}_i)\}|$ and $d^*_{gi} = |\{j \in \{1, \cdots, n\}\backslash\{i\} : d_{\mathcal{U}}(\mathbf{U}_i, \mathbf{U}_j) \leqslant R_K(\mathbf{U}_i), i \text{ and } j \in \mathbf{G}_{gm}, \exists m \in \{1, \cdots, M^0_g\}\}|$. Define $\widetilde{d}_{gi} = |\{j \in \{1, \cdots, n\}\backslash\{i\} : d_{\mathcal{U}}(\mathbf{U}_i, \mathbf{U}_j) \leqslant R_{K,\min}, i \text{ and } j \in \mathbf{G}_{gm}, \exists m \in \{1, \cdots, M^0_g\}\}|$. It is easy to know $d^*_{gi} \geqslant \widetilde{d}_{gi} \geqslant |B_{gi}| = |\{j \in \{1, \cdots, n\}\backslash\{i\} : d_{\mathcal{U}}(\mathbf{U}_i, \mathbf{U}_j) \leqslant$

$a(\frac{K}{n})^{1/r}, i$ and $j \in \mathbf{G}_{gm}, \exists m \in \{1, \cdots, M_g^0\}\}|$ when $\{R_{K,\min} \geqslant a(\frac{K}{n})^{1/r}\}$. Then,

$$\begin{aligned}
&\mathrm{pr}\left(d_{gi}^* \leqslant \frac{n-1}{2}\mathrm{pr}\{\mathbf{U}_1 \in B_{a(K/n)^{1/r}}(\mathbf{U}) \cap \mathcal{U}_{g\cdot}^0(\mathbf{U})\}\right) \\
&\leqslant \mathrm{pr}\left(\widetilde{d}_{gi} \leqslant \frac{n-1}{2}\mathrm{pr}\{\mathbf{U}_1 \in B_{a(K/n)^{1/r}}(\mathbf{U}) \cap \mathcal{U}_{g\cdot}^0(\mathbf{U})\}\right) \\
&\leqslant \mathrm{pr}\left(|B_{gi}| \leqslant \frac{n-1}{2}\mathrm{pr}\{\mathbf{U}_1 \in B_{a(K/n)^{1/r}}(\mathbf{U}) \cap \mathcal{U}_{g\cdot}^0(\mathbf{U})\}\right) + \mathrm{pr}\left(R_{K,\min} \leqslant a(\frac{K}{n})^{1/r}\right) \\
&\leqslant \exp\left(-\frac{p_{\min}c_{1,r}K}{48c_{2,r}p_{\max}}\right) + n\exp\left(-\frac{K}{3}\right). \quad \text{for } n \geqslant 2
\end{aligned}$$

The last inequality by the Lemma S3 of Madrid Padilla et al. (2020), $\mathrm{pr}\left(R_{K,\min} \leqslant a(\frac{K}{n})^{1/r}\right) \leqslant n\exp(-K/3)$. Then,

$$\mathrm{pr}\left(d_{gi}^* \geqslant \frac{n-1}{2}\mathrm{pr}\{\mathbf{U}_1 \in B_{a(K/n)^{1/r}}(\mathbf{U}) \cap \mathcal{U}_{g\cdot}^0(\mathbf{U})\}\right) \geqslant 1 - \exp\left(-\frac{p_{\min}c_{1,r}K}{48c_{2,r}p_{\max}}\right) - n\exp\left(-\frac{K}{3}\right).$$

By $\frac{n-1}{2}\mathrm{pr}\{\mathbf{U}_1 \in B_{a(K/n)^{1/r}}(\mathbf{U}) \cap \mathcal{U}_{g\cdot}^0(\mathbf{U})\} \geqslant \frac{K}{4}p_{\min}c_{1,r}a^r = \frac{c_{1,r}p_{\min}K}{8c_{2,r}p_{\max}}$. Finally, we have

$$\begin{aligned}
\mathrm{pr}\left(K' \geqslant \frac{c_{1,r}p_{\min}K}{8c_{2,r}p_{\max}}\right) &= \mathrm{pr}\left(d_{gi}^* \geqslant \frac{c_{1,r}p_{\min}K}{8c_{2,r}p_{\max}}, \forall i, g\right) \\
&\geqslant 1 - np\exp\left(-\frac{p_{\min}c_{1,r}K}{48c_{2,r}p_{\max}}\right) - n^2p\exp\left(-\frac{K}{3}\right). \qquad \text{(A.18)}
\end{aligned}$$

□

**The proof of Theorem 2:** Define

$$L_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta} - \sum_{g=1}^{p}\mathbf{Z}^g[\boldsymbol{\alpha}]_g\|^2, P_n(\boldsymbol{\alpha}) = \lambda\sum_{g=1}^{p}\sum_{i=1}^{nK}\rho(|[\nabla_G[\boldsymbol{\alpha}]_g]_i|),$$

$$L_n^{\mathcal{G}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta} - \sum_{g=1}^{p}\mathbf{Z}^g\mathbf{W}_g\boldsymbol{\theta}_g\|^2, P_n^{\mathcal{G}}(\boldsymbol{\theta}) = \lambda\sum_{g=1}^{p}\sum_{i=1}^{nK}\rho(|[\nabla_G\mathbf{W}_g\boldsymbol{\theta}_g]_i|),$$

and let

$$Q_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) = L_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) + \sum_{g=1}^{p}P_n^g(\boldsymbol{\alpha}), Q_n^{\mathcal{G}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = L_n^{\mathcal{G}}(\boldsymbol{\beta}, \boldsymbol{\theta}) + \sum_{g=1}^{p}P_n^{\mathcal{G}_g}(\boldsymbol{\theta}).$$

Let $T_g : \mathcal{M}_{\mathcal{G}}^g \to \mathbb{R}^{M_g^0}$ be the mapping that $T_g(\boldsymbol{\alpha})$ is the $M_g^0 \times 1$ vector $(g = 1, \cdots, p)$ and its $m$-th vector component equals to the common value of $\alpha_{gi}$ for $i \in G_{gm}$. Let $T_g^* : \mathbb{R}^n \to \mathbb{R}^{M_g^0}$ be the mapping that $T_g^*([\boldsymbol{\alpha}]_g) = \{|G_{gm}|^{-1}\sum_{i\in G_{gm}}\alpha_{ig}, m = 1, \cdots, M_g^0\}$. Clearly, when $[\boldsymbol{\alpha}]_g \in \mathcal{M}_{\mathcal{G}}^g$, $T_g([\boldsymbol{\alpha}]_g) = T_g([\boldsymbol{\alpha}^*]_g)$.

By calculation, for every $[\boldsymbol{\alpha}]_g \in \mathcal{M}_{\mathcal{G}}^g$, we have $P_n^g([\boldsymbol{\alpha}]_g) = P_n^{\mathcal{G}_g}(T_g([\boldsymbol{\alpha}]_g))$ and for every $\boldsymbol{\theta}_g \in$

$\mathbb{R}^{M_g^0}$, we have $P_n^g(T_g^{-1}(\boldsymbol{\theta}_g)) = P_n^{\mathcal{G}_g}(\boldsymbol{\theta}_g)$. Hence

$$Q_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) = Q_n^{\mathcal{G}}(\boldsymbol{\beta}, T(\boldsymbol{\alpha})), Q_n^{\mathcal{G}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = Q_n(\boldsymbol{\beta}, T^{-1}(\boldsymbol{\theta})). \tag{A.19}$$

where $\boldsymbol{\alpha} = ([\boldsymbol{\alpha}]_1, \cdots, [\boldsymbol{\alpha}]_p)^T, T(\boldsymbol{\alpha}) = (T_1([\boldsymbol{\alpha}]_1), \cdots, T_p([\boldsymbol{\alpha}]_p))^T$. Consider the neighborhood of $(\boldsymbol{\beta}^0, [\boldsymbol{\alpha}^0]_1, \cdots, [\boldsymbol{\alpha}^0]_p)$:

$$\Theta = \{\boldsymbol{\beta} \in \mathbb{R}^q, \boldsymbol{\alpha}_i \in \mathbb{R}^p, \forall g : \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\| \leqslant \phi_n, \sup_{i,g} |\alpha_{ig} - \alpha_{ig}^0| \leqslant \phi_n\}.$$

By the result in Theorem 1, there exists an event $E_1$ in which

$$\|\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0\| \leqslant \phi_n, \sup_i |\widehat{\alpha}_{ig}^{or} - \alpha_{ig}^0| \leqslant \phi_n$$

and $P(E_1^C) \leqslant 2(q + \sum_{g=1}^p M_g^0)n^{-1}$. Hence, $(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or}) \in \Theta$ in $E_1$. For any $[\boldsymbol{\alpha}]_g \in \mathbb{R}^n$, let $[\boldsymbol{\alpha}^*]_g = T_g^{-1}(T_g^*([\boldsymbol{\alpha}]_g))$. We show that $(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or})$ is a strictly local minimizer of the objective function (10) with probability approaching 1 through the following two steps.

(i). On the event $E_1$, $Q_n(\boldsymbol{\beta}, \boldsymbol{\alpha}^*) > Q_n(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or})$ for any $(\boldsymbol{\beta}, \boldsymbol{\alpha}) \in \Theta$ and $(\boldsymbol{\beta}, \boldsymbol{\alpha}^*) \neq (\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or})$, where $\boldsymbol{\beta}^* = (\beta_1^*, \cdots, \beta_q^*)^T, \boldsymbol{\beta} = (\beta_1, \cdots, \beta_q)^T, \widehat{\boldsymbol{\beta}}^{or} = (\widehat{\beta}_1^{or}, \cdots, \widehat{\beta}_q^{or})^T$. Similarity, $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_n)^T = ([\boldsymbol{\alpha}]_1, \cdots, [\boldsymbol{\alpha}]_p), \boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^*, \cdots, \boldsymbol{\alpha}_n^*)^T = ([\boldsymbol{\alpha}^*]_1, \cdots, [\boldsymbol{\alpha}^*]_p), \widehat{\boldsymbol{\alpha}}^{or} = (\widehat{\boldsymbol{\alpha}}_1^{or}, \cdots, \widehat{\boldsymbol{\alpha}}_n^{or})^T = ([\widehat{\boldsymbol{\alpha}}^{or}]_{(1)}, \cdots, [\widehat{\boldsymbol{\alpha}}^{or}]_{(p)})$.

(ii). There is an event $E_2$ such that $\text{pr}(E_2^C) \leqslant 2n^{-1} + np \exp\left(-\frac{p_{\min} c_{1,r} K}{48 c_{2,r} p_{\max}}\right) + n^2 p \exp\left(-\frac{K}{3}\right)$. In $E_1 \cap E_2$, there is a neighborhood of $((\widehat{\boldsymbol{\beta}}^{or})^T, (\widehat{\boldsymbol{\alpha}}^{or})^T)^T$, denoted by $\Theta_n$ such that $Q_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) \geqslant Q_n(\boldsymbol{\beta}, \boldsymbol{\alpha}^*)$ for any $(\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T \in \Theta_n \cap \Theta$ for sufficiently large $n$.

Therefore, by the results in (i) and (ii), we have $Q_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) > Q_n(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or})$ for any $(\boldsymbol{\beta}, \boldsymbol{\alpha}) \in \Theta_n \cap \Theta$ and $(\boldsymbol{\beta}, \boldsymbol{\alpha}) \neq (\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or})$ in $E_1 \cap E_2$, so that $(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or})$ is a strict local minimizer of $Q_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ over the event $E_1 \cap E_2$ with $\text{pr}(E_1 \cap E_2) \geqslant 1 - 2(q + \sum_{g=1}^p M_g^0 + 1)n^{-1} - np \exp\left(-\frac{p_{\min} c_{1,r} K}{48 c_{2,r} p_{\max}}\right) - n^2 p \exp\left(-\frac{K}{3}\right)$ for sufficiently large $n$.

In the following we prove the result in (i). We first show $\sum_{g=1}^p P_n^{\mathcal{G}_g}(T_g^*([\boldsymbol{\alpha}]_g)) = C_n$ for any $\boldsymbol{\alpha} \in \Theta$, Where $C_n$ is a constant which does not depend on $\alpha$. Let $T_g^*([\boldsymbol{\alpha}]_g) = \boldsymbol{\theta}_g = (\theta_1, \cdots, \theta_{M_g^0})^T$. It suffices to show that $|\theta_{gm} - \theta_{gm'}| > a\lambda$ for all $g, m$ and $m'$. Then by condition (C2), $\rho(|\theta_{gm} - \theta_{gm'}|)$ is a

constant, and as a result $\sum_{g=1}^{p} P_n^{\mathcal{G}_g}(T_g^*([\boldsymbol{\alpha}]_g))$ is a constant. Since

$$|\theta_{gm} - \theta_{gm'}| \geqslant |\theta^0_{gm} - \theta^0_{gm'}| - 2\sup_m |\theta_{gm} - \theta^0_{gm}|,$$

and

$$\begin{aligned}\sup_m |\theta_{gm} - \theta^0_{gm}| = & \quad \sup_m |\frac{\sum_{i\in G_{gm}} \alpha_{ig}}{|G_{gm}|} - \theta^0_{gm}| = \sup_m |\frac{\sum_{i\in G_{gm}} (\alpha_{ig} - \alpha^0_{ig})}{|G_{gm}|}| \\ \leqslant & \quad \sup_m \frac{\sum_{i\in G_{gm}} |(\alpha_{ig} - \alpha^0_{ig})|}{|G_{gm}|} \leqslant \sup_i |(\alpha_{ig} - \alpha^0_{ig})| \leqslant \phi_n, \end{aligned} \tag{A.20}$$

then for all $m$ and $m'$

$$|\theta_{gm} - \theta_{gm'}| \geqslant |\theta^0_{gm} - \theta^0_{gm'}| - 2\sup_m |\theta_{gm} - \theta^0_{gm}| \geqslant b_n - 2\phi_n > a\lambda,$$

where the last inequality follows from the assumption that $b_n > a\lambda \gg \phi_n$. Therefore, we have $\sum_{g=1}^{p} P_n^{\mathcal{G}_g}(T_g^*([\boldsymbol{\alpha}]_g)) = C_n$, and hence $Q_n^{\mathcal{G}}(\boldsymbol{\beta}, T^*(\boldsymbol{\alpha})) = L_n^{\mathcal{G}}(\boldsymbol{\beta}, T^*(\boldsymbol{\alpha})) + C_n$ for all $(\boldsymbol{\beta}^T, [\boldsymbol{\alpha}]_1^T, \cdots, [\boldsymbol{\alpha}]_p^T)^T \in \Theta$. Since $((\widehat{\boldsymbol{\beta}}^{or})^T, (\widehat{\boldsymbol{\theta}}^{or})^T)^T$ is the unique global minimizer of $L_n^{\mathcal{G}}(\boldsymbol{\beta}, \boldsymbol{\theta})$, then $L_n^{\mathcal{G}}(\boldsymbol{\beta}, T^*(\boldsymbol{\alpha})) > L_n^{\mathcal{G}}(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\theta}}^{or})$ for all $(\boldsymbol{\beta}^T, T^*(\boldsymbol{\alpha})^T) \neq ((\widehat{\boldsymbol{\beta}}^{or})^T, (\widehat{\boldsymbol{\theta}}^{or})^T)^T$ and thus $Q_n^{\mathcal{G}}(\boldsymbol{\beta}, T^*(\boldsymbol{\alpha})) > Q_n^{\mathcal{G}}(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\theta}}^{or})$ for all $T^*(\boldsymbol{\alpha}) \neq \widehat{\boldsymbol{\theta}}^{or}$. By (A.19), we have $Q_n^{\mathcal{G}}(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\theta}}^{or}) = Q_n(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or})$ and $Q_n^{\mathcal{G}}(\boldsymbol{\beta}, T^*(\boldsymbol{\alpha})) = Q_n(\boldsymbol{\beta}, T^{-1}(T^*(\boldsymbol{\alpha}))) = Q_n(\boldsymbol{\beta}, \boldsymbol{\alpha}^*)$. Therefore, $Q_n(\boldsymbol{\beta}, \boldsymbol{\alpha}^*) > Q_n(\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or})$ for all $\boldsymbol{\alpha}^* \neq \widehat{\boldsymbol{\alpha}}^{or}$, and the result in (i) is proved.

Next we prove the result in (ii). For a positive sequence $t_n$, let $\Theta_n = \{[\boldsymbol{\alpha}]_g : \sup_g \|[\boldsymbol{\alpha}]_g - [\widehat{\boldsymbol{\alpha}}]_g^{or}\| \leqslant t_n\}$. For $(\boldsymbol{\beta}^T, [\boldsymbol{\alpha}]_1^T, \cdots, [\boldsymbol{\alpha}]_p^T)^T \in \Theta_n \cap \Theta$, by Taylor's expansion, we have

$$Q_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) - Q_n(\boldsymbol{\beta}, \boldsymbol{\alpha}^*) = \Gamma_1 + \Gamma_2,$$

where

$$\begin{aligned}\Gamma_1 &= -(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \sum_{g=1}^{p} \mathbf{Z}_{(g)}[\boldsymbol{\alpha}]_g^{(t)})^T (\sum_{g=1}^{p} \mathbf{Z}_{(g)}([\boldsymbol{\alpha}]_g - [\boldsymbol{\alpha}^*]_g)), \\ \Gamma_2 &= \lambda \sum_{g=1}^{p} \sum_{i=1}^{n} \frac{\partial P_n^g([\boldsymbol{\alpha}]_g^{(t)})}{\partial \alpha_{ig}} (\alpha_{ig} - \alpha^*_{ig}).\end{aligned}$$

and $[\boldsymbol{\alpha}]_g^{(t)} = \varsigma[\boldsymbol{\alpha}]_g + (1-\varsigma)[\boldsymbol{\alpha}^*]_g$ for some $\varsigma \in (0,1)$. Morever,

$$\begin{aligned}\Gamma_2 &= \lambda \sum_{g=1}^{p}\sum_{i=1}^{n}\sum_{j\in\mathcal{N}_K^i} \overline{\rho}(\alpha_{ig}^{(t)} - \alpha_{jg}^{(t)})(\alpha_{ig} - \alpha_{ig}^*) - \lambda \sum_{g=1}^{p}\sum_{j=1}^{n}\sum_{i\in\mathcal{N}_K^j} \overline{\rho}(\alpha_{jg}^{(t)} - \alpha_{ig}^{(t)})(\alpha_{ig} - \alpha_{ig}^*) \\ &= \lambda \sum_{g=1}^{p}\sum_{i=1}^{n}\sum_{j\in\mathcal{N}_K^i} \overline{\rho}(\alpha_{ig}^{(t)} - \alpha_{jg}^{(t)})(\alpha_{ig} - \alpha_{ig}^*) - \lambda \sum_{g=1}^{p}\sum_{j=1}^{n}\sum_{j\in\mathcal{N}_K^i} \overline{\rho}(\alpha_{ig}^{(t)} - \alpha_{jg}^{(t)})(\alpha_{jg} - \alpha_{jg}^*) \\ &= \lambda \sum_{g=1}^{p}\sum_{i=1}^{n}\sum_{j\in\mathcal{N}_K^i} \overline{\rho}(\alpha_{ig}^{(t)} - \alpha_{jg}^{(t)})\{(\alpha_{ig} - \alpha_{ig}^*) - (\alpha_{jg} - \alpha_{jg}^*)\}.\end{aligned}$$

When $i, j \in G_{gm}, \alpha_{ig}^* = \alpha_{jg}^*$, and $\alpha_{ig}^{(t)} - \alpha_{jg}^{(t)}$ has the same sign as $\alpha_{ig} - \alpha_{jg}$. Hence

$$\begin{aligned}\Gamma_2 = \quad & \lambda \sum_{g=1}^{p}\sum_{m=1}^{M_g^0}\sum_{i,j\in G_{gm}, j\in\mathcal{N}_K^i} \rho'(|\alpha_{ig}^{(t)} - \alpha_{jg}^{(t)}|)|\alpha_{ig} - \alpha_{jg}| \\ & + \lambda \sum_{g=1}^{p}\sum_{m\neq m'}\sum_{i\in G_{gm}, j\in G_{gm'}} \overline{\rho}(\alpha_{ig}^{(t)} - \alpha_{jg}^{(t)})\{(\alpha_{ig} - \alpha_{ig}^*) - (\alpha_{jg} - \alpha_{jg}^*)\}.\end{aligned}$$

As shown in (A.20),

$$\sup_i |(\alpha_{ig}^* - \alpha_{ig}^0)| = \sup_m |\theta_{gm} - \theta_{gm}^0| \leqslant \phi_n. \tag{A.21}$$

Since $[\boldsymbol{\alpha}]_g^{(t)} = \varsigma[\boldsymbol{\alpha}]_g + (1-\varsigma)[\boldsymbol{\alpha}^*]_g$ for some constant $\varsigma \in (0,1)$, then for every component

$$\sup_i |\alpha_{ig}^{(t)} - \alpha_{ig}^0| \leqslant \varsigma \sup_i |\alpha_{ig} - \alpha_{ig}^0| + (1-\varsigma)\sup_i |\alpha_{ig}^* - \alpha_{ig}^0| \leqslant \phi_n, \tag{A.22}$$

and then for $m \neq m', i \in G_{gm}, j \in G_{gm'}$,

$$\begin{aligned}|\alpha_{ig}^{(t)} - \alpha_{jg}^{(t)}| \quad & \geqslant \min_{i\in G_{gm}, j\in G_{gm'}} |\alpha_{ig}^0 - \alpha_{jg}^0| - 2\max_i |\alpha_{ig}^{(t)} - \alpha_{ig}^0|, \\ & \geqslant b_n - 2\max_i |\alpha_{ig}^{(t)} - \alpha_{ig}^0| \geqslant b_n - 2\phi_n > a\lambda,\end{aligned}$$

and thus $\overline{\rho}(\alpha_{ig}^{(t)} - \alpha_{jg}^{(t)}) = 0$. Therefore,

$$\Gamma_2 = \lambda \sum_{g=1}^{p}\sum_{m=1}^{M_g^0}\sum_{i,j\in G_{gm}, j\in\mathcal{N}_K^i} \rho'(|\alpha_{ig}^{(t)} - \alpha_{jg}^{(t)}|)|\alpha_{ig} - \alpha_{jg}|.$$

Furthermore, by the same reasoning as (A.21), we have

$$\sup_i |(\alpha_{ig}^* - \widehat{\alpha}_{ig}^{or})| = \sup_m |\theta_{gm} - \widehat{\theta}_{gm}^{or}| \leqslant \sup_i |(\alpha_{ig} - \widehat{\alpha}_{ig}^{or})|.$$

Then

$$\begin{aligned}\sup_i |\alpha_{ig}^{(t)} - \alpha_{jg}^{(t)}| \leqslant \quad & 2\sup_i |\alpha_{ig}^{(t)} - \alpha_{ig}^{*}| \leqslant 2\sup_i |\alpha_{ig} - \alpha_{ig}^{*}| \\ & \leqslant 2(\sup_i |\alpha_{ig} - \widehat{\alpha}_{ig}^{or}| + \sup_i |\alpha_{ig}^{*} - \widehat{\alpha}_{ig}^{or})| \\ & \leqslant 4(\sup_i |\alpha_{ig} - \widehat{\alpha}_{ig}^{or}|) \leqslant 4t_n \end{aligned} \tag{A.23}$$

Hence, $\rho'(|\alpha_{ig}^{(t)} - \alpha_{jg}^{(t)}|) \geqslant \rho'(4t_n)$ by concavity of $\rho(\cdot)$. As a result,

$$\Gamma_2 \geqslant \lambda \sum_{g=1}^{p} \sum_{m=1}^{M_g^0} \sum_{i,j \in G_{gm}, j \in \mathcal{N}_K^i} \rho^{'}(4t_n)|\alpha_{ig} - \alpha_{jg}|. \tag{A.24}$$

Let

$$\begin{aligned}\boldsymbol{D} = (\boldsymbol{D}_1^T, \cdots, \boldsymbol{D}_p^T)^T \quad & = [(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta} - \sum_{g=1}^{p} \mathbf{Z}^g [\boldsymbol{\alpha}]_g^{(t)})^T (\mathbf{Z}_{(1)}, \cdots, \mathbf{Z}_{(p)})]^T \\ & = [(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha}^{(t)})^T (\mathbf{Z}^1, \cdots, \mathbf{Z}^p)]^T. \end{aligned}$$

Where $\mathbf{Z} = (\mathbf{Z}^1, \cdots, \mathbf{Z}^p), \boldsymbol{\alpha}^{(t)} = ([\boldsymbol{\alpha}]_1^{(t)}, \cdots, [\boldsymbol{\alpha}]_p^{(t)})$. Let $D_{gi}$ represents the $i$-th component of $\mathbf{D}_g$

Then

$$\begin{aligned}\Gamma_1 \quad & = -\boldsymbol{D}^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) = -\sum_{g=1}^{p} \sum_{m=1}^{M_g^0} \sum_{i,j \in \boldsymbol{G}_{gm}} \frac{D_{gi}(\alpha_{ig} - \alpha_{jg})}{|\boldsymbol{G}_{gm}|} \\ & = -\sum_{g=1}^{p} \sum_{m=1}^{M_g^0} \sum_{i,j \in \boldsymbol{G}_{gm}} \frac{D_{gi}(\alpha_{ig} - \alpha_{jg})}{2|\boldsymbol{G}_{gm}|} - \sum_{g=1}^{p} \sum_{m=1}^{M_g^0} \sum_{i,j \in \boldsymbol{G}_{gm}} \frac{D_{ig}(\alpha_{ig} - \alpha_{jg})}{2|\boldsymbol{G}_{gm}|} \\ & = -\sum_{g=1}^{p} \sum_{m=1}^{M_g^0} \sum_{i,j \in \boldsymbol{G}_{gm}} \frac{(D_{gj} - D_{gi})(\alpha_{jg} - \alpha_{ig})}{2|\boldsymbol{G}_{gm}|} \\ & = -\sum_{g=1}^{p} \sum_{m=1}^{M_g^0} \sum_{i,j \in \boldsymbol{G}_{gm}, i<j} \frac{(D_{gj} - D_{gi})(\alpha_{jg} - \alpha_{ig})}{|\boldsymbol{G}_{gm}|}. \end{aligned} \tag{A.25}$$

Since

$$D_{gi} = (\boldsymbol{\epsilon}_i + \mathbf{X}_i(\boldsymbol{\beta}^0 - \boldsymbol{\beta}) + \mathbf{Z}_i(\boldsymbol{\alpha}_i^0 - \boldsymbol{\alpha}_i^{(t)}))Z_{gi},$$

then

$$\begin{aligned}\max_{i,j} |D_{gj} - D_{gi}| \leqslant 2\max_i |D_{gi}| \quad & \leqslant 2\max_i |(\epsilon_i + \mathbf{X}_i(\boldsymbol{\beta}^0 - \boldsymbol{\beta}) + Z_i(\boldsymbol{\alpha}_i^0 - \boldsymbol{\alpha}_i^{(t)}))||Z_{gi}| \\ & \leqslant 2(\|\boldsymbol{\epsilon}\|_\infty + \sup_i \|\mathbf{X}_i\| \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}\| + \sup_i \|\mathbf{Z}_i\| \sup_i \|\boldsymbol{\alpha}_i^0 - \boldsymbol{\alpha}_i^{(t)}\|) \sup_i \|\mathbf{Z}_i\|. \end{aligned}$$

By the above equation $\sup_i |\alpha_{ig}^0 - \alpha_{ig}^{(t)}| \leqslant \sup_i \|\boldsymbol{\alpha}_i^0 - \boldsymbol{\alpha}_i^{(t)}\| \leqslant \phi_n$ and $\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}\| \leqslant \phi_n$, by Condition $(C1)$ that $\sup_i \|\mathbf{Z}_i\| \leqslant C_2\sqrt{p}$ and $\sup_i \|\mathbf{X}_i\| \leqslant C_3\sqrt{q}$, we have

$$\max_{i,j} |D_{gj} - D_{gi}| \leqslant 2C_2\sqrt{p}(\|\boldsymbol{\epsilon}\|_\infty + C_3\sqrt{q}\phi_n + C_2\sqrt{p}\phi_n).$$

By condition $(C3)$

$$\mathrm{pr}(\|\boldsymbol{\epsilon}\|_\infty > \sqrt{2c_1^{-1}}\sqrt{\log n}) \leqslant \sum_{i=1}^{n} P(|\epsilon| > \sqrt{2c_1^{-1}}\sqrt{\log n}) \leqslant 2n^{-1}.$$

Thus, by equation (A.18) and Condition $(C3)$, there is an event $E_2 = \{\{\|\epsilon\|_\infty \leqslant \sqrt{2c_1^{-1}}\sqrt{\log n}\} \cap \{K' \geqslant \frac{c_{1,d}g_{\min K}}{8c_{2,d}g_{\max}}\}\}$ such that

$$\mathrm{pr}(E_2^c) \leqslant 2n^{-1} + np\exp\left(-\frac{p_{\min}c_{1,r}K}{48c_{2,r}p_{\max}}\right) + n^2p\exp\left(-\frac{K}{3}\right),$$

and on the event $E_2$, by (A.25),

$$\begin{aligned}
\Gamma_1 &= -\sum_{g=1}^{p}\sum_{m=1}^{M_g^0}\sum_{i,j\in G_{gm},i<j}\frac{(D_{gj}-D_{gi})(\alpha_{jg}-\alpha_{ig})}{|G_{gm}|}\\
&\geqslant -\sum_{g=1}^{p}\sum_{m=1}^{M_g^0}\sum_{i,j\in G_{gm},i<j}\frac{\max_{i,j}|D_{gj}-D_{gi}|}{|G_{gm}|}|\alpha_{jg}-\alpha_{ig}|\\
&= -\sum_{g=1}^{p}\sum_{m=1}^{M_g^0}\sum_{i,j\in G_{gm},j\in\mathcal{N}_K^i} O(\frac{|G_{gm}|}{2K'})\frac{2\max_{i,j}|D_{gj}-D_{gi}|}{|G_{gm}|}|\alpha_{jg}-\alpha_{ig}|
\end{aligned}$$

$$\max_{i,j} |D_{gj} - D_{gi}| \leqslant 2C_2\sqrt{p}(\sqrt{2c_1^{-1}}\sqrt{\log n} + C_3\sqrt{q}\phi_n + C_2\sqrt{p}\phi_n).$$

Therefore, by (A.24) and (A.25), we have

$$\begin{aligned}
&Q_n(\boldsymbol{\beta},\boldsymbol{\alpha}) - Q_n(\boldsymbol{\beta},\boldsymbol{\alpha}^*) = \Gamma_1 + \Gamma_2\\
&\geqslant \sum_{g=1}^{p}\sum_{m=1}^{M_g^0}\sum_{i,j\in G_{gm},j\in\mathcal{N}_K^i}[\lambda\rho'(4t_n) - O(\frac{|G_{gm}|}{2K'})\frac{2\max_{i,j}|D_{gj}-D_{gi}|}{|G_{gm}|}]|\alpha_{ig}-\alpha_{jg}|\\
&\geqslant \sum_{g=1}^{p}\sum_{m=1}^{M_g^0}\sum_{i,j\in G_{gm},j\in\mathcal{N}_K^i}[\lambda\rho'(4t_n) - O(\frac{\max_{i,j}|D_{gj}-D_{gi}|}{K'})]|\alpha_{ig}-\alpha_{jg}|\\
&\geqslant \sum_{g=1}^{p}\sum_{m=1}^{M_g^0}\sum_{i,j\in G_{gm},j\in\mathcal{N}_K^i}\{\lambda\rho'(4t_n) - 2C_2O(K'^{-1})\sqrt{p}(\sqrt{2c_1^{-1}}\sqrt{\log n} + C_3\sqrt{q}\phi_n + C_2\sqrt{p}\phi_n)\}|\alpha_{ig}-\alpha_{jg}|.
\end{aligned}$$

Let $t_n = o(1)$, then $\rho'(4t_n) \to 1$. Since $|G_{min}| \gg \sqrt{(q + \sum_{g=1}^{p} M_g^0) n \log n}$, $p = o(n)$ and $K' \gg \max\{\sqrt{p \log n}, p\phi_n, \log n^2 p, \sqrt{pq}\phi_n\}$, then $\lambda \gg \frac{\sqrt{p \log n}}{K'}$, $\lambda \gg \frac{\sqrt{pq}\phi_n}{K'}$, $\lambda \gg \frac{p\phi_n}{K'}$ and $\lambda \gg \phi_n$. Thus, $Q_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) - Q_n(\boldsymbol{\beta}, \boldsymbol{\alpha}^*) \geqslant 0$ for sufficiently large $n$, the result in (ii) is proved. □

Remark 2: From the oracle property in the Theorem 2, we have that $\text{pr}(\widehat{M}_1 = M_1^0, \cdots, \widehat{M}_p = M_p^0) \to 1$. Moreover, since $[\widehat{\boldsymbol{\alpha}}]_g^{or} = \mathbf{W}_g \widehat{\boldsymbol{\theta}}_g^{or}$ and $[\widehat{\boldsymbol{\alpha}}]_g = \widehat{\mathbf{W}}_g \widehat{\boldsymbol{\theta}}_g$ for $g = 1, \cdots, p$, then $\text{pr}(\widehat{\mathbf{W}}_g = \mathbf{W}_g, g = 1, \cdots, p) \to 1$. Hence, the subgroup memberships can be recovered with a high probability. Corollary 1 can be easily proof.

### A3.8 *Proof of Theorem 3*

The proof of Theorem 3 follows analogously from Theorems 1-2 herein and Theorem 2 on Fan and Peng (2004), and is thus omitted.

### A3.9 *Proof of Theorem 4*

In this section, we show the results in Theorem 4 and corollary 2. Similiar to the Proposition 2 of the Wu and Zhou (2006), we first give the following definitions are introduced

$$\widehat{f}_{gz}^{or}(\mathbf{U}) = \arg \min_{f_g \in \overline{\mathcal{H}}_{\mathcal{K}}} \left\{ \mathcal{E}_z(f_g(\mathbf{U})) + \lambda_n \| f_g^*(\mathbf{U}) \|_{\mathcal{K}}^2 \right\},$$

where $\mathcal{E}_z(f_g(\mathbf{U})) = \frac{1}{n} \sum_{i=1}^{n} V(\widetilde{Y}_{gi}^{0*}, f_g(\mathbf{U}_i))$. Besides, let

$$f_g^{or}(\mathbf{U}) = \arg \min_{f_g \in \overline{\mathcal{H}}_{\mathcal{K}}} \left\{ \mathcal{E}(f_g(\mathbf{U})) + \lambda_n \| f_g^*(\mathbf{U}) \|_{\mathcal{K}}^2 \right\}.$$

where $\mathcal{E}(f_g)$ is the generalization error. And $\widetilde{Y}_{gi}^{0*}$ represents that Pseudo-labeling are given based on the real subgroup memberships. If we use $\widehat{M}_g$ the estimator in the first stage replacing unknown $M_g$. Then we can get

$$\widehat{f}_g(\mathbf{U}) = \arg \min_{f_g \in \overline{\mathcal{H}}_{\mathcal{K}}} \left\{ \widehat{\mathcal{E}}_z(f_g(\mathbf{U})) + \lambda_n \| f_g^* \|_{\mathcal{K}}^2 \right\},$$

where $\widehat{\mathcal{E}}_z(f_g(\mathbf{U})) = \frac{1}{n}\sum_{i=1}^{n} V(\widetilde{Y}^*_{gi}, f_g(\mathbf{U}_i))$ and $\widehat{\mathcal{E}}(f_g) = E(\widetilde{Y}^*_{gi}, f_g(\mathbf{U}))$. $\widetilde{Y}^*_{gi}$ represents that Pseudo-labeling are given based on the subgroup memberships by the first step. In which

$$\widehat{f}_{gz}(\mathbf{U}) = \arg\min_{f_g \in \overline{\mathcal{H}}_{\mathcal{K}}} \left\{ \widehat{\mathcal{E}}(f_g) + \lambda_n \|f_g^*\|^2_{\mathcal{K}} \right\}.$$

In order to make $\widehat{f}_g(\mathbf{U})$ consistent with the notation in the paper, our definition of $\widehat{f}_{gz}(\mathbf{U})$ violates common conventions. Note for simplicity, we let $\widehat{f}^{or}_{gz}(\mathbf{U}), \widehat{f}^{or}_{gc}(\mathbf{U}), f^{or}_g(\mathbf{U}), \widehat{f}_{gz}(\mathbf{U}), \widehat{f}_{gc}(\mathbf{U}), \widehat{f}_g(\mathbf{U})$ abbreviated as $\widehat{f}^{or}_{gz}, \widehat{f}^{or}_{gc}, \widehat{f}^{or}_g, \widehat{f}_{gz}, \widehat{f}_{gc}, \widehat{f}_g$. We first given the lemma 2 about the error analysis.

Lemma 2: *For every $\lambda_n > 0$, there holds*

$$\mathcal{R}(\widehat{f}^{or}_{gz}) - \mathcal{R}(\widehat{f}^{or}_{gc}) \leqslant \mathcal{S}(n, \lambda_n) + \mathcal{D}(\lambda_n),$$

*where*

$$\mathcal{S}(n, \lambda_n) = \left\{ \mathcal{E}(\widehat{f}^{or}_{gz}) - \mathcal{E}_z(\widehat{f}^{or}_{gz}) \right\} + \left\{ \mathcal{E}_z(\widehat{f}^{or}_{g}) - \mathcal{E}((f^{or}_g)) \right\}.$$

*and*

$$\mathcal{D}(\lambda_n) = \inf_{f_g \in \overline{\mathcal{H}}_{\mathcal{K}}} \left\{ \mathcal{E}(f_g) - \mathcal{E}(\widehat{f}^{or}_{gc}) + \lambda_n \|f_g^*\|^2_{\mathcal{K}} \right\}.$$

**Proof**. We can write

$$\begin{aligned}\mathcal{E}(\widehat{f}^{or}_{gz}) - \mathcal{E}(\widehat{f}^{or}_{gc}) = \quad & \left\{ \mathcal{E}(\widehat{f}^{or}_{gz}) - \mathcal{E}_z(\widehat{f}^{or}_{gz}) \right\} + \left\{ \left( \mathcal{E}_z(\widehat{f}^{or}_{gz}) + \lambda_n \|\widehat{f}^{*or}_{gz}\|^2_{\mathcal{K}} \right) - \left( \mathcal{E}_z(f^{or}_g) + \lambda_n \|f^{*or}_g\|^2_{\mathcal{K}} \right) \right\} \\ & + \left\{ \mathcal{E}_z(f^{or}_g) - \mathcal{E}(f^{or}_g) \right\} + \left\{ \mathcal{E}(f^{or}_g) - \mathcal{E}(\widehat{f}^{or}_{gc}) + \lambda_n \|f^{*or}_g\|^2_{\mathcal{K}} \right\} - \lambda_n \|\widehat{f}^{*or}_{gz}\|^2_{\mathcal{K}}.\end{aligned}$$

By the definition of $\widehat{f}^{or}_{gz}$, the second term is $\leqslant 0$. By the definition of $f^{or}_g$ we see the fourth term is just $\mathcal{D}(\lambda_n)$. A bridge between $\mathcal{R}(f)$ and $\mathcal{E}(f)$ was establish by Zhang (2004) that $\mathcal{R}(f) - \mathcal{R}(f_c) \leqslant \mathcal{E}(f) - \mathcal{E}(f_c)$ for any $f$. Hence, the lemma is obvious satisfied. □

Lemma 3: *For any $\lambda_n > 0$, $n \in N$ ,there hold $\|\widehat{f}^{*or}_{gz}\|_{\mathcal{K}} \leqslant \frac{1}{\sqrt{\lambda_n}}$, $|\widehat{\gamma}^{or}_g| \leqslant 1 + \frac{\kappa}{\sqrt{\lambda_n}}$ and $\mathcal{E}_z(\widehat{f}^{or}_{gz}) \leqslant 1$. Note $\kappa = \sup_{u \in \mathcal{U}} \sqrt{K(u,u)}$.*

**Proof**. By the definition of $\widehat{f}^{or}_{gz}$, and the choice $f = 0 + 0$, we have

$$\mathcal{E}_z(\widehat{f}^{or}_{gz}) + \lambda_n \|f^*_{gz}\|^2_{\mathcal{K}} \leqslant \frac{1}{n}\sum_{i=1}^{n} V(\widetilde{Y}^{0*}_{gi}, 0) + 0 = 1.$$

Thus, $\mathcal{E}_z(\widehat{f}_{gz}^{or}) \leqslant 1$ and $\|\widehat{f}_{gz}^{*or}\|_{\mathcal{K}} \leqslant \frac{1}{\sqrt{\lambda_n}}$. By the reproducing property, we have $\|\widehat{f}_{gz}^{*or}\|_\infty \leqslant \kappa\|\widehat{f}_{gz}^{*or}\|_{\mathcal{K}}$.

$$|\widehat{\gamma}_g^{or}| \leqslant \min_{1\leqslant i\leqslant n} |f_{gz}(\mathbf{U}_i)| + \|\widehat{f}_{gz}^{*or}\|_\infty \leqslant 1 + \frac{\kappa}{\sqrt{\lambda_n}}.$$

Therfore, we only need to prove $\min_{1\leqslant i\leqslant n} |f_{gz}(\mathbf{U}_i)| \leqslant 1$. Supppose $\min_{1\leqslant i\leqslant n} |f_{gz}(\mathbf{U}_i)| > 1$, then for each $i$, either $\widetilde{Y}_{gi}^* f_{gz}(\mathbf{U}_i) \geqslant \min_{1\leqslant i\leqslant n} |f_{gz}(\mathbf{U}_i)| > 1$ or $\widetilde{Y}_{gi}^* f_{gz}(\mathbf{U}_i) \leqslant -\min_{1\leqslant i\leqslant n} |f_{gz}(\mathbf{U}_i)| < -1$. For $\epsilon \in \{-1, 1\}$, set

$$I_\epsilon = \{i \in \{1, \cdots, n\} : \widetilde{Y}_{gi}^* = \epsilon, \widetilde{Y}_{gi}^* f_{gz}(\mathbf{U}_i) \leqslant -\min_{1\leqslant i\leqslant n} |f_{gz}(\mathbf{U}_i)|\}$$

Let $\min_{1\leqslant i\leqslant n} |f_{gz}(\mathbf{U}_i)| = r = f_{gz}(\mathbf{U}_{i0})$ and $\#I_\epsilon$ represent the number of elements in the set $I_\epsilon$. If $\#I_1 = \#I_{-1}$ (possibly zero), then the function $\widetilde{f}_{gz} = f_{gz} - d$ with $d = (r-1)sgnf_{gz}(\mathbf{U}_{i0})$ satisfies $|\widetilde{f}_{gz}(\mathbf{U}_{i0})| = 1$ and $|\widetilde{f}_{gz}(\mathbf{U}_i)| \geqslant 1$ for each $i$. Hence,

$$\mathcal{E}(\widetilde{f}_{gz}) = \sum_{i\in I_1\cup I_{-1}} (1 - \widetilde{Y}_{gi}^* f_{gz}(\mathbf{U}_i) + \widetilde{Y}_{gi}^* d) = \mathcal{E}(f_{gz}) + \sum_{i\in I_1} d - \sum_{i\in I_{-1}} d = \mathcal{E}(f_{gz}).$$

This means that $\widetilde{f}_{gz} = f_{gz}$ and $\min_{1\leqslant i\leqslant n} |f_{gz}(\mathbf{U}_i)| \leqslant 1$. □

Lemma 4: *For every $\lambda_n > 0$ and $\epsilon > 0$,there holds*

$$Prob\{\mathcal{S}(n, \lambda_n) > \epsilon\} \leqslant \exp\left\{-\frac{3n\epsilon^2}{256B}\right\} + \left(\frac{32B}{\epsilon} + 1\right)\mathcal{N}\left(\frac{\epsilon\sqrt{\lambda_n}}{32}\right)\exp\left\{-\frac{3n\epsilon^2}{2^{14}B}\right\},$$

*where $B = B_{\lambda_n} = 1 + \frac{\kappa}{\sqrt{\lambda_n}}$.*

**Proof**. Recall the one-sided Bernstein inequality: Suppose a random variable $\xi$ has mean $\mu$ and variance $\sigma^2$ and satisfies $|\xi - \mu| \leqslant M$. Let $(\xi_i)_{i=1}^n$ be independent samples. We have

$$\text{pr}\left\{\mu - \frac{1}{n}\sum_{i=1}^n \xi_i > \epsilon\right\} \leqslant \exp\left\{-\frac{n\epsilon^2}{2(\sigma^2 + \frac{1}{3}M\epsilon))}\right\}.$$

Then, we ask random variable $\xi$ satisfies $0 \leqslant \xi \leqslant M$, $\mu = E\xi > 0$, and $|\xi - \mu| \leqslant M$ also satisfied. By Bernstein inequality, we have

$$\text{pr}\left\{\frac{\mu - \frac{1}{n}\sum_{i=1}^n \xi_i}{\mu + \epsilon} > \alpha\right\} \leqslant \exp\left\{-\frac{n\alpha^2(\mu+\epsilon)^2}{2(\sigma^2 + \frac{1}{3}M\alpha(\mu+\epsilon))}\right\}.$$

where $0< \alpha \leqslant 1$ and $\epsilon > 0$. Here $\sigma^2 \leqslant E(\xi^2) \leqslant ME(\xi) = M\mu$ since $0 \leqslant \xi \leqslant M$. Then we find that

$$\sigma^2 + \frac{1}{3}M\alpha(\mu+\epsilon) \leqslant \frac{4}{3}M(\mu+\epsilon) \leqslant \frac{4M(\mu+\epsilon)^2}{3\epsilon}.$$

Therefore, we have

$$\mathrm{pr}\left\{\frac{\mu - \frac{1}{n}\sum_{i=1}^{n}\xi_i}{\mu+\epsilon} > \alpha\right\} \leqslant \exp\left\{-\frac{3n\alpha^2\epsilon}{8M}\right\}. \tag{A.26}$$

In addition, it suffices to show that

$$\frac{\mu - \frac{1}{m}\sum_{i=1}^{m}\xi_i}{\mu+\epsilon} \leqslant \frac{\alpha}{2} \Rightarrow \frac{\mu - \frac{1}{m}\sum_{i=1}^{m}\xi_i}{\frac{1}{m}\sum_{i=1}^{m}\xi_i + \epsilon} \leqslant \alpha.$$

By (A.26) and above suffices inequality, we can have

$$\mathrm{pr}\left\{\frac{\mu - \frac{1}{n}\sum_{i=1}^{n}\xi_i}{\frac{1}{m}\sum_{i=1}^{m}\xi_i+\epsilon} > \alpha\right\} \leqslant \exp\left\{-\frac{3n\alpha^2\epsilon}{32M}\right\}. \tag{A.27}$$

To bound the sample error $\mathcal{S}(n,\lambda_n)$, we need to use the concentration inequalities concerning the uniform convergence. To use the technique, we need the concept of covering numbers to measure the capacity of the hypothesis space. we use the oridnary definition in machine learning.

Definition 1: For a compact set $\mathcal{F}$ in a metric space and $\epsilon > 0$, the covering number $\mathcal{N}(\mathcal{F},\epsilon)$ is defined to be the minimal integer $\ell \in N$ such that there exist $\ell$ balls with radius $\epsilon$ covering $\mathcal{F}$.

Let $\mathcal{F}$ be a subset of $\lambda(U)$ such that $\|f_g^{or}\|_\infty \leqslant M$ for every $f_g^{or} \in \mathcal{F}$, and $\{f_{gj}^{or}\}_{j=1}^{N} \subset \mathcal{F}$ with $N = \mathcal{N}(\mathcal{F},\alpha\epsilon)$ such that $\mathcal{F}$ is covered by balls centered at $f_{gj}^{or}$ wirh radius $\alpha\epsilon$. Note that for every $f_g^{or} \in \mathcal{F}$ the random variable $\xi = V(\widetilde{Y}_g^{*0}, f_g^{or}(U))$ satisfies $0 \leqslant \xi \leqslant 1 + \|f_g^{or}\|_\infty \leqslant 1 + M$. Then for each $j$, we can get following inequality by (A.27)

$$\mathrm{pr}\left\{\frac{\mathcal{E}(f_{gj}^{or}) - \mathcal{E}_z(f_{gj}^{or})}{\mathcal{E}_z(f_{gj}^{or})+\epsilon} \geqslant \alpha\right\} \leqslant \exp\left\{-\frac{3n\alpha^2\epsilon}{32(1+M)}\right\}.$$

By the Lipschitz property of the loss function $V$, we find that

$$|\mathcal{E}_z(f) - \mathcal{E}_z(g)| \leqslant \|f-g\|_\infty, \quad |\mathcal{E}(f) - \mathcal{E}(g)| \leqslant \|f-g\|_\infty,$$

where $f$ and $g \in \mathcal{F}$. For each $f_g^{or} \in \mathcal{F}$, there is some $j$ such that $\|f_g^{or} - \widehat{f}_{gj}^{or}\|_\infty \leqslant \alpha\epsilon$. Hence, we have

$$|\mathcal{E}_z(f_g^{or}) - \mathcal{E}_z(f_{gj}^{or})| \leqslant \alpha\epsilon, \quad |\mathcal{E}(f_g^{or}) - \mathcal{E}(f_{gj}^{or})| \leqslant \alpha\epsilon.$$
$$\Rightarrow \qquad \frac{|\mathcal{E}_z(f_g^{or}) - \mathcal{E}_z(f_{gj}^{or})|}{\mathcal{E}_z(f_g^{or}) + \epsilon} \leqslant \alpha \quad \text{and} \quad \frac{|\mathcal{E}(f_g^{or}) - \mathcal{E}(f_{gj}^{or})|}{\mathcal{E}_z(f_g^{or}) + \epsilon} \leqslant \alpha.$$

The former implies that $\mathcal{E}_z(f_{gj}^{or}) + \epsilon \leqslant 2[\mathcal{E}_z(f_g^{or}) + \epsilon]$. Hence, we have

$$\begin{aligned} &\text{pr}\left\{\sup_{\widehat{f}_g \in \mathcal{F}} \frac{\mathcal{E}(f_g^{or}) - \mathcal{E}_z(f_g^{or})}{\mathcal{E}_z(f_g^{or}) + \epsilon} \geqslant 4\alpha\right\} \leqslant \sum_{j=1}^{N} Prob\left\{\frac{\mathcal{E}(f_{gj}^{or}) - \mathcal{E}_z(f_{gj}^{or})}{\mathcal{E}_z(f_{gj}^{or}) + \epsilon} \geqslant \alpha\right\} \\ &\leqslant N \exp\left\{-\frac{n\alpha^2\epsilon}{32(1+M)}\right\} \leqslant \mathcal{N}(\mathcal{F}, \alpha\epsilon) \exp\left\{-\frac{n\alpha^2\epsilon}{32(1+M)}\right\}. \end{aligned} \tag{A.28}$$

Actually, $\widehat{f}_{gz}^{or}$ always lies in the set $\mathcal{F}_{R,B} = \{f_g^{or} : f_g^{or} = f_g^{*or} - \gamma_g^{or} \in \mathcal{B}_R - [-B, B]\}$ with $R = 1/\sqrt{\lambda_n}$ and $B = 1 + \frac{\kappa}{\sqrt{\lambda_n}}$ by lemma 3. Since

$$\|(f_g^{*or} - \gamma_g^{or}) - (h_g^{*or} - \gamma_h^{or})\|_\infty \leqslant \|f_g^{*or} - h_g^{*or}\|_\infty + |\gamma_g^{or} - \gamma_h^{or}|.$$

Then, $\mathcal{N}(\mathcal{F}_{R,B}, \epsilon)$ is bounded by $(\frac{2B}{\epsilon} + 1)\mathcal{N}(\mathcal{B}_R, \epsilon)$. An $\frac{\epsilon}{2R}$-covering of $\mathcal{B}_1$ is the same as an $\frac{\epsilon}{2}$-covering of $\mathcal{B}_R$. Therefore, we have

$$\mathcal{N}(\mathcal{F}_{R,B}, \epsilon) \leqslant \left(\frac{2B}{\epsilon} + 1\right) \mathcal{N}\left(\frac{\epsilon}{2R}\right). \tag{A.29}$$

Let the random variable $\xi = V(\widetilde{Y}_g^{*0}, f_g^{or})$ satisfies $0 \leqslant \xi \leqslant 2B$ by the lemma 3. By the fact $\mathcal{E}_z(\widehat{f}_{gz}^{or}) \leqslant 1$ we obtain

$$\text{pr}\left\{\mathcal{E}_z(f_g^{or}) - \mathcal{E}(f_g^{or}) > \frac{\epsilon}{2}\right\} \leqslant Prob\left\{\frac{\mathcal{E}_z(f_g^{or}) - \mathcal{E}(f_g^{or})}{\mathcal{E}(f_g^{or}) + 1} > \frac{\epsilon}{4}\right\} \leqslant \exp\left\{-\frac{3n\epsilon^2}{256B}\right\}. \tag{A.30}$$

where the second inequality by (A.26) obtained. Then for $\widehat{f}_{gz}^{or} \in \mathcal{F}_{R,B}^{or}$ and $\mathcal{E}_z(\widehat{f}_{gz}^{or}) \leqslant 1$, we have

$$\begin{aligned} \text{pr}\left\{\mathcal{E}(\widehat{f}_{gz}^{or}) - \mathcal{E}_z(\widehat{f}_{gz}^{or}) > \frac{\epsilon}{2}\right\} \quad &\leqslant Prob\left\{\frac{\mathcal{E}(\widehat{f}_{gz}^{or}) - \mathcal{E}_z(\widehat{f}_{gz}^{or})}{\mathcal{E}_z(\widehat{f}_{gz}^{or}) + 1} > \frac{\epsilon}{4}\right\} \\ &\leqslant Prob\left\{\sup_{f_g \in \mathcal{F}_{R,B}} \frac{\mathcal{E}(f_g^{or}) - \mathcal{E}_z(f_g^{or})}{\mathcal{E}_z(f_g^{or}) + 1} > \frac{\epsilon}{4}\right\} \\ &\leqslant \mathcal{N}\left(\mathcal{F}_{R,B}, \frac{\epsilon}{16}\right) \exp\left\{-\frac{3n\epsilon^2}{2^{14}B}\right\} \\ &\leqslant \left(\frac{32B}{\epsilon} + 1\right) \mathcal{N}\left(\frac{\epsilon}{32R}\right) \exp\left\{-\frac{3n\epsilon^2}{2^{14}B}\right\}. \end{aligned} \tag{A.31}$$

where the inequality by (A.26), (A.28) and (A.29) obtained. Then, by (A.30) and (A.31) we have

$$\begin{aligned}\mathrm{pr}\{\mathcal{S}(n,\lambda_n)>\epsilon\} &\leqslant Prob\left\{\mathcal{E}(\widehat{f}_{gz}^{or})-\mathcal{E}_z(\widehat{f}_{gz}^{or})>\frac{\epsilon}{2}\right\}+Prob\left\{\mathcal{E}_z(f_g^{or})-\mathcal{E}(f_g^{or})>\frac{\epsilon}{2}\right\}\\ &\leqslant\left(\frac{32B}{\epsilon}+1\right)\mathcal{N}\left(\frac{\epsilon}{32R}\right)\exp\left\{-\frac{3n\epsilon^2}{2^{14}B}\right\}+\exp\left\{-\frac{3n\epsilon^2}{256B}\right\}\\ &\leqslant\left(\frac{32B}{\epsilon}+1\right)\mathcal{N}\left(\frac{\epsilon\sqrt{\lambda_n}}{32}\right)\exp\left\{-\frac{3n\epsilon^2}{2^{14}B}\right\}+\exp\left\{-\frac{3n\epsilon^2}{256B}\right\}. \qquad (A.32)\end{aligned}$$

□

**The proof of Theorem 4 :**

$$\mathcal{R}(\mathrm{sgn}(\widehat{f}_g(\mathbf{U})))-\mathcal{R}(\widehat{f}_{gc}^{or}(\mathbf{U}))=\left(\mathcal{R}(\mathrm{sgn}(\widehat{f}_g(\mathbf{U})))-\mathcal{R}(\mathrm{sgn}(\widehat{f}_{gz}^{or}(\mathbf{U})))\right)+\left(\mathcal{R}(\mathrm{sgn}(\widehat{f}_{gz}^{or}(\mathbf{U})))-\mathcal{R}(\widehat{f}_{gc}^{or}(\mathbf{U}))\right).$$

By Corollary 1 (1), the subgroup memberships can be recpvered with a high probability. That is, the first term $\mathrm{pr}\left\{\mathcal{R}(\mathrm{sgn}(\widehat{f}_g(\mathbf{U})))-\mathcal{R}(\mathrm{sgn}(\widehat{f}_{gz}^{or}(\mathbf{U})))>\epsilon\right\}\to 0$.

By Lemma 2 and Lemma 4 we immediately obtaian that for $0<\epsilon<1$,

$$\mathrm{pr}\left\{\mathcal{R}(\mathrm{sgn}(\widehat{f}_{gz}^{or})-\mathcal{R}(\widehat{f}_{gc}^{or}))>\epsilon+\mathcal{D}(\lambda_n)\right\}\leqslant\frac{34B}{\epsilon}\mathcal{N}(\frac{\epsilon\sqrt{\lambda_n}}{32})\exp\left\{-\frac{3n\epsilon^2}{2^{14}B}\right\}. \qquad (A.33)$$

By (A.33) and the conditions in Theorem 4, it is easy have the second term $\mathrm{pr}\left\{\mathcal{R}(\mathrm{sgn}(\widehat{f}_{gz}^{or}(\mathbf{U})))-\mathcal{R}(\widehat{f}_{gc}^{or}(\mathbf{U}))>\epsilon\right\}\to 0$. Therefore,

$$\begin{aligned}\mathrm{pr}\left\{\mathcal{R}(\mathrm{sgn}(\widehat{f}_g(\mathbf{U})))-\mathcal{R}(\widehat{f}_{gc}^{or}(\mathbf{U}))>\epsilon\right\}\leqslant\ &\mathrm{pr}\left\{\mathcal{R}(\mathrm{sgn}(\widehat{f}_g(\mathbf{U})))-\mathcal{R}(\mathrm{sgn}(\widehat{f}_{gz}^{or}(\mathbf{U})))>\epsilon/2\right\}+\\ &\mathrm{pr}\left\{\mathcal{R}(\mathrm{sgn}(\widehat{f}_{gz}^{or}(\mathbf{U})))-\mathcal{R}(\widehat{f}_{gc}^{or}(\mathbf{U}))>\epsilon/2\right\}\to 0.\end{aligned}$$

□

**The proof of Theorem A.1:**

Similar to binary classification, for any $g$, $M_g^0>2$, and $M_g^0$ is a fixed constant. If the underlying group memberships $G_{g1}^0,\cdots,G_{gM_g^0}^0$ for $g=1,\cdots,p$ were konwn, the oracle estimate classifier $f_g$,

$$f_g^{or}(\mathbf{U})=\arg\min_{f_g\in\prod_{m=1}^{M_g^0}\overline{\mathcal{H}}_{\mathcal{K}},\sum_{m=1}^{M_g^0}f_{gm}=1}\left\{\mathcal{E}(f_g(\mathbf{U}))+\frac{1}{2}\lambda_n\sum_{m=1}^{M_g^0}\|f_{gm}^*(\mathbf{U})\|_{\mathcal{K}}^2\right\},$$

where $\mathcal{E}(f_g(\mathbf{U})) = EV(\widetilde{\boldsymbol{Y}}_g^{0*}, f_g(\mathbf{U}))$. Besides, let

$$\widehat{f}_{gz}^{or}(\mathbf{U}) = \arg \min_{f_g \in \prod_{m=1}^{M_g^0} \overline{\mathcal{H}}_{\mathcal{K}}, \sum_{m=1}^{M_g^0} f_{gm}=1} \left\{ \mathcal{E}_z(f_g(\mathbf{U})) + \frac{1}{2}\lambda_n \sum_{m=1}^{M_g^0} \|f_{gm}^*(\mathbf{U})\|_{\mathcal{K}}^2 \right\},$$

where $\mathcal{E}_z(f_g(\mathbf{U})) = \frac{1}{n}\sum_{i=1}^n V(\widetilde{\boldsymbol{Y}}_{gi}^{0*}, f_g(\mathbf{U}_i))$. And $\widetilde{\boldsymbol{Y}}_{gi}^{0*}$ gi represents that Pseudo-labeling are given based on the real subgroup memberships. If we use the estimator $\widehat{G}_{g1}, \cdots, \widehat{G}_{g\widehat{M}_g}$ replacing unknown group membership at $M_g^0 > 2$. Then we can get

$$\widehat{f}_g(\mathbf{U}) = \arg \min_{f_g \in \prod_{m=1}^{\widehat{M}_g} \overline{\mathcal{H}}_{\mathcal{K}}, \sum_{m=1}^{\widehat{M}_g} f_{gm}=1} \left\{ \widehat{\mathcal{E}}_z(f_g(\mathbf{U})) + \frac{1}{2}\lambda_n \sum_{m=1}^{\widehat{M}_g} \|f_{gm}^*(\mathbf{U})\|_{\mathcal{K}}^2 \right\},$$

where $\widehat{\mathcal{E}}_z(f_g(\mathbf{U})) = \frac{1}{n}\sum_{i=1}^n V(\widetilde{\boldsymbol{Y}}_{gi}^*, f_g(\mathbf{U}_i))$ and $\widehat{\mathcal{E}}(f_g) = E(\widetilde{\boldsymbol{Y}}_g^*, f_g(\mathbf{U}))$. $\widetilde{\boldsymbol{Y}}_{gi}^*$ represents that Pseudo-labeling are given based on the subgroup memberships by the first step. In which

$$\widehat{f}_{gz}(\mathbf{U}) = \arg \min_{f_{gz} \in \prod_{m=1}^{\widehat{M}_g} Ar\mathcal{H}_{\mathcal{K}}, \sum_{m=1}^{\widehat{M}_g} f_{gm}=1} \left\{ \widehat{\mathcal{E}}(f_g(\mathbf{U})) + \frac{1}{2}\lambda_n \sum_{m=1}^{\widehat{M}_g} \|f_{gm}^*(\mathbf{U})\|_{\mathcal{K}}^2 \right\},$$

In order to make $\widehat{f}_g(\mathbf{U})$ consistent with the notation in the paper, our definition of $\widehat{f}_{gz}(\mathbf{U})$ violates common conventions. Note for simplicity, we let $\widehat{f}_{gz}^{or}(\mathbf{U}), \widehat{f}_{gc}^{or}(\mathbf{U}), \widehat{f}_g^{or}(\mathbf{U}), \widehat{f}_{gz}(\mathbf{U}), \widehat{f}_{gc}(\mathbf{U}), f_g(\mathbf{U})$ abbreviated as $\widehat{f}_{gz}^{or}, \widehat{f}_{gc}^{or}, f_g^{or}, \widehat{f}_{gz}, \widehat{f}_{gc}, \widehat{f}_g$. The classifier rule $\widehat{\phi}_g(\mathbf{U}) = \arg\max_m \widehat{f}_{gm}(\mathbf{U})$ and $\widehat{\phi}_g^{or}(\mathbf{U}) = \arg\max_m \widehat{f}_{gz,m}^{or}(\mathbf{U})$.

$$\mathcal{R}(\widehat{\phi}_g(\mathbf{U})) - \mathcal{R}(\widehat{f}_{gc}^{or}(\mathbf{U})) = \left(\mathcal{R}(\widehat{\phi}_g(\mathbf{U})) - \mathcal{R}(\widehat{\phi}_g^{or}(\mathbf{U}))\right) + \left(\mathcal{R}(\widehat{\phi}_g^{or}(\mathbf{U})) - \mathcal{R}(\widehat{f}_{gc}^{or}(\mathbf{U}))\right). \quad \text{(A.34)}$$

By Corollary 1 (1), the subgroup memberships can be recpvered with a high probability. That is, the first term pr $\left\{\mathcal{R}(\widehat{\phi}_g(\mathbf{U})) - \mathcal{R}(\widehat{\phi}_g^{or}(\mathbf{U})) > \epsilon\right\} \to 0$. By the Example 3 of Tewari and Bartlett (2007), we can easily get the second term pr $\left\{\mathcal{R}(\widehat{\phi}_g^{or}(\mathbf{U})) - \mathcal{R}(\widehat{f}_{gc}^{or}(\mathbf{U})) > \epsilon\right\} \to 0$. □

## References

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Biometrics* **60,** 845–853.

Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines: Theory and

application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association* **99,** 67–81.

Li, J., Li, Y., Jin, B., and Kosorok, M. R. (2021). Multithreshold change plane model: Estimation theory and applications in subgroup identification. *Statistics in Medicine* **40,** 3440–3459.

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

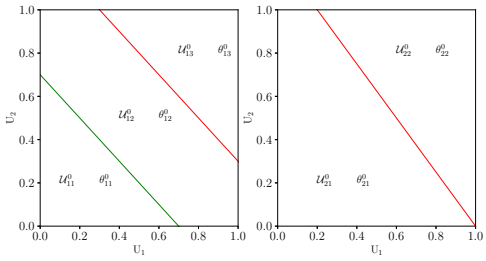
Ma, S., Huang, J., Zhang, Z., and Liu, M. (2020). Exploration of heterogeneous treatment effects via concave fusion. *The International Journal of Biostatistics* **16,** 20180026.

Madrid Padilla, O. H., Sharpnack, J., Chen, Y., and Witten, D. M. (2020). Adaptive nonparametric regression with the $k$-nearest neighbour fused lasso. *Biometrika* **107,** 293–310.

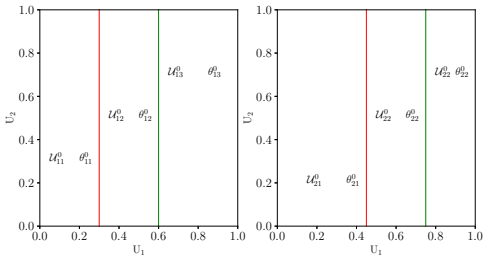
Tewari, A. and Bartlett, P. L. (2007). On the consistency of multiclass classification methods. *Journal of Machine Learning Research* **8,** 1007–1025.

Wu, Q. and Zhou, D.-X. (2006). Analysis of support vector machine classification. *Journal of Computational Analysis and Applications* **8,** 99–119.

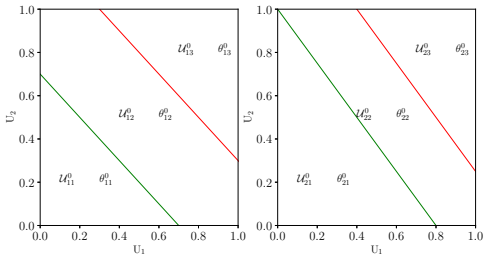
Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics* **32,** 56–85.



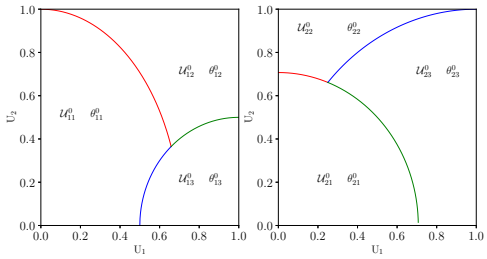
**(a)** Scenario $V$



**(b)** Scenario $VI$



**(c)** Scenario $VII$



**(d)** Scenario $VIII$

**Figure A.1:** Simple examples of group structure when $r = 2$, $p = 2$: $\{\mathcal{U}^0_{gm}, m = 1, \ldots, M^0\}$ is a partition and $\theta^0_{gm}$ is the corresponding subgroup coefficient. the solid lines are the partition boundaries. In (a) Scenario $V$: left: red line $U_1 + U_2 = 1.3$, green line $U_1 + U_2 = 0.7$, right: red line $U_1 + 0.8U_2 = 1$; In (b) Scenario $VI$: left: red line $U_1 = 0.3$, green line $U_1 = 0.6$, right: red line $U_1 = 0.45$, green line $U_1 = 0.75$; In (c) Scenario $VII$: left: red line $U_1 + U_2 = 1.3$, green line $U_1 + U_2 = 0.7$, right: red line $U_1 + 0.8U_2 = 1.2$, green line $U_1 + 0.8U_2 = 0.8$; In (d) Scenario $VIII$: left: red line $U_1^2 + 0.5U_2^2 = 0.5$, blue and green line $(U_1 - 1)^2 + U_2^2 = 0.25$, right: red and green line $U_1^2 + U_2^2 = 0.5$, blue line $(U_1 - 1)^2 + U_2^2 = 1$.

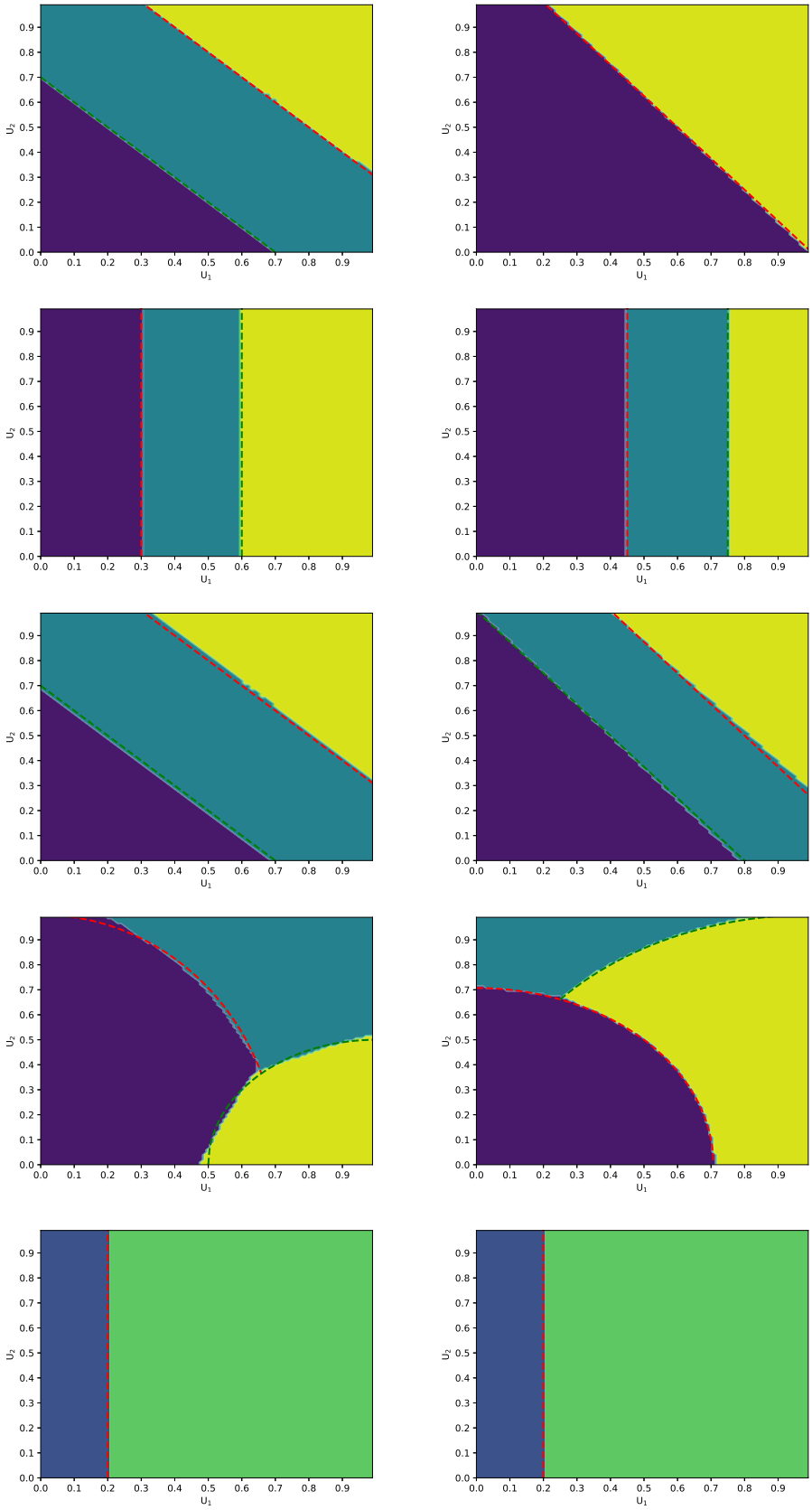



**Figure A.2:** The partition recovery results based on $n = 400$ for Case 6-10. The first column are graphs for $[\alpha]_1$ in case 6-10 while second row are for $[\alpha]_2$.

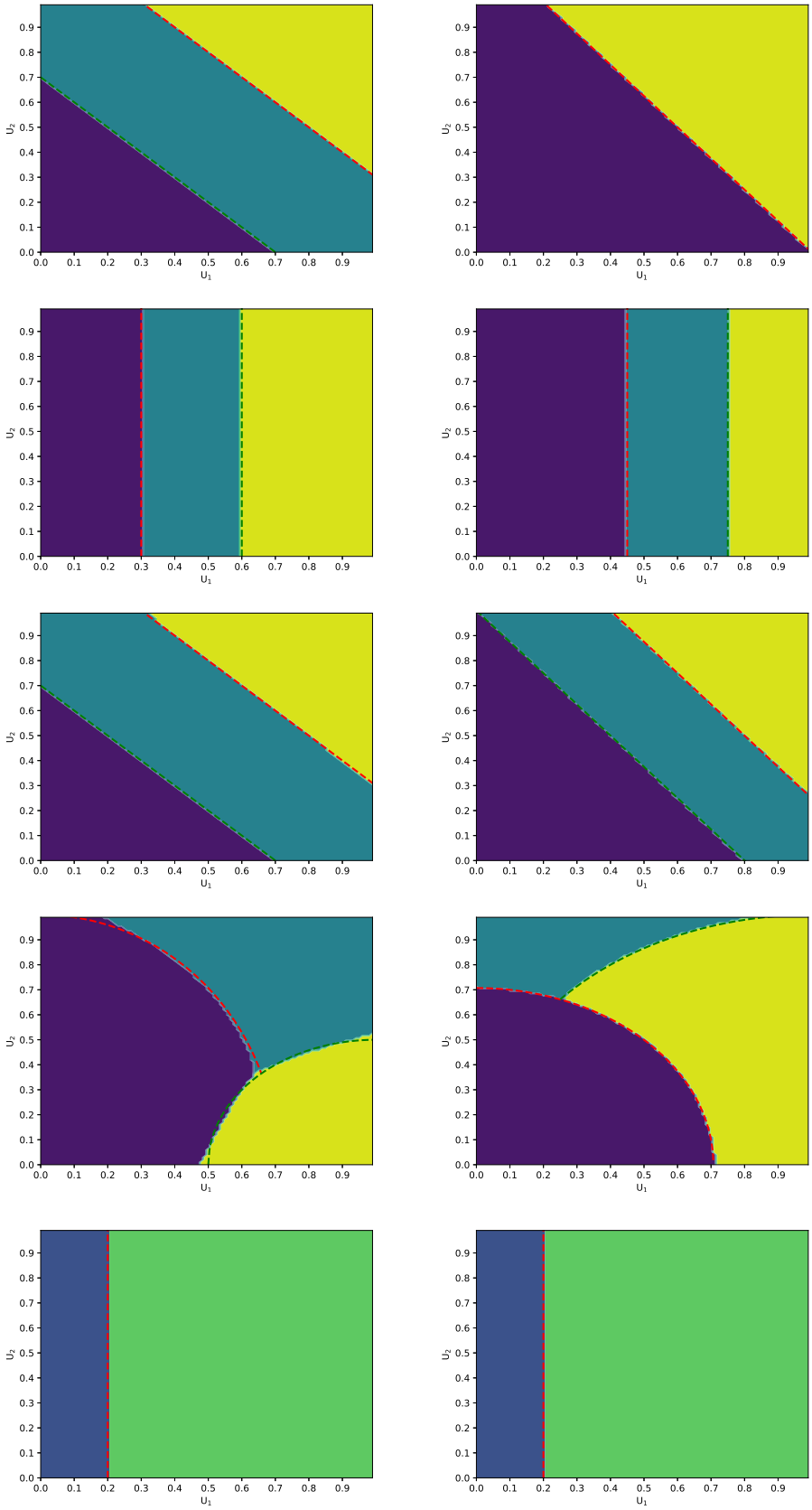



**Figure A.3:** The partition recovery results based on $n = 800$ for Case 6-10. The first column are graphs for $[\alpha]_1$ in case 6-10 while second row are for $[\alpha]_2$.

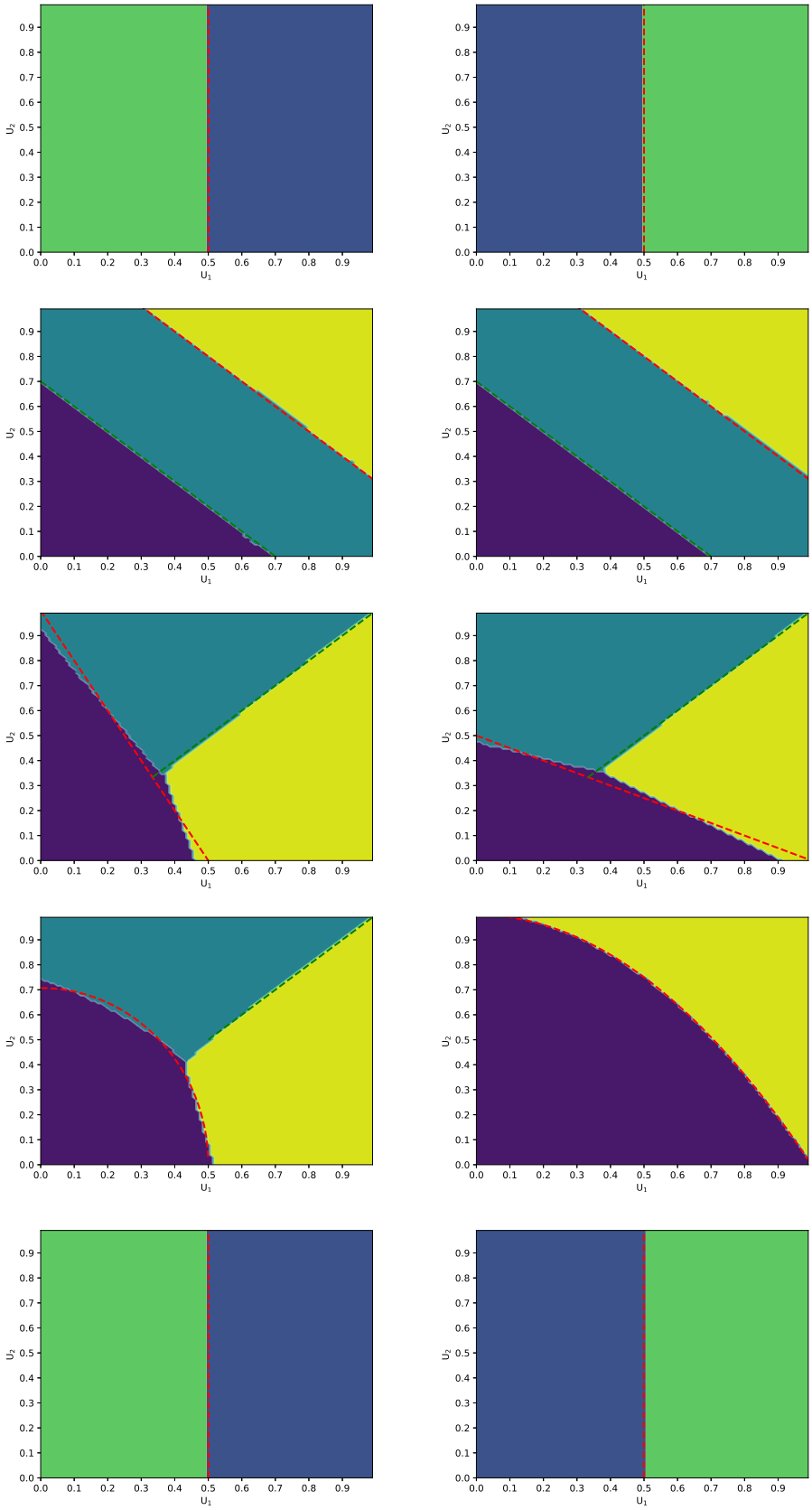



**Figure A.4:** The partition recovery results based on $n = 800$ for Case 1-5. The first column are graphs for $[\alpha]_1$ in case 1-5 while second row are for $[\alpha]_2$.

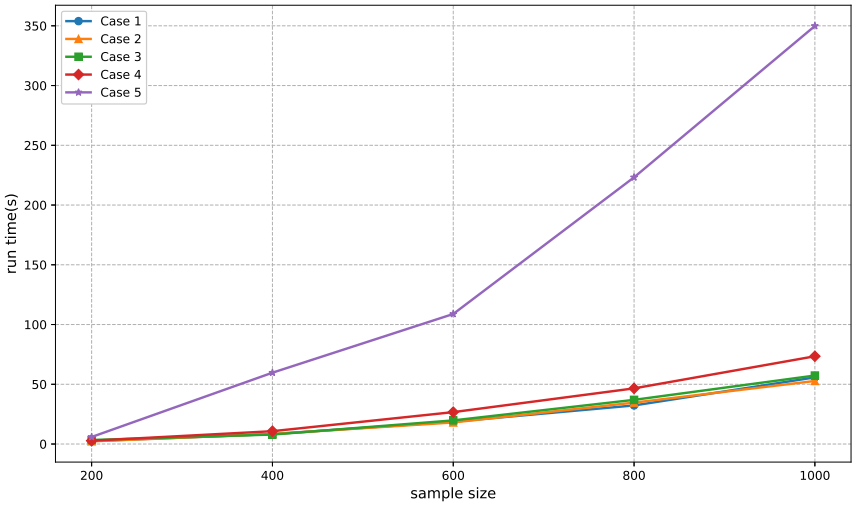
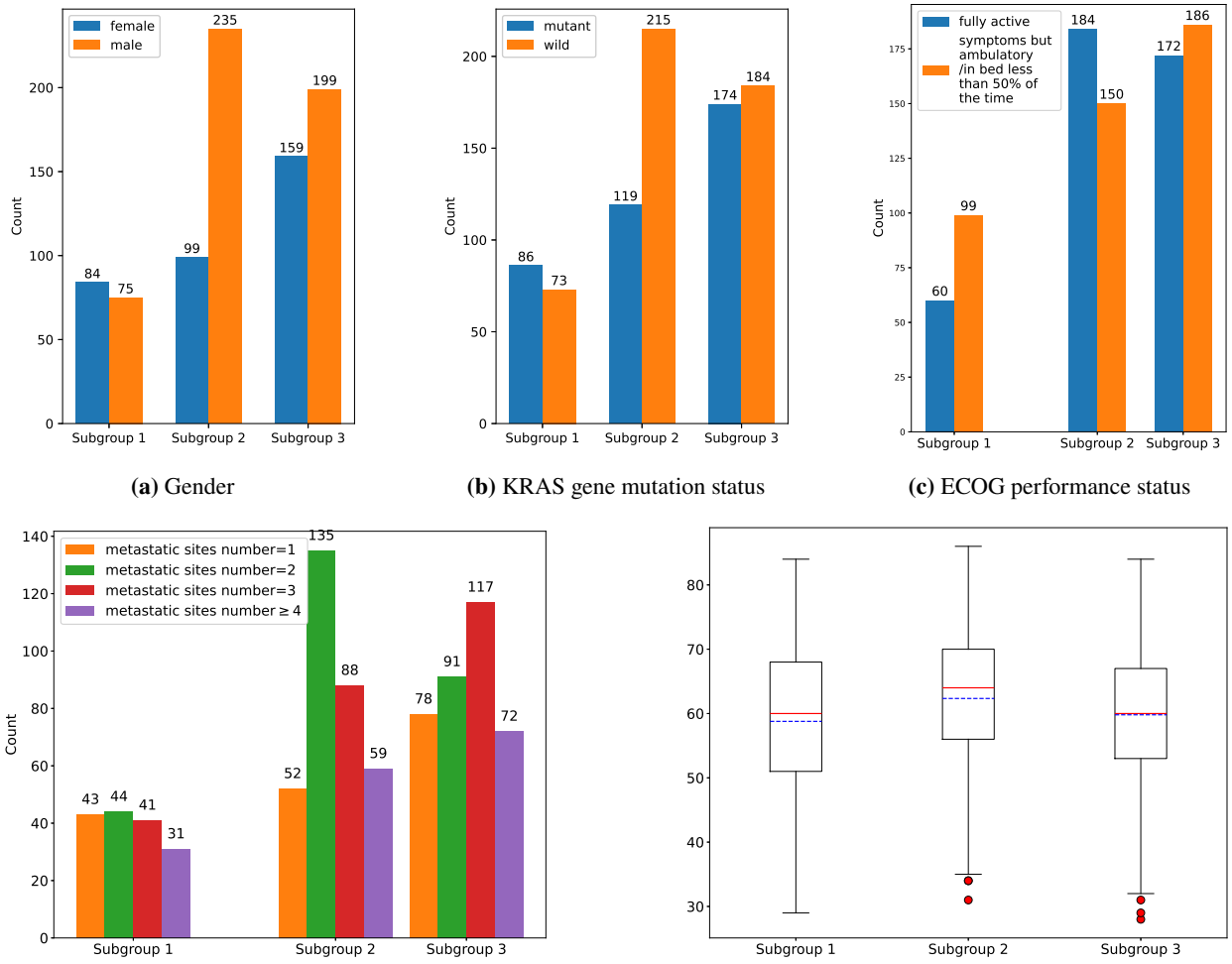



**Figure A.5:** The average computational time across 50 Monte Carlo repetitions under different sample sizes for Cases 1 to 5.




**(a)** Gender **(b)** KRAS gene mutation status **(c)** ECOG performance status

**(d)** Baseline metastatic site number **(e)** Age

**Figure A.6:** The counts of discrete threshold variables across subgroups. (a) Gender, (b) KRAS gene mutation status, (c) ECOG performance status and (d) Baseline metastatic site number. (e) Box Plots of Baseline Age Across Subgroups, solid line: mean; dashed line: median.

**Table A.1:** The sample median,bias and standard deviation (s.d.) of $\widehat{M}_1$, $\widehat{M}_2$, the Randex Index(RI) value and the percentage (per) of $\widehat{M}_1$, $\widehat{M}_2$ equaling to the true number of subgroups in case 6-9;the average accuracy (ACC) of the classifier's based on test sample $n' = 100$ in Case 6-9.

| | | n = 400 | | | | | | | n = 800 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Case | | median | bias | s.d. | RI | per | $ACC_g$ | $ACC_g^p$ | median | bias | s.d. | RI | per | $ACC_g$ | $ACC_g^p$ |
| 6 | $\widehat{M}_1$ | 3.000 | 0.000 | 0.000 | 0.958 | 1.000 | 0.968 | 0.966 | 3.000 | 0.000 | 0.000 | 0.970 | 1.000 | 0.976 | 0.969 |
| | $\widehat{M}_2$ | 2.000 | -0.005 | 0.071 | 0.969 | 0.995 | 0.989 | 0.986 | 3.000 | 0.000 | 0.000 | 0.987 | 1.000 | 0.993 | 0.981 |
| 7 | $\widehat{M}_1$ | 3.000 | 0.020 | 0.140 | 0.973 | 0.980 | 0.989 | 0.976 | 3.000 | 0.000 | 0.000 | 0.987 | 1.000 | 0.992 | 0.986 |
| | $\widehat{M}_2$ | 3.000 | 0.015 | 0.122 | 0.981 | 0.985 | 0.981 | 0.978 | 3.000 | 0.000 | 0.000 | 0.988 | 1.000 | 0.989 | 0.988 |
| 8 | $\widehat{M}_1$ | 3.000 | -0.005 | 0.071 | 0.952 | 0.995 | 0.964 | 0.961 | 3.000 | 0.000 | 0.000 | 0.965 | 1.000 | 0.976 | 0.975 |
| | $\widehat{M}_2$ | 3.000 | -0.005 | 0.071 | 0.972 | 0.995 | 0.972 | 0.968 | 3.000 | 0.000 | 0.000 | 0.943 | 1.000 | 0.987 | 0.985 |
| 9 | $\widehat{M}_1$ | 3.000 | -0.005 | 0.071 | 0.958 | 0.995 | 0.973 | 0.963 | 3.000 | -0.005 | 0.071 | 0.968 | 0.995 | 0.979 | 0.977 |
| | $\widehat{M}_2$ | 3.000 | -0.040 | 0.184 | 0.973 | 0.960 | 0.985 | 0.978 | 3.000 | -0.035 | 0.184 | 0.972 | 0.965 | 0.989 | 0.978 |
| 10 | $\widehat{M}_1$ | 2.000 | 0.030 | 0.198 | 0.952 | 0.960 | 0.983 | 0.980 | 2.000 | 0.020 | 0.140 | 0.953 | 0.980 | 0.984 | 0.991 |
| | $\widehat{M}_2$ | 2.000 | 0.010 | 0.100 | 0.978 | 0.990 | 0.991 | 0.989 | 2.000 | 0.000 | 0.000 | 0.979 | 1.000 | 0.994 | 0.992 |

**Table A.2:** The bias, empirical standard deviation (ESD), mean and squared error (MSE) of the post-group-recovery estimator in case 6-10.

| | | n = 400 | | | n = 800 | | |
|---|---|---|---|---|---|---|---|
| Case | | Bias | ESD | MSE | bias | ESD | MSE |
| 6 | $\widehat{\beta}_0^{post}$ | -0.0054 | 0.0685 | 0.0047 | 0.0018 | 0.0468 | 0.0022 |
| | $\widehat{\beta}_1^{post}$ | -0.0022 | 0.0319 | 0.0010 | -0.0011 | 0.0204 | 0.0004 |
| | $\widehat{\beta}_2^{post}$ | -0.0009 | 0.0340 | 0.0012 | 0.0020 | 0.0202 | 0.0004 |
| | $\widehat{\theta}_{11}^{post}$ | 0.0123 | 0.0455 | 0.0022 | 0.0081 | 0.0319 | 0.0011 |
| | $\widehat{\theta}_{12}^{post}$ | 0.0048 | 0.0409 | 0.0017 | 0.0009 | 0.0256 | 0.0007 |
| | $\widehat{\theta}_{13}^{post}$ | -0.0188 | 0.0541 | 0.0033 | -0.0110 | 0.0346 | 0.0013 |
| | $\widehat{\theta}_{21}^{post}$ | -0.0043 | 0.0315 | 0.0010 | -0.0027 | 0.0213 | 0.0005 |
| | $\widehat{\theta}_{22}^{post}$ | 0.0094 | 0.0392 | 0.0016 | 0.0042 | 0.0240 | 0.0006 |
| 7 | $\widehat{\beta}_0^{post}$ | 9e-05 | 0.0507 | 0.0026 | -0.0015 | 0.0352 | 0.0012 |
| | $\widehat{\beta}_1^{post}$ | -0.0013 | 0.0261 | 0.0007 | 0.0009 | 0.0190 | 0.0004 |
| | $\widehat{\beta}_2^{post}$ | -0.0008 | 0.0300 | 0.0009 | -0.0034 | 0.0201 | 0.0004 |
| | $\widehat{\theta}_{11}^{post}$ | 0.0219 | 0.1016 | 0.0108 | 0.0161 | 0.0586 | 0.0037 |
| | $\widehat{\theta}_{12}^{post}$ | 0.0091 | 0.1010 | 0.0103 | 0.0027 | 0.0663 | 0.0044 |
| | $\widehat{\theta}_{13}^{post}$ | -0.0155 | 0.0804 | 0.0067 | 0.0024 | 0.0547 | 0.0030 |
| | $\widehat{\theta}_{21}^{post}$ | 0.0036 | 0.0735 | 0.0054 | 0.0030 | 0.0500 | 0.0025 |
| | $\widehat{\theta}_{22}^{post}$ | 0.0206 | 0.0925 | 0.0090 | 0.0025 | 0.0658 | 0.0043 |
| | $\widehat{\theta}_{23}^{post}$ | -0.0100 | 0.0944 | 0.0090 | -0.0025 | 0.0603 | 0.0036 |
| 8 | $\widehat{\beta}_0^{post}$ | -0.0050 | 0.0659 | 0.0044 | 0.0018 | 0.0505 | 0.0026 |
| | $\widehat{\beta}_1^{post}$ | -0.0035 | 0.0326 | 0.0011 | -0.0012 | 0.0213 | 0.0005 |
| | $\widehat{\beta}_2^{post}$ | 0.0011 | 0.0346 | 0.0012 | 0.0020 | 0.0206 | 0.0004 |
| | $\widehat{\theta}_{11}^{post}$ | 0.0116 | 0.0483 | 0.0025 | 0.0111 | 0.0356 | 0.0014 |
| | $\widehat{\theta}_{12}^{post}$ | -0.0019 | 0.0445 | 0.0020 | -0.0042 | 0.0267 | 0.0007 |
| | $\widehat{\theta}_{13}^{post}$ | -0.0036 | 0.0576 | 0.0033 | -0.0054 | 0.0436 | 0.0019 |
| | $\widehat{\theta}_{21}^{post}$ | -0.0039 | 0.0361 | 0.0013 | -0.0047 | 0.0252 | 0.0007 |
| | $\widehat{\theta}_{22}^{post}$ | 0.0048 | 0.0347 | 0.0013 | 0.0048 | 0.0238 | 0.0006 |
| | $\widehat{\theta}_{23}^{post}$ | 0.0021 | 0.0472 | 0.0022 | -0.0006 | 0.0297 | 0.0009 |
| 9 | $\widehat{\beta}_0^{post}$ | -0.0462 | 0.1310 | 0.0193 | -0.0280 | 0.0968 | 0.0102 |
| | $\widehat{\beta}_1^{post}$ | 0.0639 | 0.1638 | 0.0309 | 0.0522 | 0.1167 | 0.0163 |
| | $\widehat{\beta}_2^{post}$ | 0.0277 | 0.1610 | 0.0267 | 0.0056 | 0.1226 | 0.0151 |
| | $\widehat{\theta}_{11}^{post}$ | 0.0039 | 0.0419 | 0.0018 | 0.0034 | 0.0305 | 0.0009 |
| | $\widehat{\theta}_{12}^{post}$ | -0.0148 | 0.0546 | 0.0032 | -0.0085 | 0.0423 | 0.0019 |
| | $\widehat{\theta}_{13}^{post}$ | 0.0027 | 0.0710 | 0.0050 | -0.0036 | 0.0435 | 0.0019 |
| | $\widehat{\theta}_{21}^{post}$ | 0.0035 | 0.0396 | 0.0016 | 0.0021 | 0.0274 | 0.0008 |
| | $\widehat{\theta}_{22}^{post}$ | 0.0076 | 0.0603 | 0.0037 | 0.0115 | 0.0409 | 0.0018 |
| | $\widehat{\theta}_{23}^{post}$ | -0.0055 | 0.0379 | 0.0015 | -0.0051 | 0.0262 | 0.0007 |
| 10 | $\widehat{\beta}_0^{post}$ | -0.0169 | 0.0995 | 0.0102 | -0.0205 | 0.0630 | 0.0044 |
| | $\widehat{\beta}_1^{post}$ | 0.0231 | 0.1314 | 0.0178 | 0.0194 | 0.0871 | 0.0080 |
| | $\widehat{\beta}_2^{post}$ | -0.0001 | 0.0263 | 0.0007 | 0.0010 | 0.0178 | 0.0003 |
| | $\widehat{\theta}_{11}^{post}$ | 0.0204 | 0.1638 | 0.0272 | 0.0258 | 0.1077 | 0.0123 |
| | $\widehat{\theta}_{12}^{post}$ | -0.0013 | 0.0936 | 0.0088 | 0.0067 | 0.0600 | 0.0036 |
| | $\widehat{\theta}_{21}^{post}$ | -0.0019 | 0.0380 | 0.0014 | -0.0034 | 0.0259 | 0.0007 |
| | $\widehat{\theta}_{22}^{post}$ | 0.0001 | 0.0371 | 0.0014 | -0.0012 | 0.0241 | 0.0006 |

**Table A.3:** The bias, empirical standard deviation (ESD), mean and squared error (MSE) of the post-group-recovery estimator in case 5.

| | Case 1 | | | Case 2 | | |
|---|---|---|---|---|---|---|
| | per | ACC | ACC$^P$ | per | ACC | ACC$^P$ |
| Li et al. (2021) | 0.9600 | 0.9983 | 0.9973 | 0.9050 | 0.9965 | 0.9934 |
| Our method | 0.9920 | 0.9871 | 0.9856 | 1.0000 | 0.9645 | 0.9653 |