Annals of the Institute of Statistical Mathematics (2022) 74:69–91
https://doi.org/10.1007/s10463-021-00794-3

Broken adaptive ridge regression for right-censored survival data

Zhihua Sun[1] · Yi Liu[1] · Kani Chen[2] · Gang Li[3]

Received: 1 February 2020 / Revised: 29 December 2020 / Accepted: 18 January 2021 /
Published online: 5 April 2021

Abstract
Broken adaptive ridge (BAR) is a computationally scalable surrogate to L_0-penalized regression, which involves iteratively performing reweighted L_2 penalized regressions and enjoys some appealing properties of both L_0 and L_2 penalized regressions while avoiding some of their limitations. In this paper, we extend the BAR method to the semi-parametric accelerated failure time (AFT) model for right-censored survival data. Specifically, we propose a censored BAR (CBAR) estimator by applying the BAR algorithm to the Leurgan's synthetic data and show that the resulting CBAR estimator is consistent for variable selection, possesses an oracle property for parameter estimation and enjoys a grouping property for highly correlation covariates. Both low- and high-dimensional covariates are considered. The effectiveness of our method is demonstrated and compared with some popular penalization methods using simulations. Real data illustrations are provided on a diffuse large-B-cell lymphoma data and a glioblastoma multiforme data.

Keywords Accelerated failure time model · Grouping effect · L_0 penalization · Right censoring · Variable selection

1 Introduction

L_0-penalized regression, which directly penalizes the cardinality of a model, has been commonly used for variable selection in the low-dimensional setting via well-known information criteria such as Mallow's C_p (Mallows 1973), Akaike's information criterion (AIC) (Akaike 1974), the Bayesian information criterion (BIC)

The research of Gang Li was partly supported by National Institute of Health Grants P30 CA-16042, P50 CA211015, and UL1TR000124-02. The research of Zhihua Sun was partly supported by Natural Science Foundation of China 11871444. The research of Yi Liu was partly supported by Natural Science Foundation of China 11801567.

✉ Gang Li
vli@ucla.edu

Extended author information available on the last page of the article

(Schwarz 1978; Chen and Chen 2008), and risk inflation criteria (RIC) (Foster and George 1994). It has also been shown to possess some optimal properties for variable selection and parameter estimation (Shen et al. 2012; Johnson et al. 2015). However, L_0-penalization is also known to have some limitations such as being computationally NP-hard, not scalable to high-dimensional data, and unstable for variable selection (Breiman 1996). To overcome these shortcomings, the broken adaptive ridge (BAR) method has been recently introduced as a surrogate to L_0 penalization for simultaneous variable selection and parameter estimation under the linear model (Dai et al. 2018, 2020). It was noted by Dai et al. (2018, 2020) that the BAR estimator, defined as the limit of an iteratively reweighted L_2 (ridge) penalization algorithm, retains some appealing properties of L_0 penalization while avoiding its pitfalls. For instance, BAR generally yields a more sparse, accurate, and interpretable model than some popular L_1-type penalization methods such as LASSO and its various variations, while maintaining comparable prediction performance. Moreover, unlike the exact L_0 penalization, BAR is computationally scalable to high-dimensional covariates and is stable for variable selection. Lastly, in addition to being consistent for variable selection and oracle for parameter estimation, the BAR estimator enjoys a grouping property for highly correlated covariates, a desirable feature not shared by most other oracle variable selection procedures.

Because of its appealing properties, the BAR penalization method has been recently extended to the Cox (1972) model with censored survival data (Zhao et al. 2019; Kawaguchi et al. 2020) via penalized likelihood. However, it is well known that the Cox (1972) proportional hazards assumption do not always hold in practice. Thus, it is desirable to extend the BAR penalization method to other common survival regression models. This paper studies an extension of the BAR penalization method to the semi-parametric accelerated failure time (AFT) model, a popular alternative to the Cox model for right censored survival data. To this end, we note that the semi-parametric AFT model is a linear model for the log-transformed survival time with a completely unspecified error distribution, for which the likelihood approach does not yield a consistent parameter estimator even for the classical uncensored linear regression model. Hence, the BAR-penalized likelihood methods of Kawaguchi et al. (2020) and Zhao et al. (2019) for the Cox (1972) do not apply to the semiparmetric AFT model. A different approach would be required.

In this paper, we propose an extension of the BAR penalization method to the semiparametric AFT model by coupling the Leurgans (1987) synthetic data approach with the BAR penalty, study its large sample properties and demonstrate it effectiveness in comparison with some popular penalization methods using simulations. Specifically, we first use the Leurgans (1987) synthetic variable method to construct a synthetic outcome variable and then apply the BAR method for uncensored linear regression (Dai et al. 2018) to the synthetic outcome variable. We then give sufficient conditions under which the proposed censored BAR (CBAR) estimator is consistent for variable selection, behaves asymptotically as well as the oracle estimator based on the true reduced model and possesses a grouping property for highly correlated covariates. We also combine BAR with a sure joint screening method to obtain a two-step variable selection and parameter estimation method for ultra-high-dimensional covariates. Not surprisingly, our simulations demonstrate that the proposed CBAR method generally

yields a more sparse and more accurate model as compared to some other popular penalization methods such as LASSO, SCAD, MCP and adaptive LASSO within the Leurgans (1987) synthetic data framework, which is consistent with the findings of Dai et al. (2018) for uncensored data. Lastly, we have implemented the proposed CBAR method in an R package, named CenBAR, and made it publicly available at https://CRAN.R-project.org/package=CenBAR.

Before going further, we note that there exist a number of other variable selection methods in the literature for the semiparametric the AFT model with right censored data. These methods are derived by combining various penalization methods such as LASSO with different extensions of the least squares principle for right censored data. For example, the Lasso, bridge, elastic net or MCP penalties have been combined with the Stute (1993) weighted least squares method (Huang et al. 2006; Huang and Ma 2010; Datta et al. 2007); and the Dantzig, elastic net, Lasso, adaptive Lasso and SCAD penalties have been combined with the Buckley and James (1979) method (Wang et al. 2008; Johnson et al. 2008; Johnson 2009; Li et al. 2014). This paper makes a unique theoretical contribution since neither the BAR penalization nor the Leurgans (1987) synthetic data method has been previously rigorously studied in the context of variable selection for the semiparametric the AFT model. We also illustrate and compare empirically the BAR penalization versus some popular penalization methods when the Leurgans (1987) synthetic data least squares method is used. We do not compare different penalization methods when they are coupled with different censored least squares methods because different censored least squares methods are derived under different conditions and none is expected to dominate another across all scenarios.

The rest of the paper is organized as follows. In Sect. 2, we define our CBAR estimator and state its theoretical properties. We also discuss how to handle ultra-high-dimensional covariates. In Sect. 3, we evaluate the finite sample performance of CBAR in comparison with other penalization methods via extensive simulations. In Sect. 4, we illustrate the CBAR method on a diffuse large-B-cell lymphoma data and a glioblastoma multiforme data with high-dimensional covariates. Proofs of the theoretical results are provided in the appendix.

2 Censored broken adaptive ridge (CBAR) regression

2.1 Notations and preliminaries

2.1.1 Model and data

Consider the linear regression model

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, ..., n, \tag{1}$$

where for the ith subject, Y_i denotes the response variable, $\mathbf{x}_i$ is the p_n-vector random covariates, $\boldsymbol{\beta} = (\beta_1, ..., \beta_{p_n})^\top$ is a vector of regression coefficients, and ε_i is i.i.d. error term with an unknown error distribution, $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2 < \infty$. Model

(1) is commonly referred to as the accelerated failure time (AFT) model when Y is the log-transformed survival time (Kalbfleisch and Prentice 2002).

Without loss of generality, assume that $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^\top, \boldsymbol{\beta}_{02}^\top)^\top$ is the true value of $\boldsymbol{\beta}$, where $\boldsymbol{\beta}_{01}$ is a $q \times 1$ nonzero vector and $\boldsymbol{\beta}_{02}$ is a $(p_n - q) \times 1$ zero vector. We further assume the columns of the design matrix $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)^\top$ have mean zero and unit L_2-norm. Throughout the paper, $\|\cdot\|$ represents the Euclidean norm for a vector and spectral norm for a matrix.

Assume that one observes a right censored data consisting of n independent and identically distributed triples $(T_i, \delta_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, where for the ith subject, $T_i = \min(Y_i, C_i)$ is the observation time, $\delta_i = I(Y_i \le C_i)$ is a censoring indicator, C_i is the i.i.d. censoring time with the distribution function H. C_i is assumed to be independent of Y_i and $\mathbf{x}_i$.

2.1.2 Broken adaptive ridge (BAR) for uncensored data

For reader convenience, we first briefly review the broken adaptive ridge (BAR) estimator of Dai et al. (2018) for simultaneous variable selection and parameter estimation with the uncensored data $\mathbf{Y}$ and $\mathbf{x}$, where $\mathbf{Y} = (Y_1, ..., Y_n)^\top$.

Following the notations of Dai et al. (2018), the BAR estimator of $\boldsymbol{\beta}$ based on $\mathbf{Y}$ and $\mathbf{x}$ is a surrogate L_0-penalized estimator defined as the limit of the following iteratively reweighted ridge regression algorithm:

$$\begin{aligned}\boldsymbol{\beta}^{(k)} &= \arg\min_{\boldsymbol{\beta}} \{\|\mathbf{Y} - \mathbf{x}\boldsymbol{\beta}\|^2 + \lambda_n \sum_{j=1}^{p_n} \frac{\beta_j^2}{\{\boldsymbol{\beta}_j^{(k-1)}\}^2}\} \\ &= \{\mathbf{x}^\top\mathbf{x} + \lambda_n \mathbf{D}(\boldsymbol{\beta}^{(k-1)})\}^{-1}\mathbf{x}^\top\mathbf{Y}, \quad k = 1, 2, ...\end{aligned} \tag{2}$$

where $\boldsymbol{\beta}^{(0)} = \arg\min_\beta \{\|\mathbf{Y} - \mathbf{x}\boldsymbol{\beta}\|^2 + \xi_n \sum_{j=1}^{p_n} \beta_j^2\} = (\mathbf{x}^\top\mathbf{x} + \xi_n\mathbf{I})^{-1}\mathbf{x}^\top\mathbf{Y}$ is an initial ridge estimator, $\xi_n > 0$ and $\lambda_n \ge 0$ are tuning penalization parameters, and for any p_n -dimensional vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_{p_n})^T$, $\mathbf{D}(\boldsymbol{\theta}) = \text{diag}\,(\frac{1}{\theta_1^2}, ..., \frac{1}{\theta_{p_n}^2})$. Note that each reweighted L_2 penalty can be regarded as an adaptive surrogate L_0 penalty and the approximation of L_0 penalization improves with each iteration. Dai et al. (2018) showed that the BAR estimator $\hat{\boldsymbol{\beta}} = \lim_{k\to\infty} \boldsymbol{\beta}^{(k)}$ is selection consistent and possesses an oracle property: if the true model is sparse with some zero coefficients, then with probability tending to 1, BAR estimates the true zero coefficients as zeros and estimates the non-zero coefficients as well as the scenario when the true sub-model is known in advance.

2.2 Broken adaptive ridge estimator for censored data (CBAR)

For right censored data, the above BAR algorithm is obviously not applicable since one only observes (T_i, δ_i) instead of Y_i. To overcome the problem, we propose to adopt the Leurgans (1987) synthetic data approach for censored linear regression to variable selection by first transforming (T_i, δ_i) into a synthetic variable Y_i^* and then applying the

BAR method to the synthetic data variable Y_i^*. Specifically, the Leurgans (1987) synthetic data Y_i^* is defined as

$$Y_i^* = \int_{-\infty}^{T^n} \left(\frac{I(T_i \geq s)}{1 - \hat{H}(s)} - I(s < 0) \right) \mathrm{d}s, \tag{3}$$

where $T^n = \max\{T_1, ..., T_n\}$ and $\hat{H}$ is the Kaplan-Meier estimator of H. To apply the BAR method to synthetic data Y_i^*, let $\mathbf{Y}^* = (Y_1^*, ..., Y_n^*)^\top$ and define an initial ridge estimator

$$\hat{\boldsymbol{\beta}}^{(0)} = (\mathbf{x}^\top \mathbf{x} + \xi_n \mathbf{I})^{-1} \mathbf{x}^\top \mathbf{Y}^*, \tag{4}$$

and then, for $k \geq 1$,

$$\hat{\boldsymbol{\beta}}^{(k)} = g(\hat{\boldsymbol{\beta}}^{(k-1)}), \tag{5}$$

where

$$g(\tilde{\boldsymbol{\beta}}) = \arg\min_{\boldsymbol{\beta}} \{ \|\mathbf{Y}^* - \mathbf{x}\boldsymbol{\beta}\|^2 + \lambda_n \sum_{j=1}^{p_n} \frac{\beta_j^2}{\tilde{\beta}_j^2} \} = \{ \mathbf{x}^\top \mathbf{x} + \lambda_n \mathbf{D}(\tilde{\boldsymbol{\beta}}) \}^{-1} \mathbf{x}^\top \mathbf{Y}^*. \tag{6}$$

Finally, the CBAR estimator is defined as

$$\hat{\boldsymbol{\beta}}^* = \lim_{k \to \infty} \hat{\boldsymbol{\beta}}^{(k)}. \tag{7}$$

In the next section, we give conditions under which the CBAR estimator $\hat{\boldsymbol{\beta}}^*$ is selection consistent and has an oracle property for estimation of the nonzero component $\boldsymbol{\beta}_{01}$ of $\boldsymbol{\beta}$.

2.3 Large sample properties of CBAR

Similar to Zhou (1992), define $F_i(t) = P\{Y_i \geq t\}$, $G_i(t) = P\{T_i \geq t\} = F_i(t)(1 - H(t))$, $K(t) = -\int_0^t \frac{1}{\lim(1/n)\sum F_i} \frac{\mathrm{d}G}{G^2}$ and denote

$$\Lambda_i^+(t) = -\int_0^t \frac{\mathrm{d}G_i(s)}{G_i(s^-)}, \ \Lambda_i^D(t) = -\int_0^t \frac{\mathrm{d}F_i(s)}{F_i(s^-)}, \ \Lambda^C(t) = \int_0^t \frac{\mathrm{d}H(s)}{1 - H(s^-)}.$$

Then,

$$M_i^+(t) = I_{[T_i \leq t]} - \int_0^t I_{[T_i \geq s]} \mathrm{d}\Lambda_i^+(s),$$

$$M_i^D(t) = I_{[T_i \leq t; \delta_i = 1]} - \int_0^t I_{[T_i \geq s]} \mathrm{d}\Lambda_i^D(s),$$

$$M_i^C(t) = I_{[T_i \le t; \delta_i = 0]} - \int_0^t I_{[T_i \ge s]} \mathrm{d}\Lambda_i^C(s)$$

are square-integrable martingales and satisfies $M_i^+ = M_i^D + M_i^C$ (Zhou 1992). Let $\mathbf{\Omega}(\tau) = (\sigma_{kl}(\tau))$ be defined by

$$\begin{aligned} \sigma_{kl}(\tau) = &\lim n \sum_{i=1}^{n} \omega_{ki}\omega_{li} \int_0^\tau \left[\int_t^\tau F_i \mathrm{d}s \right]^2 \frac{\mathrm{d}\Lambda_i^D(t)}{G_i} \\ &+ \lim n \sum_{i=1}^{n} \int_0^\tau \prod_{c_i = \omega_{ki}, \omega_{li}} \left[\frac{\sum c_j \int_t^\tau F_j \mathrm{d}s}{(1-H)\sum F_j} - \frac{c_k \int_t^\tau F_i \mathrm{d}s}{G_i} \right] G_i \mathrm{d}\Lambda^C, \end{aligned} \tag{8}$$

where $\omega_{ji} = ((\mathbf{x}^\top \mathbf{x})^{-1}\mathbf{x}^\top)_{ji}$. Let $\boldsymbol{\omega}_i$ denote the ith column of the matrix $(\mathbf{x}^\top\mathbf{x})^{-1}\mathbf{x}^\top$, $\mathbf{x}_1$ denote the first q_n columns of $\mathbf{x}$, $\mathbf{\Sigma}_n = n^{-1}\mathbf{x}^\top\mathbf{x}$ and $\mathbf{\Sigma}_{n1} = n^{-1}\mathbf{x}_1^\top\mathbf{x}_1$. Write $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\beta}}_1^{*^\top}, \hat{\boldsymbol{\beta}}_2^{*^\top})^\top$, where $\hat{\boldsymbol{\beta}}_1^*$ is a $q \times 1$ vector estimator of $\boldsymbol{\beta}_{01}$ and $\hat{\boldsymbol{\beta}}_2^*$ is a $(p_n - q) \times 1$ vector estimator of $\boldsymbol{\beta}_{02}$.

The following conditions are needed for our theoretical derivations.

(C1) $\sup_t E(\varepsilon_i - t | \varepsilon_i > t) < \infty$, and for any p_n-vector $\mathbf{b}_n$ satisfying $\|\mathbf{b}_n\| \le 1$, $\mathbf{b}_n^\top \mathbf{\Omega}(\tau)\mathbf{b}_n$ is finite for $\tau \in [K, \infty]$ and $\mathbf{b}_n^\top \mathbf{\Omega}(\tau)\mathbf{b}_n \to \mathbf{b}_n^\top \mathbf{\Omega}(\infty)\mathbf{b}_n$ as $\tau \to \infty$.

(C2) $\sup_n \int_0^\infty \sum_{i=1}^n (\mathbf{b}_n^\top \boldsymbol{\omega}_i)^2 \sum_{i=1}^n F_i^2 dK(t) < \infty$ for any p_n-vector $\mathbf{b}_n$ satisfying $\|\mathbf{b}_n\| \le 1$. X_i are bounded, and for some constants $C^* > 0$ and $S < 1$, $C^* F_i(t)^S \le 1 - H(t)$.

(C3) $\int_0^\infty \{ \frac{n \sum (\mathbf{b}_n^\top \boldsymbol{\omega}_i)^2 F_i}{1 - H(s)} \}^{\frac{1}{2}} \mathrm{d}s \le M < \infty$ and $\int_0^\infty K^{1/2}(t) | \sum \mathbf{b}_n^\top \boldsymbol{\omega}_i F_i | \mathrm{d}t < \infty$ for any p_n -vector $\mathbf{b}_n$ satisfying $\|\mathbf{b}_n\| \le 1$.

(C4) There exists a constant $\tilde{C} > 1$ such that $0 < 1/\tilde{C} < \lambda_{\min}(\mathbf{\Sigma}_n) \le \lambda_{\max}(\mathbf{\Sigma}_n) < \tilde{C} < \infty$ for every integer n.

(C5) Let $a_0 = \min_{1 \le j \le q} |\beta_{0j}|$ and $a_1 = \max_{1 \le j \le q} |\beta_{0j}|$. As $n \to \infty$, $p_n/\sqrt{n} \to 0$, $\xi_n/\sqrt{n} \to 0$ and $\lambda_n/\sqrt{n} \to 0$.

Conditions (C1)-(C3) are regularity conditions required to establish the asymptotic properties of the unpenalized synthetic data least squares estimator under diverging dimension. Conditions (C4) and (C5) are additional conditions needed to derive the selection consistency and oracle property of the synthetic data BAR estimator of this paper as stated in Theorem 1 below.

Theorem 1 (Oracle property) *Assume conditions (C1)–(C5) hold. For any q-dimensional vector $\mathbf{c}$ satisfying $\|\mathbf{c}\| \le 1$, define $z^2 = \mathbf{c}^\top \mathbf{\Omega}_1 \mathbf{c}$, where $\mathbf{\Omega}_1$ is the first $q \times q$ sub-matrix of $\mathbf{\Omega}(\infty)$. Define $f(\boldsymbol{\alpha}) = \{\mathbf{x}_1^\top \mathbf{x}_1 + \lambda_n \mathbf{D}_1(\boldsymbol{\alpha})\}^{-1} \mathbf{x}_1^\top \mathbf{Y}^*$, where $\mathbf{D}_1(\boldsymbol{\alpha}) = \mathrm{diag}(\alpha_1^{-2}, \ldots, \alpha_q^{-2})$. Then, with probability tending to 1,*

(i) $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\beta}}_1^{*^\top}, \hat{\boldsymbol{\beta}}_2^{*^\top})^\top$ *exists and is unique, with $\hat{\boldsymbol{\beta}}_2^* = 0$ and $\hat{\boldsymbol{\beta}}_1^*$ being the unique fixed point of $f(\boldsymbol{\alpha})$;*

(ii) $\sqrt{n} z^{-1} \mathbf{c}^\top (\hat{\boldsymbol{\beta}}_1^* - \boldsymbol{\beta}_{01}) \to_D N(0, 1)$.

Part (i) of the above theorem guarantees that the CBAR estimator is consistent for variable selection. Part (ii) states that the asymptotic distribution of the nonzero component of the CBAR estimator is the same as the one when the true model is known in advance. The proof of Theorem 1 is deferred to the Appendix.

2.4 Grouping effect

When the true model has a group structure, it would be desirable for a variable selection method to either retain or drop all variables that are clustered within the same group. Below we establish that the CBAR estimator possesses a grouping property in the sense that highly correlated covariates tend to be grouped together with similar coefficients. With similarly estimated coefficients, the highly correlated covariates would likely to be retained together by the CBAR method.

Theorem 2 *Assume that the columns of matrix* $\mathbf{x}$ *are standardized and* $\mathbf{Y}^*$ *is centered. Let* $\hat{\boldsymbol{\beta}}^*$ *be the CBAR estimator and* $\hat{\beta}_i^* \hat{\beta}_j^* > 0$, *then, with probability tending to* 1,

$$|\hat{\beta}_i^{*-1} - \hat{\beta}_j^{*-1}| \leq \frac{1}{\lambda_n} \|\mathbf{Y}^*\| \sqrt{2(1 - r_{ij})}, \tag{9}$$

where $r_{ij} = \mathbf{x}_i^\top \mathbf{x}_j$ *is the sample correlation of* $\mathbf{x}_i$ *and* $\mathbf{x}_j$.

The above result implies that the estimated coefficients of two highly positively-correlated variables will be similar in magnitude. The proof of Theorem 2 is given in the Appendix. Similarly, it can be shown that the estimated coefficients of two highly negatively-correlated variables will also be similar in magnitude. It is interesting to note that this grouping property of CBAR is similar to that of Elastic Net (Zou and Hastie 2005). In particular, Zou and Hastie (2005) showed that for the linear model with uncensored data, if $\hat{\beta}_i(\text{EN})\hat{\beta}_j(\text{EN}) > 0$, then

$$|\hat{\beta}_i(\text{EN}) - \hat{\beta}_j(\text{EN})| \leq \frac{1}{\lambda} \|\mathbf{Y}\| \sqrt{2(1 - r_{ij})},$$

where $\hat{\beta}_i(\text{EN})$ is the Elastic Net estimate of β_i and λ is a tuning parameter. It is also worth noting that this grouping property is different from some well known grouped variable selection methods such as grouped LASSO (Yuan and Lin 2006; Nardi and Rinaldo 2008) because the CBAR does not assume the underlying group structure is known in advance, whereas grouped LASSO makes use of a pre-specified group structure.

2.5 Ultra-high-dimensional covariates

Theorem 1 is established under a sufficient condition that $p_n < n$. In many applications, p_n can be much larger than the sample size n. For high dimensional

problems, a common strategy is to proceed a variable selection method with a sure screening dimension reduction step (Fan and Lv 2008; Zhu et al. 2011; Cui et al. 2015). This strategy also applies to the semiparametric AFT model with right censored data. For example, one can first apply the sure joint screening method BJASS of Liu et al. (2020) to obtain a lower dimensional model and then apply the CBAR method to the reduced model. We refer to the resulting two-step estimator $\hat{\boldsymbol{\beta}}^*$ as the BJASS-CBAR estimator.

Below we give some additional sufficient conditions under which the BJASS-CBAR estimator $\hat{\boldsymbol{\beta}}^*$ has an oracle property.

(D1) $\log(p) = O(n^d)$ for some $0 \le d < 1$.

(D2) $P(t \le Y_i \le C_i) \ge \tau_0 > 0$ for some positive constant τ_0 and any $t \in [0, \varsigma]$, where ς denotes the maximum follow up time. Furthermore, $\sup\{t : P(Y > t) > 0\} \ge \sup\{t : P(C > t) > 0\}$. $H(t)$ has uniformly bounded first derivative.

(D3) $\min_{j \in s^*} |\beta_j^*| \ge \omega_1 n^{-\tau_1}$ and $q < k \le \omega_2 n^{\tau_2}$ for some positive constants ω_1, ω_2 and nonnegative constants τ_1, τ_2 satisfying $\tau_1 + \tau_2 < 1/3$, where k is the size of the screened model from BJASS.

(D4) For sufficiently large n, $\lambda_{\min}(n^{-1}\mathbf{x}_s^\top \mathbf{x}_s) \ge c_1$ for some constant $c_1 > 0$ and all $s \in S_+^{2k}$, where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a matrix, and $S_+^k = \{s : s^* \subset s; \|s\|_0 \le k\}$ denotes the collection of the over-fitted models of cardinality k or smaller.

(D5) Let $\sigma_i^2 = \int\int [\frac{G_i(s \vee t)}{(1-H(s))(1-H(t))} - F_i(s)I(t < 0) - F_i(t)I(s < 0) + I(s < 0)I(t < 0)]\mathrm{d}s\mathrm{d}t - E^2(Y_i)$. There exist positive constants c_2, c_3, c_4, σ such that $|X_{ij}| \le c_2$, $|X_i^\top \boldsymbol{\beta}^*| \le c_3$, $|\sigma_i| \le \sigma$ and for sufficiently large n,

$$\max_{1 \le j \le p} \max_{1 \le i \le n} \left\{ \frac{X_{ij}^2}{\sum_{i=1}^n X_{ij}^2 \sigma_i^2} \right\} \le c_4 n^{-1}.$$

(D6) There are positive constants K_1, K_2 and τ_3 such that

$$P(|\epsilon| \ge M) \le K_1 \exp(-K_2 M^{\tau_3}),$$

for any $M = O(n^\tau) > 0$, where $\tau \ge 0$, $\tau_1 + \tau_2 + \tau < (1-d)/2$, $\tau_2 + d - \tau\tau_3 < 0$, and $2\tau_2 + 2\tau + d < 1/3$.

Theorem 3 (Oracle property of the BJASS-CBAR estimator) *Assume that conditions (D1)–(D6) hold for the full model of size p and that the assumptions of Theorem 1 hold for the BJASS reduced model of size k. Then, with probability tending to 1,*

(i) $\hat{\boldsymbol{\beta}}_2^* = 0$;

(ii) $\hat{\boldsymbol{\beta}}_1^*$ *performs as well as the oracle estimator for the true model* $\mathcal{M}_* = \{1 \le j \le q\}$ *in the sense of part (ii) of Theorem 1.*

The above result is a direct consequence of Theorems 4 of Liu et al. (2020) and the oracle property of CBAR stated in Theorem 1. In Sect. 3.2, we present a simulation study to illustrate the advantages of BJASS-BAR with $k = 2\log(n) * n^{(1/4)}$ in comparison with some other penalization methods under a high-dimensional setting.

3 Simulations

We present some simulations to illustrate the effectiveness of the proposed CBAR estimator for variable selection, prediction, parameter estimation in comparison with some popular penalization methods including Lasso (Tibshirani 1996), adaptive Lasso (Zou 2006), SCAD (Fan and Li 2001) and MCP (Zhang 2010)), in the context of the Leurgans (1987) synthetic data framework. We use the R package `glmnet` (Friedman et al. 2010) for Lasso and adaptive Lasso and R package `ncvreg` (Breheny and Huang 2011) for SCAD and MCP, performed on the Leurgans (1987) synthetic data outcome. Fivefold cross-validation (CV) is used to select tuning parameters for all methods. For CBAR, we all use 10 equally log-spaced grid points on $[a, b]$ for the paths of λ_n and ξ_n where $a = 1e^{-4}$ and $b = \max\left\{\frac{(\mathbf{x}_j^{\top}\mathbf{Y})^2}{4\mathbf{x}_j^{\top}\mathbf{x}_j}\right\}_{j=1}^{p}$.

3.1 Simulation 1: $p_n < n$

We consider the following two model settings similar to Tibshirani (1997), Fan and Li (2002), Cai et al. (2009):

Model 1: $Y_i = \mathbf{x}_i^{\top}\boldsymbol{\beta}_0 + \varepsilon_i$, where the covariate vector $\mathbf{x}_i$ is generated from a multivariate normal distribution with mean 0 and variance-covariance matrix $\boldsymbol{\Sigma} = (\rho^{|i-j|})$, and the error ε_i has the standard normal distribution and is independent of the covariates. The true parameter value is $\boldsymbol{\beta}_0 = (3, -2, 0, 0, 6, 0, \ldots, 0)^{\top}$.

Model 2: The same as Model 1 except that $\boldsymbol{\beta}_0 = (3, -2, 6, 0.3, -0.2, 0.6, 0, \ldots, 0)^{\top}$.

Note that Model 1 contains strong signals, whereas Model 2 includes both strong and weak signals. The censoring variable C_i is generated from the normal distribution $N(c, 2)$, where c is chosen to yield a desired level of censoring rate.

The variable selection performance is assessed using five measures: the mean number of misclassified non-zeros and zeros (MisC), mean of false non-zeros (FP), mean of false zeros (FN), probability that the selected model is identical to the true model (TM), and a similarity measure (SM) between the selected set $\hat{S}$ and the true active set $|S|_0$: $SM = \frac{|\hat{S}\cap S|_0}{\sqrt{|\hat{S}|_0|S|_0}}$, where $|.|_0$ denotes model size. The prediction performance is measured by the mean squared prediction error (MSPE) from the fivefold CV. The parameter estimation performance is measured by the mean of the absolute bias of the parameter estimator (MAB). We have run extensive simulations for a variety of settings by varying n, p, ρ and the censoring rate, with 1000 Monte Carlo replications for each setting. Part of the findings are presented in Table 1.

Table 1 Comparison of CBAR with Lasso, SCAD, MCP, and Adaptive Lasso (ALasso) when coupled with the Leurgans (1987) synthetic data procedure based on 1000 Monte-Carlo replications

Model	p	Method	MisC	FP	FN	TM	SM	MSPE	MAB
1	10	CBAR	**0.60**	**0.60**	0	**74%**	**0.94**	**8.86**	1.50
		Lasso	3.05	3.05	0	6.6%	0.73	9.28	2.29
		SCAD	1.11	1.11	0	46.2%	0.89	9.03	1.47
		MCP	0.76	0.76	0	63.6%	0.92	9.01	1.45
		Alasso	1.12	1.12	0	49.2%	0.89	8.91	1.68
	50	CBAR	**0.73**	**0.71**	0.02	**74.80%**	**0.94**	**8.9**	1.69
		Lasso	7.33	7.33	0	1.7%	0.58	9.77	3.36
		SCAD	2.96	2.96	0	21.3%	0.76	9.09	1.72
		MCP	1.24	1.23	0.01	47.7%	0.88	9.03	1.56
		Alasso	6.09	6.09	0	15.2%	0.67	8.69	3.04
	80	CBAR	**0.86**	**0.84**	0.02	**72.3%**	**0.93**	8.84	1.81
		Lasso	9.40	9.40	0	1.30%	0.54	10.06	3.79
		SCAD	3.90	3.90	0	15.2%	0.72	9.33	1.89
		MCP	1.41	1.40	0.01	45.6%	0.87	9.26	1.65
		Alasso	11.09	11.08	0.01	11.8%	0.59	**8.74**	4.51
	90	CBAR	**0.94**	**0.92**	0.02	**69.7%**	**0.93**	**8.93**	1.88
		Lasso	9.36	9.36	0	1.4%	0.54	10.05	3.82
		SCAD	4.09	4.09	0	13.5%	0.71	9.27	1.91
		MCP	1.44	1.43	0.01	43.2%	0.87	9.20	1.64
		Alasso	4.29	4.27	0.02	10.5%	0.70	9.13	2.69
2	10	CBAR	2.61	**0.65**	1.96	0.9%	0.77	9.36	2.36
		Lasso	3.00	2.14	**0.86**	2.2%	0.78	9.50	2.74
		SCAD	2.64	1.10	1.54	2.3%	0.78	9.35	2.35
		MCP	2.64	0.86	1.78	1.9%	0.77	9.34	2.36
		Alasso	**2.50**	0.92	1.58	**3.1%**	**0.79**	**9.14**	2.37
	50	CBAR	**3.65**	**1.03**	2.62	0.1%	**0.69**	9.41	2.92
		Lasso	9.75	7.99	**1.76**	0%	0.52	10.40	4.37
		SCAD	5.57	3.40	2.17	0%	0.61	9.84	2.77
		MCP	3.92	1.46	2.46	0%	0.67	9.80	2.64
		Alasso	9.18	7.22	1.96	0.1%	0.55	**9.21**	4.27
	80	CBAR	**3.89**	**1.19**	2.70	0%	**0.68**	9.11	3.02
		Lasso	11.61	9.69	**1.92**	0%	0.48	10.39	4.70
		SCAD	6.48	4.21	2.27	0%	0.57	9.66	2.86
		MCP	4.06	1.49	2.57	0%	0.66	9.60	2.64
		Alasso	13.82	11.78	2.04	0%	0.48	**8.99**	5.55
	90	CBAR	**3.85**	**1.16**	2.69	0%	**0.68**	9.31	3.05
		Lasso	12.44	10.47	**1.97**	0%	0.46	10.20	4.86
		SCAD	6.92	4.67	2.25	0%	0.56	9.40	2.92
		MCP	4.24	1.68	2.56	0%	0.65	9.36	2.68
		Alasso	7.00	4.50	2.50	0%	0.54	**9.27**	3.73

Data settings: $n = 100$, $p \in \{10, 50, 80, 90\}$. (MisC = mean number of misclassified non-zeros and zeros; FP = mean of false positives (non-zeros); FN = mean of false negatives (zeros); TM = probability that the selected model is exactly the true model; SM = similarity measure; MSPE = mean squared predic-

Table 1 (continued)

tion error from fivefold CV or five-jointly CV and MAB = mean of the absolute bias of the parameter estimator)

Boldface indicates the best performance among the five methods

It is seen from Table 1 that CBAR stands out as the top or top two performers with respect to almost all variable selection performance measures (MisC, FP, TM and SM). In particular, CBAR generally yields a more sparse and accurate model with the largest TM and SM, and much lower MisC and FP. Also, using fewer active features, CBAR achieves comparable prediction accuracy as other methods that use more features. For estimation, CBAR, SCAD and MCP are comparable with similar bias (MAB), whereas Lasso and Adaptive lasso can be substantially worse.

3.2 Simulation 2: $p_n >> n$

In this simulation, we consider the same models as in Simulation 1, except in a high-dimensional setting with $n = 200$, $p = 1000$. We again compared the same five penalization methods, with each method proceeded with the sure joint screening method BJASS of Liu et al. (2020) with $k = 2log(n) * n^{(1/4)}$ for the semi-parametric AFT model to yield a two-step sparse estimator. We denote these methods by BJASS-CBAR, BJASS-Lasso, BJASS-SCAD, BJASS-MCP and BJASS-ALasso. The censoring rate is 0.2. The results are summarized in Table 2.

It is observed from Table 2 that although most penalization methods had comparable performance in terms of estimation bias (MAB) and prediction error (MSPE),

Table 2 Comparison of BJASS-CBAR with CBAR with BJASS-Lasso, BJASS-SCAD, BJASS-MCP, and BJASS-ALasso when coupled with the Leurgans (1987) synthetic data procedure in a high-dimensional setting: $n = 200$, $p = 1000$

Model	Method	MisC	FP	FN	TM	SM	MAB	MSPE
1	BJASS-CBAR	**2.24**	**2.15**	0.09	**63%**	**0.93**	2.87	10.40
	BJASS-Lasso	12.61	12.55	0.06	0%	0.63	4.79	10.87
	BJASS-SCAD	4.23	4.14	0.09	20%	0.82	2.79	10.46
	BJASS-MCP	2.82	2.73	0.09	43%	0.88	**2.69**	10.45
	BJASS-ALasso	8.08	8.00	0.08	12%	0.73	4.05	10.35
2	BJASS-CBAR	**6.15**	**3.15**	3	**41%**	**0.69**	2.51	12.17
	BJASS-Lasso	17.14	14.14	3	0%	0.49	4.49	12.64
	BJASS-SCAD	8.68	5.68	3	7%	0.62	2.09	12.39
	BJASS-MCP	6.38	3.38	3	26%	0.68	**1.96**	12.38
	BJASS-ALasso	12.78	9.78	3	3%	0.54	3.75	11.91

MisC= mean number of misclassified non-zeros and zeros; FP = mean of false positives (non-zeros); FN = mean of false negatives (zeros); TM = probability that the selected model is exactly the true model; SM = similarity measures; MSPE = mean squared prediction error from fivefold CV or five-jointly CV and MAB = mean of the absolute bias of the parameter estimator

Boldface indicates the best performance among the five methods

BJASS-CBAR outperformed the other methods in the variable selection domain with the lowest MisC, FP and the largest TM and SM, which are consistent with the simulation results for the low-dimension $p_n < n$ settings in Simulation 1.

4 Real data examples

We illustrate the CBAR method on two real data sets with high-dimensional covariates.

4.1 Diffuse large-B-cell lymphoma data

The diffuse large-B-cell lymphoma (DLBCL) data includes $n = 240$ patients and $p = 7399$ gene features, which was downloaded from http://statweb.stanford.edu/~tibs/superpc/staudt.html. We first apply the BJASS sure joint screening method of Liu et al. (2020) to reduce data dimension to $k = 2\log(n)n^{\frac{1}{4}} = 43$ and then apply CBAR and four other popular penalization methods. The results are summarized in Table 3.

It is seen that BJASS-CBAR is among the most sparse model and has the smallest CV error, which is consistent with the findings in the simulation studies.

4.2 Glioblastoma multiforme data

The glioblastoma multiforme (GBM) methylation data was downloaded from the TCGA program (https://www.cancer.gov/tcga) using TCGA-Assembler 2 (TA2). The initial data consists of 577 patients and 20,156 GBM methylation variables. After removing missing data, the complete case data includes $n = 136$ patients and

Table 3 Estimated coefficients of BAJSS-CBAR, BAJSS-Lasso, BAJSS-SCAD, BAJSS-MCP and BAJSS-Alasso for the DLBCL data

Parameter	BAJSS-CBAR	BAJSS-Lasso	BAJSS-SCAD	BAJSS-MCP	BAJSS-Alasso
1456	−0.0591	−0.394	−0.609	−0.630	−0.513
1819		−0.069			
1863		−0.006			
2603		−0.025			
2672		−0.062			
3236	−0.480	−0.348	−0.394	−0.426	−0.399
5775	−0.261	−0.143	−0.133	−0.131	−0.111
6566		−0.088	−0.061	−0.004	
Tuning parameters	$\xi_n = 43$	$\lambda = 0.197$	$\gamma = 3.7$,	$\lambda = 0.260$	$\gamma = 3.598$,
	$\lambda_n = 5.721$		$\lambda = 0.211$		$\lambda = 2.058$
Number of selected	3	8	4	4	3
CV error	6.399	6.731	6.496	6.515	6.472

$p = 20,037$ methylation variables. Applying the method described in Sect. 2.5, we first performed sure joint screening using the BJASS method of Liu et al. (2020) reduce data dimension to $k = 2log(n)n^{\frac{1}{4}} = 34$ before applying the CBAR penalization method and four other penalization methods (Lasso, SCAD, MCP and Alasso). The final variable selection results are summarized in Table 4.

It is seen from Table 4 that our BJASS-CBAR selected the sparsest model with 4 variables while achieving a comparable CV error as compared to the other four methods, which is consistent with our findings in simulation studies. It is interesting to note that the four features selected by BJASS-CBAR have also been selected by three other methods. Among the four selected features, NPM2 and IRX6 have been previously discussed in the literature to possibly play critical roles with human diseases (Eirín-López et al. 2006; Box et al. 2016; Nachmani et al. 2019; Mummenhoff et al. 2001).

5 Discussion

We have rigorously extended the broken adaptive ridge (BAR) penalization method for simultaneous variable selection and parameter estimation to the semiparametric AFT model with right-censored data by coupling BAR penalization with the Leurgans (1987) synthetic data. We have established that the resulting CBAR estimator is asymptotically consistency for variable selection and have an oracle estimation property and enjoy a grouping property for highly correlated covariates. We consider both low- and high-dimensional covariate settings. Our empirical studies

Table 4 Estimated coefficients of BJASS-CBAR, BJASS-Lasso, BJASS-SCAD, BJASS-MCP and BJASS-Alasso for the TCGA GBM methylation data

Variables	BJASS-CBAR	BJASS-Lasso	BJASS-SCAD	BJASS-MCP	BJASS-Alasso
BCL2L10		0.051	0.038		0.038
CDCP2	− 0.272	− 0.077	− 0.057		− 0.068
HES5		− 0.139	− 0.153	− 0.265	− 0.162
HLA.E		0.104	0.117	0.167	0.098
HRH3		0.021			
IRX6		0.014			
KIF5C		0.004			
NIPSNAP3B		0.034			0.017
NPM2	0.230	0.087	0.065	0.089	0.078
OXGR1		0.059	0.066		0.045
SLC12A5	0.282	0.144	0.104	0.072	0.167
SMIM11A	0.417	0.349	0.469	0.507	0.418
Tuning parameters	$\xi_n = 19$	$\lambda = 0.122$	$\gamma = 3.7$,	$\lambda = 0.190$	$\lambda = 0.625$
	$\lambda_n = 1.642$		$\lambda = 0.154$		
Number of selected	4	12	9	5	9
CV error	3.793	3.832	3.804	3.835	3.620

demonstrate that CBAR generally produces a more sparse and accurate model as compared to some popular L_1-based penalization methods, which corroborates previous findings in the literature for uncensored data.

We note that coupling the BAR method with the Leurgans (1987) synthetic variable is only one of several possible ways of extending the BAR method to right-censored linear model for simultaneous variable selection and parameter estimation. For example, one may couple the BAR method with the Koul et al. (1981) synthetic data method, the Stute (1993) weighted least squares method, or the Buckley and James (1979) iterative imputation method. Our limited numerical studies (not reported here) indicate that using Koul et al. (1981) synthetic data is generally inferior to using Leurgans (1987) synthetic variable, whereas iteratively performing BAR using the Buckley and James (1979) imputation may sometimes improve the performance of the CBAR method based on the Leurgans (1987) synthetic variable. However, asymptotic properties of each of these distinct approaches require different theoretical developments. Thorough investigations and comparisons of these alternative approaches are needed in future research.

Lastly, missing data often occur in real-world applications. Although there is a vast amount literature on missing data problems, little has been done to deal with missing data in the context of variable selection for survival data. Further research in this domain is waranteed.

Appendix: Proofs of the theorems

We first introduce notations and lemmas used to prove Theorem 1.

Using Leurgans (1987) method, we transform $\mathbf{Y}$ into synthetic data $\mathbf{Y}^*$. Let $\boldsymbol{\beta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\gamma}^\top)^\top$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are $q_n \times 1$ and $(p_n - q_n) \times 1$ vector, respectively, $\boldsymbol{\Sigma}_n = \mathbf{x}^\top \mathbf{x}/n$.

$$g(\boldsymbol{\beta}) = \{\mathbf{x}^\top \mathbf{x} + \lambda_n \mathbf{D}(\boldsymbol{\beta})\}^{-1} \mathbf{x}^\top \mathbf{Y}^* = (\boldsymbol{\alpha}^*(\boldsymbol{\beta})^\top, \boldsymbol{\gamma}^*(\boldsymbol{\beta})^\top)^\top. \tag{10}$$

For simplicity, we write $\boldsymbol{\alpha}^*(\boldsymbol{\beta})$ and $\boldsymbol{\gamma}^*(\boldsymbol{\beta})$ as $\boldsymbol{\alpha}^*$ and $\boldsymbol{\gamma}^*$ hereafter. $\boldsymbol{\Sigma}_n^{-1}$ can be partitioned as

$$\boldsymbol{\Sigma}_n^{-1} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^\top & \mathbf{A}_{22} \end{pmatrix}$$

where the A_{11} is a $q \times q$ matrix. Multiplying $(\mathbf{x}^\top \mathbf{x})^{-1}(\mathbf{x}^\top \mathbf{x} + \lambda_n \mathbf{D}(\boldsymbol{\beta}))$ to equation (10)

$$\begin{pmatrix} \boldsymbol{\alpha}^* - \boldsymbol{\beta}_{01} \\ \boldsymbol{\gamma}^* \end{pmatrix} + \frac{\lambda_n}{n} \begin{pmatrix} \mathbf{A}_{11}\mathbf{D}_1(\boldsymbol{\alpha})\boldsymbol{\alpha}^* + \mathbf{A}_{12}\mathbf{D}_2(\boldsymbol{\gamma})\boldsymbol{\gamma}^* \\ \mathbf{A}_{12}^\top \mathbf{D}_1(\boldsymbol{\alpha})\boldsymbol{\alpha}^* + \mathbf{A}_{22}\mathbf{D}_2(\boldsymbol{\gamma})\boldsymbol{\gamma}^* \end{pmatrix} = (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \boldsymbol{\varepsilon}^* = \hat{\boldsymbol{\beta}}_Z - \boldsymbol{\beta}_0, \tag{11}$$

where $\boldsymbol{\varepsilon}^* = \mathbf{Y}^* - \mathbf{x}\boldsymbol{\beta}_0$, $\hat{\boldsymbol{\beta}}_Z = (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \mathbf{Y}^*$, $\mathbf{D}_1(\boldsymbol{\alpha}) = \text{diag}\,(\alpha_1^{-2}, ..., \alpha_q^{-2})$ and $\mathbf{D}_2(\boldsymbol{\gamma}) = \text{diag}\,(\gamma_1^{-2}, ..., \gamma_{p_n - q}^{-2})$.

Lemma 1 *Let δ_n be a sequence of positive real numbers satisfying $\delta_n \to \infty$ and $p_n\delta_n^2/\lambda_n \to 0$. Define $\mathbf{H}_n = \{\boldsymbol{\beta} \in \mathbb{R}^{p_n} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le \delta_n\sqrt{p_n/n}\}$ and $\mathbf{H}_{n1} = \{\boldsymbol{\alpha} \in \mathbb{R}^q : \|\boldsymbol{\alpha} - \boldsymbol{\beta}_{01}\| \le \delta_n\sqrt{p_n/n}\}$. Assume conditions (C1)–(C5) hold. Then, with probability tending to* 1, *we have*

(a) $\sup_{\boldsymbol{\beta}\in\mathbf{H}_n} \|\boldsymbol{\gamma}^*\|/\|\boldsymbol{\gamma}\| < 1/C_0$, for some constant $C_0 > 1$;
(b) *g is a mapping from $\mathbf{H}_n$ to itself.*

Proof We first prove part (a).

First, under $\lambda_n/\sqrt{n} \to 0$ and $p_n\delta_n^2/\lambda_n \to 0$, we have $\delta_n\sqrt{p_n/n} \to 0$.

Let $\hat{\boldsymbol{\beta}}_Z = (\mathbf{x}^\top\mathbf{x})^{-1}\mathbf{x}^\top\mathbf{Y}^*$, $\omega_{ji} = ((\mathbf{x}^\top\mathbf{x})^{-1}\mathbf{x}^\top)_{ji}$, $\mu_j^* = \sum_i \omega_{ji}\int_0^{T_n} F_i \mathrm{d}t$ and $\boldsymbol{\mu} = (\mu_1^*, \mu_2^*, ..., \mu_{pn}^*)$. For any p_n-vector $\mathbf{b}_n$ which $\|\mathbf{b}_n\| \le 1$, define $t_n^2 = \mathbf{b}_n^\top\boldsymbol{\Omega}(\infty))\mathbf{b}_n$. Then, we have $\sqrt{n}\, t_n^{-1}\mathbf{b}_n^\top(\hat{\boldsymbol{\beta}}_Z - \boldsymbol{\omega}) \to_D N(0,1)$. This result can be proved using similar techniques to those used in the proof of Theorem 3.1 of Zhou (1992) along the same lines as outlined below: First, we separate $\mathbf{b}_n^\top(\hat{\boldsymbol{\beta}}_Z - \boldsymbol{\omega})$ like (3.6) in Zhou (1992) with a main term $S_\beta(T^n)$ and a remainder term $SS_\beta(T^n)$, i.e., $\mathbf{b}_n^\top(\hat{\boldsymbol{\beta}}_Z - \boldsymbol{\omega}) = S_\beta(T^n) + SS_\beta(T^n)$, where $S_\beta(T^n)$ is a weighted sum of $\hat{H}(t) - H(t)$ and $\hat{G}(t) - G(t)$; and $SS_\beta(T^n)$ is a weighted sum of $(\hat{H}(t) - H(t))(\hat{G}(t) - G(t))$ and $(\hat{H}(t) - H(t))(\hat{H}(t) - H(t))$. Second, under conditions (C2) and (C3), one can show that $\sqrt{n}SS_\beta(T^n)$ is negligible. Finally, by applying the martingale central limit theorem and conditions (C1) and (C4), we establish the asymptotic normality of $\sqrt{n}S_\beta(T^n)$. By conditions (C1) and (C2), we have $\sqrt{n}t_n^{-1}\mathbf{b}_n^\top(\boldsymbol{\beta}_0 - \boldsymbol{\omega}) = o_p(1)$, for $\mathbf{b}_n = \mathbf{e}_i = (0, ..., 1, 0, ..., 0)$. Hence, we have $\|\hat{\boldsymbol{\beta}}_Z - \boldsymbol{\beta}_0\|^2 = O_p(p_n/n)$.

It then follows from (11) that

$$\sup_{\boldsymbol{\beta}\in\mathbf{H}_n} \left\|\boldsymbol{\gamma}^* + \lambda_n\mathbf{A}_{12}^\top\mathbf{D}_1(\boldsymbol{\alpha})\boldsymbol{\alpha}^*/n + \lambda_n\mathbf{A}_{22}\mathbf{D}_2(\boldsymbol{\gamma})\boldsymbol{\gamma}^*/n\right\| = O_p(\sqrt{p_n/n}). \tag{12}$$

Note that $\|\boldsymbol{\alpha} - \boldsymbol{\beta}_{01}\| \le \delta_n(p_n/n)^{1/2}$ and $\|\boldsymbol{\alpha}^*\| \le \|g(\boldsymbol{\beta})\| \le \|\hat{\boldsymbol{\beta}}_Z\| = O_p(\sqrt{p_n})$. By assumptions (C4) and (C5), we have

$$\begin{aligned}\sup_{\boldsymbol{\beta}\in\mathbf{H}_n} \left\|\lambda_n\mathbf{A}_{12}^\top\mathbf{D}_1(\boldsymbol{\alpha})\boldsymbol{\alpha}^*/n\right\| &\le \frac{\lambda_n}{n}\|\mathbf{A}_{12}^\top\| \sup_{\boldsymbol{\beta}\in\mathbf{H}_n}\|\mathbf{D}_1(\boldsymbol{\alpha})\boldsymbol{\alpha}^*\| \\ &\le \sqrt{2}\,\tilde{C}\,\frac{\lambda_n}{n}\,\frac{a_1}{a_0^2} \sup_{\boldsymbol{\beta}\in\mathbf{H}_n}\|\boldsymbol{\alpha}^*\| = o_p(\sqrt{p_n/n}),\end{aligned} \tag{13}$$

where the second inequality uses the fact $\|\mathbf{A}_{12}^\top\| \le \sqrt{2}\,\tilde{C}$, which follows from the inequality $\|\mathbf{A}_{12}\mathbf{A}_{12}^\top\| - \|\mathbf{A}_{11}^2\| \le \|\mathbf{A}_{11}^2 + \mathbf{A}_{12}\mathbf{A}_{21}\| \le \|\boldsymbol{\Sigma}_n^{-2}\| < \tilde{C}^2$. Combining (12) and (13) gives

$$\sup_{\boldsymbol{\beta}\in\mathbf{H}_n} \left\|\boldsymbol{\gamma}^* + \lambda_n\mathbf{A}_{22}\mathbf{D}_2(\boldsymbol{\gamma})\boldsymbol{\gamma}^*/n\right\| = O_p(\sqrt{p_n/n}). \tag{14}$$

Note that $\mathbf{A}_{22} = \sum_{i=1}^{p_n-q}\tau_{2i}\mathbf{u}_{2i}\mathbf{u}_{2i}^\top$ is positive definite and by the singular value decomposition, , where τ_{2i} and $\mathbf{u}_{2i}$ are eigenvalues and eigenvectors of $\mathbf{A}_{22}$. Then, since $1/\tilde{C} < \tau_{2i} < \tilde{C}$, we have

$$\frac{\lambda_n}{n}\|\mathbf{A}_{22}\mathbf{D}_2(\boldsymbol{\gamma})\boldsymbol{\gamma}^*\| = \frac{\lambda_n}{n}\left\|\sum_{i=1}^{p_n-q}\tau_{2i}\mathbf{u}_{2i}\mathbf{u}_{2i}^\top\mathbf{D}_2(\boldsymbol{\gamma})\boldsymbol{\gamma}^*\right\| = \frac{\lambda_n}{n}\left\{\sum_{i=1}^{p_n-q}\tau_{2i}^2\|\mathbf{u}_{2i}^\top\mathbf{D}_2(\boldsymbol{\gamma})\boldsymbol{\gamma}^*\|^2\right\}^{1/2}$$
$$\geq \frac{\lambda_n}{n}\frac{1}{\tilde{C}}\left\{\sum_{i=1}^{p_n-q}\|\mathbf{u}_{2i}^\top\mathbf{D}_2(\boldsymbol{\gamma})\boldsymbol{\gamma}^*\|^2\right\}^{1/2} = \frac{1}{\tilde{C}}\|\lambda_n\mathbf{D}_2(\boldsymbol{\gamma})\boldsymbol{\gamma}^*/n\|.$$

This, together with (14) and (C4), implies that with probability tending to 1,

$$\frac{1}{\tilde{C}}\|\lambda_n\mathbf{D}_2(\boldsymbol{\gamma})\boldsymbol{\gamma}^*/n\| - \|\boldsymbol{\gamma}^*\| \leq \delta_n\sqrt{p_n/n}. \tag{15}$$

Let $\mathbf{D}_{\gamma*/\gamma} = (\gamma_1^*/\gamma_1, \ldots, \gamma_{p_n-q}^*/\gamma_{p_n-q})^\top$. Because $\|\boldsymbol{\gamma}\| \leq \delta_n\sqrt{p_n/n}$, we have

$$\frac{1}{\tilde{C}}\left\|\frac{\lambda_n}{n}\mathbf{D}_2(\boldsymbol{\gamma})\boldsymbol{\gamma}^*\right\| = \frac{1}{\tilde{C}}\frac{\lambda_n}{n}\left\|\{\mathbf{D}_2(\boldsymbol{\gamma})\}^{1/2}\mathbf{D}_{\gamma*/\gamma}\right\| \geq \frac{1}{\tilde{C}}\frac{\lambda_n}{n}\frac{\sqrt{n}}{\delta_n\sqrt{p_n}}\|\mathbf{D}_{\gamma*/\gamma}\| \tag{16}$$

and

$$\|\boldsymbol{\gamma}^*\| = \|\mathbf{D}_2(\boldsymbol{\gamma})^{-1/2}\mathbf{D}_{\gamma*/\gamma}\| \leq \frac{\delta_n\sqrt{p_n}}{\sqrt{n}}\|\mathbf{D}_{\gamma*/\gamma}\|. \tag{17}$$

Combining (15), (16) and (17), we have that with probability tending to 1,

$$\|\mathbf{D}_{\gamma*/\gamma}\| \leq \frac{1}{\lambda_n/(p_n\delta_n^2\tilde{C}) - 1} < 1/C_0 \tag{18}$$

for some constant $C_0 > 1$ provided that $\lambda_n/(p_n\delta_n^2) \to \infty$.

It is worth noting that $\Pr(\|\mathbf{D}_{\gamma*/\gamma}\| \to 0) \to 1$, as $n \to \infty$. Furthermore, with probability tending to 1,

$$\|\boldsymbol{\gamma}^*\| \leq \|\mathbf{D}_{\gamma^*/\gamma}\| \max_{1\leq j\leq(p_n-q)}|\gamma_j| \leq \|\mathbf{D}_{\gamma^*/\gamma}\| \times \|\boldsymbol{\gamma}\| \leq \|\boldsymbol{\gamma}\|/C_0.$$

This proves part (a).

Next we prove part (b). First, it is easy to see from (17) and (18) that, as $n \to \infty$,

$$\Pr\left(\|\boldsymbol{\gamma}^*\| \leq \delta_n\sqrt{p_n/n}\right) \to 1. \tag{19}$$

Then, by (11), we have

$$\sup_{\boldsymbol{\beta}\in\mathbf{H}_n}\left\|\boldsymbol{\alpha}^* - \boldsymbol{\beta}_{01} + \lambda_n\mathbf{A}_{11}\mathbf{D}_1(\boldsymbol{\alpha})\boldsymbol{\alpha}^*/n + \lambda_n\mathbf{A}_{12}\mathbf{D}_2(\boldsymbol{\gamma})\boldsymbol{\gamma}^*/n\right\| = O_p(\sqrt{p_n/n}). \tag{20}$$

Similar to (13), it is easily to verify that

$$\sup_{\boldsymbol{\beta}\in\mathbf{H}_n}\left\|\lambda_n\mathbf{A}_{11}\mathbf{D}_1(\boldsymbol{\alpha})\boldsymbol{\alpha}^*/n\right\| = o_p(\sqrt{p_n/n}). \tag{21}$$

Moreover, with probability tending to 1,

$$\sup_{\boldsymbol{\beta}\in\mathbf{H}_n}\left\|\lambda_n\mathbf{A}_{12}\mathbf{D}_2(\boldsymbol{\gamma})\boldsymbol{\gamma}^*/n\right\| \le \frac{\lambda_n}{n}\sup_{\boldsymbol{\beta}\in\mathbf{H}_n}\left\|\mathbf{D}_2(\boldsymbol{\gamma})\boldsymbol{\gamma}^*\right\| \times \|\mathbf{A}_{12}\| \le 2\sqrt{2}\tilde{C}^2\delta_n\sqrt{p_n/n}, \tag{22}$$

where the last step follows from (15), (19), and the fact that $\|\mathbf{A}_{12}\| \le \sqrt{2}\tilde{C}$. It follows from (20), (21) and (22) that with probability tending to 1,

$$\sup_{\boldsymbol{\beta}\in\mathbf{H}_n}\|\boldsymbol{\alpha}^* - \boldsymbol{\beta}_{01}\| \le \left(2\sqrt{2}\tilde{C}^2 + 1\right)\delta_n n^{-1/2}\sqrt{p_n}. \tag{23}$$

Because $\delta_n\sqrt{p_n}/\sqrt{n} \to 0$, we have, as $n \to \infty$,

$$\Pr(\boldsymbol{\alpha}^* \in \mathbf{H}_{n1}) \to 1. \tag{24}$$

Combining (19) and (24) completes the proof of part (b). □

Lemma 2 *Assume that (C1)–(C5) hold. For any q-vector* $\mathbf{c}$ *satisfying* $\|\mathbf{c}\| \le 1$, *define* $z^2 = \mathbf{c}^\top\boldsymbol{\Omega}_1\mathbf{c}$ *as in Theorem 1. Define*

$$f(\boldsymbol{\alpha}) = \{\mathbf{x}_1^\top\mathbf{x}_1 + \lambda_n\mathbf{D}_1(\boldsymbol{\alpha})\}^{-1}\mathbf{x}_1^\top\mathbf{Y}^*. \tag{25}$$

Then, with probability tending to 1,

(a) $f(\boldsymbol{\alpha})$ *is a contraction mapping from* $\mathbf{b}_n \equiv \{\boldsymbol{\alpha} \in \mathbb{R}^q : \|\boldsymbol{\alpha} - \boldsymbol{\beta}_{01}\| \le \delta_n\sqrt{p_n/n}\}$ *to itself*;

(b) $\sqrt{n}\,z^{-1}\mathbf{c}^\top(\hat{\boldsymbol{\alpha}}^\circ - \boldsymbol{\beta}_{01}) \rightsquigarrow \mathcal{N}(0,1)$, *where* $\hat{\boldsymbol{\alpha}}^\circ$ *is the unique fixed point of* $f(\boldsymbol{\alpha})$ *defined by*

$$\hat{\boldsymbol{\alpha}}^\circ = \{\mathbf{x}_1^\top\mathbf{x}_1 + \lambda_n\mathbf{D}_1(\hat{\boldsymbol{\alpha}}^\circ)\}^{-1}\mathbf{x}_1^\top\mathbf{Y}^*.$$

Proof We first prove part (a). Note that (25) can be rewritten as

$$f(\boldsymbol{\alpha}) - \boldsymbol{\beta}_{01} + \frac{\lambda_n}{n}\boldsymbol{\Sigma}_{n1}^{-1}\mathbf{D}_1(\boldsymbol{\alpha})f(\boldsymbol{\alpha}) = \hat{\boldsymbol{\beta}}_{1Z} - \boldsymbol{\beta}_{01},$$

where $\hat{\boldsymbol{\beta}}_{1Z} = (\mathbf{x}_1^\top\mathbf{x}_1)^{-1}\mathbf{x}_1^\top\mathbf{Y}^*$. Then,

$$\sup_{\boldsymbol{\alpha}\in\mathbf{b}_n}\left\|f(\boldsymbol{\alpha}) - \boldsymbol{\beta}_{01} + (\lambda_n/n)\boldsymbol{\Sigma}_{n1}^{-1}\mathbf{D}_1(\boldsymbol{\alpha})f(\boldsymbol{\alpha})\right\| = O_p(1/\sqrt{n}). \tag{26}$$

$$\sup_{\boldsymbol{\alpha}\in\mathbf{b}_n}\left\|(\lambda_n/n)\boldsymbol{\Sigma}_{n1}^{-1}\mathbf{D}_1(\boldsymbol{\alpha})f(\boldsymbol{\alpha})\right\| = o_p(1/\sqrt{n}). \tag{27}$$

It follows from (26) and (27) that

$$\sup_{\boldsymbol{\alpha}\in\mathbf{b}_n}\|f(\boldsymbol{\alpha}) - \boldsymbol{\beta}_{01}\| \le \delta_n/\sqrt{n}, \tag{28}$$

where $\delta_n \to \infty$ and $\delta_n/\sqrt{n} \to 0$. Then we can get

$$\Pr\left(f(\boldsymbol{\alpha}) \in \mathbf{b}_n\right) \to 1, \text{ as } n \to \infty. \tag{29}$$

This means that f is a mapping from the region $\mathbf{b}_n$ to itself.

Rewrite (25) as $\{\mathbf{x}_1^\top \mathbf{x}_1 + \lambda_n \mathbf{D}_1(\boldsymbol{\alpha})\} f(\boldsymbol{\alpha}) = \mathbf{x}_1^\top \mathbf{Y}^*$, then, we have

$$(\boldsymbol{\Sigma}_{n1} + (\lambda_n/n)\mathbf{D}_1(\boldsymbol{\alpha}))\dot{f}(\boldsymbol{\alpha}) + (\lambda_n/n)\ \text{diag}\ \{-2f_j(\boldsymbol{\alpha})/\alpha_j^3\} = 0, \tag{30}$$

where $\dot{f}(\boldsymbol{\alpha}) = \partial f(\boldsymbol{\alpha})/\partial \boldsymbol{\alpha}^\top$ and $\text{diag}\ \{\frac{-2f_j(\boldsymbol{\alpha})}{\alpha_j^3}\} = \text{diag}\ \{\frac{-2f_1(\boldsymbol{\alpha})}{\alpha_1^3}, ..., \frac{-2f_q(\boldsymbol{\alpha})}{\alpha_q^3}\}$. With the assumption $\lambda_n/\sqrt{n} \to 0$,

$$\sup_{\boldsymbol{\alpha}\in\mathbf{b}_n} \left\| \left\{ \boldsymbol{\Sigma}_{n1} + \frac{\lambda_n}{n}\mathbf{D}_1(\boldsymbol{\alpha}) \right\} \dot{f}(\boldsymbol{\alpha}) \right\| = \sup_{\boldsymbol{\alpha}\in\mathbf{b}_n} \frac{2\lambda_n}{n} \left\| \text{diag} \left\{ \frac{f_j(\boldsymbol{\alpha})}{\alpha_j^3} \right\} \right\| = o_p(1). \tag{31}$$

Write $\boldsymbol{\Sigma}_{n1} = \sum_{i=1}^q \tau_{1i}\mathbf{u}_{1i}\mathbf{u}_{1i}^\top$, where τ_{1i} and $\mathbf{u}_{1i}$ are eigenvalues and eigenvectors of $\boldsymbol{\Sigma}_{n1}$. Then, by (C4), $1/\tilde{C} < \tau_{1i} < \tilde{C}$ for all i and

$$\begin{aligned}
\|\boldsymbol{\Sigma}_{n1}\dot{f}(\boldsymbol{\alpha})\| &= \sup_{\|\mathbf{x}\|=1, \mathbf{x}\in R^q} \|\boldsymbol{\Sigma}_{n1}\dot{f}(\boldsymbol{\alpha})\mathbf{x}\| = \sup_{\|\mathbf{x}\|=1, \mathbf{x}\in R^q} \left\| \sum_{i=1}^q \lambda_{1i}\mathbf{u}_{1i}\mathbf{u}_{1i}^\top \dot{f}(\boldsymbol{\alpha})\mathbf{x} \right\| \\
&= \sup_{\|\mathbf{x}\|=1, \mathbf{x}\in R^q} \left(\sum_{i=1}^q \lambda_{1i}^2 \|\mathbf{u}_{1i}^\top \dot{f}(\boldsymbol{\alpha})\mathbf{x}\|^2 \right)^{1/2} \\
&\geq \sup_{\|\mathbf{x}\|=1, \mathbf{x}\in R^q} \frac{1}{\tilde{C}} \left(\sum_{i=1}^q \|\mathbf{u}_{1i}^\top \dot{f}(\boldsymbol{\alpha})\mathbf{x}\|^2 \right)^{1/2} \\
&= \sup_{\|\mathbf{x}\|=1, \mathbf{x}\in R^q} \frac{1}{\tilde{C}} \|\dot{f}(\boldsymbol{\alpha})\mathbf{x}\| = \frac{1}{\tilde{C}} \|\dot{f}(\boldsymbol{\alpha})\|.
\end{aligned} \tag{32}$$

Therefore, it follows from $\boldsymbol{\alpha} \in \mathbf{b}_n$, (32) and (C4) that

$$\begin{aligned}
\left\| \left\{ \boldsymbol{\Sigma}_{n1} + (\lambda_n/n)\mathbf{D}_1(\boldsymbol{\alpha}) \right\} \dot{f}(\boldsymbol{\alpha}) \right\| &\geq \left\| \boldsymbol{\Sigma}_{n1}\dot{f}(\boldsymbol{\alpha}) \right\| - \left\| (\lambda_n/n)\mathbf{D}_1(\boldsymbol{\alpha})\dot{f}(\boldsymbol{\alpha}) \right\| \\
&\geq (1/\tilde{C})\|\dot{f}(\boldsymbol{\alpha})\| - (\lambda_n/n) \cdot a_0^{-2}\|\dot{f}(\boldsymbol{\alpha})\|.
\end{aligned}$$

This, together with (31) and the fact $\lambda_n/n \to 0$, implies that

$$\sup_{\boldsymbol{\alpha}\in\mathbf{b}_n} \|\dot{f}(\boldsymbol{\alpha})\| = o_p(1). \tag{33}$$

Finally, we can get the conclusion in part (a) from (29) and (33).

Next we prove part (b). Write

$$\begin{aligned}
n^{1/2} z^{-1}\mathbf{c}^\top(\hat{\boldsymbol{\alpha}}^\circ - \boldsymbol{\beta}_{01}) &= n^{1/2} z^{-1}\mathbf{c}^\top \left[\left\{ \boldsymbol{\Sigma}_{n1} + \frac{\lambda_n}{n}\mathbf{D}_1(\hat{\boldsymbol{\alpha}}^\circ) \right\}^{-1} \boldsymbol{\Sigma}_{n1} - \mathbf{I}_{q_n} \right] \boldsymbol{\beta}_{01} \\
&\quad + n^{-1/2} z^{-1}\mathbf{c}^\top \left\{ \boldsymbol{\Sigma}_{n1} + \frac{\lambda_n}{n}\mathbf{D}_1(\hat{\boldsymbol{\alpha}}^\circ) \right\}^{-1} \mathbf{x}_1^\top \boldsymbol{\varepsilon}^* \equiv I_1 + I_2.
\end{aligned} \tag{34}$$

By the first order resolvent expansion formula

$$(\mathbf{H}+\boldsymbol{\Delta})^{-1} = \mathbf{H}^{-1} - \mathbf{H}^{-1}\boldsymbol{\Delta}(\mathbf{H}+\boldsymbol{\Delta})^{-1},$$

the first term on the right-hand side of equation (34) can be rewritten as

$$I_1 = -z^{-1}\mathbf{c}^{\top}\boldsymbol{\Sigma}_{n1}^{-1}\frac{\lambda_n}{\sqrt{n}}\mathbf{D}_1(\hat{\boldsymbol{\alpha}}^{\circ})\left\{\boldsymbol{\Sigma}_{n1} + \frac{\lambda_n}{n}\mathbf{D}_1(\hat{\boldsymbol{\alpha}}^{\circ})\right\}^{-1}\boldsymbol{\Sigma}_{n1}\boldsymbol{\beta}_{01}.$$

Hence, by the assumption (C4) and (C5), we have

$$\|I_1\| \le \frac{\lambda_n}{\sqrt{n}}z^{-1}a_0^{-2}\|\boldsymbol{\Sigma}_{n1}^{-1}\boldsymbol{\beta}_{01}\| = O_p\left(\lambda_n/\sqrt{n}\right) \to 0. \tag{35}$$

Furthermore, applying the first order resolvent expansion formula, it can be shown that

$$\begin{aligned} I_2 &= \frac{z^{-1}}{\sqrt{n}}\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_{n1}^{-1}\mathbf{x}_1^{\mathrm{T}}\boldsymbol{\varepsilon}^* + o_p(1) \\ &= \frac{z^{-1}}{\sqrt{n}}\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_{n1}^{-1}\mathbf{x}_1^{\mathrm{T}}(\mathbf{Y}^* - \mathbf{x}_1\boldsymbol{\omega} + \mathbf{x}_1\boldsymbol{\omega} - \mathbf{x}_1\boldsymbol{\beta}_{01}) + o_p(1) \\ &= \sqrt{n}z^{-1}\mathbf{c}^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{1Z} - \boldsymbol{\omega}_1 + \boldsymbol{\omega}_1 - \boldsymbol{\beta}_{01}) + o_p(1) \end{aligned} \tag{36}$$

where $\boldsymbol{\mu}_1 = (\mu_1^*, \mu_2^*, ..., \mu_q^*)$. I_2 converges in distribution to $N(0, 1)$ by the Lindeberg-Feller central limit theorem. Finally, combining (34), (35), and (36) proves part (b). □

Proof of Theorem 1 Given the initial ridge estimator $\hat{\boldsymbol{\beta}}^{(0)}$ in (4), we have

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0 &= \left[\left(\boldsymbol{\Sigma}_n + \frac{\xi_n}{n}\mathbf{I}_{p_n}\right)^{-1}\boldsymbol{\Sigma}_n - \mathbf{I}_{p_n}\right]\boldsymbol{\beta}_0 + \left(\boldsymbol{\Sigma}_n + \frac{\xi_n}{n}\mathbf{I}_{p_n}\right)^{-1}\mathbf{x}^{\top}\boldsymbol{\varepsilon}^*/n \\ &\equiv \mathbf{T}_1 + \mathbf{T}_2. \end{aligned} \tag{37}$$

By the first-order resolvent expansion formula and $\xi_n/\sqrt{n} \to 0$,

$$\|\mathbf{T}_1\| = \left\|-\boldsymbol{\Sigma}_n^{-1}\frac{\xi_n}{n}\left(\boldsymbol{\Sigma}_n + \frac{\xi_n}{n}\mathbf{I}_{p_n}\right)^{-1}\boldsymbol{\Sigma}_n\boldsymbol{\beta}_0\right\| \le \tilde{C}^3\frac{\xi_n a_1\sqrt{p_n}}{n} = o_p\left(\sqrt{\frac{p_n}{n}}\right). \tag{38}$$

It is easy to see that $\|\mathbf{T}_2\| = O_p(\sqrt{p_n/n})$. Thus $\|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0\| = O_p((p_n/n)^{1/2})$. This, combined with part (a) of Lemma 1, implies that

$$\Pr\left(\lim_{k\to\infty}\hat{\boldsymbol{\gamma}}^{(k)} = 0\right) \to 1. \tag{39}$$

Hence, to prove part (i) of Theorem 1, it is sufficient to show that

$$\Pr\left(\lim_{k\to\infty}\|\hat{\boldsymbol{\alpha}}^{(k)}-\hat{\boldsymbol{\alpha}}^{\circ}\|=0\right)\to 1, \tag{40}$$

where $\hat{\boldsymbol{\alpha}}^{\circ}$ is the fixed point of $f(\boldsymbol{\alpha})$ defined in part (b) of Lemma 2.

Define $\boldsymbol{\gamma}^{*}=0$ if $\boldsymbol{\gamma}=0$, for any $\boldsymbol{\alpha}\in\mathbf{b}_n$,

$$\lim_{\boldsymbol{\gamma}\to 0}\boldsymbol{\gamma}^{*}(\boldsymbol{\alpha},\boldsymbol{\gamma})=0. \tag{41}$$

Combining (41) with the fact

$$\begin{pmatrix}\mathbf{x}_1^{\top}\mathbf{x}_1+\lambda_n\mathbf{D}_1(\boldsymbol{\alpha}) & \mathbf{x}_1^{\top}\mathbf{X}_2\\ \mathbf{x}_2^{\top}\mathbf{x}_1 & \mathbf{x}_2^{\top}\mathbf{x}_2+\lambda_n\mathbf{D}_2(\boldsymbol{\gamma})\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}^{*}\\ \boldsymbol{\gamma}^{*}\end{pmatrix}=\begin{pmatrix}\mathbf{x}_1^{\top}\mathbf{Y}^{*}\\ \mathbf{x}_2^{\top}\mathbf{Y}^{*}\end{pmatrix},$$

implies that for any $\boldsymbol{\alpha}\in\mathbf{b}_n$,

$$\lim_{\boldsymbol{\gamma}\to 0}\boldsymbol{\alpha}^{*}(\boldsymbol{\alpha},\boldsymbol{\gamma})=\{\mathbf{x}_1^{\mathrm{T}}\mathbf{x}_1+\lambda_n\mathbf{D}_1(\boldsymbol{\alpha})\}^{-1}\mathbf{x}_1\mathbf{Y}^{*}=f(\boldsymbol{\alpha}). \tag{42}$$

Therefore, $g(\cdot)$ is continuous and thus uniformly continuous on the compact set $\boldsymbol{\beta}\in\mathbf{H}_n$. This, together with (39) and (42), implies that as $k\to\infty$,

$$\eta_k\equiv\sup_{\boldsymbol{\alpha}\in\mathbf{b}_n}\left\|f(\boldsymbol{\alpha})-\boldsymbol{\alpha}^{*}(\boldsymbol{\alpha},\hat{\boldsymbol{\gamma}}^{(k)})\right\|\longrightarrow 0, \tag{43}$$

with probability tending to 1.

Note that

$$\begin{aligned}\|\hat{\boldsymbol{\alpha}}^{(k+1)}-\hat{\boldsymbol{\alpha}}^{\circ}\|=&\left\|\boldsymbol{\alpha}^{*}(\hat{\boldsymbol{\beta}}^{(k)})-\hat{\boldsymbol{\alpha}}^{\circ}\right\|\le\left\|\boldsymbol{\alpha}^{*}(\hat{\boldsymbol{\beta}}^{(k)})-f(\hat{\boldsymbol{\alpha}}^{(k)})\right\|+\|f(\hat{\boldsymbol{\alpha}}^{(k)})-\hat{\boldsymbol{\alpha}}^{\circ}\|\\ \le&\eta_k+\frac{1}{\tilde{C}}\|\hat{\boldsymbol{\alpha}}^{(k)}-\hat{\boldsymbol{\alpha}}^{\circ}\|,\end{aligned} \tag{44}$$

where the last step follows from $\|f(\hat{\boldsymbol{\alpha}}^{(k)})-\hat{\boldsymbol{\alpha}}^{\circ}\|=\|f(\hat{\boldsymbol{\alpha}}^{(k)})-f(\hat{\boldsymbol{\alpha}}^{\circ})\|\le(1/\tilde{C})\|\hat{\boldsymbol{\alpha}}^{(k)}-\hat{\boldsymbol{\alpha}}^{\circ}\|$. Let $a_k=\|\hat{\boldsymbol{\alpha}}^{(k)}-\hat{\boldsymbol{\alpha}}^{\circ}\|$, for all $k\ge 0$. From (43), we can induce that with probability tending to 1, for any $\epsilon>0$, there exists an positive integer N such that for all $k>N$, $|\eta_k|<\epsilon$ and

$$\begin{aligned}a_{k+1}\le&\frac{a_{k-1}}{\tilde{C}^2}+\frac{\eta_{k-1}}{\tilde{C}}+\eta_k\\ \le&\frac{a_1}{\tilde{C}^k}+\frac{\eta_1}{\tilde{C}^{k-1}}+\cdots+\frac{\eta_N}{\tilde{C}^{k-N}}+\left(\frac{\eta_{N+1}}{\tilde{C}^{k-N-1}}+\cdots+\frac{\eta_{k-1}}{\tilde{C}}+\eta_k\right)\\ \le&(a_1+\eta_1+\ldots+\eta_N)\frac{1}{\tilde{C}^{k-N}}+\frac{1-(1/\tilde{C})^{k-N}}{1-1/\tilde{C}}\epsilon\to 0,\ \text{as } k\to\infty.\end{aligned}$$

This proves (40).

Therefore, it immediately follows from (39) and (40) that the with probability tending to 1, $\lim_{k\to\infty}\boldsymbol{\beta}^{(k)}=\lim_{k\to\infty}(\hat{\boldsymbol{\alpha}}^{(k)\top},\hat{\boldsymbol{\gamma}}^{(k)\top})^{\top}=(\hat{\boldsymbol{\alpha}}^{\circ\top},0)^{\mathrm{T}}$, which completes the proof of part (i). This, in addition to part (b) of Lemma 2, proves part (ii) of Theorem 1. □

Proof of Theorem 2 Recall that $\hat{\boldsymbol{\beta}}^* = \lim_{k\to\infty} \hat{\boldsymbol{\beta}}^{(k+1)}$ and $\hat{\boldsymbol{\beta}}^{(k+1)} = \arg\min_{\boldsymbol{\beta}}\{Q(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^{(k)})\}$, where

$$Q(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^{(k)}) = \|\mathbf{Y}^* - \mathbf{x}\boldsymbol{\beta}\|^2 + \lambda_n \sum_{\ell=1}^{p_n} \beta_\ell^2/\{\hat{\beta}_\ell^{(k)}\}^2.$$

If $\beta_\ell^* \neq 0$ for $\ell \in \{i,j\}$, then $\hat{\boldsymbol{\beta}}^*$ must satisfy the following normal equations for $\ell \in \{i,j\}$:

$$-2\mathbf{x}_\ell^\top\{\mathbf{Y}^* - \mathbf{x}\hat{\boldsymbol{\beta}}^{(k+1)}\} + 2\lambda_n\hat{\beta}_\ell^{(k+1)}/\{\hat{\beta}_\ell^{(k)}\}^2 = 0.$$

Thus, for $\ell \in \{i,j\}$,

$$\hat{\beta}_\ell^{(k+1)}/\{\hat{\beta}_\ell^{(k)}\}^2 = \mathbf{x}_\ell^\top\hat{\boldsymbol{\varepsilon}}^{*(k+1)}/\lambda_n, \tag{45}$$

where $\hat{\boldsymbol{\varepsilon}}^{*(k+1)} = \mathbf{Y}^* - \mathbf{x}\hat{\boldsymbol{\beta}}^{(k+1)}$. Moreover, because

$$\|\hat{\boldsymbol{\varepsilon}}^{*(k+1)}\|^2 + \lambda_n \sum_{i=1}^{p_n} \frac{\hat{\beta}_i^2}{\tilde{\beta}_i^2} = Q(\hat{\boldsymbol{\beta}}^{(k+1)}|\hat{\boldsymbol{\beta}}^{(k)}) \le Q(0|\hat{\boldsymbol{\beta}}^{(k)}) = \|\mathbf{Y}^*\|^2,$$

we have

$$\|\hat{\boldsymbol{\varepsilon}}^{*(k+1)}\| \le \|\mathbf{Y}^*\|. \tag{46}$$

Letting $k \to \infty$ in (45) and (46), we have, for $\ell \in \{i,j\}$ and $\|\hat{\boldsymbol{\varepsilon}}^*\| \le \|\mathbf{Y}^*\|$, $\hat{\beta}_\ell^{*-1} = \mathbf{x}_\ell^\top\hat{\boldsymbol{\varepsilon}}^*\lambda_n$, where $\hat{\boldsymbol{\varepsilon}}^* = \mathbf{Y}^* - \mathbf{x}\hat{\boldsymbol{\beta}}^*$. Therefore,

$$\left|\hat{\beta}_i^{*-1} - \hat{\beta}_j^{*-1}\right| \le \frac{1}{\lambda_n}\|\mathbf{Y}^*\| \times \|\mathbf{x}_i - \mathbf{x}_j\| = \frac{1}{\lambda_n}\|\mathbf{Y}^*\|\sqrt{2(1-\rho_{ij})}.$$

□

Acknowledgements We are grateful to the referees, the associate editor and the editor for their helpful comments. The Glioblastoma multiforme data used in Sect. 4.2 are generated by the TCGA Research Network: https://www.cancer.gov/tcga.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723.

Box, J. K., Paquet, N., Adams, M. N., Boucher, D., Bolderson, E., Obyrne, K. J., Richard, D. J. (2016). Nucleophosmin: From structure and function to disease development. *BMC Molecular Biology, 17*(19), 1–12.

Breheny, P., Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics, 5*(1), 232–253.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics, 24,* 2350–2383.

Buckley, J., James, I. (1979). Linear regression with censored data. *Biometrika, 66*(3), 429–436.

Cai, T., Huang, J., Tian, L. (2009). Regularized estimation for the accelerated failure time model. *Biometrics, 65*(2), 394–404.
Chen, J., Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika, 95,* 759–771.
Cox, B. D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological), 34*(2), 187–220.
Cui, H., Li, R., Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association, 110*(510), 630–641.
Dai, L., Chen, K., Sun, Z., Liu, Z., Li, G. (2018). Broken adaptive ridge regression and its asymptotic properties. *Journal of Multivariate Analysis, 168,* 334–351.
Dai, L., Chen, K., Li, G. (2020). The broken adaptive ridge procedure and its applications. *Statistica Sinica, 30*(2), 1069–1094.
Datta, S., Le-Rademacher, J., Datta, S. (2007). Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and lasso. *Biometrics, 63*(1), 259–271.
Eirín-López, J. M., Frehlick, L. J., Ausió, J. (2006). Long-term evolution and functional diversification in the members of the nucleophosmin/nucleoplasmin family of nuclear chaperones. *Genetics, 173*(4), 1835–1850.
Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348–1360.
Fan, J., Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *Annals of Statistics, 30*(1), 74–99.
Fan, J., Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Methodological), 70*(5), 849–911.
Foster, D., George, E. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics, 22,* 1947–1975.
Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1–22.
Huang, J., Ma, S. (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analysis, 16*(2), 176–95.
Huang, J., Ma, S., Xie, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics, 62*(3), 813–820.
Johnson, B. A. (2009). On lasso for censored data. *Electronic Journal of Statistics, 3*(2009), 485–506.
Johnson, B. A., Lin, D. Y., Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association, 103*(482), 672–680.
Johnson, K. D., Lin, D., Ungar, L. H., Foster, D., Stine, R. (2015). A risk ratio comparison of l_0 and l_1 penalized regression. arXiv:1510.06319 [math.ST].
Kalbfleisch, J. D., Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). Hoboken: Wiley.
Kawaguchi, E. S., Suchard, M. A., Liu, Z., Li, G. (2020). A surrogate *l*0 sparse cox's regression with applications to sparse high-dimensional massive sample size time-to-event data. *Statistics in Medicine, 39*(6), 675–686.
Koul, H., Susarla, V., Ryzin, J. V. (1981). Regression analysis with randomly right-censored data. *Annals of Statistics, 9*(6), 1276–1288.
Leurgans, S. (1987). Linear models, random censoring and synthetic data. *Biometrika, 74*(2), 301–309.
Li, Y., Dicker, L., Zhao, S. D. (2014). The dantzig selector for censored linear regression models. *Statistica Sinica, 24*(1), 251–2568.
Liu, Y., Chen, X., Li, G. (2020). A new joint screening method for right-censored time-to-event data with ultra-high dimensional covariates. *Statistical Methods in Medical Research, 29*(6), 1499–1513.
Mallows, C. (1973). Some comments on c_p. *Technometrics, 15,* 661–675.
Mummenhoff, J., Houweling, A. C., Peters, T., Christoffels, V. M., Rther, U. (2001). Expression of Irx6 during mouse morphogenesis. *Mechanisms of Development, 103*(1–2), 193–195.
Nachmani, D., Bothmer, A. H., Grisendi, S., Mele, A., Pandolfi, P. P. (2019). Germline NPM1 mutations lead to altered rRNA 2-O-methylation and cause dyskeratosis congenita. *Nature Genetics, 51*(10), 1518–1529.
Nardi, Y., Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics, 2,* 605–633.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Shen, X., Pan, W., Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association, 107,* 223–232.
Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis, 45*(1), 89–103.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.
Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine, 16*(4), 385–395.
Wang, S., Nan, B., Zhu, J., Beer, D. G. (2008). Doubly penalized Buckley–James method for survival data with high-dimensional covariates. *Biometrics, 64*(1), 132–140.
Yuan, M., Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68*(1), 49–67.
Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics, 38*(2), 894–942.
Zhao, H., Wu, Q., Li, G., Sun, J. (2019). Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. *Journal of the American Statistical Association, 115*(529), 204–216.
Zhou, M. (1992). Asymptotic normality of the synthetic data regression estimator for censored survival data. *Annals of Statistics, 20*(2), 1002–1021.
Zhu, L., Li, L., Li, R., Zhu, L. (2011). Model-free feature screening for ultrahigh dimensional data. *Journal of the American Statistical Association, 106*(496), 1464–1475.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.
Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320.

Authors and Affiliations

Zhihua Sun[1] · Yi Liu[1] · Kani Chen[2] · Gang Li[3]

Zhihua Sun
zhihuasun@ouc.edu.cn

Yi Liu
liuyi@amss.ac.cn

Kani Chen
makchen@ust.hk

[1] Department of Mathematics, Ocean University of China, Qingdao 266000, China

[2] Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

[3] Biostatistics and Computational Medicine, University of California, Los Angeles, CA 90095-1772, USA