MULTI-THRESHOLD ACCELERATED FAILURE TIME MODEL[1]

BY JIALIANG LI AND BAISUO JIN

National University of Singapore and
University of Science and Technology of China

A two-stage procedure for simultaneously detecting multiple thresholds and achieving model selection in the segmented accelerated failure time (AFT) model is developed in this paper. In the first stage, we formulate the threshold problem as a group model selection problem so that a concave 2-norm group selection method can be applied. In the second stage, the thresholds are finalized via a refining method. We establish the strong consistency of the threshold estimates and regression coefficient estimates under some mild technical conditions. The proposed procedure performs satisfactorily in our simulation studies. Its real world applicability is demonstrated via analyzing a follicular lymphoma data.

1. Introduction. Applied economists routinely test their models for the presence of structural change. If the evidence supports it, a threshold model is constructed and one needs to detect the thresholds (also called break points) at which to split the sample. The threshold variable may be an element of regressors. One such example is the well-known threshold autoregressive model [see Tong (2012)]. Sometimes the threshold variable is simply the index of observed samples (e.g., time in a time series model) and the model is commonly referred to as the change point model or segmented regression model [see Yao and Au (1989), Perron (2006), Fearnhead and Vasileiou (2009)].

In this paper, we focus on the setting of multiple thresholds which is a much more challenging problem than a single break-point detection [Hansen (2000)]. A number of issues arise in the presence of multiple change points. These include the determination of the number of breaks, estimation of the thresholds given the number and statistical analysis of the resulting estimators. There exists a rich literature on this subject. For example, Inclán and Tiao (1994) identified multiple change-points of variance using the iterated cumulative sum of squares (ICSS) algorithm. Bai and Perron (2003) developed the dynamic programming principle

for the estimation of multiple change-points in linear regression. Following the familiar idea of penalized estimation, Harchaoui and Lévy-Leduc (2010) used the least absolute shrinkage and selection operator (LASSO) algorithm to estimate the locations of multiple change-points, in one-dimensional piecewise constant signals. Davis, Lee and Rodriguez-Yam (2006) proposed a genetic algorithm to detect multiple break points, while recently Jin, Shi and Wu (2013) considered nonconcave penalty functions including the smoothly clipped absolute deviation (SCAD) penalty and minimax concave penalty (MCP) penalty in piecewise stationary autoregressive processes. However, computational procedure and theoretical justification in Jin, Shi and Wu (2013) cannot be easily extended to survival analysis. We will consider a more difficult setting with censored event time in this paper. The theoretical results established in this paper are also more general than those in Jin, Shi and Wu (2013).

In life-testing research studies, single change point problem has been addressed by many authors. Specifically, Luo, Turnbull and Clark (1997) considered the Cox model with a change point at an unknown time and established asymptotic results for maximum partial likelihood estimates. Pons (2003) studied unknown threshold of a predictor variable under Cox model using counting processes theory while Kosorok and Song (2007) further examined the more sophisticated linear transformation models and provided necessary inference tools. Very few authors considered the problem of estimating multiple thresholds for survival regression analysis. Censored lifetime data break down the usual estimation framework for completely observed data. We usually cannot attain a closed-form solution for regression estimates and, therefore, face an increased complexity of estimating thresholds accurately. To address this under-developed issue, we consider the accelerated failure time (AFT) model as a typical example of regression models in this paper and contribute a new methodology on change-point problem for survival data analysis.

The AFT model permits a direct assessment of the covariate effects on the survival time, facilitating the interpretation of regression coefficients for the mean response. There are many estimation methods available for AFT model in the literature, including Buckley and James (1979), Prentice (1978), Tsiatis (1990), Ying (1993), Lin, Wei and Ying (1998). Many estimation methods for right censored data are rank-based and in practice the estimating functions may be discontinuous, producing challenges to the computation. In contrast, Stute (1993, 1996) proposed a weighted least squares estimator for AFT model, and established the consistency and asymptotic normality under technical conditions. Huang, Ma and Xie (2006) carried out a variable selection procedure and estimation in AFT model with high-dimensional covariates based on Stute's estimator. Xia et al. (2016) also employed Stute's estimator for nonparametric variable screening and selection. In this paper, we formally adopt the Stute estimator to study the parameter estimation for the AFT model with s thresholds, where $s \geq 0$ is unspecified. Using a two-stage procedure proposed in this paper, we may estimate the thresholds and the regression

coefficients simultaneously. Interestingly, our procedure may be straightforwardly extended to incorporate variable selection for high-dimensional data analysis.

The paper is arranged as follows. In Section 2, the multiple break-point problem for the AFT model is formulated. We then propose a two-stage procedure to detect the thresholds and estimate model parameters. In Section 3, we establish theoretical properties of our procedure. Next, simulation studies are conducted in Section 4 to examine the performance of our methods. An empirical application to a follicular lymphomais data is presented in Section 5.

Throughout the paper, $\mathbf{1}_q = (1, \ldots, 1)^\top$ is a q-dimensional constant vector, I_q is the $q \times q$ identity matrix, $1_{\{\cdot\}}$ is an indicator function, $A^\top$ is the transpose of a matrix A, and $\lfloor c \rfloor$ is the integer part of a real number c. For a vector $\mathbf{a}$, $\mathbf{a}^\top$ is its transpose, a_j is its jth component, $|\mathbf{a}|$ and $\|\mathbf{a}\|$ are respectively its L_1-norm and L_2-norm. If $\mathcal{A}$ is a set, its complement and its size are denoted by $\mathcal{A}^c$ and $\sharp\mathcal{A}$, respectively. In addition, "$\to_{\text{a.s.}}$" denotes convergence with probability 1.

2. Methodology for multiple thresholds under the AFT model.

Let T_i, $i = 1, \ldots, n$ be the independent logarithm of the failure time and $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. p-dimensional regressors. Assume $(T_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, satisfy the following AFT model with s thresholds located at $a_1 < \cdots < a_s$:

$$
\begin{aligned}
T_i &= \sum_{j=1}^{s+1} \mathbf{X}_i^\top \boldsymbol{\beta}_j^* 1_{\{a_{j-1} < Z_i \leq a_j\}} + \varepsilon_i \\
&= \mathbf{X}_i^\top \left[\boldsymbol{\beta}_1^* + \sum_{\ell=1}^{s} \mathbf{d}_\ell^* 1_{\{a_\ell < Z_i \leq a_{s+1}\}} \right] + \varepsilon_i, \qquad i = 1, \ldots, n
\end{aligned}
$$

(2.1)
$$
= \begin{cases}
\mathbf{X}_i^\top \boldsymbol{\beta}_1^* + \varepsilon_i & \text{if } a_0 < Z_i \leq a_1, \\
\mathbf{X}_i^\top (\boldsymbol{\beta}_1^* + \mathbf{d}_1^*) + \varepsilon_i & \text{if } a_1 < Z_i \leq a_2, \\
\cdots & \cdots \\
\mathbf{X}_i^\top \left(\boldsymbol{\beta}_1^* + \sum_{\ell=1}^{s} \mathbf{d}_\ell^* \right) + \varepsilon_i & \text{if } a_s < Z_i \leq a_{s+1},
\end{cases}
$$

where $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_{s+1}^*$ are unknown p-dimensional regression coefficients for $s + 1$ subgroups, $\mathbf{d}_\ell^* = \boldsymbol{\beta}_{\ell+1}^* - \boldsymbol{\beta}_\ell^*$, $\ell = 1, \ldots, s$, are the increments of coefficients between two adjacent subgroups, $s \geq 0$ is the unknown number of thresholds, Z_i is the thresholding variable, $a_1, \ldots, a_s$ are unknown threshold locations, $a_0 = 0$, $a_{s+1} = n$, $\varepsilon_1, \ldots, \varepsilon_n$ are independent random errors and $\varepsilon_{a_{\ell-1}+1}, \ldots, \varepsilon_{a_\ell}$, the subgroup in the series $\{\varepsilon_i, i = 1, \ldots, n\}$ separated by the threshold locations $\{a_\ell, \ell = 1, \ldots, s\}$, are identically distributed with mean zero and variance σ_ℓ^2, $\ell = 1, \ldots, s + 1$. Notice that if $s = 0$, the model (2.1) does not involve threshold. If $s \geq 1$, we denote $P(Z_i \leq a_j) = \tau_j$, $j = 1, \ldots, s$ where $0 < \tau_1 < \tau_2 < \cdots < \tau_s < 1$. If $Z_i = i$, we retain the usual change point model, and the condition $P(Z_i \leq a_j) =$
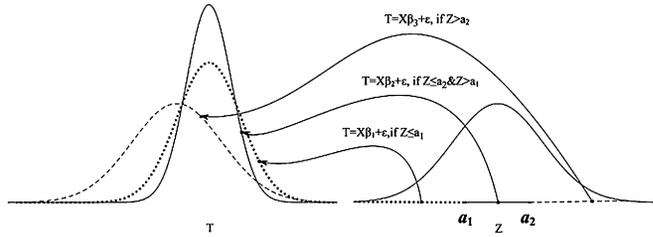
FIG. 1.   *An example of model* (2.1).

τ_j will be replaced by $a_j/n \to \tau_j$ since change point locations actually depend on the sample size in this setting. Figure 1 is an example of model (2.1), where $s = 2$, Z is generated from a normal distribution and T is generated from a mixture normal distribution.

Let C_i, $i = 1, \ldots, n$ be logarithm of the censoring time which are assumed to be independent and identically distributed (i.i.d.). In practice, we only observe $(Y_i, \delta_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, where $Y_i = \min(T_i, C_i)$ is the censored logarithm of the failure time and $\delta_i = 1_{\{T_i \leq C_i\}}$ is the censoring indicator.

Throughout this paper, let $Y_{(1,\mathcal{I})} \leq \cdots \leq Y_{(b,\mathcal{I})}$ be the ordered failure times in the index set $\mathcal{I}$ and $b = \sharp \mathcal{I}$. Let $\delta_{[\ell,\mathcal{I}]}$ be the concomitant of the ℓth-ordered $Y_{(\ell,\mathcal{I})}$, $\ell = 1, \ldots, b$. Define the Kaplan–Meier weights as

$$w_{[1,\mathcal{I}]} = \frac{\delta_{[1,\mathcal{I}]}}{b}, \qquad w_{[\ell,\mathcal{I}]} = \frac{\delta_{[\ell,\mathcal{I}]}}{b - \ell + 1} \prod_{k=1}^{\ell-1} \left(\frac{b-k}{b-k+1} \right)^{\delta_{[k,\mathcal{I}]}}, \qquad \ell = 2, \ldots, b.$$

Such weights are constructive elements of the Kaplan–Meier estimator for the survival function [see Lawless (2011)]. Let $r_{Y_i,\mathcal{I}} \in [1, b]$ be the rank of Y_i among $\{Y_i : i \in \mathcal{I}\}$, that is, $Y_i = Y_{(r_{Y_i,\mathcal{I}},\mathcal{I})}$, $i \in \mathcal{I}$. For simplicity, we denote $Y_{(r_{Y_i,\mathcal{I}})} = Y_{(r_{Y_i,\mathcal{I}},\mathcal{I})} = Y_i$ and $w_{[r_{Y_i,\mathcal{I}}]} = w_{[r_{Y_i,\mathcal{I}},\mathcal{I}]}$.

If a_j, $j = 1, \ldots, s$ are known, define $\mathcal{I}_j^* = \{i : a_{j-1} < Z_i \leq a_j\}$ and $b_j^* = \sharp \mathcal{I}_j^*$, we can use the Stute estimator to fit the AFT model. The resulting estimator $\hat{\boldsymbol{\beta}}_1^*$, $\hat{\mathbf{d}}_1^*, \ldots, \hat{\mathbf{d}}_s^*$ may be represented by a weighted least squares estimator that minimizes

$$(2.2) \qquad \sum_{j=1}^{s+1} \frac{b_j^*}{2n} \sum_{i \in \mathcal{I}_j^*} w_{[r_{Y_i,\mathcal{I}_j^*}]} \left(Y_i - \mathbf{X}_i^\top \left(\boldsymbol{\beta}_1^* + \sum_{k=1}^{j-1} \mathbf{d}_k^* \right) \right)^2.$$

We note that the weights $\{w_{[r_{Y_i,\mathcal{I}_j^*}]}\}$ depend on the order of the random failure times, and hence the related asymptotic argument is not as standard as the independent and identically distributed (i.i.d.) case in linear regression.

We intend to estimate s, $a_1, \ldots, a_s$ and then use the detected change point(s) to obtain the Stute estimator for regression coefficients. To this end, first we need to decide the total number of thresholds and the approximate distances between

pairs of thresholds. After that, we may proceed to determine the exact locations of thresholds. We propose a two-stage procedure in the following sections.

2.1. *The splitting stage.* First, we split the data sequence into $q_n + 1$ segments where q_n tends to infinity as $n \to \infty$. Let $\{Z_i : \delta_i = 1, i = 1, \ldots, n\} = \{\tilde{Z}_1, \ldots, \tilde{Z}_{n^*}\}$ where $n^* = \sum_{i=1}^n \delta_i$ is the total number of events. In order to allow each segment to include enough failure event observations, the data sequence is split such that the first segment $\mathcal{I}_1 = \{i : Z_i \leq \tilde{Z}_{(n^*-q_n m)}\}$ involves $n^* - q_n m$ events, and each of the other q_n segments $\mathcal{I}_j = \{i : \tilde{Z}_{(n^*-(q_n-j+2)m)} < Z_i \leq \tilde{Z}_{(n^*-(q_n-j+1)m)}\}$, $j = 2, \ldots, q_n + 1$ involve m events where $m = \lceil n^*/q_n \rceil$ and $\tilde{Z}_{(1)} \leq \tilde{Z}_{(2)} \cdots \leq \tilde{Z}_{(n^*)}$ are the ordered threshold variables associated with subjects no failed. Define $b_j = \sharp \mathcal{I}_j$, $j = 1, \ldots, q_n + 1$.

To select these intervals covering thresholds $\{a_j\}$, we propose a concave 2-norm group selection method, such that an estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1^\top, \hat{\mathbf{d}}_1^\top, \ldots, \hat{\mathbf{d}}_{q_n}^\top)^\top$ is given by minimizing

$$(2.3) \quad \sum_{j=1}^{q_n+1} \frac{b_j}{2n} \sum_{i \in \mathcal{I}_j} w_{[r_{Y_i}, \mathcal{I}_j]} \left(Y_i - \mathbf{X}_i^\top \left(\boldsymbol{\beta}_1 + \sum_{k=1}^{j-1} \mathbf{d}_k\right)\right)^2 + \sum_{k=1}^{q_n} p_{\lambda_n, \gamma_n}(\|\mathbf{d}_k\|),$$

where $\lambda_n > 0$, $\gamma_n > 1$ are tuning parameters, and the penalty function $p_{\lambda_n, \gamma_n}(|u|) > 0$ is concave in $|u|$. For the simplicity of presentation, we shall only consider two well-studied penalty functions in this paper, namely the smoothly clipped absolute deviation (SCAD) penalty and the minimax concave penalty (MCP). SCAD was introduced in Fan and Li (2001) and defined as $p_{\lambda, \gamma}(u) = \lambda u I_{[0,\lambda]}(u) + \frac{\gamma \lambda u - 0.5(u^2 + \lambda^2)}{\gamma - 1} I_{(\lambda, \gamma \lambda]}(u) + \frac{\lambda^2(\gamma^2 - 1)}{2(\gamma - 1)} I_{(\gamma \lambda, \infty)}(u)$, $\gamma > 1$. MCP was introduced in Zhang (2010) and defined as $p_{\lambda, \gamma}(u) = (\lambda u - \frac{u^2}{2\gamma}) I_{[0, \gamma \lambda]}(u) + \frac{1}{2}\gamma \lambda^2 I_{(\gamma \lambda, \infty)}(u)$, $\gamma > 2$.

Let $\mathbf{Y}_{(j)} = (Y_i, i \in \mathcal{I}_j)^\top$, $\mathbf{X}_{(j)} = (\mathbf{X}_i, i \in \mathcal{I}_j)^\top$, $\tilde{\mathbf{w}}_{(j)} = b_j(w_{[r_{Y_i}, \mathcal{I}_j]}, i \in \mathcal{I}_j)^\top$, $\tilde{\mathbf{y}}_{(j)} = \text{diag}(\tilde{\mathbf{w}}_{(j)})^{1/2}\mathbf{Y}_{(j)}$, $\tilde{\mathbf{X}}_{(j)} = \text{diag}(\tilde{\mathbf{w}}_{(j)})^{1/2}\mathbf{X}_{(j)}$, $j = 1, \ldots, q_n + 1$. Denote $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}}_{(1)}^\top, \ldots, \tilde{\mathbf{y}}_{(q_n+1)}^\top)^\top$, $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}^{(1)}, \ldots, \tilde{\mathbf{X}}^{(q_n+1)})$ where $\tilde{\mathbf{X}}^{(1)} = (\tilde{\mathbf{X}}_{(1)}^\top, \ldots, \tilde{\mathbf{X}}_{(q_n+1)}^\top)^\top$ and $\tilde{\mathbf{X}}^{(j)} = (\mathbf{0}_{p \times \sum_{i=1}^{j-1} b_i}, \tilde{\mathbf{X}}_{(j)}^\top, \ldots, \tilde{\mathbf{X}}_{(q_n+1)}^\top)^\top$, $j = 2, \ldots, q_n + 1$. The estimator $\hat{\boldsymbol{\theta}}$ in (2.3) can be written as

$$(2.4) \quad \hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \left\{ \frac{1}{2n} \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\theta}\|^2 + \sum_{j=1}^{q_n} p_{\lambda_n, \gamma_n}(\|\mathbf{d}_j\|) \right\}.$$

We apply the group coordinate descent (GCD) algorithm to estimate $\hat{\theta}_n$ from (2.4).

The properties of $\hat{\boldsymbol{\theta}}$ will be given by Theorem 3.2 in the next section. For simplicity, we write $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1^\top, \ldots, \hat{\boldsymbol{\theta}}_{q_n+1}^\top)^\top$ such that $\hat{\boldsymbol{\theta}}_1 = \hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\theta}}_j = \hat{\mathbf{d}}_{j-1}$, $j = 2, \ldots, q_n + 1$. Let $\hat{\mathcal{A}} = \{j : \hat{\boldsymbol{\theta}}_j \neq 0, j = 1, \ldots, q_n + 1\}$, and $\hat{\mathcal{A}}^*$ be a subset of $\hat{\mathcal{A}}$

such that $\hat{\mathcal{A}}^* = \{j : j \in \hat{\mathcal{A}}, j - 1 \notin \hat{\mathcal{A}}, j = 2, \dots, q_n + 1\} = \{\hat{k}_1, \dots, \hat{k}_{\hat{s}}\}$. If $\hat{s} = 0$, we declare there is no threshold. If $\hat{s} > 0$, by Theorem 3.2, the true threshold a_j is highly likely to be located in $(\tilde{Z}_{(n^* - (q_n - \hat{k}_j + 3)m)}, \tilde{Z}_{(n^* - (q_n - \hat{k}_j + 1)m)}], j = 1, \dots, \hat{s}$.

2.2. *The refining stage.* If $\hat{s} > 0$, by Theorem 3.2, we can estimate the threshold a_j in $(\tilde{Z}_{(n^* - (q_n - \hat{k}_j + 3)m)}, \tilde{Z}_{(n^* - (q_n - \hat{k}_j + 1)m)}]$. To obtain the estimates of the thresholds $\{a_j\}$, we denote $\hat{\mathcal{I}}_j = \{i : \tilde{Z}_{(n^* - (q_n - \hat{k}_j + 3)m)} < Z_i \le \tilde{Z}_{(n^* - (q_n - \hat{k}_j + 1)m)}\}$, $\hat{\mathcal{I}}_{j,\varsigma^-} = \{i : \tilde{Z}_{(n^* - (q_n - \hat{k}_j + 3)m)} < Z_i \le \varsigma\}$, $\hat{\mathcal{I}}_{j,\varsigma^+} = \{i : \varsigma < Z_i \le \tilde{Z}_{(n^* - (q_n - \hat{k}_j + 1)m)}\}$

$$Q_j(\varsigma^-, \boldsymbol{\beta}) = \frac{\sharp \hat{\mathcal{I}}_{j,\varsigma^-}}{\sharp \hat{\mathcal{I}}_j} \sum_{i \in \hat{\mathcal{I}}_{j,\varsigma^-}} w_{[r_{Y_i}, \hat{\mathcal{I}}_{j,\varsigma^-}]} (Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})^2,$$

$$Q_j(\varsigma^+, \boldsymbol{\beta}) = \frac{\sharp \hat{\mathcal{I}}_{j,\varsigma^+}}{\sharp \hat{\mathcal{I}}_j} \sum_{i \in \hat{\mathcal{I}}_{j,\varsigma^+}} w_{[r_{Y_i}, \hat{\mathcal{I}}_{j,\varsigma^+}]} (Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})^2,$$

and use the following method to estimate a_j:

(2.5) $\hat{a}_j = \operatorname{argmin}_{\varsigma \in (\tilde{Z}_{(n^* - (q_n - \hat{k}_j + 3)m)}, \tilde{Z}_{(n^* - (q_n - \hat{k}_j + 1)m)}] \cap \{Z_{(1)}, \dots, Z_{(n)}\}} \{Q_j(\varsigma)\},$

where $Q_j(\varsigma) = \min_{\boldsymbol{\beta}} Q_j(\varsigma^-, \boldsymbol{\beta}) + \min_{\boldsymbol{\beta}} Q_j(\varsigma^+, \boldsymbol{\beta})$, $j = 1, \dots, \hat{s}$, and $Z_{(1)} \le Z_{(2)} \cdots \le Z_{(n)}$ are the order of Z_i, $i = 1, \dots, n$. The regions separated by the thresholds achieve the overall minimum least squares errors. The consistency of $\hat{a}_j$ will be provided in Theorem 3.3.

2.3. *Remarks.* After we obtain $\hat{a}_j$, $j = 1, \dots, \hat{s}$ by (2.5), it is sensible to use the weighted least squares to obtain a final estimate of the coefficients in model (2.1). Let $\hat{\mathcal{I}}_j^* = \{i : \hat{a}_{j-1} < Z_i \le \hat{a}_j\}$ and $\hat{b}_j^* = \sharp \hat{\mathcal{I}}_j^*$, $j = 1, \dots, \hat{s} + 1$ where $\hat{a}_0 = -\infty$ and $\hat{a}_{\hat{s}+1} = +\infty$. Then the coefficient $\boldsymbol{\theta}^* = ((\boldsymbol{\beta}_1^*)^\top, (\mathbf{d}_1^*)^\top, \dots, (\mathbf{d}_{\hat{s}}^*)^\top)^\top = (\theta_1^*, \dots, \theta_{p(\hat{s}+1)}^*)^\top$ can be estimated by minimizing the following penalized least squares:

$$M(\boldsymbol{\theta}) = \sum_{j=1}^{\hat{s}_n + 1} \frac{\hat{b}_j^*}{2n} \sum_{i \in \hat{\mathcal{I}}_j^*} w_{[r_{Y_i}, \hat{\mathcal{I}}_j^*]} \left(Y_i - \mathbf{X}_i^\top \left(\boldsymbol{\beta}_1 + \sum_{k=1}^{j-1} \mathbf{d}_k \right) \right)^2$$

(2.6)

$$+ \sum_{i=1}^{p(\hat{s}_n + 1)} p_{\lambda_n, \gamma_n} (|\theta_i|),$$

where the penalty function $p_{\lambda_n, \gamma_n}(|u|) > 0$ is the same as in (2.4) and $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\top, \mathbf{d}_1^\top, \dots, \mathbf{d}_{\hat{s}}^\top)^\top = (\theta_1, \dots, \theta_{p(\hat{s}+1)})^\top$. The inclusion of the penalty functions may lead to a sparse solution since the number of parameters could be quite large

for all the subgroups. The oracle property of $\hat{\boldsymbol{\theta}}^* = \arg\min_{\boldsymbol{\theta}} M(\boldsymbol{\theta})$ will be given by Theorem 3.3.

Similar to Jin, Shi and Wu (2013), γ_n is set as 2.4 for SCAD and MCP penalties. The regularization parameter λ_n can be chosen by the Bayesian information criterion (BIC).

The performance of $\hat{\boldsymbol{\theta}}^*$ is further dependent on the segment length m. The selection of an optimal m can be carried out as follows: Apply the splitting stage to the data sequence L times with the common segment length of events (excluding the first segment) m_ℓ, $\ell = 1, \ldots, L$. In order to meet the assumption, that is, $m = \lfloor c\sqrt{n} \rfloor$ made in Theorem 3.2, we set $m_\ell = \lfloor \kappa_\ell \sqrt{n} \rfloor$, $\ell = 1, \ldots, L$, where κ_ℓ takes values from L grid-points in an interval, say $[0.1, 2.0]$. For each m_ℓ, applying the proposed two-stage procedure, we obtain the set of estimated thresholds $\hat{\mathcal{M}}_\ell = \{\hat{a}_{1,\ell}, \ldots, \hat{a}_{\hat{s}_\ell, \ell}\}$, $\ell = 1, \ldots, L$. We use the BIC to choose the best index

$$(2.7) \qquad \hat{\ell} = \operatorname{argmin}_{\ell=1,\ldots,L}\{BIC_{\hat{\mathcal{M}}_\ell}\},$$

where

$$(2.8) \qquad BIC_{\hat{\mathcal{M}}_\ell} = n\log\left(\sum_{j=1}^{\hat{s}_\ell+1} \frac{\sharp\hat{\mathcal{I}}_{j,\ell}}{n}\left(\min_{\boldsymbol{\beta}} \sum_{i\in\hat{\mathcal{I}}_{j,\ell}} w_{[r_{Y_i},\hat{\mathcal{I}}_{j,\ell}]}(Y_i - \mathbf{X}_i^\top\boldsymbol{\beta})^2\right)\right)$$

$$+ q(\hat{s}_\ell + 1)\log(n)$$

with $\hat{\mathcal{I}}_{j,\ell} = \{i : \hat{a}_{j-1,\ell} < Z_i \le \hat{a}_{j,\ell}\}$, $\hat{a}_{0,\ell} = -\infty$, $\hat{a}_{\hat{s}_\ell+1,\ell} = +\infty$. The optimal $m_{\text{opt}} = \lfloor \kappa_{\hat{\ell}} \sqrt{n} \rfloor$ and the optimal estimated thresholds $\hat{\mathcal{M}}_{\text{opt}} = \{\hat{a}_{1,\hat{\ell}}, \ldots, \hat{a}_{\hat{s}_{\hat{\ell}},\hat{\ell}}\}$.

We refer to the proposed two-stage procedure as *T*wo *S*tage *M*ultiple *C*hange-points *D*etection (TSMCD) from now on. Since two regularization methods, that is, MCP, and SCAD, will be utilized in the splitting stage, we refer to the corresponding TSMCD as TSMCD$_{(\text{MCP})}$ and TSMCD$_{(\text{SCAD})}$, respectively. The detailed algorithm of TSMCP is described in Algorithm 1.

3. Asymptotic theory. Under model (2.1), we further denote the distributions of $\{T_i, i \in \mathcal{I}_j^*\}$ and $\{Y_i, i \in \mathcal{I}_j^*\}$ to be F_j and H_j, $j = 1, \ldots, s+1$. Denote the distribution of log censoring times $\{c_1, \ldots, c_n\}$ by G. We define U_{F_j}, U_{H_j} and U_G to be the least upper bound for the support of F_j, H_j and G, respectively.

Throughout the paper, the following assumption will be made:

(A1) $P(Z_i \le a_j) = \tau_j$ where $0 < \tau_1 < \tau_2 < \cdots < \tau_s < 1$, $E(\varepsilon_i|\mathbf{X}_i) = 0$, $E(T_i^2) < \infty$, $E(\mathbf{X}_i\mathbf{X}_i^\top) = \Sigma_0$ is finite and nonsingular, and Z_i and ε_i are independent, $i = 1, \ldots, n$.

(A2) T_i and C_i are independent, H_j is continuous, and censoring mechanism is independent of covariates.

(A3) $U_{F_j} < U_G$ or $U_{F_j} = U_G = \infty$.

Algorithm 1 TSMCD

1: **for** $\ell = 1, 2, \ldots, 20$ **do**

      Step 1: Splitting stage

2:     Set $m = \lfloor 0.1\ell\sqrt{n^*} \rfloor$ and $q_n = \lfloor n^*/m \rfloor - 1$ where n^* is the number of events;

3:     Split the data sequence into $q_n + 1$ segments $\mathcal{I}_j$, $j = 1, \ldots, q_n + 1$ as Section 2.1;

4:     Estimate $\hat{\boldsymbol{\theta}} = ((\hat{\boldsymbol{\theta}}_1)^\top, \ldots, (\hat{\boldsymbol{\theta}}_{q_n+1})^\top)^\top$ by minimizing (2.3) or (2.4);

5:     Compute the index sets $\hat{\mathcal{A}} = \{j : \hat{\boldsymbol{\theta}}_j \neq 0, j = 1, \ldots, q_n + 1\}$ and $\hat{\mathcal{A}}^* = \{j : j \in \hat{\mathcal{A}}, j - 1 \notin \hat{\mathcal{A}}, j = 2, \ldots, q_n + 1\} \equiv \{\hat{k}_1, \ldots, \hat{k}_{\hat{s}}\}$ where $\hat{k}_1 < \hat{k}_2 < \cdots < \hat{k}_{\hat{s}}$ and $\hat{s} = \sharp \hat{\mathcal{A}}^*$;

      Step 2: Refining stage

6:    **if** $\hat{s} = 0$ **then** go to the step 9;

7:    **else if** $\hat{s} > 0$ **then** estimate the threshold a_j in $(\tilde{Z}_{(n^*-(q_n-\hat{k}_j+3)m)}, \tilde{Z}_{(n^*-(q_n-\hat{k}_j+1)m)}]$ by (2.5);

8:    **end if**

9:     Estimate the coefficient $\boldsymbol{\theta}^* = ((\boldsymbol{\beta}_1^*)^\top, (\mathbf{d}_1^*)^\top, \ldots, (\mathbf{d}_{\hat{s}}^*)^\top)^\top$ by minimizing (2.6);

10:    Return the set of estimated thresholds $\hat{\mathcal{M}}_\ell = \{\hat{a}_{1,\ell}, \ldots, \hat{a}_{\hat{s}_\ell, \ell}\}$ in step 7 and the estimator of the coefficient $\hat{\boldsymbol{\theta}}_\ell^*$ in step 9, and compute $BIC_{\hat{\mathcal{M}}_\ell}$ by (2.8);

11: **end for**

12: Choose $\hat{\ell}$ that minimizes the $BIC_{\hat{\mathcal{M}}_\ell}$ and obtain the final estimators $\hat{\mathcal{M}}_{\text{opt}} = \hat{\mathcal{M}}_{\hat{\ell}}$ and $\hat{\boldsymbol{\theta}}_{\text{opt}}^* = \hat{\boldsymbol{\theta}}_{\hat{\ell}}^*$.

Assumption (A1) allows heteroscedastic error. For example, we may relax the error assumption to be $\varepsilon_i = \sigma(\mathbf{X}_i^\top \boldsymbol{\beta}_j^*)\varepsilon_i^*$ where ε_i^* is independent with $\mathbf{X}_i$ and $\varepsilon_1^*, \ldots, \varepsilon_n^*$ are i.i.d. with mean zero and variance σ^2. (A2) assumes that δ_i is conditionally independent of $\mathbf{X}_i$ given T_i. (A3) implies that $U_{H_j} = U_{F_j}$.

We first present a result for the simple case where $s = 1$ is known. Tentatively we rewrite a_1 and $\mathbf{d}_1^*$ as a and $\mathbf{d}^*$, respectively, in model (2.1), and replace τ_1 with τ in assumption (A1) for simplicity.

THEOREM 3.1. *Assume conditions* (A1)–(A3) *hold for model* (2.1), $s = 1$ *is known and* $\|\mathbf{d}^*\| > 0$. *Let* $\hat{a} = \text{argmin}_{\varsigma \in \{Z_{(1)}, \ldots, Z_{(n)}\}} \{\min_{\boldsymbol{\beta}} \frac{\sharp \hat{\mathcal{I}}_{\varsigma^-}}{n} \times \sum_{i \in \hat{\mathcal{I}}_{\varsigma^-}} w_{[r_{Y_i, \hat{\mathcal{I}}_{\varsigma^-}}]}(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})^2 + \min_{\boldsymbol{\beta}} \frac{\sharp \hat{\mathcal{I}}_{\varsigma^+}}{n} \sum_{i \in \hat{\mathcal{I}}_{\varsigma^+}} w_{[r_{Y_i, \hat{\mathcal{I}}_{\varsigma^+}}]}(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})^2\}$ *where* $\hat{\mathcal{I}}_{\varsigma^-} = \{i : Z_i \leq \varsigma\}$ *and* $\hat{\mathcal{I}}_{\varsigma^+} = \{i : \varsigma < Z_i\}$, *then we have* $\frac{1}{n} \sum_{i=1}^n 1_{\{Z_i \leq \hat{a}\}} \to_{\text{a.s.}} \tau$ *as* $n \to \infty$. *Furthermore, if* Z_i *is a continuous random variable, we have* $\hat{a} \to_{\text{a.s.}} a$; *if* $Z_i = i$, *we have* $\hat{a}/n \to_{\text{a.s.}} \tau$.

The proof of Theorem 3.1 is given in the Appendix. The result itself may be of interest since many econometric studies deal with a single change point.

If $s \geq 0$ is unknown, we use the proposed TSMCD to estimate s, and the change points a_j, $j = 1, \ldots, s$ when $s > 0$. The following technical conditions are also needed:

(A4) $m \to \infty$ and $m = O((n^*)^r)$, where $0 < r \leq 1/2$ is a constant.

(A5) The penalty function $p_{\lambda_n, \gamma_n}(|u|)$ of (2.4) satisfies $p_{\lambda_n, \gamma_n}(0) = 0$, $p'_{\lambda_n, \gamma_n}(u) = 0$ if $u > \gamma_n \lambda_n$, $p'_{\lambda_n, \gamma_n}(0) = \lambda_n$, $\lambda_n \to 0$ and $\lambda_n \sqrt{n} / \log n \to \infty$ as $n \to \infty$.

(A6) $\varepsilon_i = \sigma(\mathbf{X}_i^\top \boldsymbol{\beta}_j^*) \varepsilon_i^*$, $i \in \mathcal{I}_j^*$, $j = 1, \ldots, s + 1$, ε_i^* is independent of $\mathbf{X}_i$, $\varepsilon_1^*, \ldots, \varepsilon_n^*$ are i.i.d. random errors with distribution F_0. Suppose $U_{F_0} < U_{F_j}$, and there exists some positive constant $U_{F_j}^* < U_{F_j} - U_{F_0}$, such that $\sigma(\mathbf{X}_i^\top \boldsymbol{\beta}_j^*) = 0$ if $\mathbf{X}^\top \boldsymbol{\beta}_j^* > U_{F_j}^*$, otherwise $\sigma(\mathbf{X}_i^\top \boldsymbol{\beta}_j^*) = 1$, $j = 1, \ldots, s + 1$.

By the law of large numbers and Assumption (A1), $P(Z_i \leq a_j) = \tau_j$ implies

$$\sum_{i=1}^n 1_{\{a_{j-1} < Z_i \leq a_j\}} / n \to_{\text{a.s.}} \tau_j - \tau_{j-1} > 0.$$

Assumption (A4) implies with probability 1

$$\sum_{i=1}^n 1_{\{\tilde{Z}_{(n^* - (q_n - j + 2)m)} < Z_i \leq \tilde{Z}_{(n^* - (q_n - j + 1)m)}\}} / n = m/n^* \to 0.$$

Thus Assumptions (A1) and (A4) ensure that there is at most one threshold in each segment $\{Z_i : \tilde{Z}_{(n^* - (q_n - j + 2)m)} < Z_i \leq \tilde{Z}_{(n^* - (q_n - j + 1)m)}\}$ for sufficiently large n where $\tilde{Z}_{(n^* - (q_n - j + 2)m)}$ and $\tilde{Z}_{(n^* - (q_n - j + 1)m)}$, $j = 1, \ldots, q_n + 1$, are defined in Section 2.1. Both SCAD and MCD satisfy the penalty assumption (A5). (A6) implies $T_i = \mathbf{X}_i^\top \boldsymbol{\beta}_j^*$ if $U_{F_j}^* + U_{F_0} < Y_i \leq U_{F_j}$.

Define $\mathcal{A}^* = \{i_1, \ldots, i_a\}$ to be the group index set and $\tilde{\mathbf{X}}_{\mathcal{A}^*} = (\tilde{\mathbf{X}}^{(i_1)}, \ldots, \tilde{\mathbf{X}}^{(i_a)})$. Similarly, we can define $\hat{\boldsymbol{\theta}}_{\mathcal{A}^*}$ and $\hat{\boldsymbol{\theta}}_{(\mathcal{A}^*)^c}$. Let $\mathcal{A} = \{1, k_{j,n}, k_{j,n} + 1 : j = 1, \ldots, s\}$.

THEOREM 3.2. *If Assumptions* (A1)–(A6) *hold,* $\min\{\|\boldsymbol{\beta}_1^*\|, \|\mathbf{d}_1^*\|, \ldots, \|\mathbf{d}_s^*\|\} > 2\sqrt{p}\gamma\lambda_n$, *then with probability* 1, (2.4) *has a local minimizer* $\hat{\boldsymbol{\theta}}$ *such that* $\hat{\boldsymbol{\theta}}_{\mathcal{A}^*} = (\tilde{\mathbf{X}}_{\mathcal{A}^*}^\top \tilde{\mathbf{X}}_{\mathcal{A}^*})^{-1} \tilde{\mathbf{X}}_{\mathcal{A}^*}^\top \tilde{\mathbf{y}}$ *and* $\hat{\boldsymbol{\theta}}_{(\mathcal{A}^*)^c} = \mathbf{0}$, *where*

(3.1)
$$\mathcal{A}^* \subseteq \mathcal{A} \quad \text{and} \quad 1 \in \mathcal{A}^*,$$
$$k_{\ell,n} \in \mathcal{A}^* \text{ or } k_{\ell,n} + 1 \in \mathcal{A}^* \text{ for all } \ell = 1, \ldots, s.$$

The proof of Theorem 3.2 is given in the Appendix. Theorem 3.2 provides the existence of the solution of (2.4). In this theorem, $\mathcal{A}^*$ is the group index set of no-zero group elements of $\hat{\boldsymbol{\theta}}$. By (3.1), $\mathcal{A}^*$ is not unique and may take 2^s possible forms.

Let $\mathbf{a} = (a_1, \ldots, a_s)^\top$ be the vector of threshold locations and $\hat{\mathbf{a}} = (\hat{a}_1, \ldots, \hat{a}_{\hat{s}})^\top$ is obtained by (2.5). Similar to the definition of $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{X}}$, replacing $\mathcal{I}_j$ with $\mathcal{I}_j^*$ or $\hat{\mathcal{I}}_j^*$ and then we may define $\tilde{\mathbf{y}}_{\mathbf{a}}$ or $\tilde{\mathbf{y}}_{\hat{\mathbf{a}}}$ and $\tilde{\mathbf{X}}_{\mathbf{a}}$ or $\tilde{\mathbf{X}}_{\hat{\mathbf{a}}}$. By Assumption (A1), we have $\frac{1}{n}\tilde{\mathbf{X}}_{\mathbf{a}}^\top\tilde{\mathbf{X}}_{\mathbf{a}} \to_{\text{a.s.}} \Gamma$, where Γ is a positive definite matrix.

To obtain the limiting property of $\hat{\boldsymbol{\theta}}^*$ in (2.6), we will further need the following assumption:

(A7) $\max_{u \geq 0}\{p''_{\lambda_n, \gamma_n}(u)\} + \lambda_{(s+1)p}(\Gamma) > 0$ where $\lambda_{(s+1)p}(\Gamma)$ is the minimal eigenvalue of Γ.

For MCP, (A7) is equivalent to $\lambda_{(s+1)p}(\Gamma) > 1/\gamma$, and for SCAD, (A7) is equivalent to $\lambda_{(s+1)p}(\Gamma) > 1/(\gamma - 1)$. Let $(\theta_1^*, \ldots, \theta_{(s+1)p}^*)^\top = ((\boldsymbol{\beta}_1^*)^\top, (\mathbf{d}_1^*)^\top, \ldots, (\mathbf{d}_s^*)^\top)^\top$ and $\mathcal{S} = \{j : \theta_j^* \neq 0, j = 1, \ldots, (s+1)p\}$.

Let $\hat{\mathcal{A}} = \{j : \hat{\boldsymbol{\theta}}_j \neq 0, j = 1, \ldots, q_n + 1\}$, and $\hat{\mathcal{A}}^*$ be a subset of $\hat{\mathcal{A}}$ such that $\hat{\mathcal{A}}^* = \{j : j \in \hat{\mathcal{A}}, j - 1 \notin \hat{\mathcal{A}}, j = 2, \ldots, q_n + 1\} = \{\hat{k}_1, \ldots, \hat{k}_{\hat{s}}\}$ where $\hat{s}$ is the size of $\hat{\mathcal{A}}^*$.

THEOREM 3.3. *Under the Assumptions* (A1)–(A7), $\min_{j \in \mathcal{S}}\{|\theta_j|\} > \gamma\lambda_n$, $\hat{\boldsymbol{\theta}}$ *is the local minimizer in Theorem* 3.2, Z_i *is a continuous random variable and* $\hat{\boldsymbol{\theta}}^*$ *is a minimizer of* $M(\theta)$ *given by* (2.6), *then we have, with probability* 1:

   (1) $\hat{s} = s$,
   (2) $\hat{\mathbf{a}} \to \mathbf{a}$,
   (3) $\hat{\boldsymbol{\theta}}^* = \hat{\boldsymbol{\theta}}^o$,

where $\hat{\boldsymbol{\theta}}^o = \arg\min_{\boldsymbol{\theta}}\{\|\tilde{\mathbf{y}}_{\hat{\mathbf{a}}} - \tilde{\mathbf{X}}_{\hat{\mathbf{a}}}\boldsymbol{\theta}\|^2 : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_{(s+1)p})^\top, \theta_j = 0 \, \forall j \notin \mathcal{S}\}$ *is the oracle estimator when thresholds and the set* $\mathcal{S}$ *are known.*

The proof of Theorem 3.3 is also given in the Appendix. We attain the consistency of the TSMCD estimators in this theorem. Our estimators work as well as the oracle estimators in large samples.

4. Simulation study. We generate random samples from model (2.1) with $s = 2$ thresholds in the following. Specifically, we generate the regressors $\mathbf{X}_i = (x_{1,i}, x_{2,i}, \ldots, x_{6,i})^\top$ with $x_{1,i} = 1$, $x_{j,i} \sim N(0,1)$, $j = 2, \ldots, 6$, and $\varepsilon_i \sim N(0, 0.5)$. We specify the coefficients $(\theta_1^*, \ldots, \theta_{18}^*)^\top = ((\boldsymbol{\beta}_1^*)^\top, (\mathbf{d}_1^*)^\top, (\mathbf{d}_2^*)^\top)^\top = (2, 1, 1, 1, 1, 1, -1, 0, 0, -1, -1, -1, 0, -1, 1, 0, 0, 0)^\top$, the threshold variable $Z_i = x_{2,i}$ and true thresholds $a_1 = -0.5244$, $a_2 = 0.2533$. The two thresholds are the 30% and 60% lower percentiles of the standard normal distribution. We first design the following four cases:

Example 1: $n = 150$, $x_{j,i}$, $j = 2, \ldots, 6$ are independent and censoring variable $C_i \sim N(2, 16)$.

Example 2: The same setting as Example 1 except $n = 300$.
Example 3: The same setting as Example 2 except $C_i \sim N(\sum_{j=2}^{6} x_{j,i}, 16)$.
Example 4: The same setting as Example 3 except covariance between $x_{j_1,i}$ and $x_{j_2,i}$ is $0.5^{|j_1 - j_2|}$.

The censoring rates are about 40% for all cases. Assumption (A2) stipulates that T_i and C_i must be independent. This assumption is satisfied in the first two examples. However, a more realistic assumption is that T_i and C_i are conditionally independent given $\mathbf{X}_i$. Example 3 is thus introduced to test the robustness of our method. Example 4 is the case where the regressors are moderately correlated.

In order to examine the performance of TSMCP when the model has no threshold. We design the following two cases:

Example 5: The same as Example 1 except $s = 0$ and $\boldsymbol{\beta}_1^* = (1, 0, 2, 0, 0, 0)$.
Example 6: The same as Example 4 except $s = 0$ and $\boldsymbol{\beta}_1^* = (1, 0, 2, 0, 0, 0)$.

All numerical studies are performed on a computer [Intel(R) Core(TM) i7 930 2.80 GHz 8 M Caches, 8 GB Memory]. The R program can be downloaded from the first author's website: http://www.stat.nus.edu.sg/~stalj/. The means and standard deviations of computing time by TSMCP are reported in Table 1 based on 100 simulations. The average computing time for MCP is slightly less than that for SCAD. With the increase of model complexity and sample size, the average computing time also tends to increase.

The estimation results for s are reported in Table 2 based on 1000 simulations. Our methods can correctly identify the number of thresholds with very high probability. While their performance is comparable, SCAD seems to identify slightly more unnecessary change points than MCP. Both methods improve as sample size increases.

Figure 2 displays the histograms of the estimators of thresholds, indicating the empirical estimates are symmetrically distributed around the true change points. Table 3 summarizes the estimation performance of the estimated thresholds for the

TABLE 1
The means and standard deviations of computation time (*seconds*)

		Example					
		s = 2				s = 0	
		1	**2**	**3**	**4**	**5**	**6**
TSMCP-MCP	Mean	7.55	19.85	21.90	29.56	2.77	7.59
	Stand deviation	1.97	3.15	3.56	5.14	1.59	3.34
TSMCP-SCAD	Mean	9.76	22.58	24.51	30.81	3.61	9.82
	Stand deviation	2.05	3.59	3.50	5.66	1.99	4.04

TABLE 2
Frequency of estimated ŝ in 1000 *simulations*

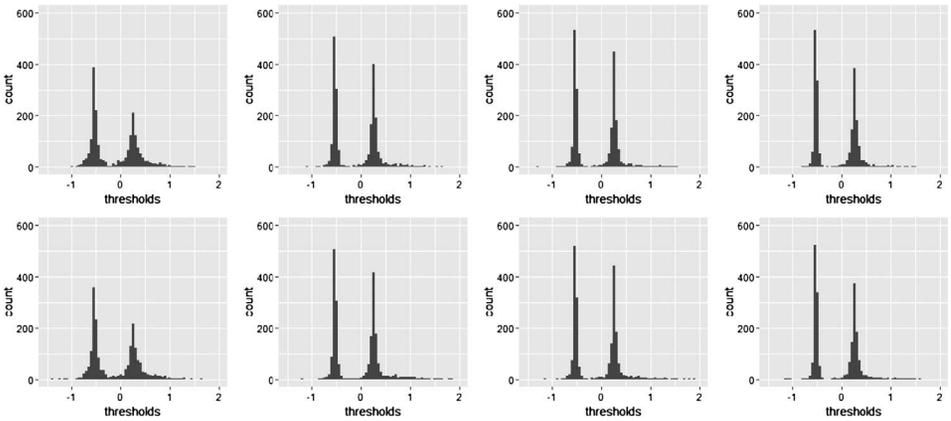| | | $\hat{s}$ | | | | | | | | | | | | |
		TSMCP-MCP							TSMCP-SCAD					
		0	**1**	**2**	**3**	**4**	**5**	**6**	**0**	**1**	**2**	**3**	**4**	**5**
$s=2$	Example 1	2	150	778	65	4	1	0	0	153	773	70	4	0
	Example 2	0	13	919	64	3	0	1	0	12	869	116	3	0
	Example 3	0	8	937	52	3	0	0	0	8	916	71	4	1
	Example 4	0	24	947	26	3	0	0	0	33	920	43	4	0
$s=0$	Example 5	914	82	4	0	0	0	0	899	88	11	1	1	0
	Example 6	945	49	6	0	0	0	0	936	58	6	0	0	0

FIG. 2. *Histograms of the estimated thresholds. The results for MCP and SCAD are displayed in the first and the second row, respectively. Examples* 1, 2, 3 *and* 4 *are from the left to the right, respectively.*

TABLE 3
Estimation performance for the threshold estimation. Bias is the average of estimated parameter minus the true value. RMSE refers to the relative mean squared errors

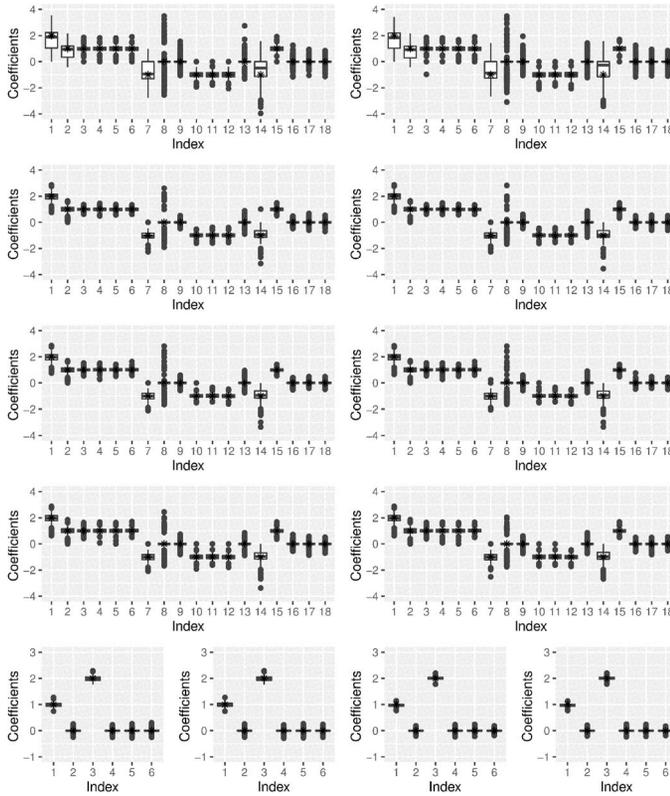| | TSMCP-MCP | | | | TSMCP-SCAD | | | |
	$\hat{a}_1$		$\hat{a}_2$		$\hat{a}_1$		$\hat{a}_2$	
	Bias	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**
Example 1	-0.014	0.031	0.037	0.476	-0.013	0.037	0.039	0.428
Example 2	-0.007	0.007	0.010	0.171	-0.008	0.007	0.012	0.158
Example 3	-0.009	0.006	0.008	0.145	-0.007	0.006	0.010	0.159
Example 4	-0.006	0.005	0.013	0.166	-0.006	0.005	0.012	0.158

FIG. 3. *Box plots of coefficients estimated from MCP (left) and SCAD (right). Results of Examples 1, 2, 3 and 4 are displayed in the four rows, respectively. Box plots of coefficients estimated from MCP and SCAD of Examples 5 and 6 are reported in the five row, respectively. "∗"s are the true values.*

cases with correct estimation of $\hat{s} = s$. Again the two penalty methods give quite similar results. The estimation bias and mean squared error for a_1 are relatively smaller than those for a_2. In fact, we note that the jumps at the two change points are $\|\mathbf{d}_1^*\|^2 = 4$ and $\|\mathbf{d}_2^*\|^2 = 2$, respectively, under our model. In general, it is easier for our methods to detect a break point with greater jump.

Finally, we report the estimation performance of the regression coefficients $\hat{\theta}_j$ using box-plots in Figure 3 when the number of thresholds are correctly estimated. In Examples 2–6, the estimated coefficients are quite consistent to the true parameter values. We can see that the variances of $\hat{\theta}_j^*$, $j = 1, 2, 7, 8, 13, 14$, which are the estimators of the coefficients of the intercept $x_{1,i}$ and $x_{2,i} = Z_i$, are larger than the others for Examples 1–4. The zero coefficients θ_j^*, $j = 8, 9, 13, 16, 17, 18$ in Examples 1–4 and θ_j^*, $j = 2, 4, 5, 6$ in Examples 5–6, can be successfully identified by our method, suggesting a satisfactory variable selection performance.

It is noted in all simulations that results from Examples 3–4 and Example 6 are comparable to those from Example 2 and Example 5, respectively. Even though Examples 3–4 and Example 6 violate the independence censoring assumption, numerical results in this section suggest our methods are quite robust and may still work steadily under conditional independence.

5. Real data analysis. Follicular lymphoma is the second most common form of non-Hodgkins lymphoma, accounting for about 22% of all documented cases. Dave et al. (2004) investigated whether the survival risks of patients with follicular lymphoma can be predicted by the gene-expression profiles of tumors and standard clinical risk factors at diagnosis. Fresh-frozen tumor-biopsy specimens and clinical data from 191 untreated patients who had received a diagnosis of follicular lymphoma between 1974 and 2001 were obtained. The median age at diagnosis was 51 years (range 23–81), and the median follow-up time was 6.6 years (range: less than 1.0–28.2). The median follow-up time among patients alive at last follow-up was 8.1 years.

A total of 156 subjects are included in analysis after excluding cases with missing information. Many authors analyzed this data in earlier works. We revisit this data set in this section and consider an AFT model between the failure time and 5 most significant genetic markers selected in Yu, Li and Ma (2012). According to Yu, Li and Ma (2012), genes 357, 2345, 6267, 6271 and 3653 in the original sample are the most important markers for the survival risk prediction when clinical information is adjusted. We thus set $x_{1,i} = 1$ as the intercept, and chose the five gene expressions as the regressors $x_{2,i}, \ldots, x_{6,i}$. We pick gene 357 as the index variable in the following analysis and set $Z_i = x_{2,i}$ since it is recognized as the most predictive marker for the failure time. We have also attempted to detect change points for other covariates but our numerical program returned an estimate $\hat{s} = 0$ for them.

Applying TSMCD$_{(MCP)}$ and TSMCD$_{(SCAD)}$, we obtain the same estimation results, yielding $\hat{s} = 2$ and two thresholds $\hat{a}_1 = -0.483$ and $\hat{a}_2 = 0.907$. These two change points divide the sample into three groups with cumulative group sizes $\hat{\tau}_1 = \sum_{i=1}^{156} 1_{\{Z_i \leq -0.483\}}/156 = 0.263$, and $\hat{\tau}_2 = \sum_{i=1}^{156} 1_{\{Z_i \leq 0.907\}}/156 = 0.814$. Figure 4 displays the Kaplan–Meier survival curves for the three groups separated by genes 357. The survival curves for Group 3 drops rapidly since the baseline and represents a high-risk group in this sample. Such an observation is not available without considering the threshold AFT model in this paper. Practitioners may adapt our model easily to discover more meaningful subpopulations with defining features.

Traditionally, when the goal is to identify subgroups from the empirical sample one may resort to clustering analysis [Gordon (1981)]. Such unsupervised learning methods produce segmentation or grouping based on the variation of covariates and ignore the censored response variable. One drawback is that those learned groupings may not be relevant to the survival response. We have implemented
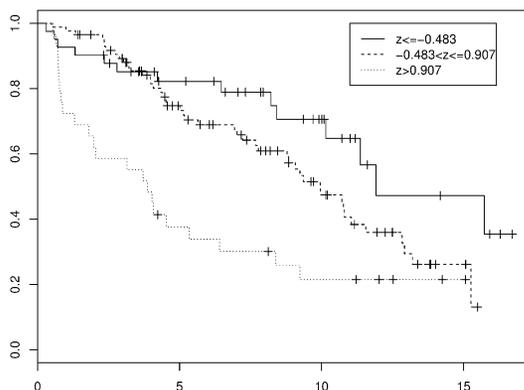
FIG. 4. *Survival curves for the three groups separated by the two thresholds of Gene* 357. + *indicates a censored observation. Groups* 1, 2 *and* 3 *refer to* $Z \leq -0.482$, $-0.482 < Z \leq 0.907$ *and* $Z > 0.907$, *respectively.*

a hierarchical clustering procedure (R function `hclust`) using only the covariate matrix and the identified clusters are quite different from our change point analysis. The comparison of cluster sizes between clustering analysis and our method is given in Table 4. There is only slight agreement between the two methods with Cohen's kappa value equal to 0.36. Furthermore, it is hard to interpret the groups resulted from the hierarchical clustering algorithm. In particular, it is unclear if the three groups suggest any practical difference in survival probability. Examining the Kaplan–Meier curves for the three groups are shown in Figure 5, we notice that the survival functions for groups 2 and 3 severely overlap for the observation period. Such findings might be less meaningful for clinical investigation.

The estimated regression coefficients are $\hat{\boldsymbol{\theta}}^* = ((\hat{\boldsymbol{\beta}}_1^*)^\top, (\hat{\mathbf{d}}_1^*)^\top, (\hat{\mathbf{d}}_2^*)^\top)^\top$ where $\hat{\boldsymbol{\beta}}_1^* = (0.791, -0.492, 0.291, 0.990, -0.589, 1.115)^\top$, $\hat{\mathbf{d}}_1^* = (1.166, 0, -0.451, -1.055, 0.540, -0.991)^\top$, $\hat{\mathbf{d}}_2^* = (0, -0.381, 0, 0, 0, 0.770)^\top$. Such results suggest at the first threshold $\hat{a}_1 = -0.483$, only the coefficient of $x_{2,i}$ does not change; and at the second threshold $\hat{a}_2 = 0.907$, only coefficients of $x_{2,i}$ and $x_{6,j}$ change and all other coefficients remain the same.

TABLE 4
Cluster sizes of real data analysis

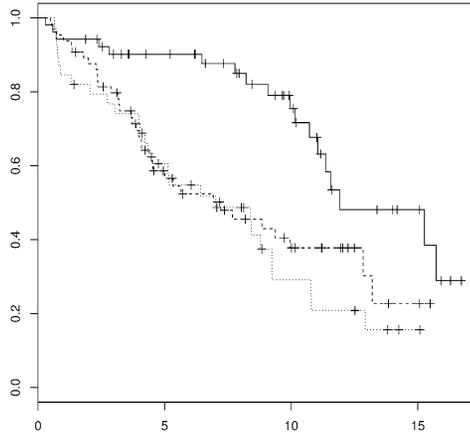| TSMCD	Cluster analysis		
	Group 1	**Group 2**	**Group 3**
Group 1	32	5	4
Group 2	20	46	20
Group 3	0	14	15

FIG. 5.   *Kaplan–Meier curves from the hierarchical clustering algorithm.*

Using the identified break point structure, we refitted the AFT model and reported the estimation results for the three subgroups in Table 5, along with 95% bootstrap confidence intervals [Huang, Ma and Xie (2006)]. The effects of genes on the survival outcome are quite different for the three groups. For example, Gene 2345 is not significant for the first group but are negatively associated with the failure time for the other two groups. Using our empirical findings, investigators may evaluate the effects of various genetic biomarkers on mortality more specifically for different subpopulations.

6. Discussion.   The Stute estimator for the AFT model is mainly developed for right-censored survival data. When the failure time is interval censored, usually a likelihood-based objective function is constructed to facilitate the regression estimation. One can incorporate change point structure in such framework and carry out the two-step estimation similar to our proposal. However, we note that

TABLE 5
*Estimates of regression coefficients for the three groups, along with bootstrap standard errors (S.E.)
and P-values by Wald test*

Covariates	$Z \leq -0.482$			$-0.482 < Z \leq 0.907$			$Z > 0.907$		
	Coef.	**S.E.**	***P*-value**	**Coef.**	**S.E.**	***P*-value**	**Coef.**	**S.E.**	***P*-value**
Intercept	0.791	0.507	0.056	1.957	0.083	0	1.957	0.083	0
Gene 357	−0.492	0.179	0.003	−0.492	0.179	0.003	−0.874	0.144	6e−10
Gene 2345	0.291	0.627	0.321	−0.159	0.084	0.029	−0.159	0.084	0.029
Gene 6267	0.990	0.725	0.086	−0.065	0.110	0.277	−0.065	0.110	0.277
Gene 6271	−0.589	0.568	0.150	−0.049	0.120	0.341	−0.049	0.120	0.341
Gene 3653	1.115	0.556	0.022	0.124	0.078	0.056	0.894	0.250	2e−4

the functional form for the likelihood under interval censoring may be quite complicated and the numerical solution may not be straightforward. More extensions in theory and computation are needed to provide a complete solution.

We consider independent survival times in this paper. Individuals observed over time may experience multiple events of the same type, or a number of events of different types. Correlated survival time data require further model structures and technical assumptions to achieve efficient results. Two familiar approaches are the marginal model and the frailty model [Kalbfleisch and Prentice (2002), Chapter 10]. In both approaches, one may still consider Stute estimator for the AFT regression and our TSMCD adapts directly. However, the asymptotic results for such analysis are not trivial since one must take care of the chapter-dependence of the repeated measures.

The TSMCD involves the penalization approaches, and thus can be easily extended to incorporate high-dimensional data analysis. Not many authors examined the change point problem under such settings and, therefore, the results in this paper could contribute significantly toward this goal. However, when the dimension of feature space is ultra-high with an exponential order of the sample size, usually a screening step must be conducted before the application of penalized estimation. In particular, one may adopt TSMCD at the screening step and estimate the effects of individual markers using the change point structure. This differs from the existing literature where the marginal effects are usually modeled as a smooth function at the initial screening. The detailed methodology construction is beyond the scope of this paper. We will carry out necessary theoretical and empirical studies for this topic.

APPENDIX A: MATRIX REPRESENTATION OF MODEL (2.1)

Let $\mathcal{C}_i = \bigcup_{j=i}^{q_n+1} \mathcal{I}_j$, $i = 1, \ldots, q_n + 1$. We split the segment $\mathcal{I}_{k_{j,n}}$ into two segments: $\mathcal{I}_{k_{j,n}}^{(1)} = \{i : \tilde{Z}_{(n^*-(q_n-k_{j,n}+2)m))} < Z_i \le a_j\}$ and $\mathcal{I}_{k_{j,n}}^{(2)} = \{i : a_j < Z_i \le \tilde{Z}_{(n^*-(q_n-k_{j,n}+1)m))}\}$. We rewrite the model (2.1) as the following linear regression model that takes the segmentation of the data sequence into consideration:

$$(A.1) \qquad T_i = \mathbf{X}_i^\top \left(\boldsymbol{\beta}_1^* + \sum_{j=1}^{s} \mathbf{d}_j^* 1_{\{i \in \mathcal{C}_{k_{j,n}}\}} \right) - \mathbf{X}_i^\top \sum_{j=1}^{s} \mathbf{d}_j^* 1_{\{i \in \mathcal{I}_{k_{j,n}}^{(1)}\}} + \varepsilon_i,$$

where $i = 1, \ldots, n$, $\mathbf{d}_j^* = \boldsymbol{\beta}_{j+1}^* - \boldsymbol{\beta}_j^*$, $j = 1, \ldots, s$; to include them in (A.1) is to facilitate the proof of Theorem 3.2. In order to estimate $k_{j,n}$, $j = 1, \ldots, s$ and the regression coefficients simultaneously, we expand the model (A.1) as the following model:

$$(A.2) \qquad T_i = \mathbf{X}_i^\top \left[\boldsymbol{\beta}_1 + \sum_{\ell=1}^{q_n} \mathbf{d}_\ell 1_{\{i \in \mathcal{C}_{\ell+1}\}} + \boldsymbol{\omega}_i \right] + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $\{\mathbf{d}_\ell\}$ and $\{\boldsymbol{\omega}_i\}$ are defined as follows:

$$\mathbf{d}_\ell = \begin{cases} \mathbf{d}_j^*, & \ell = k_{j,n}, j = 1, \dots, s, \\ \mathbf{0}_p, & \text{otherwise}; \end{cases}$$

$$\boldsymbol{\omega}_i = \begin{cases} -\mathbf{d}_j^*, & i \in \mathcal{I}_{k_{j,n}}^{(1)}, j = 1, \dots, s, \\ \mathbf{0}_p, & \text{otherwise}. \end{cases}$$

Let $\boldsymbol{\varepsilon}_{(j)} = (\varepsilon_i, i \in \mathcal{I}_j)^\top$, $\mathbf{x}_{\omega,(j)} = (\mathbf{X}_i^\top \boldsymbol{\omega}_i, i \in \mathcal{I}_j)^\top$, $\tilde{\boldsymbol{\varepsilon}}_{(j)} = \text{diag}(\tilde{\mathbf{w}}_{(j)})^{1/2} \boldsymbol{\varepsilon}_{(j)}$, $\tilde{\mathbf{x}}_{\omega,(j)} = \text{diag}(\tilde{\mathbf{w}}_{(j)})^{1/2} \mathbf{x}_{\omega,(j)}$, $\tilde{\boldsymbol{\varepsilon}} = (\boldsymbol{\varepsilon}_{(1)}^\top, \dots, \boldsymbol{\varepsilon}_{(q_n+1)}^\top)^\top$, $\tilde{\mathbf{x}}_\omega = (\tilde{\mathbf{x}}_{\omega,(1)}^\top, \dots, \tilde{\mathbf{x}}_{\omega,(q_n+1)}^\top)^\top$, $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_{q_n+1}^\top)^\top = (\boldsymbol{\beta}_1^\top, \mathbf{d}_1^\top, \dots, \mathbf{d}_{q_n}^\top)^\top$ and $\mathbf{d}_j = (d_{j1}, \dots, d_{jp})^\top$, $j = 1, \dots, q_n$. The model (A.2) has the matrix form as follows:

$$\text{(A.3)} \qquad \tilde{\mathbf{y}} = \tilde{\mathbf{X}} \boldsymbol{\theta}^* + \tilde{\mathbf{x}}_\omega + \tilde{\boldsymbol{\varepsilon}},$$

Remove the zero elements of $\boldsymbol{\theta}^*$, model (A.3) can be rewritten as

$$\text{(A.4)} \qquad \tilde{\mathbf{y}} = \tilde{\mathbf{X}}_{\mathcal{A}} \boldsymbol{\theta}_{\mathcal{A}}^* + \tilde{\mathbf{x}}_\omega + \tilde{\boldsymbol{\varepsilon}},$$

which corresponds to the matrix form of model (A.1).

APPENDIX B: PROOF OF THEOREMS 3.1–3.3

B.1. Proof of Theorem 3.1. Let $\varsigma = Z_{(k)}$, $\mathbf{Y}_{1k} = (Y_i, i \in \mathcal{I}_{Z_{(k)}^-})^\top$, $\mathbf{X}_{1k} = (\mathbf{X}_i, i \in \mathcal{I}_{Z_{(k)}^-})^\top$, $\boldsymbol{\varepsilon}_{1k} = (\varepsilon_i, i \in \mathcal{I}_{Z_{(k)}^-})^\top$, $\mathbf{Y}_{2k} = (Y_i, i \in \mathcal{I}_{Z_{(k)}^+})^\top$, $\mathbf{X}_{2k} = (\mathbf{X}_i, i \in \mathcal{I}_{Z_{(k)}^+})^\top$, $\boldsymbol{\varepsilon}_{2k} = (\varepsilon_i, i \in \mathcal{I}_{Z_{(k)}^+})^\top$, $W_{1k} = \text{diag}((w_{[r_{Y_i, \mathcal{I}_{Z_{(k)}^-}}]}, i \in \mathcal{I}_{Z_{(k)}^-}))$, $W_{2k} = \text{diag}((w_{[r_{Y_i, \mathcal{I}_{Z_{(k)}^+}}]}, i \in \mathcal{I}_{Z_{(k)}^+}))$, $Q_1(k, \boldsymbol{\beta}) = Q(Z_{(k)}^-, \boldsymbol{\beta})$ and $Q_2(k, \boldsymbol{\beta}) = Q(Z_{(k)}^+, \boldsymbol{\beta})$. Then we have

$$\hat{\boldsymbol{\beta}}_{1k} = \arg\min_{\boldsymbol{\beta}} Q_1(k, \boldsymbol{\beta}) = (\mathbf{X}_{1k}^\top W_{1k} \mathbf{X}_{1k})^{-1} \mathbf{X}_{1k}^\top W_{1k} \mathbf{Y}_{1k},$$

$$\hat{\boldsymbol{\beta}}_{2k} = \arg\min_{\boldsymbol{\beta}} Q_2(k, \boldsymbol{\beta}) = (\mathbf{X}_{2k}^\top W_{2k} \mathbf{X}_{2k})^{-1} \mathbf{X}_{2k}^\top W_{2k} \mathbf{Y}_{2k}.$$

Let $k/n \to \tau'$, $k^*/n \to \tau$ (e.g., $k = \lfloor n\tau' \rfloor$, $k^* = \lfloor n\tau \rfloor$), $E(\mathbf{X}_i \mathbf{X}_i^\top | Z_i \leq \varsigma) = \Sigma_{0,\varsigma^-}$ and $E(\mathbf{X}_i \mathbf{X}_i^\top | Z_i > \varsigma) = \Sigma_{0,\varsigma^+}$. First, we consider the case of $0 \leq \tau' < \tau$. For $\sharp \mathcal{I}_{Z_{(k)}^-} = k$, we get

$$\begin{aligned} Q_1(k, \hat{\boldsymbol{\beta}}_{1k}) &= \frac{k}{n} (\mathbf{Y}_{1k} - \mathbf{X}_{1k} \hat{\boldsymbol{\beta}}_{1k})^\top W_{1k} (\mathbf{Y}_{1k} - \mathbf{X}_{1k} \hat{\boldsymbol{\beta}}_{1k}) \\ &= \frac{k}{n} \mathbf{Y}_{1k}^\top \big(I_k - W_{1k} \mathbf{X}_{1k} (\mathbf{X}_{1k}^\top W_{1k} \mathbf{X}_{1k})^{-1} \mathbf{X}_{1k}^\top\big) \\ &\quad \times W_{1k} \big(I_k - \mathbf{X}_{1k} (\mathbf{X}_{1k}^\top W_{1k} \mathbf{X}_{1k})^{-1} \mathbf{X}_{1k} W_{1k}\big) \mathbf{Y}_{1k} \end{aligned}$$

$$= \frac{k}{n}\boldsymbol{\varepsilon}_{1k}^{\top}\boldsymbol{W}_{1k}^{1/2}\big(I_k - \boldsymbol{W}_{1k}^{1/2}\mathbf{X}_{1k}(\mathbf{X}_{1k}^{\top}\boldsymbol{W}_{1k}\mathbf{X}_{1k})^{-1}\mathbf{X}_{1k}^{\top}\boldsymbol{W}_{1k}^{1/2}\big)\boldsymbol{W}_{1k}^{1/2}\boldsymbol{\varepsilon}_{1k}$$

$$= \frac{k}{n}\boldsymbol{\varepsilon}_{1k}^{\top}\boldsymbol{W}_{1k}\boldsymbol{\varepsilon}_{1k} - \frac{k}{n}\boldsymbol{\varepsilon}_{1k}^{\top}\boldsymbol{W}_{1k}\mathbf{X}_{1k}(\mathbf{X}_{1k}^{\top}\boldsymbol{W}_{1k}\mathbf{X}_{1k})^{-1}\mathbf{X}_{1k}^{\top}\boldsymbol{W}_{1k}\boldsymbol{\varepsilon}_{1k}.$$

By Corollary 1.8 of Stute (1993) and the assumption that Z_i, ε_i are independent,

$$\boldsymbol{\varepsilon}_{1k}^{\top}\boldsymbol{W}_{1k}\boldsymbol{\varepsilon}_{1k} = \sum_{i \in \mathcal{I}_{Z_{(k)}^-}} w_{[r_{Y_i}, \mathcal{I}_{Z_{(k)}^-}]}(Y_i - \mathbf{X}_i^{\top}\boldsymbol{\beta}_1^*)^2 \to_{\text{a.s.}} E\big[(Y_i - \mathbf{X}_i^{\top}\boldsymbol{\beta}_1^*)^2 | Z_i \le Z_{(k)}\big]$$

$$= E\varepsilon_i^2 = \sigma_1^2,$$

$$\mathbf{X}_{1k}^{\top}\boldsymbol{W}_{1k}\mathbf{X}_{1k} = \sum_{i \in \mathcal{I}_{Z_{(k)}^-}} w_{[r_{Y_i}, \mathcal{I}_{Z_{(k)}^-}]}\mathbf{X}_i\mathbf{X}_i^{\top} \to_{\text{a.s.}} E\big(\mathbf{X}_i\mathbf{X}_i^{\top} | Z_i \le Z_{(k)}\big) = \Sigma_{0, Z_{(k)}^-} \le \Sigma_0,$$

$$\boldsymbol{\varepsilon}_{1k}^{\top}\boldsymbol{W}_{1k}\mathbf{X}_{1k} = \sum_{i \in \mathcal{I}_{Z_{(k)}^-}} w_{[r_{Y_i}, \mathcal{I}_{Z_{(k)}^-}]}\varepsilon_i\mathbf{X}_i^{\top} \to_{\text{a.s.}} E\big(\varepsilon_i\mathbf{X}_i^{\top} | Z_i \le Z_{(k)}\big)$$

$$= E\big[E\big(\varepsilon_i\mathbf{X}_i^{\top} | \mathbf{X}_i, Z_i \le Z_{(k)}\big)\big] = \mathbf{0}_p^{\top}.$$

Thus

$$\text{(B.1)} \qquad\qquad Q_1(k, \hat{\boldsymbol{\beta}}_{1k}) \to_{\text{a.s.}} \tau'\sigma_1^2.$$

Let $\mathbf{d}^* = \boldsymbol{\beta}_2^* - \boldsymbol{\beta}_1^*$, $\mathbf{T}_{2k} = (T_{k+1}, \ldots, T_n)^{\top}$, $\mathbf{0}_{kk^*} = \mathbf{0}_{(k^*-k) \times (p+1)}$ and $\mathbf{X}_{2k^*} = (\mathbf{X}_{k^*+1}, \ldots, \mathbf{X}_n)^{\top}$. We have

$$\mathbf{T}_{2k} = \mathbf{X}_{2k}\boldsymbol{\beta}_1^* + \begin{pmatrix} \mathbf{0}_{kk^*} \\ \mathbf{X}_{2k^*} \end{pmatrix}\mathbf{d}^* + \boldsymbol{\varepsilon}_{2k}.$$

Notice that $\sharp\mathcal{I}_{Z_{(k)}^+} = n - k$, we obtain

$$Q_2(k, \hat{\boldsymbol{\beta}}_{2k}) = \frac{n-k}{n}(\mathbf{Y}_{2k} - \mathbf{X}_{2k}\hat{\boldsymbol{\beta}}_{2k})^{\top}\boldsymbol{W}_{2k}(\mathbf{Y}_{2k} - \mathbf{X}_{2k}\hat{\boldsymbol{\beta}}_{2k})$$

$$= \frac{n-k}{n}\mathbf{Y}_{2k}^{\top}\boldsymbol{W}_{2k}^{1/2}\boldsymbol{P}_{n-k}\boldsymbol{W}_{2k}^{1/2}\mathbf{Y}_{2k}$$

$$= \frac{n-k}{n}\boldsymbol{\varepsilon}_{2k}^{\top}\boldsymbol{W}_{2k}^{1/2}\boldsymbol{P}_{n-k}\boldsymbol{W}_{2k}^{1/2}\boldsymbol{\varepsilon}_{2k}$$

$$+ \frac{n-k}{n}(\mathbf{d}^*)^{\top}\begin{pmatrix} \mathbf{0}_{kk^*} \\ \mathbf{X}_{2k^*} \end{pmatrix}^{\top}\boldsymbol{W}_{2k}^{1/2}\boldsymbol{P}_{n-k}\boldsymbol{W}_{2k}^{1/2}\begin{pmatrix} \mathbf{0}_{kk^*} \\ \mathbf{X}_{2k^*} \end{pmatrix}\mathbf{d}^*$$

$$+ \frac{2(n-k)}{n}(\mathbf{d}^*)^{\top}\begin{pmatrix} \mathbf{0}_{kk^*} \\ \mathbf{X}_{2k^*} \end{pmatrix}^{\top}\boldsymbol{W}_{2k}^{1/2}\boldsymbol{P}_{n-k}\boldsymbol{W}_{2k}^{1/2}\boldsymbol{\varepsilon}_{2k},$$

where $\boldsymbol{P}_{n-k} = I_{n-k} - \boldsymbol{W}_{2k}^{1/2}\mathbf{X}_{2k}(\mathbf{X}_{2k}^{\top}\boldsymbol{W}_{2k}\mathbf{X}_{2k})^{-1}\mathbf{X}_{2k}^{\top}\boldsymbol{W}_{2k}^{1/2}$.

Recall that $k/n \to \tau'$, $k^*/n \to \tau$, $0 \le \tau' < \tau$. Let $\alpha_{1k} = \lim_{n\to\infty}(k^* - k)/(n - k) = (\tau - \tau')/(1 - \tau')$. In the following proof, we can consider that $\{T_i, i \in \mathcal{I}_{Z_{(k)}^+}\}$ are independent samples with the same mixture distribution as $\tilde{F}_k = \alpha_{1k}F_1 + (1 - \alpha_{1k})F_2$. Let the distribution of $Y_i = \min\{T_i, C_i\}$, $i \in \mathcal{I}_{Z_{(k)}^+}$ be $\tilde{H}_k$. By (A2),

$$
\begin{aligned}
1 - \tilde{H}_k(y) &= \left(1 - \tilde{F}_k(y)\right)\left(1 - G(y)\right) \\
&= \alpha_{1k}\left(1 - F_1(y)\right)\left(1 - G(y)\right) + (1 - \alpha_{1k})\left(1 - F_2(y)\right)\left(1 - G(y)\right) \\
&= \alpha_{1k}\left(1 - H_1(y)\right) + (1 - \alpha_{1k})\left(1 - H_2(y)\right),
\end{aligned}
$$

which shows that Y_i, $i \in \mathcal{I}_{z_{(k)}^+}$ are independent samples with the same mixture distribution $\tilde{H}_k(y) = \alpha_{1k}H_1(y) + (1 - \alpha_{1k})H_2(y)$.

Let $\tilde{\boldsymbol{\beta}}^* = \boldsymbol{\beta}_1^* 1_{\{T \overset{d}{=} F_1\}} + \boldsymbol{\beta}_2^* 1_{\{T \overset{d}{=} F_2\}}$. By Corollary 1.8 of Stute (1993), we obtain

$$
\begin{aligned}
\boldsymbol{\varepsilon}_{2k}^\top \boldsymbol{W}_{2k} \boldsymbol{\varepsilon}_{2k} &= \sum_{i \in \mathcal{I}_{Z_{(k)}^+}} w_{[r_{Y_i}, \mathcal{I}_{Z_{(k)}^+}]} (Y_i - \mathbf{X}_i^\top \tilde{\boldsymbol{\beta}}^*)^2 \\
&\to_{\text{a.s.}} E\left[(T_i - \mathbf{X}_i^\top \tilde{\boldsymbol{\beta}}^*)^2 | Z_i > Z_{(k)}\right] = \alpha_{1k}\sigma_1^2 + (1 - \alpha_{1k})\sigma_2^2,
\end{aligned}
$$

$$
\mathbf{X}_{2k}^\top \boldsymbol{W}_{2k} \mathbf{X}_{2k} \to_{\text{a.s.}} E(\mathbf{X}_i \mathbf{X}_i^\top | Z_i > Z_{(k)}) = \Sigma_{0, Z_{(k)}^+} \le \Sigma_0,
$$

$$
\boldsymbol{\varepsilon}_{2k}^\top \boldsymbol{W}_{2k} \mathbf{X}_{2k} \to_{\text{a.s.}} E(\varepsilon_i \mathbf{X}_i^\top | Z_i > Z_{(k)}) = \mathbf{0}_p^\top.
$$

By Theorem 1 of Stute (1993), we have

$$
\begin{aligned}
\mathbf{X}_{2k}^\top \boldsymbol{W}_{2k} \begin{pmatrix} \mathbf{0}_{kk^*} \\ \mathbf{X}_{2k^*} \end{pmatrix} \mathbf{d}^* &= \sum_{i \in \mathcal{I}_{Z_{(k^*)}^+}} w_{[r_{Y_i}, \mathcal{I}_{Z_{(k)}^+}]} \mathbf{X}_i \mathbf{X}_i^\top \mathbf{d}^* \\
&= \sum_{i \in \mathcal{I}_{Z_{(k)}^+}} w_{[r_{Y_i}, \mathcal{I}_{Z_{(k)}^+}]} \mathbf{X}_i \mathbf{X}_i^\top 1_{\{T_i \overset{d}{=} F_2\}} \mathbf{d}^* \\
&\to_{\text{a.s.}} E\left[(\mathbf{X}_i \mathbf{X}_i^\top 1_{\{T_i \overset{d}{=} F_2\}}) | Z_i > Z_{(k^*)}\right] \mathbf{d}^* \\
&= (1 - \alpha_{1k}) \Sigma_{0, Z_{(k^*)}^+} \mathbf{d}^*.
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
(\mathbf{d}^*)^\top \begin{pmatrix} \mathbf{0}_{kk^*} \\ \mathbf{X}_{2k^*} \end{pmatrix}^\top \boldsymbol{W}_{2k} \begin{pmatrix} \mathbf{0}_{kk^*} \\ \mathbf{X}_{2k^*} \end{pmatrix} \mathbf{d}^* &= (\mathbf{d}^*)^\top \sum_{i \in \mathcal{I}_{Z_{(k^*)}^+}} w_{[r_{Y_i}, \mathcal{I}_{Z_{(k)}^+}]} \mathbf{X}_i \mathbf{X}_i^\top \mathbf{d}^* \\
&\to_{\text{a.s.}} (1 - \alpha_{1k}) (\mathbf{d}^*)^\top \Sigma_{0, Z_{(k^*)}^+} \mathbf{d}^*,
\end{aligned}
$$

$$
\begin{aligned}
(\mathbf{d}^*)^\top \begin{pmatrix} \mathbf{0}_{ka_n} \\ \mathbf{X}_{2k^*} \end{pmatrix}^\top \boldsymbol{W}_{2k} \boldsymbol{\varepsilon}_{2k} &= (\mathbf{d}^*)^\top \sum_{i \in \mathcal{I}_{Z_{(k^*)}^+}} w_{[r_{Y_i}, \mathcal{I}_{Z_{(k)}^+}]} \mathbf{X}_i \varepsilon_i \\
&\to_{\text{a.s.}} (1 - \alpha_{1k}) (\mathbf{d}^*)^\top E(\mathbf{X}_i \varepsilon_i | Z_i > Z_{(k^*)}) = 0.
\end{aligned}
$$

Thus

$$(B.2) \quad \begin{aligned} Q_2(k, \hat{\beta}_{2k}) \to_{\text{a.s.}} (\tau - \tau')\sigma_1^2 + (1 - \tau)\sigma_2^2 \\ + \frac{(1 - \tau)(\tau - \tau')}{1 - \tau'} (\mathbf{d}^*)^\top \Sigma_{0, z_{(k^*)}^+} \mathbf{d}^*. \end{aligned}$$

Combining (B.1) and (B.2), if $0 \le \tau' < \tau$, we have, with probability 1,

$$\begin{aligned} Q(Z_{(k)}) &\to \tau\sigma_1^2 + (1 - \tau)\sigma_2^2 + (1 - \tau)(\tau - \tau')(1 - \tau')^{-1}(\mathbf{d}^*)^\top \Sigma_{0, Z_{(k^*)}^+} \mathbf{d}^* \\ &> \tau\sigma_1^2 + (1 - \tau)\sigma_2^2. \end{aligned}$$

Similarly, if $1 \ge \tau' > \tau$, with probability 1, we can obtain

$$\begin{aligned} Q(Z_{(k)}) &\to \tau\sigma_1^2 + (1 - \tau)\sigma_2^2 + \tau'(1 - \tau')^{-1}(\mathbf{d}^*)^\top \Sigma_{0, Z_{(k^*)}^-} \mathbf{d}^* \\ &> \tau\sigma_1^2 + (1 - \tau)\sigma_2^2, \end{aligned}$$

and $Q(Z_{(k^*)}) \to_{\text{a.s.}} \tau\sigma_1^2 + (1 - \tau)\sigma_2^2$. Thus if $\hat{a} = \arg\min_{\varsigma \in \{Z_{(1)}, \dots, Z_{(n)}\}} \{Q(\varsigma)\} = Z_{(\hat{k})}$, we have $\hat{k}/n \to_{\text{a.s.}} \tau$, that is, $\frac{1}{n} \sum_{i=1}^n 1_{\{Z_i \le \hat{a}\}} \to_{\text{a.s.}} \tau$. If Z_i is a continuous random variable, by continuous mapping theorem, we have $\hat{a} \to_{\text{a.s.}} a$. Therefore, the proof is completed.

B.2. Proof of Theorem 3.2. By the Karush–Kuhn–Tucker (KKT) conditions, we only need to show with probability 1,

$$(B.3) \quad \max_{j \notin \mathcal{A}} \| n^{-1} (\tilde{\mathbf{X}}^{(j)})^\top (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_{\mathcal{A}^*} \hat{\boldsymbol{\theta}}_{\mathcal{A}^*}) \| \le \sqrt{p}\lambda_n,$$

$$(B.4) \quad \min_{\ell=1,\dots,s} \{ \|\hat{\boldsymbol{\theta}}_1\|, \max\{\|\hat{\boldsymbol{\theta}}_{k_{\ell,n}}\|, \|\hat{\boldsymbol{\theta}}_{k_{\ell,n}+1}\|\} \} \ge \sqrt{p}\gamma\lambda_n.$$

First, we prove (B.3). Note that α_j is arbitrary in (A.4). By the condition (A7) and (A8), we put $\alpha_j = 1$ if $k_{j,n} \in \mathcal{A}^*$ and $k_{j,n} + 1 \notin \mathcal{A}^*$, and we put $\alpha_j = 0$ if $k_{j,n} \notin \mathcal{A}^*$ and $k_{j,n} + 1 \in \mathcal{A}^*$. Thus, (A.4) can be rewritten as

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}_{\mathcal{A}^*} \tilde{\boldsymbol{\theta}}_{\mathcal{A}^*}^* + \tilde{\mathbf{x}}_\omega + \tilde{\boldsymbol{\varepsilon}}.$$

By $\hat{\boldsymbol{\theta}}_{\mathcal{A}^*} = (\tilde{\mathbf{X}}_{\mathcal{A}^*}^\top \tilde{\mathbf{X}}_{\mathcal{A}^*})^{-1} \tilde{\mathbf{X}}_{\mathcal{A}^*}^\top \tilde{\mathbf{y}}_{\mathcal{A}^*}$, we obtain

$$\begin{aligned} \|(\tilde{\mathbf{X}}^{(j)})^\top &(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_{\mathcal{A}^*} \hat{\boldsymbol{\theta}}_{\mathcal{A}^*})\| \\ &= \|(\tilde{\mathbf{X}}^{(j)})^\top (I_n - \tilde{\mathbf{X}}_{\mathcal{A}^*} (\tilde{\mathbf{X}}_{\mathcal{A}^*}^\top \tilde{\mathbf{X}}_{\mathcal{A}^*})^{-1} \tilde{\mathbf{X}}_{\mathcal{A}^*}^\top) \tilde{\mathbf{y}}\| \\ &= \|(\tilde{\mathbf{X}}^{(j)})^\top (I_n - \tilde{\mathbf{X}}_{\mathcal{A}^*} (\tilde{\mathbf{X}}_{\mathcal{A}^*}^\top \tilde{\mathbf{X}}_{\mathcal{A}^*})^{-1} \tilde{\mathbf{X}}_{\mathcal{A}^*}^\top)(\tilde{\mathbf{x}}_\omega + \tilde{\boldsymbol{\varepsilon}})\| \\ &\le \|(\tilde{\mathbf{X}}^{(j)})^\top (\tilde{\mathbf{x}}_\omega + \tilde{\boldsymbol{\varepsilon}})\| \le \|(\tilde{\mathbf{X}}^{(j)})^\top \tilde{\mathbf{x}}_\omega\| + \|(\tilde{\mathbf{X}}^{(j)})^\top \tilde{\boldsymbol{\varepsilon}}\|, \end{aligned}$$

where since $\tilde{\mathbf{X}}_{\mathcal{A}^*}(\tilde{\mathbf{X}}_{\mathcal{A}^*}^\top \tilde{\mathbf{X}}_{\mathcal{A}^*})^{-1}\tilde{\mathbf{X}}_{\mathcal{A}^*}^\top$ is a projection matrix, the first inequality comes from Theorem 2.22 of Puntanen (2011), and the second inequality comes from triangle inequality.

By the definition of $\boldsymbol{\omega}_i$, with probability 1, we have

$$
\begin{aligned}
\max_j & \left\| \frac{1}{\sqrt{n}}(\tilde{\mathbf{X}}^{(j)})^\top \tilde{\mathbf{x}}_\omega \right\| \\
&\leq \sum_{j=1}^{q_n+1} \left\| \tilde{\mathbf{X}}_{(j)}^\top \tilde{\mathbf{x}}_{\omega,(j)} \right\| \\
&\leq \sum_{\ell=1}^{s} \left\| \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_{k_{\ell,n}}^{(1)}} \tilde{\mathbf{X}}_i \tilde{\mathbf{X}}_i^\top (-\alpha_\ell) \boldsymbol{\delta}_\ell^* + \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_{k_{\ell,n}}^{(2)}} \tilde{\mathbf{X}}_i \tilde{\mathbf{X}}_i^T (1-\alpha_\ell) \boldsymbol{\delta}_\ell^* \right\| \\
&\leq \sum_{\ell=1}^{s} \|\boldsymbol{\delta}_\ell^*\| \left\| \frac{1}{\sqrt{n}} \tilde{\mathbf{X}}_{(k_{\ell,n})}^\top \tilde{\mathbf{X}}_{(k_{\ell,n})} \right\| \leq \frac{s \max_\ell \sharp\mathcal{I}_{k_{\ell,n}}}{\sqrt{n}} \|\Sigma_0\| \max_{1\leq\ell\leq s} \|\mathbf{d}_\ell^*\|.
\end{aligned}
$$

By (A4), we have $\frac{\max_\ell \sharp\mathcal{I}_{k_{\ell,n}}}{\sqrt{n}} = O(1)$. Thus $\max_j \|\frac{1}{\sqrt{n}}(\tilde{\mathbf{X}}^{(j)})^\top \tilde{\mathbf{x}}_\omega\|$ is bounded with probability 1.

Assume $\boldsymbol{\varphi}_j(\mathbf{X}, Y) = \mathbf{X}^\top(Y - \mathbf{X}^\top \beta_j^*)$. By (A6), $\boldsymbol{\varphi}_j = 0$ for all $Y \in (U_{F_j}^* + U_{E_j}, U_{F_j}]$ which is the condition (2.3) of Stute (1995). From (A1), we have $E\|\boldsymbol{\varphi}_j\|^2 < \infty$. Denote $\tilde{\mathbf{X}}_t = (\tilde{x}_{t,1}, \ldots, \tilde{x}_{t,p})^\top$. By (2.5) in Stute (1995), we have

$$
\text{(B.5)} \qquad \frac{1}{\sqrt{\sharp\mathcal{I}_j}} \sum_{t\in\mathcal{I}_j} \tilde{x}_{t,i} \tilde{\varepsilon}_t = \frac{1}{\sqrt{\sharp\mathcal{I}_j}} \sum_{t\in\mathcal{I}_j} U_{j,i,t} + R_{j,i},
$$

where the $U_{j,i,t}$, $t \in \mathcal{I}_j$ are i.i.d. with mean zero and $V_{j,i} = \text{Var}(U_{j,i,t}) < \infty$, and

$$
|R_{j,i}| = O_{\text{a.s.}}\big((\sharp\mathcal{I}_j)^{-1/2}\log(\sharp\mathcal{I}_j)\big).
$$

The expression of $V_{j,i}$ can be found in Theorem 1.1 of Stute (1996).

Because there are $s + 1$ difference models in each of which T_ℓ has the distribution F_ℓ, $\ell = 1, \ldots, s + 1$, and by $a_\ell \in \mathcal{I}_{k_{\ell,n}}$, $\ell = 1, \ldots, s$, for all $V_{j,i}$, $j \in [k_{\ell-1,n} + 1, k_{\ell,n} - 1]$ have the same value and set $V_{j,i} = V_{\ell,i}^*$, for all $j \in [k_{\ell-1,n} + 1, k_{\ell,n} - 1]$, $\ell = 1, \ldots, s + 1$, where $k_{0,n} = 0$, $k_{s+1,n} = q_n + 2$. Assume $\sharp\mathcal{I}_{k_{\ell,n}}^{(1)}/\sharp\mathcal{I}_{k_{\ell,n}} \to \alpha_\ell$, and then we can consider that $\{t_i, i \in \mathcal{I}_{k_{\ell,n}}\}$ are independent samples with the same mixture distribution as $\tilde{F}_\ell = \alpha_\ell F_\ell + (1 - \alpha_\ell) F_{\ell+1}$. Thus when $j = k_{\ell,n}$, $V_{j,i} = \tilde{V}_{\ell,i} = \alpha_\ell V_{\ell,i}^* + (1 - \alpha_\ell) V_{\ell+1,i}^*$, $\ell = 1, \ldots, s$. Therefore, $V_{j,i}$, $j = 1, \ldots, q_n + 1$, $i = 1, \ldots, p$ have at most $(2s + 1)p$ different values and we denote the greatest value among them by V.

By (B.5), we get

$$\max_{1\leq j\leq q_n+1}\left\|\frac{1}{\sqrt{n}}(\tilde{\mathbf{X}}^{(j)})^\top\tilde{\boldsymbol{\varepsilon}}\right\|$$

$$\leq \sum_{i=1}^{p}\max_{1\leq k\leq q_n+1}\left|\frac{1}{\sqrt{n}}\sum_{j=k}^{q_n+1}\sum_{t\in\mathcal{I}_j}\tilde{x}_{t,i}\tilde{\varepsilon}_t\right|$$

$$\leq \sum_{i=1}^{p}\max_{1\leq k\leq q_n+1}\left|\frac{1}{\sqrt{n}}\sum_{j=k}^{q_n+1}\sum_{t\in\mathcal{I}_j}U_{j,i,t}\right|+\sum_{i=1}^{p}\sum_{j=1}^{q_n+1}\frac{\sqrt{\sharp\mathcal{I}_j}}{\sqrt{n}}|R_{j,i}|.$$

For $n^* = O(n)$, $q_n = O(m) = O(n^{1/2})$, we have $\sum_{i=1}^{p}\sum_{j=1}^{q_n+1}\frac{\sqrt{\sharp\mathcal{I}_j}}{\sqrt{n}}|R_{j,i}| = O_{\text{a.s.}}(\log n)$.

By the Hájek and Rényi inequality [Hájek and Rényi (1955)], for any given $\epsilon > 0$, we have

$$\sum_{n=2}^{\infty}\frac{1}{n}P\left(\max_{1\leq k\leq q_n+1}\left|\frac{1}{\sqrt{n}}\sum_{j=k}^{q_n+1}\sum_{t\in\mathcal{I}_j}U_{j,i,t}\right|\geq\epsilon\log n\right)$$

$$\leq\sum_{n=2}^{\infty}\frac{1}{\epsilon^2 n(\log n)^2}\sum_{j=1}^{q_n+1}\frac{\sharp\mathcal{I}_j}{n}V_{j,i}<\sum_{n=2}^{\infty}\frac{V}{\epsilon^2 n(\log n)^2}<\infty.$$

By Theorem 2.1 of Yang, Su and Yu (2008), we have

$$\sum_{i=1}^{p}\max_{1\leq k\leq q_n+1}\left|\frac{1}{\sqrt{n}}\sum_{j=k}^{q_n+1}\sum_{t\in\mathcal{I}_j}U_{j,i,t}\right|=o_{\text{a.s.}}(\log n).$$

Combining the above result, we have $\frac{1}{\sqrt{n}}\max_{j\notin\mathcal{A}}\|(\tilde{\mathbf{X}}^{(j)})^\top(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}_{\mathcal{A}^*}\hat{\boldsymbol{\theta}}_{\mathcal{A}^*})\| = O_{\text{a.s.}}(\log n)$, but $\sqrt{n}\lambda_n/\log n\to\infty$, and then (B.3) holds with probability 1.

Let $\mathcal{A}^* = \{j_i, i = 1,\ldots,\hat{b}\}$ where $j_1 < j_2 < \cdots < j_{\hat{b}}$. By the condition (A7) and (A8), we have $j_1 = 1$ and $\hat{b} \leq 2s+1$. Put $\tilde{\mathbf{X}}_{(c:d)} = (\tilde{\mathbf{X}}_{(c)}^\top,\tilde{\mathbf{X}}_{(c+1)}^\top,\ldots,\tilde{\mathbf{X}}_{(d)}^\top)^\top$ and $\tilde{\mathbf{y}}_{(c:d)} = (\tilde{\mathbf{y}}_{(c)}^\top,\tilde{\mathbf{y}}_{(c+1)}^\top,\ldots,\tilde{\mathbf{y}}_{(d)}^\top)^\top$ for $c \leq d$, and $\tilde{\mathbb{X}}_i = \tilde{\mathbf{X}}_{(j_i:j_{i+1}-1)}^\top\tilde{\mathbf{X}}_{((j_i:j_{i+1}-1)} = \sum_{k=j_i}^{j_{i+1}-1}\tilde{\mathbf{X}}_{(k)}^\top\tilde{\mathbf{X}}_{(k)}, i = 1,\ldots,\hat{b}, j_{\hat{b}+1} = q_n + 2$. Then we have

$$(\tilde{\mathbf{X}}_{\mathcal{A}^*}^\top\tilde{\mathbf{X}}_{\mathcal{A}^*})^{-1}$$

$$=\begin{pmatrix} \tilde{\mathbb{X}}_1^{-1} & -\tilde{\mathbb{X}}_1^{-1} & 0 & \ldots & 0 & 0 & 0 \\ -\tilde{\mathbb{X}}_1^{-1} & \tilde{\mathbb{X}}_1^{-1}+\tilde{\mathbb{X}}_2^{-1} & -\tilde{\mathbb{X}}_2^{-1} & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -\tilde{\mathbb{X}}_{\hat{b}-2}^{-1} & \tilde{\mathbb{X}}_{\hat{b}-2}^{-1}+\tilde{\mathbb{X}}_{\hat{b}-1}^{-1} & -\tilde{\mathbb{X}}_{\hat{b}-1}^{-1} \\ 0 & 0 & 0 & \ldots & 0 & -\tilde{\mathbb{X}}_{\hat{b}-1}^{-1} & \tilde{\mathbb{X}}_{\hat{b}-1}^{-1}+\tilde{\mathbb{X}}_{\hat{b}}^{-1} \end{pmatrix}.$$

Thus

$$\hat{\boldsymbol{\theta}}_{\mathcal{A}^*} = \left(\tilde{\mathbf{X}}_{\mathcal{A}^*}^\top \tilde{\mathbf{X}}_{\mathcal{A}^*}\right)^{-1} \tilde{\mathbf{X}}_{\mathcal{A}^*}^\top \tilde{\mathbf{y}}_{\mathcal{A}^*}$$

$$(\text{B.6}) \qquad = \begin{pmatrix} \tilde{\mathbb{X}}_1^{-1} \tilde{\mathbf{X}}_{(j_1:j_2-1)}^\top \tilde{\mathbf{y}}_{(j_1:j_2-1)} \\ \tilde{\mathbb{X}}_2^{-1} \tilde{\mathbf{X}}_{(j_2:j_3-1)}^\top \tilde{\mathbf{y}}_{(j_2:j_3-1)} - \tilde{\mathbb{X}}_1^{-1} \tilde{\mathbf{X}}_{(j_1:j_2-1)}^\top \tilde{\mathbf{y}}_{(j_1:j_2-1)} \\ \vdots \\ \tilde{\mathbb{X}}_{\hat{b}}^{-1} \tilde{\mathbf{X}}_{(j_{\hat{b}}:j_{\hat{b}+1}-1)}^\top \tilde{\mathbf{y}}_{(j_{\hat{b}}:j_{\hat{b}+1}-1)} - \tilde{\mathbb{X}}_{\hat{b}-1}^{-1} \tilde{\mathbf{X}}_{(j_{\hat{b}-1}:j_{\hat{b}}-1)}^\top \tilde{\mathbf{y}}_{(j_{\hat{b}-1}:j_{\hat{b}}-1)} \end{pmatrix}.$$

By (2.1), $m = \lfloor c\sqrt{n^*}\rfloor$ and (B.6), we have the following results:

1. $\hat{\boldsymbol{\theta}}_1 \to_{\text{a.s.}} \boldsymbol{\beta}_1^*$.
2. If $k_{j,n} \in \mathcal{A}^*$ and $k_{j+1,n} \notin \mathcal{A}^*$, $\hat{\boldsymbol{\theta}}_{k_{j,n}} \to_{\text{a.s.}} \mathbf{d}_j^*$.
3. If $k_{j,n} \notin \mathcal{A}^*$ and $k_{j+1,n} \in \mathcal{A}^*$, $\hat{\boldsymbol{\theta}}_{k_{j+1,n}} \to_{\text{a.s.}} \mathbf{d}_j^*$.
4. If $k_{j,n} \in \mathcal{A}^*$ and $k_{j+1,n} \in \mathcal{A}^*$, $\hat{\boldsymbol{\theta}}_{k_{j,n}} + \hat{\boldsymbol{\theta}}_{k_{j+1,n}} \to_{\text{a.s.}} \mathbf{d}_j^*$.

By assuming $\min\{\|\beta_1\|_2, \|\delta_j\|_2\} > 2\sqrt{p}\gamma\lambda$, (B.4) holds with probability 1. Hence the proof of Theorem 3.2 is complete.

B.3. Proof of Theorem 3.3. Following the definition of $\hat{s}$ and the condition (3.1), the result (1) can be directly proved by Theorem 3.2. Based on the condition (3.1), the result (2) can be proved by applying Theorem 3.1 to each single change point in disjoint intervals. Under Assumption (A7) and the result (2), $\tilde{\mathbf{X}}_{\hat{\mathbf{a}}}^T \tilde{\mathbf{X}}_{\hat{\mathbf{a}}}/n \to_{\text{a.s.}} \Gamma > 0$. Thus the criterion

$$\frac{1}{2n}\|\tilde{\mathbf{y}}_{\hat{\mathbf{a}}} - \tilde{\mathbf{X}}_{\hat{\mathbf{a}}}\boldsymbol{\theta}\|^2 + \sum_{i=1}^{(s+1)p} p_{\lambda_n,\gamma}(|\theta_i|)$$

is strictly convex. By the KKT condition and result (2), (3) holds with probability 1 in the following events:

$$\max_{j \notin S}\left|n^{-1}(\tilde{\mathbf{X}}_j)^\top(\tilde{\mathbf{y}}_{\hat{\mathbf{a}}} - \tilde{\mathbf{X}}_{\hat{\mathbf{a}}}\hat{\boldsymbol{\theta}}^o)\right| \le \lambda_n \quad \text{and} \quad \min_{j \in S}\{\|\hat{\theta}_j^o\|\} \ge \gamma\lambda_n,$$

which can be proven similarly as the proof of Theorem 3.2. Hence the proof is complete.

REFERENCES

BAI, J. and PERRON, P. (2003). Computation and analysis of multiple structural change models. *J. Appl. Econometrics* **18** 1–22.

BUCKLEY, J. and JAMES, I. (1979). Linear regression with censored data. *Biometrika* **66** 429–436.

DAVE, S. S., WRIGHT, G., TAN, B., ROSENWALD, A., GASCOYNE, R. D., CHAN, W. C. et al. (2004). Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *N. Engl. J. Med.* **351** 2159–2169.

DAVIS, R. A., LEE, T. C. M. and RODRIGUEZ-YAM, G. A. (2006). Structural break estimation for nonstationary time series models. *J. Amer. Statist. Assoc.* **101** 223–239.

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

FEARNHEAD, P. and VASILEIOU, D. (2009). Bayesian analysis of isochores. *J. Amer. Statist. Assoc.* **104** 132–141.

GORDON, A. D. (1981). *Classification*: *Methods for the Exploratory Analysis of Multivariate Data*. Chapman & Hall, New York. MR0637465

HÁJEK, J. and RÉNYI, A. (1955). Generalization of an inequality of Kolmogorov. *Acta Math. Acad. Sci. Hung.* **6** 281–283.

HANSEN, B. E. (2000). Sample splitting and threshold estimation. *Econometrica* **68** 575–603. DOI:10.1111/1468-0262.00124. MR1769379

HARCHAOUI, Z. and LÉVY-LEDUC, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.* **105** 1480–1493.

HUANG, J., MA, S. and XIE, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics* **62** 813–820.

INCLÁN, C. and TIAO, G. C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *J. Amer. Statist. Assoc.* **89** 913–923. MR1294735

JIN, B., SHI, X. and WU, Y. (2013). A novel and fast methodology for simultaneous multiple structural break estimation and variable selection for nonstationary time series models. *Stat. Comput.* **23** 221–231. DOI:10.1007/s11222-011-9304-6. MR3016940

KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

KOSOROK, M. R. and SONG, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *Ann. Statist.* **35**. MR2341694

LAWLESS, J. F. (2011). *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, New York.

LIN, D. Y., WEI, L. J. and YING, Z. (1998). Accelerated failure time models for counting processes. *Biometrika* **85** 605–618. MR1665810

LUO, X., TURNBULL, B. W. and CLARK, L. C. (1997). Likelihood ratio tests for a changepoint with survival data. *Biometrika* **84** 555–565. MR1603981

PERRON, P. (2006). Dealing with structural breaks. In *Palgrave Handbook of Econometrics*, *Vol.* 1: *Econometric Theory* (K. Patterson and T. C. Mills, eds.) 278–352. Palgrave Macmillan, Basingstoke, UK.

PONS, O. (2003). Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *Ann. Statist.* **31** 442–463.

PRENTICE, R. L. (1978). Linear rank test with right censored data. *Biometrika* **65** 167–179.

PUNTANEN, S. (2011). Projection matrices, generalized inverse matrices, and singular value decomposition by Haruo Yanai, Kei Takeuchi, Yoshio Takane. *Int. Stat. Rev.* **79** 503–504.

STUTE, W. (1993). Consistent estimation under random censorship when covariables are present. *J. Multivariate Anal.* **45** 89–103. MR1222607

STUTE, W. (1995). The central limit theorem under random censorship. *Ann. Statist.* **23** 422–439.

STUTE, W. (1996). Distributional convergence under random censorship when covariables are present. *Scand. J. Stat.* **23** 461–471.

TONG, H. (2012). *Threshold Models in Non-linear Time Series Analysis*. Springer, Berlin.

TSIATIS, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18** 354–372.

XIA, X., JIANG, B., LI, J. and ZHANG, W. (2016). Low-dimensional confounder adjustment and high-dimensional penalized estimation for survival analysis. *Lifetime Data Anal.* **22** 547–569.

YANG, S., SU, C. and YU, K. (2008). A general method to the strong law of large numbers and its applications. *Statist. Probab. Lett.* **78** 794–803. MR2409544

YAO, Y.-C. and AU, S. T. (1989). Least-squares estimation of a step function. *Sankhya*, *Ser. A* **51** 370–381. MR1175613

YING, Z. (1993). A large sample study of rank estimation for censored regression data. *Ann. Statist.* **21** 76–99.

YU, T., LI, J. and MA, S. (2012). Adjusting confounders in ranking biomarkers: amodel-based ROC approach. *Brief. Bioinform.* **13** 513–523.

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

DEPARTMENT OF STATISTICS AND
  APPLIED PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE
SINGAPORE, 119077
SINGAPORE
E-MAIL: stalj@nus.edu.sg

DEPARTMENT OF STATISTICS AND FINANCE
UNIVERSITY OF SCIENCE AND TECHNOLOGY
  OF CHINA
HEFEI, ANHUI 230026
CHINA
E-MAIL: jbs@ustc.edu.cn