Statistica Sinica Preprint No: SS-2018-0319	
Title	Subgroup Analysis in Censored Linear Regression
Manuscript ID	SS-2018-0319
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0319
Complete List of Authors	Xiaodong Yan Guosheng Yin and Xingqiu Zhao
Corresponding Author	Xingqiu Zhao
E-mail	xingqiu.zhao@polyu.edu.hk
Notice: Accepted version subject to English editing.	

Statistica Sinica

Subgroup Analysis in Censored Linear Regression

Xiaodong Yan

School of Economics, Shandong University, Jinan, China

Guosheng Yin

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

Xingqiu Zhao

Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

Abstract: In the presence of treatment heterogeneity due to unknown grouping information, standard methods assuming homogeneous treatment effects cannot capture the subgroup structure in the population. To accommodate heterogeneity, we propose a concave fusion approach to identifying the subgroup structures as well as estimating treatment effects for semiparametric linear regression with censored data. In particular, the treatment effects are subject-dependent and subgroup-specific, and our concave fusion penalized method conducts the subgroup analysis without the need to know the individual subgroup memberships in advance. The proposed estimation procedure can automatically identify the subgroup structure and simultaneously estimate the subgroup-specific treatment effects. Our new algorithm proceeds through combining the Buckley–James iterative procedure and the alternating direction method of multipliers. The resulting estimators enjoy the oracle property, and simulation studies and real data application demonstrate the good performance of the new method.

Key words and phrases: Concave penalization; Oracle property; Subgroup analysis; Survival data; Treatment heterogeneity.

1. Introduction

With the rapid development of precision medicine, subgroup analysis has become commonplace in clinical trials for tailoring disease treatment and prevention to subgroups of patients with similar characteristics. Heterogeneity of treatment effects may arise due to underlying differences among groups of patients in the risk, pathology, biology, genetics, severity of disease, among others. Subgroup identification in a heterogeneous population is a crucial step for promoting individualized treatment strategies, which in turn can contribute to deeper understanding of the genetic basis of diseases, more accurate diagnosis, and better survival prediction.

When treatment heterogeneity is present, the average treatment effect obtained by the standard methods can lead to bias and incorrect conclusions. Subgroup analysis, on the other hand, is specifically developed to model potential heterogeneity in the population, which requires a rigorous statistical framework (Kravitz, Duan and Breslow 2004; Rothwell 2005; Lagakos 2006). From the finite mixture modeling perspective, Shen and He (2015) proposed a structured logistic–normal mixture model by quantifying the subgroup membership with logistic regression and the response with normal linear regression. Wu, Zheng and Yu (2016) made an extension to a logistic–Cox mixture model to accommodate censored outcomes. The mixture models typically require specifying the number of components and a parametric model for grouping, which may not be feasible in practice. By contrast, Ma and Huang (2017) developed a pairwise fusion penalized approach using concave penalty functions, such as the smoothly clipped absolute deviations (SCAD) penalty in Fan and Li (2001) and the minimax concave penalty (MCP) in Zhang (2010), which can automatically identify subgroups as well as estimating subgroup-specific intercepts. Furthermore, Ma and Huang (2016) adopted the concave fusion penalized method to estimate the grouping structures and the subgroup-specific treatment effects. This automatic fusion approach to identifying the subgroups is based on complete observations, and thus it is not directly applicable to handling treatment heterogeneity with censored data. Subgroup analysis for censored heterogeneous data brings new theoretical and computational challenges due to censoring and complexity of survival models. As survival data represent one of the most important clinical endpoints, the inference and analysis methods accommodating treatment heterogeneity across subgroups with censored observations play an increasingly critical role in precision medicine. However, most existing survival models are developed for statistical inference on average effects (e.g., Kalbeisch and Prentice, 1980; Fleming and Harrington, 1991; Andersen, Borgan, Keiding and Gill, 1993). In addition, penalized methods have been proposed for variable selection in the Cox proportional hazards model (e.g., Tibshirani, 1997; Fan and Li, 2002; Bradic, Fan and Jiang, 2011; Huang et al., 2013). When the proportional hazards assumption does not hold, alternative models are developed to handle sparsity in regression. For example, Cai, Huang and Tian (2009) proposed a regularization estimation approach for the linear or accelerated failure time model. Liu and Zeng (2013) studied variable selection in transformation survival models with possibly time-varying covariates. Lin and Lv (2013) investigated a high-dimensional sparse additive hazards model with survival data.

To conduct a more systematical subgroup analysis with heterogeneous survival models, we propose a censored linear regression model with heterogeneous treatment effects and assume a sparsity structure for treatment effects. Specifically, the regression model considered allows the effects to be subgroup-specific with unknown grouping information. To estimate the subgroup structures and subgroup-specific treatment effects, we utilize a concave fusion penalized method to shrink pairwise differences of treatment effects, where the tuning parameter is selected by a modified Bayesian information criterion (BIC). Our numerical algorithm combines the Buckley–James iterative procedure (Buckley and James, 1979) and the alternating direction method of multipliers (BJ-ADMM) using concave penalties such as the SCAD or MCP. Under some canonical conditions, the oracle Buckley–James least squares estimator with *a priori* knowledge of the true subgroups is a local minimizer of the proposed objective function with high probability. That means, our proposed estimator can approximate the oracle estimator with high probability.

The rest of this paper is organized as follows. Section 2 describes the censored linear regression model under heterogeneity, the Buckley–James least squares objective function, and the concave fusion penalization method. To compute the penalized Buckley–James least squares estimator, we propose a BJ-ADMM algorithm with concave fusion penalties in Section 3. The theoretical properties of the resulting estimator are established in Section 4. The finite-sample properties of the proposed method are evaluated through simulation studies in Section 5 and our method is illustrated through a real data example in Section 6. Concluding remarks are provided in Section 7, and the technical proofs are given in the Supplemental Materials.

2. Model and Method

2.1 Censored Linear Model with Heterogeneity

Consider a clinical trial with survival endpoint. For each subject, let Y and C denote the transformed survival and censoring times respectively, and let $Z = (z_1, \ldots, z_q)^\top$ be a q-dimensional nuisance covariate vector and $X = (x_1, \ldots, x_p)^\top$ be a p-dimensional covariate vector of interest. The observed data consist of $\{Y_i^*, \delta_i, X_i, Z_i; i = 1, \ldots, n\}$, independent copies of $\{Y^*, \delta, X, Z\}$, with $Y^* = \min(Y, C)$ and $\delta = I(Y \leq C)$.

Under the homogeneous treatment effects, the semiparametric linear regression model takes the form of

$$Y_i = Z_i^\top \eta + X_i^\top \beta + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\eta = (\eta_1, \ldots, \eta_q)^\top$, $\beta = (\beta_1, \ldots, \beta_p)^\top$, and ϵ_i's are assumed to be independent and identically distributed with an unknown distribution F. The corresponding probability density function of ϵ_i is f, $F^{-1}(1) < \infty$ and $E|\epsilon_i| < \infty$, while $E(\epsilon_i)$ need not to be 0. Further, we assume that ϵ_i is independent of (Z_i, X_i, C_i) and conditional on Z_i and X_i, Y_i and C_i are independent.

If individuals are from multi-subgroups with different treatment effects, the homogeneity assumption in model (2.1) is violated. To estimate subgroup-specific effects, we consider a heterogenous linear regression model,

$$Y_i = Z_i^\top \eta + X_i^\top \beta_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.2}$$

where the key difference between models (2.1) and (2.2) lies in the individual-specific treatment effects β_i.

To estimate each individual-specific treatment effect β_i, we assume that all subjects can be classified into R subgroups $\mathcal{G}_1, \ldots, \mathcal{G}_R$, and the regression coefficients satisfy the fused sparse structure,

$$\|\beta_i - \beta_j\| = 0, \quad i, j \in \mathcal{G}_r, \ r = 1, \ldots, R. \tag{2.3}$$

Under the sparsity assumption (2.3), the treatment effects are the same within each subgroup but different across subgroups. Suppose that for $i \in \mathcal{G}_r$, $\beta_i = \rho_r$, where ρ_r is the common value of β_i's in subgroup $\mathcal{G}_r$. Our goal is to estimate the subgroup-specific treatment effects ρ_r's (i.e., β_i's) and identify the fused sparse structure $\mathcal{G}_r$'s (i.e., R) simultaneously.

2.2 Penalized Method via Concave Fusion

Penalized procedures are commonly used for parameter estimation and variable selection. For estimating the parameters $\boldsymbol{\beta} = (\beta_1^\top, \ldots, \beta_n^\top)^\top$ and η, as well as selecting the proper grouping structure of $\boldsymbol{\beta}$ under the sparse assumption (2.3), we develop a penalized Buckley–James least squares method. Let $\boldsymbol{\theta} = (\eta^\top, \boldsymbol{\beta}^\top)^\top$ and $\theta_i = (\eta^\top, \beta_i^\top)^\top$.

As Y_i cannot be completely observed due to censoring, we impute Y_i by its conditional expectation given the observed data,

$$\begin{aligned}\widetilde{Y}_i(\theta_i, F) &= E(Y_i \mid X_i, Z_i, Y_i^*, \delta_i) \\ &= \delta_i Y_i^* + (1-\delta_i)\Big\{Z_i^\top \eta + X_i^\top \beta_i + \frac{\int_{Y_i^* - Z_i^\top \eta - X_i^\top \beta_i}^{\infty} t dF(t)}{1 - F(Y_i^* - Z_i^\top \eta - X_i^\top \beta_i)}\Big\}. \qquad (2.4)\end{aligned}$$

Let $\epsilon_i(\theta_i) = Y_i - Z_i^\top \eta - X_i^\top \beta_i$, $\zeta_i(\theta_i) = C_i - Z_i^\top \eta - X_i^\top \beta_i$, and $\upsilon_i(\theta_i) = \min(\zeta_i(\theta_i), \epsilon_i(\theta_i))$. For a given $\boldsymbol{\theta}$, based on $\{(\upsilon_i(\theta_i), \delta_i), i = 1, \ldots, n\}$ the Kaplan–Meier estimator of the unknown error distribution F in (2.4) is given by

$$\widetilde{F}_{\boldsymbol{\theta}}(t) = 1 - \prod_{i:\upsilon_i(\theta_i) \le t} \left\{1 - \frac{1}{G_n(\boldsymbol{\theta}, \upsilon_i(\theta_i))}\right\}^{\delta_i}, \qquad (2.5)$$

where $G_n(\boldsymbol{\theta}, u) = \sum_{i=1}^n I(\upsilon_i(\theta_i) \ge u)$.

Motivated by the Buckley–James least squares method (Buckley and James, 1979; Miller and Halpern, 1982), we propose a penalized Buckley–James least squares objective function,

$$\begin{aligned}\ell_P(\boldsymbol{\theta};\lambda) &= \frac{1}{2}\sum_{i=1}^{n}\left[\{\widetilde{Y}_i(\theta_i,\widetilde{F}_{\boldsymbol{\theta}}) - Z_i^\top\eta - X_i^\top\beta_i\} - \frac{1}{n}\sum_{i=1}^{n}\{\widetilde{Y}_i(\theta_i,\widetilde{F}_{\boldsymbol{\theta}}) - Z_i^\top\eta - X_i^\top\beta_i\}\right]^2\\ &\quad + \sum_{1\le i<j\le n} P_\lambda(\|\beta_i-\beta_j\|) \qquad (2.6)\end{aligned}$$

where $P_\lambda(\cdot)$ is a penalty function and $\lambda \geq 0$ is a tuning parameter that controls the amount of penalty on $\|\beta_i-\beta_j\|$'s. The tuning parameter λ determines an estimation path of individual-specific treatment effects, and it can shrink $\|\beta_i-\beta_j\|$ towards zero with a large enough value of λ. For a given λ, we define

$$\widehat{\boldsymbol{\theta}}(\lambda) = \text{argmin}_{\boldsymbol{\theta}\in\mathcal{R}^{q+np}}\ \ell_P(\boldsymbol{\theta};\lambda), \qquad (2.7)$$

and the optimal value of λ can be selected via a properly constructed BIC. In particular, we partition the support of λ into a grid of $\lambda_{\min} = \lambda_0 < \lambda_1 < \cdots < \lambda_J = \lambda_{\max}$, and for each λ_j we compute a solution path of $\widehat{\boldsymbol{\theta}}(\lambda_j)$, and obtain the estimated number of subgroups $\widehat{R}(\lambda_j)$ and subgroup-specific effects $\{\widehat{\rho}_1(\lambda_j),\ldots,\widehat{\rho}_{\widehat{R}(\lambda_j)}(\lambda_j)\}$. The optimal $\widehat{\lambda}$ is selected by minimizing a data-driven BIC, i.e., $\widehat{\lambda} = \text{argmin}_{\lambda_j;j=1,\ldots,J}\text{BIC}(\lambda_j)$. Subsequently, we obtain the estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\widehat{\lambda})$, and the individuals can be separated into $\widehat{R} = \widehat{R}(\widehat{\lambda})$ subgroups accordingly, i.e., $\widehat{\mathcal{G}}_r = \{i : \widehat{\beta}_i = \widehat{\rho}_r, i = 1,\ldots n\}$, $r = 1,\ldots,\widehat{R}$.

The commonly used sparsity-inducing penalties include:

(i) Lasso penalty (Tibshirani 1996) with $P_\lambda(t) = \lambda|t|$;

(ii) Smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001) with $P_\lambda(t) = \lambda\int_0^{|t|}\min\{1,(a\lambda-x)_+/a(\lambda-1)\}dx$, $a>2$;

(iii) Minimax concave penalty (MCP) (Zhang 2010) with $P_\lambda(t) = \lambda\int_0^{|t|}\{1 - x/(a\lambda)\}_+ dx$, $a > 2$.

However, Lasso generally assigns a small penalty to a small difference of $\|\beta_i - \beta_j\|$ and consequently the resulting subgroups tend to be dense, which may include too many spurious subgroups with very small differences among them.

3. Computational Procedure

3.1 The BJ-ADMM Algorithm

We propose to use the Buckley–James iterative procedure in conjunction with the ADMM algorithm to obtain the estimator $\widehat{\boldsymbol{\theta}}$. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^\top$, $\boldsymbol{X} = \mathrm{diag}(X_1^\top, \ldots, X_n^\top)$, and $\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$. Let $\bar{\mathbb{Z}}$ be an $n \times q$ matrix with every row equal to $\bar{Z}$, and $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$, where $\boldsymbol{X}_i$ is the ith row of $\boldsymbol{X}$. Let $\bar{\mathbb{X}}$ be an $n \times np$ matrix with every row equal to $\bar{\boldsymbol{X}}$. Define $\widetilde{\mathbb{Z}} = \boldsymbol{Z} - \bar{\mathbb{Z}}$, $\widetilde{\mathbb{X}} = \boldsymbol{X} - \bar{\mathbb{X}}$, and $\mathcal{Q}_Z = I_n - \boldsymbol{Z}(\widetilde{\mathbb{Z}}^\top \boldsymbol{Z})^{-1}\widetilde{\mathbb{Z}}^\top$ where I_n is an $n \times n$ identity matrix. Let $\widetilde{\boldsymbol{Y}}(\boldsymbol{\theta}, \widetilde{F}_{\boldsymbol{\theta}}) = (\widetilde{Y}_1(\theta_1, \widetilde{F}_{\boldsymbol{\theta}}), \ldots, \widetilde{Y}_n(\theta_n, \widetilde{F}_{\boldsymbol{\theta}}))^\top$, $\bar{Y}(\boldsymbol{\theta}, \widetilde{F}_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^n Y_i(\theta_i, \widetilde{F}_{\boldsymbol{\theta}})$, $\bar{\mathbb{Y}}(\boldsymbol{\theta}, \widetilde{F}_{\boldsymbol{\theta}})$ be an n-vector with each component $\bar{Y}(\boldsymbol{\theta}, \widetilde{F}_{\boldsymbol{\theta}})$, and $\widetilde{\mathbb{Y}}(\boldsymbol{\theta}, \widetilde{F}_{\boldsymbol{\theta}}) = \widetilde{\boldsymbol{Y}}(\boldsymbol{\theta}, \widetilde{F}_{\boldsymbol{\theta}}) \; - \bar{\mathbb{Y}}(\boldsymbol{\theta}, \widetilde{F}_{\boldsymbol{\theta}})$. Let $\Omega = \mathcal{E} \otimes I_p$, where $\mathcal{E} = \{(e_i - e_j), i < j\}^\top_{\frac{n(n-1)}{2} \times n}$ with e_i being the ith $n \times 1$ unit vector whose ith element is 1 and the remaining elements are 0, I_p is a $p \times p$ identity matrix and $\otimes$ represents the Kronecker product. Let $\langle a, b\rangle = a^\top b$ represent the inner product of two vectors a and b of the same dimension. With the notation $\alpha_{ij} = \beta_i - \beta_j$, the objective function in

(2.6) can be written as

$$\begin{aligned}\widetilde{\ell}_P(\eta,\boldsymbol{\beta},\boldsymbol{\alpha}) &= \frac{1}{2}\sum_{i=1}^{n}\{\widetilde{Y}_i(\theta_i,\widetilde{F}_{\boldsymbol{\theta}})-Z_i^{\top}\eta-X_i^{\top}\beta_i\}^2-\frac{1}{2n}\Big\{\sum_{i=1}^{n}(\widetilde{Y}_i(\theta_i,\widetilde{F}_{\boldsymbol{\theta}})-Z_i^{\top}\eta-X_i^{\top}\beta_i)\Big\}^2\\ &\quad+\sum_{1\le i<j\le n}P_\lambda(\|\alpha_{ij}\|), \qquad \text{subject to } \beta_i-\beta_j-\alpha_{ij}=0, \qquad (3.1)\end{aligned}$$

where $\boldsymbol{\alpha}=\{\alpha_{ij}^{\top}, i<j\}^{\top}$. Under the constraints in (3.1), the augmented Lagrangian equation is

$$Q(\eta,\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\nu})=\widetilde{\ell}_P(\eta,\boldsymbol{\beta},\boldsymbol{\alpha})+\sum_{i<j}\langle\nu_{ij},\beta_i-\beta_j-\alpha_{ij}\rangle+\frac{\varphi}{2}\sum_{i<j}\|\beta_i-\beta_j-\alpha_{ij}\|^2, \qquad (3.2)$$

where the dual variables $\boldsymbol{\nu}=\{\nu_{ij}^{\top}, i<j\}^{\top}$ are the Lagrange multipliers and φ is a penalty parameter. Given the parameter values $\boldsymbol{\theta}^{(k)}=(\eta^{(k)\top},\boldsymbol{\beta}^{(k)\top})^{\top}$ and $\boldsymbol{\nu}^{(k)}$ at the kth step, our BJ-ADMM iterative algorithm proceeds as follows:

$$\begin{aligned}\boldsymbol{\alpha}^{(k+1)} &= \arg\min_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha},\boldsymbol{\beta}^{(k)},\boldsymbol{\nu}^{(k)}), & (3.3)\\ \nu_{ij}^{(k+1)} &= \nu_{ij}^{(k)}+\varphi(\beta_i^{(k)}-\beta_j^{(k)}-\alpha_{ij}^{(k+1)}), & (3.4)\\ (\eta^{(k+1)},\boldsymbol{\beta}^{(k+1)}) &= \arg\min_{\eta,\boldsymbol{\beta}} Q(\eta,\boldsymbol{\beta},\boldsymbol{\alpha}^{(k+1)},\boldsymbol{\nu}^{(k+1)}\mid\boldsymbol{\theta}^{(k)}), & (3.5)\end{aligned}$$

where $L(\boldsymbol{\alpha},\boldsymbol{\beta}^{(k)},\boldsymbol{\nu}^{(k)})$ is the simplified version of the objective function $Q(\eta^{(k)},\boldsymbol{\beta}^{(k)},\boldsymbol{\alpha},\boldsymbol{\nu}^{(k)})$ after discarding the terms independent of α,

$$\begin{aligned}L(\boldsymbol{\alpha},\boldsymbol{\beta}^{(k)},\boldsymbol{\nu}^{(k)}) &= \frac{\varphi}{2}\sum_{i<j}\|\beta_i^{(k)}-\beta_j^{(k)}+\varphi^{-1}\nu_{ij}^{(k)}-\alpha_{ij}\|^2+\sum_{i<j}P_\lambda(\|\alpha_{ij}\|)(3.6)\\ Q(\eta,\boldsymbol{\beta},\boldsymbol{\alpha}^{(k+1)},\boldsymbol{\nu}^{(k+1)}\mid\boldsymbol{\theta}^{(k)}) &= \widetilde{\ell}_P(\eta,\boldsymbol{\beta},\boldsymbol{\alpha}^{(k+1)}\mid\boldsymbol{\theta}^{(k)})+\sum_{i<j}\Big\langle\nu_{ij}^{(k+1)},\beta_i-\beta_j-\alpha_{ij}^{(k+1)}\Big\rangle\\ &\quad+\frac{\varphi}{2}\sum_{i<j}\|\beta_i-\beta_j-\alpha_{ij}^{(k+1)}\|^2, \qquad (3.7)\end{aligned}$$

and

$$
\begin{aligned}
&\widetilde{\ell}_P(\eta, \boldsymbol{\beta}, \boldsymbol{\alpha}^{(k+1)} \mid \boldsymbol{\theta}^{(k)}) \\
= \;& \frac{1}{2}\sum_{i=1}^{n}\{\widetilde{Y}_i(\theta_i^{(k)}, \widetilde{F}_{\boldsymbol{\theta}^{(k)}}) - Z_i^\top \eta - X_i^\top \beta_i\}^2 - \frac{1}{2n}\Big\{\sum_{i=1}^{n}(\widetilde{Y}_i(\theta_i^{(k)}, \widetilde{F}_{\boldsymbol{\theta}^{(k)}}) - Z_i^\top \eta - X_i^\top \beta_i)\Big\}^2 \\
&+ \sum_{1\le i<j\le n} P_\lambda(\|\alpha_{ij}^{(k+1)}\|).
\end{aligned}
$$

Note that the element $\alpha_{ij}^{(k+1)}$ of $\boldsymbol{\alpha}^{(k+1)}$ is the minimizer of $\frac{\varphi}{2}\|\xi_{ij}^{(k)} - \alpha_{ij}\|^2 + P_\lambda(\|\alpha_{ij}\|)$, where $\xi_{ij}^{(k)} = \beta_i^{(k)} - \beta_j^{(k)} + \varphi^{-1}\nu_{ij}^{(k)}$. Different groupwise thresholding operators $P_\lambda(\cdot)$ would yield different estimates $\alpha_{ij}^{(k+1)}$:

(i) for the Lasso penalty (Tibshirani 1996),

$$
\alpha_{ij}^{(k+1)} = S(\xi_{ij}^{(k)}, \lambda/\varphi), \quad \text{where} \quad S(w,t) = \begin{cases} (1 - t/\|w\|)w, & \text{if } t/\|w\| < 1, \\ 0, & \text{otherwise;} \end{cases}
$$

(ii) for the SCAD penalty (Fan and Li 2001) with $a > 1/\varphi + 1$,

$$
\alpha_{ij}^{(k+1)} = \begin{cases} S(\xi_{ij}^{(k)}, \lambda/\varphi), & \text{if } \|\xi_{ij}^{(k)}\| \le \lambda + \lambda/\varphi, \\ \xi_{ij}^{(k)}, & \text{if } \|\xi_{ij}^{(k)}\| > a\lambda, \\ \frac{S(\xi_{ij}^{(k)}, a\lambda/((a-1)\varphi))}{1 - 1/((a-1)\varphi)}, & \text{otherwise;} \end{cases}
$$

(iii) for the MCP (Zhang 2010) with $a > 1/\varphi$,

$$
\alpha_{ij}^{(k+1)} = \begin{cases} \frac{S(\xi_{ij}^{(k)}, \lambda/\varphi)}{1 - 1/(a\varphi)}, & \text{if } \|\xi_{ij}^{(k)}\| \le a\lambda, \\ \xi_{ij}^{(k)}, & \text{otherwise.} \end{cases}
$$

Via some algebraic manipulation, the problem in (3.7) is equivalent to minimizing

$$\begin{aligned} & h(\eta, \boldsymbol{\beta}, \boldsymbol{\alpha}^{(k+1)}, \boldsymbol{\nu}^{(k+1)} \mid \boldsymbol{\theta}^{(k)}) \\ = \; & \frac{1}{2}\|\widetilde{\boldsymbol{Y}}(\boldsymbol{\theta}^{(k)}, \widetilde{F}_{\boldsymbol{\theta}^{(k)}}) - \boldsymbol{Z}\eta - \boldsymbol{X}\boldsymbol{\beta}\|^2 - \frac{n}{2}\{\bar{Y}(\boldsymbol{\theta}^{(k)}, \widetilde{F}_{\boldsymbol{\theta}^{(k)}}) - \bar{Z}^\top\eta - \bar{\boldsymbol{X}}^\top\boldsymbol{\beta}\}^2 \\ & + \frac{\varphi}{2}\|\Omega\boldsymbol{\beta} - \boldsymbol{\alpha}^{(k+1)} + \varphi^{-1}\boldsymbol{\nu}^{(k+1)}\|^2. \end{aligned}$$

Thus, for given values of $\boldsymbol{\alpha}^{(k+1)}$, $\boldsymbol{\nu}^{(k+1)}$ and $\boldsymbol{\theta}^{(k)}$, we update $\boldsymbol{\beta}^{(k+1)}$ and $\eta^{(k+1)}$ as follows:

$$\begin{aligned} \boldsymbol{\beta}^{(k+1)} &= (\widetilde{\mathbb{X}}^\top\mathcal{Q}_Z\boldsymbol{X} + \varphi\Omega^\top\Omega)^{-1}\{\widetilde{\mathbb{X}}^\top\mathcal{Q}_Z\widetilde{\boldsymbol{Y}}(\boldsymbol{\theta}^{(k)}, \widetilde{F}_{\boldsymbol{\theta}^{(k)}}) + \varphi\Omega^\top(\boldsymbol{\alpha}^{(k+1)} - \varphi^{-1}\boldsymbol{\nu}^{(k+1)})\}, \\ \eta^{(k+1)} &= (\widetilde{\mathbb{Z}}^\top\boldsymbol{Z})^{-1}\widetilde{\mathbb{Z}}^\top\{\widetilde{\boldsymbol{Y}}(\boldsymbol{\theta}^{(k)}, \widetilde{F}_{\boldsymbol{\theta}^{(k)}}) - \boldsymbol{X}\boldsymbol{\beta}^{(k+1)}\}. \end{aligned}$$

The BJ-ADMM algorithm is terminated until the primal residual $\boldsymbol{r}^{(k)} = \Omega\boldsymbol{\beta}^{(k)} - \boldsymbol{\alpha}^{(k)}$ is close enough to zero, such as $\|\boldsymbol{r}^{(k)}\| < 0.01$. Once convergence is reached, subjects i and j with $\widehat{\alpha}_{ij} = 0$ can be grouped into one subgroup $\widehat{\mathcal{G}}_r$ and estimate the rth subgroup-specific treatment effect through $\widehat{\rho}_r = \frac{1}{|\widehat{\mathcal{G}}_r|}\sum_{i\in\widehat{\mathcal{G}}_r}\widehat{\beta}_i$, where $|\mathcal{G}_r|$ denotes the number of elements in $\mathcal{G}_r$. Note that when $\mathcal{Q}_Z = I_n$, the proposed algorithm reduces to an estimation procedure for the model $Y_i = X_i^\top\beta_i + \epsilon_i$.

3.2 Initial Values

To facilitate the $(k+1)$th update of $(\boldsymbol{\alpha}^{(k+1)}, \boldsymbol{\nu}^{(k+1)}, \eta^{(k+1)}, \boldsymbol{\beta}^{(k+1)})$ in (3.3) to (3.5) of the BJ-ADMM iterative algorithm, we need to specify proper initial values. Motivated by the Buckley–James iterative procedure (Miller and Halpern 1982), we obtain the regression estimators $\eta^{(m+1)}$ and $\boldsymbol{\beta}^{(m+1)} = (\beta_1^{(m+1)\top}, \ldots, \beta_n^{(m+1)\top})^\top$ at the $(m+1)$th step by minimizing a ridge fusion criterion

$$\begin{aligned}\ell(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}) \;&=\; \frac{1}{2}\sum_{i=1}^{n}\{\widetilde{Y}_i(\theta_i^{(m)}, \widetilde{F}_{\boldsymbol{\theta}^{(m)}}) - Z_i^\top\eta - X_i^\top\beta_i\}^2 - \frac{1}{2n}\{\sum_{i=1}^{n}(\widetilde{Y}_i(\theta_i^{(m)}, \widetilde{F}_{\boldsymbol{\theta}^{(m)}}) - Z_i^\top\eta - X_i^\top\beta_i)\}^2 \\ &\quad + \frac{\lambda^*}{2}\sum_{1\le i<j\le n}\|\beta_i - \beta_j\|^2, \end{aligned} \tag{3.8}$$

where $\boldsymbol{\theta}^{(m)} = (\eta^{(m)\top}, \boldsymbol{\beta}^{(m)\top})^\top$ are the parameter estimates at the mth step, and we set $\lambda^* = 0.001$.

Using the matrix notation, (3.8) can be written as

$$\begin{aligned}\ell(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}) \;&=\; \frac{1}{2}\|\widetilde{\boldsymbol{Y}}(\boldsymbol{\theta}^{(m)}, \widetilde{F}_{\boldsymbol{\theta}^{(m)}}) - \boldsymbol{Z}\eta - \boldsymbol{X}\boldsymbol{\beta}\|^2 - \frac{n}{2}\{\bar{Y}(\boldsymbol{\theta}^{(m)}, \widetilde{F}_{\boldsymbol{\theta}^{(m)}}) - \bar{Z}^\top\eta - \bar{\boldsymbol{X}}^\top\boldsymbol{\beta}\}^2 \\ &\quad + \frac{\lambda^*}{2}\|\Omega\boldsymbol{\beta}\|^2, \end{aligned}$$

which leads to

$$\begin{aligned}\boldsymbol{\beta}^{(m+1)} \;&=\; (\widetilde{\mathbb{X}}^\top\mathcal{Q}_Z\boldsymbol{X} + \lambda^*\Omega^\top\Omega)^{-1}\widetilde{\mathbb{X}}^\top\mathcal{Q}_Z\widetilde{\boldsymbol{Y}}(\boldsymbol{\theta}^{(m)}, \widetilde{F}_{\boldsymbol{\theta}^{(m)}}), \\ \eta^{(m+1)} \;&=\; (\widetilde{\mathbb{Z}}^\top\boldsymbol{Z})^{-1}\widetilde{\mathbb{Z}}^\top\{\widetilde{\boldsymbol{Y}}(\boldsymbol{\theta}^{(m)}, \widetilde{F}_{\boldsymbol{\theta}^{(m)}}) - \boldsymbol{X}\boldsymbol{\beta}^{(m+1)}(\lambda^*)\}.\end{aligned}$$

In each iterative step, we also update $\widetilde{\boldsymbol{Y}}(\boldsymbol{\theta}^{(m)}, \widetilde{F}_{\boldsymbol{\theta}^{(m)}})$, and the iteration is continued until $\boldsymbol{\theta}^{(m)}$ converges to the limit value, which is then used as the initial value for the BJ-ADMM iterative procedure.

3.3 Tuning Parameter

From a grid of λ values, we select the optimal tuning parameter $\widehat{\lambda}$ by minimizing a modified BIC,

$$\mathrm{BIC}(\lambda) = \log\left\{\frac{1}{n}\|\widetilde{\mathbb{Y}}(\widehat{\boldsymbol{\theta}}(\lambda), \widetilde{F}_{\widehat{\boldsymbol{\theta}}(\lambda)}) - \widetilde{\mathbb{Z}}\widehat{\eta}(\lambda) - \widetilde{\mathbb{X}}\widehat{\boldsymbol{\beta}}(\lambda)\|^2\right\} + C_n\frac{\log n}{n}\left\{\widehat{R}(\lambda)p + q\right\}, \tag{3.9}$$

where C_n is a positive number dependent on n. By default, we take $C_n = \log(np+q)$, $\varphi = 1$, and $a = 3$.

3.4 Convergence of the BJ-ADMM Algorithm

The convergence of the BJ-ADMM algorithm can be demonstrated by showing that both the primal residual and dual residual approach zero in the iterative procedure.

Proposition 1. *If $\{\boldsymbol{\alpha}^{(k)}\}_{k=1}^{\infty}$ is bounded and $\|\boldsymbol{\nu}^{(k+1)} - \boldsymbol{\nu}^{(k)}\| \to 0$, then $\{\eta^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\nu}^{(k)}\}_{k=1}^{\infty}$ is bounded. Moreover, there exists a subsequence $\{\eta^{(k_j)}, \boldsymbol{\beta}^{(k_j)}, \boldsymbol{\alpha}^{(k_j)}, \boldsymbol{\nu}^{(k_j)}\}_{k_j=1}^{\infty}$, such that*

$$\lim_{k_j\to\infty} \|\eta^{(k_j+1)} - \eta^{(k_j)}\| + \|\boldsymbol{\beta}^{(k_j+1)} - \boldsymbol{\beta}^{(k_j)}\| + \|\boldsymbol{\alpha}^{(k_j+1)} - \boldsymbol{\alpha}^{(k_j)}\| + \|\boldsymbol{\nu}^{(k_j+1)} - \boldsymbol{\nu}^{(k_j)}\| = 0$$

and thus $\{\eta^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\nu}^{(k)}\}_{k=1}^{\infty}$ has a subsequence which converges to a stationary point.

The proof is given in the Supplementary Materials. This proposition guarantees that the BJ-ADMM algorithm applied to the objective function (3.2) converges to a minimum point which is locally optimal.

4. Asymptotic Results

4.1 Notation and Conditions

To study the consistency and oracle property of the proposed concave-penalized Buckley–James estimator, we first introduce some notation and regularity conditions. Let $\widetilde{\Pi} = \{\pi_{ir}\}$ denote an $n \times R$ matrix with $\pi_{ir} = 1$ for $i \in \mathcal{G}_r$ and $\pi_{ir} = 0$ for $i \notin \mathcal{G}_r$. Let $\Pi = \widetilde{\Pi} \otimes I_p$, $\boldsymbol{U} = (\boldsymbol{Z}, \boldsymbol{X}\Pi)_{n\times(q+Rp)}$, and U_i is the ith row vector of $\boldsymbol{U}$, i.e., $U_i = [Z_i^\top, X_i^\top \pi_{i1}, \cdots, X_i^\top \pi_{iR}]^\top$ and $\bar{U} = \frac{1}{n}\sum_{i=1}^{n} U_i$. Define $\boldsymbol{\phi} = (\eta^\top, \boldsymbol{\rho}^\top)^\top$, $\boldsymbol{\rho} = (\rho_1^\top, \ldots, \rho_R^\top)^\top$, where ρ_r is the rth subgroup-specific parameter vector of dimension p. Then $\boldsymbol{\beta} = \Pi\boldsymbol{\rho}$, and the corresponding true parameters are $\boldsymbol{\phi}_0 = (\eta_0^\top, \boldsymbol{\rho}_0^\top)^\top$ and $\boldsymbol{\beta}_0 = \Pi\boldsymbol{\rho}_0$. Note that $\Pi^\top\Pi = \text{diag}(|\mathcal{G}_1|, \ldots, |\mathcal{G}_R|) \otimes I_p$, and let $\mathcal{G}_{\min} = \min_{1\leq r\leq R} |\mathcal{G}_r|$ and $\mathcal{G}_{\max} = \max_{1\leq r\leq R} |\mathcal{G}_r|$, which represent the minimum and maximum group sizes, respectively. Let $\epsilon_i(\boldsymbol{\phi}) = Y_i - U_i^\top\boldsymbol{\phi}$, $\zeta_i(\boldsymbol{\phi}) = C_i - U_i^\top\boldsymbol{\phi}$, and $\upsilon_i(\boldsymbol{\phi}) = \min(\epsilon_i(\boldsymbol{\phi}), \zeta_i(\boldsymbol{\phi}))$. In the sequel, we restrict $\boldsymbol{\phi}$ in a bounded interval $\|\boldsymbol{\phi}\| \leq \kappa$, and then $\max_i \|\theta_i\| \leq \kappa$. Based on $\{(\upsilon_i(\boldsymbol{\phi}), \delta_i), i = 1, \ldots, n\}$, we have

$$\widetilde{F}_{\boldsymbol{\phi}}(t) = 1 - \prod_{i:\upsilon_i(\boldsymbol{\phi})\leq t} \left[1 - \frac{1}{G_n(\boldsymbol{\phi}, \upsilon_i(\boldsymbol{\phi}))}\right]^{\delta_i},$$

where $G_n(\boldsymbol{\phi}, u) = \sum_{i=1}^n I(\upsilon_i(\boldsymbol{\phi}) \geq u)$. For a given vector $b = (b_1, \ldots, b_t)^\top \in \mathcal{R}^t$ and a symmetric matrix $A_{t\times t}$, define $\|b\|_\infty = \max_{1\leq s\leq t} |b_s|$, $\|A\|_\infty = \max_{1\leq i\leq t} \sum_{j=1}^t |A_{ij}|$, and $\|A\| = \|A\|_2 = \max_{b\in\mathcal{R}^t, \|b\|=1} \|Ab\|$. Let $\mathbb{E}_{\min}(A)$ and $\mathbb{E}_{\max}(A)$ be the smallest and largest eigenvalues of A respectively, and further let

$$\underline{\rho} = \min_{i\in\mathcal{G}_r, j\in\mathcal{G}_{r'}, r\neq r'} \|\beta_{0i} - \beta_{0j}\| = \min_{r\neq r'} \|\rho_{0r} - \rho_{0r'}\|$$

which is the minimum difference of the common treatment effects between two subgroups.

Define $\mathbb{D}_{\boldsymbol{\phi},i}(u) = (D_{\boldsymbol{\phi}}^{(1)\top}(u), D_{\boldsymbol{\phi}}^{(2)\top}(u)\pi_{i1}, \cdots, D_{\boldsymbol{\phi}}^{(2)\top}(u)\pi_{iR})^\top$, where $D_{\boldsymbol{\phi}}^{(1)}(u) = E[Z_i \mid Y_i^* - U_i^\top\boldsymbol{\phi} \geq u]$ and $D_{\boldsymbol{\phi}}^{(2)}(u) = E[X_i \mid Y_i^* - U_i^\top\boldsymbol{\phi} \geq u]$, and they can be respectively estimated by

$$\begin{aligned} \widehat{D}_{\boldsymbol{\phi}}^{(1)}(u) &= \sum_{i=1}^n Z_i I(\upsilon_i(\boldsymbol{\phi}) \geq u) / \sum_{i=1}^n I(\upsilon_i(\boldsymbol{\phi}) \geq u), \\ \widehat{D}_{\boldsymbol{\phi}}^{(2)}(u) &= \sum_{i=1}^n X_i I(\upsilon_i(\boldsymbol{\phi}) \geq u) / \sum_{i=1}^n I(\upsilon_i(\boldsymbol{\phi}) \geq u). \end{aligned}$$

Also define

$$W_F(t) = t - \frac{\int_t^\infty s dF(s)}{1 - F(t)} \quad \text{and} \quad W_F(t, h) = h(t) - \frac{\int_t^\infty h(s) dF(s)}{1 - F(t)}. \tag{4.1}$$

Let

$$\Sigma_n = \sum_{i=1}^{n} \int I(\zeta_i(\boldsymbol{\phi}_0) \geq u)(U_i - \mathbb{D}_{\boldsymbol{\phi}_0,i}(u))(U_i - \mathbb{D}_{\boldsymbol{\phi}_0,i}(u))^\top W_F^2(u) dF(u)$$

$$\text{and} \qquad V_n = \sum_{i=1}^{n} \int I(\zeta_i(\boldsymbol{\phi}_0) \geq u) U_i (U_i - \mathbb{D}_{\boldsymbol{\phi}_0,i}(u))^\top W_F(u) W_F(u, f'/f) dF(u),$$

where f' is the first derivative of density function f. Let $\mathcal{V}_n = E(V_n^{-1}\Sigma_n V_n^{-1})$. Based on the composition of U, we correspondingly decompose $\mathcal{V}_n$ as

$$\mathcal{V}_n = \begin{pmatrix} \mathcal{V}_{n11} & \mathcal{V}_{n12} \\ \mathcal{V}_{n21} & \mathcal{V}_{n22} \end{pmatrix},$$

where $\mathcal{V}_{n11}$ is a $q \times q$ matrix.

For convenience, we rewrite the penalty function as $p_\lambda(\cdot) = \lambda \varrho_\lambda(\cdot)$ and $\varrho_\lambda(\cdot)$ as $\varrho(\cdot)$ when it is free of λ. Hereafter, $P_\lambda(s)$ is taken to be the folded-concave penalty studied by Lv and Fan (2009) and defined in condition (C1). Let c and c_j's denote some positive constants. We impose three regularity conditions as follows:

(C1) $\varrho_\lambda(s)$ is symmetric, non-decreasing and concave in $s \in [0, \infty)$, and the derivative $\varrho_\lambda^{'}(s)$ is continuous on $(0, \infty)$. It is constant for $s \geq a\lambda$ for some $a > 0$, and $\varrho_\lambda(0) = 0$. In addition, $\varrho_\lambda^{'}(s)$ is increasing in λ and $\varrho_\lambda^{'}(0+) \equiv \varrho^{'}(0+) = c > 0$ is independent of λ.

(C2) Let $\mathcal{S}(s, t \mid F) = tI(t \leq s) + \frac{\int_s^\infty u dF(u)}{1-F(s)} I(t > s)$, $\zeta_i = \zeta_i(\theta_{0i})$, and $\epsilon_i = \epsilon_i(\theta_{0i})$. The imputed noise vector

$$\mathbb{S} = (\mathcal{S}(\zeta_1, \epsilon_1 \mid F), \ldots, \mathcal{S}(\zeta_n, \epsilon_n \mid F))^\top$$

has sub-Gaussian tails such that

$$P(|\boldsymbol{a}^\top \{\mathbb{S} - E(\boldsymbol{\epsilon})\}| > \|\boldsymbol{a}\| x) \leq 2\exp(-c_1 x^2)$$

for any vector $\boldsymbol{a} \in \mathcal{R}^n$ and $x > 0$, where $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^\top$.

(C3) (i) $\sup_i \|X_i\| \leq c_2$ and $\sup_i \|Z_i\| \leq c_3$; (ii) $\mathbb{E}_{\min}(\boldsymbol{U}^\top \boldsymbol{U}) \geq c_4 \mathcal{G}_{\min}$ and $\mathbb{E}_{\max}(\boldsymbol{U}^\top \boldsymbol{U}) \leq c_5 n$.

The penalty criterion in condition (C1) indicates that the singularity at the region ensures sparsity; the concavity reduces the amount of penalty for large parameters; and the increase of $\varrho_\lambda^{'}(s)$ with respect to λ allows λ to effectively control the overall strength of the penalty. The sub-Gaussian tail behavior of the error term in condition (C2) is an extension of Ma and Huang (2016) for the fact that $E(\epsilon_i)$ may not be 0.

4.2 Censored Heterogeneous Model

We first study the theoretical properties of the oracle estimators $\widehat{\boldsymbol{\phi}}^{or} = (\widehat{\eta}^{or\top}, \widehat{\boldsymbol{\rho}}^{or\top})^\top$ in the censored heterogenous linear model. If we had known the underlying subgroup structure (2.3), i.e., the matrix Π is known, then the oracle estimator of $\boldsymbol{\phi}$ would be

$$\widehat{\boldsymbol{\phi}}^{or} = \text{argmin}_{\boldsymbol{\phi} \in \mathcal{R}^{Lp+q}} \Big\{ \frac{1}{2} \|\widetilde{\boldsymbol{Y}}(\boldsymbol{\phi}, \widetilde{F}_{\boldsymbol{\phi}}) - \boldsymbol{U}\boldsymbol{\phi}\|^2 - \frac{n}{2} [\bar{Y}(\boldsymbol{\phi}, \widetilde{F}_{\boldsymbol{\phi}}) - \bar{U}^\top \boldsymbol{\phi}]^2 \Big\}. \tag{4.2}$$

Since the group membership of the subjects, Π, is typically unknown in advance, the oracle estimators are not obtainable in practice, which however can shed light on the theoretical properties of the proposed estimators. Let $v_n = \max(n^{1/2}/\mathcal{G}_{\min}, n^{4\varsigma}/\mathcal{G}_{\min})$ and

$$\widetilde{\Psi}_n(\boldsymbol{\phi}) = n^{-1/2} \sum_{i=1}^n \int I(\zeta_i(\boldsymbol{\phi}) \geq u)(U_i - \mathbb{D}_{\boldsymbol{\phi},i}(u)) W_{F_{\boldsymbol{\phi}}}(u) d\mathcal{M}(u, \epsilon_i(\boldsymbol{\phi}) \mid F_{\boldsymbol{\phi}}),$$

where

$$\mathcal{M}(s, t \mid F) = I(t \leq s) - \frac{\int_{-\infty}^{s} I(t \geq u) dF(u)}{1 - F(s-)}.$$

Theorem 1. *(Large sample properties for oracle estimators). Under conditions (C2)–(C3) and*

$$P\left(\lim_{n\to\infty} n^{1/2-4\varsigma}\left\{\inf_{\boldsymbol{\phi}\leq\kappa,\|\boldsymbol{\phi}-\boldsymbol{\phi}_0\|\geq n^{-\gamma}} \|\widetilde{\Psi}_n(\boldsymbol{\phi})\|\right\} = \infty\right) = 1,$$

and $4\varsigma+\gamma>1$ *with* $\frac{1}{8}\leq\varsigma<1$*, we have*

(i) (Consistency) $\|\widehat{\boldsymbol{\phi}}^{or}-\boldsymbol{\phi}_0\| = o(v_n)$ *a.s.,* $\|\widehat{\boldsymbol{\beta}}^{or}-\boldsymbol{\beta}_0\| = o(\sqrt{\mathcal{G}_{\max}}v_n)$ *a.s., and* $\sup_i\|\widehat{\beta}_i^{or}-\beta_{0i}\| = o(v_n)$ *a.s.*

(ii) (Asymptotic normality) If $v_n\to 0$*, then* $G_n\mathcal{V}_n^{-1/2}(\widehat{\boldsymbol{\phi}}^{or}-\boldsymbol{\phi}_0)\xrightarrow{D}\mathcal{N}(0,1)$*, where* G_n *is a* $1\times(q+Rp)$ *row vector such that* $\|G_n\|=1$*, and* $\xrightarrow{D}$ *denotes convergence in distribution.*

Since $|\mathcal{G}_{\min}|\leq n/R$ and $v_n\to 0$, we conclude $R=o\{\min(n^{1/2},n^{1-4\varsigma})\}$, and thus Theorem 1 indicates that the number of subgroups L is assumed to grow slower than $\min(n^{1/2},n^{1-4\varsigma})$. Let $\mathcal{G}_{\min}=n^{\psi}$ with $0<\psi\leq 1$, and the bound can be rewritten as $v_n=\min(n^{1/2-\psi},n^{4\varsigma-\psi})$.

Theorem 2. *Under conditions (C1)–(C3) and* $\underline{\rho}>c\lambda$ *with* $\lambda\gg\max(\sqrt{\log(n)}/\mathcal{G}_{\min},n^{-1/2+4\varsigma}/\mathcal{G}_{\min})$ *for some constant* $c>0$*, there exists a local minimizer* $\widehat{\boldsymbol{\theta}}(\lambda)$ *of the objective function* $\ell_P(\boldsymbol{\theta};\lambda)$ *given in (2.6) satisfying*

$$P\{\widehat{\boldsymbol{\theta}}(\lambda)=\widehat{\boldsymbol{\theta}}^{or}\}\to 1.$$

Theorem 2 implies that if the minimal difference of the common treatment effects between two subgroups satisfies $\underline{\rho}\gg\max(\sqrt{\log(n)}/\mathcal{G}_{\min},n^{-1/2+4\varsigma}/\mathcal{G}_{\min})$, the oracle estimator $\widehat{\boldsymbol{\theta}}^{or}$ is a local minimizer of the objective function with high probability, and then our method can recover the true subgroup structure with high probability.

Corollary 1. *Under conditions in Theorem 2, as* $n\to\infty$*,* $G_n\mathcal{V}_n^{-1/2}(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}_0)\xrightarrow{D}\mathcal{N}(0,1)$*. As a result, we have* $G_{n1}\mathcal{V}_{n11}^{-1/2}(\widehat{\eta}(\lambda)-\eta_0)\xrightarrow{D}\mathcal{N}(0,1)$*, and* $G_{n2}\mathcal{V}_{n22}^{-1/2}(\widehat{\boldsymbol{\rho}}(\lambda)-\boldsymbol{\rho}_0)\xrightarrow{D}\mathcal{N}(0,1)$*, where* G_{n1} *and* G_{n2} *are respectively* $1\times q$ *and* $1\times Rp$ *row vectors with* $\|G_{n1}\|=\|G_{n2}\|=1$*.*

The asymptotic distribution of the penalized estimators can be used to construct a confidence interval for each ρ_j and also to test the significance of each component of the subgroup-specific treatment effects.

4.3 Censored Homogeneous Model

When the true model contains only homogeneous treatment effects,

$$Y_i=Z_i^\top\eta+X_i^\top\rho+\epsilon_i,\quad i=1,\ldots,n,$$

we have $\beta_1=\cdots=\beta_n=\rho$ and $R=1$. The oracle estimators $\widehat{\boldsymbol{\phi}}^{or}=(\widehat{\eta}^{or\top},\widehat{\rho}^{or\top})^\top$ in the censored homogeneous linear model are

$$\begin{aligned}\widehat{\boldsymbol{\phi}}^{or} &= \text{argmin}_{\boldsymbol{\phi}\in\mathcal{R}^{p+q}}\left\{\frac{1}{2}\|\widetilde{\boldsymbol{Y}}(\boldsymbol{\phi},\widetilde{F}_{\boldsymbol{\phi}})-\boldsymbol{U}^*\boldsymbol{\phi}\|^2-\frac{n}{2}\{\bar{Y}(\boldsymbol{\phi},\widetilde{F}_{\boldsymbol{\phi}})-\bar{U}^{*\top}\boldsymbol{\phi}\}^2\right\}\\ &= \text{argmin}_{(\eta^\top,\rho^\top)^\top\in\mathcal{R}^{p+q}}\left\{\frac{1}{2}\|\widetilde{\boldsymbol{Y}}(\boldsymbol{\phi},\widetilde{F}_{\boldsymbol{\phi}})-\boldsymbol{Z}\eta-\boldsymbol{x}\rho\|^2-\frac{n}{2}\{\bar{Y}(\boldsymbol{\phi},\widetilde{F}_{\boldsymbol{\phi}})-\bar{Z}^\top\eta-\bar{X}^\top\rho\}^2\right\},\end{aligned}$$

where $\boldsymbol{x}=(X_1,\ldots,X_n)^\top$, $\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i$, $\boldsymbol{U}^*=(\boldsymbol{Z},\boldsymbol{x})$, $U^*=(Z^\top,X^\top)^\top$, $U_i^*=(Z_i^\top,X_i^\top)^\top$, and $\bar{U}^*=\frac{1}{n}\sum_{i=1}^n U_i^*$. Let $\widehat{\boldsymbol{\beta}}^{or}=(\widehat{\beta}_1^{or\top},\ldots,\widehat{\beta}_n^{or\top})^\top$ with $\widehat{\beta}_i^{or}=\widehat{\rho}^{or}$, and set $\widehat{\rho}$ and $\widehat{\boldsymbol{\theta}}=(\widehat{\eta}^\top,\widehat{\boldsymbol{\beta}}^\top)^\top$ to be the penalized estimators of ρ and $\boldsymbol{\theta}=(\eta^\top,\boldsymbol{\beta}^\top)^\top$ respectively, and η_0 and ρ_0 correspond to the true coefficient vectors and $\boldsymbol{\phi}_0=(\eta_0^\top,\rho_0^\top)^\top$. Let

$$\Sigma_n^*=\sum_{i=1}^n\int I(\zeta_i(\boldsymbol{\phi}_0)\geq u)(U_i^*-\mathbb{D}_{\boldsymbol{\phi}_0}^*(u))(U_i^*-\mathbb{D}_{\boldsymbol{\phi}_0}^*(u))^\top W_F^2(u)dF(u)$$

$$\text{and}\qquad V_n^*=\sum_{i=1}^n\int I(\zeta_i(\boldsymbol{\phi}_0)\geq u)U_i^*(U_i^*-\mathbb{D}_{\boldsymbol{\phi}_0}^*(u))^\top W_F(u)W_F(u,f'/f)dF(u),$$

where $\mathbb{D}^*_{\boldsymbol{\phi}}(u) = E(U^*|Y^* - U^{*\top}\boldsymbol{\phi} \geq u)$. Then, $\mathbb{D}^*_{\boldsymbol{\phi}}(u)$ can be estimated by $\widehat{\mathbb{D}}^*_{\boldsymbol{\phi}}(u) = \sum_{i=1}^n U_i^* I(\upsilon_i^*(\boldsymbol{\phi}) \geq u)/\sum_{i=1}^n I(\upsilon_i^*(\boldsymbol{\phi}) \geq u)$, where $\upsilon_i^*(\boldsymbol{\phi}) = \min(\epsilon_i^*(\boldsymbol{\phi}), \zeta_i^*(\boldsymbol{\phi}))$, $\epsilon_i^*(\boldsymbol{\phi}) = Y_i - U_i^{*\top}\boldsymbol{\phi}$, and $\zeta_i^*(\boldsymbol{\phi}) = C_i - U_i^{*\top}\boldsymbol{\phi}$. Let $\mathcal{V}_n^* = E(V_n^{*-1}\Sigma_n^* V_n^{*-1})$. Based on the composition of $\boldsymbol{U}^*$, $\mathcal{V}_n^*$ can be correspondingly decomposed as

$$\mathcal{V}_n^* = \begin{pmatrix} \mathcal{V}^*_{n11} & \mathcal{V}^*_{n12} \\ \mathcal{V}^*_{n21} & \mathcal{V}^*_{n22} \end{pmatrix},$$

where $\mathcal{V}^*_{n11}$ is a $q \times q$ matrix.

Moreover, we replace condition (C3) with (C3*) as follows:

(C3*) (i) $\sup_i \|X_i\| \leq c_2$ and $\sup_i \|Z_i\| \leq c_3$; (ii) $\mathbb{E}_{\min}(\boldsymbol{U}^{*\top}\boldsymbol{U}^*) \geq c_6 n$ and $\mathbb{E}_{\max}(\boldsymbol{U}^{*\top}\boldsymbol{U}^*) \leq c_7 n$.

Let $\upsilon_n' = \max(n^{-1/2}, n^{4\varsigma-1})$ and

$$\widetilde{\Psi}_n^*(\boldsymbol{\phi}) = n^{-1/2}\sum_{i=1}^n \int I(\zeta_i^*(\boldsymbol{\phi}) \geq u)(U_i^* - \mathbb{D}^*_{\boldsymbol{\phi}}(u))W_{F_{\boldsymbol{\phi}}}(u)d\mathcal{M}(u, \epsilon_i^*(\boldsymbol{\phi}) \mid F_{\boldsymbol{\phi}}).$$

Theorem 3. *If conditions (C1), (C2) and (C3*) hold, and further*

$$P\left(\lim_{n\to\infty} n^{1/2-4\varsigma}\left\{\inf_{\boldsymbol{\phi}\leq\kappa, \|\boldsymbol{\phi}-\boldsymbol{\phi}_0\|\geq n^{-\gamma}} \|\widetilde{\Psi}_n^*(\boldsymbol{\phi})\|\right\} = \infty\right) = 1,$$

and $4\varsigma + \gamma > 1$ with $\frac{1}{8} \leq \varsigma < 1$, then we have

(i) (Consistency) $\|\widehat{\boldsymbol{\phi}}^{or} - \boldsymbol{\phi}_0\| = o(\upsilon_n')$ a.s., $\|\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}_0\| = o(\sqrt{n}\upsilon_n')$ a.s., and $\sup_i \|\widehat{\beta}_i^{or} - \beta_{0i}\| = o(\upsilon_n')$ a.s. ;

(ii) (Asymptotic normality) If $\upsilon_n' \to 0$, then $G_n'\mathcal{V}_n^{*-1/2}(\widehat{\boldsymbol{\phi}}^{or} - \boldsymbol{\phi}_0) \xrightarrow{D} \mathcal{N}(0,1)$, where G_n' is a $1 \times (q+p)$ row vector with $\|G_n'\| = 1$;

(iii) If $\lambda \gg \max(\sqrt{\log(n)}/n, n^{-3/2+4\varsigma})$ *for some constant* $\varsigma > 0$*, there exists a local minimizer* $\widehat{\boldsymbol{\theta}}$ *of the objective function* $\ell_P(\boldsymbol{\theta}; \lambda)$ *given in (2.6) satisfying*

$$P\{\widehat{\boldsymbol{\theta}}(\lambda) = \widehat{\boldsymbol{\theta}}^{or}\} \to 1.$$

Corollary 2. *Under conditions in Theorem 3, as* $n \to \infty$, $G_n' \mathcal{V}_n^{*-1/2}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) \xrightarrow{D} \mathcal{N}(0, 1)$*. As a result, we have* $G_{n1}' \mathcal{V}_{n11}^{*-1/2}(\widehat{\eta}(\lambda) - \eta_0) \xrightarrow{D} \mathcal{N}(0, 1)$*, and* $G_{n2}' \mathcal{V}_{n22}^{*-1/2}(\widehat{\rho}(\lambda) - \rho_0) \xrightarrow{D} \mathcal{N}(0, 1)$*, where* G_{n1}' *and* G_{n2}' *are respectively* $1 \times q$ *and* $1 \times p$ *row vectors with* $\|G_{n1}'\| = \|G_{n2}'\| = 1$.

5. Simulation Studies

To evaluate the finite-sample performance of the proposed method, we considered three censored linear regression examples including one heterogenous treatment effect, multiple heterogenous treatment effects, and the homogeneous regression setting.

Example 1 (One treatment variable). We generated data from a censored heterogenous linear regression model,

$$Y_i = Z_i^\top \eta + X_i \beta_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $Z_i = (Z_{i1}, Z_{i2})^\top$ was generated from a bivariate standard normal distribution, X_i was generated from the standard normal distribution, and ϵ_i was taken from the normal distribution $\mathcal{N}(1, 0.2^2)$. Furthermore, we generated censoring time C_i from $\log\{\min(\tau, \mathrm{Unif}(0, \tau + 2))\}$, where $\mathrm{Unif}(\cdot,\cdot)$ denotes a uniform distribution and τ controls the censoring rate. The true coefficients were set as $\eta = (\eta_1, \eta_2)^\top = (-1, 1)^\top$. We randomly assigned the treatment coefficients to three subgroups with equal probabilities, i.e., we let $P(i \in \mathcal{G}_1) = P(i \in \mathcal{G}_2) = P(i \in \mathcal{G}_3) = 1/3$, so that $\beta_i = \rho_1$ for $i \in \mathcal{G}_1$, $\beta_i = \rho_2$ for $i \in \mathcal{G}_2$, and $\beta_i = \rho_3$ for $i \in \mathcal{G}_3$. To investigate the effect of the size of the difference among subgroup-specific treatment effects, we considered three values of $\boldsymbol{\rho}$:

Case 1: $\rho_1 = 1$, $\rho_2 = -1$ and $\rho_3 = 0$;

Case 2: $\rho_1 = 2$, $\rho_2 = -2$ and $\rho_3 = 0$;

Case 3: $\rho_1 = 4$, $\rho_2 = -4$ and $\rho_3 = 0$.

We chose sample sizes of $n = 100$ and 200 and censoring rates of 20% and 40% which corresponded to $\tau = 20$ and 1. We compared the performance of the estimators using the proposed BJ-ADMM algorithm with two concave penalties (SCAD and MCP) and the Lasso penalty. Following Ma and Huang (2016, 2017), we took $\varphi = 1$ and $a = 3$ for MCP and SCAD penalties. The optimal value of the tuning parameter λ was selected by minimizing the modified BIC in (3.9). All simulation results are based on 500 replications.

Figure 1 displays the fusiongrams, i.e., the solution paths for $\widehat{\beta}_1(\lambda), \ldots, \widehat{\beta}_n(\lambda)$ against λ using the SCAD, MCP and Lasso penalties under Case 3 of Example 1. For both SCAD and MCP, the method can provide nearly unbiased estimates, and when λ reaches around 0.8, the estimates of $(\beta_1, \ldots, \beta_n)$ are merged into the three groups at true values -4, 0 and 4. When λ exceeds 1.8, all the estimates of β_i's are shrunk into one single value. For the Lasso, the estimates of β_i's are quickly merged into one value starting from $\lambda = 0.2$ due to its tendency towards over-shrinkage.

To evaluate the proposed estimation procedure, we present the estimates $\widehat{R}$, $\widehat{\beta}_i$'s, $\widehat{\rho}_j$'s, and $\widehat{\eta}$ over 500 replications for each setting. Table 1 shows the mean, median, standard deviation of the estimated numbers of subgroups $\widehat{R}$ and the percentage of $\widehat{R}$ equal to the true number of groups by the SCAD and MCP shrinkage procedures. In Case 1 with censoring rate 20%, Case 2, and Case 3, the median of $\widehat{R}$ is always 3 which is the true number of subgroups. As sample size n increases, the mean moves closer to 3 and the standard deviation becomes smaller, and the percentage of correctly selecting the number of subgroups increases as well. The two concave penalties SCAD and MCP procedures have similar performance.

To examine the treatment effect estimates $\widehat{\beta}_i$, $i = 1, \ldots, n$, we plot $X_i\beta_i$, $X_i\widehat{\beta}_i$ and $X_i\widehat{\beta}_i^{BJ}$ against the values of X_i in Figure 2 under the SCAD method for $n = 100$ and a censoring rate of 20%, where β_i's are the true values, $\widehat{\beta}_i$'s are the estimated values by the proposed BJ-ADMM algorithm with SCAD, and $\widehat{\beta}_i^{BJ}$'s are the estimated values from the Buckley–James iterative procedure. It exhibits that the fitted lines by the BJ-ADMM with SCAD are close to the truth, while those by the Buckley–James iterative procedure center around the horizontal line $y = 0$, which deviate far away from the truth. Figure 3 exhibits the mean squared error (MSE) for the estimates of η, which also demonstrates the good performance of our method under different settings.

To further study the estimation accuracy of the subgroup-specific effects $\widehat{\rho}_r$, we compare the mean, median and standard deviation of the estimates $\widehat{\rho}_1$, $\widehat{\rho}_2$ and $\widehat{\rho}_3$ by the proposed method with the SCAD and MCP penalties and those of the oracle estimators in Table 2. Both the means and medians of the three versions of $(\widehat{\rho}_1, \widehat{\rho}_2, \widehat{\rho}_3)$ are close to the true values for all cases. As n increases, the biases decrease and the standard deviations also decrease, while the converse is true when the censoring rate increases. Moreover, the estimates using the SCAD and MCP penalties are similar, and both are close to the oracle results. In addition, the size of the difference between subgroup-specific treatment effects slightly influences the performance of the proposed method.

Example 2 (Multiple treatment variables). In this experiment, we generated data from a censored heterogenous linear regression model,

$$Y_i = Z_i^\top \eta + X_i^\top \beta_i + \epsilon_i, \quad i = 1, \dots, n,$$

where Z_i and η were generated in the same way as in Example 1, and $X_i = (X_{i1}, X_{i2})^\top$ was simulated from a bivariate standard normal distribution. We randomly assigned the responses to three groups with equal probabilities, i.e., $R = 3$ and $P(i \in \mathcal{G}_1) = P(i \in \mathcal{G}_2) = P(i \in \mathcal{G}_3) = 1/3$, so that $\beta_i = \rho_1$ for $i \in \mathcal{G}_1$, $\beta_i = \rho_2$ for $i \in \mathcal{G}_2$, and $\beta_i = \rho_3$ for $i \in \mathcal{G}_3$, where $\rho_1 = (4,4)^\top$, $\rho_2 = (-4,-4)^\top$ and $\rho_3 = (0,0)^\top$. In this experiment, we also consider the non-centralized quadratic loss function with fusion penalty, given by

$$\ell_P^*(\boldsymbol{\theta}; \lambda) = \frac{1}{2} \sum_{i=1}^{n} \{\widetilde{Y}_i(\theta_i, \widetilde{F}_{\boldsymbol{\theta}}) - Z_i^\top \eta - X_i^\top \beta_i\}^2 + \sum_{1 \le i < j \le n} P_\lambda(\|\beta_i - \beta_j\|). \tag{5.1}$$

As a result, four cases are examined to explore the effect of centralization as follows:

Case 1: $\epsilon_i \sim \mathcal{N}(0, 0.5^2)$ by $\ell_P(\boldsymbol{\theta}; \lambda)$ in (2.6);

Case 2: $\epsilon_i \sim \mathcal{N}(0, 0.5^2)$ by $\ell_P^*(\boldsymbol{\theta}; \lambda)$ in (5.1);

Case 3: $\epsilon_i \sim \mathcal{N}(1, 0.5^2)$ by $\ell_P(\boldsymbol{\theta}; \lambda)$ in (2.6);

Case 4: $\epsilon_i \sim \mathcal{N}(1, 0.5^2)$ by $\ell_P^*(\boldsymbol{\theta}; \lambda)$ in (5.1).

Figure 4 displays the fusiongrams for $\boldsymbol{\beta}_1 = (\beta_{11}, \dots, \beta_{1n})^\top$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \dots, \beta_{2n})^\top$ with $n = 100$ and a censoring rate of 20% under Case 3. It indicates that the BJ-ADMM methods with SCAD and MCP behave similarly and both are more suited for enforcing sparser subgroups in comparison to the Lasso penalty. Table 3 reports the mean, median and standard deviation of $\widehat{R}$ and the percentage of $\widehat{R}$ equal to the true number of subgroups by the BJ-ADMM procedure with the SCAD and MCP penalties based on 500 replicates in Case 3. The median of $\widehat{R}$ always matches the true number of subgroups which is 3, and the mean of $\widehat{R}$ is also close to 3. Moreover, the percentage of correctly selecting the true number of subgroups increases as the censoring rate becomes smaller or the sample size increases. Table 4 reports the mean, median and standard deviation (SD) of the root mean square errors (RMSE) of the estimator $\widehat{\boldsymbol{\rho}}$ with the formula $\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|/\sqrt{Rp}$ under the SCAD penalty over 500 replications with $n = 100, 200$, and censoring rates of $20\%, 40\%$ respectively under the four cases of Example 2. The results under Case 2 shows the best performance because the objective function $\ell_P^*(\boldsymbol{\theta}; \lambda)$ correctly reflects the parameter structure of the model, while $\ell_P^*(\boldsymbol{\theta}; \lambda)$ leads to invalid estimation in Case 4. Our centralized quadratic loss function with a fusion penalty, i.e., $\ell_P(\boldsymbol{\theta}; \lambda)$, always provides valid estimates of the group-specified coefficients. Furthermore, we evaluate the performance of the estimators $\widehat{\boldsymbol{\rho}} = (\widehat{\rho}_1^\top, \widehat{\rho}_2^\top, \widehat{\rho}_3^\top)^\top$ using the MSE with the formula $\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|/\sqrt{Rp}$. Figure 5 depicts the boxplots of the MSEs of $\widehat{\boldsymbol{\rho}}$ by the two concave penalties SCAD and MCP under censoring rates of 20% and 40% respectively under Case 3. The MSE decreases as the censoring rate decreases or the sample size increases for both SCAD and MCP, and the BJ-ADMM with SCAD and MCP procedures perform similarly in all settings.

Example 3 (Homogeneous treatment effect). In this experiment, we generated data from a censored homogeneous linear regression model,

$$Y_i = Z_i^\top \eta + X_i \beta + \epsilon_i, \quad i = 1, \ldots, n,$$

where Z_i, X_i, ϵ_i and η were generated in the same way as in Example 1. We took $\beta = 2$, sample size $n = 100$, and censoring rates of 20% and 40%. Besides the independent censoring as in the previous two examples, we also considered covariate-dependent censoring by generating C from $\mathcal{N}(\mu + X, 1)$, where μ controls the censoring rate.

Table 5 presents the simulation results of the estimate $\widehat{R}$ by the SCAD and MCP shrinkage procedures over 500 replicates. In all cases, the medians of $\widehat{R}$ are exactly 1, which implies a homogeneous treatment effect. Regardless of independent or covariate-dependent censoring mechanisms, all the means of the estimated numbers of subgroups are close to 1, and the standard deviation becomes smaller with a lower censoring rate. Moreover, the percentage of correctly selecting the true number of subgroups becomes higher as the censoring rate decreases. The two concave penalties SCAD and MCP perform equally well.

Furthermore, we considered a null hypothesis $H_0 : \beta_1 = \cdots = \beta_n = \beta^*$ with $\beta^* = 2$, for testing homogeneity, and applied the χ^2-test statistic,

$$\mathcal{T}_n^* = (\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^*)^\top (\widehat{\mathcal{V}}_{n11})^{-1} (\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^*),$$

where $\boldsymbol{\rho}^* = (1_{\widehat{R}} \otimes I_p)\beta^*$ and $1_{\widehat{R}}$ is a vector of length $\widehat{R}$ with all elements equal to 1. We calculated the average type I error rate based on 500 replications using $\frac{1}{500}\sum_{j=1}^{500} I\{\mathcal{T}_n^{*j} > \chi^2_{\widehat{R}}(0.95)\}$, where $\mathcal{T}_n^{*j}$ is the value of $\mathcal{T}_n^*$ from the jth replicate and $\chi^2_{\widehat{R}}(0.95)$ is the 0.95-quantile of the χ^2 distribution with $\widehat{R}$ degrees of freedom. We obtained the average type I error rate of 0.0552 and 0.0554 for SCAD and MCP, respectively, which are very close to the nominal significance level 0.05.

6. Application

As an illustration, we applied the proposed method to a real data set from a clinical trial in primary biliary cirrhosis (PBC) of the liver carried out in the Mayo Clinic (Fleming and Harrington 1991). Patients in the PBC data were randomized to two treatment groups:

D-penicillamine and placebo, and 16 baseline covariates were collected: age in years (z_1), sex (z_2), presence of ascites (z_3), presence of hepatomegaly (z_4), presence of spiders (z_5), presence of edema (z_6), serum bilirubin (z_7), serum cholesterol (z_8), albumin (z_9), urine copper (z_{10}), alkaline phosphatase (z_{11}), serum glutamic-oxaloacetic transaminase (z_{12}), triglycerides (z_{13}), histologic stage of disease (z_{14}), platelet count (z_{15}), and prothrombin time (z_{16}). After removing the missing data, we ended up with $n = 276$ observations. During the follow-up, 129 patients died and the other 147 patients were censored, leading to a censoring rate of 53%. We took the log-transformed survival time as the response variable Y_i, and considered a binary variable X for the two treatments ($X_i = 1$ for patients in the D-penicillamine group; $X_i = 0$ for patients in the placebo group).

To check the possible heterogeneity in treatment effects, we first fitted a censored homogeneous linear model, $Y_i = Z_i^\top \eta + \epsilon_i$, with $Z_i = (z_{i1}, \ldots, z_{i16})^\top$, using the Buckley–James estimation procedure. We then plotted the Kaplan–Meier kernel density estimate of residuals $\{(\delta_i, Y_i - Z_i^\top \widehat{\eta}^{BJ}) : X_i = 1, i = 1, \cdots, 276\}$, where $\widehat{\eta}^{BJ}$ is the Buckley–James estimator. Figure 6 shows that the distribution has multiple modes for these patients, which indicates possible heterogeneous treatment effects.

As a result, we considered the censored heterogeneous linear regression, $Y_i = Z_i^\top \eta + X_i \beta_i + \epsilon_i$. All covariates were standardized before applying the proposed method with the SCAD and MCP penalties. We selected the optimal tuning parameter $\widehat{\lambda} = 0.15$ for both SCAD and MCP penalties by minimizing the modified BIC defined in (3.9) respectively, and identified $\widehat{R} = 3$ major subgroups by our proposed BJ-ADMM algorithm. Figure 7 displays the fusiongrams for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^\top$ using the SCAD and MCP penalties, indicating the existence of heterogeneity in treatment effects.

In Table 6, we report the estimates $\widehat{\rho}_1$, $\widehat{\rho}_2$ and $\widehat{\rho}_3$ with the p-values for testing the significance of each component of the subgroup-specific treatment effects using the proposed method, and those using the standard Buckley–James method. The Buckley–James results show that the treatment had no statistically significant effect on the survival time. However, the BJ-ADMM methods with the MCP and SCAD suggest that the D-penicillamine treatment had significantly positive and negative subgroup-specific effects on the survival times of patients in the first and second groups respectively, but no effect in the third group.

7. Conclusion

To accommodate random censoring in survival data, the concave fusion penalized Buckley–James least squares approach is developed for simultaneously estimating the grouping structure and the subgroup-specific treatment effects in a heterogeneous linear regression model. Our BJ-ADMM algorithm with the SCAD or MCP penalty works well in both simulation and real data examples. It is possible to incorporate the modified Buckley–James estimator (Lai and Ying, 1991) to our method for dealing with the difficulties caused by the instability at the upper tail of the associated Kaplan–Meier estimator of the underlying error distribution. Extensions in other survival models, such as the Cox proportional hazards model (Zhang and Lu, 2007), additive hazards (Lin and Lv, 2013), or transformation models, are also worth pursuing.

Supplementary Materials

The Supplementary Materials include the proofs of Proposition 1 and Theorems 1–3.

Acknowledgements

The authors are very grateful to the Editor, the Associate Editor and the two referees for their valuable comments and suggestions that greatly improved the paper. Yin's research is supported in part by the Research Grant Council of Hong Kong (17307318). Zhao's research is supported in part by the Research Grant Council of Hong Kong (15301218), the National Natural Science Foundation of China (11771366) and The Hong Kong Polytechnic University.

References

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.

Bradic, J, Fan, J. and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.* **39**, 3092–3120.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66**, 429–36.

Cai, T. Huang, J. and Tian, L. (2009). Regularized estimation for the accelerated failure time model. *Biometrics* **65**, 394–404.

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

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

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*, Wiley, New York.

Huang, J., Sun, T., Ying, Z., Yu, Y. and Zhang, C.-H. (2013). Oracle inequalities for the LASSO in the Cox model. *Ann. Statist.* **41**, 1142–1165.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*, John Wiley, New York.

Ke, T., Fan, J. and Wu, Y. (2015). Homogeneity in regression. *J. Amer. Statist. Assoc.* **110**, 175–194.

Kravitz, R. L., D. N. and Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Q* **82**, 661–687.

Lagakos, S. W. (2006). The challenge of subgroup analysis: Reporting without distorting. *N. Engl. J. Med.* **354**, 1667–1669.

Lai, T. L. and Ying, Z. (1988). Stochastic integrals of empirical-type processes with applications to censored regression. *J. Multivar. Anal.* **27**, 334–358.

Lai, T. L. and Ying, Z. (1991). Rank regression methods for left-truncated and right-censored data. *Ann. Statist.* **19**, 31–546.

Lin, W. and Lv, J. (2013). High-dimensional sparse additive hazards regression. *J. Amer. Statist. Assoc.* **108**, 247–264.

Liu, X. and Zeng, D. (2013). Variable selection in semiparametric transformation models for right-censored data. *Biometrika* **100**, 859–876.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498–3528.

Ma, S. and Huang, J. (2017). A concave pairwise fusion approach to subgroup analysis. *J. Amer. Statist. Assoc.* **112**, 410–423.

Ma, S. and Huang, J. (2016). Estimating subgroup-specific treatment effects via concave fusion. *arXiv preprint arXiv:1607.03717*.

Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika* **69**, 521–531.

Rothwell, P. M. (2005). Subgroup analysis in randomized clinical trials: Importance, indications and interpretation. *Lancet* **365**, 176–186.

Shen, J. and He, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *J. Amer. Statist. Assoc.* **110**, 303–312.

Shen, X. and Huang, H. C. (2010). Grouping pursuit through a regularization solution surface. *J. Amer. Statist. Assoc.* **105**, 727–739.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc., Series B* **58**, 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statist. Med.* **16**, 385–395.

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

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

Zhang, H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94**, 691–703.

Xiaodong Yan, School of Economics, Shandong University, Jinan, China

E-mail: yanxiaodong@sdu.edu.cn

Guosheng Yin, Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

E-mail: gyin@hku.hk

Xingqiu Zhao, Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

E-mail: xingqiu.zhao@polyu.edu.hk

Table 1: The mean, median and standard deviation (SD) of $\widehat{R}$ and the percentage of $\widehat{R}$ equal to the true number of subgroups, $P(\widehat{R} = R)$, by the BJ-ADMM algorithm with the MCP and SCAD penalties based on 500 replications with $n = 100, 200$, and censoring rates of $20\%, 40\%$ respectively in Example 1.

			BJ-ADMM+SCAD				BJ-ADMM+MCP			
Case	n	Censoring	Mean	Median	SD	$P(\widehat{R} = R)$	Mean	Median	SD	$P(\widehat{R} = R)$
Case 1	100	20%	3.62	3	0.672	0.85	3.64	3	0.675	0.84
		40%	3.86	3.5	0.708	0.80	3.89	3.5	0.710	0.80
	200	20%	3.48	3	0.587	0.88	3.51	3	0.591	0.87
		40%	3.60	3.5	0.626	0.83	3.63	3.5	0.630	0.82
Case 2	100	20%	3.23	3	0.330	0.94	3.25	3	0.332	0.93
		40%	3.45	3	0.358	0.89	3.47	3	0.360	0.89
	200	20%	3.11	3	0.267	0.96	3.13	3	0.269	0.96
		40%	3.21	3	0.210	0.92	3.22	3	0.213	0.91
Case 3	100	20%	3.04	3	0.131	1.00	3.05	3	0.134	1.00
		40%	3.09	3	0.148	0.96	3.10	3	0.157	0.95
	200	20%	3.01	3	0.087	1.00	3.01	3	0.088	1.00
		40%	3.04	3	0.096	0.98	3.03	3	0.095	0.97

Table 2: The mean, median and standard deviation (SD) of the estimators $\widehat{\rho}_1$, $\widehat{\rho}_2$ and $\widehat{\rho}_3$ by the SCAD and MCP penalties and the oracle (OR) estimators over 500 replications with $n = 100, 200$, and censoring rates of $20\%, 40\%$ respectively in Example 1.

		$n = 100$						$n = 200$					
		Censoring = 20%			Censoring = 40%			Censoring = 20%			Censoring = 40%		
Case		Mean	Median	SD	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Case 1	$\widehat{\rho}_1$(SCAD)	1.076	1.072	0.074	1.112	1.107	0.125	1.046	1.042	0.051	1.087	1.082	0.106
	$\widehat{\rho}_1$(MCP)	1.082	1.076	0.078	1.124	1.120	0.129	1.048	1.045	0.053	0.059	1.085	0.109
	$\widehat{\rho}_1$(OR)	1.050	1.048	0.050	1.053	1.059	0.086	1.026	1.022	0.027	1.047	1.042	0.076
	$\widehat{\rho}_2$(SCAD)	-0.913	-0.924	0.092	-0.897	-0.902	0.133	-0.938	-0.943	0.072	-0.917	-0.919	0.093
	$\widehat{\rho}_2$(MCP)	-0.910	-0.916	0.095	-0.895	-0.900	0.135	-0.934	-0.939	0.076	-0.913	-0.916	0.097
	$\widehat{\rho}_2$(OR)	-1.065	-1.053	0.073	-1.098	-1.079	0.098	-1.023	-1.021	0.051	-1.049	-1.043	0.080
	$\widehat{\rho}_3$(SCAD)	0.024	0.020	0.031	0.039	0.033	0.057	0.011	0.009	0.018	0.021	0.019	0.037
	$\widehat{\rho}_3$(MCP)	0.022	0.019	0.034	0.041	0.039	0.060	-0.013	-0.011	0.019	-0.024	-0.021	0.042
	$\widehat{\rho}_3$(OR)	0.012	0.010	0.019	0.027	0.025	0.035	-0.007	-0.005	0.010	-0.011	-0.009	0.021
Case 2	$\widehat{\rho}_1$(SCAD)	2.049	2.044	0.044	2.089	2.086	0.092	2.016	2.012	0.031	2.047	2.042	0.056
	$\widehat{\rho}_1$(MCP)	2.052	2.048	0.048	2.092	2.090	0.093	2.018	2.015	0.033	2.049	2.045	0.059
	$\widehat{\rho}_1$(OR)	2.020	2.028	0.040	2.043	2.049	0.076	2.008	2.006	0.017	2.017	2.012	0.046
	$\widehat{\rho}_2$(SCAD)	-1.953	-1.964	0.052	-1.917	-1.919	0.083	-1.978	-1.983	0.032	-1.957	-1.959	0.053
	$\widehat{\rho}_2$(MCP)	-1.950	-1.956	0.055	-1.915	-1.918	0.085	-1.974	-1.979	0.036	-1.953	-1.956	0.057
	$\widehat{\rho}_2$(OR)	-2.015	-2.013	0.030	-2.058	-2.059	0.058	-2.009	-2.006	0.021	-2.029	-2.023	0.060
	$\widehat{\rho}_3$(SCAD)	0.008	0.007	0.011	0.019	0.023	0.027	0.005	0.003	0.008	0.014	0.012	0.017
	$\widehat{\rho}_3$(MCP)	0.009	0.006	0.013	0.021	0.025	0.029	-0.007	-0.005	0.009	-0.016	-0.014	0.012
	$\widehat{\rho}_3$(OR)	0.005	0.004	0.009	0.013	0.010	0.015	-0.003	-0.002	0.003	-0.06	-0.005	0.008
Case 3	$\widehat{\rho}_1$(SCAD)	3.989	3.996	0.034	3.919	3.924	0.069	3.995	3.997	0.020	3.937	3.942	0.036
	$\widehat{\rho}_1$(MCP)	3.987	3.994	0.036	3.916	3.922	0.070	3.992	3.994	0.019	3.936	3.940	0.037
	$\widehat{\rho}_1$(OR)	3.991	3.998	0.031	3.923	3.934	0.066	3.998	3.999	0.016	3.954	3.960	0.032
	$\widehat{\rho}_2$(SCAD)	-3.982	-3.984	0.041	-3.921	-3.925	0.073	-3.989	-3.991	0.023	-3.951	-3.959	0.042
	$\widehat{\rho}_2$(MCP)	-3.980	-3.983	0.045	-3.920	-3.922	0.075	-3.988	-3.992	0.026	-3.951	-3.955	0.044
	$\widehat{\rho}_2$(OR)	-3.989	-3.991	0.038	-3.931	-3.934	0.068	-3.994	-3.998	0.020	-3.976	-3.980	0.039
	$\widehat{\rho}_3$(SCAD)	-0.004	0.003	0.011	0.009	0.013	0.017	-0.002	-0.003	0.006	-0.006	-0.007	0.013
	$\widehat{\rho}_3$(MCP)	-0.002	0.003	0.014	0.010	0.011	0.018	-0.001	-0.002	0.006	-0.006	-0.005	0.012
	$\widehat{\rho}_3$(OR)	-0.000	0.001	0.009	0.004	0.005	0.015	-0.000	-0.000	0.003	-0.004	-0.003	0.009

Table 3: The mean, median and standard deviation (SD) of $\widehat{R}$ and the percentage of $\widehat{R}$ equal to the true number of subgroups, $P(\widehat{R} = R)$, by the MCP and SCAD penalties based on 500 replications with $n = 100, 200$, and censoring rates of $20\%, 40\%$ respectively in Case 3 of Example 2.

		BJ-ADMM+SCAD				BJ-ADMM+MCP			
n	Censoring	Mean	Median	SD	$P(\widehat{R} = R)$	Mean	Median	SD	$P(\widehat{R} = R)$
100	20%	3.11	3	0.423	0.91	3.26	3	0.412	0.92
	40%	3.36	3	0.502	0.85	3.40	3	0.602	0.88
200	20%	3.06	3	0.223	0.95	3.11	3	0.302	0.94
	40%	3.16	3	0.372	0.89	3.20	3	0.451	0.90

Table 4: The mean, median and standard deviation (SD) of RMSEs of the estimators $\widehat{\boldsymbol{\rho}}$ with the formula $\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|/\sqrt{Rp}$ under the SCAD penalty over 500 replications with $n = 100, 200$, and censoring rates of $20\%, 40\%$ respectively under the four Cases of Example 2.

	$n = 100$						$n = 200$					
	Censoring = 20%			Censoring = 40%			Censoring = 20%			Censoring = 40%		
Case	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Case 1	0.026	0.029	0.039	0.059	0.064	0.088	0.018	0.016	0.028	0.035	0.034	0.053
Case 2	0.017	0.016	0.021	0.038	0.044	0.052	0.009	0.006	0.014	0.023	0.021	0.031
Case 3	0.041	0.047	0.070	0.079	0.084	0.116	0.029	0.031	0.041	0.053	0.054	0.082
Case 4	1.834	1.947	2.882	2.419	2.528	3.062	1.589	1.638	2.031	2.011	1.927	2.421

Table 5: The mean, median and standard deviation (SD) of $\widehat{R}$ and the percentage of $\widehat{R}$ equal to the true number of subgroups, $P(\widehat{R} = R)$, by the MCP and SCAD penalties based on 500 replications with $n = 100$ and censoring rates of $20\%, 40\%$ respectively in Example 3.

		BJ-ADMM+SCAD				BJ-ADMM+MCP			
Mechanism	Rate	Mean	Median	SD	$P(\widehat{R} = R)$	Mean	Median	SD	$P(\widehat{R} = R)$
Independent	20%	1.12	1	0.137	0.97	1.10	1	0.132	0.96
	40%	1.19	1	0.242	0.95	1.20	1	0.237	0.94
Dependent	20%	1.17	1	0.152	0.96	1.15	1	0.149	0.96
	40%	1.25	1	0.207	0.91	1.26	1	0.196	0.93

Table 6: The estimates and p-values of $\widehat{\rho}_1$, $\widehat{\rho}_2$ and $\widehat{\rho}_3$ by the BJ-ADMM with the MCP and SCAD methods, and those of $\widehat{\beta} = \widehat{\rho}_1$ by the Buckley–James method for the PBC data.

Method	Result	$\widehat{\rho}_1$	$\widehat{\rho}_2$	$\widehat{\rho}_3$
BJ-ADMM+SCAD	Estimate	0.767	-0.567	0.003
	p-value	0.006	0.009	0.708
BJ-ADMM+MCP	Estimate	0.769	-0.566	0.003
	p-value	0.005	0.009	0.708
Buckley–James	Estimate	0.003		
	p-value	0.710		

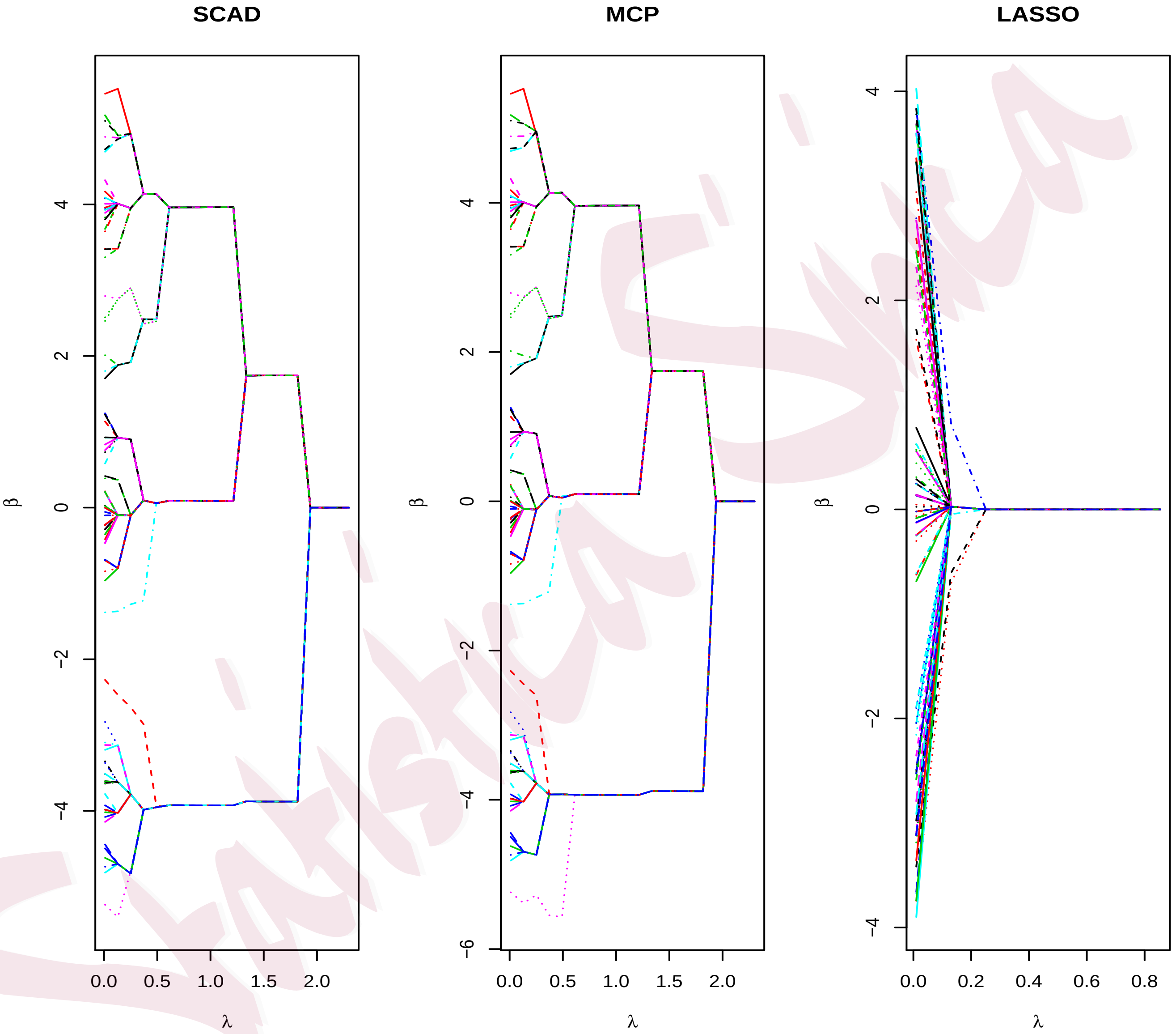

Figure 1: Fusiongrams (or solution paths) of $\widehat{\beta}_1(\lambda), \ldots, \widehat{\beta}_n(\lambda)$ versus λ for Case 3 of Example 1 with $n = 100$ and censoring rate 40% under three different penalties SCAD, MCP, and Lasso.

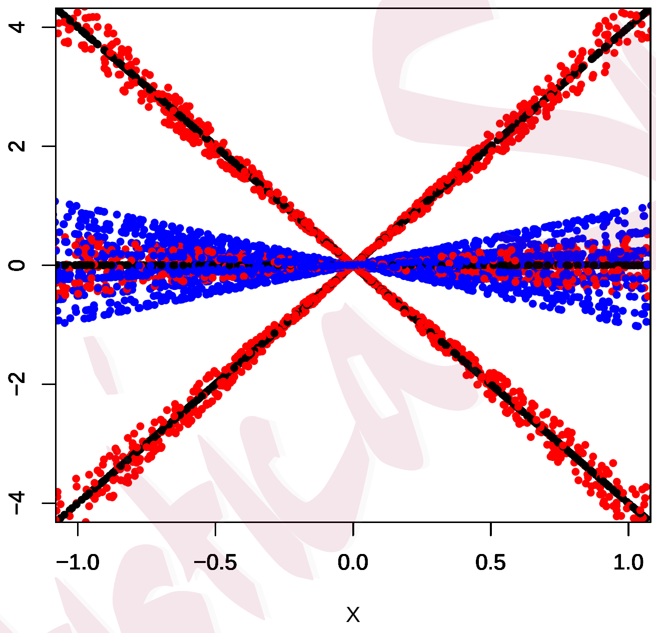

Figure 2: Plots of $X_i\beta_i$ (black dotted lines), $X_i\widehat{\beta}_i$ (red dotted lines) and $X_i\widehat{\beta}_i^{BJ}$ (blue dotted lines) versus values of X_i, where β_i's are the true values, $\widehat{\beta}_i$'s are the estimated values by BJ-ADMM+SCAD and $\widehat{\beta}_i^{BJ}$'s are the estimated values using the Buckley–James iterative procedure for Case 3 of Example 1.

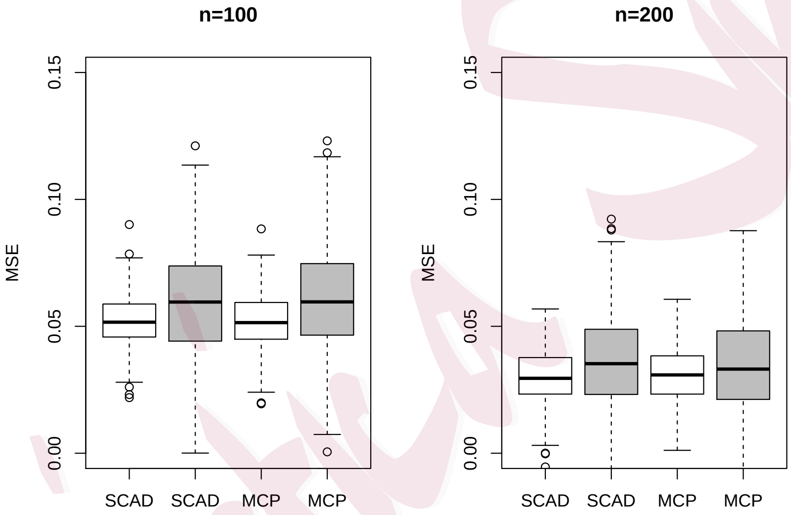

Figure 3: Boxplots of the MSEs of $\widehat{\eta}$ using BJ-ADMM+SCAD and BJ-ADMM+MCP with $n = 100, 200$, and censoring rates of 20% (white) and 40% (grey) respectively for Case 3 of Example 1.

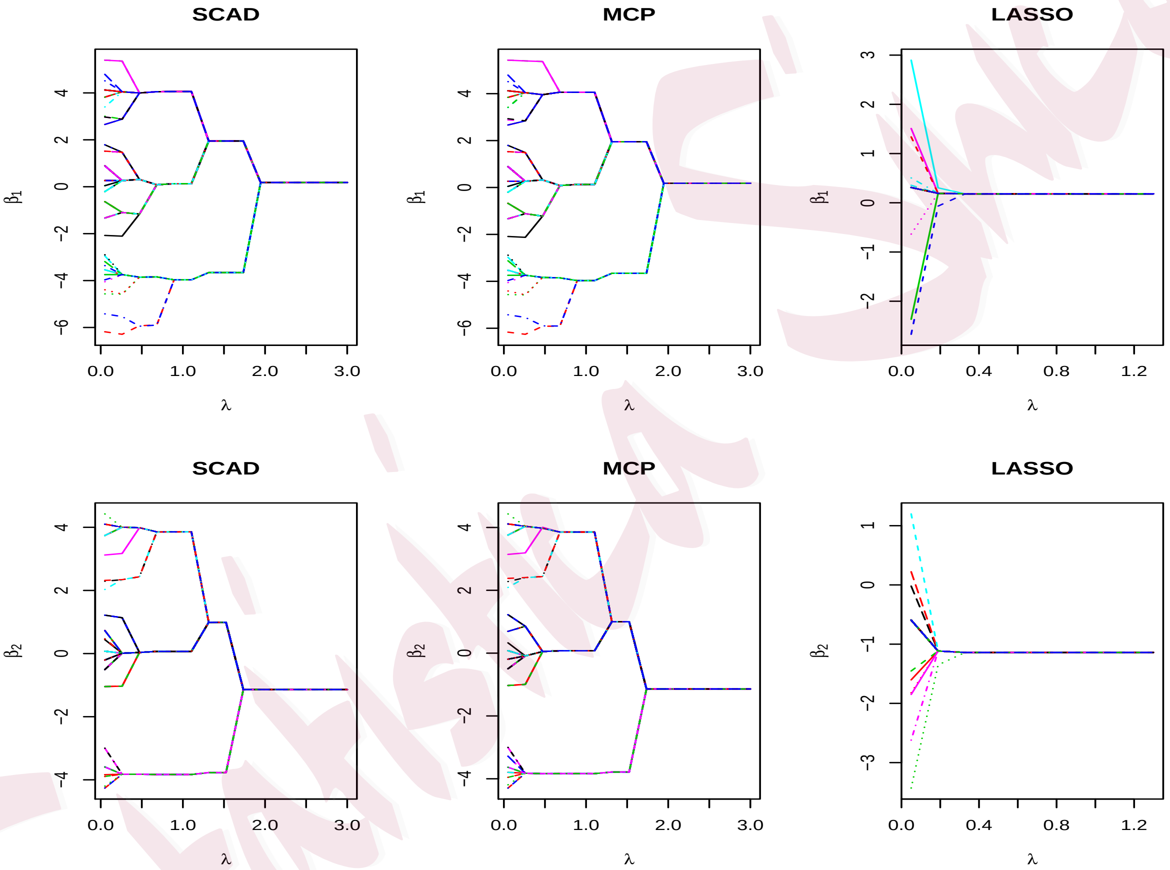

Figure 4: Fusiongrams of $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1n})^\top$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2n})^\top$ with $n = 100$ and censoring rate 20% in Example 2.

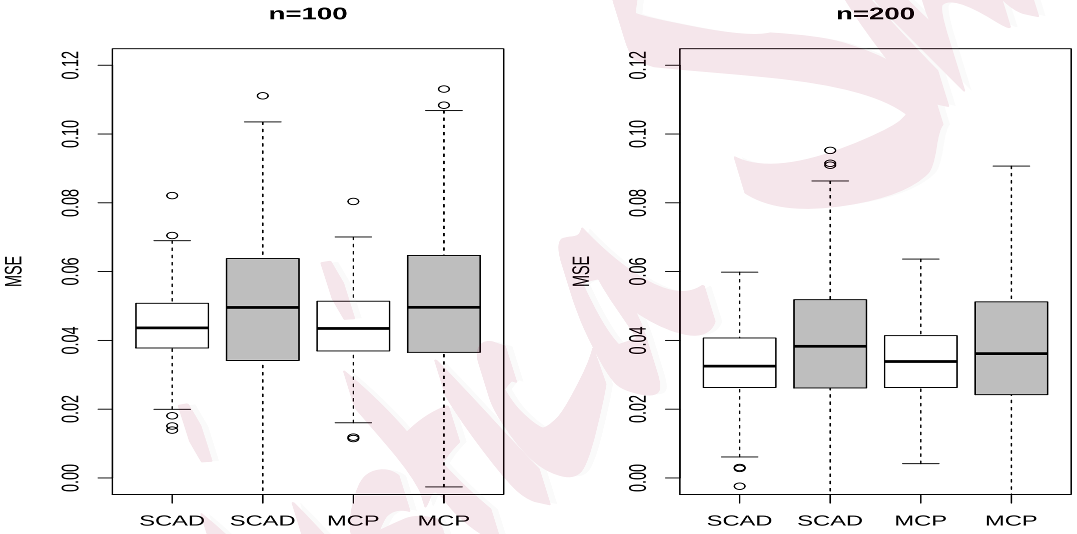

Figure 5: Boxplots of the MSEs of $\widehat{\boldsymbol{\rho}}$ using BJ-ADMM+SCAD and BJ-ADMM+MCP with $n = 100, 200$, and censoring rates of 20% (white) and 40% (grey) respectively in Example 2.

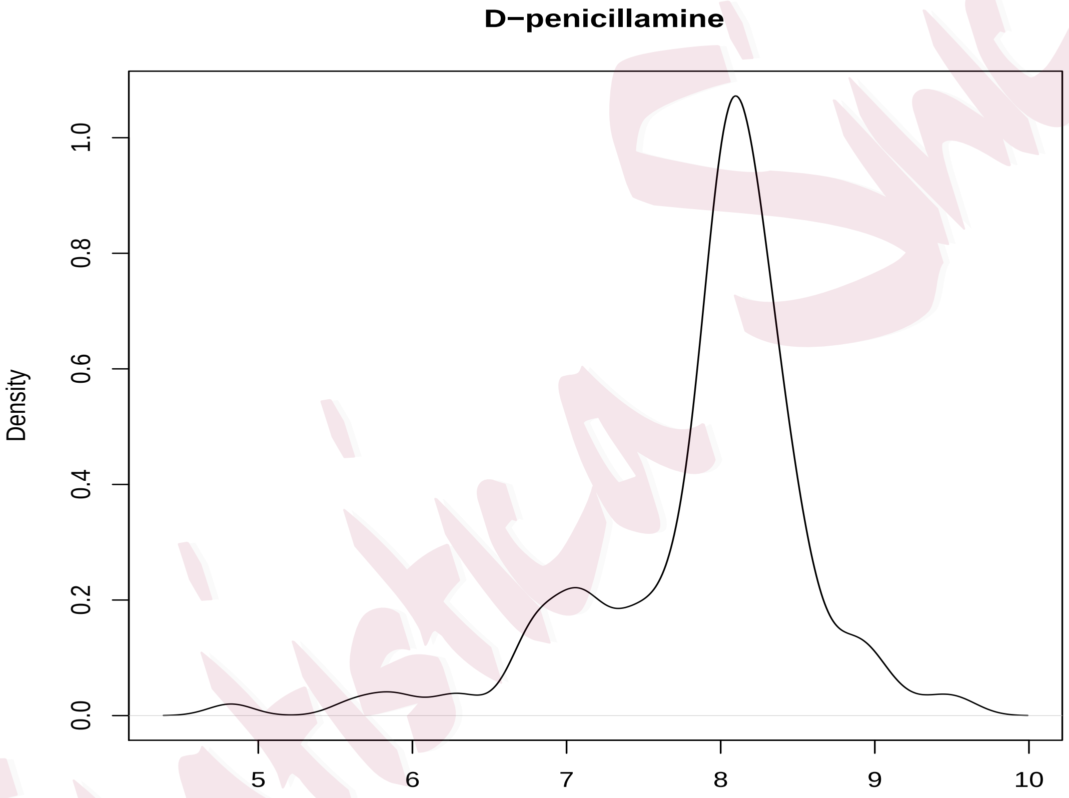

Figure 6: The kernel density plot of the residuals after controlling for the effects of the 16 baseline covariates for the patients treated with D-penicillamine in the PBC data.

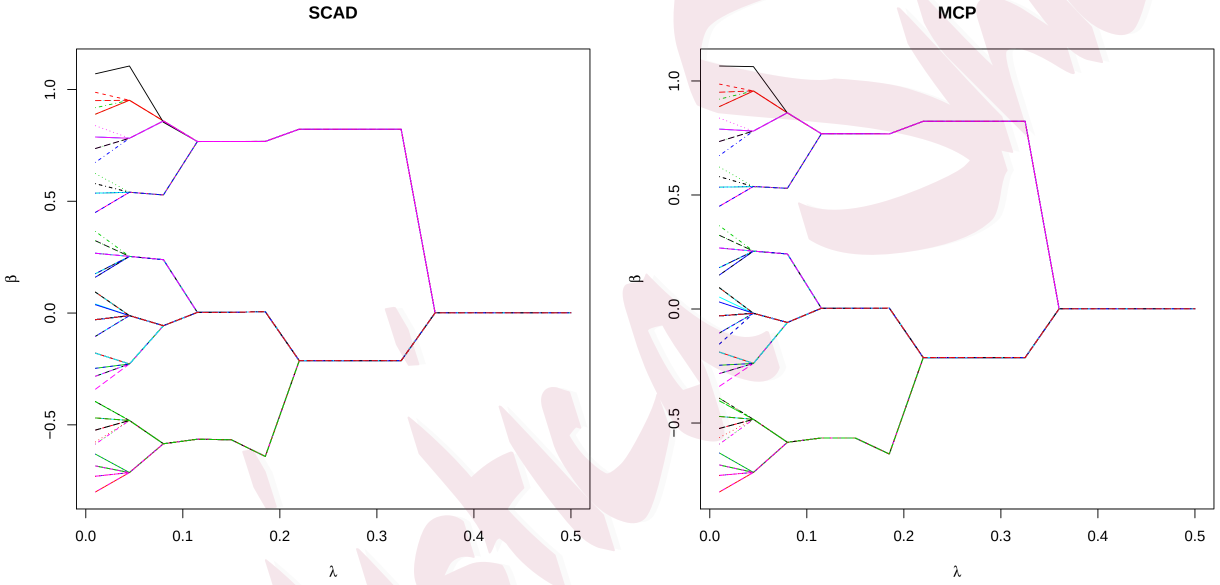

Figure 7: Fusiongrams of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^\top$ using the proposed BJ-ADMM with the SCAD and MCP penalties for the PBC data.