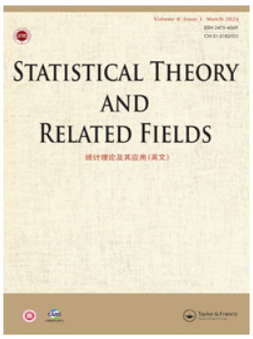

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tstf20

Variable selection and subgroup analysis for high-dimensional censored data

Yu Zhang, Jiangli Wang & Weiping Zhang

To cite this article: Yu Zhang, Jiangli Wang & Weiping Zhang (13 Mar 2024): Variable selection and subgroup analysis for high-dimensional censored data, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2024.2327113

To link to this article: https://doi.org/10.1080/24754269.2024.2327113

Published online: 13 Mar 2024.

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2024.2327113

Variable selection and subgroup analysis for high-dimensional censored data

Yu Zhang[a], Jiangli Wang[b] and Weiping Zhang[a]

[a]Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei, People's Republic of China; [b]School of Artificial Intelligence and Big Data, Hefei University, Hefei, People's Republic of China

ABSTRACT

This paper proposes a penalized method for high-dimensional variable selection and subgroup identification in the Tobit model. Based on Olsen's [(1978). Note on the uniqueness of the maximum likelihood estimator for the Tobit model. *Econometrica: Journal of the Econometric Society*, *46*(5), 1211–1215. https://doi.org/10.2307/1911445] convex reparameterization of the Tobit negative log-likelihood, we develop an efficient algorithm for minimizing the objective function by combining the alternating direction method of multipliers (ADMM) and generalised coordinate descent (GCD). We also establish the oracle properties of our proposed estimator under some mild regularity conditions. Furthermore, extensive simulations and an empirical data study are conducted to demonstrate the performance of the proposed approach.

ARTICLE HISTORY
Received 24 July 2023
Revised 12 February 2024
Accepted 29 February 2024

KEYWORDS
Tobit model; fusion; concave penalty; oracle property

1. Introduction

Subgroup analysis has broad applicability in precision medicine, economics and sociology as there is an increasing need to distinguish homogeneous subgroups of individuals, detect the subgroup structure and model the relationships between the response variable and predictors for individuals belonging to different subgroups. Thus, vast statistical methods for subgroup analysis have been developed, such as mixture models (Everitt, 2013) and regularization methods (Ma & Huang, 2017). Mixture model methods assume that the data come from a mixture of subgroups and require the specification of an underlying distribution. Shen and He (2015) proposed a structured logistic-normal mixture model to identify subgroups. However, they often require the number of subgroups to be specified to group the parameterized models, which can often be difficult to implement in practice. In contrast, Ma and Huang (2017) developed a pairwise fusion approach using concave penalty functions, such as the smoothly clipped absolute deviation (SCAD, J. Fan & Li, 2001) penalty and the minimax concave penalty (MCP, Zhang, 2010), that automatically identifies subgroup structures and estimates subgroup-specific effects. Ma et al. (2019) considered a heterogeneous treatment effects model. Wang et al. (2019) proposed a general framework of spatial subgroup analysis method for spatial data with repeated measures.

In numerous regression problems, the dependent variables can only be observed within a restricted range. For instance, when studying the influencing factors of different family expenditures in a group, some families may spend zero on items like medical insurance. Similarly, when studying individual or collective income, negative income generated by debt cannot be included in the income calculation. These scenarios involve left-censored data, which exist commonly in economics. Therefore, it encourages us to develop models specially tailored to address such situations. Tobin (1958) developed the Tobit model to study the relationship between the annual expenditure of durable goods and household income. Due to the large number of scenarios of left-censored data in economics and social sciences, the Tobit model remained popular and it has been thoroughly studied and extended to deal with other types of censored data (Amemiya, 1984). Regarding subgroup analysis with censored data, Dagne (2016) proposed a method that simultaneously addresses left-censoring and unobserved heterogeneity within longitudinal data. Additionally, Yan et al. (2021) developed a censored linear regression model with heterogeneous treatment effects.

The advent of advanced data collection techniques has led to an increase in the prevalence of high-dimensional data in the aforementioned fields. When dealing with high-dimensional problems, J. Fan and Lv (2010) provided a comprehensive overview of variable selection approaches, which incorporate methods discussed by J. Fan and Li (2001). In the context of high-dimensional censored models, Müller and van de Geer (2016) and Zhou and Liu (2016) respectively introduced lasso penalty and adaptive lasso penalty to the least absolute deviation (LAD) estimator (Powell, 1984) for variable selection in high-dimensional censored models. Johnson (2009) and

CONTACT Weiping Zhang zwp@ustc.edu.cn Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, People's Republic of China

Soret et al. (2018) proposed lasso penalties for Buckley-James estimators in right and left censored data. Moreover, Bradic et al. (2011) provided a non-concave penalized approach in Cox proportional hazards model with non-polynomial-dimensionality. Alhamzawi (2016) and Alhamzawi (2020) developed the Bayesian method of penalty censored regression. Recently, Jacobson and Zou (2023) extended the Tobit model to high-dimensional regression. However, none of these methods focus on subgroup analysis in high-dimensional censored data.

This paper focuses on subgroup identification and variable selection for a high-dimensional Tobit model. To the best of our knowledge, there have been no discussions on subgroup analysis for high-dimensional Tobit models in the existing literature. We adopt a penalized approach to identify the subgroup structures and select covariates simultaneously. The subgroup structure is determined by penalizing pairwise differences between subject-specific effects while significant covariates are chosen based on a penalty on coefficients. To ensure the sparsity and unbiasedness of the proposed estimators, we consider two commonly used concave penalties, SCAD (J. Fan & Li, 2001) and MCP (Zhang, 2010). Due to the non-convex of the negative log-likelihood in the Tobit model (Tobin, 1958), optimization becomes challenging for high-dimensional settings. To address this issue, we employ a convex reparameterization of negative log-likelihood, building upon the idea proposed by Olsen (1978). This reparameterization enables us to solve the problem using convex optimization approaches. The computational algorithm we proposed combines the alternating direction method of multipliers (ADMM) algorithm (Boyd et al., 2011) and generalized coordinate descent (GCD) algorithm (Jacobson & Zou, 2023; Yang & Zou, 2013) using two concave penalties, such as SCAD or MCP. Furthermore, we conduct a theoretical analysis of the proposed estimators and establish their oracle properties under mild conditions.

The remainder of this paper is organized as follows. Section 2 introduces the main problem and outlines the proposed method. In Section 3, we propose an algorithm for identifying the subgroup structures and performing variable selection. We state technical assumptions and establish the theoretical properties of our proposed approach in Section 4. Section 5 provides extensive simulation studies to illustrate the empirical performance of the proposed method, while Section 6 presents its application to empirical data. A summary and prospects for future research are presented in Section 7 and all technical proofs are given in Appendix.

2. Model and method

2.1. Model setting

Suppose that $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is a p-dimensional vector of covariates for the ith subject. $y_i \geq c$ is the response for a restricted range, where c is a known constant. Without loss of generality, we assume that $c = 0$ in the following. The Tobit model assumes that the observed data y satisfies $y = \max(y^*, c)$, where y^* is a latent variable. Under the homogeneous case, the classical linear model takes the form

$$y_i^* = \mu + \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where μ is the unknown intercept, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the vector of coefficients for the covariates $\mathbf{x}_i$, and ϵ_i are assumed to be independent and identically distributed with normal distribution $N(0, \sigma^2)$.

If individuals are from different groups with a unique intercept μ_i, the homogeneity assumption in the model (1) is invalid. To model the subject-specific effects, we consider the subject-specific linear model

$$y_i^* = \mu_i + \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n. \tag{2}$$

We assume $(y_1^*, \ldots, y_n^*)$ arise from K different groups with $K \geq 1$ unknown and the subjects from the same groups have the same intercept. In other words, we have $\mu_i = \pi_k$ for all $i \in \mathcal{G}_k$, where π_k is the common value of intercept μ_i in subgroup $\mathcal{G}_k$ and $\mathcal{G} = (\mathcal{G}_1, \ldots, \mathcal{G}_K)$ is a mutually partition of $\{1, \ldots, n\}$. In practice, the number of subgroups K is unknown and is smaller than the sample size n.

Define $d_i = I(y_i \geq 0)$, where $I(\cdot)$ is an indicator function. Then the observed data $(y_1, \ldots, y_n)$ satisfy the following Tobit model

$$y_i = d_i y_i^* = \begin{cases} \mu_i + \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i, & \text{if } y_i^* \geq 0, \\ 0, & \text{if } y_i^* < 0, \end{cases} \quad i = 1, \ldots, n, \tag{3}$$

with subgroup structure

$$\mu_i = \begin{cases} \pi_1, & \text{if } i \in \mathcal{G}_1, \\ \pi_2, & \text{if } i \in \mathcal{G}_2, \\ \vdots & \vdots \\ \pi_K, & \text{if } i \in \mathcal{G}_K. \end{cases} \tag{4}$$

Let $\Phi(\cdot)$ denote the standard normal cumulative distribution function (CDF). Then

$$P(y_i^* \le 0) = P(\mu_i + \mathbf{x}_i^\top \boldsymbol{\beta} < 0) = \Phi\left(-\frac{\mu_i + \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma}\right),$$

and the Tobit likelihood is given by

$$L_n(\boldsymbol{\mu}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2\right\}\right]^{d_i} \left[\Phi\left(\frac{-\mu_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma}\right)\right]^{1-d_i},$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$. Therefore, after omitting an inconsequential constant, the log-likelihood function of the Tobit model is

$$\log L_n(\boldsymbol{\mu}, \boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{n} \left(d_i \left[-\frac{1}{2}(y_i - \mu_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2/\sigma^2 - \log(\sigma)\right] + (1 - d_i) \log\left[\Phi\left(-\frac{\mu_i + \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma}\right)\right]\right).$$

It is apparent that the function $\log L_n(\boldsymbol{\mu}, \boldsymbol{\beta}, \sigma^2)$ is non-concave with respect to the parameters $(\boldsymbol{\mu}, \boldsymbol{\beta}, \sigma^2)$. By adopting the reparameterization suggested in the works of Olsen (1978) and Jacobson and Zou (2023) with $\boldsymbol{\delta} = \boldsymbol{\beta}/\sigma$, $\alpha_i = \mu_i/\sigma$ and $\gamma^2 = \sigma^{-2}$, we achieve a transformation that leads to a concave function with respect to parameters $(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma)$,

$$\log L_n(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma) = \sum_{i=1}^{n} \left\{ d_i \left[\log(\gamma) - \frac{1}{2}(\gamma y_i - \alpha_i - \mathbf{x}_i^\top \boldsymbol{\delta})^2\right] + (1 - d_i) \log\left(\Phi(-\alpha_i - \mathbf{x}_i^\top \boldsymbol{\delta})\right)\right\},$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\top$.

2.2. Method

There are usually some redundant covariates in high-dimensional scenarios, and regularization is the most commonly used method to identify the sparsity of regression coefficient vectors (Bondell & Reich, 2008; J. Fan & Lv, 2010; Y. Fan & Tang, 2013). The subgroup structure (4) can be transformed as the fusion sparse structure, $\alpha_i - \alpha_j = (\mu_i - \mu_j)/\sigma = 0$ $(i, j \in \mathcal{G}_k,\ k = 1, \ldots, K)$. In order to estimate the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and γ, and to select proper covariates through the sparsity assumption of $\boldsymbol{\delta}$, we propose a new method that combines ideas of penalized Tobit regression (Jacobson & Zou, 2023) and the subgroup analysis by concave pairwise fusion penalization (Ma & Huang, 2017; Ma et al., 2019), which can be expressed as minimizing the following loss function

$$Q(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma; \lambda_1, \lambda_2) = -\frac{1}{n} \log L_n(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma) + \sum_{i=1}^{p} P_{\lambda_1}(|\delta_i|) + \sum_{i<j} P_{\lambda_2}(|\alpha_i - \alpha_j|), \tag{5}$$

where $P_{\lambda_1}(\cdot)$ and $P_{\lambda_2}(\cdot)$ are penalty functions, $\lambda_1, \lambda_2 \ge 0$ are tuning parameters that control the strengths of regularization of $|\delta_i|$ and $|\alpha_i - \alpha_j|$, respectively. Note that the sparsity of $\boldsymbol{\delta}$ is achieved through the first penalty term, while the homogeneity detection is achieved by the second penalty term. Additionally, when $\lambda_1 = 0$ the problem reduces to the subgroup analysis in censored regression; when $\lambda_2 = 0$ the problem reduces to the penalized Tobit regression.

It is important to note that lasso estimators may fail to achieve consistent model selection unless a stringent 'irrepresentable condition' (Zhao & Yu, 2006; Zou, 2006). Particularly, the L_1 penalty tends to overshrink non-zero difference of $|\alpha_i - \alpha_j|$, which can result in an inflated number of subgroups. To address this limitation, we consider two common concave penalty functions for the purposes of identifying the subgroup structure and selecting variables, namely the smoothly clipped absolute deviation (SCAD, J. Fan & Li, 2001) and the minimax concave penalty (MCP, Zhang, 2010). These penalties provide alternative approaches to handle the challenges associated with variable selection and subgroup detection.

The SCAD is defined as follows

$$P_\lambda(t) = \lambda \int_0^t I(x \le \lambda) + \frac{(a\lambda - x)_+}{(a-1)\lambda} I(x > \lambda)\mathrm{d}x, \quad a > 2,$$

and the MCP is

$$P_\lambda(t) = \int_0^t \frac{(a\lambda - x)_+}{a}\mathrm{d}x, \quad a > 1,$$

where a is a parameter that controls the concavity of the penalty function.

3. Computational algorithm

In this section, we propose an algorithm utilizing the alternating direction method of multipliers (ADMM) (Boyd et al., 2011) in conjunction with generalized coordinate descent (GCD) (Jacobson & Zou, 2023; Yang & Zou, 2013) to address the minimization problem (5). Since the penalty function is not separable with respect to α_i, it is challenging to directly minimize the objective function (5). We introduce a new set of parameters $\eta_{ij} = \alpha_i - \alpha_j$, and then the minimizing problem can be written as the following constraint optimization problem

$$S(\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\eta}, \gamma; \lambda_1, \lambda_2) = \ell_n(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma) + \sum_{i=1}^{p} P_{\lambda_1}(|\delta_i|) + \sum_{i<j} P_{\lambda_2}(|\eta_{ij}|),$$
$$\text{subject to } \alpha_i - \alpha_j - \eta_{ij} = 0,$$

where $\ell_n(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma) = -\frac{1}{n}\log L_n(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma)$ is the negative log-likelihood function which we call it Tobit loss for short and $\boldsymbol{\eta} = \{\eta_{ij}, i < j\}^\top$. Applying the augmented Lagrangian method, the estimates of the parameters can be obtained by minimizing

$$L(\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{\varphi}, \gamma; \lambda_1, \lambda_2, \rho) = S(\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\eta}, \gamma; \lambda_1, \lambda_2) + \sum_{i<j} \varphi_{ij}(\alpha_i - \alpha_j - \eta_{ij}) + \frac{\rho}{2}\sum_{i<j}(\alpha_i - \alpha_j - \eta_{ij})^2, \tag{6}$$

where $\boldsymbol{\varphi} = \{\varphi_{ij}, i < j\}^\top$ are Lagrange multipliers, and ρ is the penalty parameter.

To obtain the minimum in (6), we propose to use the following iterative algorithm based on the ADMM. Let t denote the iteration step. We update the estimates of $(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma)$, $\boldsymbol{\eta}$, and $\boldsymbol{\varphi}$ iteratively at the $(t+1)$th iteration step as follows

$$(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \gamma^{(t+1)}) = \arg\min_{\boldsymbol{\delta}, \boldsymbol{\alpha}, \gamma} L(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma, \boldsymbol{\eta}^{(t)}, \boldsymbol{\varphi}^{(t)}), \tag{7}$$

$$\boldsymbol{\eta}^{(t+1)} = \arg\min_{\boldsymbol{\eta}} L(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \gamma^{(t+1)}, \boldsymbol{\eta}, \boldsymbol{\varphi}^{(t)}), \tag{8}$$

$$\boldsymbol{\varphi}^{(t+1)} = \boldsymbol{\varphi}^{(t)} + \rho(\Delta\boldsymbol{\alpha}^{(t+1)} - \boldsymbol{\eta}^{(t+1)}), \tag{9}$$

where $\Delta = \{(e_i - e_j), i < j\}^\top$.

It is worth noting that, given $(\boldsymbol{\eta}^{(t)}, \boldsymbol{\varphi}^{(t)})$, the objective function in the first minimization problem (7) can be simplified as

$$L(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma, \boldsymbol{\eta}, \boldsymbol{\varphi}) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\alpha_i, \boldsymbol{\delta}, \gamma) + \sum_{i=1}^{p} P_{\lambda_1}(|\delta_i|) + \sum_{i<j} \varphi_{ij}(\alpha_i - \alpha_j - \eta_{ij}) + \frac{\rho}{2}\sum_{i<j}(\alpha_i - \alpha_j - \eta_{ij})^2 + C, \tag{10}$$

where $\ell_i(\alpha_i, \boldsymbol{\delta}, \gamma) = \frac{1}{2}d_i(\gamma y_i - \alpha_i - \mathbf{x}_i^\top\boldsymbol{\delta})^2 - (1 - d_i)\log\Phi(-\mathbf{x}_i^\top\boldsymbol{\delta} - \alpha_i))$, and C is a constant independent with $(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma)$.

Due to the complexity of the function (10) with respect to $(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma)$, we apply the generalized coordinate descent (GCD) method to solve the problem. By employing the GCD method, we can iteratively update each variable while holding the others fixed.

Let $\boldsymbol{\alpha}', \boldsymbol{\delta}'$ and γ' be the current values for $\boldsymbol{\alpha}, \boldsymbol{\delta}$ and γ, respectively. For the sake of simplicity in notation, let $\mathbf{v}_{(-j)}$ denote the vector $\mathbf{v}$ with the jth element removed in the subsequent context. In order to get the estimate of α_i, we begin by expressing the Tobit loss ℓ_n with respect to α_i

$$\ell_n(\alpha_i \mid \boldsymbol{\alpha}'_{(-i)}, \boldsymbol{\delta}', \gamma') = \frac{1}{n}\left\{\frac{1}{2}d_i(\gamma y_i - \alpha_i - \mathbf{x}_i^\top \boldsymbol{\delta}')^2 - (1-d_i)\log\Phi(-\mathbf{x}_i^\top \boldsymbol{\delta}' - \alpha_i)\right\}.$$

We observe that after dropping the negligible constants, the Tobit loss ℓ_n associated with α_i solely depends on the data collected from subject i. In line with Theorem 1 presented in Jacobson and Zou (2023), the quadratic majorization function for $\ell_n(\alpha_i \mid \boldsymbol{\alpha}', \boldsymbol{\delta}', \gamma')$ takes the following form:

$$Q_\alpha(\alpha_i \mid \boldsymbol{\alpha}'_{(-i)}, \boldsymbol{\delta}', \gamma') = \ell_n(\alpha'_i \mid \boldsymbol{\delta}', \gamma') + \dot{\ell}_n(\alpha'_i \mid \boldsymbol{\delta}', \gamma')(\alpha_i - \alpha'_i) + \frac{1}{2}(\alpha_i - \alpha'_i)^2,$$

where $\dot{\ell}_n(\alpha'_i \mid \boldsymbol{\delta}', \gamma')$ represents the derivative of function with respect to α_i. Following the MM principle, the update of α_i can be obtained by minimizing the expression $Q_\alpha(\alpha_i \mid \boldsymbol{\alpha}'_{(-i)}, \boldsymbol{\delta}', \gamma') + \frac{\rho}{2}\sum_{k<j}\{(e_k - e_j)^\top \boldsymbol{\alpha} - \eta_{kj} + \rho^{-1}\varphi_{kj}\}^2$, and e_j is an $n \times 1$ vector with the jth element being 1 and the remaining elements being 0. Therefore, for fixed $\boldsymbol{\alpha}^{(t,k)}, \boldsymbol{\delta}^{(t,k)}$, and $\gamma^{(t,k)}$ at the kth step, the update of $\boldsymbol{\alpha}^{(t,k+1)}$ is as follows

$$\boldsymbol{\alpha}^{(t,k+1)} = (I + n\rho\Delta^\top\Delta)^{-1}(n\rho\Delta^\top \boldsymbol{\eta}^{(t)} - n\Delta^\top \boldsymbol{\varphi}^{(t)} + \mathbf{a}^{(t,k)}), \tag{11}$$

where I is the identity matrix, and $\mathbf{a}^{(t,k)} = \boldsymbol{\alpha}^{(t,k)} - \dot{\ell}_n(\boldsymbol{\alpha}^{(t,k)})$ with $\dot{\ell}_n(\boldsymbol{\alpha}^{(t,k)}) = (\dot{\ell}_n(\alpha_1^{(t,k)}), \ldots, \dot{\ell}_n(\alpha_n^{(t,k)}))^\top$.

Now we consider coordinate-wise updates of δ_j, $j = 1, \ldots, p$. Let $M_j = \frac{1}{n}\sum_{i=1}^n x_{ij}^2$. Similarly to α_i, we also have the quadratic majorization function of $\ell_n(\delta_j \mid \boldsymbol{\alpha}', \boldsymbol{\delta}', \gamma')$ with respect to δ_j with form

$$Q_\delta(\delta_j \mid \boldsymbol{\alpha}', \boldsymbol{\delta}'_{-j}, \gamma') = \ell_n(\delta'_j \mid \boldsymbol{\alpha}', \boldsymbol{\delta}'_{(-j)}, \gamma') + \dot{\ell}_n(\delta'_j \mid \boldsymbol{\alpha}', \boldsymbol{\delta}'_{(-j)}, \gamma')(\delta_j - \delta'_j) + \frac{M_j}{2}(\delta_j - \delta'_j)^2,$$

where $\dot{\ell}_n(\delta'_j \mid \boldsymbol{\alpha}', \boldsymbol{\delta}'_{(-j)}, \gamma')$ is the derivative of $\ell_n(\delta_j \mid \boldsymbol{\alpha}', \boldsymbol{\delta}'_{(-j)}, \gamma')$ with respect to δ_j. Here, the Tobit loss with respect to δ_j can be expressed as

$$\begin{aligned}\ell_n(\delta_j \mid \boldsymbol{\alpha}', \boldsymbol{\delta}'_{(-j)}, \gamma') = \frac{1}{n}\sum_{i=1}^n \frac{1}{2}d_i(&\gamma y_i - \alpha'_i - \mathbf{x}_{i,(-j)}^\top \boldsymbol{\delta}'_{(-j)} - x_{ij}\delta_j)^2 \\ &- (1-d_i)\log\Phi(-\mathbf{x}_{i,(-j)})^\top \boldsymbol{\delta}'_{(-j)} - \alpha'_i - x_{ij}\delta_j)).\end{aligned}$$

We can update δ_j by minimizing $Q_\delta(\delta_j \mid \boldsymbol{\alpha}', \boldsymbol{\delta}'_{-j}, \gamma') + P_{\lambda_1}(\delta_j)$ through MM principle. For $j = 1, \ldots, p$, let $\upsilon_j^{(t,k)} = \delta_j^{(t,k)} - \frac{1}{M_j}\ell'_n(\delta_j^{(t,k)} \mid \boldsymbol{\alpha}^{(t,k+1)}, \boldsymbol{\delta}^{(t,k)}, \gamma^{(t,k)})$. Hence, for SCAD penalty with $a_1 > \max_j\{M_j^{-1}\} + 1$, the update of δ_j at the $(k+1)$th step is

$$\delta_j^{(t,k+1)} = \begin{cases} \mathrm{ST}\left(\upsilon_j^{(t,k)}, \dfrac{\lambda_1}{M_j}\right), & \text{if} \quad |\upsilon_j^{(t,k)}| \le \lambda_1 + \lambda_1/M_j, \\ \dfrac{\mathrm{ST}(\upsilon_j^{(t,k)}, a_1\lambda_1/((a_1-1)M_j))}{1 - ((a_1-1)M_j)^{-1}}, & \text{if} \quad \lambda_1 + \lambda_1/M_j < |\upsilon_j^{(t,k)}| \le a_1\lambda_1, \\ \upsilon_j^{(t,k)}, & \text{if} \quad |\upsilon_j^{(t,k)}| > a_1\lambda_1, \end{cases} \tag{12}$$

where $\mathrm{ST}(t, \lambda) = \mathrm{sign}(t)(|t| - \lambda)_+$ is the soft-thresholding rule, and $(x)_+ = \max\{x, 0\}$. And when $a_1 > \max_j\{M_j^{-1}\}$ for the MCP penalty, the updated value is

$$\delta_j^{(t,k+1)} = \begin{cases} \dfrac{\mathrm{ST}(\upsilon_j^{(t,k)}, \lambda_1/M_j)}{1 - (a_1 M_j)^{-1}}, & \text{if} \quad |\upsilon_j^{(t,k)}| \le a_1\lambda_1, \\ \upsilon_j^{(t,k)}, & \text{if} \quad |\upsilon_j^{(t,k)}| > a_1\lambda_1. \end{cases} \tag{13}$$

Lastly, for given $\boldsymbol{\alpha}^{(t,k+1)}$ and $\boldsymbol{\delta}^{(t,k+1)}$, we minimize $\ell_n(\gamma \mid \boldsymbol{\alpha}^{(t,k+1)}, \boldsymbol{\delta}^{(t,k+1)})$ to update γ,

$$\gamma^{(t,k+1)} = \frac{\sum_{i=1}^n d_i y_i(\alpha_i^{(t,k+1)} + \mathbf{x}_i^\top \boldsymbol{\delta}^{(t,k+1)}) + \sqrt{\left(\sum_{i=1}^n d_i y_i(\alpha_i^{(t,k+1)} + \mathbf{x}_i^\top \boldsymbol{\delta}^{(t,k+1)})\right)^2 + 4\left(\sum_{i=1}^n d_i y_i^2\right)\sum_{i=1}^n d_i}}{2\sum_{i=1}^n d_i y_i^2}. \tag{14}$$

Once the convergence is achieved, we denote the final iteration of the GCD as $(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \gamma^{(t+1)})$.

As for the second minimization function (8), by eliminating insignificant constants that have no effect on the minimization process, the optimization function simplifies to the following form

$$\eta_{ij} = \arg\min_{\eta_{ij}} \frac{\rho}{2}(\eta_{ij} - \zeta_{ij})^2 + P_{\lambda_2}(|\eta_{ij}|), \tag{15}$$

with respect to η_{ij}, where $\zeta_{ij} = \alpha_i - \alpha_j + \rho^{-1}\varphi_{ij}$. It's worth noting that (15) is convex with respect to each η_{ij} when $a_2 > \rho^{-1}$ for MCP or $a_2 > \rho^{-1} + 1$ for SCAD. Hence, the closed-form solution for the MCP penalty at the $(t+1)$ iteration is

$$\eta_{ij}^{(t+1)} = \begin{cases} \dfrac{\mathrm{ST}(\zeta_{ij}^{(t+1)}, \lambda_2/\rho)}{1 - (a_2\rho)^{-1}}, & \text{if} \quad |\zeta_{ij}^{(t+1)}| \le a_2\lambda_2, \\ \zeta_{ij}^{(t+1)}, & \text{if} \quad |\zeta_{ij}^{(t+1)}| > a_2\lambda_2. \end{cases} \tag{16}$$

Then for the SCAD penalty, it is

$$\eta_{ij}^{(t+1)} = \begin{cases} \mathrm{ST}(\zeta_{ij}^{(t+1)}, \lambda_2/\rho), & \text{if} \quad |\zeta_{ij}^{t+1}| \le \lambda_2 + \lambda_2/\rho, \\ \dfrac{\mathrm{ST}(\zeta_{ij}^{(t+1)}, a_2\lambda_2/((a_2-1)\rho))}{1 - ((a_2-1)\rho)^{-1}}, & \text{if} \quad \lambda_2 + \lambda_2/\rho < |\zeta_{ij}^{(t+1)}| \le a_2\lambda_2, \\ \zeta_{ij}^{(t+1)}, & \text{if} \quad |\zeta_{ij}| > a_2\lambda_2. \end{cases} \tag{17}$$

We provide the complete algorithm in Algorithm (1), referred to as the ADMM-GCD algorithm for convenience.

Algorithm 1 An ADMM-GCD algorithm for high-dimensional censored regression with heterogeneous effects

Require: Initialize $\boldsymbol{\delta}^{(0)}, \boldsymbol{\alpha}^{(0)}, \gamma^{(0)}, \boldsymbol{\eta}^{(0)} \leftarrow \boldsymbol{\alpha}_i^{(0)} - \boldsymbol{\alpha}_j^{(0)}, \boldsymbol{\varphi}^{(0)} \leftarrow \mathbf{0}$;
Given $\lambda_1, \lambda_2 > 0$ and $\epsilon_a, \epsilon_b > 0$;
for $t = 0, 1, 2, \ldots$ **do**
  $\boldsymbol{\alpha}^{(t,0)} \leftarrow \boldsymbol{\alpha}^{(t)}, \boldsymbol{\delta}^{(t,0)} \leftarrow \boldsymbol{\delta}^{(t)}$;
  for $k = 0, 1, 2, \ldots$ **do**
    Compute $\boldsymbol{\alpha}^{(t,k+1)}$ using (11);
    for $j = 1$ to p **do**
      Compute $\delta_j^{(t,k+1)}$ using (12) or (13);
    end for
    Compute $\gamma^{(t,k+1)}$ using (14);
    if $\|(\boldsymbol{\alpha}^{(t,k+1)} - \boldsymbol{\alpha}^{(t,k)})^\top, (\boldsymbol{\delta}^{(t,k+1)} - \boldsymbol{\delta}^{(t,k)})^\top, (\gamma^{(t,k+1)} - \gamma^{(t,k)})\|_2^2 \le \epsilon_a$ **then**
      Stop and denote $\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t,k+1)}$, $\boldsymbol{\delta}^{(t+1)} = (\delta_1^{(t,k+1)}, \ldots, \delta_p^{(t,k+1)})^\top$, $\gamma^{(t+1)} = \gamma^{(t,k+1)}$;
    end if
  end for
  Compute $\boldsymbol{\eta}^{(t+1)}$ using (16) or (17);
  Compute $\boldsymbol{\varphi}^{(t+1)}$ using (9);
  if $\|\boldsymbol{\Delta}\boldsymbol{\alpha}^{(t+1)} - \boldsymbol{\eta}^{(t+1)}\|_2^2 < \epsilon_b$ **then**
    Stop and get $\widehat{\boldsymbol{\mu}} = \boldsymbol{\alpha}^{(t+1)}/\gamma^{(t+1)}$, $\widehat{\boldsymbol{\beta}} = \boldsymbol{\delta}^{(t+1)}/\gamma^{(t+1)}$, $\widehat{\sigma} = 1/\gamma^{(t+1)}$;
  end if
end for

Remark: In the algorithm, there are two iterative steps. In the nested GCD iteration, the iterative convergence criterion is defined as follows

$$\left\| \left(\boldsymbol{\alpha}^{(t,k+1)} - \boldsymbol{\alpha}^{(t,k)}\right)^\top, \left(\boldsymbol{\delta}^{(t,k+1)} - \boldsymbol{\delta}^{(t,k)}\right)^\top, \left(\gamma^{(t,k+1)} - \gamma^{(t,k)}\right) \right\|_2^2 \le \epsilon_a.$$

On the other hand, for ADMM, we employ the following criterion

$$\|\mathbf{r}^{(t+1)}\|^2 = \|\boldsymbol{\Delta}\boldsymbol{\alpha}^{(t+1)} - \boldsymbol{\eta}^{(t+1)}\|_2^2 < \epsilon_b.$$

It is worth mentioning that in our simulation studies, we set $\epsilon_a = 10^{-4}$ and $\epsilon_b = 10^{-5}$, respectively.

The following Proposition 3.1 presents the convergence properties of the proposed algorithm.

Proposition 3.1: *Given* $a_1 > \max\{1/M_j\}$ *and* $a_2 > 1/\rho$ *for MCP or* $a_1 > \max\{1/M_j\} + 1$ *and* $a_2 > 1/\rho + 1$ *for SCAD, any accumulation point* $(\boldsymbol{\delta}^{(t+1)}, \boldsymbol{\alpha}^{(t+1)}, \gamma^{(t+1)}, \boldsymbol{\eta}^{(t+1)})$ *generated by Algorithm* 1 *is a coordinate-wise minimum of* $L(\boldsymbol{\delta}, \boldsymbol{\alpha}, \gamma, \boldsymbol{\eta}, \boldsymbol{\varphi}^{(t)})$. *In addition, the primal residual* $\mathbf{r}^{(t+1)} = \Delta\boldsymbol{\alpha}^{(t+1)} - \boldsymbol{\eta}^{(t+1)}$ *and the dual residual* $\mathbf{s}^{(t+1)} = \rho\boldsymbol{\Delta}^\top(\boldsymbol{\eta}^{(t+1)} - \boldsymbol{\eta}^{(t)})$ *of the ADMM satisfy that* $\lim_{t\to\infty} \|\mathbf{r}^{(t)}\|_2^2 = 0$ *and* $\lim_{t\to\infty} \|\mathbf{s}^{(t)}\|_2^2 = 0$ *for both MCP and SCAD penalties.*

4. Theoretical properties

In this section, we study the theoretical properties of the proposed estimators. We first introduce $\mathcal{M}_\mathcal{G}$, a subspace of R^n, defined as

$$\mathcal{M}_\mathcal{G} = \{\boldsymbol{\alpha} \in R^n : \alpha_i = \alpha_j, \text{ for any } i, j \in \mathcal{G}_k, 1 \le k \le K\}.$$

For each $\boldsymbol{\alpha} \in \mathcal{M}_\mathcal{G}$, it can be also written as $\boldsymbol{\alpha} = \mathbf{Z}\boldsymbol{\tau}$, where $\mathbf{Z} = \{z_{ik}\}$ is the $n \times K$ indicator matrix defined by $z_{ik} = 1$ for $i \in \mathcal{G}_k$ and $z_{ik} = 0$ otherwise, and $\boldsymbol{\tau}$ is a $K \times 1$ vector of parameters. Let $|\mathcal{G}_k|$ denotes the number of elements in $\mathcal{G}_k$, we have $\mathbf{T} = \mathbf{Z}^\top\mathbf{Z} = \text{diag}(|\mathcal{G}_1|, \ldots, |\mathcal{G}_K|)$ by matrix calculation. Define $|\mathcal{G}_{\min}| = \min_{1\le k\le K} |\mathcal{G}_k|$ and $|\mathcal{G}_{\max}| = \max_{1\le k\le K} |\mathcal{G}_k|$.

First, we assume that the true values of the parameters for $\boldsymbol{\delta}$, $\boldsymbol{\alpha}$ and γ are $\boldsymbol{\delta}^*$, $\boldsymbol{\alpha}^*$ and γ^*, respectively. Let $\boldsymbol{\tau}^* = (\tau_k^*, k = 1, \ldots, K)$, where τ_k^* is the underlying common intercept for group $\mathcal{G}_k$. We let $\mathcal{A} = \{j : \delta_j \neq 0\} \subseteq \{1, \ldots, p\}$ denote the true support set of $\boldsymbol{\delta}$ and define $\mathcal{A}' = \mathcal{A} \cup \{p+1, \ldots, p+K+1\}$, $\mathcal{A}_1 = \mathcal{A} \setminus \{p+K+1\}$ and $s = |\mathcal{A}|$. Under the sparsity assumption, $s \ll p$.

To consider the true variables, we set $\boldsymbol{\delta} = (\boldsymbol{\delta_1}^\top, \boldsymbol{\delta_0}^\top)^\top$ and define $\mathcal{M}_\mathcal{B}$, where $\boldsymbol{\delta_1}$ are the true variables with $\{\delta_j \neq 0\}$ and $\boldsymbol{\delta_0}$ are the zero variables. Then we define a subspace of R^p as

$$\mathcal{M}_\mathcal{B} = \{\boldsymbol{\delta} \in R^p : \delta_i = \delta_i \text{ for } i \in \mathcal{A} \text{ and } \delta_i = 0 \text{ for } i \in \mathcal{A}^c\}.$$

Additionally, we let $\Theta =: (\boldsymbol{\alpha}^\top, \boldsymbol{\delta}^\top, \gamma)^\top$ and $\Xi =: (\boldsymbol{\tau}^\top, \boldsymbol{\delta_1}^\top, \gamma)^\top$ for notational convenience.

Since there are obvious differences between censored and uncensored observations in the Tobit likelihood, we use a more convenient expression to differentiate them clearly. When we define n_1 as the number of observations for which $y_i > 0$ and $n_0 = n - n_1$, we can re-block our observations as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_0 \\ \mathbf{X}_1 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \end{bmatrix},$$

where $\mathbf{X}_0$ is the $n_0 \times (p+1)$ matrix of predictors corresponding to the observations for which $y_i \le 0$ while $\mathbf{X}_1$ is the $n_1 \times (p+1)$ matrix of predictors corresponding to the observations for which $y_i > 0$. Similarly, $\mathbf{y}_0$ and $\mathbf{y}_1$ denote the responses greater than and not greater than 0, respectively. Then by the definition above we get the same form

$$\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha}_0 \\ \boldsymbol{\alpha}_1 \end{bmatrix} = \mathbf{Z}\boldsymbol{\tau} = \begin{bmatrix} \mathbf{Z_0}\boldsymbol{\tau} \\ \mathbf{Z_1}\boldsymbol{\tau} \end{bmatrix}.$$

4.1. Technical conditions

In this subsection, we will introduce several mild conditions and discuss their relevance in detail.

(C1): Assume $\|\mathbf{X}_j\|_2 = \sqrt{n}$ for $1 \le j \le p$, $\|\mathbf{X}\|_\infty \le C_1 s$, $\|\boldsymbol{\delta}^*\|_\infty \le C_2\sqrt{s}$, $\|\boldsymbol{\alpha}^*\|_\infty \le C_3\sqrt{n}$, and $|\gamma^*| \le C_0$ for some constant $0 \le C_0, C_1, C_2, C_3 \le \infty$.

(C2): The penalty function $P_\lambda(t)$ is symmetric of t, and it is nondecreasing and concave for $t \in [0, \infty)$. It is a constant on $t \ge a\lambda$ for the function $\rho(t) = P_\lambda(t)/\lambda$ with $0 \le a \le \infty$ and $\rho(0) = 0$. Moreover, $\rho'(t)$ exists and is continuous except for a finite number of t and $\rho'(0+) = 1$.

(C3): The error vectors $\varepsilon_i, i = 1, \ldots, n$ are i.i.d. normal distributed with mean zero and variance σ^{*2} such that $P(|\varepsilon_i| > t) \le 2c\exp(-\frac{1}{2}c^{-2}t^2)/t$, where c is a constant.

Conditions (C2) and (C3) are widely adopted in high-dimensional settings. The penalties, including MCP and SCAD mentioned in the article, satisfy (C2).

When the true group memberships $\mathcal{G}_1, \ldots, \mathcal{G}_K$ and true support set $\mathcal{A}$ are known, the oracle estimators for $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$ and γ are defined as

$$\widehat{\Theta}^{\text{or}} = (\widehat{\boldsymbol{\alpha}}^{\text{or}}, \widehat{\boldsymbol{\delta}}^{\text{or}}, \widehat{\gamma}^{\text{or}}) = \arg \max_{\boldsymbol{\alpha} \in \mathcal{M}_{\mathcal{G}}, \boldsymbol{\delta} \in \mathcal{M}_{\mathcal{B}}, \gamma \in R} \log L_n(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma).$$

After removing the redundant variables, we can write $\widetilde{\mathbf{X}} = (\mathbf{Z}, \mathbf{X}_{\mathcal{A}})$ and $\widetilde{\boldsymbol{\delta}} = (\boldsymbol{\tau}^\top, \boldsymbol{\delta}_{\mathbf{1}}^\top)$ for ease of calculation. Then, the oracle estimators for $\boldsymbol{\tau}$, $\boldsymbol{\delta}_{\mathbf{1}}$ and γ are given by

$$\begin{aligned}
\widehat{\Xi}^{\text{or}} &= (\widehat{\boldsymbol{\tau}}^{\text{or}}, \widehat{\boldsymbol{\delta}}_1^{\text{or}}, \widehat{\gamma}^{\text{or}}) = (\widetilde{\boldsymbol{\delta}}^{\text{or}}, \widehat{\gamma}^{\text{or}}) \\
&= \arg \max_{\boldsymbol{\delta}_{\mathbf{1}} \in R^{K+s}, \gamma \in R} \log L_n(\boldsymbol{\delta}_{\mathbf{1}}, \gamma) \\
&= \arg \max_{\boldsymbol{\delta}_{\mathbf{1}} \in R^{K+s}, \gamma \in R} \sum_{i=1}^{n} d_i \left[\log(\gamma) - \frac{1}{2}(\gamma y_i - \widetilde{\mathbf{x}}_i^\top \widetilde{\boldsymbol{\delta}}_1)^2 \right] + (1 - d_i) \log \Phi \left((-\widetilde{\mathbf{x}}_i^\top \widetilde{\boldsymbol{\delta}}_1) \right).
\end{aligned} \tag{18}$$

The problem (18) can be reduced to the traditional Tobit model. To obtain the estimator using the maximum likelihood method, it is necessary to calculate the matrix of second partials. We define $g(s) = \phi(s)/\Phi(s)$ and $h(s) = g(s)(s + g(s))$. The matrix can be expressed as follows

$$\begin{aligned}
\nabla^2 \log L_n(\Xi) &= - \begin{bmatrix} \widetilde{\mathbf{X}}^\top \\ -\mathbf{y}^\top \end{bmatrix} \begin{bmatrix} \mathbf{D}(\widetilde{\boldsymbol{\delta}}_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{X}} & -\mathbf{y} \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & n_1 \gamma^{-2} \end{bmatrix} \\
&= - \begin{bmatrix} \widetilde{\mathbf{X}}_0^\top \mathbf{D}(\widetilde{\boldsymbol{\delta}}_1) \widetilde{\mathbf{X}}_0 + \widetilde{\mathbf{X}}_1^\top \widetilde{\mathbf{X}}_1 & -\widetilde{\mathbf{X}}_0^\top \mathbf{D}(\widetilde{\boldsymbol{\delta}}_1) \mathbf{y}_0 - \widetilde{\mathbf{X}}_1^\top \mathbf{y}_1 \\ -\mathbf{y}_0^\top \mathbf{D}(\widetilde{\boldsymbol{\delta}}_1) \widetilde{\mathbf{X}}_0 - \mathbf{y}_1^\top \widetilde{\mathbf{X}}_1 & \mathbf{y}_0^\top \mathbf{D}(\widetilde{\boldsymbol{\delta}}_1) \mathbf{y}_0 + \mathbf{y}_1^\top \mathbf{y}_1 - n_1 \gamma^{-2} \end{bmatrix},
\end{aligned} \tag{19}$$

where $\mathbf{D}(\widetilde{\boldsymbol{\delta}}_1)$ is a $n_0 \times n_0$ diagonal matrix with $[\mathbf{D}(\widetilde{\boldsymbol{\delta}}_1)]_{ii} = h_i = h(-\widetilde{\mathbf{x}}_i^\top \widetilde{\boldsymbol{\delta}})$.

Olsen (1978) found that the matrix in (19) is negative semidefinite. Theorem 1 in Amemiya (1973) established the asymptotic result that this matrix becomes non-zero with probability one. This ensures the invertibility of the above gradient matrix. Moreover, we introduce an additional condition to support Theorem 4.2.

(C4): The tuning parameter $\lambda_1 \gg C_1 s \cdot \max\{\frac{s^{3/2}}{n_0}, \frac{1}{\sqrt{n_0}}, \sqrt{\frac{\log n}{n_1}}\}$ and $\lambda_2 \gg |\mathcal{G}_{\min}|^{-1} \cdot \max\{\frac{s^{3/2}}{n_0}, \frac{1}{\sqrt{n_0}}, \sqrt{\frac{\log n}{n_1}}\}$.

4.2. Theoretical results

Theorem 4.1: *Suppose that* $y_i^* = \mu_i^* + \mathbf{x}_i^\top \boldsymbol{\beta}^* + \epsilon_i$ *where* $\epsilon_i \overset{iid}{\sim} N(0, \sigma^{*2})$ *and define* $y_i = y_i^* \cdot I(y_i^* > 0)$ *for* $i = 1, \ldots, n$. *Let* $\widehat{\Xi}^{or}$ *denote the oracle solution to the Tobit model when the true group memberships and true support set of* $\boldsymbol{\beta}$ *are known. Suppose conditions (C1)–(C3) hold, and then* $\widehat{\Xi}^{or}$ *corresponds to the unique maximum of the likelihood function and is a consistent estimator of the true parameter values* Ξ^* *such that*

$$\sqrt{n}(\widehat{\Xi}^{or} - \Xi^*) \to N(\mathbf{0}, \boldsymbol{\Sigma}), \tag{20}$$

where $\boldsymbol{\Sigma} = \lim_{n \to \infty} [-\frac{1}{n} \nabla^2 \log L_n(\Xi)|_{\Xi = \Xi^*}]^{-1}$. *Moreover, we denote* $\lambda_{\max}$ *as the maximum eigenvalue of the matrix* $\boldsymbol{\Sigma}$. *If* $\lambda_{\max} = O(1)$ *is satisfied, we have that with probability at least* $1 - p_1 = 1 - C\sqrt{\frac{\log n}{n}} \cdot \exp\{-\frac{n}{2 \log n}\}$,

$$\left\| \widehat{\Theta}^{or} - \Theta^* \right\|_\infty \leq \phi_n, \tag{21}$$

where $\phi_n = 1/\sqrt{\log n}$ *and* C *is a constant.*

For $K \geq 2$, let

$$b_n = \min_{i \in \mathcal{G}_k, j \in \mathcal{G}_{k'}, k \neq k'} |\alpha_i^* - \alpha_j^*| = \min_{k \neq k'} |\tau_i^* - \tau_j^*|$$

be the minimal difference of the common values between the two groups.

Theorem 4.2: *Suppose the conditions in Theorem 4.1 hold and* $K \geq 2$. *If the minimum signal strength of* $\boldsymbol{\delta}^*$ *satisfies* $|\boldsymbol{\delta}_{\mathcal{A}}|_{\min} > (a+1)\lambda_1$ *and* $b_n > a\lambda_2$. *When* $\lambda_1, \lambda_2 \gg \phi_n$, *where* a *is a given constant in (C2), then there exists a local*

minimum $\widehat{\Theta}(\lambda_1,\lambda_2) = (\widehat{\boldsymbol{\alpha}}^\top, \widehat{\boldsymbol{\delta}}^\top, \gamma)^\top$ *of the objective function* $Q(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma)$ *given in (5) satisfying*

$$P\left((\widehat{\boldsymbol{\alpha}}^\top, \widehat{\boldsymbol{\delta}}^\top, \gamma)^\top = (\widehat{\boldsymbol{\alpha}}^{or})^\top, (\widehat{\boldsymbol{\delta}}^{or})^\top, (\widehat{\gamma}^{or})^\top\right) \to 1,$$

that is,

$$P\left(\widehat{\Theta}(\lambda_1,\lambda_2) = \widehat{\Theta}^{or}\right) \to 1.$$

The proofs of these theorems are given in the Appendix.

5. Simulation studies

In this section, we conduct extensive simulation studies to investigate the numerical performance of the proposed approaches. We generate data from the censored heterogeneous linear model:

$$y_i^* = \mu_i + \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i,\ i = 1, \ldots, n,$$

where x_{ij}, $j = 1, \ldots, p$ are generated from independent normal distribution $N(1, 1)$, and the error terms ϵ_i are from independent normal distribution $N(0, 0.5^2)$. We set $y_i = \max\{0, y_i^*\}$ with censoring rate q. The true coefficients are set as $\boldsymbol{\beta} = (5, 1, -2, 0.5, 0.1, 0, \ldots, 0)^\top$, which is a p-dimensional vector with $p-5$ zero elements. To investigate the effect of the magnitude of difference between subgroup-specific effects, we consider two cases for the subgroup structure:

Case 1: $K = 2$, $P(\mu_i = -2) = P(\mu_i = 2) = \frac{1}{2}$;
Case 2: $K = 3$, $P(\mu_i = -2) = P(\mu_i = 2) = P(\mu_i = 0.5) = \frac{1}{3}$.

We evaluate the performance of the estimators obtained using the proposed method using three different penalties (SCAD, MCP and Lasso), and compare them to the penalized Tobit approach (Tobit SCAD, Jacobson & Zou, 2023), which assumes a homogeneous intercept effect μ. Additionally, we present the results of Oracle estimators (Tobit Oracle) as well. For each simulation, we generate 100 datasets with sample size of $n = 100$, for every combination of $q \in \{20\%, 40\%\}$ and $p = 10$, 50, 200. Specifically, we set $\rho = 2$ and set $a_1 = a_2 = 3.7$ for the SCAD penalty and $a_1 = a_2 = 3$ for the MCP penalty. Subsequently, we conduct the simulations by selecting the optimal tuning parameters via minimizing the modified BIC (Wang et al., 2007):

$$\mathrm{BIC}(\lambda_1,\lambda_2) = -2\log L_n(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\delta}}, \widehat{\gamma}) + C_n \log(n)(\widehat{K} + s + 1). \tag{22}$$

Wang et al. (2009) used $C_n = \log(\log(d))$ in the simulation to apply the divergence of the predictor with sample size in high-dimensional scenarios. In this article, we let $C_n = c\log(\log(d))$, where $d = n + p + 1$ and c is a positive constant that we set to 1.5.

We evaluate the methods based on three aspects, accuracy of the coefficient estimates, performance of the variable selection and identifying the subgroup structures. To measure the estimation accuracy of parameters $\widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{\beta}}$, and $\widehat{\sigma}$, we use the square error of the mean squared errors. Let $\boldsymbol{\mu}^*$, $\boldsymbol{\beta}^*$ and σ^* represent the true parameters. The square roots of the mean squared errors for $\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}$ are defined by $\mathrm{err}(\widehat{\boldsymbol{\mu}}) = \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\|_2/\sqrt{n}$, $\mathrm{err}(\widehat{\boldsymbol{\beta}}) = \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ and $\mathrm{err}(\widehat{\sigma}) = |\widehat{\sigma} - \sigma^*|$, respectively.

In order to assess the variable selection of these methods, we report the number of true variables not included (NT) and the number of error variables included (NE). Moreover, to evaluate the performance of the subgroup analysis, we present the estimate of the number of groups ($\widehat{K}$), the rate of false estimation of the number of groups (FK%) and the Rand Index (RI, Rand, 1971), which is defined by

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}},$$

where true positive (TP) indicates that two observations from the same truth group are allocated to the same group, true negative (TN) means two observations from different groups are allocated to different groups, false positive (FP) denotes two observations from different groups but allocated to the same group, and false negative (FN) represents two observations from the same group are allocated to different groups. A high Rand Index indicates a substantial proportion which of individuals being assigned to the correct subgroups.

Tables 1 and 2 present the average square root of the mean squared error (RSME) of the estimates of the five methods. The cases considered in the tables involve varying values of p and censoring rates. It is evident from the

Table 1. The sample means and standard deviations of the index to be measured when $K = 2$.

Censoring	p	Method	Err($\widehat{\mu}$)	Err($\widehat{\beta}$)	Err($\widehat{\sigma}$)	$\widehat{K}$
20%	10	MCP	0.557(0.020)	0.211(0.012)	**0.213(0.013)**	2(0.000)
		SCAD	**0.513(0.069)**	**0.204(0.019)**	0.261(0.022)	2(0.000)
		Lasso	2.314(0.271)	0.306(0.121)	0.330(0.054)	1(0.000)
		Tobit SCAD	2.516(0.337)	0.825(0.240)	1.760(0.098)	–
		Tobit Oracle	0.105(0.007)	0.119(0.013)	0.097(0.005)	2(0.000)
	50	MCP	0.640(0.079)	**0.303(0.034)**	0.305(0.050)	2(0.000)
		SCAD	**0.627(0.102)**	0.314(0.021)	0.311(0.051)	2(0.000)
		Lasso	2.575(0.382)	0.379(0.103)	**0.279(0.042)**	1(0.000)
		Tobit SCAD	2.863(0.415)	0.978(0.241)	1.802(0.117)	–
		Tobit Oracle	0.112(0.006)	0.105(0.012)	0.089(0.005)	2(0.000)
	200	MCP	0.692(0.053)	0.357(0.015)	0.316(0.025)	2(0.000)
		SCAD	**0.671(0.041)**	**0.345(0.019)**	**0.305(0.031)**	2(0.000)
		Lasso	2.659(0.316)	0.501(0.129)	0.344(0.078)	1(0.000)
		Tobit SCAD	2.744(0.292)	0.941(0.312)	1.862(0.109)	–
		Tobit Oracle	0.121(0.008)	0.103(0.0109)	0.101(0.005)	2(0.000)
40%	10	MCP	0.621(0.061)	0.351(0.041)	**0.261(0.014)**	2(0.000)
		SCAD	**0.553(0.079)**	**0.259(0.032)**	0.264(0.022)	2(0.000)
		Lasso	2.448(0.343)	0.298(0.063)	0.455(0.048)	1(0.000)
		Tobit SCAD	2.481(0.384)	0.780(0.231)	1.926(0.111)	–
		Tobit Oracle	0.135(0.011)	0.134(0.014)	0.189(0.025)	2(0.000)
	50	MCP	0.774(0.130)	**0.565(0.066)**	0.425(0.088)	2(0.000)
		SCAD	**0.685(0.114)**	0.578(0.051)	**0.414(0.091)**	2(0.000)
		Lasso	2.329(0.371)	0.557(0.079)	1.325(0.131)	1(0.000)
		Tobit SCAD	2.587(0.495)	1.016(0.218)	1.577(0.079)	–
		Tobit Oracle	0.131(0.014)	0.157(0.012)	0.166(0.027)	2(0.000)
	200	MCP	**0.773(0.122)**	**0.585(0.057)**	**0.549(0.023)**	2(0.000)
		SCAD	0.809(0.118)	0.584(0.052)	0.561(0.031)	2(0.000)
		Lasso	2.515(0.298)	0.804(0.075)	1.011(0.238)	1(0.000)
		Tobit SCAD	2.604(0.531)	0.726(0.209)	1.864(0.122)	–
		Tobit Oracle	0.120(0.048)	0.144(0.008)	0.172(0.052)	2(0.000)

Table 2. The sample means and standard deviations of the index to be measured when $K = 3$.

Censoring	p	Method	Err($\widehat{\mu}$)	Err($\widehat{\beta}$)	Err($\widehat{\sigma}$)	$\widehat{K}$
20%	10	MCP	0.309(0.083)	0.228(0.072)	0.273(0.012)	3(0.000)
		SCAD	**0.245(0.044)**	**0.147(0.013)**	**0.271(0.008)**	3(0.000)
		Lasso	1.946(1.128)	0.349(0.033)	0.310(0.031)	1(0.000)
		Tobit SCAD	1.934(0.264)	0.567(0.178)	1.462(0.094)	–
		Tobit Oracle	0.064(0.004)	0.112(0.002)	0.127(0.009)	3(0.000)
	50	MCP	**0.617(0.062)**	0.485(0.028)	0.325(0.004)	3(0.000)
		SCAD	0.681(0.046)	0.501(0.031)	**0.321(0.004)**	3(0.000)
		Lasso	2.135(0.931)	0.769(0.218)	0.496(0.074)	1(0.000)
		Tobit SCAD	2.172(0.416)	**0.481(0.032)**	1.647(1.015)	–
		Tobit Oracle	0.076(0.008)	0.117(0.003)	0.118(0.013)	3(0.000)
	200	MCP	0.753(0.067)	**0.501(0.042)**	**0.379(0.013)**	3(0.000)
		SCAD	**0.693(0.059)**	0.537(0.044)	0.388(0.024)	3(0.000)
		Lasso	2.004(0.657)	0.741(0.118)	0.428(0.037)	1(0.000)
		Tobit SCAD	2.021(0.269)	0.555(0.037)	1.757(0.085)	–
		Tobit Oracle	0.073(0.054)	0.118(0.007)	0.103(0.017)	3(0.000)
40%	10	MCP	0.347(0.043)	0.346(0.172)	0.285(0.018)	2.98(0.141)
		SCAD	**0.267(0.069)**	**0.288(0.021)**	**0.280(0.032)**	3(0.000)
		Lasso	2.155(0.450)	0.355(0.094)	0.766(0.103)	1(0.000)
		Tobit Lasso	1.994(0.294)	0.642(0.191)	1.459(0.099)	–
		Tobit Oracle	0.052(0.008)	0.110(0.003)	0.058(0.025)	3(0.000)
	50	MCP	0.692(0.071)	0.574(0.054)	**0.474(0.035)**	2.97(0.171)
		SCAD	**0.645(0.080)**	**0.568(0.109)**	0.513(0.021)	2.97(0.223)
		Lasso	2.064(0.233)	0.611(0.058)	0.542(0.037)	1(0.000)
		Tobit SCAD	2.213(0.206)	0.956(0.452)	1.535(0.079)	–
		Tobit Oracle	0.051(0.048)	0.119(0.010)	0.069(0.031)	3(0.000)
	200	MCP	**0.712(0.141)**	0.674(0.128)	**0.469(0.031)**	2.99(0.100)
		SCAD	0.723(0.119)	**0.666(0.072)**	0.483(0.021)	2.97(0.171)
		Lasso	2.369(1.004)	0.810(0.142)	0.558(0.062)	1(0.000)
		Tobit SCAD	2.287(0.407)	0.722(0.195)	1.179(0.121)	–
		Tobit Oracle	0.056(0.018)	0.122(0.004)	0.081(0.054)	3(0.000)

tables that the proposed methods, namely MCP and SCAD, consistently yield smaller RMSE values compared to the Lasso and Tobit SCAD methods across all simulations. Moreover, the method utilizing the lasso penalty consistently underestimates the number of groups. Consequently, the substantial deviation of the estimator can be attributed to the loss of heterogeneous intercept information. Similarly, the Tobit SCAD method when utilized without the subgroup recognition function, shows comparable results.

In Tables 3 and 4, we report the results for variable selection and identification of subgroups. The NT and NE values indicate that our method exhibits comparable performance to other methods in identifying relevant variables

Table 3. The sample means and standard deviations of the index to be measured when $K = 2$.

Censoring	p	Method	NT	NE	RI	FK%
20%	10	MCP	0.51(0.045)	1.79(0.121)	0.943(0.006)	0
		SCAD	0.45(0.028)	**1.63(0.112)**	**0.960(0.002)**	0
		Lasso	**0.42(0.026)**	2.14(0.287)	0.328(0.004)	100
		Tobit SCAD	0.72(0.041)	1.82(0.176)	–	–
		Tobit Oracle	0.00(0.000)	0.00(0.000)	1.000(0.000)	0
	50	MCP	0.65(0.031)	**5.22(0.316)**	0.904(0.009)	0
		SCAD	**0.64(0.019)**	5.34(0.286)	**0.918(0.019)**	0
		Lasso	0.81(0.033)	7.48(0.461)	0.325(0.003)	100
		Tobit SCAD	0.87(0.027)	5.99(0.204)	–	–
		Tobit Oracle	0.00(0.000)	0.00(0.000)	1.000(0.000)	0
	200	MCP	**0.62(0.026)**	**6.55(0.197)**	**0.885(0.025)**	0
		SCAD	0.68(0.041)	6.82(0.231)	0.872(0.008)	0
		Lasso	0.73(0.045)	11.91(0.722)	0.334(0.006)	100
		Tobit SCAD	0.72(0.037)	7.22(0.504)	–	–
		Tobit Oracle	0.00(0.000)	0.00(0.000)	1.000(0.000)	0
40%	10	MCP	0.63(0.041)	1.94(0.209)	0.912(0.011)	0
		SCAD	**0.51(0.039)**	1.83(0.232)	**0.928(0.009)**	0
		Lasso	0.75(0.032)	1.88(0.176)	0.343(0.011)	100
		Tobit SCAD	0.77(0.047)	**1.28(0.244)**	–	–
		Tobit Oracle	0.00(0.000)	0.00(0.000)	1.000(0.000)	0
	50	MCP	0.68(0.035)	5.11(0.359)	**0.877(0.029)**	0
		SCAD	0.62(0.027)	**4.89(0.315)**	0.871(0.033)	0
		Lasso	**0.61(0.041)**	8.28(0.541)	0.332(0.008)	100
		Tobit SCAD	0.71(0.033)	5.81(0.391)	–	–
		Tobit Oracle	0.00(0.000)	0.00(0.000)	1.000(0.000)	0
	200	MCP	**0.66(0.051)**	7.01(0.522)	0.893(0.021)	0
		SCAD	0.68(0.045)	**6.97(0.421)**	**0.895(0.017)**	0
		Lasso	0.89(0.065)	10.13(0.762)	0.328(0.013)	100
		Tobit SCAD	0.81(0.059)	7.87(0.606)	–	–
		Tobit Oracle	0.00(0.000)	0.00(0.000)	1.000(0.000)	0

Table 4. The sample means and standard deviations of the index to be measured when $K = 3$.

Censoring	p	Method	NT	NE	RI	FK%
20%	10	MCP	0.41(0.029)	2.19(0.188)	0.895(0.017)	0
		SCAD	0.39(0.031)	**2.13(0.168)**	**0.918(0.029)**	0
		Lasso	0.54(0.027)	2.96(0.192)	0.343(0.006)	100
		Tobit SCAD	**0.36(0.028)**	2.52(0.095)	–	–
		Tobit Oracle	0.00(0.000)	0.00(0.000)	1.000(0.000)	0
	50	MCP	**0.62(0.032)**	5.39(0.212)	**0.864(0.021)**	0
		SCAD	0.66(0.031)	**5.32(0.312)**	0.848(0.027)	0
		Lasso	0.88(0.042)	7.13(0.402)	0.328(0.003)	100
		Tobit SCAD	0.71(0.035)	5.78(0.321)	–	–
		Tobit Oracle	0.00(0.000)	0.00(0.000)	1.000(0.000)	0
	200	MCP	**0.67(0.033)**	**7.14(0.363)**	0.839(0.028)	0
		SCAD	0.74(0.028)	7.21(0.353)	**0.841(0.033)**	0
		Lasso	0.93(0.048)	10.82(0.599)	0.331(0.005)	100
		Tobit SCAD	0.78(0.039)	8.06(0.448)	–	–
		Tobit Oracle	0.00(0.000)	0.00(0.000)	1.000(0.000)	0
40%	10	MCP	0.51(0.023)	2.16(0.166)	0.884(0.019)	2
		SCAD	**0.43(0.021)**	2.21(0.138)	**0.892(0.012)**	0
		Lasso	0.49(0.026)	2.30(0.243)	0.317(0.005)	100
		Tobit SCAD	0.64(0.039)	**1.91(0.174)**	–	–
		Tobit Oracle	0.00(0.000)	0.00(0.000)	1.000(0.000)	0
	50	MCP	0.73(0.024)	6.52(0.354)	0.834(0.019)	3
		SCAD	**0.71(0.042)**	**6.45(0.476)**	**0.841(0.021)**	5
		Lasso	0.81(0.052)	9.47(0.780)	0.319(0.007)	100
		Tobit SCAD	0.80(0.035)	6.68(0.467)	–	–
		Tobit Oracle	0.00(0.000)	0.00(0.000)	1.000(0.000)	0
	200	MCP	0.75(0.029)	**7.18(0.323)**	0.818(0.021)	1
		SCAD	**0.69(0.034)**	7.21(0.261)	**0.825(0.019)**	3
		Lasso	0.77(0.051)	12.34(0.652)	0.321(0.005)	100
		Tobit SCAD	0.71(0.048)	8.07(0.316)	–	–
		Tobit Oracle	0.00(0.000)	0.00(0.000)	1.000(0.000)	0

while outperforming them in efficiently screening out irrelevant variables. The proposed approaches also achieve higher RI and lower FK values, further establishing their superiority over the alternative methods. Despite facing increased challenges in model recovery due to higher censoring rates and larger parameter dimensions, our method still maintains a remarkably low prediction error rate for the number of subgroups. Incorrect estimations only occur when the deletion ratio reaches 40%, underscoring the robustness of our approach even under such demanding conditions.

6. An empirical application to the HIV drug resistance data

Antiretroviral therapy (ART) is a common medical treatment for human immunodeficiency virus (HIV). However, the high mutation rate of HIV leads to drug-resistant mutations (DRMs) in HIV-infected patients receiving ART. In response to this challenge, physicians regularly monitor HIV viral load. When the patient's treatment regimen fails to suppress the virus, genotypic testing is conducted to check for DRMs, and subsequently, they can update the patient's drug regimen appropriately.

To identify DRMs and quantify the degree of resistance they provide to different ART treatments, the proposed approach is applied to model the relationship between HIV viral load and mutations in the virus's genome. The data used in this section are from the OPTIONS trial by the AIDS Clinical Trials Group (Gandhi et al., 2020), which can be downloaded from the Stanford HIV Drug Resistance database (Shafer, 2006). Specifically, the OPTIONS trial encompassed 412 participants afflicted with HIV-infected, who were undergoing protease inhibitor (PI)-based treatment and grappling with virological failure. Each individual was administered an individualized ART regimen on the basis of their drug resistance and treatment history. Individuals exhibiting moderate drug resistance were randomly allocated to either include nucleoside reverse transcriptase inhibitors (NRTIs) into their optimized treatment regimens or to exclude NRTIs from these regimens. Individuals with high drug resistance were all provided with optimized regimens that encompassed NRTIs.

Our dataset includes $n = 407$ participants who were subjected to a comprehensive 12-week follow-up assessment. Within this dataset, there are a multitude of predictors $p = 601$, including 99 protease (PR) and 240 reverse transcriptase (RT) gene mutation indicators. Due to the technical limitations of the assays employed for its measurement, it cannot be measured when the HIV viral load is less than the threshold (50 copies/ml), the response variable is left censored. As proposed by Soret et al. (2018), we use $\log_{10}$-HIV viral load as our response due to its prevalent conformity to a normal distribution. In this trial, 35.6% of individuals have no detectable viral load, which implies a left censoring ratio of 35.6% for the data sample. We compare our proposed methods (SCAD and MCP) and Tobit SCAD (Jacobson & Zou, 2023) for modelling HIV viral load 12 weeks after drug regimen assignment as a function of several variables, including HIV genotypic mutations, current drug regimen, baseline viral load, and observation week, etc.

First, to evaluate the performance of model selection, we apply Tobit SCAD and the proposed methods (SCAD and MCP) for fitting the entire data respectively. The constant values a_1, a_2, ρ are set as in Section 5. Consequently, sparse models containing covariates M184V, the baseline viral load and RAL are consistently selected across all three approaches. Among them, M184V is an indicator of mutations in the reverse transcriptase (RT) gene, and RAL refers to whether the participant was taking raltegravir. The Stanford University HIV Resistance Database lists M184V as a major NRTIs resistance mutation (Jacobson & Zou, 2023; Shafer, 2006). Moreover, the absence of additional NRTIs being chosen as important variables aligns with the results in Gandhi et al. (2020). This congruence highlights the practical performance of the proposed approaches in model selection. As a result, the specific estimated coefficients are comprehensively shown in Table 5.

The key difference between the proposed methods and the Tobit SCAD lies in the capacity of the proposed methods to identify the subgroup structures within the intercepts. As depicted in Figure 1, we present the estimated density function of $y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}^{\mathrm{TS}}$, where $\widehat{\boldsymbol{\beta}}^{\mathrm{TS}}$ is the coefficient vector estimate for Tobit SCAD model (Jacobson & Zou, 2023). It is not difficult to see the distribution of $y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}^{\mathrm{TS}}$ is multimodal. Consequently, it appears more appropriate to employ a model with heterogeneous intercepts to fit the dataset. Subsequently, in Figure 2, we present the density functions estimates of $y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}^{\mathrm{PS}}$ for two subgroups characterized by different intercepts, where $\widehat{\boldsymbol{\beta}}^{\mathrm{PS}}$ is the coefficient vector estimate for proposed SCAD model. The density function estimates of the proposed MCP method exhibit similar characteristics and are therefore omitted here. When compared with the density functions shown in Figure 1, it is evident that each subgroup in Figure 2 displays an unimodal distribution with greater homogeneity.

The application of the proposed SCAD method leads to the detection of two subgroups with sample sizes of $n_1 = 47$ and $n_2 = 360$, respectively. It is worth noting that the patients in the OPTION design were originally categorized into two groups: a randomized group consisting of 356 patients, and a highly resistant group, comprising 51 patients. Consequently, we can consider this historical grouping as a control group structure. We calculate the rand index

Table 5. The estimators of parameters in the example.

	Tobit SCAD	Proposed(SCAD)		Proposed(MCP)	
Intercept	1.98	1.35	4.29	1.47	4.38
Baseline	0.36	0.12		0.11	
M184V	0.53	0.58		0.48	
RAL	−0.65	−0.61		−0.59	

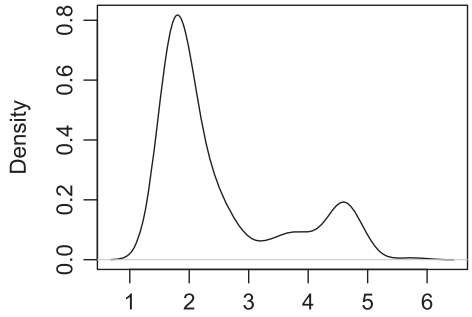

Figure 1. Density plot of the response variable after adjusting for the effects of the covariates in the empirical example.

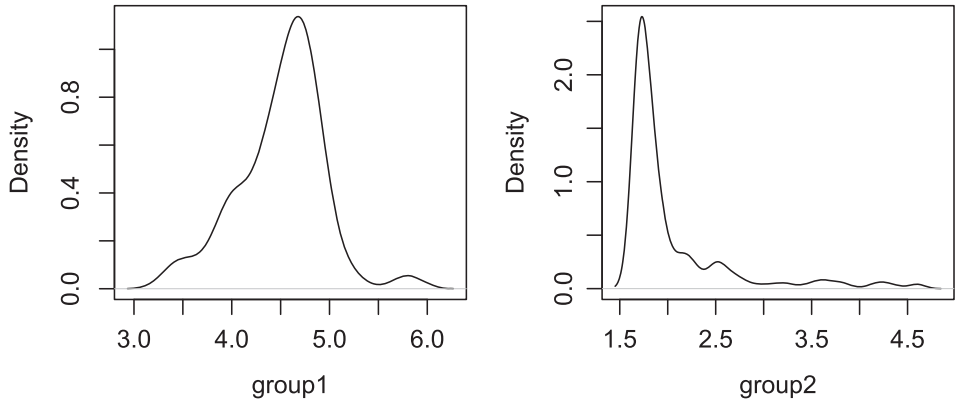

Figure 2. Density plots of response variable after adjusting for the effects of the covariates under two different intercept groups for proposed SCAD method.

Table 6. The in-sample error and out-of-sample error

Method	In-sample error	Out-of-sample error
Tobit SCAD	0.91	1.08
proposed(SCAD)	**0.18**	**0.17**
proposed(MCP)	0.21	0.23
control(SCAD)	0.34	0.38

(RI) for the proposed SCAD method based on the control group structure. The rand index (RI) is 0.757, which indicates a substantial degree of agreement between the detected subgroup structure and the historical one. This suggests that the subgroup composition may undergo changes as a result of antiretroviral therapy, a phenomenon that is quite reasonable in the context of HIV patient management.

In addition, we compare the in-sample error and out-of-sample error for the aforementioned methods and the Tobit SCAD method with known subgroup structures. These errors are defined as the mean square error of

$$(i)\ \mathrm{err}(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2, \quad (ii)\ \mathrm{err}(y_i, \hat{y}_i^{-}) = (y_i - \hat{y}_i^{-})^2, \quad i = 1, \ldots, n$$

across all samples, where $\hat{y}_i$ is the fitted response, and $\hat{y}_i^{-}$ is the predicted response using a 5-fold cross-validation approach. To ensure comparability, we employ stratified sampling to maintain a consistent left-censored ratio across each test set. Table 6 presents the two types of criteria. For both two types of error, the Tobit SCAD method, when applied with known subgroup structures, exhibits smaller errors compared to the homogeneous model. This result underscores the significance of capturing the inherent heterogeneity within the dataset, as it enables more effective modelling of the influential variables affecting the response. While the errors in the control group are slightly larger compared to the proposed methods, it is essential to consider that these discrepancies may stem from changes in drug resistance among patients during treatment. Moreover, the disparities with the control group could serve as a basis for further examination of specific patient cases. It is conceivable that the existence of distinct groups may be attributed to disparities in the treatment trajectories of patients or inherent physiological differences among individuals. In clinical terms, these findings suggest the potential for devising more precise and tailored management protocols for specific patient subgroups, improving the overall treatment efficacy.

7. Conclusion

This paper introduces a novel approach to analysing censored data with potential heterogeneity in intercept effects by combining the penalty Tobit likelihood function with a concave fusion penalty. The proposed method can automatically identify heterogeneous structures in intercept effects and conduct variable selection.

To address the optimization problem associated with the method, we propose an algorithm based on the generalized coordinate descent method and alternating direction method of multipliers. This algorithm simplifies the optimization problem and reduces computational costs by employing a quadratic optimization function instead of a complex nonlinear optimization problem. This choice ensures efficient computation, particularly in scenarios with high complexity, while still maintaining good properties for the estimator. Furthermore, we establish the oracle property of parameter estimators. It is shown that within the domain of the oracle solution, a local minimum point of the objective function can be consistent with the oracle solution, providing theoretical support for the correctness of the method. Our ADMM-GCD algorithm with SCAD or MCP performs well in both extensive simulation case studies and the application of real data.

The proposed method can also be extended to incorporate the fusion of coefficient terms. However, there are still challenges in applying this method to generalized linear models or survival models, and further research is needed to develop algorithms and establish theoretical properties in these models.

Acknowledgements

We are very grateful to the editor, managing editor and the referee for their comments that improved this article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Natural Science Foundation of China (NSF) of China [grant numbers 12171450 and 71921001.

References

Alhamzawi, A. (2020). A new Bayesian elastic net for Tobit regression. *Journal of Physics: Conference Series*, *1664*(1), 012047.

Alhamzawi, R. (2016). Bayesian elastic net Tobit quantile regression. *Communications in Statistics-Simulation and Computation*, *45*(7), 2409–2427. https://doi.org/10.1080/03610918.2014.904341

Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica: Journal of the Econometric Society*, *41*(6), 997–1016. https://doi.org/10.2307/1914031

Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics*, *24*(1–2), 3–61. https://doi.org/10.1016/0304-4076(84)90074-5

Bondell, H. D., & Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, *64*(1), 115–123. https://doi.org/10.1111/biom.2008.64.issue-1

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*(1), 1–122. https://doi.org/10.1561/2200000016

Bradic, J., Fan, J., & Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Annals of Statistics*, *39*(6), 3092. https://doi.org/10.1214/11-AOS911

Dagne, G. A. (2016). A growth mixture Tobit model: Application to AIDS studies. *Journal of Applied Statistics*, *43*(7), 1174–1185. https://doi.org/10.1080/02664763.2015.1092114

Everitt, B. (2013). *Finite mixture distributions*. Springer Science & Business Media.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*(1), 101–148.

Fan, Y., & Tang, C. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(3), 531–552. https://doi.org/10.1111/rssb.12001

Gandhi, R. T., Tashima, K. T., Smeaton, L. M., Vu, V., Ritz, J., Andrade, A., Eron, J. J., Hogg, E., & Fichtenbaum, C. J. (2020). Long-term outcomes in a large randomized trial of HIV-1 salvage therapy: 96-week results of AIDS Clinical Trials Group A5241 (OPTIONS). *The Journal of Infectious Diseases*, *221*(9), 1407–1415. https://doi.org/10.1093/infdis/jiz281

Jacobson, T., & Zou, H. (2023). High-dimensional censored regression via the penalized Tobit likelihood. *Journal of Business & Economic Statistics*, *42*(1), 286–297. https://doi.org/10.1080/07350015.2023.2182309

Johnson, B. A. (2009). On lasso for censored data. *Electronic Journal of Statistics*, *3*, 485–506. https://doi.org/10.1214/08-EJS322

Ma, S., & Huang, J. (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, *112*(517), 410–423. https://doi.org/10.1080/01621459.2016.1148039

Ma, S., Huang, J., Zhang, Z., & Liu, M. (2019). Exploration of heterogeneous treatment effects via concave fusion. *The International Journal of Biostatistics*, *16*(1), 20180026. https://doi.org/10.1515/ijb-2018-0026
Müller, P., & van de Geer, S. (2016). Censored linear model in high dimensions: Penalised linear regression on high-dimensional data with left-censored response variable. *Test*, *25*(1), 75–92. https://doi.org/10.1007/s11749-015-0441-7
Olsen, R. J. (1978). Note on the uniqueness of the maximum likelihood estimator for the Tobit model. *Econometrica: Journal of the Econometric Society*, *46*(5), 1211–1215. https://doi.org/10.2307/1911445
Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, *25*(3), 303–325. https://doi.org/10.1016/0304-4076(84)90004-6
Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850. https://doi.org/10.1080/01621459.1971.10482356
Shafer, R. W. (2006). Rationale and uses of a public HIV drug-resistance database. *The Journal of Infectious Diseases*, *194*(s1), S51–S58. https://doi.org/10.1086/jid.2006.194.issue-s1
Shen, J., & He, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *Journal of the American Statistical Association*, *110*(509), 303–312. https://doi.org/10.1080/01621459.2014.894763
Soret, P., Avalos, M., Wittkop, L., Commenges, D., & Thiébaut, R. (2018). Lasso regularization for left-censored Gaussian outcome and high-dimensional predictors. *BMC Medical Research Methodology*, *18*(1), 1–13. https://doi.org/10.1186/s12874-018-0609-4
Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society*, *26*(1), 24–36. https://doi.org/10.2307/1907382
Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, *109*(3), 475–494. https://doi.org/10.1023/A:1017501703105
Wang, H., Li, B., & Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(3), 671–683. https://doi.org/10.1111/j.1467-9868.2008.00693.x
Wang, H., Li, R., & Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, *94*(3), 553–568. https://doi.org/10.1093/biomet/asm053
Wang, X., Zhu, Z., & Zhang, H. H. (2019). Spatial automatic subgroup analysis for areal data with repeated measures. arXiv:1906.01853.
Yan, X., Yin, G., & Zhao, X. (2021). Subgroup analysis in censored linear regression. *Statistica Sinica*, *31*(2), 1027–1054.
Yang, Y., & Zou, H. (2013). An efficient algorithm for computing the HHSVM and its generalizations. *Journal of Computational and Graphical Statistics*, *22*(2), 396–415. https://doi.org/10.1080/10618600.2012.680324
Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *32*(2), 894–942.
Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research*, *7*(90), 2541–2563.
Zhou, X., & Liu, G. (2016). LAD-lasso variable selection for doubly censored median regression models. *Communications in Statistics-Theory and Methods*, *45*(12), 3658–3667. https://doi.org/10.1080/03610926.2014.904357
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735

Appendix

A.1 Proof of Proposition 3.1

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

$$L(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \gamma^{(t+1)}, \boldsymbol{\eta}^{(t+1)}, \boldsymbol{\varphi}^{(t)}) \leq L(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \gamma^{(t+1)}, \boldsymbol{\eta}, \boldsymbol{\varphi}^{(t)})$$

for any $\boldsymbol{\eta}$. Define

$$f^{(t+1)} = \inf_{\boldsymbol{\Delta\alpha}^{(t+1)} - \boldsymbol{\eta} = 0} \left\{ L(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \gamma^{(t+1)}, \boldsymbol{\eta}, \boldsymbol{\varphi}^{(t)}) \right\},$$

and then we have

$$L(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \gamma^{(t+1)}, \boldsymbol{\eta}^{(t+1)}, \boldsymbol{\varphi}^{(t)}) \leq f^{(t+1)}.$$

Let k be a non-negative integer, since $\boldsymbol{\varphi}^{(t+k-1)} = \boldsymbol{\varphi}^{(t)} + \rho \sum_{i=1}^{k-1} (\boldsymbol{\Delta\alpha}^{(t+i)} - \boldsymbol{\eta}^{(t+i)})$, so we can obtain that

$$\begin{aligned}
&L(\boldsymbol{\alpha}^{(t+k)}, \boldsymbol{\delta}^{(t+k)}, \gamma^{(t+k)}, \boldsymbol{\eta}^{(t+k)}, \boldsymbol{\varphi}^{(t+k-1)}) \\
&= \ell_n(\boldsymbol{\alpha}^{(t+k)}, \boldsymbol{\delta}^{(t+k)}, \gamma^{(t+k)}) + \sum_{j=1}^{p} P_{\lambda_1}(|\delta_j^{(t+k)}|) + \sum_{i<j} P_{\lambda_2}(|\eta_{ij}^{(t+k)}|) \\
&\quad + \boldsymbol{\varphi}^{(t+k-1)T}(\boldsymbol{\Delta\alpha}^{(t+k)} - \boldsymbol{\eta}^{(t+k)}) + \frac{\rho}{2} \|\boldsymbol{\Delta\alpha}^{(t+k)} - \boldsymbol{\eta}^{(t+k)}\|_2^2 \\
&= S(\boldsymbol{\alpha}^{(t+k)}, \boldsymbol{\delta}^{(t+k)}, \gamma^{(t+k)}, \boldsymbol{\eta}^{(t+k)}) + \frac{\rho}{2} \|\boldsymbol{\Delta\alpha}^{(t+k)} - \boldsymbol{\eta}^{(t+k)}\|_2^2 \\
&\quad + \left[\boldsymbol{\varphi}^{(t)} + \rho \sum_{i=1}^{k-1} (\boldsymbol{\Delta\alpha}^{(t+i)} - \boldsymbol{\eta}^{(t+i)}) \right] \times (\boldsymbol{\Delta\alpha}^{(t+k)} - \boldsymbol{\eta}^{(t+k)}) \\
&\leq f^{(t+k)}.
\end{aligned}$$

According to Yang and Zou (2013), we can infer that the GCD limit point $(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \gamma^{(t+1)})$ is the coordinatewise minimum of $L(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma, \boldsymbol{\eta}^{(t)}, \boldsymbol{\varphi}^{(t)})$. Additionally, the objective function is convex with respect to $\boldsymbol{\eta}$. Then according to Theorem 4.1 in Tseng (2001), the sequence $(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\delta}^{(t)}, \gamma^{(t)}, \boldsymbol{\eta}^{(t)})$ converges to a coordinatewise minimum point $(\boldsymbol{\alpha}^*, \boldsymbol{\delta}^*, \gamma^*, \boldsymbol{\eta}^*)$. Thus we have

$$\begin{aligned} f^* &= \lim_{t\to\infty} f^{(t+1)} = \lim_{t\to\infty} f^{(t+k)} \\ &= \inf_{\boldsymbol{\Delta\alpha}^*-\boldsymbol{\eta}=0} \left\{ \ell_n(\boldsymbol{\alpha}^*, \boldsymbol{\delta}^*, \gamma^*) + \sum_{j=1}^{p} P_{\lambda_1}(|\delta_j^*|) + \sum_{i<j} P_{\lambda_2}(|\eta_{ij}|) \right\}, \end{aligned}$$

and for all $k \geq 0$

$$\begin{aligned} f^* &\geq \lim_{t\to\infty} L(\boldsymbol{\alpha}^{(t+k)}, \boldsymbol{\delta}^{(t+k)}, \gamma^{(t+k)}, \boldsymbol{\eta}^{(t+k)}, \boldsymbol{\varphi}^{(t+k-1)}) \\ &= \ell_n(\boldsymbol{\alpha}^*, \boldsymbol{\delta}^*, \gamma^*) + \sum_{j=1}^{p} P_{\lambda_1}(|\delta_j^*|) + \sum_{i<j} P_{\lambda_2}(|\eta_{ij}^*|) + \lim_{t\to\infty} \boldsymbol{\varphi}^{(t)\top}(\boldsymbol{\Delta\alpha}^* - \boldsymbol{\eta}^*) \\ &\quad + \left(k - \frac{1}{2}\right) \rho \|\boldsymbol{\Delta\alpha}^* - \boldsymbol{\eta}^*\|_2^2. \end{aligned}$$

Therefore,

$$\left(k - \frac{1}{2}\right) \rho \|\boldsymbol{\Delta\alpha}^* - \boldsymbol{\eta}^*\|_2^2 \leq \inf_{\boldsymbol{\Delta\alpha}^*-\boldsymbol{\eta}=0} \left\{ \sum_{i<j} P_{\lambda_2}(|\eta_{ij}|) \right\} - \sum_{i<j} P_{\lambda_2}(|\eta_{ij}^*|) - \lim_{t\to\infty} \boldsymbol{\varphi}^{(t)\top}(\boldsymbol{\Delta\alpha}^* - \boldsymbol{\eta}^*).$$

Since the above inequality holds for all $k \geq 0$, then $\lim_{t\to\infty} \|\mathbf{r}^{(t)}\|_2^2 = \|\boldsymbol{\Delta\alpha}^* - \boldsymbol{\eta}^*\|_2^2 = 0$.

Moreover, since $(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \gamma^{(t+1)})$ minimize $L(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma, \boldsymbol{\eta}^{(t)}, \boldsymbol{\varphi}^{(t)})$, by definition, we have that

$$\begin{aligned} &\partial L(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \gamma^{(t+1)}, \boldsymbol{\eta}^{(t)}, \boldsymbol{\varphi}^{(t)})/\partial\boldsymbol{\alpha} \\ &= \partial S(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \gamma^{(t+1)}, \boldsymbol{\eta}^{(t)})/\partial\boldsymbol{\alpha} + \boldsymbol{\Delta}^\top \boldsymbol{\varphi}^{(t)} + \rho \boldsymbol{\Delta}^\top (\boldsymbol{\Delta\alpha}^{(t+1)} - \boldsymbol{\eta}^{(t)}) \\ &= \partial S(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \gamma^{(t+1)}, \boldsymbol{\eta}^{(t)})/\partial\boldsymbol{\alpha} + \boldsymbol{\Delta}^\top \left(\boldsymbol{\varphi}^{(t)} + \rho(\boldsymbol{\Delta\alpha}^{(t+1)} - \boldsymbol{\eta}^{(t)})\right) \\ &= \partial S(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \gamma^{(t+1)}, \boldsymbol{\eta}^{(t)})/\partial\boldsymbol{\alpha} + \boldsymbol{\Delta}^\top \boldsymbol{\varphi}^{(t+1)} + \rho \boldsymbol{\Delta}^\top (\boldsymbol{\eta}^{(t+1)} - \boldsymbol{\eta}^{(t)}) = 0. \end{aligned}$$

Thus, $\|\mathbf{s}^{(t+1)}\|_2^2 = \rho \boldsymbol{\Delta}^\top (\boldsymbol{\eta}^{(t+1)} - \boldsymbol{\eta}^{(t)}) = -(\partial S(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \gamma^{(t+1)}, \boldsymbol{\eta}^{(t)})/\partial\boldsymbol{\alpha} + \boldsymbol{\Delta}^\top \boldsymbol{\varphi}^{(t+1)})$ Since $\lim_{t\to\infty} \|\mathbf{r}^{(t)}\|_2^2 = \|\boldsymbol{\Delta\alpha}^* - \boldsymbol{\eta}^*\|_2^2 = 0$,

$$\begin{aligned} &\lim_{t\to\infty} \partial L(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \gamma^{(t+1)}, \boldsymbol{\eta}^{(t)}, \boldsymbol{\varphi}^{(t)})/\partial\boldsymbol{\alpha} \\ &= \lim_{t\to\infty} \partial S(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \gamma^{(t+1)}, \boldsymbol{\eta}^{(t)})/\partial\boldsymbol{\alpha} + \boldsymbol{\Delta}^\top \boldsymbol{\varphi}^{(t)} = 0. \end{aligned}$$

Therefore, $\lim_{t\to\infty} \|\mathbf{s}^{(t+1)}\|_2^2 = 0$.

A.2 Proof of Theorem 4.1

Let $\boldsymbol{\theta} = (\widehat{\Xi}^{(or)} - \widehat{\Xi}^*), q = K + s + 1$. We can see that $\sqrt{n}\boldsymbol{\theta} \sim N(0, \boldsymbol{\Sigma})$ according to (20) where $\boldsymbol{\Sigma} = [-\frac{1}{n}\nabla^2 \log L_n(\Xi)|_{\Xi=\Xi^*}]^{-1}$.

In order to get the tail probability of the event in (21), we introduce a q-dimension vector $\mathbf{a} = (a_1, \ldots, a_q)^\top$ satisfying $\|\mathbf{a}\|_2 = 1$, that is $a_1^2 + \ldots + a_q^2 = 1$. It is obvious that $\mathbf{a}^\top\boldsymbol{\theta} \sim N(0, \frac{1}{n}\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a})$. Since

$$|\mathbf{a}^\top\boldsymbol{\theta}| = |a_1\theta_1 + \ldots + a_q\theta_q| \leq |a_1\theta_1| + \ldots + |a_q\theta_q|,$$

$|\mathbf{a}^\top\boldsymbol{\theta}| \leq 1 \cdot |\theta_j|_{\max}$. Then it is easy to get $\max_{\|\mathbf{a}\|_2=1} |\mathbf{a}^\top\boldsymbol{\theta}| = \max_j |\theta_j|$. Since $\boldsymbol{\Sigma}$ is a symmetric positive definite matrix, the maximum value of the quadratic $f = \mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}$ at $\|\mathbf{a}\| = 1$ is the largest eigenvalue $\lambda_{\max}$ of the matrix $\boldsymbol{\Sigma}$.

For a normal distribution $X \sim N(0, \sigma^2)$, the two-tailed probability inequality is

$$P(|X| < c) \geq 1 - \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma}{c} \cdot \exp\left\{-\frac{c^2}{2\sigma^2}\right\}. \tag{A1}$$

So

$$P(|\mathbf{a}^\top\boldsymbol{\theta}| < \phi_n) \geq 1 - \sqrt{\frac{2}{\pi}} \cdot \frac{\sqrt{f/n}}{\phi_n} \cdot \exp\left\{-\frac{\phi_n^2}{2(f/n)}\right\}.$$

Since the right-side of the above inequality with respect to f is a decreasing function, then

$$P\left(\max_{\|\mathbf{a}\|_2=1} |\mathbf{a}^\top\boldsymbol{\theta}| < \phi_n\right) \geq 1 - \sqrt{\frac{2}{\pi}} \cdot \frac{\sqrt{\lambda_{\max}/n}}{\phi_n} \cdot \exp\left\{-\frac{\phi_n^2}{2(\lambda_{\max}/n)}\right\}.$$

Therefore,

$$P\left(\left\|\widehat{\Theta}^{\text{or}}-\Theta^*\right\|_\infty<\phi_n\right)=P\left(\left\|\widehat{\Xi}^{\text{or}}-\Xi^*\right\|_\infty<\phi_n\right)$$
$$=P\left(\max_j|\theta_j|<\phi_n\right)=P\left(\max_{\|\mathbf{a}\|_2=1}|\mathbf{a}^\top\boldsymbol{\theta}|<\phi_n\right)$$
$$\geq 1-\sqrt{\frac{2}{\pi}}\cdot\frac{\sqrt{\lambda_{\max}/n}}{\phi_n}\cdot\exp\left\{-\frac{\phi_n^2}{2(\lambda_{\max}/n)}\right\}. \tag{A2}$$

Since $\lambda_{\max}=O(1)$, we set $\phi_n=1/\sqrt{\log n}$ and $t=\sqrt{\log n/n}$. Then $P(\|\widehat{\Theta}^{\text{or}}-\Theta^*\|_\infty<\phi_n)\geq 1-C\sqrt{\frac{\log n}{n}}\cdot\exp\{-\frac{n}{2\log n}\}=1-C\cdot t\exp\{-\frac{1}{2t^2}\}$, where C is a constant. Obviously we have $t\to 0$ and then $P(\|\widehat{\Theta}^{\text{or}}-\Theta^*\|_\infty<\phi_n)\to 1$ when $n\to\infty$.

A.3 Proof of Theorem 4.2

First, with the underlying group division $\mathcal{G}_1,\ldots,\mathcal{G}_K$ and true support set $\mathcal{A}$ we define

$$L_n(\boldsymbol{\alpha},\boldsymbol{\delta},\gamma)=\ell_n(\boldsymbol{\alpha},\boldsymbol{\delta},\gamma),\quad P_{\lambda_1}(\boldsymbol{\delta})=\lambda_1\sum_{j=1}^{p}\rho(|\delta_j|),\quad P_{\lambda_2}(\boldsymbol{\alpha})=\lambda_2\sum_{i<j}\rho(|\alpha_i-\alpha_j|),$$
$$L_n^{\mathcal{O}}(\boldsymbol{\tau},\boldsymbol{\delta_1},\gamma)=\ell_n^{\mathcal{O}}(\boldsymbol{\tau},\boldsymbol{\delta_1},\gamma),\quad P_{\lambda_1}^{\mathcal{O}}(\boldsymbol{\delta_1})=\lambda_1\sum_{j\in\mathcal{A}}\rho(|\delta_j|),\quad P_{\lambda_2}^{\mathcal{O}}(\boldsymbol{\tau})=\lambda_2\sum_{k<k'}|\mathcal{G}_k|\,|\mathcal{G}_{k'}|\rho(|\tau_i-\tau_j|),$$

and denote

$$Q_n(\boldsymbol{\alpha},\boldsymbol{\delta},\gamma)=L_n(\boldsymbol{\alpha},\boldsymbol{\delta},\gamma)+P_{\lambda_1}(\boldsymbol{\delta})+P_{\lambda_2}(\boldsymbol{\alpha}),$$
$$Q_n^{\mathcal{O}}(\boldsymbol{\tau},\boldsymbol{\delta_1},\gamma)=L_n^{\mathcal{O}}(\boldsymbol{\tau},\boldsymbol{\delta_1},\gamma)+P_{\lambda_1}^{\mathcal{O}}(\boldsymbol{\delta_1})+P_{\lambda_2}^{\mathcal{O}}(\boldsymbol{\tau}).$$

Let $T:\mathcal{M}_{\mathcal{G}}\to R^K$ be the mapping such as that $T(\boldsymbol{\alpha})$ is the $K\times 1$ vector whose kth coordinate equals to the common value of α_i for $i\in\mathcal{G}_k$. And let $T_0:R^n\to R^K$ be the mapping such that $T_0(\boldsymbol{\alpha})=\{|\mathcal{G}_k|^{-1}\sum_{i\in\mathcal{G}_k}\alpha_i\}_{k=1}^K$. Let $S:R^p\to R^s$ be the mapping such that $S(\boldsymbol{\delta})$ retains only the part of $\boldsymbol{\delta}$ whose corner is labelled $\mathcal{A}$, that is $S(\boldsymbol{\delta})=\boldsymbol{\delta}_{\mathcal{A}}$. And let $S^{-1}:R^s\to\mathcal{M}_{\mathcal{B}}$ be the mapping such that $S^{-1}(\boldsymbol{\delta_1})=(\boldsymbol{\delta_1}^\top,\mathbf{0}_{\mathcal{A}^c}^\top)^\top$. Obviously, when $\boldsymbol{\alpha}\in\mathcal{M}_{\mathcal{G}}$ and $\boldsymbol{\delta}\in\mathcal{M}_{\mathcal{B}}$, $T(\boldsymbol{\alpha})=T_0(\boldsymbol{\alpha})$ and $\boldsymbol{\delta}_{\mathcal{A}}=S(\boldsymbol{\delta})$. Moreover, for every $\boldsymbol{\delta}\in\mathcal{M}_{\mathcal{B}}$ and $\boldsymbol{\alpha}\in\mathcal{M}_{\mathcal{G}}$, we have $P_{\lambda_1}(\boldsymbol{\delta})=P_{\lambda_1}^{\mathcal{O}}(\boldsymbol{\delta}_{\mathcal{A}})$ and $P_{\lambda_2}(\boldsymbol{\alpha})=P_{\lambda_2}^{\mathcal{O}}(T(\boldsymbol{\alpha}))$. For every $\boldsymbol{\delta_1}\in R^s$ and $\boldsymbol{\tau}\in R^K$, we have $P_{\lambda_1}(S^{-1}(\boldsymbol{\delta_1}))=P_{\lambda_1}^{\mathcal{O}}(\boldsymbol{\delta_1})$ and $P_{\lambda_2}(T^{-1}(\boldsymbol{\tau}))=P_{\lambda_2}^{\mathcal{O}}(\boldsymbol{\tau})$. Hence

$$Q_n(\boldsymbol{\alpha},\boldsymbol{\delta},\gamma)=Q_n^{\mathcal{O}}(T(\boldsymbol{\alpha}),\boldsymbol{\delta}_{\mathcal{A}},\gamma),\ Q_n^{\mathcal{O}}(\boldsymbol{\tau},\boldsymbol{\delta_1},\gamma)=Q_n(T^{-1}(\boldsymbol{\tau}),S^{-1}(\boldsymbol{\delta_1}),\gamma). \tag{A3}$$

Consider the neighbourhood of $(\boldsymbol{\alpha^*},\boldsymbol{\delta^*},\gamma^*)$

$$\Psi=\{\boldsymbol{\alpha}\in R^n,\boldsymbol{\delta}\in R^p,\gamma\in R:\left\|((\boldsymbol{\alpha}-\boldsymbol{\alpha^*})^\top,(\boldsymbol{\delta}-\boldsymbol{\delta^*})^\top,\gamma-\gamma^*)^\top\right\|_\infty\leq\phi_n\}.$$

Define the event $E_1=\{\widehat{\Theta}\in\Psi\}$. By Theorem 1 we have $P(E_1^c)\leq p_1$. For any $\boldsymbol{\alpha}\in R^n$ and $\boldsymbol{\delta}\in R^p$ let $\boldsymbol{\alpha^0}=T^{-1}(T_0(\boldsymbol{\alpha}))$ and $\boldsymbol{\delta^0}=S^{-1}(\boldsymbol{\delta}_{\mathcal{A}})$. We will prove that $(\widehat{\boldsymbol{\alpha}}^{\text{or}},\widehat{\boldsymbol{\delta}}^{\text{or}},\widehat{\gamma}^{\text{or}})$ is a local minimizer of the objective function $Q_n(\boldsymbol{\alpha},\boldsymbol{\delta},\gamma)$ with probability approaching 1 through the following two steps.

(i) On the event E_1, $Q_n(\boldsymbol{\alpha^0},\boldsymbol{\delta^0},\gamma)\geq Q_n(\widehat{\boldsymbol{\alpha}}^{\text{or}},\widehat{\boldsymbol{\delta}}^{\text{or}},\widehat{\gamma}^{\text{or}})$ for any $(\boldsymbol{\alpha}^\top,\boldsymbol{\delta}^\top,\gamma)^\top\in\Psi$.

(ii) There is an event E_2 such that $P(E_2^c)\leq p_2=\frac{n_1}{n\sqrt{\log n}}$. On $E_1\cap E_2$, there is a neighbourhood of $((\widehat{\boldsymbol{\alpha}}^{\text{or}})^\top,(\widehat{\boldsymbol{\delta}}^{\text{or}})^\top,\widehat{\gamma}^{\text{or}})^\top$, denoted by $\Psi_n=\{\boldsymbol{\alpha},\boldsymbol{\delta}:\|((\boldsymbol{\alpha}-\widehat{\boldsymbol{\alpha}}^{\text{or}})^\top,(\boldsymbol{\delta}-\widehat{\boldsymbol{\delta}}^{\text{or}})^\top)^\top\|_\infty\leq t_n\}$ such that $Q_n(\boldsymbol{\alpha},\boldsymbol{\delta},\gamma)\geq Q_n(\boldsymbol{\alpha^0},\boldsymbol{\delta^0},\gamma)$ for any $(\boldsymbol{\alpha}^\top,\boldsymbol{\delta}^\top,\gamma)^\top\in\Psi\cap\Psi_n$ for sufficiently large n.

Therefore, by the result of (i) and (ii), we have $Q_n(\boldsymbol{\alpha},\boldsymbol{\delta},\gamma)\geq Q_n(\widehat{\boldsymbol{\alpha}}^{\text{or}},\widehat{\boldsymbol{\delta}}^{\text{or}},\widehat{\gamma}^{\text{or}})$ for any $(\boldsymbol{\alpha}^\top,\boldsymbol{\delta}^\top,\gamma)^\top\in\Psi\cap\Psi_n$, so that $((\widehat{\boldsymbol{\alpha}}^{\text{or}})^\top,(\widehat{\boldsymbol{\delta}}^{\text{or}})^\top,\widehat{\gamma}^{\text{or}})^\top$ is a strict local minimizer of $Q_n(\boldsymbol{\alpha},\boldsymbol{\delta},\gamma)$ on the event $E_1\cap E_2$ with $P(E_1\cap E_2)\geq 1-p_1-p_2$ for sufficiently large n.

Step (i): Since $(\widehat{\boldsymbol{\tau}}^{\text{or}},\widehat{\boldsymbol{\delta}}_1^{\text{or}},\widehat{\gamma}^{\text{or}})$ is a global minimizer of $L_n^{\mathcal{O}}(\boldsymbol{\tau},\boldsymbol{\delta_1},\gamma)$, $L_n^{\mathcal{O}}(T_0(\boldsymbol{\alpha}),S(\boldsymbol{\delta}),\gamma)\geq L_n^{\mathcal{O}}(\widehat{\boldsymbol{\tau}}^{or},\widehat{\boldsymbol{\delta}}_1^{or},\widehat{\gamma}^{or})$ for all $(\boldsymbol{\alpha}^\top,\boldsymbol{\delta}^\top,\gamma)\in\Psi$. Then we derive that $P_{\lambda_1}^{\mathcal{O}}(S(\boldsymbol{\delta}))$ is a constant which does not depend on $\boldsymbol{\delta}$ for $\boldsymbol{\delta}\in\Psi$ and $P_{\lambda_2}^{\mathcal{O}}(T_0(\boldsymbol{\alpha}))$ is also a constant which does not depend on $\boldsymbol{\alpha}$ for $\boldsymbol{\alpha}\in\Psi$. Let $T_0(\boldsymbol{\alpha})=\boldsymbol{\tau}=(\tau_1,\ldots,\tau_K)^\top$. For any $k\neq k'$, since

$$|\tau_k^*-\tau_{k'}^*|=|\tau_k^*-\tau_{k'}^*+\tau_k-\tau_k+\tau_{k'}-\tau_{k'}|$$
$$\leq|\tau_k-\tau_{k'}|+|\tau_k^*-\tau_k|+|\tau_{k'}-\tau_{k'}^*|,$$

so

$$|\tau_k-\tau_{k'}|\geq|\tau_k^*-\tau_{k'}^*|-2\sup_k|\tau_k-\tau_k^*|,$$

and

$$\sup_k |\tau_k - \tau_k^*| = \sup_k \left| \sum_{i\in\mathcal{G}_k} \alpha_i/|\mathcal{G}_k| - \tau_k^* \right| = \sup_k \left| \sum_{i\in\mathcal{G}_k} (\alpha_i - \alpha_k^*)/|\mathcal{G}_k| \right|$$
$$\leq \sup_k \sup_{i\in\mathcal{G}_k} |\alpha_i - \alpha_i^*| = \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_\infty. \tag{A4}$$

Therefore, for all k and k',

$$|\tau_k - \tau_{k'}| \geq |\tau_k^* - \tau_{k'}^*| - 2\|\alpha_k - \alpha_k^*\|_\infty \geq b_n - 2\phi_n > a\lambda_2,$$

which indicates $\rho(|\tau_k - \tau_{k'}|)$ is a constant by Condition (C2), and as a result $P^{\mathcal{G}}_{\lambda_2}(T_0(\boldsymbol{\alpha}))$ is also a constant. Similarly, for any $j \in \mathcal{A}$, let $S(\boldsymbol{\alpha}) = \boldsymbol{\delta_1} = (\delta_1, \ldots, \delta_{p_1})$, since

$$|\delta_j^*| = |\delta_j^* - \delta_j + \delta_j| \leq |\delta_j^* - \delta_j| + |\delta_j|$$

By the condition $|\boldsymbol{\delta}_{\mathcal{A}}|_{\min} > (a+1)\lambda_1$, we have

$$|\delta_j| \geq \|\boldsymbol{\delta}^*\|_{\min} - \|\boldsymbol{\delta}^* - \boldsymbol{\delta}\|_\infty \geq \|\boldsymbol{\delta}^*\|_{\min} - \|\Theta^* - \Theta\|_\infty \geq (a+1)\lambda_1 - \phi_n \geq a\lambda_1.$$

As a result, both $\rho(|\delta_j|)$ and $P^{\mathcal{O}}_{\lambda_1}(S(\boldsymbol{\delta}))$ are constants.

On conclusion, we have $Q_n^{\mathcal{O}}(T_0(\boldsymbol{\alpha}), S(\boldsymbol{\delta}), \gamma) \geq Q_n^{\mathcal{O}}(\widehat{\boldsymbol{\tau}}^{\text{or}}, \widehat{\boldsymbol{\delta}}_1^{\text{or}}, \widehat{\gamma}^{\text{or}})$ for all $(\boldsymbol{\alpha}^\top, \boldsymbol{\delta}^\top, \gamma) \in \Psi$. In addition, $Q_n^{\mathcal{O}}(\widehat{\boldsymbol{\tau}}^{\text{or}}, \widehat{\boldsymbol{\delta}}_1^{\text{or}}, \widehat{\gamma}^{or}) = Q_n(\widehat{\boldsymbol{\alpha}}^{\text{or}}, \widehat{\boldsymbol{\delta}}^{\text{or}}, \widehat{\gamma}^{\text{or}})$ and $Q_n^{\mathcal{O}}(T_0(\boldsymbol{\alpha}), S(\boldsymbol{\delta}), \gamma) = Q_n(T^{-1}(T_0(\boldsymbol{\alpha})), S^{-1}(\boldsymbol{\delta}_{\mathcal{A}}), \gamma) = Q_n(\boldsymbol{\alpha^0}, \boldsymbol{\delta^0}, \gamma)$. Hence, we get $Q_n(\boldsymbol{\alpha^0}, \boldsymbol{\delta^0}, \gamma) \geq Q_n(\widehat{\boldsymbol{\alpha}}^{\text{or}}, \widehat{\boldsymbol{\delta}}^{\text{or}}, \widehat{\gamma}^{\text{or}})$, and the result in (i) is proved.

Step (ii): First, we introduce a neighbourhood $\Psi_n = \{\boldsymbol{\alpha}, \boldsymbol{\delta} : \|((\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}^{\text{or}})^\top, (\boldsymbol{\delta} - \widehat{\boldsymbol{\delta}}^{\text{or}})^\top)^\top\|_\infty \leq t_n\}$ for a positive sequence t_n. For $(\boldsymbol{\alpha}^\top, \boldsymbol{\delta}^\top, \gamma) \in \Psi \cap \Psi_n$, by Taylor's expansion at $(\boldsymbol{\alpha^0}, \boldsymbol{\delta^0})$, we have

$$\begin{aligned} &Q_n(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma) - Q_n(\boldsymbol{\alpha^0}, \boldsymbol{\delta^0}, \gamma) \\ &= -\mathbf{w}(\boldsymbol{\alpha} - \boldsymbol{\alpha^0}) + \sum_{i=1}^n \frac{\partial P_{\lambda_2}(\boldsymbol{\alpha}^m)}{\partial \alpha_i}(\alpha_i - \alpha_i^0) - \mathbf{v}(\boldsymbol{\delta} - \boldsymbol{\delta^0}) + \sum_{j=1}^p \frac{\partial P_{\lambda_1}(\boldsymbol{\delta}^m)}{\partial \delta_j}(\delta_j - \delta_j^0) \\ &= \Gamma_1 + \Gamma_2 + \Gamma_3 + \Gamma_4, \end{aligned}$$

where $\mathbf{w} = [\frac{1}{n_1}(\gamma \mathbf{y}_1 - \boldsymbol{\alpha}_1^m - \mathbf{X}_1\boldsymbol{\delta})^\top, -\frac{1}{n_0}g(-\boldsymbol{\alpha}_0^m - \mathbf{X}_0\boldsymbol{\delta})^\top]^\top = (\frac{1}{n_1}\Lambda_1^\top, -\frac{1}{n_0}\Lambda_2^\top)^\top$ and $\mathbf{v} = \frac{1}{n_1}\mathbf{X}_1^\top(\gamma \mathbf{y}_1 - \boldsymbol{\alpha}_1^m - \mathbf{X}_1\boldsymbol{\delta}^m) - \frac{1}{n_0}\mathbf{X}_0^\top g(-\boldsymbol{\alpha}_0^m - \mathbf{X}_0^\top\boldsymbol{\delta}^m) = \frac{1}{n_1}\mathbf{X}_1^\top\Lambda_1 + \frac{1}{n_0}\mathbf{X}_0^\top\Lambda_2$ in which $\boldsymbol{\alpha}^m = \zeta_1\boldsymbol{\alpha} + (1-\zeta_1)\boldsymbol{\alpha^0}$ and $\boldsymbol{\delta}^m = \zeta_2\boldsymbol{\delta} + (1-\zeta_2)\boldsymbol{\delta^0}$ for some $\zeta_1, \zeta_2 \in (0,1)$. Firstly,

$$\begin{aligned} \Gamma_1 = -\mathbf{w}^\top(\boldsymbol{\alpha} - \boldsymbol{\alpha^0}) &= -\sum_{k=1}^K \sum_{\{i,j\in\mathcal{G}_k\}} \frac{w_i(\alpha_i - \alpha_j)}{|\mathcal{G}_k|} \\ &= -\sum_{k=1}^K \sum_{\{i,j\in\mathcal{G}_k\}} \frac{w_i(\alpha_i - \alpha_j)}{2|\mathcal{G}_k|} - \sum_{k=1}^K \sum_{i,j\in\mathcal{G}_k} \frac{w_i(\alpha_i - \alpha_j)}{2|\mathcal{G}_k|} \\ &= -\sum_{k=1}^K \sum_{\{i,j\in\mathcal{G}_k\}} \frac{(w_j - w_i)(\alpha_j - \alpha_i)}{2|\mathcal{G}_k|} \\ &= -\sum_{k=1}^K \sum_{\{i,j\in\mathcal{G}_k, i<j\}} \frac{(w_j - w_i)(\alpha_j - \alpha_i)}{|\mathcal{G}_k|}. \end{aligned} \tag{A5}$$

As shown in (A4),

$$\|\boldsymbol{\alpha}^0 - \boldsymbol{\alpha}^*\|_\infty = \|\boldsymbol{\tau} - \boldsymbol{\tau}^*\|_\infty \leq \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_\infty. \tag{A6}$$

Since $\boldsymbol{\alpha}^m = \zeta_1\boldsymbol{\alpha} + (1-\zeta_1)\boldsymbol{\alpha}^0$,

$$\|\boldsymbol{\alpha}^m - \boldsymbol{\alpha}^*\|_\infty \leq \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_\infty \leq \phi_n. \tag{A7}$$

As the same steps in (A4) and (A6),

$$\|\boldsymbol{\delta^0} - \boldsymbol{\delta}^*\|_\infty = \|\boldsymbol{\delta_1} - \boldsymbol{\delta_1}^*\|_\infty \leq \|\boldsymbol{\delta} - \boldsymbol{\delta}^*\|_\infty. \tag{A8}$$

Since $\boldsymbol{\delta}^m = \zeta_2\boldsymbol{\delta} + (1-\zeta_2)\boldsymbol{\delta^0}$,

$$\|\boldsymbol{\delta}^m - \boldsymbol{\delta}^*\|_\infty \leq \|\boldsymbol{\delta} - \boldsymbol{\delta}^*\|_\infty \leq \phi_n. \tag{A9}$$

Then by Condition (C1),

$$\begin{aligned}\Lambda_1 &= \frac{\gamma}{\gamma^*}(\boldsymbol{\alpha}_1^* + \mathbf{X}_1\boldsymbol{\delta}^* + \boldsymbol{\varepsilon}_1\gamma^*) - \mathbf{X}_1\boldsymbol{\delta}^m - \boldsymbol{\alpha}_1^m \\ &= \frac{\gamma-\gamma^*}{\gamma^*}(\boldsymbol{\alpha}_1^* + \mathbf{X}_1\boldsymbol{\delta}^*) + \boldsymbol{\alpha}_1^* - \boldsymbol{\alpha}_1^m + \mathbf{X}_1(\boldsymbol{\delta}^* - \boldsymbol{\delta}^m) + (\gamma - \gamma^* + \gamma^*)\boldsymbol{\varepsilon}_1, \\ \|\Lambda_1\|_\infty &\le \frac{\|\gamma-\gamma^*\|_\infty}{\|\gamma^*\|_\infty}(\|\boldsymbol{\alpha}_1^*\|_\infty + \|\mathbf{X}_1\boldsymbol{\delta}^*\|_\infty) + \|\boldsymbol{\alpha}_1^* - \boldsymbol{\alpha}_1^m\|_\infty \\ &\quad + \|\mathbf{X}_1(\boldsymbol{\delta}^* - \boldsymbol{\delta}^m)\|_\infty + (\|\gamma-\gamma^*\|_\infty + \|\gamma^*\|_\infty)\|\boldsymbol{\varepsilon}_1\|_\infty \\ &\le \frac{\phi_n}{C_0}(C_3\sqrt{n_1} + C_1 s\cdot C_2\sqrt{s}) + \phi_n + C_1 s\cdot\phi_n + (\phi_n + C_0)\|\boldsymbol{\varepsilon}_1\|_\infty.\end{aligned}$$

Consider the ith element of the vector in Λ_2, define a positive constant ξ_i satisfying

$$\begin{aligned}g_i(-\mathbf{X}_0\boldsymbol{\delta}^m - \boldsymbol{\alpha}_0^m) &\le |-\mathbf{x}_i^\top\boldsymbol{\delta}^m - \alpha_i^m| + \xi_i \\ &\le |-\mathbf{x}_i^\top(\boldsymbol{\delta}^m - \boldsymbol{\delta}^*)| + |\alpha_i^m - \alpha_i^*| + |-\mathbf{x}_i^\top\boldsymbol{\delta}^* - \alpha_i^*| + \xi_i.\end{aligned}$$

Define $\xi = \max\{\xi_1, \dots m\xi_n\}$, so we have

$$\begin{aligned}\|\Lambda_2\|_\infty &\le \|-\mathbf{X}_0(\boldsymbol{\delta}^m - \boldsymbol{\delta}^*)\|_\infty + \|\boldsymbol{\alpha}_0^m - \boldsymbol{\alpha}_0^*\|_\infty + \|\mathbf{X}_0\boldsymbol{\delta}^* + \boldsymbol{\alpha}_0^*\|_\infty + \xi \\ &\le \|\mathbf{X}_0\|_\infty\|\boldsymbol{\delta}^m - \boldsymbol{\delta}^*\|_\infty + \|\boldsymbol{\alpha}_0^m - \boldsymbol{\alpha}_0^*\|_\infty + \|\mathbf{X}_0\|_\infty\|\boldsymbol{\delta}^*\|_\infty + \|\boldsymbol{\alpha}_0^*\|_\infty + \xi \\ &\le C_1 s\cdot\phi_n + \phi_n + C_1 s\cdot C_2\sqrt{s} + C_3\sqrt{n_0} + \xi.\end{aligned}$$

Therefore,

$$\max_{i,j}|\mathbf{w}_j - \mathbf{w}_i| \le 2\|\mathbf{w}\|_\infty \le 2\max\left\{\frac{\|\Lambda_1\|_\infty}{n_1}, \frac{\|\Lambda_2\|_\infty}{n_0}\right\}.$$

By Condition (C3), set a constant c_1

$$P\left(\|\boldsymbol{\varepsilon}_1\|_\infty > \sqrt{\frac{\log n}{c_1}}\right) \le \sum_{i=1}^{n_1} P\left(|\varepsilon_i| > \sqrt{\frac{\log n}{c_1}}\right) \le \frac{n_1}{n\sqrt{\log n}}.$$

Thus there is an event E_2 such that $P(E_2^c) \le \frac{n_1}{n\sqrt{\log n}}$, and on the event E_2,

$$\begin{aligned}&|\mathcal{G}_{\min}|^{-1}\max_{i,j}|w_j - w_i| \\ &\le 2|\mathcal{G}_{\min}|^{-1}\max\left\{\frac{\phi_n}{n_1}\left(C_1'\sqrt{n_1} + C_2' s^{\frac{3}{2}} + C_3' s + C_4'\sqrt{\log n}\right) + C_5'\frac{\sqrt{\log n}}{n_1},\right. \\ &\left.\frac{1}{n_0}\left(C_6'\sqrt{n_0} + C_7' s^{\frac{3}{2}} + C_8'\phi_n s + C_9'\phi + \xi\right)\right\}.\end{aligned}$$

Under the conditon (C4), it is easy to get $\lambda_2 \gg |\mathcal{G}_{\min}|^{-1}\max(\frac{s^{\frac{3}{2}}}{n_0}, \frac{1}{\sqrt{n_0}}, \sqrt{\frac{\log n}{n_1}})$, and hence

$$\lambda_2 \gg |\mathcal{G}_{\min}|^{-1}\max_{i,j}|w_j - w_i|. \tag{A10}$$

Then, denote $\bar{\rho}(t) = \rho'(|t|)\mathrm{sgn}(t)$,

$$\begin{aligned}\Gamma_2 &= \lambda_2\sum_{i=1}^{n}\sum_{j\ne i}\bar{\rho}(\alpha_i^m - \alpha_j^m)(\alpha_i - \alpha_i^0) \\ &= \lambda_2\sum_{i<j}\bar{\rho}(\alpha_i^m - \alpha_j^m)(\alpha_i - \alpha_i^0) + \lambda_2\sum_{i>j}\bar{\rho}(\alpha_i^m - \alpha_j^m)(\alpha_i - \alpha_i^0).\end{aligned}$$

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

$$\begin{aligned}\Gamma_2 &= \lambda_2\sum_{i<j}\bar{\rho}(\alpha_i^m - \alpha_j^m)(\alpha_i - \alpha_i^0) + \lambda_2\sum_{j>i}\bar{\rho}(\alpha_j^m - \alpha_i^m)(\alpha_j - \alpha_j^0) \\ &= \lambda_2\sum_{i<j}\bar{\rho}(\alpha_i^m - \alpha_j^m)(\alpha_i - \alpha_i^0) - \lambda_2\sum_{i<j}\bar{\rho}(\alpha_i^m - \alpha_j^m)(\alpha_j - \alpha_j^0) \\ &= \lambda_2\sum_{i<j}\bar{\rho}(\alpha_i^m - \alpha_j^m)\{(\alpha_i - \alpha_i^0) - (\alpha_j - \alpha_j^0)\}.\end{aligned} \tag{A11}$$

When $i,j \in \mathcal{G}_k, \alpha_i^0 = \alpha_j^0$, and $\alpha_i^m - \alpha_j^m$ has the same sign as $\alpha_i - \alpha_j$, and hence

$$\Gamma_2 = \lambda_2 \sum_{i=1}^{K} \sum_{i,j\in\mathcal{G}_k, i<j} \rho'(|\alpha_i^m - \alpha_j^m|)|\alpha_i - \alpha_j| + \lambda_2 \sum_{k<k'} \sum_{i\in\mathcal{G}_k, j\in\mathcal{G}_{k'}} \bar{\rho}(\alpha_i^m - \alpha_j^m)\{(\alpha_i - \alpha_i^0) - (\alpha_j - \alpha_j^0)\}.$$

Then, for $k \neq k', i \in \mathcal{G}_k, j \in \mathcal{G}_{k'}$, since

$$|\alpha_i^* - \alpha_j^*| \leq |\alpha_i^m - \alpha_j^m| + |\alpha_i^* - \alpha_i^m| + |\alpha_j^m - \alpha_j^*|,$$

we have

$$\begin{aligned} |\alpha_i^m - \alpha_j^m| &\geq \min_{i\in\mathcal{G}_k, j\in\mathcal{G}_{k'}} |\alpha_i^* - \alpha_j^*| - 2\|\boldsymbol{\alpha}^m - \boldsymbol{\alpha}^*\|_\infty \\ &\geq b_n - 2\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_\infty \geq b_n - 2\phi_n \geq a\lambda_2, \end{aligned}$$

and thus $\bar{\rho}(\alpha_i^m - \alpha_j^m) = 0$. Therefore,

$$\Gamma_2 = \lambda_2 \sum_{i=1}^{K} \sum_{i,j\in\mathcal{G}_k, i<j} \rho'(|\alpha_i^m - \alpha_j^m|)|\alpha_i - \alpha_j|. \tag{A12}$$

Moreover, by the same reasoning as (A4), for $i,j \in \mathcal{G}$ we have

$$\|\boldsymbol{\alpha}^{\mathbf{0}} - \widehat{\boldsymbol{\alpha}}^{\mathrm{or}}\|_\infty \leq \|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}^{\mathrm{or}}\|_\infty.$$

Then

$$\begin{aligned} |\alpha_i^m - \alpha_j^m| &\leq |\alpha_i^m - \alpha_i^0| + |\alpha_j^m - \alpha_j^0| \\ &\leq 2\|\boldsymbol{\alpha}^m - \boldsymbol{\alpha}^0\|_\infty \leq 2\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^{\mathbf{0}}\|_\infty \\ &\leq 2(\|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}^{\mathrm{or}}\|_\infty + \|\boldsymbol{\alpha}^0 - \widehat{\boldsymbol{\alpha}}^{\mathrm{or}}\|_\infty) \\ &\leq 4\|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}^{\mathrm{or}}\|_\infty \leq 4t_n. \end{aligned} \tag{A13}$$

Since $\rho(\cdot)$ is concave, $\rho'(|\alpha_i^m - \alpha_j^m|) \geq \rho'(4t_n)$. As a result,

$$\Gamma_2 \geq \lambda_2 \sum_{k=1}^{K} \sum_{i,j\in\mathcal{G}_k, i<j} \rho'(4t_n)|\alpha_i - \alpha_j|. \tag{A14}$$

On the other hand, we have

$$\begin{aligned} \Gamma_3 &= -\mathbf{v}(\boldsymbol{\delta} - \boldsymbol{\delta}^{\mathbf{0}}) \\ &= -\left(\sum_{j\in\mathcal{A}} v_j(\delta_j - \delta_j^0) + \sum_{j\in\mathcal{A}^c} v_j(\delta_j - \delta_j^0)\right) = \sum_{j\in\mathcal{A}^c} v_j\delta_j. \end{aligned} \tag{A15}$$

Since $\Lambda_3 = \mathbf{X}_1^\top \Lambda_1$ and $\Lambda_4 = \mathbf{X}_0^\top \Lambda_2$, on the event E_2,

$$\begin{aligned} \max|v_j| &\leq \left(\frac{\|\mathbf{X}_1\|_\infty}{n_1}\|\Lambda_1\|_\infty + \frac{\|\mathbf{X}_0\|_\infty}{n_0}\|\Lambda_2\|_\infty\right) \\ &\leq C_1 s\left(\frac{1}{n_1}\|\Lambda_1\|_\infty + \frac{1}{n_0}\|\Lambda_2\|_\infty\right) \\ &\leq C_1 s\left\{\frac{\phi_n}{n_1}\left(C_1'\sqrt{n_1} + C_2' s^{\frac{3}{2}} + C_3' s + C_4'\sqrt{\log n}\right) + C_5'\frac{\sqrt{\log n}}{n_1}\right. \\ &\quad \left. + \frac{1}{n_0}\left(C_6'\sqrt{n_0} + C_7' s^{\frac{3}{2}} + C_8'\phi_n s + C_9'\phi + \xi\right)\right\}. \end{aligned} \tag{A16}$$

Under the condition (C4), we can get

$$\lambda_1 \gg \max_j |\delta_j|. \tag{A17}$$

Then,

$$\Gamma_4 = \lambda_1 \sum_{j=1}^{p} \bar{\rho}(\delta_j^m)(\delta_j - \delta_j^0)$$

$$= \lambda_1 \left(\sum_{j \in \mathcal{A}} \bar{\rho}(\delta_j^m)(\delta_j - \delta_j^0) + \sum_{j \in \mathcal{A}^c} \bar{\rho}(\delta_j^m)(\delta_j - \delta_j^0) \right). \tag{A18}$$

When $j \in \mathcal{A}^c$, $\delta_j^0 = 0$, and δ_j^m has the same sign as δ_j. Hence

$$\Gamma_4 = \lambda_1 \left(\sum_{j \in \mathcal{A}'^c} \rho'(|\delta_j^m|)|\delta_j| + \sum_{j \in \mathcal{A}} \bar{\rho}(\delta_j^m)(\delta_j - \delta_j^0) \right). \tag{A19}$$

For $j \in \mathcal{A}$, by (A9),

$$|\delta_j^m| \geq \min_{j \in \mathcal{A}} |\delta_j^*| - \|\boldsymbol{\delta}^* - \boldsymbol{\delta}^m\|_\infty \geq (a+1)\lambda_1 - \phi_n \geq a\lambda_1. \tag{A20}$$

Thus $\bar{\rho}(\delta_j^m) = 0$. Therefore,

$$\Gamma_4 = \lambda_1 \sum_{j \in \mathcal{A}^c} \rho'(|\delta_j^m|)|\delta_j|. \tag{A21}$$

Furthermore, by the same process as (A13), for $j \in \mathcal{A}^c$

$$\begin{aligned} |\delta_j^m| &\leq \|\boldsymbol{\delta}^m - \boldsymbol{\delta}^{\mathbf{0}}\|_\infty \leq \|\boldsymbol{\delta} - \boldsymbol{\delta}^{\mathbf{0}}\|_\infty \\ &\leq \|\boldsymbol{\delta} - \widehat{\boldsymbol{\delta}}^{\text{or}}\|_\infty + \|\boldsymbol{\delta}^{\mathbf{0}} - \widehat{\boldsymbol{\delta}}^{\text{or}}\|_\infty \\ &\leq 2\|\boldsymbol{\delta} - \widehat{\boldsymbol{\delta}}^{\text{or}}\|_\infty \leq 2t_n. \end{aligned} \tag{A22}$$

Let $t_n = o(1)$. Then $\rho'(4t_n) \to 1$, $\rho'(2t_n) \to 1$. Therefore, by (A5), (A10) and (A14),

$$\Gamma_1 + \Gamma_2 \geq \sum_{k=1}^{K} \sum_{i,j \in \mathcal{G}_k, i<j} \left[\lambda_2 \rho'(4t_n) - |\mathcal{G}_{\min}|^{-1} \max_{i,j} |w_j - w_i| \right] |\alpha_i - \alpha_j| \geq 0. \tag{A23}$$

And by (A15), (A17) and (A21),

$$\Gamma_3 + \Gamma_4 \geq \sum_{j \in \mathcal{A}^c} \left[\lambda_1 \rho'(2t_n) - \max_j |v_j| \right] |\delta_j| \geq 0. \tag{A24}$$

Therefore, for sufficiently large n,

$$Q_n(\boldsymbol{\alpha}, \boldsymbol{\delta}, \gamma) - Q_n(\boldsymbol{\alpha}^{\mathbf{0}}, \boldsymbol{\delta}^{\mathbf{0}}, \gamma) = \Gamma_1 + \Gamma_2 + \Gamma_3 + \Gamma_4 \geq 0,$$

so that the result (ii) is proved.