IMS Collections
Borrowing Strength: Theory Powering Applications –
A Festschrift for Lawrence D. Brown
Vol. 6 (2010) 70–86

DOI: 10.1214/10-IMSCOLL606

High-dimensional variable selection for Cox's proportional hazards model

Jianqing Fan[1,*], Yang Feng[1] and Yichao Wu[2,†]

Princeton University and North Carolina State University

Abstract: Variable selection in high dimensional space has challenged many contemporary statistical problems from many frontiers of scientific disciplines. Recent technological advances have made it possible to collect a huge amount of covariate information such as microarray, proteomic and SNP data via bioimaging technology while observing survival information on patients in clinical studies. Thus, the same challenge applies in survival analysis in order to understand the association between genomics information and clinical information about the survival time. In this work, we extend the sure screening procedure [6] to Cox's proportional hazards model with an iterative version available. Numerical simulation studies have shown encouraging performance of the proposed method in comparison with other techniques such as LASSO. This demonstrates the utility and versatility of the iterative sure independence screening scheme.

Contents

*Partially supported by the NSF grants DMS-0704337, DMS-0714554 and NIH grant R01-GM072611.
†Partially supported by the NSF grant DMS-0905561 and the NIH/NCI grant R01-CA149569.
[1]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, e-mail: jqfan@princeton.edu; yangfeng@princeton.edu
[2]Department of Statistics, North Carolina State University, Raleigh, NC 27695, e-mail: wu@stat.ncsu.edu
Keywords and phrases: Cox's proportional hazards model, variable selection.
AMS 2000 subject classifications: Primary 62N02; secondary 62J99.

1. Introduction

Survival analysis is a commonly-used method for the analysis of failure time such as biological death, mechanical failure, or credit default. Within this context, death or failure is also referred to as an "event". Survival analysis tries to model time-to-event data, which is usually subject to censoring due to the termination of study. The main goal is to study the dependence of the survival time T on covariate variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$, where p denotes the dimensionality of the covariate space. One common way of achieving this goal is hazard regression, which studies how the conditional hazard function of T depends on the covariate $\mathbf{X} = \mathbf{x}$, which is defined as

$$h(t|\mathbf{x}) = \lim_{\triangle t \downarrow 0} \frac{1}{\triangle t} P\{t \le T < t + \triangle t | T \ge t, \mathbf{X} = \mathbf{x}\}.$$

According to the definition, the conditional hazard function is nothing but the instantaneous rate of failure at time t given a particular value $\mathbf{x}$ for the covariate $\mathbf{X}$. The proportional hazards model is very popular, partially due to its simplicity and its convenience in dealing with censoring. The model assumes that

$$h(t|\mathbf{x}) = h_0(t)\Psi(\mathbf{x}),$$

in which $h_0(t)$ is the baseline hazard function and $\Psi(\mathbf{x})$ is the covariate effect. Note that this model is not uniquely determined in that $ch_0(t)$ and $\Psi(\mathbf{x})/c$ give the same model for any $c > 0$. Thus one identifiability condition needs to be specified. When the identifiability condition $\Psi(\mathbf{0}) = 1$ is enforced, the function $h_0(t)$, the conditional hazard function of T given $\mathbf{X} = \mathbf{0}$, is called the baseline hazard function.

By taking the reparametrization $\Psi(\mathbf{x}) = e^{\psi(\mathbf{X})}$, Cox [1, 2] introduced the proportional hazards model

$$h(t|\mathbf{x}) = h_0(t)e^{\psi(\mathbf{X})}.$$

See [11] and references therein for more detailed literature on Cox's proportional hazards model.

Here the baseline hazard function $h_0(t)$ is typically completely unspecified and needs to be estimated nonparametrically. A linear model assumption, $\psi(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}$, may be made, as is done in this paper. Here $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the regression parameter vector. While conducting survival analysis, we not only need to estimate $\boldsymbol{\beta}$ but also have to estimate the baseline hazard function $h_0(t)$ nonparametrically. Interested readers may consult [11] for more details.

Recent technological advances have made it possible to collect a huge amount of covariate information such as microarray, proteomic and SNP data via bioimaging technology while observing survival information on patients in clinical studies. However it is quite likely that not all available covariates are associated with the clinical outcome such as the survival time. In fact, typically a small fraction of covaraites are associated with the clinical time. This is the notion of sparsity and consequently calls for the identification of important risk factors and at the same time quantifying their risk contributions when we analyze time-to-event data with many predictors. Mathematically, it means that we need to identify which β_js are nonzero and also estimate these nonzero β_js.

Most classical model selection techniques have been extended from linear regression to survival analysis. They include the best-subset selection, stepwise selection, bootstrap procedures [14], Bayesian variable selection [9, 10]. Please see references therein. Similarly, other more modern penalization approaches have been extended as well. Tibshirani [15] applied the LASSO penalty to survival analysis. Fan and Li

[5] considered survival analysis with the SCAD and other folded concave penalties. Zhang and Lu [19] proposed the adaptive LASSO penalty while studying time-to-event data. Among many other considerations is Li and Dicker [12]. Available theory and empirical results show that these penalization approaches work well with a moderate number of covariates.

Recently we have seen a surge of interest in variable selection with an ultra-high dimensionality. By ultra-high, Fan and Lv [6] meant that the dimensionality grows exponentially in the sample size, i.e., $\log(p) = O(n^a)$ for some $a \in (0, 1/2)$. For ultra-high dimensional linear regression, Fan and Lv [6] proposed sure independence screening (SIS) based on marginal correlation ranking. Asymptotic theory is proved to show that, with high probability, SIS keeps all important predictor variables with vanishing false selection rate. An important extension, iterative SIS (ISIS), was also proposed to handle difficult cases such as when some important predictors are marginally uncorrelated with the response. In order to deal with more complex real data, Fan, Samworth and Wu [7] extended SIS and ISIS to more general loss based models such as generalized linear models, robust regression, and classification and improved some important steps of the original ISIS. In particular, they proposed the concept of conditional marginal regression and a new variant of the method based on splitting samples. A non-asymptotic theoretical result shows that the splitting based new variant can reduce false discovery rate. Although the extension of Fan, Samworth and Wu [7] covers a wide range of statistical models, it has not been explored whether the iterative sure independence screening method can be extended to hazard regression with censoring event time. In this work, we will focus on Cox's proportional hazards model and extend SIS and ISIS accordingly. Other extensions of SIS include Fan and Song [8] and Fan, Feng and Song [3] to generalized linear models and nonparametric additive models, in which new insights are provided via elegant mathematical results and carefully designed simulation studies.

The rest of the article is organized as follows. Section 2 details the Cox's proportional hazards model. An overview of variable selection via penalized approach is given in Section 3 for Cox's proportional hazards model. We extend the SIS and ISIS procedures to Cox's model in Section 4. Simulation results in Section 5 and real data analysis in Section 6 demonstrate the effectiveness of the proposed SIS and ISIS methods.

2. Cox's proportional hazards models

Let T, C, and $\mathbf{X}$ denote the survival time, the censoring time, and their associated covariates, respectively. Correspondingly, denote by $Y = \min\{T, C\}$ the observed time and $\delta = I(T \le C)$ the censoring indicator. For simplicity we assume that T and C are conditionally independent given $\mathbf{X}$ and that the censoring mechanism is non-informative. Our observed data set $\{(\mathbf{x}_i, y_i, \delta_i) : \mathbf{x}_i \in I\!R^p, y_i \in I\!R^+, \delta_i \in \{0, 1\}, i = 1, 2, \ldots, n\}$ is an independently and identically distributed random sample from a certain population $(\mathbf{X}, Y, \delta)$. Define $\mathcal{C} = \{i : \delta_i = 0\}$ and $\mathcal{U} = \{i : \delta_i = 1\}$ to be the censored and uncensored index sets, respectively. Then the complete likelihood of the observed data set is given by

$$L = \prod_{i \in \mathcal{U}} f(y_i|\mathbf{x}_i) \prod_{i \in \mathcal{C}} \bar{F}(y_i|\mathbf{x}_i) = \prod_{i \in \mathcal{U}} h(y_i|\mathbf{x}_i) \prod_{i=1}^{n} \bar{F}(y_i|\mathbf{x}_i),$$

where $f(t|\mathbf{x})$, $\bar{F}(t|\mathbf{x}) = \int_t^\infty f(s|\mathbf{x})\, ds$, and $h(t|\mathbf{x}) = f(t|\mathbf{x})/\bar{F}(t|\mathbf{x})$ are the conditional density function, the conditional survival function, and the conditional

hazard function of T given $\mathbf{X} = \mathbf{x}$, respectively.

Let $t_1^0 < t_2^0 < \cdots < t_N^0$ be the ordered distinct observed failure times. Let (j) index its associate covariates $\mathbf{x}_{(j)}$ and $\mathcal{R}(t)$ be the risk set right before the time t: $\mathcal{R}(t) = \{i : y_i \geq t\}$. Consider the proportional hazards model,

$$h(t|\mathbf{x}) = h_0(t) \exp(\mathbf{x}^T \boldsymbol{\beta}), \tag{2.1}$$

where $h_0(t)$ is the baseline hazard function. In this model, both $h_0(t)$ and $\boldsymbol{\beta}$ are unknown and have to be estimated. Under model (2.1), the likelihood becomes

$$L = \prod_{j=1}^{N} h_0(y_{(j)}) \exp(\mathbf{x}_{(j)}^T \boldsymbol{\beta}) \prod_{i=1}^{n} \exp\{-H_0(y_i) \exp(\mathbf{x}_i^T \boldsymbol{\beta})\},$$

where $H_0(t) = \int_0^t h_0(s)\, ds$ is the corresponding cumulative baseline hazard function.

Following Breslow's idea, consider the "least informative" nonparametric modeling for $H_0(\cdot)$, in which $H_0(t)$ has a possible jump h_j at the observed failure time t_j^0, namely, $H_0(t) = \sum_{j=1}^{N} h_j I(t_j^0 \leq t)$. Then

$$H_0(y_i) = \sum_{j=1}^{N} h_j I(i \in \mathcal{R}(t_j^0)). \tag{2.2}$$

Consequently the log-likelihood becomes

$$\sum_{j=1}^{N} \{\log(h_j) + \mathbf{x}_{(j)}^T \boldsymbol{\beta}\} - \sum_{i=1}^{n} \left\{ \sum_{j=1}^{N} h_j I(i \in \mathcal{R}(t_j^0)) \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \right\}.$$

Maximizer h_j is given by

$$\hat{h}_j(\boldsymbol{\beta}) = \left\{ \sum_{i \in \mathcal{R}(t_j^0)} \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \right\}^{-1}. \tag{2.3}$$

Putting this maximizer back to the log-likelihood, we get

$$\sum_{j=1}^{N} \left[\mathbf{x}_{(j)}^T \boldsymbol{\beta} - \log \left\{ \sum_{i \in \mathcal{R}(t_j^0)} \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \right\} \right],$$

which is equivalent to

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \mathbf{x}_i^T \boldsymbol{\beta} - \sum_{i=1}^{n} \delta_i \log \left\{ \sum_{j \in \mathcal{R}(y_i)} \exp(\mathbf{x}_j^T \boldsymbol{\beta}) \right\}. \tag{2.4}$$

by using the censoring indicator δ_i. This is the partial likelihood due to [2].

Maximizing $\ell(\boldsymbol{\beta})$ in (2.4) with respect to $\boldsymbol{\beta}$, we can get an estimate $\hat{\boldsymbol{\beta}}$ of the regression parameter. Once $\hat{\boldsymbol{\beta}}$ is available, we may plug it into (2.3) to get $\hat{h}_j(\hat{\boldsymbol{\beta}})$. These newly obtained $\hat{h}_j(\hat{\boldsymbol{\beta}})$s can be plugged into (2.2) to obtain our nonparametric estimate of the baseline cumulative hazard function.

3. Variable selection for Cox's proportional hazards model via penalization

In the estimation scheme presented in the previous section, none of the estimated regression coefficients is exactly zero, leaving all covariates in the final model. Consequently it is incapable of selecting important variables and handling the case with $p > n$. To achieve variable selection, classical techniques such as the best-subset selection, stepwise selection, and bootstrap procedures have been extended accordingly to handle Cox's proportional hazards model.

In this section, we will focus on some more advanced techniques for variable selection via penalization. Variable selection via penalization has received lots of attention recently. Basically it uses some variable selection-capable penalty function to regularize the objective function while performing optimization. Many variable selection-capable penalty functions have been proposed. A well known example is the L_1 penalty $\lambda \sum_{j=1}^{p} |\beta_j|$, which is also known as the LASSO penalty [16]. Among many others are the SCAD penalty [4], elastic-net penalty [22], adaptive L_1 [20, 19], and minimax concave penalty [18].

Denote a general penalty function by $p_\lambda(\cdot)$, where $\lambda > 0$ is a regularization parameter. From derivations in the last section, penalized likelihood is equivalent to penalized partial likelihood: While maximizing $\ell(\boldsymbol{\beta})$ in (2.4), one may regularize it using $\sum_{j=1}^{p} p_\lambda(\beta_j)$. Equivalently we solve

$$\min -\ell(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(\beta_j)$$

by including a negative sign in front of $\ell(\boldsymbol{\beta})$. In the literature, Tibshirani [15], Fan and Li [5], and Zhang and Lu [19] considered the L_1, SCAD, and adaptive L_1 penalties while studying time-to-event data, respectively, among many others.

In this paper, we will use the SCAD penalty for our extensions of SIS and ISIS whenever necessary. The SCAD function is a quadratic spline and symmetric around the origin. It can be defined in terms of its first order derivative

$$p'_\lambda(|\beta|) = \lambda \left\{ \mathbb{1}_{\{|\beta| \le \lambda\}} + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} \mathbb{1}_{\{|\beta| > \lambda\}} \right\},$$

for some $a > 2$ and $\beta \neq 0$. Here a is a parameter and Fan and Li [4] recommend to use $a = 3.7$ based on a Bayesian argument. The SCAD penalty is plotted in Figure 1 for $a = 3.7$ and $\lambda = 2$. The SCAD penalty is non-convex, leading to non-convex optimization. For the non-convex SCAD penalized optimization, Fan and Li [4] proposed the local quadratic approximation; Zou and Li [21] proposed the local linear approximation; Wu and Liu [17] presented the difference convex algorithm. In this work, whenever necessary we use the local linear approximation algorithm to solve the SCAD penalized optimization.

4. SIS and ISIS for Cox's proportional hazard model

The penalty based variable selection techniques work great with a moderate number of covariates. However their usefulness is limited while dealing with an ultra-high dimensionality as shown in Fan and Lv [6]. In the linear regression case, [6] proposed to rank covariates according to the absolute value of their marginal correlation with

the response variable and select the top ranked covariates. They provided theoretical result to guarantee that this simple correlation ranking retains all important covariates with high probability. Thus they named their method sure independence screening (SIS). In order to handle difficult problems such as the one with some important covariates being marginally uncorrelated with the response, they proposed iterative SIS (ISIS). ISIS begins with SIS, then it regresses the response on covariates selected by SIS and uses the regression residual as a "working" response to recruit more covariates with SIS. This process can be repeated until some convergence criterion has been met. Empirical improvement over SIS has been observed for ISIS. In order to increase the power of the sure independence screening technique, Fan, Samworth and Wu [7] has extended SIS and ISIS to more general models such as generalized linear models, robust regression, and classification and made several important improvements. We now extend the key idea of SIS and ISIS to handle Cox's proportional hazards model.

Let $\mathcal{M}^*$ be the index set of the true underlying sparse model, namely, $\mathcal{M}^* = \{j : \beta_j^* \neq 0 \text{ and } 1 \leq j \leq p\}$, where β_j^*s are the true regression coefficients in the Cox's proportional hazards model (2.1).

4.1. Ranking by marginal utility

First let us review the definition of sure screening property.

Definition 1 (Sure screening property). We say a model selection procedure satisfies sure screening property if the selected model $\hat{\mathcal{M}}$ with model size $o_p(n)$ includes the true model $\mathcal{M}^*$ with probability tending to one.

For each covariate X_m $(1 \leq m \leq p)$, define its marginal utility as the maximum of the partial likelihood of the single covariate:

$$u_m = \max_{\beta_m} \left(\sum_{i=1}^{n} \delta_i x_{im} \beta_m - \sum_{i=1}^{n} \delta_i \log \left\{ \sum_{j \in \mathcal{R}(y_i)} \exp(x_{jm} \beta_m) \right\} \right).$$

Here x_{im} is the m-th element of $\mathbf{x}_i$, namely $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$. Intuitively speaking, the larger the marginal utility is, the more information the corresponding

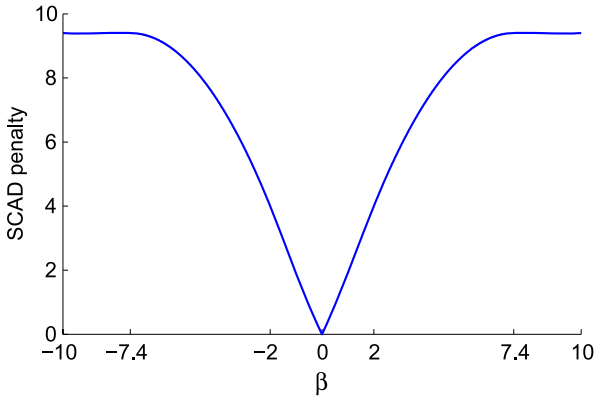

FIG 1. *Plot of the SCAD penalty with* $a = 3.7$ *and* $\lambda = 2$.

covariate contains the information about the survival outcome. Once we have obtained all marginal utilities u_m for $m = 1, 2, \ldots, p$, we rank all covariates according to their corresponding marginal utilities from the largest to the smallest and select the d top ranked covariates. Denote by $\mathcal{I}$ the index set of these d covariates that have been selected.

The index set $\mathcal{I}$ is expected to cover the true index set $\mathcal{M}^*$ with a high probability, especially when we use a relative large d. This is formally shown by Fan and Lv [6] for the linear model with Gaussian noise and Gaussian covariates and significantly expanded by Fan and Song [8] to generalized linear models with non-Gaussian covariates. The parameter d is usually chosen large enough to ensure the sure screening property. However the estimated index set $\mathcal{I}$ may also include a lot of unimportant covariates. To improve performance, the penalization based variable selection approach can be applied to the selected subset of the variables $\{X_j, j \in \mathcal{I}\}$ to further delete unimportant variables. Mathematically, we then solve the following penalized partial likelihood problem:

$$\min_{\boldsymbol{\beta}_{\mathcal{I}}} \left(-\sum_{i=1}^{n} \delta_i \mathbf{x}_{\mathcal{I},i}^T \boldsymbol{\beta}_{\mathcal{I}} + \sum_{i=1}^{n} \delta_i \log\Bigg\{ \sum_{j \in \mathcal{R}(y_i)} \exp(\mathbf{x}_{\mathcal{I},j}^T \boldsymbol{\beta}_{\mathcal{I}}) \Bigg\} + \sum_{m \in \mathcal{I}} p_\lambda(\beta_m) \right),$$

where $\mathbf{x}_{\mathcal{I},i}$ denotes a sub-vector of $\mathbf{x}_i$ with indices in $\mathcal{I}$ and similarly for $\boldsymbol{\beta}_{\mathcal{I}}$. It will lead to sparse regression parameter estimate $\hat{\boldsymbol{\beta}}_{\mathcal{I}}$. Denote the index set of nonzero components of $\hat{\boldsymbol{\beta}}_{\mathcal{I}}$ by $\hat{\mathcal{M}}$, which will serve as our final estimate of $\mathcal{M}^*$.

4.2. Conditional feature ranking and iterative feature selection

Fan and Lv [6] pointed out that SIS can fail badly for some challenging scenarios such as the case that there exist jointly related but marginally unrelated covariates or jointly uncorrelated covariates having higher marginal correlation with the response than some important predictors. To deal with such difficult scenarios, iterative SIS (ISIS) has been proposed. Comparing to SIS which is based on marginal information only, ISIS tries to make more use of joint covariates' information.

The iterative SIS begins with using SIS to select an index set $\hat{\mathcal{I}}_1$, upon which a penalization based variable selection step is applied to get regression parameter estimate $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{I}}_1}$. A refined estimate of the true index set is obtained and denoted by $\hat{\mathcal{M}}_1$, the index set corresponding to nonzero elements of $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{I}}_1}$.

As in [7], we next define the conditional utility of each covariate m that is not in $\hat{\mathcal{M}}_1$ as follows:

$$u_{m|\hat{\mathcal{M}}_1} = \max_{\beta_m, \boldsymbol{\beta}_{\hat{\mathcal{M}}_1}} \left(\sum_{i=1}^{n} \delta_i (x_{im}\beta_m + \mathbf{x}_{\hat{\mathcal{M}}_1,i}^T \boldsymbol{\beta}_{\hat{\mathcal{M}}_1}) - \sum_{i=1}^{n} \delta_i \log\Bigg\{ \sum_{j \in \mathcal{R}(y_i)} \exp(x_{jm}^T \beta_m + \mathbf{x}_{\hat{\mathcal{M}}_1,j}^T \boldsymbol{\beta}_{\hat{\mathcal{M}}_1}) \Bigg\} \right).$$

This conditional utility measures the additional contribution of the mth covariate given that all covariates with indices in $\hat{\mathcal{M}}_1$ have been included in the model.

Once the conditional utilities have been defined for each covariate that is not in $\hat{\mathcal{M}}_1$, we rank them from the largest to the smallest and select these covariates

with top rankings. Denote the index set of these selected covariates by $\hat{\mathcal{I}}_2$. With $\hat{\mathcal{I}}_2$ having been identified, we minimize

$$
\begin{aligned}
&-\sum_{i=1}^{n}\delta_i(\mathbf{x}^T_{\hat{\mathcal{M}}_1\cup\hat{\mathcal{I}}_2,i}\boldsymbol{\beta}_{\hat{\mathcal{M}}_1\cup\hat{\mathcal{I}}_2})+\sum_{i=1}^{n}\delta_i\log\Big\{\sum_{j\in\mathcal{R}(y_i)}\exp(\mathbf{x}^T_{\hat{\mathcal{M}}_1\cup\hat{\mathcal{I}}_2,j}\boldsymbol{\beta}_{\hat{\mathcal{M}}_1\cup\hat{\mathcal{I}}_2})\Big\}\\
&\qquad+\sum_{m\in\hat{\mathcal{M}}_1\cup\hat{\mathcal{I}}_2}p_\lambda(\beta_j)
\end{aligned}
\tag{4.1}
$$

with respect to $\boldsymbol{\beta}_{\hat{\mathcal{M}}_1\cup\hat{\mathcal{I}}_2}$ to get sparse estimate $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}_1\cup\hat{\mathcal{I}}_2}$. Denote the index set corresponding to nonzero components of $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}_1\cup\hat{\mathcal{I}}_2}$ to be $\hat{\mathcal{M}}_2$, which is our updated estimate of the true index set $\mathcal{M}^*$. Note that this step can delete some variables $\{X_j\in\hat{\mathcal{M}}_1\}$ that were previously selected. This idea was proposed in [7] and is an improvement of the idea in [6].

The above iteration can be repeated until some convergence criterion is reached. We adopt the criterion of either having identified d covariates or $\hat{\mathcal{M}}_j=\hat{\mathcal{M}}_{j-1}$ for some j.

4.3. New variants of SIS and ISIS for reducing FSR

Fan, Samworth and Wu [7] noted that the idea of sample spliting can also be used to reduce the false selection rate. Without loss of generality, we assume that the sample size n is even. We randomly split the sample into two halves. Then apply SIS or ISIS separately to the data in each partition to obtain two estimates $\hat{\mathcal{I}}^{(1)}$ and $\hat{\mathcal{I}}^{(2)}$ of the true index set $\mathcal{M}^*$. Both these two estimates could have high FSRs because they are based on a simple and crude screening method. Yet each of them should include all important covariates with high probabilities. Namely, important covariates should appear in both sets with probability tending to one asymptotically. Define a new estimate by intersection $\hat{\mathcal{I}}=\hat{\mathcal{I}}^{(1)}\cap\hat{\mathcal{I}}^{(2)}$. The new estimate $\hat{\mathcal{I}}$ should include all important covaraites with high probability as well due to properties of each individual estimate. However by construction, the number of unimportant covariates in the new estimate $\hat{\mathcal{I}}$ is much smaller. The reason is that, in order for an unimportant covariate to appear in $\hat{\mathcal{I}}$, it has to be included in both $\hat{\mathcal{I}}^{(1)}$ and $\hat{\mathcal{I}}^{(2)}$ randomly.

For the new variant method based on random splitting, Fan, Samworth and Wu [7] obtained some non-asymptotic probability bound for the event that r unimportant covaraites are included in the intersection $\hat{\mathcal{I}}$ for any natural number r under some exchangeability condition on all unimportant covariates. The probability bound is decreasing in the dimensionality, showing a "blessing of dimensionality". Please consult Fan, Samworth and Wu [7] for more details. We want to remark that their theoretical bound is applicable to our setting as well while studying time-to-event data because theoretical bound is based on splitting the sample into two halves and only requires the independence between these two halves.

While defining new variants, we may use the same d as used in the original SIS and ISIS. However it will lead to a very aggressive screening. We call the corresponding variant the first variant of (I)SIS. Alternatively, in each step we may choose larger $\hat{\mathcal{I}}^{(1)}$ and $\hat{\mathcal{I}}^{(2)}$ to ensure that their intersection $\hat{\mathcal{I}}^{(1)}\cap\hat{\mathcal{I}}^{(2)}$ has d covariates, which is called the second variant. The second variant ensures that there are at least d covariates included before applying penalization in each step and is thus much less aggressive. Numerical examples will be used to explore their performance and prefer to the first variant.

5. Simulation

5.1. Design of simulations

In this section, we conduct simulation studies to show the power of the (I)SIS and its variants by comparing them with LASSO [15] in the Cox's proportional hazards model. Here the regularization parameter for LASSO is tuned via five fold cross validation. Most of the settings are adapted from Fan and Lv [6] and Fan, Samworth and Wu [7]. Four different configurations are considered with $n = 300$ and $p = 400$. And two of them are revisited with a different pair of sample size $n = 400$ and dimensionality $p = 1000$. Covariates in different settings are generated as follows.

Case 1: $X_1, \ldots, X_p$ are independent and identically distributed $N(0, 1)$ random variables.

Case 2: $X_1, \ldots, X_p$ are multivariate Gaussian, marginally $N(0, 1)$, and with serial correlation $\text{corr}(X_i, X_j) = \rho$ if $i \neq j$. Here we take $\rho = 0.5$.

Case 3: $X_1, \ldots, X_p$ are multivariate Gaussian, marginally $N(0, 1)$, and with correlation structure $\text{corr}(X_i, X_4) = 1/\sqrt{2}$ for all $i \neq 4$ and $\text{corr}(X_i, X_j) = 1/2$ if i and j are distinct elements of $\{1, \cdots, p\}\backslash\{4\}$.

Case 4: $X_1, \ldots, X_p$ are multivariate Gaussian, marginally $N(0, 1)$, and with correlation structure $\text{corr}(X_i, X_5) = 0$ for all $i \neq 5$, $\text{corr}(X_i, X_4) = 1/\sqrt{2}$ for all $i \notin \{4, 5\}$, and $\text{corr}(X_i, X_j) = 1/2$ if i and j are distinct elements of $\{1, \cdots, p\}\backslash\{4, 5\}$.

Case 5: Same as Case 2 except $n = 400$ and $p = 1000$.

Case 6: Same as Case 4 except $n = 400$ and $p = 1000$.

Here, Case 1 with independent predictors is the most straightforward for variable selection. In Cases 2-6, however, we have serial correlation such that $\text{corr}(X_i, X_j)$ does not decay as $|i - j|$ increases. We will see later that for Cases 3, 4 and 6, the true coefficients are specially chosen such that the response is marginally independent but jointly dependent of X_4. We therefore expect variable selection in these situations to be much more challenging, especially for the non-iterated versions of SIS. Notice that in the asymptotic theory of SIS in [6], this type of dependence is ruled out by their Condition (4).

In our implementation, we choose $d = \lfloor \frac{n}{4 \log n} \rfloor$ for both the vanilla version of SIS (Van-SIS) and the second variant (Var2-SIS). For the first variant (Var1-SIS), however, we use $d = \lfloor \frac{n}{\log n} \rfloor$ (note that since the selected variables for the first variant are in the intersection of two sets of size d, we typically end up with far fewer than d variables selected by this method). For any type of SIS or ISIS, we apply SCAD with these selected predictors to get a final estimate of the regression coefficients at the end of the screening step. Whenever necessary, the BIC is used to select the best tuning parameter in the regularization framework.

In all setting, the censoring time is generated from exponential distribution with mean 10. This corresponds to choosing the baseline hazard function $h_0(t) = 0.1$ for $t \geq 0$. The true regression coefficients and censoring rate in each of the six cases are as follows:

Case 1: $\beta_1 = -1.6328, \beta_2 = 1.3988, \beta_3 = -1.6497, \beta_4 = 1.6353, \beta_5 = -1.4209, \beta_6 = 1.7022$, and $\beta_j = 0$ for $j > 6$. The corresponding censoring rate is 33%.

Case 2: The coefficients are the same as Case 1. The corresponding censoring rate is 27%.

Case 3: $\beta_1 = 4, \beta_2 = 4, \beta_3 = 4, \beta_4 = -6\sqrt{2}$, and $\beta_j = 0$ for $j > 4$. The corresponding censoring rate is 30%.

Case 4: $\beta_1 = 4, \beta_2 = 4, \beta_3 = 4, \beta_4 = -6\sqrt{2}, \beta_5 = 4/3$ and $\beta_j = 0$ for $j > 5$. The corresponding censoring rate is 31%.

Case 5: $\beta_1 = -1.5140, \beta_2 = 1.2799, \beta_3 = -1.5307, \beta_4 = 1.5164, \beta_5 = -1.3020, \beta_6 = 1.5833$, and $\beta_j = 0$ for $j > 6$. The corresponding censoring rate is 23%.

Case 6: The coefficients are the same as Case 4. The corresponding censoring rate is 36%.

In Cases 1, 2 and 5 the coefficients were chosen randomly, and were generated as $(4 \log n/\sqrt{n} + |Z|/4)U$ with $Z \sim N(0, 1)$ and $U = 1$ with probability 0.5 and -1 with probability 0.5, independent of Z. For Cases 3, 4, and 6, the choices ensure that even though $\beta_4 \neq 0$, we have that X_4 and Y are marginally independent. The fact that X_4 is marginally independent of the response is designed to make it difficult for the common independent learning to select this variable. In Cases 4 and 6, we add another important variable X_5 with a small coefficient to make it even more difficult.

5.2. Results of simulations

We report our simulation results based on 100 Monte Carlo repetitions for each setting in Tables 1-7. To present our simulation results, we use several different performance measures. In the rows labeled $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1$ and $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2$, we report the median L_1 and squared L_2 estimation errors $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1 = \sum_{j=1}^p |\hat{\beta}_j - \beta_j|$ and $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2 = \sum_{j=1}^p |\hat{\beta}_j - \beta_j|^2$, respectively, where the median is over the 100 repetitions. In the row with label P_1, we report the proportion of the 100 repetitions that the (I)SIS procedure under consideration includes all of the important variables in the model, while the row with label P_2 reports the corresponding proportion of times that the final variables selected, after further application of the SCAD penalty, include all of the important ones. We also report the median model size (MMS) of the final model among 100 repetitions in the row labeled MMS.

We report results of Cases 1 and 2 in Table 1. Recall that the covariates in Case 1 are all independent. In this case, the Van-SIS performs reasonably well. Yet,

TABLE 1

Results for Cases 1 and 2. Here P_1 stands for the probability that (I)SIS includes the true model, i.e., has the sure screening property. P_2 stands for the probability that the final model has the sure screening property. MMS stands for Median Model Size among 100 repetitions. The sample size $n = 300$ and the number of covariates is $p = 400$

	Van-SIS	**Van-ISIS**	**Var1-SIS**	**Var1-ISIS**	**Var2-SIS**	**Var2-ISIS**	**LASSO**
Case 1: independent covariates							
$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1$	0.79	0.57	0.73	0.61	0.76	0.62	4.23
$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2$	0.13	0.09	0.15	0.1	0.15	0.1	0.98
P_1	1	1	0.99	1	0.99	1	–
P_2	1	1	0.99	1	0.99	1	1
MMS	7	6	6	6	6	6	68.5
Case 2: Equi-correlated covariates with $\rho = 0.5$							
$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1$	2.2	0.64	4.22	0.8	3.95	0.78	4.38
$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2$	1.74	0.11	4.71	0.29	4.07	0.28	0.98
P_1	0.71	1	0.42	0.99	0.46	0.99	–
P_2	0.71	1	0.42	0.99	0.46	0.99	1
MMS	7	6	6	6	7	6	57

it does not perform well for the dependent case, Case 2. Note the only difference between Case 1 and Case 2 is the covariance structure of the covariates. For both cases, vanilla-ISIS and its second variant perform very well. It is worth noticing that the ISIS improves significantly over SIS, when covariates are dependent, in terms of both the probability of including all the true variables and in reducing the estimation error. This comparison indicates that the ISIS performs much better when there is serious correlation among covariates.

While implementing the LASSO penalized Cox's proportional hazards model, we adapted the Fortran source code in the R package "glmpath". Recall that the objective function in the LASSO penalized Cox's proportional hazards model is convex and nonlinear. What the Fortran code does is to call a MINOS subroutine to solve the corresponding nonlinear convex optimization problem. Here MINOS is an optimization software developed by Systems Optimization Laboratory at Stanford University. This nonlinear convex optimization problem is much more complicated than a general quadratic programming problem. Thus generally it takes much longer time to solve, especially so when the dimensionality is high as confirmed by Table 3. However the algorithm we used does converge as the objective function is strictly convex.

Table 1 shows that LASSO has the sure screening property as the ISIS, however, the median model size is ten times as large as that of ISIS. As a consequence, it also has larger estimation errors in terms of $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1$ and $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2$. The fact that the median absolute deviation error is much larger than the median square error indicates that the LASSO selects many small nonzero coefficients for those unimportant variables. This is also verified by the fact that LASSO has a very large median model size. The explanation is the bias issue noted by Fan and Li [4]. In order for LASSO to have a small bias for nonzero coefficients, a smaller λ should be chosen. Yet, a small λ recruits many small coefficients for unimportant variables. For Case 2, the LASSO has a similar performance as in Case 1 in that it includes all the important variables but has a much larger model size.

Results of Cases 3 and 4 are reported in Table 2. Note that, in both cases, the design ensures that X_4 is marginally independent of but jointly dependent on Y. This special design disables the SIS to include X_4 in the corresponding identified model as confirmed by our numerical results. However, by using ISIS, we are able to select X_4 for each repetition. Surprisingly, LASSO rarely includes X_4 even if it is not a marginal screening based method. Case 5 is even more challenging. In addition to the same challenge as case 4, the coefficient β_5 is 3 times smaller than the

TABLE 2
Results for Cases 3 and 4. The same caption as Table 1 is used

	Van-SIS	**Van-ISIS**	**Var1-SIS**	**Var1-ISIS**	**Var2-SIS**	**Var2-ISIS**	**LASSO**
Case 3: An important predictor that is independent of survival time							
$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1$	20.1	1.03	20.01	0.99	20.09	1.08	20.53
$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2$	94.72	0.49	100.42	0.47	94.77	0.55	76.31
P_1	0	1	0	1	0	1	–
P_2	0	1	0	1	0	1	0.06
MMS	13	4	8	4	13	4	118.5
Case 4: Two very hard variables to be selected							
$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1$	20.87	1.15	20.95	1.4	20.96	1.41	21.04
$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2$	96.46	0.51	102.14	1.77	97.15	1.78	77.03
P_1	0	1	0	0.99	0	0.99	–
P_2	0	1	0	0.99	0	0.99	0.02
MMS	13	5	9	5	13	5	118

TABLE 3
The average running time (in seconds) comparison for Van-ISIS and LASSO

	Case 1	**Case 2**	**Case 3**	**Case 4**
Van-ISIS	379.29	213.44	402.94	231.68
LASSO	37730.82	26348.12	46847	28157.71

first four variables. Through the correlation with the first 4 variables, unimportant variables $\{X_j, j \geq 6\}$ have a larger marginal utility with Y than X_5. Nevertheless, the ISIS works very well and demonstrates once more that it uses adequately the joint covariate information.

We also compare the computational cost of van-ISIS and LASSO in Table 3 for Cases 1-4. Table 3 shows that it takes LASSO several hours for each repetition, while van-ISIS can finish it in just several minutes. This is a huge improvement. For this reason, for Cases 5 and 6 where $p = 1000$, we only report the results for ISIS since it takes LASSO over several days to complete a single repetition. Results of Cases 5 and 6 are reported in Table 4. The table demonstrates similar performance as Cases 2 and 4 even with more covariates.

To conclude the simulation section, we demonstrate the difficulty of our simulated models by showing the distribution, among 100 simulations, of the minimum $|t|$-statistic for the estimates of the true nonzero regression coefficients in the oracle model with only true important predictors included. More explicitly, during each repetition of each simulation setting, we pretend to know the index set $\mathcal{M}^*$ the true underlying sparse model, fit the Cox's proportional hazards model using only predictors with indices in $\mathcal{M}^*$ by calling function "coxph" of R package "survival", and report the smallest absolute value of the t-statistic for the regression estimates. For example, for case 1, the model size is only 6 and the minimum $|t|$-statistic is computed based on these 6 estimates for each simulation. This shows the difficulty to recover all significant variables even in the oracle model with the minimum model size. The corresponding boxplot for each case is shown in Figure 2. To demonstrate the effect of including unimportant variables, the minimum $|t|$-statistic for the estimates of the true nonzero regression coefficients is recalculated and shown by the boxplots in Figure 3 for the model with the true important variables and 20 unimportant variables.

As expected, cases 1 and 2 are relatively easy cases, whereas cases 3 and 4 are relatively harder in the oracle model. When we are not in the oracle model with 20 noisy variables are added, the difficulty increases as shown in Figure 3. It has more impact on cases 3, 4 and 6.

TABLE 4
Results for Cases 5 and 6. The same caption as that of Table 1 is used

	Van-SIS	**Van-ISIS**	**Var1-SIS**	**Var1-ISIS**	**Var2-SIS**	**Var2-ISIS**
Case 5: The same as case 2 with $p = 1000$ and $n = 400$						
$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1$	1.53	0.52	3.55	0.55	2.95	0.51
$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2$	0.9	0.07	3.48	0.08	2.5	0.07
P_1	0.82	1	0.39	1	0.5	1
P_2	0.82	1	0.39	1	0.5	1
MMS	8	6	6	6	7	6
Case 6: The same as case 4 with $p = 1000$ and $n = 400$						
$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1$	20.88	0.99	20.94	1.1	20.94	1.29
$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2$	93.53	0.39	104.76	0.44	94.02	1.35
P_1	0	1	0	1	0	0.99
P_2	0	1	0	1	0	0.99
MMS	16	5	8	5	16	5

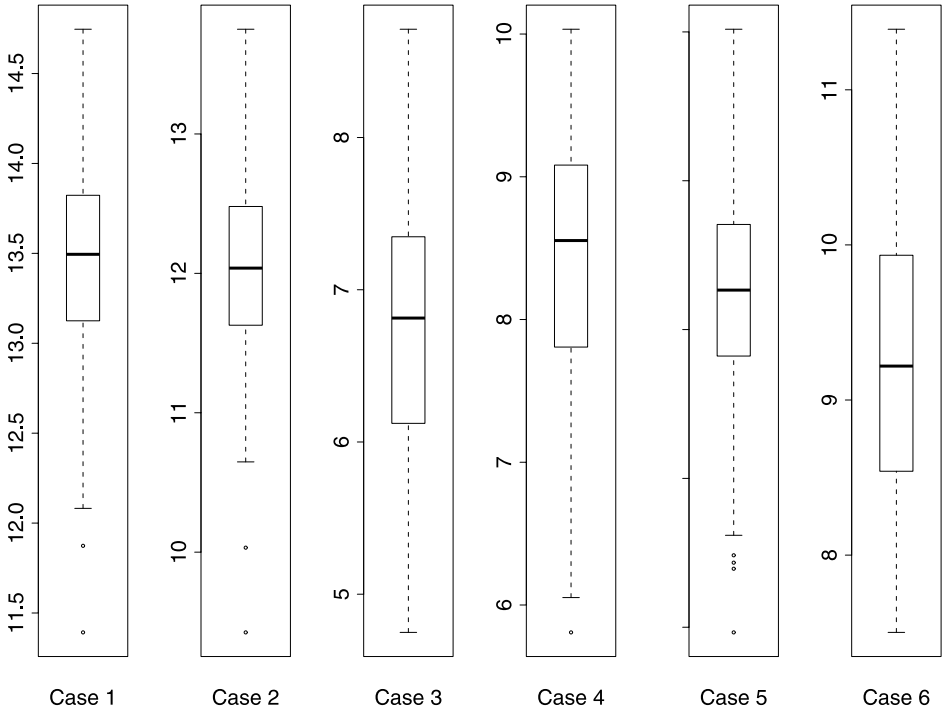

FIG 2. *The boxplots of the minimum* $|t|$*-statistic in the oracle models among 100 simulations.*

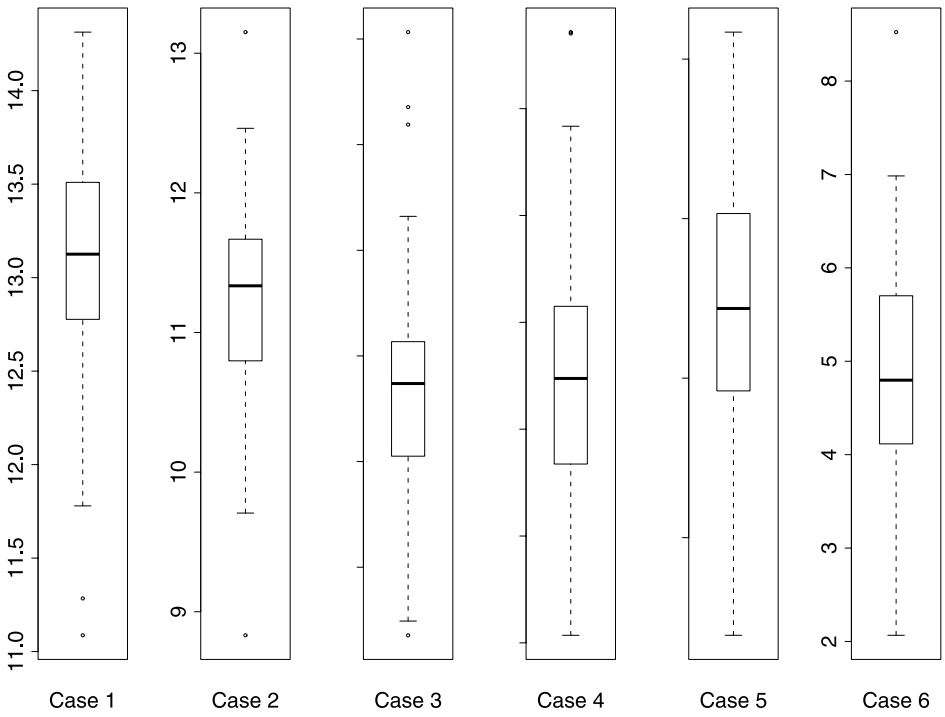

FIG 3. *The boxplots of the minimum* $|t|$*-statistic in the models where 20 noise variables are added among 100 simulations.*

6. Real data

In this section, we use one real data set to demonstrate the power of the proposed method. The Neuroblastoma data set is due to Oberthuer et al. [13]. It was used in [7] for classification studies. Neuroblastoma is an extracranial solid cancer. It is most common in childhood and even in infancy. The annual number of incidences is about several hundreds in the United States. Neuroblastoma is a malignant pediatric tumor originating from neural crest elements of the sympathetic nervous system.

The study includes 251 patients of the German Neuroblastoma Trials NB90-NB2004, who were diagnosed between 1989 and 2004. The patients' ages range from 0 to 296 months at diagnosis with a median age of 15 months. Neuroblastoma specimens of these 251 patients were analyzed using a customized oligonucleotide microarray. The goal is to study the association of gene expression with variable clinical information such as survival time and 3-year event free survival, among others.

We obtained the neuroblastoma data from the MicroArray Quality Control phase-II (MAQC-II) project conducted by the Food Drug Administration (FDA). The complete data set includes gene expression at 10,707 probe sites. It also includes the survival information of each patient. In this example, we focus on the overall survival. There are five outlier arrays. After removing outlier arrays from our consideration, there are 246 patients. The (overall) survival information is available for all 246 patients. The censoring rate is 205/246, which is very heavy. The survival times of those 246 patients are summarized in Figure 4.

As real data are always complex, there may be some genes that are marginally unimportant but work jointly with other genes. Thus it is more appropriate to apply iterative SIS instead of SIS, since the former is more powerful. We standardize each predictor to have mean zero and standard deviation 1 and apply van-ISIS to the standardized data with $d = \lfloor n/\log(n) \rfloor = 43$. ISIS followed with SCAD penalized Cox regression selects 8 genes with probe site names: A_23_P31816, A_23_P31816, A_23_P31816, A_32_P424973, A_32_P159651, Hs61272.2, Hs13208.1, and Hs150167.1.

Now we try to provide some understanding to the significance of these selected genes in predicting the survival information in comparison to other genes that are not selected. We first fitted the Cox's proportional hazard model with all these eight genes. Estimated coefficients are given in Table 5, estimated baseline survival

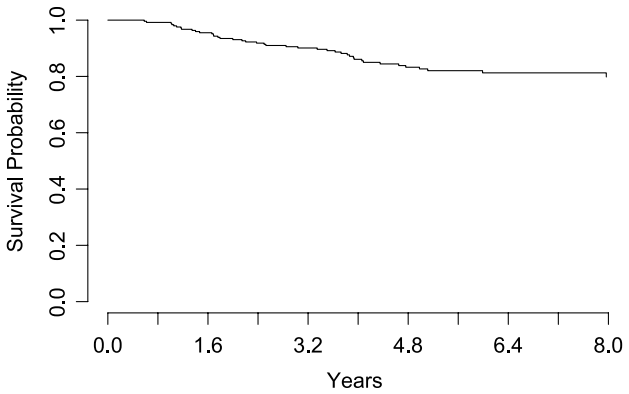

FIG 4. *Estimated survivor function for 246 patients.*

TABLE 5
Estimated coefficients for Neuroblastoma data

Probe ID	Estimated coefficient	Standard error	p-value
A_23_P31816	0.864	0.203	2.1e−05
A_23_P31816	−0.940	0.314	2.8e−03
A_23_P31816	−0.815	1.704	6.3e−01
A_32_P424973	−1.957	0.396	7.8e−07
A_32_P159651	−1.295	0.185	2.6e−12
Hs61272.2	1.664	0.249	2.3e−11
Hs13208.1	−0.789	0.149	1.1e−07
Hs150167.1	1.708	1.687	3.1e−01

function is plotted in Figure 5, and the corresponding log-(partial)likelihood (2.4) is −129.3517. The log-likelihood corresponding to the null model without any predictor is −215.4561. A χ^2 test shows the obvious significance of the model with the eight selected genes. Table 5 shows that there are two estimated coefficients that are statistically insignificant at $\alpha = 1\%$.

Next for each one of these eight genes, we remove it, fit Cox's proportional hazard model with the other seven genes, and get the corresponding log-likelihood. The eight log-likelihoods are −137.5785, −135.1846, −129.4621, −142.4066, −156.4644, −158.3799, −141.0432, and −129.8390. Their average is −141.2948, a decrease of log-likelihood by 11.9431, which is very significant with reduction one gene (the reduction of the degree of freedom by 1). In comparison to the model with the eight selected genes, χ^2 tests shows significance for all selected genes except A_23_P31816 and Hs150167.1. This matches the p-values reported in Table 5.

Finally we randomly select 2 genes out of the genes that are not selected, fit the Cox's proportional hazard model with the above eight genes plus these two randomly selected genes, and record the corresponding log-likelihood. We repeat this process 20 times. We find that the average of these 20 new log-likelihoods is −128.3933, an increase of the log-likelihood merely by 0.9584 with two extra variables included. Comparing to the model with the eight selected genes, χ^2 test shows no significance for the model corresponding to any of the 20 repetitions.

The above experiments show that the selected 8 genes are very important. Deleting one reduces a lot of log-likelihood, while adding two random genes do not increase very much the log-likelihood.

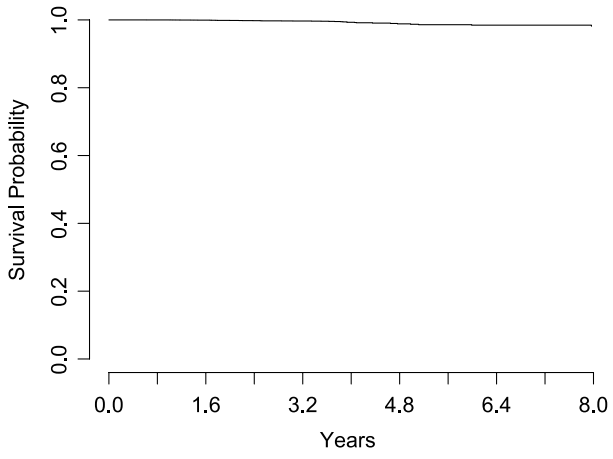

FIG 5. *Estimated baseline survivor function.*

7. Conclusion

We have developed a variable selection technique for the survival analysis with the dimensionality that can be much larger than sample size. The focus is on the iterative sure independence screening, which iteratively applies a large-scale screening that filters unimportant variables by using the conditional marginal utility, and a moderate-scale selection by using penalized partial likelihood method, which selects further the unfiltered variables. The methodological power of the vanilla ISIS has been demonstrated via carefully designed simulation studies. It has sure independence screening with very small false selection. Comparing with the version of LASSO we used, it is much more computationally efficient and far more specific in selecting important variables. As a result, it has much smaller absolute deviation error and mean square error.

References

[1] Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

[2] Cox, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–76.

[3] Fan, J., Feng, Y. and Song, R. (2010). Nonparametric independence screening in sparse ultra-high dimensional additive models. Submitted.

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

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

[6] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B* **70** 849–911.

[7] Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional variable selection: beyond the lienar model. *J. Mach. Learn. Res.* To appear.

[8] Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *Ann. Statist.* To appear.

[9] Faraggi, D. and Simon, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics* **54** 1475–5.

[10] Ibrahim, J. G., Chen, M.-H. and Maceachern, S. N. (1999). Bayesian variable selection for proportional hazards models. *Canand. J. Statist.* **27** 70117.

[11] Klein, J. P. and Moeschberger, M. L. (2005). *Survival Analysis*, 2nd ed. Springer.

[12] Li, Y. and Dicker, L. (2009). Dantzig selector for censored linear regression. Technical report, Harvard Univ. Biostatistics.

[13] Oberthuer, A., Berthold, F., Warnat, P., Hero, B., Kahlert, Y., Spitz, R., Ernestus, K., König, R., Haas, S., Eils, R., Schwab, M., Brors, B., Westermann, F. and Fischer, M. (2006). Customized oligonucleotide microarray gene expressionbased classification of neuroblastoma patients outperforms current clinical risk stratification. *Journal of Clinical Oncology* **24** 5070–5078.

[14] Sauerbrei, W. and Schumacher, M. (1992). A bootstrap resampling procedure for model building: Application to the cox regression model. *Statist. Med.* **11** 2093–2109.

[15] Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statist. Med.* **16** 385–95.

[16] Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.
[17] Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statist. Sinica* **19** 801–817.
[18] Zhang, C.-H. (2009). Penalized linear unbiased selection. *Ann. Statist.* To appear.
[19] Zhang, H. H. and Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika* **94** 691–703.
[20] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429.
[21] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36** 1509–1566.
[22] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67** 301–320.