Annals of the Institute of Statistical Mathematics (2019) 71:1007–1031
https://doi.org/10.1007/s10463-018-0675-8

Joint feature screening for ultra-high-dimensional sparse additive hazards model by the sparsity-restricted pseudo-score estimator

Xiaolin Chen[1] · **Yi Liu**[2] · **Qihua Wang**[3,4]

Received: 23 November 2016 / Revised: 29 May 2018 / Published online: 22 June 2018

Abstract
Due to the coexistence of ultra-high dimensionality and right censoring, it is very challenging to develop feature screening procedure for ultra-high-dimensional survival data. In this paper, we propose a joint screening approach for the sparse additive hazards model with ultra-high-dimensional features. Our proposed screening is based on a sparsity-restricted pseudo-score estimator which could be obtained effectively through the iterative hard-thresholding algorithm. We establish the sure screening property of the proposed procedure theoretically under rather mild assumptions. Extensive simulation studies verify its improvements over the main existing screening approaches for ultra-high-dimensional survival data. Finally, the proposed screening method is illustrated by dataset from a breast cancer study.

Keywords Additive hazards model · Joint feature screening · Iterative hard-thresholding algorithm · Sure screening property

✉ Qihua Wang
qhwang@amss.ac.cn

Xiaolin Chen
xlchen@amss.ac.cn

Yi Liu
liuyi@amss.ac.cn

1 School of Statistics, Qufu Normal University, Qufu 273165, China

2 College of Science, China University of Petroleum (East China), Qingdao 266580, China

3 Department of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018, China

4 Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

1 Introduction

As the rapid development of science and technology, high-dimensional data are more frequently encountered in various scientific fields, such as molecular biology, clinical genomics, brain image research and so on. A distinguishing feature of high-dimensional data is that the number of predictors p is larger or much larger than that of observations n. When the dimension of data p grows exponentially with the sample size n, they have been called ultra-high-dimensional data in the literature. Furthermore, they are called ultra-high-dimensional survival data with the outcome being a time-to-event subject to right censoring. One basic task of statistical inference for such kind of data is to select only a small number of important variables to establish a parsimonious model. It is very challenging to develop reliable procedures for this target both methodologically and theoretically due to the coexistence of ultra-high dimensionality and right censoring, where regularized methods (Tibshirani 1997; Fan and Li 2002; Cai et al. 2005; Bradic et al. 2011; Huang et al. 2013; Leng and Ma 2007; Martinussen and Scheike 2009; Lin and Lv 2013) are inapplicable due to their computational expediency, statistical accuracy and algorithmic stability (Fan et al. 2009). A useful alternative method is feature screening or variable screening, which could handle the ultra-high dimensionality efficiently.

In the seminal work of Fan and Lv (2008), a marginal Pearson correlation screening method was proposed for linear model to reduce the ultra-high dimensionality of the feature space to a moderate size. This approach enjoys the sure screening property, which means that all important features are selected out with a probability tending to 1. So they named it sure independence screening (SIS). To improve the efficiency further, an iterative SIS (ISIS) method was further suggested. Inspired by Fan and Lv (2008), many researchers have paid attention to the investigation of feature screening methods in recent years. See, for example, Fan and Song (2010), Li et al. (2012), He et al. (2013), Chang et al. (2013), Fan et al. (2014), Song et al. (2014), Xu and Chen (2014), Chen et al. (2017), Liu and Chen (2018), Chen (2018) and so on.

In the past several years, investigation of feature screening for ultra-high-dimensional survival data has achieved rapid development. Literature about this aspect includes Fan et al. (2010), Zhao and Li (2012), Gorst-Rasmussen and Scheike (2013), He et al. (2013), Zhao and Li (2014), Song et al. (2014), Wu and Yin (2015), Zhou and Zhu (2017), Zhang et al. (2017), Chen et al. (2018) and so on. Fan et al. (2010) extended the SIS/ISIS to the Cox model without the theoretical justification. Zhao and Li (2012) proposed a principled sure independence screening method under the Cox model and established the corresponding sure screening property. Motivated by the pseudo-score estimator of the additive hazards model, Gorst-Rasmussen and Scheike (2013) proposed a method called the feature aberration at survival times (FAST) for the single-index hazard rate models. Zhao and Li (2014) suggested a score test variable screening approach, which could be used for several models, including the Cox model, the additive hazards model, the accelerated failure time model and so on. He et al. (2013) considered the variable screening for the quantile regression model based on the marginal spline estimator and inverse probability censoring weighting. Subsequently, Wu and Yin (2015) investigated an analogous model by combining a marginal utility and the technique of redistribution of the mass. Zhou and Zhu (2017) extended the

model-free feature screening method of Zhu et al. (2011) to the survival data. Zhang et al. (2017) proposed a correlation rank screening method based on the correlation of survival function of the response and each predictor. Chen et al. (2018) studied two robust feature screening procedures through the distance correlation.

It is noted that most of the existing methods are based on some marginal utility measures and thus have obvious shortcomings. For example, they fail to pick up features who are jointly related to the response with other covariates but marginally uncorrelated with the response. They could also recruit predictors who have strong marginal correlation but are actually unrelated with the response. To overcome marginal screening procedure's disadvantages, Xu and Chen (2014) proposed a novel, interesting and efficient joint screening procedure for linear and generalized linear model based on sparsity-restricted maximum likelihood estimator, which estimates the ultra-high-dimensional model coefficients on a low-dimensional subspace. The joint screening approach screens out features with zero-estimated coefficients and thus could naturally takes the joint effects of features in the screening process and improves the performances of marginal screening methods. Because of the superiority of this kind of approach, Yang et al. (2016) and Yang et al. (2018) generalized the joint screening idea from linear and generalized linear model to the Cox model and the additive Cox model for survival data, respectively.

In this paper, we consider the feature screening for ultra-high-dimensional survival data under the additive hazards model (Lin and Ying 1994), which takes the following form

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t) + \boldsymbol{Z}^T\boldsymbol{\beta}^*, \tag{1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_p)^T$ is the p-dimensional covariate vector, and $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*, \ldots, \beta_p^*)^T$ is the true regression coefficient vector. For $j = 1, \ldots, p$, the jth feature Z_j is deemed to be important if $\beta_j^* \neq 0$, otherwise unimportant. According to the sparsity principle, there are only a small number of nonzero β_j^*'s. Let M denote an arbitrary subset of $\{1, \ldots, p\}$. Denote by M_0 the set of indices of all important features, i.e., $M_0 = \{j : \beta_j^* \neq 0\}$. We assume that the cardinality of M_0 is small. Several papers have investigated regularized variable selection for this model, see, Leng and Ma (2007), Martinussen and Scheike (2009), Lin and Lv (2013) for example. However, there is little relevant research for feature screening. Although the approaches in Gorst-Rasmussen and Scheike (2013) and Zhao and Li (2014) could be used for that purpose, the current literature still lacks special methods tailored for the additive hazards model. Motivated by the success of joint screening in linear and generalized linear model (Xu and Chen 2014) and Cox model (Yang et al. 2016) based on the maximum likelihood estimator and maximum partial likelihood estimator, respectively, we proposed a joint screening method for the additive hazards model based on the sparsity-restricted pseudo-score estimator, which could be efficiently obtained by the iterative hard-thresholding algorithm. This proposed procedure is referred to as sparsity-restricted pseudo-score estimator-based screening (SPES for short). Compared with Gorst-Rasmussen and Scheike (2013) and Zhao and Li (2014), the SPES method has two significant advantages. First, it

could adequately capture the association between the features and thus improve the efficiency significantly, which has been verified in our simulation studies. Second, the proposed method could achieve the purpose of simultaneous feature screening and estimation of regression coefficients numerically, in the sense that it not only identifies the important features, but also gives empirically consistent estimates of the corresponding regression coefficients.

The remainder of this article is organized as follows. In Sect. 2, we introduce the SPES method for the sparse additive hazards model and the iterative hard-thresholding algorithm alongside a convergence theorem. The sure screening properties are also provided in this section. Section 3 reports the simulation results, while analysis of data from a breast cancer study illustrates the proposed method in Sect. 4. A brief discussion is given in Sect. 5. The proofs of the theorems are relegated to "Appendix."

2 Methodology

2.1 Sparsity-restricted pseudo-score estimator

Let T and C denote the survival time and censoring time, respectively. In addition, denote by $X = T \wedge C$ the observed time and by $\delta = I(T \leq C)$ the censoring indicator, where $I(A)$ is the indicator function of the set A. Throughout this article, we assume that T and C are conditionally independent given the p-dimensional covariates $\boldsymbol{Z}$. Let $\{X_i, \delta_i, \boldsymbol{Z}_i\}$, $i = 1, \ldots, n$, be a sample of n independent copies of $\{X, \delta, \boldsymbol{Z}\}$. Denote $N_i(t) = I(T_i \leq t, \delta_i = 1)$ and $Y_i(t) = I(X_i \geq t)$ to be the counting process and at-risk process, respectively.

For the purpose of estimating regression coefficient $\boldsymbol{\beta}^*$ when p is small, Lin and Ying (1994) proposed the following pseudo-score estimating function

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t)\}\{\mathrm{d}N_i(t) - Y_i(t)\boldsymbol{\beta}^T\boldsymbol{Z}_i\mathrm{d}t\}, \tag{2}$$

where τ is the maximum follow-up time, $\bar{\boldsymbol{Z}}(t) = \boldsymbol{S}^{(1)}(t)/\boldsymbol{S}^{(0)}(t)$ and $\boldsymbol{S}^{(k)}(t) = n^{-1}\sum_{i=1}^{n} Y_i(t)\boldsymbol{Z}_i^{\otimes k}$ for $k = 0, 1, 2$ with $\boldsymbol{a}^{\otimes 0} = 1$, $\boldsymbol{a}^{\otimes 1} = \boldsymbol{a}$ and $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^T$ for a vector $\boldsymbol{a}$. The estimating equation $\boldsymbol{U}_n(\boldsymbol{\beta}) = 0$ gives an explicit estimator for $\boldsymbol{\beta}^*$, denoted by $\check{\boldsymbol{\beta}} = \boldsymbol{V}_n^{-1}\boldsymbol{b}_n$, where

$$\boldsymbol{V}_n = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t)\}^{\otimes 2}Y_i(t)\mathrm{d}t$$

and

$$\boldsymbol{b}_n = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t)\}\mathrm{d}N_i(t).$$

However, in the scenario where the dimension p is large or very large compared with the sample size n, it is not feasible to make inference for $\boldsymbol{\beta}^*$ in this way. To identify the truly important features in the ultra-high-dimensional covariates, Gorst-Rasmussen and Scheike (2013) proposed a method, called FAST, which used $b_{n,j}$ (the jth component of $\boldsymbol{b}_n$) to measure the influence of jth covariate on the survival time for $j = 1, \ldots, p$. The utility can be understood from the fact that $\boldsymbol{b}_n$ is a (scaled) estimator vector of the regression coefficients from p univariate marginal additive hazards model. Although this approach escapes from estimating the singular matrix $\boldsymbol{V}_n$, it makes itself to be a marginal screening means. Thus, it has the disadvantages inherent in the marginal screening procedures. For example, FAST may miss truly active features which are marginally independent of the survival time, but contribute to the survival time jointly with other predictors. In addition, when there are many irrelevant predictors which are highly correlated with some strongly active predictors, FAST may fail to identify other active predictors with relatively weak marginal signals. To overcome these drawbacks and enhance the finite sample performance, by combining the FAST and regularized variable selection, Gorst-Rasmussen and Scheike (2013) further proposed an iterative version of FAST, named IFAST. However, this kind of iterative screening approaches still lacks theoretical support.

Different from the method of Gorst-Rasmussen and Scheike (2013), we propose a new joint screening procedure based on the constrained pseudo-score estimator with an explicit constraint on the number of nonzero regression coefficients. Specifically, for some known positive integer k, we define the sparsity-restricted pseudo-score estimator as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathcal{B}(k)}{\operatorname{argmax}} \{L_n(\boldsymbol{\beta})\}, \tag{3}$$

where $\mathcal{B}(k) = \{\boldsymbol{\beta} | \boldsymbol{\beta} \in \mathcal{B}, \|\boldsymbol{\beta}\|_0 \leq k\}$ with $\mathcal{B}$ being a compact set containing true regression coefficient, $\|\cdot\|_0$ being the number of nonzero coordinates and the objective function $L_n(\boldsymbol{\beta}) = \boldsymbol{b}_n^T \boldsymbol{\beta} - \frac{1}{2}\boldsymbol{\beta}^T \boldsymbol{V}_n \boldsymbol{\beta}$. Then, we identify the important features by the nonzero components of $\hat{\boldsymbol{\beta}}$, i.e., the selected model is $\hat{M} = \{j : \hat{\beta}_j \neq 0\}$. It is easy to see that $\hat{M}$ may depend on k. However, we omit this notation for simplicity. In order to detect all important features, it is necessary that k is not less than the true number of nonzero regression coefficient of $\boldsymbol{\beta}^*$, say q. The proposed sparsity-restricted pseudo-score estimator is motivated by two facts. Firstly, solving estimating equation $\boldsymbol{U}_n(\boldsymbol{\beta}) = 0$ is equivalent to maximizing the objective function $L_n(\boldsymbol{\beta})$. This fact has been used in the regularized variable selection for additive hazards model by several researchers, such as Leng and Ma (2007), Lin and Lv (2013) and so on. Secondly, by the sparsity principle, the number of nonzero components of $\boldsymbol{\beta}^*$ is small. Thus, it is more reasonable to solve the estimating equation $\boldsymbol{U}_n(\boldsymbol{\beta}) = 0$ under the constraint on the number of nonzero components. It is easy to see that the suggested screening procedure is significantly different from the method of Gorst-Rasmussen and Scheike (2013) in that our procedure is a joint screening means which could take the joint effects of all the covariates into account naturally. Therefore, it is anticipated that the SPES approach could improve the FAST and IFAST remarkably, which has been verified in our simulation studies in Sect. 3.

For the sake of presenting the sure screening property of the SPES approach, we introduce some notations and assumptions. For a model M, denote $\boldsymbol{Z}_M = \{Z_j, j \in M\}$ and $\boldsymbol{\beta}_M = \{\beta_j, j \in M\}$. Define $\mathbf{M}_+^k = \{M | M_0 \subset M, \|M\|_0 \leq k\}$ and $\mathbf{M}_-^k = \{M | M_0 \not\subset M, \|M\|_0 \leq k\}$. Let $\boldsymbol{s}^{(k)}(t) = E\{Y(t)\boldsymbol{Z}^{\otimes k}\}$ for $k = 0, 1, 2$, $\bar{\boldsymbol{z}}(t) = \boldsymbol{s}^{(1)}(t)/\boldsymbol{s}^{(0)}(t)$.

Assumption 1 (i) There exists a compact neighborhood $\mathcal{B}$ such that $\boldsymbol{\beta}^* \in \mathcal{B}$; (ii) $\int_0^\tau \lambda_0(t)dt < \infty$; (iii) $P\{Y(\tau) = 1\} > 0$; (iv) $\boldsymbol{Z}$ is bounded in probability 1.

Assumption 2 Denote $d_n = \sup_{t\in[0,\tau]} \|\bar{\boldsymbol{Z}}(t) - \bar{\boldsymbol{z}}(t)\|_\infty$ and $e_n = \sup_{t\in[0,\tau]} |\boldsymbol{S}^{(0)}(t) - \boldsymbol{s}^{(0)}(t)|$. The random sequences d_n and e_n are bounded almost surely.

Assumption 3 Define $\varepsilon_{ij} = \int_0^\tau \{Z_{ij} - \bar{z}_j(t)\} \mathrm{d}M_i(t)$, where $\bar{z}_j(t)$ is the jth component of $\bar{\mathbf{z}}(t)$ and $M_i(t) = N_i(t) - \int_0^t Y_i(s)\{\lambda_0(s) + \boldsymbol{Z}_i^T \boldsymbol{\beta}^*\}\mathrm{d}s$. Suppose that the Cramér condition holds for ε_{ij}, i.e., $E|\varepsilon_{ij}|^l \leq 2^{-1} l! c_1^{l-2} \sigma_j^2$ for all j, where c_1 is a positive constant, $l \geq 2$, and $\sigma_j^2 = \mathrm{var}(\varepsilon_{ij}) < \infty$.

Assumption 4 There exist positive constants c_2, c_3, τ_1 and τ_2 such that $\min_{j\in M_0} |\beta_j^*| \geq c_2 n^{-\tau_1}$ and $q \leq k \leq c_3 n^{\tau_2}$.

Assumption 5 For any $M \in \mathbf{M}_+^{2k}$, it holds that $\rho_{\min}(\boldsymbol{V}_{n,M}) \geq c_4$, where c_4 is a positive constant, $\rho_{\min}(\boldsymbol{A})$ is the minimum eigenvalue of the matrix $\boldsymbol{A}$ and $\boldsymbol{V}_{n,M}$ is the version of $\boldsymbol{V}_n$ based on model M.

Assumption 1 is very mild and widely used in the survival analysis literature. Assumptions 2 and 3 are presented here to prove a large deviation result for martingales under the additive hazards model, which is useful in proving the sure screening property of the SPES procedure. Let $\mathcal{F}_{ti} = \sigma\{N_i(u), \mathbf{Z}_i(u+), Y_i(u+), 0 \leq u \leq t\}$ for $i = 1, \ldots, n$. Then $M_i(t), i = 1, \ldots, n$ are orthogonal local square integrable martingales with respect to filtrations $\mathcal{F}_{ti}, i = 1, \ldots, n$. In Assumption 4, we assume that the minimum true signal cannot be too small and it is in the order of $n^{-\tau_1}$ which allows the minimum true signal to vanish to zero as the sample size n approaches the infinity. Such an assumption is regular in the feature screening literature. In addition, Assumption 4 also allows the dimension of important features to diverge to infinity at a polynomial speed of the sample size n. Assumption 5 is easily to be satisfied in view that k is not large.

In the following theorem, we provide the desirable sure screening property of SPES.

Theorem 1 *Under Assumptions 1–5, if* $2\tau_1 + \tau_2 < 1$, $\log(p) = O(n^m)$, $2\tau_1 + 3\tau_2 < 1 - 2m$, *where* $0 < m < \frac{1}{2}$ *and* $\max_{1\leq j\leq p} \sigma_j^2 = O(n^{\frac{1}{2}-\tau_1-\frac{\tau_2}{2}})$, *we have*

$$\mathrm{pr}(M_0 \subset \hat{M}) \to 1,$$

as n goes to infinity.

2.2 Implementation

Although the sure screening property has been obtained under rather mild assumptions, there still exists a great challenge in the computation of the sparsity-restricted pseudo-score estimator. As noted by Xu and Chen (2014), this kind of estimator resembles the best-subset selection procedure, and thus can be computationally expensive, especially for high-dimensional problems. Following Xu and Chen (2014), we propose the iterative hard-thresholding (IHT) algorithm to solve the suggested sparsity-restricted pseudo-score estimator.

Specifically, given the current estimate $\hat{\boldsymbol{\beta}}^{(t)}$, next iterative value $\hat{\boldsymbol{\beta}}^{(t+1)}$ could be obtained by

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \underset{\boldsymbol{\beta}\in\mathcal{B}(k)}{\operatorname{argmax}} \left\{ L_n\left(\hat{\boldsymbol{\beta}}^{(t)}\right) + \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(t)}\right)^T \boldsymbol{U}_n\left(\hat{\boldsymbol{\beta}}^{(t)}\right) - \frac{u}{2}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(t)}\|_2^2 \right\}, \quad (4)$$

for $t = 0, 1, 2, \ldots$. The first two terms in the right side of (4) come from the Taylor's expansion of $L_n(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}^{(t)}$ and $\frac{u}{2}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(t)}\|_2^2$ is a regularization term which controls how far next iteration moves from the current iterate in terms of Euclidean norm. For the $(t+1)$th iteration, $\hat{\boldsymbol{\beta}}^{(t+1)}$ can be computed by a two-step procedure. We firstly solve (4) without the L_0 norm constraint and denote the solution by $\tilde{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + u^{-1}\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}^{(t)})$. If $\|\tilde{\boldsymbol{\beta}}^{(t+1)}\|_0 \le k$, set $\hat{\boldsymbol{\beta}}^{(t+1)} = \tilde{\boldsymbol{\beta}}^{(t+1)}$. Otherwise, pick out the kth largest component of the absolute $\tilde{\boldsymbol{\beta}}^{(t+1)}$, denoted by $\tilde{\beta}_{(k)}^{(t+1)}$. Then retain the components of $\tilde{\boldsymbol{\beta}}^{(t+1)}$, whose absolute values are no less than the absolute value of $\tilde{\beta}_{(k)}^{(t+1)}$, and set the other components to be zeros. Formally,

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \mathbf{H}(\tilde{\boldsymbol{\beta}}^{(t+1)}; k) = \left(H\left(\tilde{\beta}_1^{(t+1)}; \tilde{\beta}_{(k)}^{(t+1)}\right), \ldots, H\left(\tilde{\beta}_p^{(t+1)}; \tilde{\beta}_{(k)}^{(t+1)}\right)\right)^T, \quad (5)$$

where $H(\tilde{\beta}_j^{(t+1)}; \tilde{\beta}_{(k)}^{(t+1)}) = \tilde{\beta}_j^{(t+1)} I(|\tilde{\beta}_j^{(t+1)}| \ge |\tilde{\beta}_{(k)}^{(t+1)}|)$.

In the following theorem, we provide the convergence property of the IHT algorithm for the solution sequence of the sparsity-restricted pseudo-score estimator.

Theorem 2 *For the iterative sequence (5), if $u \ge \rho_{\max}(\boldsymbol{V}_n)$, we have that $L_n(\hat{\boldsymbol{\beta}}^{(t+1)}) \ge L_n(\hat{\boldsymbol{\beta}}^{(t)})$.*

This theorem guarantees the ascent property of the IHT algorithm if step length is properly selected. From this theorem, we also could conclude that it is necessary that the step length u should be no less than $\rho_{\max}(\boldsymbol{V}_n)$ to guarantee the convergence of the suggested IHT algorithm. However, as is known to all, the smaller the step length u is, the faster of the IHT algorithm is. Besides, the largest eigenvalue of $\boldsymbol{V}_n$ may be large since the dimension of matrix $\boldsymbol{V}_n$ is high. Therefore, it is likely that the convergence speed of proposed IHT algorithm is very slow. Nevertheless, $u \ge \rho_{\max}(\boldsymbol{V}_n)$ is only a sufficient condition to guarantee the IHT algorithm's convergence, but not a necessary condition. This fact motivates us to utilize a well-known adaptive step length double

scheme (Bertsekas 2016) to increase the convergence, and then reduce the computation burden. We call this algorithm as the adaptive IHT algorithm, which is summarized as follows:

Step 1 Obtain an initial estimator $\hat{\boldsymbol{\beta}}^{(0)}$. Compute $L_n(\hat{\boldsymbol{\beta}}^{(0)})$ and set $t = 0$;

Step 2 Compute $\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}^{(t)})$ and update $\boldsymbol{\beta}$ according to (5). Denote by $\hat{\boldsymbol{\beta}}^{(t+1)}$ the updated $\boldsymbol{\beta}$;

Step 3 Compute $L_n(\hat{\boldsymbol{\beta}}^{(t+1)})$. If $L_n(\hat{\boldsymbol{\beta}}^{(t+1)}) > L_n(\hat{\boldsymbol{\beta}}^{(t)})$, go to Step 4; otherwise, double the current u and go back to Step 2;

Step 4 Repeat Steps 2 and 3 until convergence.

As an iterative algorithm, the IHT algorithm needs an initial estimator to accomplish the whole iteration process. In our experiences, a good initial estimator speeds up the convergence and the LASSO estimator is a suitable choice for the initial estimator. Theoretically, we have proven that for the LASSO initial estimator with some appropriate tuning parameters, the adaptive IHT algorithm has the sure screening property for any finite number of iterations. This gives a justification for the use of LASSO estimator as an initial estimator. Before presenting the formal result, one more assumption is required:

Assumption 6 There exists a positive constant c_5 such that, for sufficiently large n,

$$\boldsymbol{\eta}^T \boldsymbol{V}_n \boldsymbol{\eta} \geq c_5 \|\boldsymbol{\eta}_{M_0}\|_2^2,$$

for any $\boldsymbol{\eta} \neq \mathbf{0}_p$, $\|\boldsymbol{\eta}_{M_0^c}\|_1 \leq 3\|\boldsymbol{\eta}_{M_0}\|_1$, where M_0^c is the complement of M_0 in $\{1, 2, \ldots, p\}$.

Assumption 6 is needed for deriving an error bound for the LASSO estimator. There are many similar assumptions in the literature, such as Bickel et al. (2009) for the linear model, Huang et al. (2013) for the Cox model, Xu and Chen (2014) for the generalized linear model, and so on. Thus, this assumption is rather mild. Throughout the paper, let $\|\cdot\|_\infty$ be the ∞ norm for a vector or matrix.

Theorem 3 *Define* $\hat{\boldsymbol{\beta}}^{(0)} = \operatorname{argmax}_{\boldsymbol{\beta}}\{L_n(\boldsymbol{\beta}) - \lambda\|\boldsymbol{\beta}\|_1\}$, *where* λ *satisfies* $\lambda n^{\frac{1}{2}-m} \to \infty$, $\lambda n^{\tau_1+\tau_2} \to 0$. *Under Assumptions 1–6, if* $\log(p) = O(n^m)$, $\tau_1 + \tau_2 < \frac{1}{2} - m$, $\|\boldsymbol{V}_n\|_\infty = O_p(n^{\tau_3})$, *and* $u > c_6 r$ *with* $r = O(n^{\tau_3})$ *and* c_6 *being a positive constant, we have*

$$\operatorname{pr}(M_0 \subset \hat{M}^{(t)}) \to 1,$$

for any finite $t \geq 1$ *as* n *goes to infinity, where* $\hat{M}^{(t)} = \{j : \hat{\beta}_j^{(t)} \neq 0\}$ *is the set of screened features by* $\hat{\boldsymbol{\beta}}^{(t)}$.

This theorem provides solid theoretical ground for the use of LASSO initial estimator. But it is still hard to determine which estimator to be employed in the entire solution path of the LASSO. In our simulation studies, we use λ that recruits the first

k features in the solution path of LASSO as the chosen tuning parameter value. This choice works well for all the examples in our simulation studies. Although the exact sparsity-restricted pseudo-score estimator $\hat{\boldsymbol{\beta}}$ could not be obtained and the IHT solver is only an approximate solution, the sure screening property is not affected by this approximation. This observation is easily seen from Theorem 3. In addition, from the proof of Theorem 3, we have $\|\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|_\infty = o_p(w)$ with $w = \min_{j \in M_0} \|\beta_j^*\|$ for any t, which hints that the sparsity-restricted pseudo-score estimator may have desirable estimation accuracy, which is validated in our simulation studies.

3 Simulation studies

In this section, we will present extensive simulation studies to evaluate the performance of the proposed SPES procedure. In addition, we also report the comparisons with PSIS (principled sure independence screening) of Zhao and Li (2012), FAST and IFAST (iterative FAST) of Gorst-Rasmussen and Scheike (2013), CRIS (censored rank independence screening) of Song et al. (2014), CSIRS (censored sure independent ranking and screening) of Zhou and Zhu (2017), CRSIS (correlation rank sure independent screening) of Zhang et al. (2017) and popular LASSO to study the SPES's improvements. For LASSO and FAST/IFAST, we directly use the functions in R package **ahaz**. Specifically, we obtain LASSO's entire regularization path by the function "ahazpen" and select the LASSO estimator as the first solution with nonzero number being no larger than the prespecified k according to order of the tuning parameter. In addition, FAST is carried out by the function "ahazisis" with the features recruited being k. Moreover, IFAST is also executed by this function with LASSO penalty being in its iteration process. The iteration are performed five loops.

Let $[a]$ denote the integer part of a. As for the screening bound k, Fan and Lv (2008) recommended $k = [n/\log(n)]$ as a sensible choice under the linear regression model, while Li et al. (2012) used $k = a[n/\log(n)]$ with a being a positive integer for model-free screening. Xu and Chen (2014) set $k = a\log(n)n^{1/3}$ with $a = 1, 2/3$ and $1/3$ for linear model, Poisson model and logistic model, respectively. As discussed in Fan et al. (2009), choosing a larger value of k increases the probability that screening method will include all of the important variables, but including more inactive variables will tend to have a slight detrimental effect on the performance of the final variable selection and parameter estimation method. In addition, it is reasonable to chose a small k for models in which the response provides less information about the covariates. So Fan et al. (2009) applied $k = [n/(4\log(n))]$ for logistic regression and $k = [n/(2\log(n))]$ for poisson regression. As noted, the estimating function (2) provides less information compared with likelihood based method. Besides, a small k is beneficial to our sparsity-restricted pseudo-score estimation. So we use $k = [n/(5\log(n))]$ throughout our simulation. And we find that this option works well in all our examples.

Denote by L the number of our Monte Carlo repetitions. For $l = 1, \ldots, L$, let $\hat{M}_l$ and $\hat{\boldsymbol{\beta}}_l = (\hat{\beta}_{l,1}, \hat{\beta}_{l,2}, \ldots, \hat{\beta}_{l,p})^T$ be model selected and estimator of $\boldsymbol{\beta}$ in the lth repetition for each method, respectively. To summarize our simulation results, we will report the following performance measures: $\mathrm{RC}_{\mathrm{all}}$, the retaining capac-

ity of all important features, $L^{-1}\sum_{l=1}^{L} I(M_0 \subset \hat{M}_l)$; RC_j, the retaining capacity of jth important features, $L^{-1}\sum_{l=1}^{L} I(\hat{\beta}_{l,j} \neq 0)$ for $j = 1, \ldots, q$; PSR, the positive selection rate, $L^{-1}\sum_{l=1}^{L} \|M_0 \cap \hat{M}_l\|_0/q$; FDR, the false discovery rate, $L^{-1}\sum_{l=1}^{L} \|\hat{M}_l - M_0\|_0/\|\hat{M}_l\|_0$; AMS, the average model size, $L^{-1}\sum_{l=1}^{L} \|\hat{M}_l\|_0$; TP, the average number of important feature selected, $L^{-1}\sum_{l=1}^{L}\sum_{j=1}^{q} I(\hat{\beta}_{l,j} \neq 0)$; L1.err, $L^{-1}\sum_{l=1}^{L} \|\hat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}^*\|_1 = L^{-1}\sum_{l=1}^{L}\sum_{j=1}^{p} |\hat{\beta}_{l,j} - \beta_j^*|$; L2.err, $L^{-1}\sum_{l=1}^{L} \|\hat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}^*\|_2 = L^{-1}\sum_{l=1}^{L} \sqrt{\sum_{j=1}^{p}(\hat{\beta}_{l,j} - \beta_j^*)^2}$. A good procedure should show high $\mathrm{RC}_{\mathrm{all}}$, RC_j, PSR, TP and low FDR, L1.err and L2.err. In our simulation results, we do not present L1.err and L2.err for the method of FAST, because the function "ahazisis" in R packages **ahaz** does not compute the marginal regression coefficients, and the corresponding estimators from IFAST must be more accurate than those obtained from FAST. In addition, we also do not give L1.err and L2.err for PSIS, CRIS, CSIRS and CRSIS because PSIS is tailed for the Cox model and CRIS, CSIRS and CRSIS are model-free. It is meaningless to list results of L1.err and L2.err for these approaches.

In the following five examples, survival times were generated according to the additive hazards model (1) with baseline hazard function $\lambda_0(t) = 1$, and covariates and regression coefficients described in detail in every example. It should be noted that in all these examples the covariates are generated under the constraint $\boldsymbol{Z}^T\boldsymbol{\beta}^* > -1$ to ensure the positivity of conditional hazard function. Besides, we simulated the censoring times according to the uniform distribution on $(0, c_0)$, where different c_0's were set to produce approximate 35% censoring rate in each example. Throughout, we set the total number of repetitions of each scenario 300 times. Throughout the paper, $\mathbf{0}$ represents a vector with all the components being zeros. And the according dimension is indicated by the subscript. Some other detailed elements of simulation setup are given as follows:

Example 1 $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^T$ follows the multivariate normal distribution with mean $\mathbf{0}_p$ and the covariance matrix $\boldsymbol{\Sigma} = (\sigma_{rs})_{p\times p}$ with $\sigma_{rr} = 1$ and $\sigma_{rs} = 0$ for $r \neq s$. $M_0 = \{1, 2, 3, 4\}$, $\boldsymbol{\beta}_{M_0} = (3, 3, -3, -3)^T$, $\boldsymbol{\beta}_{M_0^c} = \mathbf{0}_{p-4}$, and $(n, p) = (400, 5000)$.

Example 2 Same as Example 1 except that $\sigma_{rs} = \rho^{|r-s|}$ with $\rho = 0.2$ and $(n, p) = (400, 3000)$.

Example 3 Same as Example 2 except that $\rho = 0.5$ and $(n, p) = (400, 2000)$.

Example 4 $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^T$ follows the multivariate normal distribution with mean $\mathbf{0}_p$ and the covariance matrix $\boldsymbol{\Sigma} = (\sigma_{rs})_{p\times p}$ with $\sigma_{rr} = 1$. $M_0 = \{1, 2, 3, 4\}$. When $r \neq s$, $\sigma_{rs} = 0.15$ for $r, s \in M_0$ and $\sigma_{rs} = 0.3$ for r or $s \in M_0^c$. $\boldsymbol{\beta}_{M_0} = (3, 3, -3, -3)^T$, $\boldsymbol{\beta}_{M_0^c} = \mathbf{0}_{p-4}$, and $(n, p) = (400, 2000)$.

Example 5 $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^T$ follows the multivariate normal distribution with mean $\mathbf{0}_p$ and the covariance matrix $\boldsymbol{\Sigma} = (\sigma_{rs})_{p\times p}$ with $\sigma_{rs} = \rho^{|r-s|}$ and $\rho = 0.5$. $M_0 = \{1, 2, 3, 4\}$, $\boldsymbol{\beta}_{M_0} = (3, -3, 3, -3\rho + 3\rho^2 - 3\rho^3)^T$, $\boldsymbol{\beta}_{M_0^c} = \mathbf{0}_{p-4}$, and $(n, p) = (400, 1000)$.

Example 1, with the independent covariates, is the most straightforward for feature screening. In Examples 2 and 3, there exists serial correlation in the covariates with the correlation decaying as $|r - s|$ increases. $\rho = 0.2$ and $\rho = 0.5$ represent the low and

Table 1 Summary statistics for Example 1

Method	PSR	FDR	AMS	TP	L1.err	L2.err	RC_1	RC_2	RC_3	RC_4	RC_{all}
PSIS	0.826	0.745	13	3.306	–	–	0.836	0.853	0.786	0.830	0.416
FAST	0.687	0.788	13	2.750	–	–	0.716	0.670	0.636	0.726	0.136
IFAST	0.958	0.705	13	3.833	8.056	6.958	0.953	0.960	0.960	0.960	0.883
CRIS	0.237	0.926	13	0.950	–	–	0.343	0.326	0.093	0.186	0.000
CSIRS	0.606	0.813	13	2.426	–	–	0.653	0.606	0.536	0.630	0.060
CRSIS	0.473	0.854	13	1.893	–	–	0.483	0.456	0.450	0.503	0.013
LASSO	0.927	0.687	12	3.710	10.590	25.782	0.956	0.930	0.916	0.906	0.756
SPES	0.989	0.695	13	3.956	10.432	9.937	0.986	0.990	0.986	0.993	0.983

Table 2 Summary statistics for Example 2

Method	PSR	FDR	AMS	TP	L1.err	L2.err	RC_1	RC_2	RC_3	RC_4	RC_{all}
PSIS	0.859	0.735	13	3.436	–	–	0.950	0.740	0.783	0.963	0.500
FAST	0.755	0.767	13	3.020	–	–	0.860	0.590	0.653	0.916	0.200
IFAST	0.980	0.698	13	3.923	7.519	5.730	0.990	0.966	0.980	0.986	0.926
CRIS	0.303	0.906	13	1.213	–	–	0.476	0.253	0.176	0.306	0.000
CSIRS	0.684	0.789	13	2.736	–	–	0.823	0.503	0.580	0.830	0.110
CRSIS	0.590	0.818	13	2.360	–	–	0.693	0.446	0.466	0.753	0.056
LASSO	0.963	0.675	12	3.853	9.794	22.063	0.996	0.940	0.926	0.990	0.870
SPES	0.998	0.692	13	3.993	10.149	9.305	1.000	0.996	0.996	1.000	0.996

moderate correlations, respectively. Example 4 is a much more difficult setting, which describes the scenario that every feature is either important or associated with some other important features. It is very challenging to distinguish the important features from the unimportant ones under this construction. This example is borrowed from Xu and Chen (2014). Example 5 depicts a challenging situation where Z_4 is marginally unimportant, but jointly significant. Thus, the marginal screening procedure is likely to lose it.

Simulation results for Examples 1–3 are presented in Tables 1, 2 and 3, from which we can see that the SPES method outperforms the other methods in terms of the above-mentioned measures overall. The proposed method has the highest PSRs, TPs, RC_{all}s and second lowest FDRs in all the three examples. Only LASSO has better FDR values than SPES. However, the advantages are so slight that they could be ignored. In addition, SPES has almost the highest RC_js except RC_3s in Table 3. As for L1.errs and L2.errs, IFAST obtains better results for our proposed method. This phenomenon is reasonable by noting that IFAST performs the regularized variable selection step in the iteration process and thus could obtain very accurate estimators.

Tables 4 and 5 report the simulation results for Examples 4 and 5. For Example 4, we could see that the SPES has the highest PSRs, TPs and second lowest FDRs and RC_{all}s. We believe that the difference of RC_{all}s of LASSO and SPES are most probably

Table 3 Summary statistics for Example 3

Method	PSR	FDR	AMS	TP	L1.err	L2.err	RC_1	RC_2	RC_3	RC_4	RC_{all}
PSIS	0.755	0.767	13	3.020	–	–	0.983	0.513	0.530	0.993	0.190
FAST	0.688	0.788	13	2.753	–	–	0.913	0.426	0.450	0.963	0.063
IFAST	0.943	0.709	13	3.773	7.755	7.084	0.996	0.883	0.893	1.000	0.786
CRIS	0.301	0.907	13	1.206	–	–	0.566	0.210	0.086	0.343	0.000
CSIRS	0.642	0.802	13	2.570	–	–	0.856	0.403	0.400	0.910	0.063
CRSIS	0.534	0.835	13	2.136	–	–	0.790	0.260	0.276	0.810	0.003
LASSO	0.871	0.707	12	3.486	10.076	23.519	1.000	0.766	0.720	1.000	0.580
SPES	0.945	0.709	13	3.780	10.425	10.423	0.996	0.896	0.886	1.000	0.873

Table 4 Summary statistics for Example 4

Method	PSR	FDR	AMS	TP	L1.err	L2.err	RC_1	RC_2	RC_3	RC_4	RC_{all}
PSIS	0.825	0.746	13	3.300	–	–	0.846	0.803	0.820	0.830	0.446
FAST	0.746	0.770	13	2.986	–	–	0.746	0.730	0.743	0.766	0.290
IFAST	0.955	0.705	13	3.823	7.712	6.673	0.960	0.950	0.950	0.963	0.823
CRIS	0.355	0.890	13	1.420	–	–	0.443	0.470	0.233	0.273	0.003
CSIRS	0.700	0.784	13	2.800	–	-	0.700	0.663	0.730	0.706	0.193
CRSIS	0.606	0.813	13	2.426	-	-	0.630	0.586	0.600	0.610	0.083
LASSO	0.998	0.670	12	3.993	8.956	18.083	0.996	1.000	1.000	0.996	0.993
SPES	0.996	0.693	13	3.986	9.829	8.787	0.996	1.000	0.996	0.993	0.990

Table 5 Summary statistics for Example 5

Method	PSR	FDR	AMS	TP	L1.err	L2.err	RC_1	RC_2	RC_3	RC_4	RC_{all}
PSIS	0.495	0.847	13	1.980	–	–	0.996	0..040	0.930	0.013	0.000
FAST	0.475	0.853	13	1.903	–	–	0.986	0.040	0.860	0.016	0.000
IFAST	0.806	0.751	13	3.226	6.720	6.356	0.983	0.993	0.833	0.416	0.413
CRIS	0.319	0.901	13	1.276	–	–	0.733	0.003	0.530	0.010	0.000
CSIRS	0.453	0.860	13	1.813	–	–	0.970	0.026	0.810	0.006	0.000
CRSIS	0.406	0.874	13	1.626	–	–	0.906	0.013	0.690	0.016	0.000
LASSO	0.579	0.806	12	2.316	9.852	24.260	0.986	0.416	0.883	0.030	0.023
SPES	0.965	0.703	13	3.860	6.211	3.750	0.990	0.983	0.986	0.900	0.900

caused by the randomness. It is understandable that LASSO performs comparably with SPES by noting that it could be deemed as a joint screening procedure. However, in our limited experience, LASSO is not very stable under complex data structure. When we decrease the sample from 400 to 300, the RC_{all} of LASSO reduces significantly, while SPES only has a small loss in RC_{all}. To save space, we did not list the simulation results for Example 4 with $n = 300$. This point is further verified by Example 5, in which a marginally unimportant, but jointly significant predictor exists. From Table 5, we can

Table 6 Biases of coefficient estimators for Examples 1–5

Setup	Method	β_1	β_2	β_3	β_4
Example 1	LASSO	− 2.506	− 2.522	2.536	2.500
	IFAST	− 0.494	− 0.512	0.516	0.460
	SPES	0.208	0.168	− 0.159	− 0.226
Example 2	LASSO	− 2.241	− 2.413	2.392	2.222
	IFAST	− 0.316	− 0.476	0.413	0.282
	SPES	0.260	0.204	− 0.240	− 0.275
Example 3	LASSO	− 2.081	− 2.678	2.693	2.067
	IFAST	− 0.212	− 0.765	0.790	0.227
	SPES	0.263	− 0.119	0.158	− 0.297
Example 4	LASSO	− 2.090	− 2.106	2.096	2.089
	IFAST	− 0.437	− 0.469	0.445	0.422
	SPES	0.236	0.210	− 0.214	− 0.212
Example 5	LASSO	− 2.587	2.906	− 2.777	1.123
	IFAST	− 0.712	0.890	− 1.058	0.680
	SPES	0.129	− 0.089	0.094	0.051

see that LASSO could not identify this marginally unimportant, but jointly significant variable, even if it could be seen as a joint screening method. In this example, we also could see that SPES has the highest PSRs, TPs, RC_{all}s and lowest FDRs and L1.errs and L2.errs. These results indicate the reliability of SPES under complex data structure.

From Tables 1, 2, 3, 4 and 5, we could see that the L1.errs and L2.errs of SPES are comparable to those of IFAST, which estimates the regression coefficients by the penalized method in the iterative loop. This motivates us to examine how accurate the sparsity-restricted pseudo-score estimator of regression coefficients could be. Table 6 presents the biases of coefficient estimators of truly important features, from which we can see that the sparsity-restricted pseudo-score estimator and estimators from IFAST are consistent numerically, while estimators from other methods are not. Furthermore, estimators from SPES have smaller biases than those from IFAST. This phenomenon indicates that the sparsity-restricted pseudo-score estimator-based screening can also be used for estimation. In other words, our suggested methods could achieve the goal of simultaneous feature screening and coefficient estimation.

As suggested by one reviewer, we conducted further simulation studies to investigate and compare the performances of SPES and the examined approaches under smaller absolute regression coefficients and misspecified survival models. The detailed elements of simulation setup are provided below.

Example 6 Same as Example 5 except that $\boldsymbol{\beta}_{M_0} = (1, -1, 1, -\rho + \rho^2 - \rho^3)^T$.

Example 7 Survival times were generated from the following Cox model:

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t)\exp(\boldsymbol{Z}^T\boldsymbol{\beta}^*),$$

Table 7 Summary statistics for Example 6

Method	PSR	FDR	AMS	TP	L1.err	L2.err	RC_1	RC_2	RC_3	RC_4	RC_{all}
PSIS	0.495	0.847	13	1.980	–	–	0.996	0.030	0.953	0.000	0.000
FAST	0.488	0.849	13	1.953	–	–	0.986	0.030	0.936	0.000	0.000
IFAST	0.830	0.744	13	3.320	2.507	0.663	0.990	1.000	0.913	0.416	0.406
CRIS	0.331	0.897	13	1.326	–	-	0.776	0.013	0.516	0.020	0.000
CSIRS	0.465	0.856	13	1.863	–	–	0.973	0.023	0.856	0.010	0.000
CRSIS	0.414	0.872	13	1.656	–	–	0.913	0.016	0.723	0.003	0.000
LASSO	0.618	0.794	12	2.473	3.206	2.479	0.996	0.526	0.923	0.026	0.006
SPES	0.935	0.712	13	3.740	2.806	0.708	1.000	0.986	0.986	0.766	0.766

Table 8 Summary statistics for Example 7

Method	PSR	FDR	AMS	TP	RC_1	RC_2	RC_3	RC_4	RC_{all}
PSIS	0.923	0.261	5	3.693	0.920	0.923	0.916	0.933	0.713
FAST	0.942	0.246	5	3.770	0.940	0.953	0.933	0.943	0.780
IFAST	0.988	0.209	5	3.953	0.983	0.986	0.993	0.990	0.953
CRIS	0.667	0.466	5	2.670	0.653	0.636	0.733	0.646	0.173
CSIRS	0.938	0.249	5	3.753	0.933	0.930	0.956	0.933	0.763
CRSIS	0.913	0.269	5	3.653	0.923	0.920	0.923	0.886	0.666
LASSO	0.996	0.165	5	3.986	0.996	0.996	0.993	1.000	0.986
SPES	0.989	0.208	5	3.956	0.990	0.990	0.993	0.983	0.956

where $\lambda_0(t)$ is an unspecified baseline hazard function, $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^T$ is the p-dimensional covariate vector, and $\boldsymbol{\beta}^*$ is the regression coefficient vector. Let $\lambda_0(t) = 1$ and $Z_1, \ldots, Z_p$ be independent and identically distributed $N(0, 1)$ random variables. $M_0 = \{1, 2, 3, 4\}$, $\boldsymbol{\beta}_{M_0} = (1, 1, 1, 1)^T$ and $\boldsymbol{\beta}_{M_0^c} = \mathbf{0}_{p-4}$. $(n, p) = (150, 1000)$.

Example 6 modifies Example 5 by reducing the absolute regression coefficients by three times. Thus, it is more challenging to recover all the important features. In Example 7, survival times were generated from the popular Cox model, which is misspecified by the additive hazards model (1).

Tables 7 and 8 present the simulation results for Examples 6 and 7. By comparison with Table 5, it can be seen that performances of SPES under Example 6 are worse than those under Example 5, which is within our expectation. However, our proposed means is still the best among all of the methods and has overwhelming advantages over them. We believe that the behavior of SPES could become better as the sample size increases. As for Example 7, IFAST, LASSO and SPES give satisfactory results, while the other methods' behavior is not well based on the results in Table 8. Specifically, CRIS behaves the worst among all the methods. Through this example, we can conclude that our SPES have robustness to some extent.

4 The breast cancer study

Dataset for this study contains 24,885 genes for 295 female patients diagnosed with breast cancer between 1984 and 1995 at the Netherlands Cancer Institute. After initial screening by the Rosetta error model, 4919 genes were filtered out as significantly regulated genes, refer to Annest et al. (2009). The time of interest to event is the survival time. Among the 295 patients, 216 were still alive at the end of follow-up, producing 73.2% censoring rate. Patients' times to death or censoring ranged from 0.05 to 18.3 years, with a median of 7.2 years.

This dataset contains 5 samples with 21 missing gene expressions. Instead of discarding these samples or setting the missing values to be zeros directly, we firstly impute them by the weighted K-nearest neighbor method of Troyanskaya et al. (2001) with $K = 15$. Then, we apply the SPES method alongside the marginal screening methods considered in Sect. 3 to this imputed microarray data with the threshold being $k = [295/(5\log(295))] = 10$. The selected genes by various methods are listed in Table 9. From this table, we can observe that fifty-five unique genes are discovered by various methods in total. Among them, the top six frequently selected genes are NM.001168, NM.001333, Contig38288.RC, D43950, NM.006607 and U96131, which are selected by five, four, four, three, three and three approaches, respectively. In addition, four genes (NM.001333, Contig38288.RC, U96131 and D43950) identified by our SPES are also selected by other methods. Specifically, NM.001333 and Contig38288.RC are selected by 3 methods except our SPES, respectively, while U96131 and D43950 are selected by 2 methods except our SPES, respectively. Our SPES method obtains 6 genes which are missed by any of the other 7 methods. They are NM.020974, Contig56390.RC, NM.000909, Contig46937.RC, NM.000125 and Contig14284.RC. This result may provide new insights for practitioners to understand the effects of these 6 genes on the survival time of patients with breast cancer.

By consulting the literature, we discovered that Contig38288.RC had been identified as a predictive gene in van't Veer et al. (2002). After screening, our proposed SPES ranks Contig38288.RC at the second position, which is the highest ranking among the eight approaches. It means that it is easier to discover this informative gene by utilizing our proposed SPES method.

To further reduce the number of genes to a smaller size, we use adaptive LASSO with generalized cross-validation criteria of Leng and Ma (2007). Finally, we fitted the additive hazards model (1) with each group of adaptive LASSO selected genes. The final selected genes and estimated regression coefficients from finally fitted model are given in Table 10. It is easy to see that after utilizing adaptive LASSO, SPES, CRIRS, CRSIS and LASSO identify the gene Contig38288.RC and the proposed SPES procedure with adaptive LASSO established the most parsimonious model.

5 Discussion

In this paper, we proposed a joint feature screening procedure for the sparse additive hazards model based on the sparsity-restricted pseudo-score estimator, which could be solved approximately via the IHT algorithm. We also established the sure screening

Table 9 The names of selected genes by various methods

PSIS	FAST	IFAST	CRIS	CSIRS	CRSIS	LASSO	SPES
Contig29100.RC	NM.003981	NM.003258	Contig45588.RC	U96131	NM.001605	NM.001605	NM.001333
NM.007042	NM.004701	NM.003430	NM.001168	D43950	Contig58368.RC	NM.002811	Contig38288.RC
D87078	NM.003258	NM.003981	NM.002106	M96577	NM.001333	Contig48270.RC	U96131
NM.000286	NM.016359	AL110226	Contig48270.RC	Contig38288.RC	NM.018410	D43950	NM.020974
NM.016479	NM.001168	NM.006622	NM.006579	D14678	NM.006607	NM.006607	D43950
AF148505	NM.004217	Contig3359.RC	NM.001124	Contig57584.RC	Contig38288.RC	Contig38288.RC	Contig56390.RC
Contig59134.RC	NM.013277	NM.016359	NM.006623	NM.004701	NM.001809	U96131	NM.000909
Contig34634.RC	M96577	Contig41383.RC	Contig55725.RC	NM.001168	NM.020142	D14678	Contig46937.RC
AL137615	NM.006607	NM.018181	NM.014575	NM.001333	Contig31288.RC	NM.001168	NM.000125
NM.012110	Contig57584.RC	NM.001168	NM.016109	D38553	NM.000270	NM.001333	Contig14284.RC

The genes are ranked by the order of screening result

Table 10 The names of selected genes by various methods after ALASSO and regression coefficient estimators and corresponding standard deviations of fitted additive hazards model

PSIS		FAST		IFAST		CRIS	
GENE	EST.	GENE	EST.	GENE	EST.	GENE	EST.
Contig29100.RC	− 10.024	NM.003258	1.906	NM.003258	5.210	Contig48270.RC	9.234
NM.007042	− 11.280	NM.016359	2.605	NM.003430	− 5.157	NM.006579	7.322
NM.000286	− 5.452	NM.004217	1.845	AL110226	2.202	Contig55725.RC	3.467
NM.016479	5.187	M96577	2.499	NM.006622	− 6.252	NM.016109	4.563
AF148505	− 11.322	NM.006607	2.813	Contig3359.RC	2.727	–	–
Contig34634.RC	− 4.465	Contig57584.RC	1.888	NM.016359	7.306	–	–
AL137615	− 3.265	–	–	Contig41383.RC	− 6.790	–	–
NM.012110	5.815	–	–	NM.018181	6.227	–	–
CSIRS		CRSIS		LASSO		SPES	
GENE	EST.	GENE	EST.	GENE	EST.	GENE	EST.
D43950	6.435	NM.001605	4.694	NM.001605	4.482	NM.001333	4.839
Contig38288.RC	4.040	Contig58368.RC	23.409	NM.002811	4.059	Contig38288.RC	5.224
D14678	4.505	NM.001333	3.572	Contig48270.RC	6.446	D43950	7.885
NM.001333	4.051	NM.006607	1.952	D43950	3.397	–	–
–	–	Contig38288.RC	1.992	Contig38288.RC	2.844	–	–
–	–	NM.020142	− 12.598	NM.001333	4.766	–	–
–	–	Contig31288.RC	4.859	–	–	–	–
–	–	NM.000270	2.779	–	–	–	–

The genes are ranked by the order of screening result. Estimators of regression coefficients (denoted by EST.) are presented by $\times 100$.

property of the proposed screening procedure, which are verified through our simulation study. It is shown that the SPES approach behaves well under various situation and better than them in some complex circumstances by comparing with the main existing screening methods for ultra-high-dimensional survival data. Besides the ability of identifying all the important features, we also found that our method could achieve the purpose of simultaneous feature screening and coefficient estimation.

Although we address the question of feature screening by the IHT algorithm for the additive hazards model in this article, there are still some problems worthy of further research. We have found that the sparsity-restricted pseudo-score estimator is very accurate estimator of regression coefficient in our limited numerical studies. Thus, it is very interesting to explore convergence rate and error bound for sparsity-restricted pseudo-score estimator in the high-dimensional sparse additive hazards model. Although Zhang and Zhang (2012) investigated the error bound of l_0 penalized least squares estimate for complete data, it is more challenging to study it for the approximate solution from IHT algorithm under the right censored data. We will explore this problem in the near future.

Acknowledgements Chen's research was supported by the National Natural Science Foundation of China (11501573, 11326184 and 11771250) and and National Social Science Foundation of China (17BTJ019). Liu's research was supported by the Fundamental Research Funds for the Central Universities (17CX02035A). Wang's research was supported by the National Natural Science Foundation of China (General program 11171331, Key program 11331011 and program for Creative Research Group in China 61621003), a grant from the Key Lab of Random Complex Structure and Data Science, CAS and a grant from Zhejiang Gongshang University.

Appendix: Proofs of the theorems

To prove Theorem 1, we firstly give a large deviation result for martingale under the additive hazards model. This result could be proved along the similar line as those for Theorem 3.1 in Bradic et al. (2011). So we omit the proof here for simplicity.

Lemma 1 *Under Assumptions 1–3, for any positive sequence* $\{u_n\}$ *bounded away from zero, if* $\max_{1\le j\le p} \sigma_j^2 = O(u_n)$, *there exist positive constants* c_7 *and* c_8 *such that*

$$\mathrm{pr}(|\sqrt{n}U_{n,j}(\boldsymbol{\beta}^*)| > u_n) \le c_7 \exp(-c_8 u_n)$$

uniformly over j, *where* $U_{n,j}(\boldsymbol{\beta}^*)$ *is the* j*th component of* $\boldsymbol{U}_n(\boldsymbol{\beta}^*)$.

This large deviation result represents a uniform, nonasymptotic exponential inequality for martingales under the additive hazards model and is independent of dimensionality p. *So it will be very useful for the high-dimensional additive hazards model.*

Proof of Theorem 1 Denote $\hat{\boldsymbol{\beta}}_M$ to be the (unrestricted) pseudo-score estimator of $\boldsymbol{\beta}$ based on model M. In order to establish the sure screening property, we just need to prove

$$\mathrm{pr}\big(\hat{M} \in \mathbf{M}_+^k\big) \longrightarrow 1,$$

as n goes to ∞. It suffices to show

$$\mathrm{pr}\left(\max_{M\in\mathbf{M}_-^k} L_n(\hat{\boldsymbol{\beta}}_M) \geq \min_{M\in\mathbf{M}_+^k} L_n(\hat{\boldsymbol{\beta}}_M)\right) \longrightarrow 0,$$

as n goes to ∞.

For any $M \in \mathbf{M}_-^k$, let $M' = M \cup M_0 \in \mathbf{M}_+^{2k}$.

Firstly, consider $\boldsymbol{\beta}_{M'}$ being close to $\boldsymbol{\beta}^*_{M'}$ such that $\|\boldsymbol{\beta}_{M'} - \boldsymbol{\beta}^*_{M'}\|_2 = c_2 n^{-\tau_1}$. After some algebraic manipulations, we have

$$\begin{aligned} &L_n(\boldsymbol{\beta}_{M'}) - L_n(\boldsymbol{\beta}^*_{M'}) \\ &= (\boldsymbol{\beta}_{M'} - \boldsymbol{\beta}^*_{M'})^T \boldsymbol{U}_{n,M'}(\boldsymbol{\beta}^*_{M'}) - \frac{1}{2}(\boldsymbol{\beta}_{M'} - \boldsymbol{\beta}^*_{M'})^T \boldsymbol{V}_{n,M'}(\boldsymbol{\beta}_{M'} - \boldsymbol{\beta}^*_{M'}). \end{aligned}$$

Then, by the Cauchy–Schwartz inequality and Assumption 5, we can conclude that

$$\begin{aligned} &L_n(\boldsymbol{\beta}_{M'}) - L_n(\boldsymbol{\beta}^*_{M'}) \\ &\leq \|\boldsymbol{\beta}_{M'} - \boldsymbol{\beta}^*_{M'}\|_2 \|\boldsymbol{U}_{n,M'}(\boldsymbol{\beta}^*_{M'})\|_2 - \frac{c_4}{2}\|\boldsymbol{\beta}_{M'} - \boldsymbol{\beta}^*_{M'}\|_2^2 \\ &\leq c_2 n^{-\tau_1} \|\boldsymbol{U}_{n,M'}(\boldsymbol{\beta}^*_{M'})\|_2 - \frac{1}{2} c_4 c_2^2 n^{-2\tau_1}. \end{aligned}$$

Thus, we have

$$\begin{aligned} &\mathrm{pr}\left(L_n\left(\boldsymbol{\beta}_{M'}\right) - L_n\left(\boldsymbol{\beta}^*_{M'}\right) \geq 0\right) \\ &\leq \mathrm{pr}\left(\|\boldsymbol{U}_{n,M'}\left(\boldsymbol{\beta}^*_{M'}\right)\|_2 \geq \frac{1}{2} c_2 c_4 n^{-\tau_1}\right) \\ &\leq \sum_{j\in M'} \mathrm{pr}\left(U_{n,j}^2\left(\boldsymbol{\beta}^*_{M'}\right) \geq \frac{1}{2k}\left(\frac{1}{2} c_2 c_4 n^{-\tau_1}\right)^2\right) \\ &= \sum_{j\in M'} \mathrm{pr}\left(|U_{n,j}\left(\boldsymbol{\beta}^*_{M'}\right)| \geq \left(\frac{1}{2k}\right)^{1/2} \frac{1}{2} c_2 c_4 n^{-\tau_1}\right), \end{aligned}$$

where the second inequality is obtained by Bonferroni inequality.

Because $M_0 \subset M'$, we can get that $U_{n,j}(\boldsymbol{\beta}^*_{M'}) = U_{n,j}(\boldsymbol{\beta}^*)$. Then under the conditions in Theorem 1 and by Lemma 1, we have

$$\begin{aligned} &\mathrm{pr}\left(|U_{n,j}\left(\boldsymbol{\beta}^*_{M'}\right)| \geq \left(\frac{1}{2k}\right)^{1/2} \frac{1}{2} c_2 c_4 n^{-\tau_1}\right) \\ &= \mathrm{pr}\left(|\sqrt{n} U_{n,j}\left(\boldsymbol{\beta}^*_{M'}\right)| \geq \left(\frac{n}{2k}\right)^{1/2} \frac{1}{2} c_2 c_4 n^{-\tau_1}\right) \\ &= \mathrm{pr}\left(|\sqrt{n} U_{n,j}\left(\boldsymbol{\beta}^*_{M'}\right)| \geq \frac{1}{2\sqrt{2}} c_2 c_4 c_3^{-\frac{1}{2}} n^{\frac{1}{2}-\tau_1-\frac{\tau_2}{2}}\right) \\ &\leq c_7 \exp(-c_8 n^{\frac{1}{2}-\tau_1-\frac{\tau_2}{2}}). \end{aligned}$$

Then

$$\mathrm{pr}(L_n(\boldsymbol{\beta}_{M'}) - L_n(\boldsymbol{\beta}^*_{M'}) \geq 0) \leq 2kc_7 \exp\left(-c_8 n^{\frac{1}{2}-\tau_1-\frac{\tau_2}{2}}\right).$$

Then, by the Bonferroni inequality and assumptions in Theorem 1, we can arrive at

$$\begin{aligned}
&\mathrm{pr}\left(\max_{M\in\mathbf{M}_-^k} L_n\left(\boldsymbol{\beta}_{M'}\right) \geq L_n\left(\boldsymbol{\beta}^*_{M'}\right)\right) \\
&\leq \sum_{M\in\mathbf{M}_-^k} \mathrm{pr}\left(L_n\left(\boldsymbol{\beta}_{M'}\right) \geq L_n\left(\boldsymbol{\beta}^*_{M'}\right)\right) \\
&\leq p^k 2kc_7 \exp\left(-c_8 n^{\frac{1}{2}-\tau_1-\frac{\tau_2}{2}}\right) \\
&\leq 2c_7 \exp\left(\log(c_3) + \tau_2 \log(n) + c_9 n^{m+\tau_2} - c_8 n^{\frac{1}{2}-\tau_1-\frac{\tau_2}{2}}\right) \\
&= o(1),
\end{aligned}$$

where c_9 is a positive constant.

By the concavity of $L_n(\boldsymbol{\beta}_{M'})$, we can conclude that the above result holds for any $\boldsymbol{\beta}_{M'}$ that $\|\boldsymbol{\beta}_{M'} - \boldsymbol{\beta}^*_{M'}\| \geq c_2 n^{-\tau_1}$.

For any $M \in \mathbf{M}_-^k$, let $\breve{\boldsymbol{\beta}}_{M'}$ being $\hat{\boldsymbol{\beta}}_M$ augmented with zeros corresponding to the elements in M'/M_0. It is easy to see that $\|\breve{\boldsymbol{\beta}}_{M'} - \boldsymbol{\beta}^*_{M'}\| \geq \|\boldsymbol{\beta}^*_{M_0/M}\| \geq c_2 n^{-\tau_1}$. So we have

$$\begin{aligned}
&\mathrm{pr}\left(\max_{M\in\mathbf{M}_-^k} L_n(\hat{\boldsymbol{\beta}}_M) \geq \min_{M\in\mathbf{M}_+^k} L_n(\hat{\boldsymbol{\beta}}_M)\right) \\
&= \mathrm{pr}\left(\max_{M\in\mathbf{M}_-^k} L_n\left(\breve{\boldsymbol{\beta}}_{M'}\right) \geq \min_{M\in\mathbf{M}_+^k} L_n(\hat{\boldsymbol{\beta}}_M)\right) \\
&\leq \mathrm{pr}\left(\max_{M\in\mathbf{M}_-^k} L_n\left(\breve{\boldsymbol{\beta}}_{M'}\right) \geq L_n\left(\boldsymbol{\beta}^*_{M'}\right)\right) \\
&= o(1).
\end{aligned}$$

Then the proof is finished. □

Proof of Theorem 2 Denote $Q_n(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(t)}) = L_n(\hat{\boldsymbol{\beta}}^{(t)}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(t)})^T \boldsymbol{U}_n(\hat{\boldsymbol{\beta}}^{(t)}) - \frac{u}{2}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(t)}\|_2^2$. Then $\hat{\boldsymbol{\beta}}^{(t+1)} = \mathrm{argmin}_{\boldsymbol{\beta}\in\mathcal{B}(k)}\{-Q_n(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(t)})\}$.

After some algebraic manipulations, it is easy to see that

$$\begin{aligned}
&L_n\left(\hat{\boldsymbol{\beta}}^{(t)}\right) \\
&= Q_n\left(\hat{\boldsymbol{\beta}}^{(t)} | \hat{\boldsymbol{\beta}}^{(t)}\right) \\
&\leq Q_n\left(\hat{\boldsymbol{\beta}}^{(t+1)} | \hat{\boldsymbol{\beta}}^{(t)}\right)
\end{aligned}$$

$$
\begin{aligned}
&= L_n\left(\hat{\boldsymbol{\beta}}^{(t)}\right) + \left(\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\right)^T \boldsymbol{U}_n(\hat{\boldsymbol{\beta}}^{(t)}) - \frac{u}{2}\|\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\|_2^2 \\
&= L_n\left(\hat{\boldsymbol{\beta}}^{(t+1)}\right) - \frac{u}{2}\|\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\|_2^2 + \left(\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\right)^T \boldsymbol{U}_n\left(\hat{\boldsymbol{\beta}}^{(t)}\right) \\
&\quad + L_n\left(\hat{\boldsymbol{\beta}}^{(t)}\right) - L_n\left(\hat{\boldsymbol{\beta}}^{(t+1)}\right) \\
&= L_n\left(\hat{\boldsymbol{\beta}}^{(t+1)}\right) - \frac{u}{2}\|\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\|_2^2 + \frac{1}{2}\left(\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\right)^T \boldsymbol{V}_n\left(\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\right).
\end{aligned}
$$

It is easy to see that

$$
\left(\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\right)^T \boldsymbol{V}_n\left(\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\right) \le \rho_{\max}(\boldsymbol{V}_n)\|\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\|_2^2.
$$

So under the assumptions in Theorem 2, we have

$$
\begin{aligned}
&L_n\left(\hat{\boldsymbol{\beta}}^{(t)}\right) \\
&\le L_n\left(\hat{\boldsymbol{\beta}}^{(t+1)}\right) - \frac{u}{2}\|\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\|_2^2 + \frac{1}{2}\rho_{\max}(\boldsymbol{V}_n)\|\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\|_2^2 \\
&= L_n\left(\hat{\boldsymbol{\beta}}^{(t+1)}\right) - \frac{1}{2}(u - \rho_{\max}(\boldsymbol{V}_n))\|\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\|_2^2 \\
&\le L_n\left(\hat{\boldsymbol{\beta}}^{(t+1)}\right).
\end{aligned}
$$

This ends up the proof. □

Before presenting the proof of Theorem 2, let's introduce a lemma firstly.

Lemma 2 *Define* $\hat{\boldsymbol{\beta}}^{(0)} = \text{argmax}_{\boldsymbol{\beta}}\{L_n(\boldsymbol{\beta}) - \lambda\|\boldsymbol{\beta}\|_1\}$, *where* λ *satisfies* $\lambda n^{\frac{1}{2}-m} \to \infty$, $\lambda n^{\tau_1+\tau_2} \to 0$. *Under Assumptions 1–3 and 6, if* $\max_{1\le j\le p} \sigma_j^2 = O(\lambda n^{\frac{1}{2}})$, *we have*

$$
\text{pr}\left(\|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_1 \le 8c_5^{-1}\lambda q\right) \to 1,
$$

where c_5 *is defined in Assumption 6.*

Proof It is easy to see that

$$
L_n\left(\hat{\boldsymbol{\beta}}^{(0)}\right) - \lambda\|\hat{\boldsymbol{\beta}}^{(0)}\|_1 - \left(L_n(\boldsymbol{\beta}^*) - \lambda\|\boldsymbol{\beta}^*\|_1\right) \ge 0,
$$

or equivalently

$$
L_n(\boldsymbol{\beta}^*) - L_n(\hat{\boldsymbol{\beta}}^{(0)}) \le \lambda\|\boldsymbol{\beta}^*\|_1 - \lambda\|\hat{\boldsymbol{\beta}}^{(0)}\|_1.
$$

Define $\boldsymbol{\delta} = (\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*) = (\delta_1, \ldots, \delta_p)^T$. By some algebraic manipulations, we have

$$\begin{aligned}
&L_n(\hat{\boldsymbol{\beta}}^{(0)}) - L_n(\boldsymbol{\beta}^*) \\
&= \boldsymbol{b}_n^T \hat{\boldsymbol{\beta}}^{(0)} - \frac{1}{2}\hat{\boldsymbol{\beta}}^{(0)T} \boldsymbol{V}_n \hat{\boldsymbol{\beta}}^{(0)} - \left\{ \boldsymbol{b}_n^T \boldsymbol{\beta}^* - \frac{1}{2}\boldsymbol{\beta}^{*T} \boldsymbol{V}_n \boldsymbol{\beta}^* \right\} \\
&= (\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*)^T \left\{ \boldsymbol{b}_n - \frac{1}{2}\boldsymbol{V}_n \boldsymbol{\beta}^* \right\} - \frac{1}{2}(\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*)^T \boldsymbol{V}_n (\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*) \\
&= \boldsymbol{\delta}^T U_n(\boldsymbol{\beta}^*) - \frac{1}{2}\boldsymbol{\delta}^T \boldsymbol{V}_n \boldsymbol{\delta}.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\boldsymbol{\delta}^T \boldsymbol{V}_n \boldsymbol{\delta} \\
&= 2\boldsymbol{\delta}^T \boldsymbol{U}_n(\boldsymbol{\beta}^*) + L_n(\boldsymbol{\beta}^*) - L_n(\hat{\boldsymbol{\beta}}^{(0)}) \\
&\leq 2\boldsymbol{\delta}^T \boldsymbol{U}_n(\boldsymbol{\beta}^*) + \lambda \|\boldsymbol{\beta}^*\|_1 - \lambda \|\hat{\boldsymbol{\beta}}^{(0)}\|_1.
\end{aligned}$$

Denote $\mathcal{A} = \{\max_{1 \leq j \leq p} |U_{n,j}(\boldsymbol{\beta}^*)| \leq \frac{\lambda}{4}\}$. Because $\max_{1 \leq j \leq p} \sigma_j^2 = O(\lambda n^{\frac{1}{2}})$, then by Lemma 1, we have

$$\begin{aligned}
&\mathrm{pr}(\mathcal{A}^c) \\
&\leq \sum_{j=1}^{p} \mathrm{pr}\left(|U_{n,j}(\boldsymbol{\beta}^*)| > \frac{\lambda}{4} \right) \\
&= \sum_{j=1}^{p} \mathrm{pr}\left(|\sqrt{n} U_{n,j}(\boldsymbol{\beta}^*)| > \frac{\sqrt{n}\lambda}{4} \right) \\
&\leq p c_7 \exp\left(-c_8 \frac{\sqrt{n}\lambda}{4} \right) \\
&\leq c_7 \exp\left(c_{10} n^m - c_8 \frac{\sqrt{n}\lambda}{2} \right) \\
&\to 0,
\end{aligned}$$

where c_{10} is a positive constant. So we obtain that $\mathrm{pr}(\mathcal{A}) \to 1$ and $\|\boldsymbol{U}_n(\boldsymbol{\beta}^*)\|_\infty = O_p(\lambda)$. Under the event $\mathcal{A}$, it is easy to see that

$$\begin{aligned}
&\boldsymbol{\delta}^T \boldsymbol{V}_n \boldsymbol{\delta} \\
&\leq \frac{1}{2}\lambda \|\boldsymbol{\delta}\|_1 + \lambda \|\boldsymbol{\beta}^*\|_1 - \lambda \|\hat{\boldsymbol{\beta}}^{(0)}\|_1.
\end{aligned}$$

Thus

$$\boldsymbol{\delta}^T \boldsymbol{V}_n \boldsymbol{\delta} + \frac{1}{2}\lambda \|\boldsymbol{\delta}\|_1$$

$$
\begin{aligned}
&\le \lambda\|\boldsymbol{\delta}\|_1 + \lambda\|\boldsymbol{\beta}^*\|_1 - \lambda\|\hat{\boldsymbol{\beta}}^{(0)}\|_1 \\
&\le \lambda \sum_{j=1}^{p} \left(|\hat{\beta}_j^{(0)} - \beta_j^*| + |\beta_j^*| - |\hat{\beta}_j^{(0)}|\right) \\
&= \lambda \sum_{j\in M_0} \left(|\hat{\beta}_j^{(0)} - \beta_j^*| + |\beta_j^*| - |\hat{\beta}_j^{(0)}|\right) \\
&\le 2\lambda \sum_{j\in M_0} |\delta_j| \\
&\le 2\lambda\|\boldsymbol{\delta}_{M_0}\|_1.
\end{aligned}
$$

It is easy to see that $\boldsymbol{V}_n$ is semipositive definite. Thus $\|\boldsymbol{\delta}\|_1 \le 4\|\boldsymbol{\delta}_{M_0}\|_1$, and furthermore $\|\boldsymbol{\delta}_{M_0^c}\|_1 \le 3\|\boldsymbol{\delta}_{M_0}\|_1$. By the Cauchy–Schwarz inequality and Assumption 6,

$$
\|\boldsymbol{\delta}_{M_0}\|_1^2 \le q\|\boldsymbol{\delta}_{M_0}\|_2^2 \le q c_5^{-1}\boldsymbol{\delta}^T \boldsymbol{V}_n\boldsymbol{\delta} \le 2c_5^{-1}\lambda q\|\boldsymbol{\delta}_{M_0}\|_1.
$$

So $\|\boldsymbol{\delta}_{M_0}\|_1 \le 2c_5^{-1}\lambda q$. Then finally we arrive at

$$
\|\boldsymbol{\delta}\|_1 = \|\boldsymbol{\delta}_{M_0^c}\|_1 + \|\boldsymbol{\delta}_{M_0}\|_1 \le 4\|\boldsymbol{\delta}_{M_0}\|_1 \le 8c_5^{-1}\lambda q.
$$

This finishes the proof. □

Proof of Theorem 3 Recall that $w = \min_{j\in M_0}\|\beta_j^*\|$. We just need to show $\mathrm{pr}(\|\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|_\infty < \frac{w}{2}) \to 1$. It suffices to prove $\|\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|_\infty = o_p(w)$. As in Xu and Chen (2014), we use the method of mathematical induction to get this result.

When $t = 0$, by Lemma 2, we have

$$
\mathrm{pr}(\|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_1 \le 8c_5^{-1}\lambda q) \to 1.
$$

Because $\lambda = o(n^{-(\tau_1+\tau_2)})$, $q = O(n^{\tau_2})$, $w^{-1} = O(n^{\tau_1})$, $\lambda q w^{-1} = o(n^{-(\tau_1+\tau_2)})$ $O(n^{\tau_2})O(n^{\tau_1}) = o(1)$. Thus $\lambda q = o(w)$. So we have $\|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_1 = o_p(w)$. It is noted that $\|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_\infty \le \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_1$. Then the desired result is obtained for $t = 0$.

Suppose that $\|\hat{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty = o_p(w)$. In the following, we will show that $\|\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|_\infty = o_p(w)$ is also true. From the adaptive iterative hard-thresholding algorithm, it is noted that $\hat{\boldsymbol{\beta}}^{(t)} = \mathbf{H}(\tilde{\boldsymbol{\beta}}^{(t-1)}; k)$, where $\tilde{\boldsymbol{\beta}}^{(t-1)} = \hat{\boldsymbol{\beta}}^{(t-1)} + u^{-1}\dot{L}_n(\hat{\boldsymbol{\beta}}^{(t-1)})$. If $\|\tilde{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty = o_p(w)$ holds, it can be seen that elements of $\tilde{\boldsymbol{\beta}}_{M_0}^{(t-1)}$ are among the ones with top k largest absolute values in probability. Thus $\|\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|_\infty \le \|\tilde{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty = o_p(w)$. So what remains is to prove $\|\tilde{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty = o_p(w)$. Note that $\|\tilde{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty \le \|\hat{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty + u^{-1}\|\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}^{(t-1)})\|_\infty$. By some algebraic

manipulations, we could obtain that

$$\begin{aligned}
&\|\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}^{(t-1)})\|_\infty \\
&= \|\boldsymbol{U}_n(\boldsymbol{\beta}^*) + \boldsymbol{U}_n(\hat{\boldsymbol{\beta}}^{(t-1)}) - \boldsymbol{U}_n(\boldsymbol{\beta}^*)\|_\infty \\
&\leq \|\boldsymbol{U}_n(\boldsymbol{\beta}^*)\|_\infty + \|\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}^{(t-1)}) - \boldsymbol{U}_n(\boldsymbol{\beta}^*)\|_\infty \\
&= \|\boldsymbol{U}_n(\boldsymbol{\beta}^*)\|_\infty + \|\boldsymbol{V}_n(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{(t-1)})\|_\infty \\
&\leq \|\boldsymbol{U}_n(\boldsymbol{\beta}^*)\|_\infty + \|\boldsymbol{V}_n\|_\infty \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{(t-1)}\|_\infty.
\end{aligned}$$

Thus

$$\begin{aligned}
&u^{-1}\|\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}^{(t-1)})\|_\infty \\
&\leq u^{-1}O_p(\lambda) + u^{-1}\|\boldsymbol{V}_n\|_\infty o_p(w) \\
&\leq (c_6 r)^{-1}\lambda O_p(1) + (c_6 r)^{-1} n^{\tau_3} o_p(w) O_p(1) \\
&= c_6^{-1} O(n^{-\tau_3}) o(n^{-(\tau_1+\tau_2)}) O_p(1) + c_6^{-1} O_p(n^{-\tau_3}) n^{\tau_3} o_p(w) \\
&= o_p(w).
\end{aligned}$$

This ends up the proof. □

References

Annest, A., Bumgarner, R., Raftery, A., Yeung, K. (2009). Iterative Bayesian model averaging: A method for the application of survival analysis to high-dimensional microarray data. *BMC Bioinformatics*, *10*, 72.

Bertsekas, D. (2016). *Nonlinear programming* (3rd ed.). Nashua: Athena Scientific.

Bickel, P., Ritov, Y., Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, *37*, 1705–1732.

Bradic, J., Fan, J., Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *The Annals of Statistics*, *39*, 3092–3120.

Cai, J., Fan, J., Li, R., Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika*, *92*, 303–316.

Chang, J., Tang, C., Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *The Annals of Statistics*, *41*, 2123–2148.

Chen, X. (2018). Model-free conditional feature screening for ultra-high dimensional right censored data. *Journal of Statistical Computation and Simulation*. https://doi.org/10.1080/00949655.2018.1466142.

Chen, X., Chen, X., Liu, Y. (2017). A note on quantile feature screening via distance correlation. *Statistical Papers*. https://doi.org/10.1007/s00362-017-0894-8.

Chen, X., Chen, X., Wang, H. (2018). Robust feature screening for ultra-high dimensional right censored data via distance correlation. *Computational Statistics and Data Analysis*, *119*, 118–138.

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

Fan, J., Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of Royal Statistical Society, Series B*, *70*, 849–911.

Fan, J., Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, *38*, 3567–3604.

Fan, J., Feng, Y., Wu, Y. (2010). Ultrahigh dimensional variable selection for Cox's proportional hazards model. *Institute of Mathematical Statistics Collections*, *6*, 70–86.

Fan, J., Samworth, R., Wu, Y. (2009). Ultrahigh dimensional variable selection: Beyond the linear model. *Journal of Machine Learning Research*, *10*, 1829–1853.
Fan, J., Ma, Y., Dai, W. (2014). Nonparametric independent screening in sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association*, *109*, 1270–1284.
Gorst-Rasmussen, A., Scheike, T. (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of Royal Statistical Society, Series B*, *75*, 217–245.
He, X., Wang, L., Hong, H. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics*, *41*, 342–369.
Huang, J., Sun, T., Ying, Z., Yu, Y., Zhang, C. (2013). Oracle inequalities for the lasso in the Cox model. *The Annals of Statistics*, *41*, 1142–1165.
Leng, C., Ma, S. (2007). Path consistent model selection in additive risk model via lasso. *Statistics in Medicine*, *26*, 3753–3770.
Li, G., Peng, H., Zhang, J., Zhu, L. (2012a). Robust rank correlation based screening. *The Annals of Statistics*, *40*, 1846–1877.
Li, R., Zhong, W., Zhu, L. (2012b). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, *107*, 1129–1139.
Lin, D., Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, *81*, 61–71.
Lin, W., Lv, J. (2013). High-dimensional sparse additive hazards regression. *Journal of the American Statistical Association*, *108*, 247–264.
Liu, Y., Chen, X. (2018). Quantile screening for ultra-high-dimensional heterogeneous data conditional on some variables. *Journal of Statistical Computation and Simulational*, *88*, 329–342.
Martinussen, T., Scheike, T. (2009). The additive hazards model with high-dimensional regressors. *The Annals of Statistics*, *15*, 330–342.
Song, R., Lu, W., Ma, S., Jessie Jeng, X. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika*, *101*, 799–814.
Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, *16*, 385–395.
Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., et al. (2001). Missing value estimation methods for DNA microarray. *Bioinformatics*, *17*, 520–525.
van't Veer, L., Dai, H., van de Vijver, M., He, Y., Hart, A., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*, 530–536.
Wu, Y., Yin, G. (2015). Conditional quantile screening in ultrahigh-dimensional heterogeneous data. *Biometrika*, *102*, 65–76.
Xu, C., Chen, J. (2014). The sparse MLE for ultra-high-dimensional feature screening. *Journal of the American Statistical Association*, *109*, 1257–1269.
Yang, G., Yu, Y., Li, R., Buu, A. (2016). Feature screening in ultrahigh dimensional Cox's model. *Statistics Sinica*, *26*, 881–901.
Yang, G., Hou, S., Wang, L., Sun, Y. (2018). Feature screening in ultrahigh-dimensional additive Cox model. *Journal of Statistical Computation and Simulation*, *88*, 1117–1133.
Zhang, C., Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, *27*, 576–593.
Zhang, J., Liu, Y., Wu, Y. (2017). Correlation rank screening for ultrahigh-dimensional survival data. *Computational Statistics & Data Analysis*, *108*, 121–132.
Zhao, S., Li, Y. (2012). Principled sure independence screening for Cox model with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, *105*, 397–411.
Zhao, S., Li, Y. (2014). Score test variable screening. *Biometrics*, *70*, 862–871.
Zhou, T., Zhu, L. (2017). Model-free features screening for ultrahigh dimensional censored regression. *Statistics and Computing*, *27*, 947–961.
Zhu, L., Li, L., Li, R., Zhu, L. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, *106*, 1464–1475.