Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Subgroup Analysis in the Heterogeneous Cox Model

Xiangbin Hu[1] | Jian Huang[2] | Li Liu[3] | Defeng Sun[1] | Xingqiu Zhao*[1]

[1]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China
[2]School of Mathematics and Statistics, Wuhan University, Wuhan, China
[3]Department of Statistics and Actuarial Science, University of Iowa, Iowa, U.S.A.

Correspondence
*Xingqiu Zhao, Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China. Email: xingqiu.zhao@polyu.edu.hk

Abstract

In the analysis of censored survival data, to avoid a biased inference of treatment effects on the hazard function of the survival time, it is important to consider the treatment heterogeneity. Without requiring any prior knowledge about the subgroup structure, we propose a data driven subgroup analysis procedure for the heterogenous Cox model by constructing a pairwise fusion penalized partial likelihood-based objective function. The proposed method can determine the number of subgroups, identify the group structure and estimate the treatment effect simultaneously and automatically. A majorized alternating direction method of multipliers algorithm is then developed to deal with the numerically challenging high-dimensional problems. We also establish the oracle properties and the model selection consistency for the proposed penalized estimator. Our proposed method is evaluated by simulation studies and further illustrated by the analysis of the breast cancer data.

KEYWORDS:
Subgroup analysis; Cox model; Treatment heterogeneity; Majorized ADMM algorithm; Oracle property

1 | INTRODUCTION

In the study of survival analysis, one of the main purposes is to estimate the covariate effects on survival times. Various important topics on classical Cox's proportional hazards model have been widely discussed by specifying that the covariates have log-linear effects on the hazard function of survival time. For example, Fan and Li,[1] Zhang and Lu,[2] and Zhao et al[3] developed variable selection approaches for the Cox model. Bradic et al,[4] Huang et al,[5] and Fang et al[6] investigated the asymptotic properties of the penalized partial likelihood estimators for high dimensional Cox models. Chen et al,[7] Qu et al,[8] and Kong et al[9] studied functional Cox regression models. All these studies are based on the assumption that the covariate effects possess homogeneity.

In clinical medicine applications, treatment effects are usually heterogenous, i.e., the same treatment may result in different effects over different groups of patients with similar characteristics. In these situations, the homogeneous assumption in classical model would lead to biased estimates. Thus, identifying the group-specific treatment effect is the key in the process of precision medicine treatment. Some subgroup analysis methods have been developed. Among others, Kravitz et al,[10] Rothwell,[11] and Lagakos[12] used descriptive statistics to analyze heterogenous experimental data; Wei and Kosorok,[13] Shen and He,[14] and Wu et al[15] studied the problem of treatment heterogeneity based on the finite mixture models, such as Gaussian mixture model, logistic-normal mixture model, and logistic-Cox mixture model. Recently, Ma and Huang[16] and Ma et al[17] developed a kind of regularization method to identify the grouping structure and estimate the treatment effect simultaneously based on a data driven process. Furthermore, Zhang et al[18] extended the regularization method to the quantile regression model and proposed a robust subgroup identification method. Chen et al[19] utilized this method to analyze the zero-inflated Poisson regression model.

The works mentioned above focus on complete observation data except Wu et al.[15] For censored survival data, the incomplete data information and complexity of survival models bring challenges for subgroup analysis. In this paper, we consider the

subgroup analysis in the heterogenous Cox model under the assumption of sparsity subgroup structure. Based on the objective function constructed through combining the negative logarithmic partial likelihood function and a concave fusion penalty function, we can identify the subgroup structure and estimate treatment effects simultaneously without any prior knowledge about the group structure. The likelihood-based regularization approaches make the statistical inference of identifying the subgroup structure and estimating treatment effects become an automated procedure and so it is easy to implement.

To overcome the computational difficulties caused from the complicated nature of the likelihood-based objective function, we borrow the ideas of the majorized alternating direction method of multipliers (ADMM) algorithm.[20] Compared to the classical ADMM algorithm suggested by Ma and Huang,[16] this algorithm is able to efficiently handle large scale problems to get more accurate solutions by transforming an objective function into a majorized convex function with a pairwise fusion penalty. We take the ridge solution of the negative log-likelihood function as the initial solution of the algorithm, and find that the initial solution performs well in identifying the subgroup structure in our simulation studies.

Using the oracle estimator as a bridge, we obtain the oracle property of the proposed estimator. Concretely, we obtain the consistency and asymptotic normality of the oracle estimator at first. Then we show that the oracle estimator and the proposed estimator are asymptotically equivalent. Thus, the latter is consistent and possesses the asymptotic normality. This property also illustrates that the proposed method can identify the subgroup structure of the model as if we knew it in advance.

The rest of this article is organized as follows. In Section 2, we introduce the heterogenous Cox model with right censored data and propose a penalized estimation approach. Section 3 presents the majorized ADMM algorithm for computing the proposed estimators. In Section 4, we establish the consistency and the asymptotic normality of the proposed estimator. We then conduct simulation studies to demonstrate the performance of the proposed method in Section 5, and use the method to analyze a real data example in Section 6. Section 7 provides some concluding remarks. The proofs of the theoretical results are relegated to the Appendix.

2 | HETEROGENOUS COX MODEL AND ESTIMATION PROCEDURE

Consider a survival study containing n independent subjects. For subject i, let U_i and C_i denote the failure time and the censoring time, respectively. Then the observed data consist of $\{(T_i, \Delta_i) : i = 1, \cdots, n\}$, where $T_i = U_i \wedge C_i$ and $\Delta_i = 1_{\{U_i \le C_i\}}$. Let X_i and Z_i denote covariates with dimensions p and q, respectively. Let $\lambda(t|X_i, Z_i)$ be the conditional hazard rate function of U given X_i and Z_i. Then the homogeneous Cox model is

$$\lambda(t|X_i, Z_i) = \lambda_0(t)\exp(Z_i^T\eta + X_i^T\beta), i = 1, \cdots, n, \tag{1}$$

where $\lambda_0(t)$ is the baseline hazard function, η and β are unknown regression parameters denoting the average effects. However, the homogeneous assumption about covariate effects is not satisfied when the effects of X_i are different among subjects. To describe the treatment heterogeneity, we propose the heterogeneous Cox model as follows:

$$\lambda(t|X_i, Z_i) = \lambda_0(t)\exp(Z_i^T\eta + X_i^T\beta_i), i = 1, \cdots, n, \tag{2}$$

where β_i is subject-specific effect of X_i on the hazard function. We suppose that n subjects are divided into K potential subgroups according to set $\mathcal{G} = (\mathcal{G}_1, \ldots, \mathcal{G}_K)$, and $\beta_i \equiv \alpha_k$ for all $i \in \mathcal{G}_k$, $k = 1, \cdots, K$. For this model, we focus on identifying the subgroup set $\mathcal{G}$ and estimating parameters $\{\alpha_1, \cdots, \alpha_K\}$ and η.

For the coefficient of X, define $\boldsymbol{\alpha} = (\alpha_1^T, \ldots, \alpha_K^T)^T$ and $\boldsymbol{\beta} = (\beta_1^T, \ldots, \beta_n^T)^T$. The negative partial log-likelihood function is

$$\ell_n(\eta, \boldsymbol{\beta}) = -\sum_{i=1}^{n}\Delta_i(Z_i^T\eta + X_i^T\beta_i) + \sum_{i=1}^{n}\Delta_i \log\Big(\sum_{j\in R(T_i)}\exp(Z_j^T\eta + X_j^T\beta_j)\Big), \tag{3}$$

where $R(T_i) = \{j : T_j \ge T_i\}$ is the risk set. For the purpose of identifying the subgroup structure, we use a concave pairwise penalty $p_\gamma(||\beta_i - \beta_j||, \lambda)$ to shrink small value of $||\beta_i - \beta_j||$ to 0, where $||\cdot||$ is the L_2-norm of a vector. Then the criterion function is

$$Q_n(\eta, \boldsymbol{\beta}) = \ell_n(\eta, \boldsymbol{\beta}) + \sum_{i<j} p_\gamma(||\beta_i - \beta_j||, \lambda), \tag{4}$$

where $\lambda \ge 0$ is a tuning parameter. Thus, we can obtain the estimator $(\hat{\eta}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda))$ by minimizing the objective function (4) with a given turning parameter λ. Finally, the estimator for $\boldsymbol{\alpha}$ is the distinct value of $\widehat{\boldsymbol{\beta}}(\lambda)$, denoted by $\widehat{\boldsymbol{\alpha}}(\lambda) = (\hat{\alpha}_1^T(\lambda), \ldots, \hat{\alpha}_{\widehat{K}}^T(\lambda))^T$. The identified subgroup structure is $\widehat{\mathcal{G}}_k(\lambda) = \{i : \hat{\beta}_i(\lambda) = \hat{\alpha}_k(\lambda), 1 \le i \le n\}$, where $1 \le k \le \widehat{K}(\lambda)$.

The penalty function can be naively chosen as the L_1 penalty function $p_\gamma(t,\lambda) = \lambda|t|$, but L_1 penalty tends to choose too many subgroups. Following Ma and Huang,[16] a better choice of the penalty function is the smoothly clipped absolute deviation (SCAD)[21] with

$$p_\gamma(t,\lambda) = \lambda \int_0^{|t|} \min\{1, (\gamma - x/\lambda)_+/(\gamma-1)\}dx,$$

or the minimax concave penalty (MCP)[22] with

$$p_\gamma(t,\lambda) = \lambda \int_0^{|t|} (1 - x/(\gamma\lambda))_+ dx.$$

3 | MAJORIZED ADMM ALGORITHM

In this section, we present the algorithm to find the solution path $(\hat{\eta}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda))$. Introducing a new set of parameters $u_{ij} = \beta_i - \beta_j$, we can reformulate the criterion function $Q_n(\eta, \boldsymbol{\beta})$ as

$$Q_n(\eta, \boldsymbol{\beta}, \mathbf{u}) = \ell_n(\eta, \boldsymbol{\beta}) + \sum_{i<j} p_\gamma(||u_{ij}||, \lambda)$$

subject to $\beta_i - \beta_j - u_{ij} = 0$, where $\mathbf{u} = (u_{ij}^T, i<j)^T$. Following Ma et al,[17] we can solve this minimization problem using the standard ADMM algorithm by approximating $\ell_n(\eta, \boldsymbol{\beta})$ as the quadratic function

$$\begin{aligned}\ell_n(\eta,\boldsymbol{\beta}) \approx & \ell_n(\eta^{(m-1)}, \boldsymbol{\beta}^{(m-1)}) + \nabla\ell_n(\eta^{(m-1)}, \boldsymbol{\beta}^{(m-1)})^T\big((\eta,\boldsymbol{\beta}) - (\eta^{(m-1)}, \boldsymbol{\beta}^{(m-1)})\big) \\ & + \big((\eta,\boldsymbol{\beta})^T - (\eta^{(m-1)}, \boldsymbol{\beta}^{(m-1)})^T\big)\nabla^2\ell_n(\eta^{(m-1)}, \boldsymbol{\beta}^{(m-1)})\big((\eta,\boldsymbol{\beta}) - (\eta^{(m-1)}, \boldsymbol{\beta}^{(m-1)})\big),\end{aligned}$$

where $(\eta^{(m-1)}, \boldsymbol{\beta}^{(m-1)})$ is the value of parameter in the mth iteration step. However, the quadratic approximation is only accurate when $(\eta, \boldsymbol{\beta})$ is close to $(\eta^{(m-1)}, \boldsymbol{\beta}^{(m-1)})$, and the calculation of the second order derivative $\nabla^2\ell_n(\eta, \boldsymbol{\beta})$ is time consuming. Hence, it motivates us to utilize the idea of the majorized ADMM algorithm.[20]

Introduce another set of parameters $Y_i = Z_i^T\eta + X_i^T\beta_i$, and let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$. The negative log partial-likelihood function $l_n(\eta, \boldsymbol{\beta})$ can be rewritten as

$$g(\mathbf{Y}) = -\sum_{i=1}^n \Delta_i Y_i + \sum_{i=1}^n \Delta_i \log\Big(\sum_{j\in R(T_i)} \exp(Y_j)\Big).$$

Then we need to minimize

$$Q_n(\eta, \boldsymbol{\beta}, \mathbf{u}, \mathbf{Y}) = g(\mathbf{Y}) + \sum_{i<j} p_\gamma(||u_{ij}||, \lambda) \tag{5}$$

subject to $\beta_i - \beta_j - u_{ij} = 0$ and $Y_i = Z_i^T\eta + X_i^T\beta_i$. Since $\nabla^2 g(\mathbf{Y}) \preceq \widetilde{G}$ for $\widetilde{G} = \frac{1}{2}\mathrm{diag}\{\tilde{g}_1, \ldots, \tilde{g}_n\}$ and $\tilde{g}_j = \sum_{i=1}^n \Delta_i I_{j\in R(T_i)}$, we have

$$g(\mathbf{Y}) \le \tilde{g}(\mathbf{Y}; \mathbf{Y}') := g(\mathbf{Y}') + \langle \mathbf{Y} - \mathbf{Y}', \nabla g(\mathbf{Y}')\rangle + \frac{1}{2}\|\mathbf{Y} - \mathbf{Y}'\|_{\widetilde{G}}^2$$

for any $\mathbf{Y}$ and $\mathbf{Y}'$ with $\|\mathbf{x}\|_{\widetilde{G}}^2 = \langle \mathbf{x}, \widetilde{G}\mathbf{x}\rangle$. The objective function (5) is then tansformed to the majorized augmented Lagrangian function as follows

$$\begin{aligned}Q_n'(\eta, \boldsymbol{\beta}, \mathbf{Y}, \mathbf{u}; \mathbf{w}, \mathbf{v}, \mathbf{Y}') = & \tilde{g}(\mathbf{Y}; \mathbf{Y}') + \sum_{i<j} p_\gamma(\|u_{ij}\|, \lambda) + \sum_{i=1}^n \langle w_i, Y_i - Z_i^T\eta - X_i^T\beta_i\rangle \\ & + \sum_{i<j}\langle v_{ij}, \beta_i - \beta_j - u_{ij}\rangle + \frac{\vartheta}{2}\sum_{i=1}^n (Y_i - Z_i^T\eta - X_i^T\beta_i)^2 + \frac{\vartheta}{2}\sum_{i<j}||\beta_i - \beta_j - u_{ij}||^2,\end{aligned}$$

where the dual variables $\mathbf{w} = (w_i, i = 1, \ldots, n)^T$ and $\mathbf{v} = (v_{ij}^T, i<j)^T$ are the Lagrange multipliers, and ϑ is the penalty parameter. We then compute the estimators $\widehat{\boldsymbol{\beta}}$ and $\hat{\eta}$ through the following majorized ADMM algorithm.

At the mth iteration, for a given value of parameter $(\eta^{(m-1)}, \beta^{(m-1)}, \mathbf{Y}^{(m-1)}, \mathbf{u}^{(m-1)}; \mathbf{w}^{(m-1)}, \mathbf{v}^{(m-1)}, \mathbf{Y}'^{(m-1)})$, cluster size $K^{(m-1)}$, and subgroup set $\mathcal{G}^{(m-1)}$, the iteration goes as follows:

Step 1. Update $(\eta^{(m)}, \boldsymbol{\beta}^{(m)})$ by minimizing

$$Q'_n(\eta, \boldsymbol{\beta}, \mathbf{Y}^{(m-1)}, \mathbf{u}^{(m-1)}; \mathbf{w}^{(m-1)}, \mathbf{v}^{(m-1)}, \mathbf{Y}'^{(m-1)});$$

Step 2. Update $(\mathbf{Y}^{(m)}, \mathbf{u}^{(m)})$ by minimizing

$$Q'_n(\eta^{(m)}, \boldsymbol{\beta}^{(m)}, \mathbf{Y}, \mathbf{u}; \mathbf{w}^{(m-1)}, \mathbf{v}^{(m-1)}, \mathbf{Y}'^{(m-1)})$$

and update

$$Y_i'^{(m)} = Z_i^T \eta^{(m)} + X_i^T \beta_i^{(m)} \tag{6}$$

for $i = 1, \ldots, n$;

Step 3. Update $\mathbf{w}^{(m)}$ and $\mathbf{v}^{(m)}$ by

$$\begin{aligned} w_i^{(m)} =& w_i^{(m-1)} + \varrho\vartheta(Y_i^{(m)} - Z_i^T \eta^{(m)} - X_i^T \beta_i^{(m)}), \\ v_{ij}^{(m)} =& v_{ij}^{(m-1)} + \varrho\vartheta(\beta_i^{(m)} - \beta_j^{(m)} - u_{ij}^{(m)}), \end{aligned} \tag{7}$$

where the constant $\varrho \in (0, (1+\sqrt{5})/2)$;

Step 4. Update $K^{(m)}$ and $\mathcal{G}^{(m)}$ by clustering $\boldsymbol{\beta}^{(m)}$.

At Step 1, for fixed $(\mathbf{Y}, \mathbf{u}, \mathbf{w}, \mathbf{v}, \mathbf{Y}')$, it suffices to minimize the following objective function in order to update $\boldsymbol{\beta}$ and η:

$$\begin{aligned} &\sum_{i=1}^{n} \langle w_i, Y_i - Z_i^T \eta - X_i^T \beta_i \rangle + \sum_{i<j} \langle v_{ij}, \beta_i - \beta_j - u_{ij} \rangle \\ &+ \frac{\vartheta}{2} \sum_{i=1}^{n} (Y_i - Z_i^T \eta - X_i^T \beta_i)^2 + \frac{\vartheta}{2} \sum_{i<j} ||\beta_i - \beta_j - u_{ij}||^2. \end{aligned} \tag{8}$$

Define $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$, $\mathbf{X} = \mathrm{diag}(X_1^T, \ldots, X_n^T)$ and $\mathbf{A} = \mathbf{D} \otimes I_p$, where $\mathbf{D} = \{(e_i - e_j), i < j\}^T$ with e_i being an $n \times 1$ vector whose ith entry is 1 and the remaining ones are 0, I_p is a $p \times p$ identity matrix, and $\otimes$ is a Kronecker product. For given K and $\mathcal{G}$, let $\mathbf{W}_{\mathcal{G}} = \{\omega_{ik}\}$ be an $n \times K$ matrix, where the entry ω_{ik} takes 1 if $i \in \mathcal{G}_k$ and 0 otherwise. In addition, we define $\widetilde{\mathbf{W}}_{\mathcal{G}} = \mathbf{W}_{\mathcal{G}} \otimes I_p$, $\widetilde{\mathbf{X}} = \mathbf{X}\widetilde{\mathbf{W}}_{\mathcal{G}}$ and $\widetilde{\mathbf{A}} = \mathbf{A}\widetilde{\mathbf{W}}_{\mathcal{G}}$. Thus, after removing the terms irrelevant to $\boldsymbol{\beta}$ and η, the minimal point of (8) is obtained equivalently by minimizing

$$\frac{1}{2}\|\mathbf{Y} - \mathbf{Z}\eta - \widetilde{\mathbf{X}}\boldsymbol{\alpha} + \frac{\mathbf{w}}{\vartheta}\|^2 + \frac{1}{2}\|\widetilde{\mathbf{A}}\boldsymbol{\alpha} - \mathbf{u} + \frac{\mathbf{v}}{\vartheta}\|^2.$$

At the mth iteration, the parameters $\boldsymbol{\beta}$ and η are updated through the following equations

$$\begin{aligned} \boldsymbol{\alpha}^{(m)} =& \mathbf{H}_{\mathcal{G}}^{-1} \mathbf{S}_{\mathcal{G}}^{(m-1)}, \\ \boldsymbol{\beta}^{(m)} =& \widetilde{\mathbf{W}}_{\mathcal{G}} \boldsymbol{\alpha}^{(m)}, \\ \eta^{(m)} =& (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{Y}^{(m-1)} - \mathbf{X}\boldsymbol{\beta}^{(m)} + \vartheta^{-1}\mathbf{w}^{(m-1)}), \end{aligned} \tag{9}$$

where $\mathbf{H}_{\mathcal{G}} = \widetilde{\mathbf{X}}^T\mathbf{Q}_Z\widetilde{\mathbf{X}} + \widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}}$, and $\mathbf{S}_{\mathcal{G}}^{(m-1)} = \widetilde{\mathbf{X}}^T\mathbf{Q}_Z(\mathbf{Y}^{(m-1)} + \vartheta^{-1}\mathbf{w}^{(m-1)}) + \widetilde{\mathbf{A}}^T(\mathbf{u}^{(m-1)} - \vartheta^{-1}\mathbf{v}^{(m-1)})$. It deserves to note that the updated solution of parameter $\boldsymbol{\beta}^{(m)}$ includes the integrated information of $\boldsymbol{\alpha}^{(m)}$, $\mathcal{G}^{(m-1)}$ and $K^{(m-1)}$.

At Step 2, for fixed $(\eta, \beta, \mathbf{w}, \mathbf{v}, \mathbf{Y}')$, we need to get the minimal points

$$\begin{aligned} \arg\min_{\mathbf{Y}} \quad & \langle \mathbf{Y}, \nabla g(\mathbf{Y}') \rangle + \frac{1}{2}\|\mathbf{Y} - \mathbf{Y}'\|_{\widetilde{G}}^2 \\ & + \sum_{i=1}^{n} \langle w_i, Y_i - Z_i^T\eta - X_i^T\beta_i \rangle + \frac{\vartheta}{2}\sum_{i=1}^{n}(Y_i - Z_i^T\eta - X_i^T\beta_i)^2, \end{aligned} \tag{10}$$

$$\arg\min_{u_{ij}} \quad \frac{1}{2}\|\beta_i - \beta_j + \frac{v_{ij}}{\vartheta} - u_{ij}\|^2 + \frac{1}{\vartheta}p_\gamma(\|u_{ij}\|, \lambda). \tag{11}$$

At the mth iteration, for (10), it can be solved that for $i = 1, \ldots, n$,

$$Y_i^{(m)} = (\tilde{g}_i + \vartheta)^{-1}\Big[-\nabla_i g(\mathbf{Y}'^{(m-1)}) + \tilde{g}_i Y_i'^{(m-1)} - w_i^{(m-1)} + \vartheta(Z_i^T\eta^{(m)} + X_i^T\beta_i^{(m)}) \Big]. \tag{12}$$

For (11), we can get the closed form of $u_{ij}^{(m)}$ for some commonly used penalties, such as group MCP and group SCAD. For the group SCAD penalty with parameter γ, i.e.,

$$p'_\gamma(\|u_{ij}\|, \lambda) = \lambda I(\|u_{ij}\| \le \lambda) + \frac{(\gamma\lambda - \|u_{ij}\|)_+}{\gamma - 1} I(\|u_{ij}\| > \lambda),$$

we have

$$u_{ij}^{(m)} = \begin{cases} S(c_{ij}^{(m-1)}; \lambda/\vartheta), & \|c_{ij}^{(m-1)}\| \le \lambda + \lambda/\vartheta, \\ \frac{\left(\vartheta(\gamma-1)-\lambda\gamma/\|c_{ij}^{(m-1)}\|\right)c_{ij}^{(m-1)}}{\vartheta\gamma-\vartheta-1}, & \lambda+\lambda/\vartheta < \|c_{ij}^{(m-1)}\| \le \lambda\gamma, \\ c_{ij}^{(m-1)}, & \|c_{ij}^{(m-1)}\| > \lambda\gamma, \end{cases} \tag{13}$$

where $c_{ij}^{(m-1)} = \beta_i^{(m)} - \beta_j^{(m)} + \frac{v_{ij}^{(m-1)}}{\vartheta}$ and $S(c;\lambda) = (1-\lambda/\|c\|)_+ c$. For group MCP penalty with parameter γ, i.e.,

$$p'_\gamma(\|u_{ij}\|, \lambda) = \frac{(\gamma\lambda - \|u_{ij}\|)_+}{\gamma},$$

we have

$$u_{ij}^{(m)} = \begin{cases} S\left(\frac{\vartheta c_{ij}^{(m-1)}}{\vartheta - 1/\gamma}; \frac{\lambda}{\vartheta-1/\gamma}\right), & \|c_{ij}^{(m-1)}\| \le \lambda\gamma, \\ c_{ij}^{(m-1)}, & \|c_{ij}^{(m-1)}\| > \lambda\gamma. \end{cases} \tag{14}$$

At Step 4, we first solve the following optimization problem

$$\tilde{u}_{ij}^{(m)} = \arg\min_{\tilde{u}_{ij}} \frac{1}{2}\|\beta_i^{(m)} - \beta_j^{(m)} - \tilde{u}_{ij}\|^2 + p_\gamma(\|\tilde{u}_{ij}\|, \lambda), \tag{15}$$

and then update $K^{(m)}$ and $\mathcal{G}^{(m)}$ by clustering individuals i and j into the same group if $\tilde{u}_{ij} = 0$. This step is critical to clustering analysis of the regression coefficient $\boldsymbol{\beta}$ so that Step 1 can be carried out smoothly in the recursive process. The performance of the algorithm depends on the choice of the penalty function and the tuning parameter λ.

The initial points in the algorithm are taken as follows. Since covariate Z has no subgroup effect, we simply take the estimator $\hat{\eta}$ as $\eta^{(0)}$ by treating the hazard function as a homogeneous effect model. As a reasonable initial point of parameter $\boldsymbol{\beta}$, it should reflect not only the form of the assumed hazard function but also the subgroup relation among different individuals. So we consider the ridge solution of the negative log-likelihood function as $\boldsymbol{\beta}^{(0)}$. Concretely, we define

$$\boldsymbol{\beta}^{(0)} = \arg\min_{\boldsymbol{\beta}} l_n(\eta^{(0)}, \boldsymbol{\beta}) + \frac{\lambda^*}{2}\sum_{i<j}\|\beta_i - \beta_j\|^2,$$

where tuning parameter λ^* is taken as 0.001 in our simulation studies, and utilize a majorized algorithm to find the solution of $\boldsymbol{\beta}^{(0)}$ through (5). We take $K^{(0)} = \lfloor\sqrt{n}\rfloor$ to ensure that there are enough groups at the beginning of the iteration. A cluster analysis method can then be applied to $\boldsymbol{\beta}^{(0)}$ for determining $\mathcal{G}^{(0)} = (\mathcal{G}_1^{(0)}, \ldots, \mathcal{G}_{K^{(0)}}^{(0)})$. Take $\mathbf{Y}^{(0)} = \mathbf{Y}'^{(0)} = \mathbf{Z}\eta^{(0)} + \mathbf{X}\boldsymbol{\beta}^{(0)}$, $\mathbf{u}^{(0)} = \mathbf{A}\boldsymbol{\beta}^{(0)}$ and $\mathbf{w}^{(0)} = \mathbf{v}^{(0)} = \mathbf{0}$.

Denote the primal residual as

$$r^{(m)} = \sum_{i=1}^{n}(y_i^{(m)} - \mathbf{z}_i^T\eta^{(m)} - \mathbf{x}_i^T\beta_i^{(m)})^2 + \sum_{i<j}||\beta_i^{(m)} - \beta_j^{(m)} - \mathbf{u}_{ij}^{(m)}||^2. \tag{16}$$

We stop the iteration when $r^{(m)}$ is small enough.

We summarize the above descriptions in Algorithm 1.

Algorithm 1 Majorized ADMM algorithm

Initialize $(\eta^{(0)}, \beta^{(0)}, \mathbf{Y}^{(0)}, \mathbf{u}^{(0)}; \mathbf{w}^{(0)}, \mathbf{v}^{(0)}, \mathbf{Y}'^{(0)})$, $K^{(0)}$, and $\mathcal{G}^{(0)}$
for $m = 1, 2, \cdots$ **do**
  Update $(\boldsymbol{\beta}^{(m)}, \eta^{(m)})$ using (9)
  Update $(\mathbf{Y}^{(m)}, \mathbf{u}^{(m)})$ using (12) (13), and (14)
  Update $\mathbf{Y}'^{(m)}$ using (6)
  Update $(\mathbf{w}^{(m)}, \mathbf{v}^{(m)})$ using (7)
  Compute $\tilde{u}_{ij}^{(m)}$ using (15), and update $(K^{(m)}, \mathcal{G}^{(m)})$ according to $\tilde{u}_{ij}^{(m)}$
  Compute $r^{(m)}$ using (16)
  if $r^{(m)}$ is small enough **then**
    Stop and denote the last interation by $(\hat{\boldsymbol{\beta}}, \hat{\eta})$
  end if
end for

4 | ASYMPTOTIC RESULTS

Let $N_i(t) = 1_{(T_i \le t, \Delta_i = 1)}$, $Y_i(t) = 1_{(T_i \ge t)}$, and τ be the end time of study. Suppose that $\int_0^\tau \lambda_0(t)dt < \infty$. The negative partial log-likelihood function can be rewritten as

$$\ell_n(\eta, \boldsymbol{\beta}) = -\sum_{i=1}^{n} \int_0^\tau \left[\left(Z_i^T \eta + X_i^T \beta_i \right) - \log \left\{ \sum_{j=1}^{n} Y_j(t) \exp(Z_j^T \eta + X_j^T \beta_j) \right\} \right] d N_i(t).$$

The objective function is $Q_n(\eta, \boldsymbol{\beta}) = \ell_n(\eta, \boldsymbol{\beta}) + P_n(\boldsymbol{\beta})$, where $P_n(\boldsymbol{\beta}) = \sum_{i<j} p_\gamma(||\beta_i - \beta_j||, \lambda)$. Denote the true subgroup set as $\mathcal{G}_0 = (\mathcal{G}_{0,1}, \cdots, \mathcal{G}_{0,K_0})$. Define $\widetilde{\mathbf{W}}_{\mathcal{G}_0} = \mathbf{W}_{\mathcal{G}_0} \otimes I_p$, $\widetilde{\mathbf{X}}_{\mathcal{G}_0} = \mathbf{X}\widetilde{\mathbf{W}}_{\mathcal{G}_0}$, $\mathbf{B} = (\mathbf{Z}, \widetilde{\mathbf{X}}_{\mathcal{G}_0})$, and let $\mathbf{B}_i$ be the i-th column of $\mathbf{B}^T$. Let $\boldsymbol{\theta} = (\eta^T, \boldsymbol{\alpha}^T)^T$, and $S^{(0)}(\boldsymbol{\theta}, \mathbf{B}, t) = n^{-1} \sum_{i=1}^n Y_i(t) \exp(\mathbf{B}_i^T \boldsymbol{\theta})$. Thus, with the prior information of $\mathcal{G}_0$, we write the negative partial log-likelihood function as

$$\tilde{\ell}_n(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \int_0^\tau \left[\mathbf{B}_i^T \boldsymbol{\theta} - \log[n S^{(0)}(\boldsymbol{\theta}, \mathbf{B}, t)] \right] d N_i(t).$$

Then the oracle estimator $\widehat{\boldsymbol{\theta}}^{or} = (\hat{\eta}^{or}, \widehat{\boldsymbol{\alpha}}^{or})$ is the minimizer of $\tilde{\ell}_n(\boldsymbol{\theta})$.

Now we present the asymptotic results of the proposed estimators.

Theorem 1. Suppose that Conditions (C1)-(C3) given in the Appendix hold. Let $\boldsymbol{\theta}_0$ be the true value of parameter $\boldsymbol{\theta}$. Then

(i) $\widehat{\boldsymbol{\theta}}^{or} \xrightarrow{p} \boldsymbol{\theta}_0$;

(ii) $\sqrt{n}(\widehat{\boldsymbol{\theta}}^{or} - \boldsymbol{\theta}_0)$ converges in distribution to the multivariate normal distribution with zero mean and covariance matrix $\Sigma^{-1}(\boldsymbol{\theta}_0)$, where $\Sigma(\boldsymbol{\theta}_0)$ is given in the Appendix.

Theorem 1 shows that when the grouping structure is known, the oracle estimator is consistent and asymptotically normal. Next, when the true subgroup set $\mathcal{G}_0$ is known, we define the oracle parameter space of $\boldsymbol{\beta}$ as

$$\mathcal{M}_{\mathcal{G}_0} = \{\boldsymbol{\beta} \in R^{np} : \beta_i = \beta_j = \alpha_k, \text{ for any } i, j \in \mathcal{G}_{0,k}, 1 \le k \le K_0\}.$$

Define $(\hat{\eta}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ as the minimizer of $\ell_n(\eta, \boldsymbol{\beta})$ with subject to $\boldsymbol{\beta} \in \mathcal{M}_{\mathcal{G}_0}$. Set $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$ to be the true parameter. We first consider the case of $K_0 \ge 2$ and have the following result.

Theorem 2. Suppose that Conditions (C1)-(C4) given in the Appendix hold. Let $b = \min_{i \in \mathcal{G}_{0,k}, j \in \mathcal{G}_{0,k'}, k \ne k'} ||\beta_{0i} - \beta_{0j}|| = \min_{k \ne k'} ||\alpha_{0k} - \alpha_{0k'}||$. Assume that $b > a\lambda$ for constant a in Condition (C4) and $b \gg \phi_2$. Then there exists a local minimizer $(\hat{\eta}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda))$ of the objective function $Q_n(\eta, \boldsymbol{\beta}; \lambda)$ satisfying $P\big((\hat{\eta}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda)) = (\hat{\eta}^{or}, \widehat{\boldsymbol{\beta}}^{or})\big) \to 1$.

Next, we consider the case of a homogeneous model in which $K_0 = 1$ and $\beta_{01} = \cdots = \beta_{0n} \equiv \boldsymbol{\alpha}_0$.

Theorem 3. Suppose that Conditions (C1)-(C4) given in the Appendix hold. When there is only one group, we define the oracle parameter space of $\boldsymbol{\beta}$ as $\mathcal{M} = \{\boldsymbol{\beta} \in R^{np} : \beta_i \equiv \boldsymbol{\alpha}, i = 1, \cdots, n\}$, and the oracle estimator $(\hat{\eta}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ as the minimizer of $\ell_n(\eta, \boldsymbol{\beta})$ with $\boldsymbol{\beta} \in \mathcal{M}$. Then there exists a local minimizer $(\hat{\eta}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda))$ of the objective function $Q_n(\eta, \boldsymbol{\beta}; \lambda)$ satisfying $P\big((\hat{\eta}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda)) = (\hat{\eta}^{or}, \widehat{\boldsymbol{\beta}}^{or})\big) \to 1$.

Let $\widehat{\boldsymbol{\alpha}}(\lambda)$ be the distinct value of $\widehat{\boldsymbol{\beta}}(\lambda)$ and $\widehat{\boldsymbol{\alpha}}^{or}$ be the distinct value of $\widehat{\boldsymbol{\beta}}^{or}$. By Theorems 1–3, we conclude that $n^{1/2}(\widehat{\boldsymbol{\theta}}(\lambda) - \boldsymbol{\theta}_0)$ converges in distribution to the multivariate normal distribution with mean 0 and covariance matrix $\Sigma^{-1}(\boldsymbol{\theta}_0)$.

5 | SIMULATION STUDIES

We conducted simulation studies to evaluate the performance of the proposed method. The data were generated from model (2) with censoring rate 0.20, where $\lambda_0(t) = 1$, $\eta = (-1, 1)^T$, and $Z_i = (Z_{i1}, Z_{i2})^T$ was generated from multivariate normal with mean 0, variance 1 and correlation 0.4. We considered four examples: (i) one treatment variable with two latent subgroups of equal size; (ii) multi-treatment variable with two subgroups of unequal size; (iii) one treatment variable with three latent subgroups of equal size; (iv) one treatment variable with a homogeneous effect. Two penalties, group SCAD and group MCP, were used in the examples to compare their performance with oracle estimators. The parameter γ was taken as 3.7 and 2.5 for SCAD and MCP, respectively. We set sample size $n = 100$ or 200 in Examples 1, 2 and 4 and $n = 150$ or 300 in Example 3, and let $\vartheta = 1$ in the majorized ADMM algorithm.

To implement the algorithm, we adopt the warm start to update the solution path of $\boldsymbol{\beta}$ and η along different values of λ, and use the modified BIC criterion in Lee et al[23] to select the optimal tuning parameter λ by minimizing

$$BIC(\lambda) = l_n(\hat{\eta}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda)) + C_n \frac{\log n}{n}(\widehat{K}(\lambda)p + q),$$

where $C_n = \log(n\widehat{K}(\lambda) + q)$. The simulation results are based on 100 replications.

Example 1. We first generated X_i from Bernoulli(0.5)+1. Let $\mathcal{G}_1 = \{1, \ldots, n/2\}$ and $\mathcal{G}_2 = \{n/2+1, \ldots, n\}$, and the effects of variable X on the survival time were divided into 2 groups with equal size. We considered the following two cases to investigate the effect of the size of the difference between the subgroup-specific treatment effects:

Case 1: $\beta_i = -1.5$ for $i \in \mathcal{G}_1$ and $\beta_i = 1.5$ for $i \in \mathcal{G}_2$, that is, $\boldsymbol{\alpha} = (-1.5, 1.5)^T$.

Case 2: $\beta_i = -3$ for $i \in \mathcal{G}_1$ and $\beta_i = 3$ for $i \in \mathcal{G}_2$, that is $\boldsymbol{\alpha} = (-3, 3)^T$.

We also compared our approach with the subgroup analysis approach under the logistic-Cox mixture model[15] in Example 1.

TABLE 1 Simulation results for estimation of group size K in Example 1.

n	METHOD	MEAN	MEDIAN	SD	TPR
		Case 1: $\alpha = (-1.5, 1.5)$ and $\eta = (-1, 1)$			
100	GMCP	2.10	2	0.333	0.911
	GSCAD	2.09	2	0.321	0.909
200	GMCP	2.13	2	0.367	0.922
	GSCAD	2.08	2	0.273	0.923
		Case 2: $\alpha = (-3, 3)$ and $\eta = (-1, 1)$			
100	GMCP	2	2	0	0.978
	GSCAD	2	2	0	0.979
200	GMCP	2	2	0	0.980
	GSCAD	2	2	0	0.984

The true value of K is $K = 2$. SD represents standard deviation; TPR represents rate of individuals selected into the subgroups correctly.

The simulation results for Exapmle 1 are summarized in Tables 1 and 2 and Figure 1. Figure 1 includes two kinds of fusiongrams for GMCP when $n = 100$, where one is from one simulated dataset and the other is based on the median estimate of 100 replications for each fixed tuning parameter. The plots from one dataset show how the group size and estimates change as the tuning parameter value increases. It is clear that regression coefficients will be estimated as one group for large enough value of the tuning parameter. As a comparison, the estimates in the fusiongram based on 100 replications are more concentrated. This implies that our ridge initial solution can statistically subgroup the regression coefficients to some degree. The fusiongram for GSCAD and the fusiongram for $n = 200$ are similar and so omitted here. Table 1 reports the estimates of group size K in Example 1. The means and medians of $\widehat{K}$ under both GMCP and GSCAD selectors are close to the true value. When the difference of treatment effects between two subgroups increases, the true positive rate (TPR) becomes larger and are closer to 1, indicating identification of the subgroup structure more accurate. Table 2 further shows the estimates of regression coefficients. We can see that the MEANs and MEDIANs are close to the true values of the parameters, and the standard deviations reduce as the sample size increases. Noting that the logistic-Cox mixture model assumes that the parameter $K = 2$ is given and the grouping membership satisfies a logistic model, its parameter space is much smaller than our model. Table 2 shows the biases and standard errors of our estimators are comparable to those obtained by fitting the logistic-Cox mixture model.

Example 2. Suppose $X_i = (X_{i1}, X_{i2})^T$, where X_{i1} and X_{i2} were generated from Bernoulli(0.5) + 1 and Uniform(1, 3), respectively. Set $\beta_i = (-2, 0.5)^T$ for $i \in \mathcal{G}_1$, and $\beta_i = (2, 3)^T$ for $i \in \mathcal{G}_2$, where $\mathcal{G}_1 = \{1, \ldots, 2n/5\}$, and $\mathcal{G}_2 = \{2n/5 + 1, \ldots, n\}$. Thus, $\boldsymbol{\alpha} = (\alpha_1^T, \alpha_2^T)^T$ with $\alpha_1 = (-2, 2)^T$ and $\alpha_2 = (0.5, 3)^T$.

Example 3. Suppose that X_i was generated from Bernoulli(0.5) + 1. Set $\mathcal{G}_1 = \{1, \ldots, n/3\}$, $\mathcal{G}_2 = \{n/3 + 1, \ldots, 2n/3\}$, and $\mathcal{G}_3 = \{2n/3 + 1, \ldots, n\}$.We set $\beta_i = -3$ for $i \in \mathcal{G}_1$, $\beta_i = 0$ for $i \in \mathcal{G}_2$, and $\beta_i = 3$ for $i \in \mathcal{G}_3$. That is $\boldsymbol{\alpha} = (-3, 0, 3)^T$.

Example 4. Consider the homogeneous model where X_i was generated from Bernoulli(0.5) + 1, and $\beta_i \equiv 1$ for all i.

TABLE 2 Simulation results for estimation of regression coefficients in Example 1.

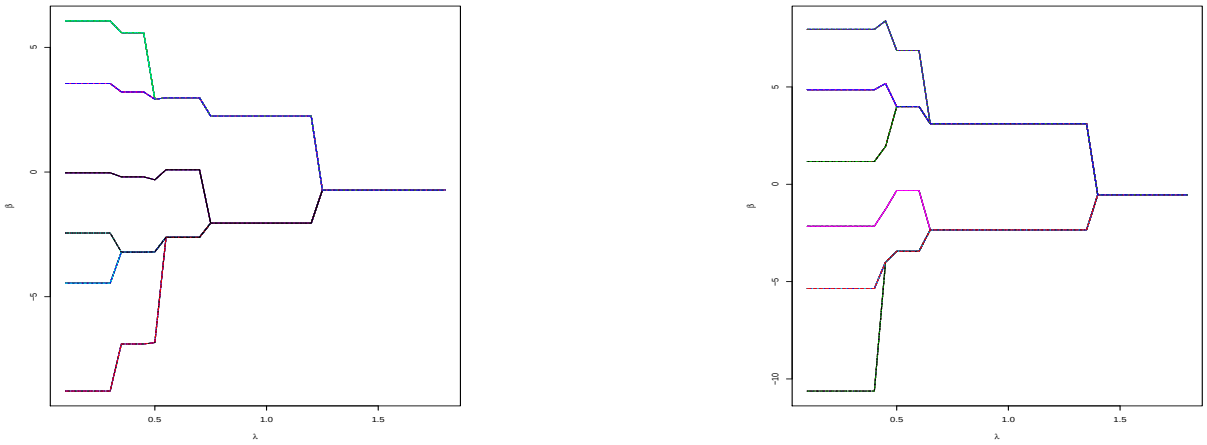
n	PARAMETER	METHOD	MEAN	MEDIAN	SD
			Case 1: $\alpha = (-1.5, 1.5)$ and $\eta = (-1, 1)$		
100	α	GMCP	$(-1.760, 1.773)$	$(-1.782, 1.789)$	$(0.413, 0.421)$
		GSCAD	$(-1.735, 1.791)$	$(-1.773, 1.778)$	$(0.409, 0.406)$
		Mixture	$(-1.545, 1.543)$	$(-1.520, 1.570)$	$(0.500, 0.487)$
		Oracle	$(-1.518, 1.586)$	$(-1.505, 1.594)$	$(0.339, 0.287)$
	η	GMCP	$(-0.850, 0.844)$	$(-0.869, 0.854)$	$(0.232, 0.236)$
		GSCAD	$(-0.843, 0.841)$	$(-0.859, 0.852)$	$(0.229, 0.234)$
		Mixture	$(-1.021, 1.013)$	$(-1.020, 1.026)$	$(0.235, 0.242)$
		Oracle	$(-1.025, 1.027)$	$(-1.015, 1.012)$	$(0.175, 0.168)$
200	α	GMCP	$(-1.704, 1.667)$	$(-1.715, 1.698)$	$(0.299, 0.287)$
		GSCAD	$(-1.750, 1.625)$	$(-1.773, 1.671)$	$(0.307, 0.328)$
		Mixture	$(-1.521, 1.550)$	$(-1.516, 1.554)$	$(0.271, 0.256)$
		Oracle	$(-1.532, 1.538)$	$(-1.531, 1.522)$	$(0.215, 0.215)$
	η	GMCP	$(-0.924, 0.925)$	$(-0.917, 0.910)$	$(0.158, 0.149)$
		GSCAD	$(-0.918, 0.910)$	$(-0.916, 0.903)$	$(0.162, 0.162)$
		Mixture	$(-1.032, 1.033)$	$(-1.031, 1.033)$	$(0.140, 0.137)$
		Oracle	$(-1.020, 1.019)$	$(-1.014, 1.013)$	$(0.112, 0.113)$
			Case 2: $\alpha = (-3, 3)$ and $\eta = (-1, 1)$		
100	α	GMCP	$(-2.969, 3.171)$	$(-3.013, 3.175)$	$(0.642, 0.471)$
		GSCAD	$(-2.976, 3.175)$	$(-3.019, 3.175)$	$(0.645, 0.475)$
		Mixture	$(-2.846, 2.879)$	$(-2.944, 3.109)$	$(0.896, 1.124)$
		Oracle	$(-3.077, 3.167)$	$(-3.013, 3.137)$	$(0.545, 0.450)$
	η	GMCP	$(-0.957, 0.965)$	$(-0.932, 0.969)$	$(0.217, 0.215)$
		GSCAD	$(-0.960, 0.968)$	$(-0.936, 0.969)$	$(0.215, 0.218)$
		Mixture	$(-0.965, 0.968)$	$(-0.975, 1.004)$	$(0.281, 0.296)$
		Oracle	$(-1.025, 1.028)$	$(-1.015, 1.015)$	$(0.178, 0.172)$
200	α	GMCP	$(-2.815, 2.931)$	$(-2.831, 2.955)$	$(0.493, 0.436)$
		GSCAD	$(-2.856, 2.944)$	$(-2.897, 2.987)$	$(0.487, 0.464)$
		Mixture	$(-3.002, 3.013)$	$(-3.008, 3.058)$	$(0.538, 0.572)$
		Oracle	$(-3.077, 3.069)$	$(-3.081, 3.040)$	$(0.339, 0.328)$
	η	GMCP	$(-0.977, 0.966)$	$(-1.000, 1.006)$	$(0.186, 0.199)$
		GSCAD	$(-0.994, 0.976)$	$(-1.005, 1.006)$	$(0.171, 0.186)$
		Mixture	$(-1.012, 1.015)$	$(-1.024, 1.020)$	$(0.158, 0.164)$
		Oracle	$(-1.021, 1.018)$	$(-1.018, 1.009)$	$(0.114, 0.114)$

SD represents standard deviation; Mixture denotes the subgroup analysis results under the logistic-Cox mixture model[15].

The simulation results for Examples 2–4 are summarized in Tables 3–7 and Figures 2–5. The figures show the fusiongram for estimation in Examples 2–4, respectively. Tabels 3, 5 and 7 display the estimates of group size K and the TPR in Examples 2–4, respectively. The means and medians of $\widehat{K}$ under both GMCP and GSCAD selectors are close to the true value, and the TPR are close to 1, which reflect that our methods can identify the group structure correctly with high probability. As the sample size increases, the standard deviation of $\widehat{K}$ decreases and the TPR increases, which demonstrate the good performances of our approaches. Furthermore, Tables 4, 6 and 7 report the estimates of the regression coefficients. The MEAN and MEDIAN of estimators are very close to the true value, and standard deviation (SD) for parameters reduce as the sample size increases.

(a) Fusiongram based on one dataset

Case 1: $\alpha = (-1.5, 1.5)$ Case 2: $\alpha = (-3, 3)$



(b) Fusiongram based on 100 replications

Case 1: $\alpha = (-1.5, 1.5)$ Case 2: $\alpha = (-3, 3)$

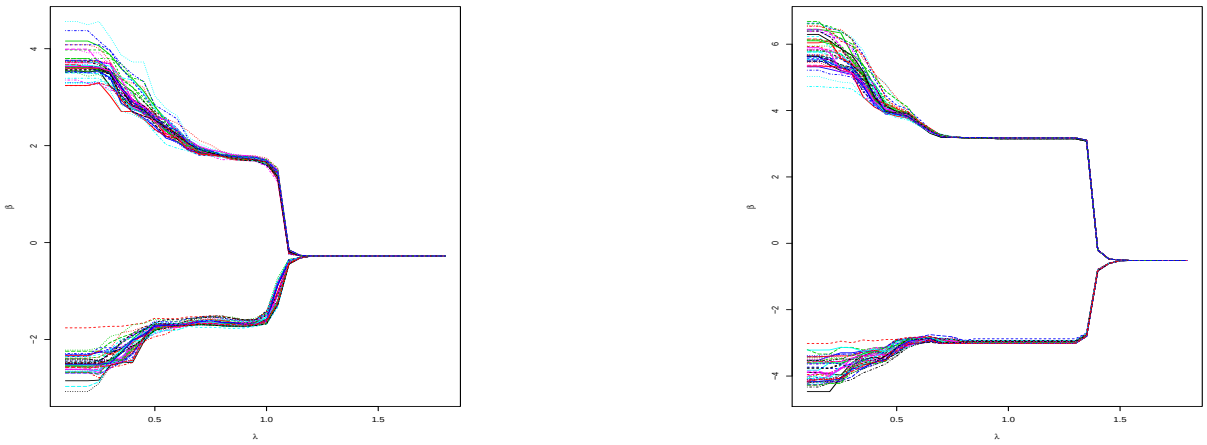


FIGURE 1 Fusiongram for estimation of parameter $\boldsymbol{\beta}$ for GMCP in Example 1 when $n = 100$.

TABLE 3 Simulation results for estimation of group size K in Example 2.

n	METHOD	MEAN	MEDIAN	SD	TPR
100	GMCP	2.10	2	0.362	0.982
	GSCAD	2.08	2	0.339	0.984
200	GMCP	2.07	2	0.256	0.991
	GSCAD	2.07	2	0.256	0.991

The true value of K is $K = 2$. SD represents standard deviation; TPR represents rate of individuals selected into the subgroups correctly.

TABLE 4 Simulation results for estimation of regression coefficients in Example 2.

n	PARAMETER	METHOD	MEAN	MEDIAN	SD
			True Value: $\alpha_1 = (-2, 2)$, $\alpha_2 = (0.5, 3)$ and $\eta = (-1, 1)$		
100	$\hat{\alpha}_1$	GMCP	$(-2.062, 2.110)$	$(-1.996, 2.107)$	$(0.558, 0.465)$
		GSCAD	$(-2.059, 2.108)$	$(-1.996, 2.120)$	$(0.555, 0.463)$
		Oracle	$(-2.108, 2.111)$	$(-2.064, 2.058)$	$(0.606, 0.351)$
	$\hat{\alpha}_2$	GMCP	$(0.496, 2.995)$	$(0.507, 3.004)$	$(0.475, 0.755)$
		GSCAD	$(0.497, 3.012)$	$(0.507, 3.004)$	$(0.474, 0.727)$
		Oracle	$(0.486, 3.244)$	$(0.529, 3.247)$	$(0.453, 0.470)$
	$\hat{\eta}$	GMCP	$(-0.942, 0.957)$	$(-0.972, 0.978)$	$(0.279, 0.283)$
		GSCAD	$(-0.946, 0.959)$	$(-0.977, 0.978)$	$(0.271, 0.274)$
		Oracle	$(-1.058, 1.060)$	$(-1.053, 1.071)$	$(0.150, 0.173)$
200	$\hat{\alpha}_1$	GMCP	$(-1.989, 1.973)$	$(-1.976, 1.974)$	$(0.389, 0.250)$
		GSCAD	$(-1.989, 1.973)$	$(-1.976, 1.974)$	$(0.389, 0.250)$
		Oracle	$(-2.108, 2.070)$	$(-2.095, 2.065)$	$(0.413, 0.249)$
	$\hat{\alpha}_2$	GMCP	$(0.485, 2.929)$	$(0.479, 2.981)$	$(0.329, 0.456)$
		GSCAD	$(0.485, 2.929)$	$(0.479, 2.981)$	$(0.329, 0.456)$
		Oracle	$(0.516, 3.090)$	$(0.501, 3.066)$	$(0.273, 0.275)$
	$\hat{\eta}$	GMCP	$(-0.984, 1.003)$	$(-0.989, 1.024)$	$(0.163, 0.165)$
		GSCAD	$(-0.984, 1.003)$	$(-0.989, 1.024)$	$(0.163, 0.165)$
		Oracle	$(-1.021, 1.009)$	$(-1.001, 0.996)$	$(0.123, 0.114)$

SD represents standard deviation.

TABLE 5 Simulation results for estimation of group size K in Example 3.

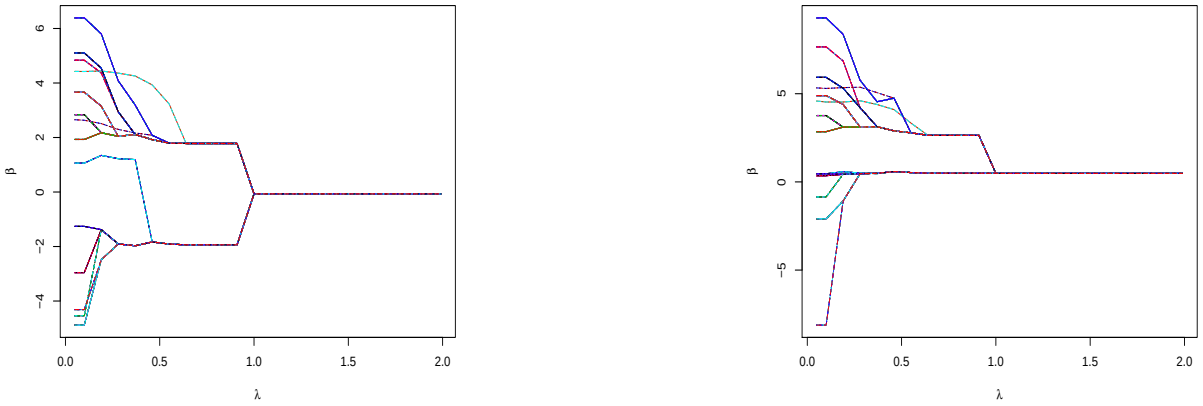
n	METHOD	MEAN	MEDIAN	SD	TPR
150	GMCP	2.99	3	0.225	0.866
	GSCAD	3.03	3	0.264	0.866
300	GMCP	3.01	3	0.100	0.874
	GSCAD	3	3	0	0.876

The true value of K is $K = 3$. SD represents standard deviation; TPR represents rate of individuals selected into the subgroups correctly.

6 | APPLICATION

We applied the proposed method to analyzing the breast cancer data,[24,25] which can be found in the "nki" data set in the R package "dynpred". This trail was carried out in the Dutch Cancer Institute, where 295 patients with breast cancer were put into two treatment groups by the type of surgery (excision and mastectomy), some of them accompanying with two kinds of adjuvant therapies, chemotherapy or hormonal therapy. The main goal is to investigate effects of different surgical treatments on patients' hazard. Hence we focused on the observed data from 255 patients who were not treated with the hormonal therapy for the analysis. Let U_i and C_i be survival and censoring times for the ith patient, $i = 1, \ldots, n$ where $n = 255$. Let X denote the treatment group indicator defined as 1 for patients treated with excision and 0 for patients treated with mastectomy. According to the iterative sure independence screening result[26], we took 5 additional baseline covariates $Z_1, \cdots, Z_5$ into consideration, including age (age), the logarithmic intensity ratio for estrogen-receptor status ($mlratio$), histological grade ($histolgrade = 1$ if well differentiated; 0 otherwise), vascular invasion ($vasc.inv = 1$ for more than 3 vessels; 0 otherwise), and the cross-validated version of the prognostic index ($PICV$). All the continuous covariates were standardized for convenience.

(a) Fusiongram based on one dataset



(b) Fusiongram based on 100 replications

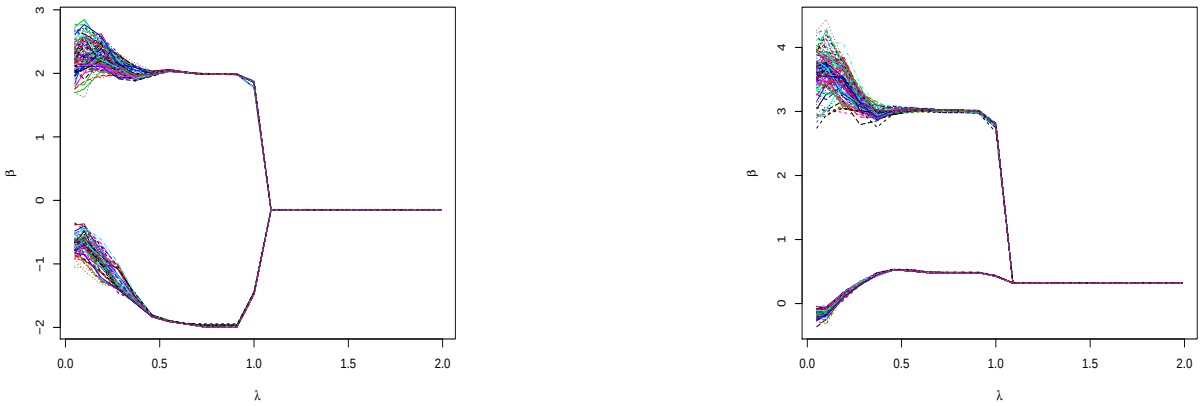


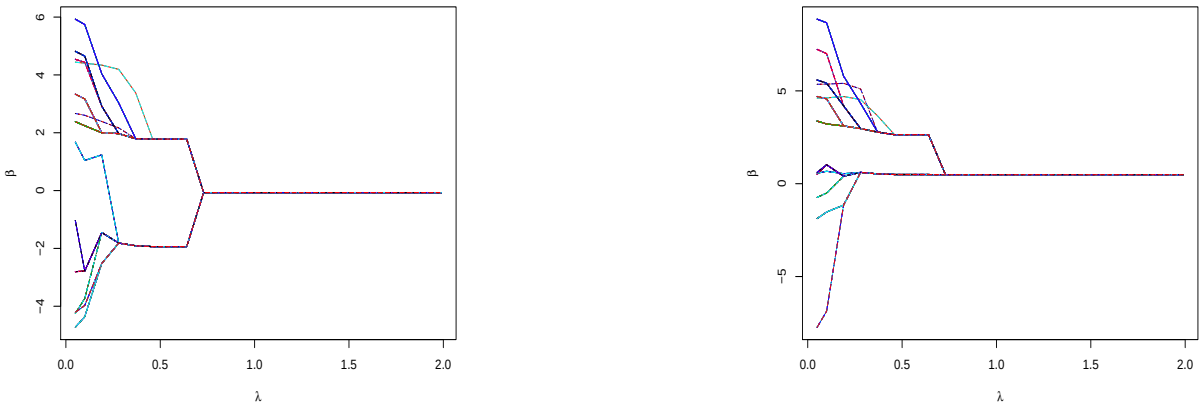
FIGURE 2 Fusiongram for estimation of parameter $\boldsymbol{\beta}$ by GMCP selector in Example 2 when $n = 200$.

To check for the possible heterogeneity of treatment effects, we first fitted the homogeneous Cox model based on the excision treatment group. Figure 6 displays the plot of the kernel density estimate of the martingale residual. We observed that the distribution has multiple modes, indicating the existence of heterogeneous treatment effects.

To demonstrate the heterogeneity of treatment effects, we fitted the proposed heterogeneous Cox model in (2) using our subgroup analysis procedure with group MCP and group SCAD penalties, where the optimal tuning parameter was determined by the modified BIC criterion. Figure 7 displays the fusiongram for the estimate of $\boldsymbol{\beta}$. The grouping and parameter estimation results with GMCP are summarized in Table 8, while the results with GSCAD are similar and so are omitted. For comparison, we also provide the estimation results by fitting both the homogeneous Cox model and the logistic-Cox mixture model[15] in the table. It can be seen from the table that the fitted homogenous Cox model could not detect any significant treatment effect, while both the logistic-Cox mixture approach and the proposed subgroup analysis approach identified the significant subgroup-specific treatment effects.

Furthermore, we present the grouping result in Table 9 according to the type of surgery. It can be seen from the table that our subgroup analysis approach identifies 90% of the patients with the excision and 4% of the patients with the mastectomy as one subgroup and 96% of the patients with the mastectomy and 10% of the patients with excision as another subgroup. For the patients in subgroup 1, the excision can reduce the hazard and prolong the lifetime significantly; while for the patients in subgroup

(a) Fusiongram based on one dataset



(b) Fusiongram based on 100 replications

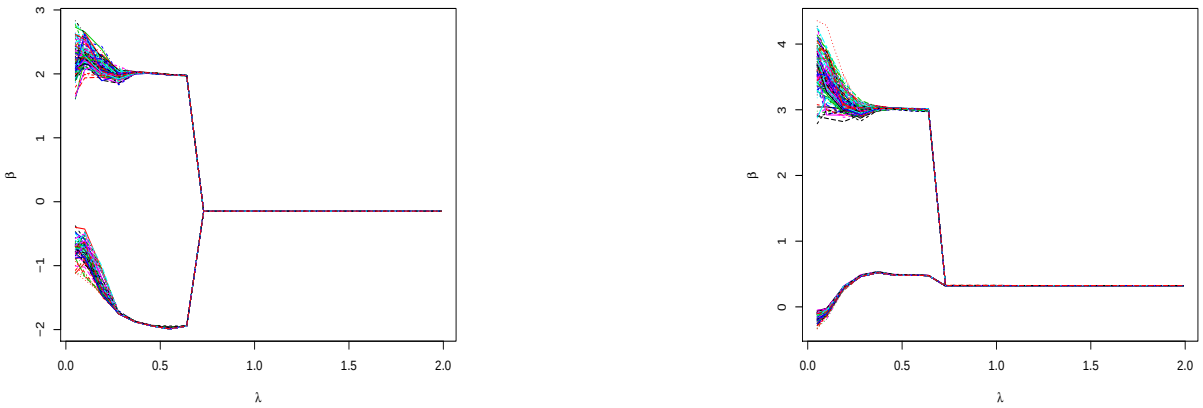


FIGURE 3 Fusiongram for estimation of parameter $\boldsymbol{\beta}$ by GSCAD selector in Example 2 when $n = 200$.

2, the mastectomy is better than the excision. The subgroup analysis approach[15] provides the estimates of the probabilities that patients belong to each subgroup under the logistic model.

The key difference between our approach and the subgroup analysis approach[15] is that the number of the potential subgroups K and the grouping structure are left completely unspecified in our proposed model, while Wu et al.[15] assumed that $K = 2$ and the subgroup membership satisfies a logistic model. Our subgroup analysis method is more flexible and applicable.

7 | DISCUSSION

In this paper, we conduct the subgroup analysis for the heterogenous Cox model using the concave fusion penalized partial likelihood approach. The proposed approach can identify the grouping structure and estimate the heterogeneous covariate effects involved in the model simultaneously and automatically. To obtain an efficient solution to the objective function, we apply the majorized ADMM algorithm which not only converges faster but also calculates more accurately than the local quadratic approximated ADMM algorithm suggested by Ma el al.[17] Our simulation and real data analysis demonstrate that the proposed method performs well. We expect that the proposed approach can be extensively used for subgroup analysis with survival data.

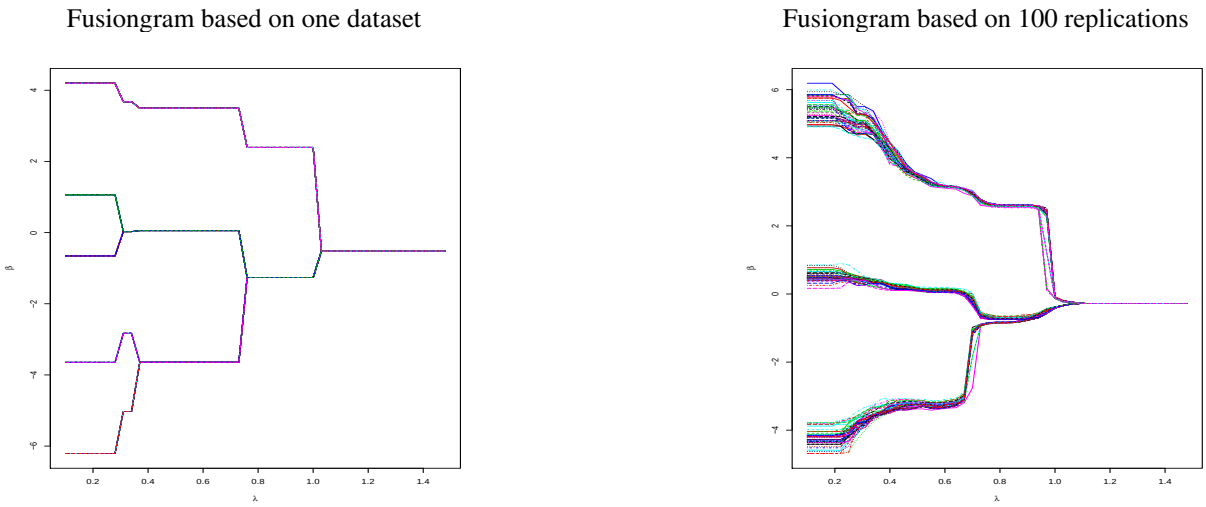


FIGURE 4 Fusiongram for estimation of parameter $\boldsymbol{\beta}$ for GMCP in Example 3 when $n = 150$.

TABLE 6 Simulation results for estimation of regression coefficients in Example 3.

n	PARAMETER	METHOD	MEAN	MEDIAN	SD
			True Value: $\alpha = (-3, 0, 3)$, $\eta = (-1, 1)$		
150	α_1	GMCP	-3.422	-3.453	0.556
		GSCAD	-3.429	-3.387	0.553
		Oracle	-3.087	-3.043	0.429
	α_2	GMCP	0.062	0.074	0.393
		GSCAD	0.033	0.041	0.420
		Oracle	-0.007	0.001	0.233
	α_3	GMCP	3.309	3.219	0.552
		GSCAD	3.280	3.204	0.578
		Oracle	3.076	3.040	0.364
	η	GMCP	$(-0.747, 0.727)$	$(-0.777, 0.720)$	$(0.228, 0.224)$
		GSCAD	$(-0.716, 0.713)$	$(-0.714, 0.716)$	$(0.234, 0.233)$
		Oracle	$(-1.033, 1.024)$	$(-1.029, 1.016)$	$(0.129, 0.133)$
300	α_1	GMCP	-3.283	-3.289	0.388
		GSCAD	-3.288	-3.283	0.411
		Oracle	-3.065	-3.031	0.298
	α_2	GMCP	-0.068	-0.067	0.314
		GSCAD	-0.068	-0.072	0.298
		Oracle	-0.007	-0.007	0.152
	α_3	GMCP	3.088	3.016	0.525
		GSCAD	3.133	3.158	0.523
		Oracle	3.033	3.038	0.238
	η	GMCP	$(-0.775, 0.788)$	$(-0.797, 0.816)$	$(0.182, 0.186)$
		GSCAD	$(-0.785, 0.790)$	$(-0.795, 0.805)$	$(0.183, 0.185)$
		Oracle	$(-1.012, 1.017)$	$(-1.009, 1.013)$	$(0.099, 0.094)$

SD represents standard deviation.

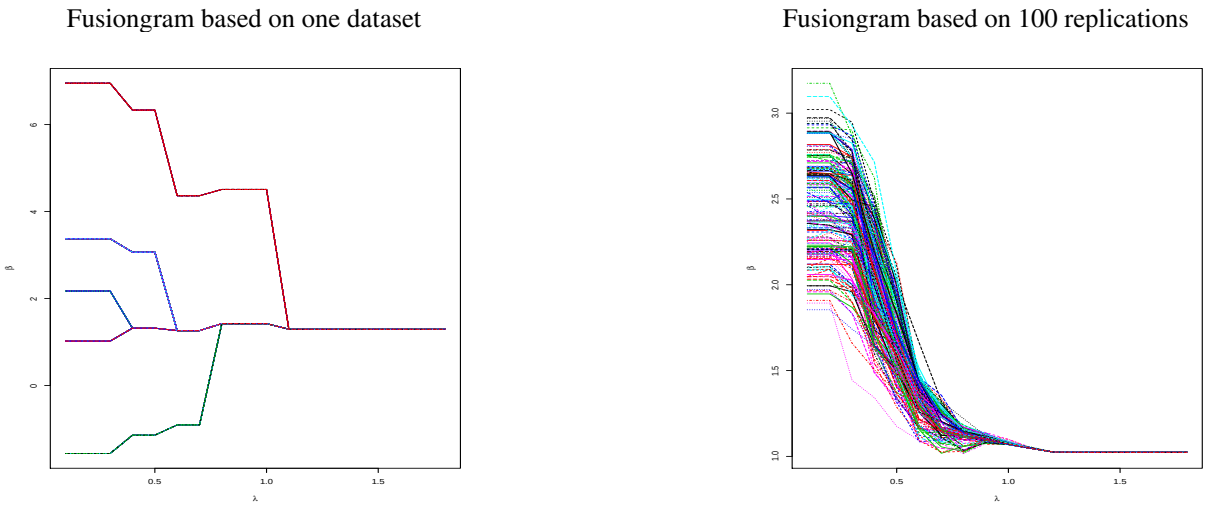


FIGURE 5 Fusiongram for estimation of parameter $\boldsymbol{\beta}$ for GMCP in Example 4 when $n = 200$.

TABLE 7 Simulation results for estimation of K and regression coefficients in Example 4.

n	PARAMETER	METHOD	MEAN	MEDIAN	SD
100	K	GMCP	1.08	1	0.273
		GSCAD	1.09	1	0.288
		Oracle	–	–	–
	α	GMCP	1.022	0.999	0.221
		GSCAD	1.017	0.995	0.221
		Oracle	1.042	1.026	0.245
	η	GMCP	$(-1.004, 0.999)$	$(-0.995, 0.991)$	$(0.178, 0.171)$
		GSCAD	$(-1.000, 0.997)$	$(-0.990, 0.988)$	$(0.180, 0.170)$
		Oracle	$(-1.026, 1.026)$	$(-1.017, 1.008)$	$(0.170, 0.170)$
200	K	GMCP	1.04	1	0.197
		GSCAD	1.01	1	0.100
		Oracle	–	–	–
	α	GMCP	1.022	1.021	0.184
		GSCAD	1.024	1.023	0.183
		Oracle	1.020	1.019	0.172
	η	GMCP	$(-1.029, 1.030)$	$(-1.039, 1.027)$	$(0.114, 0.118)$
		GSCAD	$(-1.029, 1.030)$	$(-1.032, 1.024)$	$(0.113, 0.117)$
		Oracle	$(-1.019, 1.018)$	$(-1.015, 1.008)$	$(0.109, 0.111)$

SD represents standard deviation.

The main differences between our method and Ma et al's [17] are threefold. First, we deal with the Cox model with heterogeneity and censoring, while they consider the heterogenous linear model with complete data. Second, we use the negative partial likelihood-based loss function, while they use the least squares-based loss function. Third, to solve the minimization problem, we utilize the majorized ADMM algorithm, while they apply the local quadratic approximated ADMM algorithm.

Further, the proposed method can be extended to handling the case where the unknown number of subgroups and the dimension of covariates can increase with sample size in the proposed heterogenous Cox model. For this situation, we propose to use

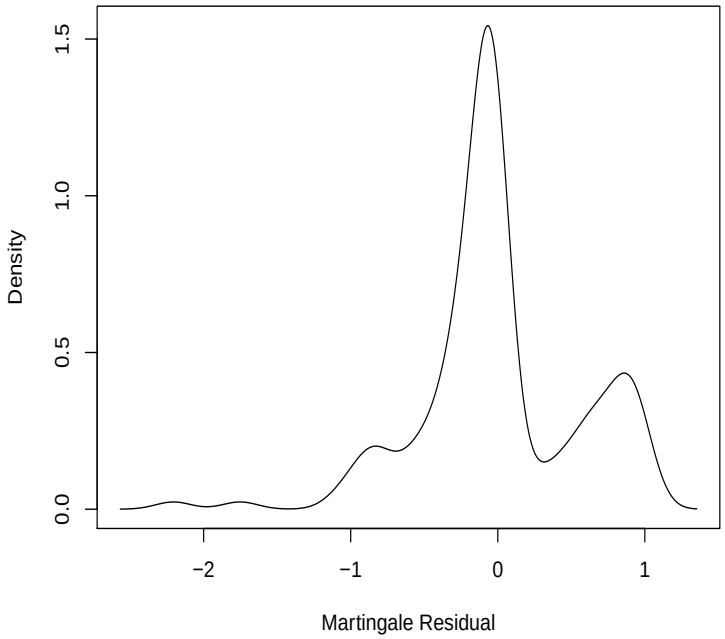


FIGURE 6 The kernel density plot of the residuals after controlling for the effects of the 5 covariates for the patients treated with the excision in the Breast Cancer data.

TABLE 8 Analysis results for Breast Cancer Data.

	PL		Mixture		GMCP	
Covariate	Estimate (ESE)	$p-$value	Estimate (ESE)	p-value	Estimate (ESE)	p-value
Subgroup 1	−0.311(0.244)	.203	−1.571(0.409)	< .001*	−3.981(0.575)	< .001*
Subgroup 2	−(−)	−	1.374(0.425)	.001*	1.917(0.343)	< .001*
age	−0.323(0.110)	.003*	−0.058(0.146)	.689	−0.320(0.109)	.003*
mlratio	−0.285(0.152)	.060	−0.347(0.179)	.053	−0.420(0.155)	.006*
histolgrade	−1.110(0.542)	.041*	−1.004(0.587)	.087	−1.289(0.551)	.019*
vasc.inv	0.642(0.250)	.010*	0.046(0.324)	.889	1.081(0.274)	< .001*
PICV	0.421(0.165)	.011*	0.534(0.166)	.001*	0.505(0.171)	.003*

PL represents partial likelihood approach; Mixture denotes the subgroup analysis results under the logistic-Cox mixture model [15]; * represents significance at 0.05 level.

the criterion function

$$Q_n(\eta, \boldsymbol{\beta}) = \ell_n(\eta, \boldsymbol{\beta}) + \sum_{i<j} p_\gamma^{(1)}(||\beta_i - \beta_j||, \lambda_1) + \sum_{j=1}^{q} p_\gamma^{(2)}(\eta_j, \lambda_2).$$

With the penalty functions $p_\gamma^{(1)}(\cdot, \lambda_1)$ and $p_\gamma^{(2)}(\cdot, \lambda_2)$, we can conduct subgroup analysis and variable selection simultaneously.

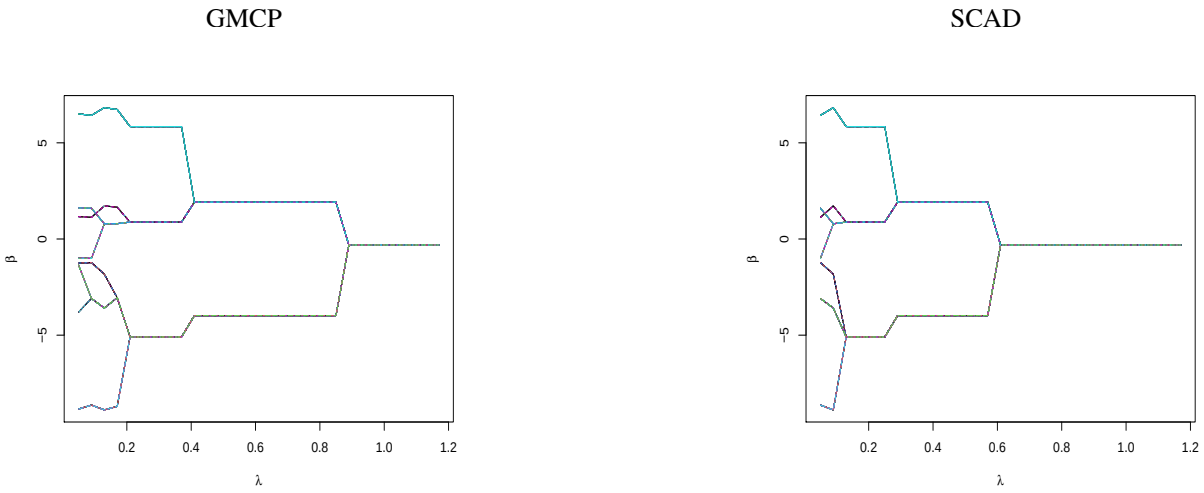


FIGURE 7 Fusiongram for estimation of parameter $\boldsymbol{\beta}$ in Breast Cancer Data analysis.

TABLE 9 The number of patients with different type of surgery in two subgroups.

Treatment	Subgroup 1	Subgroup 2	Total
Excision	128	15	143
Mastectomy	5	107	112
Total	133	122	255

ACKNOWLEDGEMENTS

The authors are grateful to the Editor, Associate Editor, and two referees for their valuable comments and suggestions that greatly improved the paper. This research is supported in part by National Natural Science Foundation of China (11771366), and the Research Grant Council of Hong Kong (15301218, 15303319).

References

1. Fan J, Li R. Variable selection for Cox's proportional hazards model and frailty model. *Ann Stat.* 2002;30(1):74–99.

2. Zhang HH, Lu W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika.* 2007;94(3):691–703.

3. Zhao H, Wu Q, Li G, Sun J. Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. *J Am Stat Assoc.* 2020;115(529):204–216.

4. Bradic J, Fan J, Jiang J. Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann Stat.* 2011;39(6):3092–3120.

5. Huang J, Sun T, Ying Z, Yu Y, Zhang C. Oracle inequalities for the lasso in the Cox model. *Ann Stat.* 2013;41(3):1142–1165.

6. Fang EX, Ning Y, Liu H. Testing and confidence intervals for high dimensional proportional hazards models. *J R Stat Soc Series B Stat Methodol.* 2017;79(5):1415–1437.

7. Chen K, Chen K, Muller HG, Wang JL. Stringing high-dimensional data for function analysis. *J Am Stat Assoc.* 2011;106(493):275–284.

8. Qu S, Wang JL, Wang X. Optimal estimation for the functional cox model. *Ann Stat.* 2016;44(4):1708–1738.

9. Kong D, Ibrahim JG, Lee E, Zhu H. FLCRM: Functional linear cox regression model. *Biometrics.* 2018;74(1):109–117.

10. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Q.* 2004;82(4):661–687.

11. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet.* 2005;365(9454):176–186.

12. Lagakos SW. The challenge of subgroup analyses-reporting without distorting. *N Engl J Med.* 2006;354(16):1667–1669.

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

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

15. Wu R, Zheng M, Yu W. Subgroup analysis with time-to-event data under a logistic-Cox mixture model. *Scand J Stat.* 2016;43(3):863–878.

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

17. Ma S, Huang J, Zhang Z, Liu M. Exploration of heterogeneous treatment effects via concave fusion. *Int J Biostat.* 2019;16(1):20180026.doi:10.1515/ijb-2018-0026.

18. Zhang Y, Wang JH, Zhu Z. Robust subgroup identification. [published ahead of print, 2019]. *Stat Sin.* doi:10.5705/ss.202017.0179.

19. Chen K, Huang R, Chan NH, Yau CY. Subgroup analysis of zero-inflated Poisson regression model with applications to insurance data. *Insur Math Econ.* 2019;86(5):8–18.

20. Li M, Sun D, Toh KC. A majorized ADMM with indefinite proximal terms for linearly constrained convex composite optimization. *SIAM J Optim.* 2016;26(2):922–950.

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

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

23. Lee ER, Noh H, Park BU. Model selection via Bayesian Information Criterion for quantile regression models. *J Am Stat Assoc.* 2014;109(505):216–229.

24. van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* 2002;415(6871):530–536.

25. van de Vijver MJ, He Y, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med.* 2002;347(25):1999–2009.

26. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Series B Stat Methodol.* 2008;70(5):849–911.

APPENDIX

A PROOFS OF THEOREMS

To establish the asymptotic properties of the proposed estimator, we need the following regularity conditions.

(C1) The end time of study τ satisfies that $\int_0^\tau \lambda_0(t)dt < \infty$.

(C2) The covariates X_i and Z_i satisfy that $||X_i|| \le c_1$ and $||Z_i|| \le c_2$ with probability 1.

(C3) The dimension of covariates p, q and the true cluster size K_0 are constants. The sizes of $\mathcal{G}_{0,k}$ satisfy that $|\mathcal{G}_{0,k}|/n \to p_k$ for $k = 1, \cdots, K_0$ when n goes to infinity.

(C4) Set the penalty function $\rho_\gamma(t) = \lambda^{-1} p_\gamma(t, \lambda)$. Suppose that $\rho_\gamma(t)$ is symmetric, non-decreasing and concave on $[0, \infty)$. $\rho_\gamma(t)$ is constant when $t \ge a\lambda$, where a is a positive constant. Furthermore, $\rho_\gamma(0) = 0$ and the derivative $\rho'_\gamma(t)$ satisfies that $\rho'_\gamma(0^+) = 1$.

We introduce more notation before proving the theorems.

Let $S^{(l)}(\boldsymbol{\theta}, \mathbf{B}, t) = n^{-1} \sum_{i=1}^n Y_i(t)\mathbf{B}_i^{\otimes l} \exp(\mathbf{B}_i^T\boldsymbol{\theta})$, where $\mathbf{a}^{\otimes l} = 1, \mathbf{a}, \mathbf{a}\mathbf{a}^T$ for $l = 0, 1, 2$. Define the score function

$$\widetilde{U}_n(\boldsymbol{\theta}) = -\sum_{i=1}^n \int_0^\tau \left[\mathbf{B}_i - \frac{S^{(1)}(\boldsymbol{\theta}, \mathbf{B}, t)}{S^{(0)}(\boldsymbol{\theta}, \mathbf{B}, t)}\right] dN_i(t),$$

and the Hessian matrix

$$\widetilde{H}_n(\boldsymbol{\theta}) = \sum_{i=1}^n \int_0^\tau \left[\frac{S^{(2)}(\boldsymbol{\theta}, \mathbf{B}, t)}{S^{(0)}(\boldsymbol{\theta}, \mathbf{B}, t)} - \left\{\frac{S^{(1)}(\boldsymbol{\theta}, \mathbf{B}, t)}{S^{(0)}(\boldsymbol{\theta}, \mathbf{B}, t)}\right\}^{\otimes 2}\right] dN_i(t).$$

Let $S^{(k,l)}(\boldsymbol{\theta}, \mathbf{B}, t) = \frac{1}{|\mathcal{G}_{0,k}|} \sum_{i \in \mathcal{G}_{0,k}} Y_i(t)\mathbf{B}_i^{\otimes l} \exp(\mathbf{B}_i^T\boldsymbol{\theta})$, where $l = 0, 1, 2$ and $k = 1, \cdots, K_0$. Then we have

$$S^{(l)}(\boldsymbol{\theta}, \mathbf{B}, t) = \frac{1}{n}\sum_{i=1}^n Y_i(t)\mathbf{B}_i^{\otimes l} \exp(\mathbf{B}_i^T\boldsymbol{\theta}) = \sum_{k=1}^{K_0} \frac{|\mathcal{G}_{0,k}|}{n} S^{(k,l)}(\boldsymbol{\theta}, \mathbf{B}, t).$$

Note that $\mathbf{B}_i$, $i \in \mathcal{G}_{0,k}$ are independent and identically distributed random vectors. Denote the expectation of $S^{(k,l)}(\boldsymbol{\theta}, \mathbf{B}, t)$ by $s^{(k,l)}(\boldsymbol{\theta}, t)$, and $s^{(l)}(\boldsymbol{\theta}, t) = \sum_{k=1}^K p_k s^{(k,l)}(\boldsymbol{\theta}, t)$, where $|\mathcal{G}_{0,k}|/n \to p_k$ when $n \to \infty$. Then we have

$$\sup_{t \in [0,\tau]} |S^{(k,l)}(\boldsymbol{\theta}, \mathbf{B}, t) - s^{(k,l)}(\boldsymbol{\theta}, t)|_\infty \xrightarrow{p} 0,$$

and $\sup_{t\in[0,\tau]} |S^{(l)}(\boldsymbol{\theta}, \mathbf{B}, t) - s^{(l)}(\boldsymbol{\theta}, t)|_\infty \xrightarrow{p} 0$, where $|\cdot|_\infty$ denotes the maximum norm. Define

$$\Sigma(\boldsymbol{\theta}_0) = \int_0^\tau \left\{\frac{s^{(2)}(\boldsymbol{\theta}_0, t)}{s^{(0)}(\boldsymbol{\theta}_0, t)} - \left(\frac{s^{(1)}(\boldsymbol{\theta}_0, t)}{s^{(0)}(\boldsymbol{\theta}_0, t)}\right)^{\otimes 2}\right\} s^{(0)}(\boldsymbol{\theta}_0, t)\lambda_0(t)dt.$$

A.1 Proof of Theorem 1

(i) The proof of the first part is based on the techniques for the consistency of the M-estimator. Note that

$$\frac{1}{n}\left(\tilde{\ell}_n(\boldsymbol{\theta}) - \tilde{\ell}_n(\boldsymbol{\theta}_0)\right) = -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \left[\mathbf{B}_i^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \log\frac{S^{(0)}(\boldsymbol{\theta}, \mathbf{B}, t)}{S^{(0)}(\boldsymbol{\theta}_0, \mathbf{B}, t)}\right] dN_i(t).$$

Define

$$\begin{aligned} A_n(\boldsymbol{\theta}) &= -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \left[\mathbf{B}_i^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \log\frac{S^{(0)}(\boldsymbol{\theta}, \mathbf{B}, t)}{S^{(0)}(\boldsymbol{\theta}_0, \mathbf{B}, t)}\right] Y_i(t)\exp(\mathbf{B}_i^T\boldsymbol{\theta}_0)\lambda_0(t)dt \\ &= -\int_0^\tau \left[S^{(1)}(\boldsymbol{\theta}_0, \mathbf{B}, t)^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \log\left\{\frac{S^{(0)}(\boldsymbol{\theta}, \mathbf{B}, t)}{S^{(0)}(\boldsymbol{\theta}_0, \mathbf{B}, t)}\right\} S^{(0)}(\boldsymbol{\theta}_0, \mathbf{B}, t)\right] \lambda_0(t)dt \end{aligned}$$

as the compensator of $\frac{1}{n}\big(\tilde{\ell}_n(\boldsymbol{\theta}) - \tilde{\ell}_n(\boldsymbol{\theta}_0)\big)$, and $M_i(t) = N_i(t) - \int_0^t Y_i(u)\exp(\mathbf{B}_i^T\boldsymbol{\theta}_0)\lambda_0(u)du$. Since $M_i(t)$ is a locally square integrable martingale, then

$$\frac{1}{n}\big(\tilde{\ell}_n(\boldsymbol{\theta}) - \tilde{\ell}_n(\boldsymbol{\theta}_0)\big) - A_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\Big[\mathbf{B}_i^T(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - \log\frac{S^{(0)}(\boldsymbol{\theta},\mathbf{B},t)}{S^{(0)}(\boldsymbol{\theta}_0,\mathbf{B},t)}\Big]dM_i(t)$$

is also a locally square integrable martingale. Hence $\frac{1}{n}\big(\tilde{\ell}_n(\boldsymbol{\theta}) - \tilde{\ell}_n(\boldsymbol{\theta}_0)\big) - A_n(\boldsymbol{\theta})$ has a predictable variation process

$$\begin{aligned}
&\Big\langle \frac{1}{n}\big(\tilde{\ell}_n(\boldsymbol{\theta}) - \tilde{\ell}_n(\boldsymbol{\theta}_0)\big) - A_n(\boldsymbol{\theta}), \frac{1}{n}\big(\tilde{\ell}_n(\boldsymbol{\theta}) - \tilde{\ell}_n(\boldsymbol{\theta}_0)\big) - A_n(\boldsymbol{\theta})\Big\rangle\\
=&\frac{1}{n^2}\sum_{i=1}^{n}\int_0^{\tau}\left[\left\{\mathbf{B}_i^T(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - \log\frac{\sum_{i=1}^n Y_i(t)\exp(\mathbf{B}_i^T\boldsymbol{\theta})}{\sum_{i=1}^n Y_i(t)\exp(\mathbf{B}_i^T\boldsymbol{\theta})}\right\}^2 Y_i(t)\exp(\mathbf{B}_i^T\boldsymbol{\theta}_0)\lambda_0(t)\right]dt\\
=&\frac{1}{n}\int_0^{\tau}\left[(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T S^{(2)}(\boldsymbol{\theta},\mathbf{B},t)(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - 2(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T S^{(1)}(\boldsymbol{\theta},\mathbf{B},t)\log\frac{S^{(0)}(\boldsymbol{\theta},\mathbf{B},t)}{S^{(0)}(\boldsymbol{\theta}_0,\mathbf{B},t)} + \left\{\log\frac{S^{(0)}(\boldsymbol{\theta},\mathbf{B},t)}{S^{(0)}(\boldsymbol{\theta}_0,\mathbf{B},t)}\right\}^2\right]\lambda_0(t)dt.
\end{aligned}$$

By Conditions (C2) and (C3), for any k and l, $s^{(k,l)}(\boldsymbol{\theta},t)$ and $s^{(l)}(\boldsymbol{\theta},t)$ are bounded. Then, by Condition (C1), the predictable variation process has a finite limit. This gives that $\lim_{n\to\infty}\frac{1}{n}\big(\tilde{\ell}_n(\boldsymbol{\theta}) - \tilde{\ell}_n(\boldsymbol{\theta}_0)\big) = A(\boldsymbol{\theta})$, where

$$A(\boldsymbol{\theta}) = \lim_{n\to\infty} A_n(\boldsymbol{\theta}) = -\int_0^{\tau}\left[s^{(1)}(\boldsymbol{\theta}_0,t)^T(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - \log\left\{\frac{s^{(0)}(\boldsymbol{\theta},t)}{s^{(0)}(\boldsymbol{\theta}_0,t)}\right\}s^{(0)}(\boldsymbol{\theta}_0,t)\right]\lambda_0(t)dt.$$

Noting that $\widehat{\boldsymbol{\theta}}^{or}$ is the global minimizer of $\tilde{\ell}_n(\boldsymbol{\theta})$, it is also the global minimizer of $\frac{1}{n}\big(\tilde{\ell}_n(\boldsymbol{\theta}) - \tilde{\ell}_n(\boldsymbol{\theta}_0)\big)$. Since $A(\boldsymbol{\theta})$ is a convex function about $\boldsymbol{\theta}$ and has a global minimizer $\boldsymbol{\theta}_0$, it follows that $\widehat{\boldsymbol{\theta}}^{or}\xrightarrow{p}\boldsymbol{\theta}_0$.

(ii) To prove this part, it suffices to show that $\frac{1}{\sqrt{n}}\widetilde{U}_n(\boldsymbol{\theta}_0)$ converges to a zero mean multivariate normal distribution with covariance matrix $\Sigma(\boldsymbol{\theta}_0)$, and $|\frac{1}{n}\widetilde{H}_n(\widehat{\boldsymbol{\theta}}^{or}) - \Sigma(\boldsymbol{\theta}_0)|_\infty\xrightarrow{p}0$. For this, we only need to verify the conditions of Theorem 8.2.1 of Fleming and Harrington (1991). Recall that

$$\sup_{0\le t\le\tau}|S^{(l)}(\boldsymbol{\theta}_0,\mathbf{B},t) - s^{(l)}(\boldsymbol{\theta}_0,t)|_\infty\xrightarrow{p}0.$$

Noting that $\frac{\partial}{\partial\boldsymbol{\theta}}S^{(k,0)}(\boldsymbol{\theta},\mathbf{B},t) = S^{(k,1)}(\boldsymbol{\theta},\mathbf{B},t)$ and $\frac{\partial}{\partial\boldsymbol{\theta}}S^{(k,1)}(\boldsymbol{\theta},\mathbf{B},t) = S^{(k,2)}(\boldsymbol{\theta},\mathbf{B},t)$, we have $\frac{\partial}{\partial\boldsymbol{\theta}}s^{(k,0)}(\boldsymbol{\theta},t) = s^{(k,1)}(\boldsymbol{\theta},t)$ and $\frac{\partial}{\partial\boldsymbol{\theta}}s^{(k,1)}(\boldsymbol{\theta},t) = s^{(k,2)}(\boldsymbol{\theta},t)$, $k = 1,\cdots,K$. Since $s^{(l)}(\boldsymbol{\theta},t)$ is a linear combination of $s^{(k,l)}(\boldsymbol{\theta},t)$, it follows that $\frac{\partial}{\partial\boldsymbol{\theta}}s^{(0)}(\boldsymbol{\theta},t) = s^{(1)}(\boldsymbol{\theta},t)$ and $\frac{\partial}{\partial\boldsymbol{\theta}}s^{(1)}(\boldsymbol{\theta},t) = s^{(2)}(\boldsymbol{\theta},t)$. By Condition (C2), $s^{(l)}(\boldsymbol{\theta},t)$ is bounded. In addition, as the composition of continuous functions is continuous, we then get that $s^{(l)}(\boldsymbol{\theta}_0,t)$, $0<t<\tau$ are equicontinuous for $l = 0,1,2$.

Condition (C2) gives that $||\mathbf{B}_i|| \le \sqrt{c_1^2+c_2^2}$ with probability 1. Noting that Y_i is a decreasing counting process from 1 to 0, and $\mathbf{B}_i^T\boldsymbol{\theta}_0 > -||\mathbf{B}_i||\cdot||\boldsymbol{\theta}_0||$, we have

$$n^{-1/2}\sup_{1\le i\le n,0\le t\le\tau}||\mathbf{B}_i||Y_i(t)1_{\{\mathbf{B}_i^T\boldsymbol{\theta}_0>-||\mathbf{B}_i||\cdot||\boldsymbol{\theta}_0||\}}\xrightarrow{p}0.$$

Finally, the convexity of negative partial log-likelihood ensures that $\frac{1}{n}\widetilde{H}_n(\boldsymbol{\theta}_0)$ is positive definite and so its limit is

$$\Sigma(\boldsymbol{\theta}_0) = \int_0^{\tau}\left\{\frac{s^{(2)}(\boldsymbol{\theta}_0,t)}{s^{(0)}(\boldsymbol{\theta}_0,t)} - \left(\frac{s^{(1)}(\boldsymbol{\theta}_0,t)}{s^{(0)}(\boldsymbol{\theta}_0,t)}\right)^{\otimes 2}\right\}s^{(0)}(\boldsymbol{\theta}_0,t)\lambda_0(t)dt.$$

By Theorem 8.2.1 in Fleming and Harrington(1991), we conclude the asymptotic normality of $\frac{1}{\sqrt{n}}\widetilde{U}_n(\boldsymbol{\theta}_0)$ and $|\frac{1}{n}\widetilde{H}_n(\widehat{\boldsymbol{\theta}}^{or}) - \Sigma(\boldsymbol{\theta}_0)|_\infty\xrightarrow{p}0$.

By the Taylor's expansion, we get that $\widetilde{U}_n(\widehat{\boldsymbol{\theta}}^{or}) = \widetilde{U}_n(\boldsymbol{\theta}_0) - \widetilde{H}_n(\widetilde{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}^{or}-\boldsymbol{\theta}_0)$, where $\widetilde{\boldsymbol{\theta}}$ is a vector between $\widehat{\boldsymbol{\theta}}^{or}$ and $\boldsymbol{\theta}_0$. Noting that $\widetilde{U}_n(\widehat{\boldsymbol{\theta}}^{or}) = 0$, we have

$$\frac{1}{n}\widetilde{H}_n(\widetilde{\boldsymbol{\theta}})\sqrt{n}(\widehat{\boldsymbol{\theta}}^{or}-\boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}}\widetilde{U}_n(\boldsymbol{\theta}_0).$$

Using the fact that both $\frac{1}{n}\widetilde{H}_n(\widehat{\theta}^{or})$ and $\frac{1}{n}\widetilde{H}_n(\theta_0)$ converge to $\Sigma(\theta_0)$ in probability, $\widetilde{H}_n(\widetilde{\theta})$ also converges to $\Sigma(\theta_0)$ in probability. Besides, as $\frac{1}{\sqrt{n}}\widetilde{U}_n(\theta_0)$ converges to a zero mean normal distribution with covariance matrix $\Sigma(\theta_0)$, we conclude that $\sqrt{n}(\widehat{\theta}^{or}-\theta_0)$ converges to a normal distribution with zero mean and covariance matrix $\Sigma^{-1}(\theta_0)$.

A.2 Proof of Theorem 2

Define the mapping $T^* : R^{np} \to R^{K_0 p}$ as

$$T^*(\boldsymbol{\beta}) = \{|\mathcal{G}_{0,k}|^{-1} \sum_{i\in\mathcal{G}_{0,k}} \beta_i^T, k = 1, \cdots, K_0\}^T,$$

and let the one-to-one mapping $T : \mathcal{M}_{\mathcal{G}_0} \to R^{K_0 p}$ satisfying $T(\boldsymbol{\beta}) = T^*(\boldsymbol{\beta})$. For any vector $\boldsymbol{\beta} \in R^{np}$, set $\boldsymbol{\alpha} = T^*(\boldsymbol{\beta})$ and $\boldsymbol{\beta}^* = T^{-1}(T^*(\boldsymbol{\beta})) = T^{-1}(\boldsymbol{\alpha})$. Noting that for any vector $\eta \in R^q$ and $\boldsymbol{\beta}^* \in \mathcal{M}_{\mathcal{G}_0}$, we have $\ell_n(\eta, \boldsymbol{\beta}^*) = \tilde{\ell}_n((\eta^T, \boldsymbol{\alpha}^T)^T)$. Hence, $\widehat{\theta}^{or}$ defined in Theorem 1 equals to $\left((\hat{\eta}^{or})^T, T(\widehat{\boldsymbol{\beta}}^{or})^T\right)^T$. Consider the neighbourhood of $(\eta_0, \boldsymbol{\beta}_0)$, i.e.,

$$\Theta = \{\eta \in R^q, \boldsymbol{\beta} \in R^{np} : ||\eta - \eta_0|| \le \phi_n, \max_i ||\beta_i - \beta_{0i}|| \le \phi_n\},$$

where $\phi_n \to 0$ as n goes to infinity. To conclude the theorem, it suffices to clarify the following two steps.

(i) For any $(\eta^T, \boldsymbol{\beta}^T)^T \in \Theta$, if $(\eta^T, (\boldsymbol{\beta}^*)^T)^T \neq ((\hat{\eta}^{or})^T, (\widehat{\boldsymbol{\beta}}^{or})^T)^T$, then $Q_n(\eta, \boldsymbol{\beta}^*) > Q_n(\hat{\eta}^{or}, \widehat{\boldsymbol{\beta}}^{or})$.

(ii) For any $(\eta^T, \boldsymbol{\beta}^T)^T \in \Theta$ and large enough n, $Q_n(\eta, \boldsymbol{\beta}) \ge Q_n(\eta, \boldsymbol{\beta}^*)$.

In fact, by Theorem 1, we have $P\left((\hat{\eta}^{or}, \widehat{\boldsymbol{\beta}}^{or}) \in \Theta\right) \to 1$. If (i) and (ii) hold, for any $(\eta^T, \boldsymbol{\beta}^T)^T \in \Theta$ satisfying $(\eta^T, (\boldsymbol{\beta}^*)^T)^T \neq ((\hat{\eta}^{or})^T, (\widehat{\boldsymbol{\beta}}^{or})^T)^T$ and large enough n, we have $Q_n(\eta, \boldsymbol{\beta}) > Q_n(\hat{\eta}^{or}, \widehat{\boldsymbol{\beta}}^{or})$. That means that there is a local minimizer of $Q_n(\eta, \boldsymbol{\beta}; \lambda)$ satisfying that $(\hat{\eta}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda)) = (\hat{\eta}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ with probability tend to 1.

For (i), since $\ell_n(\eta, \boldsymbol{\beta}^*) = \tilde{\ell}_n\left((\eta^T, \boldsymbol{\alpha}^T)^T\right) > \tilde{\ell}_n\left(((\hat{\eta}^{or})^T, (\widehat{\boldsymbol{\alpha}}^{or})^T)^T\right) = \ell_n(\hat{\eta}^{or}, \widehat{\boldsymbol{\beta}}^{or})$, we only need to consider the penalty function $P_n(\beta) = \lambda \sum_{i<j} \rho_\gamma(||\beta_i - \beta_j||)$. Note that $\beta_i^* = \beta_j^*$ when subjects i and j are from the same group. Thus,

$$P_n(\beta^*) = \lambda \sum_{i<j, i\in\mathcal{G}_{0,k}, j\in\mathcal{G}_{0,k'}} \rho_\gamma(||\beta_i^* - \beta_j^*||) = \lambda \sum_{k\neq k'} \frac{|\mathcal{G}_{0,k}||\mathcal{G}_{0,k'}|}{2} \rho_\gamma(||\alpha_k - \alpha_{k'}||).$$

For any $(\eta^T, \boldsymbol{\beta}^T)^T \in \Theta$, we have $\max_i ||\beta_i - \beta_{0i}|| \le \phi_n$. Then for any $k \neq k'$,

$$\begin{aligned}
&||\alpha_k - \alpha_{k'}|| \\
\ge &||\alpha_{0k} - \alpha_{0k'}|| - ||\alpha_k - \alpha_{0k}|| - ||\alpha_{0k'} - \alpha_{k'}|| \ge ||\alpha_{0k} - \alpha_{0k'}|| - 2\max_k ||\alpha_k - \alpha_{0k}|| \\
\ge &b - 2\max_k \left|\left||\mathcal{G}_{0,k}|^{-1} \sum_{i\in\mathcal{G}_{0,k}} (\beta_i - \beta_{0i})\right|\right| \ge b_n - 2|\mathcal{G}_{0,k}|^{-1} \max_k \sum_{i\in\mathcal{G}_{0,k}} \left|\left|(\beta_i - \beta_{0i})\right|\right| \\
\ge &b - 2\max_i ||\beta_i - \beta_{0i}|| \ge b - 2\phi_n > a\lambda.
\end{aligned} \tag{A1}$$

The last inequality follows since $b > a\lambda$ and $b \gg \phi_n$. By Condition (C4), $\rho_\gamma(||\alpha_k - \alpha_{k'}||)$ is a constant, and $P_n(\boldsymbol{\beta}^*)$ is only dependent on sample size n for any $(\eta^T, \boldsymbol{\beta}^T)^T \in \Theta$, which can be denoted as C_n. By the fact that $(\hat{\eta}^{or}, \widehat{\boldsymbol{\alpha}}^{or})$ is the unique global minimizer of $\tilde{\ell}_n(\eta, \boldsymbol{\alpha})$, we get

$$Q_n(\eta, \boldsymbol{\beta}^*) = \ell_n(\eta, \boldsymbol{\beta}^*) + C_n > \ell_n(\hat{\eta}^{or}, \widehat{\boldsymbol{\beta}}^{or}) + C_n = Q_n(\hat{\eta}^{or}, \widehat{\boldsymbol{\beta}}^{or})$$

when $(\eta^T, (\boldsymbol{\beta}^*)^T)^T \neq ((\hat{\eta}^{or})^T, (\widehat{\boldsymbol{\beta}}^{or})^T)^T$. Thus, (i) is concluded.

For (ii), by the Taylor's expansion, we have

$$Q_n(\eta, \boldsymbol{\beta}) - Q_n(\eta, \boldsymbol{\beta}^*) = \left.\frac{\partial \ell_n(\eta, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right|_{\beta=\widetilde{\beta}} (\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \left.\frac{\partial P_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right|_{\beta=\widetilde{\beta}} (\boldsymbol{\beta} - \boldsymbol{\beta}^*) =: \Gamma_1 + \Gamma_2,$$

where $\widetilde{\boldsymbol{\beta}}$ is a vector between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$.

We first consider the second term Γ_2. Note that $P_n(\boldsymbol{\beta}) = \lambda \sum_{i<j, i\in\mathcal{G}_{0,k}, j\in\mathcal{G}_{0,k'}} \rho_\gamma(||\beta_i - \beta_j||)$. Then

$$\begin{aligned}\Gamma_2 =& \left.\frac{\partial P_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right|_{\beta=\widetilde{\beta}} (\boldsymbol{\beta} - \boldsymbol{\beta}^*)\\ =&\lambda \sum_{n\geq j>i\geq 1} \rho'_\gamma(||\tilde{\beta}_i - \tilde{\beta}_j||)\frac{(\tilde{\beta}_i - \tilde{\beta}_j)^T}{||\tilde{\beta}_i - \tilde{\beta}_j||}(\beta_i - \beta_i^*) + \lambda \sum_{1\leq j<i\leq n} \rho'_\gamma(||\tilde{\beta}_j - \tilde{\beta}_i||)\frac{-(\tilde{\beta}_j - \tilde{\beta}_i)^T}{||\tilde{\beta}_j - \tilde{\beta}_i||}(\beta_i - \beta_i^*)\\ =&\lambda \sum_{1\leq i<j\leq n} \rho'_\gamma(||\tilde{\beta}_i - \tilde{\beta}_j||)\frac{(\tilde{\beta}_i - \tilde{\beta}_j)^T}{||\tilde{\beta}_i - \tilde{\beta}_j||}\left\{(\beta_i - \beta_i^*) - (\beta_j - \beta_j^*)\right\}.\end{aligned}$$

On one hand, when subjects i and j are from different groups, that is $i \in \mathcal{G}_{0,k}$ and $j \in \mathcal{G}_{0,k'}$, $k \neq k'$, we have

$$||\tilde{\beta}_i - \tilde{\beta}_j|| \geq ||\beta_{0i} - \beta_{0j}|| - 2\max_i ||\tilde{\beta}_i - \beta_{0i}|| = ||\alpha_{0k} - \alpha_{0k'}|| - 2\max_i ||\tilde{\beta}_i - \beta_{0i}||.$$

Since $(\eta, \boldsymbol{\beta}) \in \Theta$, we can see that $\max_i ||\beta_i - \beta_{0i}|| \leq \phi_n$. By (A1), we have $\max_k ||\alpha_k - \alpha_{0k}|| \leq \phi_n$ for $\boldsymbol{\alpha} = T^*(\boldsymbol{\beta})$. Then $\boldsymbol{\beta}^*$ satisfies that $\max_i ||\beta_i^* - \beta_{0i}|| \leq \phi_n$. By the definition of $\widetilde{\boldsymbol{\beta}}$, we have $\max_i ||\tilde{\beta}_i - \beta_{0i}|| \leq \phi_n$, and $||\tilde{\beta}_i - \tilde{\beta}_j|| \geq b - 2\phi_n > a\lambda$. By Condition (C4), $\rho_\gamma(t)$ is a constant when $t > a\lambda$ and $\rho'_\gamma(t) \equiv 0$ when $t > a\lambda$. Thus, when subjects i and j are from different groups, $\rho'_\gamma(||\tilde{\beta}_i - \tilde{\beta}_j||) \equiv 0$. On the other hand, $\beta_i^* = \beta_j^*$ when i and j are from the same group. Hence $\frac{(\tilde{\beta}_i - \tilde{\beta}_j)^T}{||\tilde{\beta}_i - \tilde{\beta}_j||} = \frac{(\beta_i - \beta_j)^T}{||\beta_i - \beta_j||}$ and

$$\rho'_\gamma(||\tilde{\beta}_i - \tilde{\beta}_j||)\frac{(\tilde{\beta}_i - \tilde{\beta}_j)^T}{||\tilde{\beta}_i - \tilde{\beta}_j||}\left\{(\beta_i - \beta_i^*) - (\beta_j - \beta_j^*)\right\} = \rho'_\gamma(||\tilde{\beta}_i - \tilde{\beta}_j||)||\tilde{\beta}_i - \tilde{\beta}_j||.$$

Note that

$$\begin{aligned}&\max_k \max_{i,j\in\mathcal{G}_{0,k}} ||\tilde{\beta}_i - \tilde{\beta}_j|| = \max_k \max_{i,j\in\mathcal{G}_{0,k}} ||\tilde{\beta}_i - \beta_i^* + \beta_i^* - \beta_j^* + \beta_j^* - \tilde{\beta}_j||\\ \leq& 2\max_i ||\tilde{\beta}_i - \beta_i^*|| \leq 2\max_i(||\tilde{\beta}_i - \beta_{0i}|| + ||\beta_i^* - \beta_{0i}||) \leq 4\phi_n.\end{aligned}$$

By Condition (C4), we have

$$\Gamma_2 = \sum_{k=1}^{K_0} \sum_{\{i,j\in\mathcal{G}_{0,k}, i<j\}} \lambda\rho'_\gamma(||\tilde{\beta}_i - \tilde{\beta}_j||)||\beta_i - \beta_j|| \geq \sum_{k=1}^{K_0} \sum_{\{i,j\in\mathcal{G}_{0,k}, i<j\}} \lambda\rho'_\gamma(4\phi_n)||\beta_i - \beta_j||.$$

Now we turn to the first term Γ_1. Let

$$\mathbf{U}_i = \left.\frac{\partial \ell_n(\eta, \boldsymbol{\beta})}{\partial \beta_i}\right|_{\beta=\widetilde{\beta}} = -\int_0^\tau X_i d N_i(t) + \int_0^\tau \frac{Y_i(t)X_i \exp(Z_i^T\eta + X_i^T\tilde{\beta}_i)}{\frac{1}{n}\sum_{j=1}^n Y_j(t)\exp(Z_j^T\eta + X_j^T\tilde{\beta}_j)} d\bar{N}(t), \tag{A2}$$

where $\bar{N}(t) = \frac{1}{n}\sum_{i=1}^n N_i(t)$. Then after some calculation, we have

$$\begin{aligned}\Gamma_1 =& \sum_{i=1}^n \mathbf{U}_i^T(\beta_i - \beta_i^*) = \sum_{k=1}^{K_0}\sum_{i\in\mathcal{G}_{0,k}} \mathbf{U}_i^T(\beta_i - \beta_i^*) = \sum_{k=1}^{K_0}\sum_{i,j\in\mathcal{G}_{0,k}} \frac{\mathbf{U}_i^T(\beta_i - \beta_j)}{|\mathcal{G}_{0,k}|} = \sum_{k=1}^{K_0}\sum_{i,j\in\mathcal{G}_{0,k}} \frac{\mathbf{U}_i^T(\beta_i - \beta_j)}{2|\mathcal{G}_{0,k}|} + \sum_{k=1}^{K_0}\sum_{i,j\in\mathcal{G}_{0,k}} \frac{\mathbf{U}_j^T(\beta_j - \beta_i)}{2|\mathcal{G}_{0,k}|}\\ =& \sum_{k=1}^{K_0}\sum_{i,j\in\mathcal{G}_{0,k}} \frac{(\mathbf{U}_i - \mathbf{U}_j)^T(\beta_i - \beta_j)}{2|\mathcal{G}_{0,k}|} = \sum_{k=1}^{K_0}\sum_{\{i,j\in\mathcal{G}_{0,k}, i<j\}} \frac{(\mathbf{U}_i - \mathbf{U}_j)^T(\beta_i - \beta_j)}{|\mathcal{G}_{0,k}|} \geq -\sum_{k=1}^{K_0}\sum_{\{i,j\in\mathcal{G}_{0,k}, i<j\}} \frac{2\max_i ||\mathbf{U}_i|| \cdot ||\beta_i - \beta_j||}{|\mathcal{G}_{\min}|},\end{aligned}$$

where $|\mathcal{G}_{\min}| = \min_{k=1,\cdots,K_0} |\mathcal{G}_{0,k}|$. Following the same clues as before, for any $(\eta, \boldsymbol{\beta}) \in \Theta$, we have $(\eta, \widetilde{\boldsymbol{\beta}}) \in \Theta$. Then, by Condition (C2) and (A2), we can find a constant C_U such that $\max_i ||\mathbf{U}_i|| \leq C_U$ with probability 1.

Note that $\lim_{n\to\infty} \rho'_\gamma(4\phi_n) = 1$ and $|\mathcal{G}_{\min}|$ goes to infinity as $n \to \infty$. For large enough n, we can get that

$$Q_n(\eta, \boldsymbol{\beta}) - Q_n(\eta, \boldsymbol{\beta}^*) = \Gamma_1 + \Gamma_2 \geq \sum_{k=1}^{K_0} \sum_{\{i,j\in\mathcal{G}_{0,k}, i<j\}} ||\beta_i - \beta_j||\left[\lambda\rho'_\gamma(4\phi_n) - 2C_U/|\mathcal{G}_{\min}|\right] \geq 0.$$

Thus, (ii) is concluded.

A.3 Proof of Theorem 3

Proof. Similar to the proof of Theorem 2, we define the mapping T and T^* when $K_0 = 1$ and $\mathcal{M}_{\mathcal{G}_0} = \mathcal{M}$. For any vector $\boldsymbol{\beta} \in R^{np}$, set $\boldsymbol{\alpha} = T^*(\boldsymbol{\beta}) \in R^p$ and $\boldsymbol{\beta}^* = T^{-1}(\boldsymbol{\alpha}) \in \mathcal{M}$. The neighbourhood of true parameter Θ and ϕ_n are the same as those in Theorem 2. Then we only need to show the following two steps.

(i) For any $(\eta^T, \boldsymbol{\beta}^T)^T \in \Theta$, if $(\eta^T, (\boldsymbol{\beta}^*)^T)^T \neq ((\hat{\eta}^{or})^T, (\widehat{\boldsymbol{\beta}}^{or})^T)^T$, then $Q_n(\eta, \boldsymbol{\beta}^*) > Q_n(\hat{\eta}^{or}, \widehat{\boldsymbol{\beta}}^{or})$.

(ii) For any $(\eta^T, \boldsymbol{\beta}^T)^T \in \Theta$ and large enough n, $Q_n(\eta, \boldsymbol{\beta}) \geq Q_n(\eta, \boldsymbol{\beta}^*)$.

For (i), when there is only one group, we have $\beta_i^* \equiv \boldsymbol{\alpha}$ and so $P_n(\boldsymbol{\beta}^*) = P_n(\widehat{\boldsymbol{\beta}}^{or}) \equiv 0$. Since $\ell_n(\eta, \boldsymbol{\beta}^*) = \ell_n(\hat{\eta}^{or}, \widehat{\boldsymbol{\beta}}^{or})$, it follows that $Q_n(\eta, \boldsymbol{\beta}^*) > Q_n(\hat{\eta}^{or}, \widehat{\boldsymbol{\beta}}^{or})$.

For (ii),

$$Q_n(\eta, \boldsymbol{\beta}) - Q_n(\eta, \boldsymbol{\beta}^*) = \frac{\partial \ell_n(\eta, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} (\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \frac{\partial P_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} (\boldsymbol{\beta} - \boldsymbol{\beta}^*) =: \Gamma_1 + \Gamma_2,$$

where $\widetilde{\boldsymbol{\beta}}$ is a vector between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$. We first consider the second term

$$\Gamma_2 = \frac{\partial P_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} (\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \lambda \sum_{1 \leq i < j \leq n} \rho_\gamma'(||\tilde{\beta}_i - \tilde{\beta}_j||) \frac{(\tilde{\beta}_i - \tilde{\beta}_j)^T}{||\tilde{\beta}_i - \tilde{\beta}_j||} \big\{(\beta_i - \beta_i^*) - (\beta_j - \beta_j^*)\big\}.$$

Since i and j are from the same group, we have $\beta_i^* = \beta_j^*$ and $\frac{(\tilde{\beta}_i - \tilde{\beta}_j)^T}{||\tilde{\beta}_i - \tilde{\beta}_j||} = \frac{(\beta_i - \beta_j)^T}{||\beta_i - \beta_j||}$. Furthermore, $\max_{i,j} ||\tilde{\beta}_i - \tilde{\beta}_j|| \leq 4\phi_n$. Then by Condition (C4), we get that

$$\Gamma_2 = \lambda \sum_{1 \leq i < j \leq n} \rho_\gamma'(||\tilde{\beta}_i - \tilde{\beta}_j||) ||\tilde{\beta}_i - \tilde{\beta}_j|| \geq \lambda \sum_{1 \leq i < j \leq n} \rho_\gamma'(4\phi_n) ||\tilde{\beta}_i - \tilde{\beta}_j||.$$

For the first term Γ_1, we have

$$\mathbf{U}_i = \frac{\partial \ell_n(\eta, \boldsymbol{\beta})}{\partial \beta_i}\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} = -\int_0^\tau X_i d N_i(t) + \int_0^\tau \frac{Y_i(t) X_i \exp(Z_i^T \eta + X_i^T \tilde{\beta}_i)}{\frac{1}{n}\sum_{j=1}^n Y_j(t) \exp(Z_j^T \eta + X_j^T \tilde{\beta}_j)} d\bar{N}(t),$$

where $\bar{N}(t) = \frac{1}{n}\sum_{i=1}^n N_i(t)$. Since there is a constant C_U such that $\max_i ||\mathbf{U}_i|| \leq C_U$ with probability 1, it yields that

$$\Gamma_1 \geq -\sum_{1 \leq i < j \leq n} \frac{2 \max_i ||\mathbf{U}_i|| \cdot ||\beta_i - \beta_j||}{|n|} \geq -\sum_{1 \leq i < j \leq n} \frac{2C_U ||\beta_i - \beta_j||}{|n|}.$$

Noting that $\lim_{n \to \infty} \rho_\gamma'(4\phi_n) = 1$, we obtain that for large enough n,

$$Q_n(\eta, \boldsymbol{\beta}) - Q_n(\eta, \boldsymbol{\beta}^*) = \Gamma_1 + \Gamma_2 \geq \sum_{1 \leq i < j \leq n} ||\beta_i - \beta_j|| \big[\lambda \rho_\gamma'(4\phi_n) - 2C_U/|n|\big] \geq 0.$$

Hence, (ii) is concluded. □