Independent screening for single-index hazard rate models with ultra-high dimensional features

Gorst-Rasmussen, Anders; Scheike, Thomas

Publication date:
2011

Document Version
Early version, also known as pre-print

Link to publication from Aalborg University

Citation for published version (APA):
Gorst-Rasmussen, A., & Scheike, T. (2011). *Independent screening for single-index hazard rate models with ultra-high dimensional features*. Department of Mathematical Sciences, Aalborg University. Research Report Series No. R-2011-06

AALBORG UNIVERSITY

Independent screening for single-index hazard rate models with ultra-high dimensional features

by

Anders Gorst-Rasmussen and Thomas Scheike

R-2011-06

May 2011

Department of Mathematical Sciences
Aalborg University
Fredrik Bajers Vej 7 G ▪ DK - 9220 Aalborg Øst ▪ Denmark
Phone: +45 99 40 80 80 ▪ Telefax: +45 98 15 81 29
URL: http://www.math.aau.dk

ISSN 1399–2503 ▪ On-line version ISSN 1601–7811

Independent screening for single-index hazard rate models with ultra-high dimensional features

Anders Gorst-Rasmussen†

Department of Mathematical Sciences, Aalborg University, Denmark

Thomas Scheike

Department of Biostatistics, University of Copenhagen, Denmark

Summary. In data sets with many more features than observations, independent screening based on all univariate regression models leads to a computationally convenient variable selection method. Recent efforts have shown that in the case of generalized linear models, independent screening may suffice to capture all relevant features with high probability, even in ultra-high dimension. It is unclear whether this formal sure screening property is attainable when the response is a right-censored survival time. We propose a computationally very efficient independent screening method for survival data which can be viewed as the natural survival equivalent of correlation screening. We state conditions under which the method admits the sure screening property within a general class of single-index hazard rate models with ultra-high dimensional features. An iterative variant is also described which combines screening with penalized regression in order to handle more complex feature covariance structures. The methods are evaluated through simulation studies and through application to a real gene expression data set.

Keywords: Independent screening; survival data; additive hazards model; variable selection; ultra-high dimension

1. Introduction

With the increasing proliferation of biomarker studies, there is a need for efficient methods for relating a survival time response to a large number of features. In typical genetic microarray studies, the sample size n is measured in hundreds whereas the number of features p per sample can be in excess of millions. Sparse regression techniques such as lasso (Tibshirani, 1997) and SCAD (Fan and Li, 2001) have proved useful for dealing with such high-dimensional features but their usefulness diminishes when p becomes extremely large compared to n. The notion of NP-dimensionality (Fan and Lv, 2009) has been conceived to describe such ultra-high dimensional settings which are formally analyzed in an asymptotic regime where p grows at a non-polynomial rate with n. Despite recent progress (Bradic *et al.*, 2011), theoretical knowledge about sparse regression techniques under NP-dimensionality is still in its infancy. Moreover, NP-dimensionality poses substantial computational challenges. When for example pairwise interactions among gene expressions in a genetic microarray study are of interest, the dimension of the feature space will trouble even the most efficient algorithms for fitting sparse regression models. A popular ad hoc solution is to simply pretend that feature correlations are ignorable and resort to computationally swift univariate regression methods; so-called independent screening methods.

In an important paper, Fan and Lv (2008) laid the formal foundation for using independent screening to distinguish 'relevant' features from 'irrelevant' ones. For the linear regression model they showed that, when the design is close to orthogonal, a superset of the true set of nonzero regression coefficients can be estimated consistently by simple hard-thresholding of feature-response correlations. This sure independent screening (SIS) property of correlation screening is a rather trivial one, if not for the fact that it holds true in the asymptotic regime of NP-dimensionality. Thus, when the feature covariance structure is sufficiently simple, SIS methods can overcome the noise accumulation in extremely high dimension. In order to accommodate more complex feature covariance structures Fan and Lv (2008) and Fan *et al.* (2009) developed heuristic, iterated methods combining independent screening with forward selection techniques. Recently, Fan and Song (2010) extended the formal basis for SIS to generalized linear models.

In biomedical applications, the response of interest is often a right-censored survival time, making the study of screening methods for survival data an important one. Fan *et al.* (2010) investigated SIS methods

†*Address for correspondence:* Anders Gorst-Rasmussen, Department of Mathematical Sciences, Aalborg University, 9220 Aalborg, Denmark. Email: gorst@math.aau.dk

for the Cox proportional hazards model based on ranking features according to the univariate partial log-likelihood but gave no formal justification. Tibshirani (2009) suggested soft-thresholding of univariate Cox score statistics with some theoretical justification but under strong assumptions. Indeed, independent screening methods for survival data are apt to be difficult to justify theoretically due to the presence of censoring which can confound marginal associations between the response and the features. Recent work by Zhao and Li (2010) contains ideas which indicate that independent screening based on the Cox model may have the SIS property in the absence of censoring.

In the present paper, we depart from the standard approach of studying SIS as a rather specific type of model misspecification in which the univariate versions of a particular regression model are used to infer the structure of the joint version of the same particular regression model. Instead, we propose a survival variant of independent screening based on a model-free statistic which we call the 'Feature Aberration at Survival Times' (FAST) statistic. The FAST statistic is a simple linear statistic which aggregates across survival times the aberration of each feature relative to its time-varying average. Independent screening based on this statistic can be regarded as a natural survival equivalent of correlation screening. We study the SIS property of FAST screening in ultra-high dimension for a general class of single-index hazard rate regression models in which the risk of an event depends on the features through some linear functional. A key aim has been to derive simple and operational sufficient conditions for the SIS property to hold. Accordingly, our main result states that the FAST statistic has the SIS property in an ultra-high dimensional setting under covariance assumptions as in Fan *et al.* (2009), provided that censoring is essentially random and that features satisfy a technical condition which holds when they follow an elliptically contoured distribution. Utilizing the fact that the FAST statistic is related to the univariate regression coefficients in the semiparametric additive hazards model (Lin and Ying (1994); McKeague and Sasieni (1994)), we develop methods for iterated SIS. The techniques are evaluated in a simulation study where we also compare with screening methods for the Cox model (Fan *et al.*, 2010). Finally, an application to a real genetic microarray data set is presented.

2. The FAST statistic and its motivation

Let T be a survival time which is subject to right-censoring by some random variable C. Denote by $N(t) := 1(T \wedge C \leq t \wedge T \leq C)$ the counting process which counts events up to time t, let $Y(t) := 1(T \wedge C \geq t)$ be the at-risk process, and let $\mathbf{Z} \in \mathbb{R}^p$ denote a random vector of explanatory variables or features. It is assumed throughout that $\mathbf{Z}$ has finite variance and is standardized, i.e. centered and with a covariance matrix $\mathbf{\Sigma}$ with unit diagonal. We observe n independent and identically distributed (i.i.d.) replicates of $\{(N_i, Y_i, \mathbf{Z}_i) : 0 \leq t \leq \tau\}$ for $i = 1, \ldots, n$ where $[0, \tau]$ is the observation time window.

Define the 'Feature Aberration at Survival Times' (FAST) statistic as follows:

$$\mathbf{d} := n^{-1} \int_0^\tau \sum_{i=1}^n \{\mathbf{Z}_i - \bar{\mathbf{Z}}(t)\} \mathrm{d}N_i(t); \tag{1}$$

where $\bar{\mathbf{Z}}$ is the at-risk-average of the $\mathbf{Z}_i$s,

$$\bar{\mathbf{Z}}(t) := \frac{\sum_{i=1}^n \mathbf{Z}_i Y_i(t)}{\sum_{i=1}^n Y_i(t)}.$$

Components of the FAST statistic define basic measures of the marginal association between each feature and survival. In the following, we provide two motivations for using the FAST statistic for screening purposes. The first, being model-based, is perhaps the most intuitive – the second shows that, even in a model-free setting, the FAST statistic may provide valuable information about marginal associations.

2.1. A model-based interpretation of the FAST statistic

Assume in this section that the T_is have hazard functions of the form

$$\lambda_j(t) = \lambda_0(t) + \mathbf{Z}_j^\top \boldsymbol{\alpha}^0; \qquad j = 1, 2, \ldots, n; \tag{2}$$

with λ_0 an unspecified baseline hazard rate and $\boldsymbol{\alpha}^0 \in \mathbb{R}^p$ a vector of regression coefficients. This is the so-called semiparametric additive hazards model (Lin and Ying (1994); McKeague and Sasieni (1994)), henceforth simply the Lin-Ying model. The Lin-Ying model corresponds to assuming for each N_j an intensity function of the form $Y_j(t)\{\lambda_0(t) + \mathbf{Z}_j^\top \boldsymbol{\alpha}^0\}$. From the Doob-Meyer decomposition $\mathrm{d}N_j(t) = \mathrm{d}M_j(t) + Y_j(t)\{\lambda_0(t) + \mathbf{Z}_j^\top \boldsymbol{\alpha}^0\}\mathrm{d}t$ with M_j a martingale, it is easily verified that

$$\sum_{i=1}^n \{\mathbf{Z}_i - \bar{\mathbf{Z}}(t)\} \mathrm{d}N_i(t) = \Big[\sum_{i=1}^n \{\mathbf{Z}_i - \bar{\mathbf{Z}}(t)\}^{\otimes 2} Y_i(t) \mathrm{d}t\Big] \boldsymbol{\alpha}^0 + \sum_{i=1}^n \{\mathbf{Z}_i - \bar{\mathbf{Z}}(t)\} \mathrm{d}M_i(t), \quad t \in [0, \tau]. \tag{3}$$

This suggests that $\boldsymbol{\alpha}^0$ is estimable as the solution to the $p \times p$ linear system of equations

$$\mathbf{d} = \mathbf{D}\boldsymbol{\alpha}; \tag{4}$$

where

$$\mathbf{d} := n^{-1} \sum_{i=1}^{n} \int_0^\tau \{\mathbf{Z}_i - \bar{\mathbf{Z}}(t)\} \mathrm{d}N_i(t), \quad \text{and } \mathbf{D} := n^{-1} \sum_{i=1}^{n} \int_0^\tau Y_i(t) \{\mathbf{Z}_i - \bar{\mathbf{Z}}(t)\}^{\otimes 2} \mathrm{d}t. \tag{5}$$

Suppose $\hat{\boldsymbol{\alpha}}$ solves (4). Standard martingale arguments (Lin and Ying, 1994) imply root n consistency of $\hat{\boldsymbol{\alpha}}$,

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0) \xrightarrow{d} \mathrm{N}(0, \mathbf{D}^{-1}\mathbf{B}\mathbf{D}^{-1}), \quad \text{where } \mathbf{B} = n^{-1} \sum_{i=1}^{n} \int_0^\tau \{\mathbf{Z}_i - \bar{\mathbf{Z}}(t)\}^{\otimes 2} \mathrm{d}N_i(t). \tag{6}$$

For now, simply observe that the left-hand side of (4) is exactly the FAST statistic; whereas $d_j D_{jj}^{-1}$ for $j = 1, 2, \ldots, p$ estimate the regression coefficients in the corresponding p univariate Lin-Ying models. Hence we can interpret $\mathbf{d}$ as a (scaled) estimator of the univariate regression coefficients in a working Lin-Ying model.

A nice heuristic interpretation of $\mathbf{d}$ results from the pointwise signal/error decomposition (3) which is essentially a reformulated linear regression model $\mathbf{X}^\top \mathbf{X}\boldsymbol{\alpha}^0 + \mathbf{X}^\top \boldsymbol{\varepsilon} = \mathbf{X}^\top \mathbf{y}$ with 'responses' $y_j := \mathrm{d}N_j(t)$ and 'explanatory variables' $\mathbf{X}_j := \{\mathbf{Z}_j - \bar{\mathbf{Z}}(t)\} Y_j(t)$. The FAST statistic is given by the time average of $E\{\mathbf{X}^\top \mathbf{y}\}$ and may accordingly be viewed as a survival equivalent of the usual predictor-response correlations.

2.2. *A model-free interpretation of the FAST statistic*

For a feature to be judged (marginally) associated with survival in any reasonable interpretation of survival data, one would first require that the feature is correlated with the probability of experiencing an event – second, that this correlation persists throughout the time window. The FAST statistic can be shown to reflect these two requirements when the censoring mechanism is sufficiently simple.

Specifically, assume administrative censoring at time τ (so that $C_1 \equiv \tau$). Set $V(t) := \mathrm{Var}\{F(t|\mathbf{Z}_1)\}^{1/2}$ where $F(t|\mathbf{Z}_1) := P(T_1 \leq t|\mathbf{Z}_1)$ denotes the conditional probability of death before time t. For each j, denote by δ_j the population version of d_j (the in probability limit of d_j when $n \to \infty$). Then

$$\begin{aligned}
\delta_j &= E\Big(\Big[Z_{1j} - \frac{E\{Z_{1j}Y_1(t)\}}{E\{Y_1(t)\}}\Big] 1(T_1 \leq t \wedge \tau)\Big) \\
&= E\{Z_{1j}F(\tau|\mathbf{Z}_1)\} - \int_0^\tau \frac{E\{Z_{1j}Y_1(t)\}}{E\{Y_1(t)\}} E\{\mathrm{d}F(t|\mathbf{Z}_1)\} \\
&= V(\tau)\mathrm{Cor}\{Z_{1j}, F(\tau|\mathbf{Z}_1)\} + \int_0^\tau \mathrm{Cor}\{Z_{1j}, F(t|\mathbf{Z}_1)\} \frac{V(t)}{E\{Y_1(t)\}} E\{\mathrm{d}F(t|\mathbf{Z}_1)\}.
\end{aligned}$$

We can make the following observations:

(i) If $\mathrm{Cor}\{Z_{1j}, F(t|\mathbf{Z}_1)\}$ has constant sign throughout $[0, \tau]$, then $|\delta_j| \geq |V(\tau)\mathrm{Cor}\{Z_{1j}, F(\tau|\mathbf{Z}_1)\}|$.
(ii) Conversely, if $\mathrm{Cor}\{Z_{1j}, F(t|\mathbf{Z}_1)\}$ changes sign, so that the the direction of association with $F(t|\mathbf{Z}_1)$ is not persistent throughout $[0, \tau]$, then this will lead to a smaller value of $|\delta_j|$ compared to (i).
(iii) Lastly, if $\mathrm{Cor}\{Z_{1j}, F(t|\mathbf{Z}_1)\} \equiv 0$ then $\delta_j = 0$.

In other words, the sample version d_j estimates a time-averaged summary of the correlation function $t \mapsto \mathrm{Cor}\{Z_{1j}, F(t|\mathbf{Z}_1)\}$ which takes into account both magnitude and persistent behavior throughout $[0, \tau]$. This indicates that the FAST statistic is relevant for judging marginal association of features with survival beyond the model-specific setting of Section 2.1.

3. Independent screening with the FAST statistic

In this section, we extend the heuristic arguments of the previous section and provide theoretical justification for using the FAST statistic to screen for relevant features when the data-generating model belongs to a class of single-index hazard rate regression models.

3.1. The general case of single-index hazard rate models

In the notation of Section 2, we assume survival times T_j to have hazard rate functions of single-index form:

$$\lambda_j(t) = \lambda(t, \mathbf{Z}_j^\top \boldsymbol{\alpha}^0), \quad j = 1, 2, \ldots, n. \tag{7}$$

Here $\lambda\colon [0,\infty) \times \mathbb{R} \to [0,\infty)$ is a continuous function, $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are random vectors in $\mathbb{R}^{p_n}$, $\boldsymbol{\alpha}^0 \in \mathbb{R}^{p_n}$ is a vector of regression coefficients, and $\mathbf{Z}_j^\top \boldsymbol{\alpha}^0$ defines a risk score. We subscript p by n to indicate that the dimension of the feature space can grow with the sample size. Censoring will always be assumed at least independent so that C_j is independent of T_j conditionally on $\mathbf{Z}_j$. We impose the following assumption on the hazard 'link function' λ:

Assumption 1. The survival function $\exp\{-\int_0^t \lambda(s,\cdot)\mathrm{d}s\}$ is strictly monotonic for each $t \geq 0$.

Requiring the survival function to depend monotonically on $\mathbf{Z}_j^\top \boldsymbol{\alpha}^0$ is natural in order to enable interpretation of the components of $\boldsymbol{\alpha}^0$ as indicative of positive or negative association with survival. Note that it suffices that $\lambda(t,\cdot)$ is strictly monotonic for each $t \geq 0$. Assumption 1 holds for a range of popular survival regression models. For example, $\lambda(t,x) := \lambda_0(t) + x$ with λ_0 some baseline hazard yields the Lin-Ying model (2); $\lambda(t,x) := \lambda_0(t)\mathrm{e}^x$ is a Cox model; and $\lambda(t,x) := \mathrm{e}^x \lambda_0(t\mathrm{e}^x)$ is an accelerated failure time model.

Denote by $\boldsymbol{\delta}$ the population version of the FAST statistic under the model (7) which, by the Doob-Meyer decomposition $\mathrm{d}N_1(t) = \mathrm{d}M_1(t) + Y_1(t)\lambda(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)\mathrm{d}t$ with M_1 a martingale, takes the form

$$\boldsymbol{\delta} = E\Big[\int_0^\tau \{\mathbf{Z}_1 - \mathbf{e}(t)\} Y_1(t)\lambda(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)\mathrm{d}t\Big]; \qquad \text{where } \mathbf{e}(t) := \frac{E\{\mathbf{Z}_1 Y_1(t)\}}{E\{Y_1(t)\}}. \tag{8}$$

Our proposed FAST screening procedure is as follows: given some (data-dependent) threshold $\gamma_n > 0$,

(i). calculate the FAST statistic $\mathbf{d}$ from the available data and
(ii). declare the 'relevant features' to be the set $\{1 \leq j \leq p_n : |d_j| > \gamma_n\}$.

By the arguments in Section 2, this procedure defines a natural survival equivalent of correlation screening. Define the following sets of features:

$$\begin{aligned} \widehat{\mathrm{M}}_d^n &:= \{1 \leq j \leq p_n : |d_j| > \gamma_n\}, \\ \mathrm{M}^n &:= \{1 \leq j \leq p_n : \alpha_j^0 \neq 0\}, \\ \mathrm{M}_\delta^n &:= \{1 \leq j \leq p_n : \delta_j \neq 0\}. \end{aligned}$$

The problem of establishing the SIS property of FAST screening amounts to determining when $\mathrm{M}^n \subseteq \widehat{\mathrm{M}}_\mathrm{d}^n$ holds with large probability for large n. This translates into two questions: first, when do we have $\mathrm{M}_\delta^n \subseteq \widehat{\mathrm{M}}_d^n$; second, when do we have $\mathrm{M}^n \subseteq \mathrm{M}_\delta^n$? The first question is essentially model-independent and requires establishing an exponential bound for $n^{1/2}|d_j - \delta_j|$ as $n \to \infty$. The second question is strongly model-dependent and is answered by manipulating expectations under the single-index model (7).

We state the main results here and relegate proofs to the appendix where we also state various regularity conditions. The following principal assumptions, however, deserve separate attention:

Assumption 2. There exists $\mathbf{c} \in \mathbb{R}^{p_n}$ such that $E(\mathbf{Z}_1 | \mathbf{Z}_1^\top \boldsymbol{\alpha}^0) = \mathbf{c}\mathbf{Z}_1^\top \boldsymbol{\alpha}^0$.

Assumption 3. The censoring time C_1 depends on $T_1, \mathbf{Z}_1$ only through Z_{1j}, $j \notin \mathrm{M}^n$.

Assumption 4. Z_{1j}, $j \in \mathrm{M}^n$ is independent of Z_{1j}, $j \notin \mathrm{M}^n$.

Assumption 2 is a 'linear regression' property which holds true for Gaussian features and, more generally, for features following an elliptically contoured distribution (Hardin, 1982). In view of Hall and Li (1993) which states that most low dimensional projections of high dimensional features are close to linear, Assumption 2 may not be unreasonable a priori even for general feature distributions when p_n is large.

Assumption 3 restricts the censoring mechanism to be partially random in the sense of depending only on irrelevant features. As we will discuss in detail below, such rather strong restrictions on the censoring distribution seem necessary for obtaining the SIS property; Assumption 3 is both general and convenient.

Assumption 4 is the partial orthogonality condition also used by Fan and Song (2010). Under this assumption and Assumption 3, it follows from (8) that $\delta_j = 0$ whenever $j \notin \mathrm{M}^n$, implying $\mathrm{M}_\delta^n \subseteq \mathrm{M}^n$. Provided that we also have $\delta_j \neq 0$ for $j \in \mathrm{M}^n$ (that is, $\mathrm{M}^n \subseteq \mathrm{M}_{\mathrm{pre}}^n$), there exists a threshold $\zeta_n > 0$ such that

$$\min_{j \in \mathrm{M}^n} |\delta_j| \geq \zeta_n \qquad \max_{j \notin \mathrm{M}^n} |\delta_j| = 0.$$

Consequently, Assumptions 3-4 are needed to enable consistent model selection via independent screening. Although model selection consistency is not essential in order to capture just some superset of the relevant features via independent screening, it is pertinent in order to limit the size of such a superset.

The following theorem on FAST screening (FAST-SIS) is our main theoretical result. It states that the screening property $\mathrm{M}^n \subseteq \widehat{\mathrm{M}}_d^n$ may hold with large probability even when p_n grows exponentially fast in a certain power of n which depends on the tail behavior of features. The covariance condition in the theorem is analogous to that of Fan and Song (2010) for SIS in generalized linear models with Gaussian features.

Theorem 1. Suppose that Assumptions 1-3 hold alongside the regularity conditions of the appendix and that $P(|Z_{1j}| > s) \leq l_0 \exp(-l_1 s^{\eta})$ for some positive constants l_0, l_1, η and sufficiently large s. Suppose moreover that for some $c_1 > 0$ and $\kappa < 1/2$,

$$|\mathrm{Cov}[Z_{1j}, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0\}]| \geq c_1 n^{-\kappa}, \quad j \in \mathrm{M}^n. \tag{9}$$

Then $\mathrm{M}^n \subseteq \mathrm{M}_\delta^n$. Suppose in addition that $\gamma_n = c_2 n^{-\kappa}$ for some constant $0 < c_2 \leq c_1/2$ and that $\log p_n = o\{n^{(1-2\kappa)\eta/(\eta+2)}\}$. Then the SIS property holds, $P(\mathrm{M}^n \subseteq \widehat{\mathrm{M}}_d^n) \to 1$ when $n \to \infty$.

Observe that with bounded features, we may take $\eta = \infty$ and handle dimension of order $\log p_n = o(n^{1-2\kappa})$.

We may dispense with Assumption 2 on the feature distribution by revising (9). By Lemma 5 in the appendix, taking $\tilde{e}_j(t) := E\{Z_{1j} P(T_1 \geq t | \mathbf{Z}_1)\}/E\{P(T_1 \geq t|\mathbf{Z}_1)\}$, it holds generally under Assumption 3 that

$$\delta_j = E\{\tilde{e}_j(T_1 \wedge C_1 \wedge \tau)\}, \quad j \in \mathrm{M}^n.$$

Accordingly, if we replace (9) with the assumption that $E|Z_{1j} P(T_1 \geq t|\mathbf{Z}_1)| \geq c_1 n^{-\kappa}$ uniformly in t for $j \in \mathrm{M}^n$, the conclusions of Theorem 1 still hold. In other words, we can generally expect FAST-SIS to detect features which are ‘correlated with the chance of survival’, much in line with Section 2. While this is valuable structural insight, the covariance assumption (9) seems a more operational condition.

Assumption 3 is crucial to the proof of Theorem 1 and to the general idea of translating a model-based feature selection problem into a problem of hard-thresholding $\boldsymbol{\delta}$. A weaker assumption is not possible in general. For example, suppose that only Assumption 2 holds and that the censoring time also follows some single-index model of the form (7) with regression coefficients $\boldsymbol{\beta}^0$. Applying Lemma 2.1 of Cheng and Wu (1994) to (8), there exists finite constants ζ_1, ζ_2 (depending on n) such that

$$\boldsymbol{\delta} = \boldsymbol{\Sigma}(\zeta_1 \boldsymbol{\alpha}^0 + \zeta_2 \boldsymbol{\beta}^0). \tag{10}$$

It follows that unrestricted censoring will generally confound the relationship between $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}\boldsymbol{\alpha}^0$, hence $\boldsymbol{\alpha}^0$. The precise impact of such unrestricted censoring seems difficult to discern, although (10) suggests that FAST-SIS may still be able to capture the underlying model (unless $\zeta_1 \boldsymbol{\alpha}^0 + \zeta_2 \boldsymbol{\beta}^0$ is particularly ill-behaved). We will have more to say about unrestricted censoring in the next section.

Theorem 1 shows that FAST-SIS can consistently capture a superset of the relevant features. A priori, this superset can be quite large; indeed, ‘perfect’ screening would result by simply including all features. For FAST-SIS to be useful, it must substantially reduce feature space dimension. Below we state a survival analogue of Theorem 5 in Fan and Song (2010), providing an asymptotic rate on the FAST-SIS model size.

Theorem 2. Suppose that Assumptions 1-3 hold alongside the regularity conditions of the appendix and that $P(|Z_{1j}| > s) \leq l_0 \exp(-l_1 s^{\eta})$ for positive constants l_0, l_1, η and sufficiently large s. If $\gamma_n = c_4 n^{-2\kappa}$ for some $\kappa < 1/2$ and $c_4 > 0$, there exists a positive constant c_5 such that

$$P[|\widehat{\mathrm{M}}_d^n| \leq O\{n^{2\kappa} \lambda_{\max}(\boldsymbol{\Sigma})\}] \geq 1 - O(p_n \exp\{-c_5 n^{(1-2\kappa)\eta/(\eta+2)}\});$$

where $\lambda_{\max}(\Sigma)$ denotes the maximal eigenvalue of the covariance matrix $\boldsymbol{\Sigma}$ of the feature distribution.

Informally, the theorem states that, under similar assumptions as in Theorem 1 and the partial orthogonality condition (Assumption 4), if features are not too strongly correlated (as measured by the maximal eigenvalue of the covariance matrix) and p_n grows sufficiently fast, we can choose a threshold γ_n for hard-thresholding such that the false selection rate becomes asymptotically negligible.

Our theorems say little about how to actually select the hard-thresholding parameter γ_n in practice. Following Fan and Lv (2008) and Fan *et al.* (2009), we would typically choose γ_n such that $|\mathrm{M}_{\mathrm{pre}}^n|$ is of order $n/\log n$. Devising a general data-adaptive way of choosing γ_n is an open problem; false-selection-based criteria are briefly mentioned in Section 3.3.

3.2. The special case of the Aalen model

Additional insight into the impact of censoring on FAST-SIS is possible within the more restrictive context of the nonparametric Aalen model with Gaussian features (Aalen (1980); Aalen (1989)). This particular model asserts a hazard rate function for T_i of the form

$$\lambda_j(t) = \lambda_0(t) + \mathbf{Z}_j^\top \boldsymbol{\alpha}^0(t), \quad j = 1, 2, \ldots, n; \tag{11}$$

for some baseline hazard rate function λ_0 and $\boldsymbol{\alpha}^0$ a vector of continuous regression coefficient functions. The Aalen model extends the Lin-Ying model of Section 2 by allowing time-varying regression coefficients. Alternatively, it can be viewed as defining an expansion to the first order of a general hazard rate function in the class (7) in the sense that

$$\lambda\left(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0\right) \approx \lambda(t, 0) + \mathbf{Z}_1^\top \boldsymbol{\alpha}^0 \frac{\partial \lambda(t,x)}{\partial x}\Big|_{x=0}. \tag{12}$$

For Aalen models with Gaussian features, we have the following analogue to Theorem 1.

Theorem 3. Suppose that Assumptions 1-2 hold alongside the regularity conditions of the appendix. Suppose moreover that the $\mathbf{Z}_1$ is mean zero Gaussian and that T_1 follows a model of the form (11) with regression coefficients $\boldsymbol{\alpha}^0$. Assume that C_1 also follows a model of the form (11) conditionally on $\mathbf{Z}_1$ and that censoring is independent. Let $\mathbf{A}^0(t) := \int_0^t \boldsymbol{\alpha}^0(s)\mathrm{d}s$. If for some $\kappa < 1/2$ and $c_1 > 0$, we have

$$|\mathrm{Cov}[Z_{1j}, \mathbf{Z}_1^\top E\{\mathbf{A}^0(T_1 \wedge C_1 \wedge \tau)\}]| \geq c_1 n^{-\kappa}, \quad j \in \mathrm{M}^n, \tag{13}$$

then the conclusions of Theorem 1 hold with $\eta = 2$.

In view of (12), Theorem 3 can be viewed as establishing, within the model class (7), conditions for first-order validity of FAST-SIS under a general (independent) censoring mechanism and Gaussian features. The expectation term in (13) is essentially the 'expected regression coefficients at the exit time' which is strongly dependent on censoring through the symmetric dependence on survival and censoring time.

In fact, general independent censoring is a nuisance even in the Lin-Ying model which would otherwise seem the 'natural model' in which to use FAST-SIS. Specifically, assuming only independent censoring, suppose that T_1 follows a Lin-Ying model with regression coefficients $\boldsymbol{\alpha}^0$ conditionally on $\mathbf{Z}_1$ and that C_1 also follows some Lin-Ying model conditionally on $\mathbf{Z}_1$. If $\mathbf{Z}_1 = \boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{Z}}_1$ where the components of $\tilde{\mathbf{Z}}_1$ are i.i.d. with mean zero and unit variance, there exists a $p_n \times p_n$ diagonal matrix $\mathbf{C}$ such that

$$\boldsymbol{\delta} = \boldsymbol{\Sigma}^{1/2}\mathbf{C}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}^0. \tag{14}$$

See Lemma 6 in the appendix. It holds that $\mathbf{C}$ has constant diagonal iff features are Gaussian; otherwise the diagonal is nonconstant and depends nontrivially on the regression coefficients of the censoring model. A curious implication is that, under Gaussian features, FAST screening has the SIS property for this 'double' Lin-Ying model irrespective of the (independent) censoring mechanism. Conversely, sufficient conditions for a SIS property to hold here under more general feature distributions would require the jth component of $\boldsymbol{\Sigma}^{1/2}\mathbf{C}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}^0$ to be 'large' whenever α_j^0 is 'large'; hardly a very operational assumption. In other words, even in the simple Lin-Ying model, unrestricted censoring complicates analysis of FAST-SIS considerably.

3.3. Scaling the FAST statistic

The FAST statistic is easily generalized to incorporate scaling. Inspection of the results in the appendix immediately shows that multiplying the FAST statistic by some strictly positive, deterministic weight does not alter its asymptotic behavior. Under suitable assumptions, this also holds when weights are stochastic. In the notation of Section 2, the following two types of scaling are immediately relevant:

$$d_j^Z = d_j B_{jj}^{-1/2} \text{ (Z-FAST);} \tag{15}$$

$$d_j^{\mathrm{LY}} = d_j D_{jj}^{-1} \quad \text{(Lin-Ying-FAST).} \tag{16}$$

The Z-FAST statistic corresponds to standardizing $\mathbf{d}$ by its estimated standard deviation; screening with this statistic is equivalent to the standard approach of ranking features according to univariate Wald p-values. Various forms of asymptotic false-positive control can be implemented for Z-FAST, courtesy of the central limit theorem. Note that Z-FAST is model-independent in the sense that its interpretation (and asymptotic normality) does not depend on a specific model. In contrast, the Lin-Ying-FAST statistic is model-specific and corresponds to calculating the univariate regression coefficients in the Lin-Ying model, thus leading to an analogue of the idea of 'ranking by absolute regression coefficients' of Fan and Song (2010) .

We may even devise a scaling of $\mathbf{d}$ which mimicks the 'ranking by marginal likelihood ratio' screening of Fan and Song (2010) by considering univariate versions of the natural loss function $\boldsymbol{\beta} \mapsto \boldsymbol{\beta}^\top \mathbf{D}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{d}$ for the Lin-Ying model. The components of the resulting statistic are rather similar to (16), taking the form

$$d_j^{\text{loss}} = d_j B_{jj}^{-1/2} \text{ (loss-FAST).} \tag{17}$$

Additional flexibility can be gained by using a time-dependent scaling where some strictly positive (stochastic) weight is multiplied on the integrand in (1). This is beyond the scope of the present paper.

4. Beyond simple independent screening – iterated FAST screening

The main assumption underlying any SIS method, including FAST-SIS, is that the design is close to orthogonal. This assumption is easily violated: a relevant feature may have a low marginal association with survival; an irrelevant feature may be indirectly associated with survival through associations with relevant features etc. To address such issues, Fan and Lv (2008) and Fan *et al.* (2009) proposed various heuristic iterative SIS (ISIS) methods which generally work as follows. First, SIS is used to recruit a small subset of features within which an even smaller subset of features is selected using a (multivariate) variable selection method such as penalized regression. Second, the (univariate) relevance of each feature not selected in the variable selection step is re-evaluated, adjusted for all the selected features. Third, a small subset of the most relevant of these new features is joined to the set of already selected features, and the variable selection step is repeated. The last two steps are iterated until the set of selected features stabilizes or some stopping criterion of choice is reached.

We advocate a similar strategy to extend the application domain of FAST-SIS. In view of Section 2.1, a variable step using a working Lin-Ying model is intuitively sensible. We may also provide some formal justification. Firstly, estimation in a Lin-Ying model corresponds to optimizing the loss function

$$L(\boldsymbol{\beta}) := \boldsymbol{\beta}^\top \mathbf{D}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{d}; \tag{18}$$

where $\mathbf{D}$ was defined in Section 2.1. As discussed by Martinussen and Scheike (2009), the loss function (18) is meaningful for general hazard rate models: it is the empirical version of the mean squared prediction error for predicting, with a working Lin-Ying model, the part of the intensity which is orthogonal to the at-risk indicator. In the present context, we are mainly interested in the model selection properties of a working Lin-Ying model. Suppose that T_1 conditionally on $\mathbf{Z}_1$ follows a single-index model of the form (7) and that Assumptions 3-4 hold. Suppose that $\boldsymbol{\Delta}\boldsymbol{\beta}^0 = \boldsymbol{\delta}$ with $\boldsymbol{\Delta}$ the in probability limit of $\boldsymbol{D}$. Then $\alpha_j^0 \equiv 0$ implies $\beta_j^0 = 0$ (Hattori, 2006) so that a working Lin-Ying model will yield conservative model selection in a quite general setting. Under stronger assumptions, the following result, related to work by Brillinger (1983) and Li and Duan (1989), is available.

Theorem 4. Assume that T_1 conditionally on $\mathbf{Z}_1$ follows a single-index model of the form (7). Suppose moreover that Assumption 2 holds and that C_1 is independent of $T_1, \mathbf{Z}_1$ (random censoring). If $\boldsymbol{\beta}^0$ defined by $\boldsymbol{\Delta}\boldsymbol{\beta}^0 = \boldsymbol{\delta}$ is the vector of regression coefficients of the associated working Lin-Ying model and $\boldsymbol{\Delta}$ is nonsingular, then there exists a nonzero constant ν depending only on the distributions of $\mathbf{Z}_1^\top \boldsymbol{\alpha}^0$ and C_1 such that $\boldsymbol{\beta}^0 = \nu\boldsymbol{\alpha}^0$.

Thus a working Lin-Ying model can consistently estimate regression coefficient signs under misspecification. From the efforts of Zhu *et al.* (2009) and Zhu and Zhu (2009) for other types of single-index models, it seems conceivable that variable selection methods designed for the Lin-Ying model will enjoy certain consistency properties within the model class (7). The conclusion of Theorem 4 continues to hold when $\boldsymbol{\Delta}$ is replaced by any matrix proportional to the feature covariance matrix $\boldsymbol{\Sigma}$. This is a consequence of Assumption 2 and underlines the considerable flexibility available when estimating in single-index models.

Variable selection based on the Lin-Ying loss (18) can be accomplished by optimizing a penalized loss function of the form

$$\boldsymbol{\beta} \mapsto L(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|); \tag{19}$$

where $p_\lambda : \mathbb{R} \to \mathbb{R}$ is some nonnegative penalty function, singular at the origin to facilitate model selection (Fan and Li, 2001) and depending on some tuning parameter λ controlling the sparsity of the penalized estimator. A popular choice is the lasso penalty (Tibshirani, 2009) and its adaptive variant (Zou, 2006), corresponding to penalty functions $p_\lambda(|\beta_j|) = \lambda|\beta_j|$ and $p_\lambda(|\beta_j|) = \lambda|\beta_j|/|\hat{\beta}_j|$ with $\hat{\boldsymbol{\beta}}$ some root n consistent estimator of $\boldsymbol{\beta}^0$, respectively. These penalties were studied by Ma and Leng (2007) and Martinussen and Scheike (2009) for the Lin-Ying model. Empirically, we have had better success with the one-step SCAD (OS-SCAD) penalty of Zou and Li (2008) than with lasso penalties. Letting

$$w_\lambda(x) := \lambda 1(x \leq \lambda) + \frac{(a\lambda - x)_+}{a-1} 1(x > \lambda), \quad a > 2 \tag{20}$$

an OS-SCAD penalty function for the Lin-Ying model can be defined as follows:

$$p_\lambda(|\beta_j|) := w_\lambda(\bar{D}|\hat{\beta}_j|)|\beta_j|. \tag{21}$$

Here $\hat{\boldsymbol{\beta}} := \operatorname{argmin}_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$ is the unpenalized estimator and $\bar{D} := n^{-1}\operatorname{tr}(\mathbf{D})$ is the average diagonal element of $\mathbf{D}$; this particular re-scaling is just one way to lessen dependency of the penalization on the time scale. If $\mathbf{D}$ has approximately constant diagonal (which is often the case for standardized features), then re-scaling by $\bar{D}$ leads to a similar penalty as for OS-SCAD in the linear regression model with standardized features. The choice $a = 3.7$ in (20) was recommended by Fan and Li (2001). OS-SCAD has not previously been explored for the Lin-Ying model but its favorable performance in ISIS for other regression models is well known (Fan *et al.*, 2009, 2010). OS-SCAD can be implemented efficiently using, for example, coordinate descent methods for fitting the lasso (Gorst-Rasmussen and Scheike, 2011; Friedman *et al.*, 2007). For fixed p, the OS-SCAD penalty (21) has the oracle property if the Lin-Ying model holds true. A proof is beyond scope but follows by adapting Zou and Li (2008) along the lines of Martinussen and Scheike (2009).

In the basic FAST-ISIS algorithm proposed below, the initial recruitment step corresponds to ranking the regression coefficients in the univariate Lin-Ying models. This is a convenient generic choice because it enables interpretation of the algorithm as standard 'vanilla ISIS' (Fan *et al.*, 2009) for the Lin-Ying model.

Algorithm 1 (Lin-Ying-FAST-ISIS). Set $\mathrm{M} := \{1, \ldots, p\}$, let $r_{\max}$ be some pre-defined maximal number of iterations of the algorithm.

(a) (*Initial recruitment*). Perform SIS by ranking $|d_j D_{jj}^{-1}|$, $j = 1, \ldots, p_n$, according to decreasing order of magnitude and retaining the $k_0 \leq d$ most relevant features $\mathrm{A}_1 \subseteq \mathrm{M}$.

(b) For $r = 1, 2, \ldots$ do:

(i) (*Feature selection*). Define $\omega_j := \infty$ if $j \in \mathrm{A}_r$ and $\omega_j := 0$ otherwise. Estimate

$$\hat{\boldsymbol{\beta}} := \operatorname{argmin}_{\boldsymbol{\beta}} \Big\{ L(\boldsymbol{\beta}) + \sum_{j=1}^{p_n} \omega_j p_{\hat{\lambda}}(|\beta_j|) \Big\};$$

with p_λ defined in (21) for some suitable tuning parameter $\hat{\lambda}$. Set $\mathrm{B}_r := \{j : \hat{\beta}_j \neq 0\}$.

(ii) If $r > 1$ and $\mathrm{B}_r = \mathrm{B}_{r-1}$, or if $r = r_{\max}$; return B_r.

(iii) (*Re-recruitment*). Otherwise, re-evaluate relevance of features in $\mathrm{M}\backslash\mathrm{B}_r$ according to the absolute value of their regression coefficient $|\tilde{\beta}_j|$ in the $|\mathrm{M}\backslash\mathrm{B}_r|$ unpenalized Lin-Ying models including each feature in $\mathrm{M}\backslash\mathrm{B}_r$ and all features in B_r, i.e.

$$\tilde{\beta}_j := \hat{\beta}_1^{(j)}, \quad \text{where } \hat{\boldsymbol{\beta}}^{(j)} = \operatorname{argmin}_{\boldsymbol{\beta}_{\{j\}\cup \mathrm{B}_r}} L(\boldsymbol{\beta}_{\{j\}\cup \mathrm{B}_r}), \quad j \in \mathrm{M}\backslash\mathrm{B}_r. \tag{22}$$

Take $\mathrm{A}_{r+1} := \mathrm{C}_r \cup \mathrm{B}_r$ where C_r is the set of the k_r most relevant features in $\mathrm{M}\backslash\mathrm{A}_r$, ranked according to decreasing order of magnitude of $|\tilde{\beta}_j|$.

Fan and Lv (2008) recommended choosing d to be of order $n/\log n$. Following Fan *et al.* (2009), we may take $k_0 = \lfloor 2d/3 \rfloor$ and $k_l = d - |\mathrm{A}_l|$ at each step. This k_0 ensures that we complete at least one iteration of the algorithm; the choice of k_r for $r > 0$ ensures that at most d features are included in the final solution.

Algorithm 1 defines an iterated variant of SIS with the Lin-Ying-FAST statistic (16). We can devise an analogous iterated variant of Z-FAST-SIS in which the initial recruitment is performed by ranking based on the statistic (15), and the subsequent re-recruitments are performed by ranking $|Z|$-statistics in the multivariate Lin-Ying model according to decreasing order of magnitude, using the variance estimator (6). A third option would be to base recruitment on (17) and re-recruitments on the decrease in the multivariate loss (18) when joining a given feature to the set of features picked out in the variable selection step.

The re-recruitment step (b.iii) in Algorithm 1 resembles that of Fan *et al.* (2009). Its naive implementation will be computationally burdensome when p_n is large, requiring a low-dimensional matrix inversion per feature. Significant speedup over the naive implementation is possible via the matrix identity

$$\mathbf{D} = \begin{pmatrix} e & \mathbf{f}^\top \\ \mathbf{f} & \tilde{\mathbf{D}} \end{pmatrix} \Rightarrow \mathbf{D}^{-1} = \begin{pmatrix} k^{-1} & -k^{-1}\mathbf{f}^\top\tilde{\mathbf{D}}^{-1} \\ -k^{-1}\tilde{\mathbf{D}}^{-1}\mathbf{f} & (\tilde{\mathbf{D}} - e^{-1}\mathbf{f}\mathbf{f}^\top)^{-1} \end{pmatrix} \quad \text{where } k = e - \mathbf{f}^\top\tilde{\mathbf{D}}^{-1}\mathbf{f}. \tag{23}$$

Note that only the first row of $\mathbf{D}^{-1}$ is required for the re-recruitment step so that (22) can be implemented using just a single low-dimensional matrix inversion alongside $O(p_n)$ matrix/vector multiplications. Combining (23) with (6), a similarly efficient implementation applies for *Z*-FAST-ISIS.

The variable selection step (b.i) of Algorithm 1 requires the choice of an appropriate tuning parameter. This is traditionally a difficult part of penalized regression, particularly when the aim is model selection where methods such as cross-validation are prone to overfitting (Leng *et al.*, 2007). Previous work on ISIS used the Bayesian information criterion (BIC) for tuning parameter selection (Fan *et al.*, 2009). Although BIC is based on the likelihood, we may still define the following 'pseudo BIC' based on the loss (18):

$$\mathrm{PBIC}(\lambda) = \kappa\{L(\hat{\boldsymbol\beta}_\lambda) - L(\hat{\boldsymbol\beta})\} + n^{-1}\mathrm{df}_\lambda \log n. \tag{24}$$

Here $\hat{\boldsymbol\beta}_\lambda$ is the penalized estimator, $\hat{\boldsymbol\beta}$ is the unpenalized estimator, $\kappa > 0$ is a scaling constant of choice, and df_λ estimates the degrees of freedom of the penalized estimator. A computationally convenient choice is $\mathrm{df}_\lambda = \|\hat{\boldsymbol\beta}_\lambda\|_0$ (Zou *et al.*, 2007). It turns out that choosing $\hat\lambda = \operatorname{argmin}_\lambda \mathrm{PBIC}_\lambda$ may lead to model selection consistency. Specifically, the loss (18) for the Lin-Ying model is of the least-squares type. Then we can repeat the arguments of Wang and Leng (2007) and show that, under suitable consistency assumptions for the penalized estimator, there exists a sequence $\lambda_n \to 0$ yielding selection consistency for $\hat{\boldsymbol\beta}_{\lambda_n}$ and satisfying

$$P\Big\{\inf_{\lambda\in S} \mathrm{PBIC}(\lambda) > \mathrm{PBIC}(\lambda_n)\Big\} \to 1, \qquad n \to \infty; \tag{25}$$

with S the union of the set of tuning parameters λ which lead to overfitted (strict supermodels of the true model), respectively underfitted models (any model which do not include the true model). While (25) holds independently of the scaling constant κ, the finite-sample behavior of PBIC depends strongly on κ. A sensible value may be inferred heuristically as follows: the range of a 'true' likelihood BIC is asymptotically equivalent to a Wald statistic in the sense that (for fixed p),

$$\mathrm{BIC}(0) - \mathrm{BIC}(\infty) = \hat{\boldsymbol\beta}_{\mathrm{ML}}^\top \mathrm{I}(\boldsymbol\beta_0)\hat{\boldsymbol\beta}_{\mathrm{ML}} + o_p(n^{-1/2}); \tag{26}$$

with $\hat{\boldsymbol\beta}_{\mathrm{ML}}$ the maximum likelihood estimator and $\mathrm{I}(\boldsymbol\beta_0) \approx n^{-1}\mathrm{Var}(\hat{\boldsymbol\beta}_{\mathrm{ML}} - \boldsymbol\beta_0)^{-1}$ the information matrix. We may specify κ by requiring that $\mathrm{PBIC}(0) - \mathrm{PBIC}(\infty)$ admits an analogous interpretation as a Wald statistic. Since $\mathrm{PBIC}(0) - \mathrm{PBIC}(\infty) = \kappa\mathbf{d}^\top\mathbf{D}^{-1}\mathbf{d} + o_p(n^{-1/2})$, it follows from (6) that we should choose

$$\kappa := \frac{\mathbf{d}^\top\mathbf{B}^{-1}\mathbf{d}}{\mathbf{d}^\top\mathbf{D}^{-1}\mathbf{d}}.$$

This choice of κ also removes the dependency of PBIC on the time scale.

5. Simulation studies

In this section, we investigate the performance of FAST screening on simulated data. Rather than comparing with popular variable selection methods such as the lasso, we will compare with analogous screening methods based on the Cox model (Fan *et al.*, 2010). This seems a more pertinent benchmark since previous work has already demonstrated that (iterated) SIS can outperform variable selection based on penalized regression in a number of cases (Fan and Lv (2008); Fan *et al.* (2009)).

For all the simulations, survival times were generated from three different conditionally exponential models of the generic form (7); that is, a time-independent hazard 'link function' applied to a linear functional of features. For suitable constants c, the link functions were as follows (see also Figure 1):

$$\begin{aligned} \text{Logit}: \quad & \lambda_{\mathrm{logit}}(t,x) &:=& \quad \{1+\exp(c_{\mathrm{logit}}x\}^{-1} \\ \text{Cox}: \quad & \lambda_{\mathrm{cox}}(t,x) &:=& \quad \exp(c_{\mathrm{cox}}x) \\ \text{Log}: \quad & \lambda_{\mathrm{log}}(t,x) &:=& \quad \log\{\mathrm{e}+(c_{\mathrm{log}}x)^2\}\{1+\exp(c_{\mathrm{log}}x)\}^{-1}. \end{aligned}$$

The link functions represent different characteristic effects on the feature functional, ranging from uniformly bounded (logit) over fast decay/increase (Cox), to fast decay/slow increase (log). We took $c_{\text{logit}} = 1.39$, $c_{\text{cox}} = 0.68$, and $c_{\log} = 1.39$ and, unless otherwise stated, survival times were right-censored by independent exponential random variables with rate parameters 0.12 (logit link), 0.3 (Cox link) and 0.17 (log link). These constants were selected to provide a crude 'calibration' to make the simulation models more comparable: for a univariate standard Gaussian feature Z_1, a regression coefficient $\beta = 1$, and a sample size of $n = 300$, the expected $|Z|$-statistic was 8 for all three link functions with an expected censoring rate of 25%, as evaluated by numerical integration based on the true likelihood.

Methods for FAST screening have been implemented in the R package 'ahaz' (Gorst-Rasmussen, 2011).

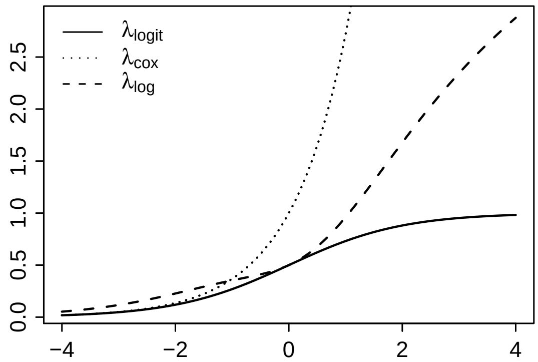

Figure 1. The three hazard rate link functions used in the simulation studies

5.1. *Performance of FAST-SIS*

We first considered the performance of basic, non-iterated FAST-SIS. Features were generated as in scenario 1 of Fan and Song (2010). Specifically, let ε be univariate standard Gaussian. Define

$$\mathbf{Z}_j := \frac{\boldsymbol{\varepsilon}_j + a_j\varepsilon}{\sqrt{1+a_j^2}}, \quad j = 1,\dots,p; \tag{27}$$

where $\boldsymbol{\varepsilon}_j$ is independently distributed as a standard Gaussian for $j = 1,2,\dots,\lfloor p/3 \rfloor$: independently distributed according to a double exponential distribution with location parameter zero and scale parameter 1 for $j = \lfloor p/3 \rfloor + 1,\dots,\lfloor 2p/3 \rfloor$; and independently distributed according to a Gaussian mixture $0.5N(-1,1)+0.5N(1,0.5)$ for $j = \lfloor 2p/3 \rfloor + 1,\dots,p$. The constants a_j satisfy $a_1 = \cdots = a_{15}$ and $a_j = 0$ for $j > 15$. With the choice $a_1 = \sqrt{\rho/(1-\rho)}$, $0 \le \rho \le 1$, we obtain $\mathrm{Cor}(Z_{1i}, Z_{1j}) = \rho$ for $i \ne j$, $i,j \le 15$, enabling crude adjustment of the correlation structure of the feature distribution. Regression coefficients were chosen to be of the generic form $\boldsymbol{\alpha}^0 = (1,1.3,1,1.3,\dots)^\top$ with exactly the first s components nonzero.

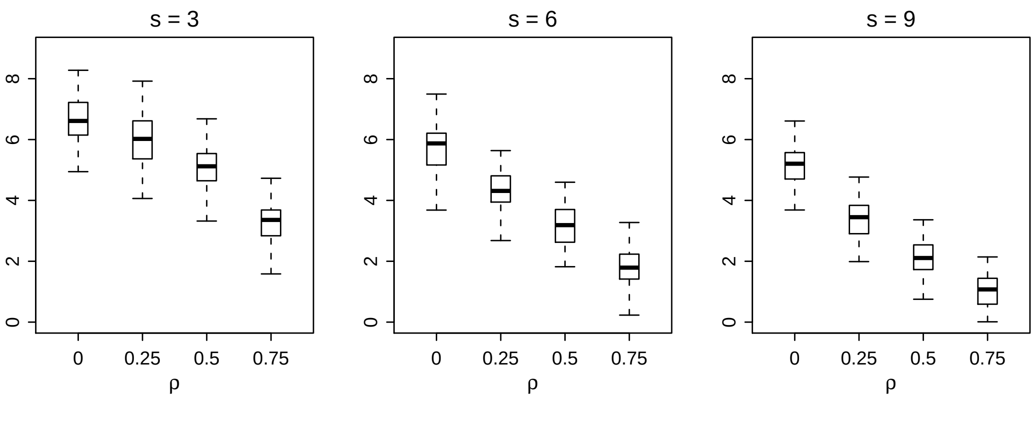

Figure 2. Minimum observed $|Z|$-statistics in the oracle model under $\lambda_{\log}$, for the SIS simulation study.

Table 1. MMMS and RSD (in parentheses) for basic SIS with $n = 300$ and $p = 20{,}000$ (100 simulations).

		λ_{logit}			λ_{cox}			$\lambda_{\log}$		
ρ		$s=3$	$s=6$	$s=9$	$s=3$	$s=6$	$s=9$	$s=3$	$s=6$	$s=9$
0	$\mathbf{d}$	3 (1)	32 (53)	530 (914)	3 (0)	7 (5)	45 (103)	3 (0)	22 (44)	202 (302)
	$\mathbf{d}^{\text{LY}}$	4 (1)	66 (95)	678 (939)	3 (0)	11 (14)	96 (176)	3 (1)	41 (87)	389 (466)
	$\mathbf{d}^{\text{Z}}$	3 (1)	40 (71)	522 (873)	3 (0)	7 (7)	48 (105)	3 (0)	22 (45)	262 (318)
	Cox	3 (1)	44 (68)	572 (928)	3 (0)	7 (4)	40 (117)	3 (0)	26 (51)	280 (306)
0.25	$\mathbf{d}$	3 (0)	6 (1)	11 (1)	3 (0)	6 (0)	9 (1)	3 (0)	6 (1)	10 (1)
	$\mathbf{d}^{\text{LY}}$	3 (0)	7 (1)	11 (2)	3 (0)	6 (1)	10 (1)	3 (0)	7 (1)	11 (1)
	$\mathbf{d}^{\text{Z}}$	3 (0)	6 (1)	11 (1)	3 (0)	6 (0)	10 (1)	3 (0)	6 (1)	10 (1)
	Cox	3 (0)	6 (1)	11 (1)	3 (0)	6 (0)	9 (1)	3 (0)	6 (1)	10 (1)
0.5	$\mathbf{d}$	3 (0)	7 (2)	12 (2)	3 (0)	6 (1)	10 (1)	3 (0)	7 (1)	11 (2)
	$\mathbf{d}^{\text{LY}}$	3 (0)	9 (3)	13 (1)	3 (0)	8 (2)	13 (2)	3 (0)	8 (2)	12 (2)
	$\mathbf{d}^{\text{Z}}$	3 (0)	8 (3)	12 (1)	3 (0)	7 (2)	12 (2)	3 (0)	7 (2)	12 (2)
	Cox	3 (1)	9 (3)	13 (2)	3 (0)	6 (1)	11 (2)	3 (0)	8 (2)	12 (2)
0.75	$\mathbf{d}$	3 (1)	9 (2)	13 (1)	3 (0)	8 (2)	12 (1)	3 (1)	9 (3)	12 (2)
	$\mathbf{d}^{\text{LY}}$	4 (2)	11 (3)	14 (2)	4 (1)	11 (3)	14 (1)	4 (2)	10 (2)	13 (1)
	$\mathbf{d}^{\text{Z}}$	4 (1)	10 (2)	13 (1)	3 (1)	10 (3)	13 (1)	3 (1)	9 (2)	13 (1)
	Cox	5 (3)	12 (2)	14 (1)	3 (0)	7 (2)	12 (2)	4 (1)	11 (3)	14 (2)

For each combination of hazard link function, non-sparsity level s, and correlation ρ, we performed 100 simulations with $p = 20{,}000$ features and $n = 300$ observations. Features were ranked using the vanilla FAST statistic, the scaled FAST statistics (15) and (16), and SIS based on a Cox working model (Cox-SIS), the latter ranking features according their absolute univariate regression coefficient. Results are shown in Table 1. As a performance measure, we report the median of the minimum model size (MMS) needed to detect all relevant features alongside its relative standard deviation (RSD), the interquartile range divided by 1.34. The MMS is a useful performance measure for this type of study since it eliminates the need to select a threshold parameter for SIS. The censoring rate in the simulations was typically 30%-40%.

For all methods, the MMMS is seen to increase with feature correlation ρ and non-sparsity s. As also noted by Fan and Song (2010) for the case of SIS for generalized linear models, some correlation among features can actually be helpful since it increases the strength of marginal signals. Overall, the statistic $\mathbf{d}^{\text{LY}}$ seems to perform slightly worse than both $\mathbf{d}$ and $\mathbf{d}^{\text{Z}}$ whereas the latter two statistics perform similarly to Cox-SIS. In our basic implementation, screening with any of the FAST statistics was more than 100 times faster than Cox-SIS, providing a rough indication of the relative computational efficiency of FAST-SIS.

To gauge the relative difficulty of the different simulation scenarios, Figure 2 shows box plots of the minimum of the observed $|Z|$-statistics in the oracle model (the joint model with only the relevant features included and estimation based on the likelihood under the true link function) for the link function $\lambda_{\log}$. This particular link function represents an 'intermediate' level of difficulty; with $|Z|$-statistics for λ_{cox} generally being somewhat larger and $|Z|$-statistics for λ_{logit} being slightly smaller. Even with oracle information and the correct working model, these are evidently difficult data to deal with.

5.2. FAST-SIS with non-Gaussian features and nonrandom censoring

We next investigated FAST-SIS with non-Gaussian features and a more complex censoring mechanism. The simulation scenario was inspired by the previous section but with all features generated according to either a standard Gaussian distribution, a t-distribution with 4 degrees of freedom, or a unit rate exponential distribution. Features were standardized to have mean zero and variance one, and the feature correlation structure was such that $\text{Cor}(Z_{1i}, Z_{1j}) = 0.125$ for $i, j < 15$, $i \neq j$ and $\text{Cor}(Z_{1i}, Z_{1j}) = 0$ otherwise. Survival times were generated according to the link function $\lambda_{\log}$ with regression coefficients $\boldsymbol{\beta} = (1, 1.3, 1, 1.3, 1, 1.3, 0, 0, \ldots)$ while censoring times were generated according to the same model (link function $\lambda_{\log}$ and conditionally on the same feature realizations) with regression coefficients $\tilde{\boldsymbol{\beta}} = k\boldsymbol{\beta}$. The constant k controls the association between censoring and survival times, leading to a basic example of nonrandom censoring (competing risks).

Using $p = 20{,}000$ features and $n = 300$ observations, we performed 100 simulations under each of the three feature distributions, for different values of k. Table 2 reports the MMMS and RSD for the four

Table 2. MMMS and RSD (in parentheses) for SIS under nongaussian features/nonrandom censoring with $n = 300$ and $p = 20,000$ (100 simulations).

			k			
Feature distr.		$k = 0$	-0.5	-0.25	0.25	0.5
Gaussian	$\mathbf{d}$	6 (1)	8 (8)	7 (4)	6 (1)	7 (3)
	$\mathbf{d}^{\mathrm{LY}}$	6 (1)	8 (6)	7 (3)	7 (2)	8 (5)
	$\mathbf{d}^{Z}$	6 (1)	7 (6)	7 (2)	6 (1)	7 (2)
	$\mathbf{d}^{\mathrm{loss}}$	6 (1)	8 (6)	7 (3)	6 (1)	7 (3)
	Cox	6 (1)	8 (5)	7 (2)	6 (1)	7 (2)
$t\ (df = 4)$	$\mathbf{d}$	6 (1)	13 (17)	7 (5)	6 (1)	7 (3)
	$\mathbf{d}^{\mathrm{LY}}$	11 (7)	12 (8)	9 (7)	48 (136)	99 (185)
	$\mathbf{d}^{Z}$	7 (3)	17 (20)	8 (5)	7 (2)	7 (3)
	$\mathbf{d}^{\mathrm{loss}}$	6 (1)	8 (7)	7 (4)	8 (15)	10 (10)
	Cox	7 (4)	15 (23)	8 (10)	8 (4)	9 (5)
Exponential	$\mathbf{d}$	6 (1)	6 (2)	6 (1)	7 (4)	8 (7)
	$\mathbf{d}^{\mathrm{LY}}$	6 (1)	11 (12)	7 (3)	6 (1)	6 (1)
	$\mathbf{d}^{Z}$	15 (10)	34 (36)	24 (17)	22 (28)	26 (29)
	$\mathbf{d}^{\mathrm{loss}}$	6 (0)	7 (4)	6 (1)	6 (1)	6 (1)
	Cox	8 (4)	22 (31)	14 (11)	9 (6)	9 (8)

different screening methods of the previous section, as well as for the statistic $\mathbf{d}^{\mathrm{loss}}$ in (17). The censoring rate in all scenarios was around 50%.

From the column with $k = 0$ (random censoring), the heavier tails of the t-distribution increases the MMMS, particularly for $\mathbf{d}^{\mathrm{LY}}$. The vanilla FAST statistic $\mathbf{d}$ seems the least affected here, most likely because it does not directly involve second-order statistics which are poorly estimated due to the heavier tails. While $\mathbf{d}^{Z}$ and $\mathbf{d}^{\mathrm{loss}}$ are also scaled by second-order statistics, the impact of the tails is dampened by the square-root transformation in the scaling factors. In contrast, the more distinctly non-Gaussian exponential distribution is problematic for $\mathbf{d}^{Z}$. Overall, the statistics $\mathbf{d}$ and $\mathbf{d}^{\mathrm{loss}}$ seems to have the best and most consistent performance across feature distributions. Nonrandom censoring generally increases the MMMS and RSD, particularly for the non-Gaussian distributions. There appears to be no clear difference between the effect of positive and negative values of k. We found that the effect of $k \neq 0$ diminished when the sample size was increased (results not shown), suggesting that nonrandom censoring in the present example leads to a power rather than bias issue. This may not be surprising in view of the considerations below (14). However, the example still shows the dramatic impact of nonrandom censoring on the performance of SIS.

5.3. Performance of FAST-ISIS

We lastly evaluated the ability of FAST-ISIS (Algorithm 1) to cope with scenarios where FAST-SIS fails. As in the previous sections, we compare our results with the analogous ISIS screening method for the Cox model. To perform Cox-ISIS, we used the R package 'SIS', with (re)recruitment based on the absolute Cox regression coefficients and variable selection based on OS-SCAD. We also compared with Z-FAST-ISIS variant described below Algorithm 1 in which (re)recruitment is based on the Lin-Ying model $|Z|$-statistics (results for FAST-ISIS with (re)recruitment based on the loss function were very similar).

For the simulations, we adopted the structural form of the feature distributions used by Fan *et al.* (2010). We considered $n = 300$ observations and $p = 500$ features which were jointly Gaussian and marginally standard Gaussian. Only regression coefficients and feature correlations differed between cases as follows:

(a) The regression coefficients are $\beta_1 = -0.96$, $\beta_2 = 0.90$, $\beta_3 = 1.20$, $\beta_4 = 0.96$, $\beta_5 = -0.85$, $\beta_6 = 1.08$ and $\beta_j = 0$ for $j > 6$. Features are independent, $\mathrm{Cor}(Z_{1i}, Z_{1j}) = 0$ for $i \neq j$.
(b) The regression coefficients are the same as in (a) while $\mathrm{Corr}(Z_{1i}, Z_{1j}) = 0.5$ for $i \neq j$.
(c) The regression coefficients are $\beta_1 = \beta_2 = \beta_3 = 4/3$, $\beta_4 = -2\sqrt{2}$. The correlation between features is given by $\mathrm{Cor}(Z_{1,4}, Z_{1j}) = 1/\sqrt{2}$ for $j \neq 4$ and $\mathrm{Cor}(Z_{1i}, Z_{1j}) = 0.5$ for $i \neq j$, $i, j \neq 4$.
(d) The regression coefficients are $\beta_1 = \beta_2 = \beta_3 = 4/3$, $\beta_4 = -2\sqrt{2}$ and $\beta_5 = 2/3$. The correlation between features is $\mathrm{Cor}(Z_{1,4}, Z_{1j}) = 1/\sqrt{2}$ for $j \notin \{4,5\}$, $\mathrm{Cor}(Z_{1,5}, Z_{1j}) = 0$ for $j \neq 5$, and $\mathrm{Cor}(Z_{1i}, Z_{1j}) = 0.5$ for $i \neq j$, $i, j \notin \{4,5\}$.

Table 3. Simulation results for ISIS with $n = 300$, $p = 500$ and $d = 17$ (100 simulations). Numbers in parentheses are standard deviations (or relative standard deviation, for the MMMS).

			Average no. true positives (ISIS)			*Average model size (ISIS)*		
Link	*Case*	*MMMS (RSD)*	LY-FAST	*Z*-FAST	Cox	LY-FAST	*Z*-FAST	Cox
λ_{logit}	(a)	7 (3)	6.0 (0)	6.0 (0)	5.5 (1)	7.8 (1)	7.9 (2)	6.3 (2)
	(b)	500 (1)	5.5 (1)	5.5 (1)	3.4 (1)	7.0 (2)	6.7 (2)	4.3 (2)
	(c)	240 (125)	3.7 (1)	3.8 (1)	3.0 (2)	5.2 (2)	5.7 (3)	4.5 (4)
	(d)	230 (124)	4.8 (1)	4.7 (1)	3.5 (2)	5.9 (2)	6.2 (3)	4.9 (4)
λ_{cox}	(a)	7 (1)	6.0 (0)	6.0 (0)	6.0 (0)	7.5 (1)	7.5 (1)	6.2 (1)
	(b)	500 (1)	5.8 (1)	5.8 (1)	5.6 (1)	7.2 (2)	6.8 (1)	6.4 (2)
	(c)	218 (120)	3.7 (1)	3.6 (1)	3.0 (2)	5.1 (3)	5.3 (3)	4.9 (4)
	(d)	258 (129)	4.9 (1)	4.8 (1)	3.8 (2)	6.3 (2)	6.0 (2)	6.4 (5)
λ_{log}	(a)	6 (1)	6.0 (0)	6.0 (0)	6.0 (0)	7.3 (1)	7.4 (1)	6.3 (1)
	(b)	500 (1)	5.8 (1)	5.7 (1)	4.9 (1)	7.2 (2)	6.7 (1)	5.7 (2)
	(c)	252 (150)	3.9 (0)	3.9 (1)	3.4 (1)	5.3 (2)	4.9 (2)	5.5 (5)
	(d)	223 (132)	4.9 (1)	4.8 (1)	4.0 (2)	6.0 (2)	6.1 (2)	5.9 (5)

Case (a) serves as a basic benchmark whereas case (b) is harder because of the correlation between relevant and irrelevant features. Case (c) introduces a strongly relevant feature Z_4 which is not marginally associated with survival; lastly, case (d) is similar to case (c) but also includes a feature Z_5 which is weakly associated with survival and does not 'borrow' strength from its correlation with other relevant features.

Following Fan *et al.* (2010), we took $d = \lfloor n/\log n/3 \rfloor = 17$ for the initial dimension reduction; performance did not depend much on the detailed choice of d of order $n/\log n$. For the three different screening methods, ISIS was run for maximum of 5 iterations. (P)BIC was used for tuning the variable selection steps. Results are shown in Table 3, summarized over 100 simulations. We report the average number of truly relevant features selected by ISIS and the average final model size, alongside standard deviations in parentheses. To provide an idea of the improvement over basic SIS, we also report the median of the minimum model size (MMMS) for the initial SIS step (based on vanilla FAST-SIS only). The censoring rate in the different scenarios was 25%-35%.

The overall performance of the three ISIS methods is comparable between the different cases. All methods deliver a dramatic improvement over non-iterated SIS, but no one method performs significantly better than the others. The two FAST-ISIS methods have a surprisingly similar performance. As one would expect, Cox-ISIS does particularly well under the link function λ_{cox} but does not appear to be uniformly better than the two FAST-ISIS methods even in this ideal setting. Under the link function λ_{logit}, both FAST-ISIS methods outperform Cox-ISIS in terms of the number of true positives identified, as do they for the link function λ_{log}, although less convincingly. On the other hand, the two FAST-ISIS methods generally select slightly larger models than Cox-ISIS and their false-positive rates (not shown) are correspondingly slightly larger. FAST-ISIS was 40-50 times faster than Cox-ISIS, typically completing calculations in 0.5-1 seconds in our specific implementation. Figure 3 shows box plots of the minimum of the observed $|Z|$-statistics in the oracle model (based on the likelihood undebr the true model).

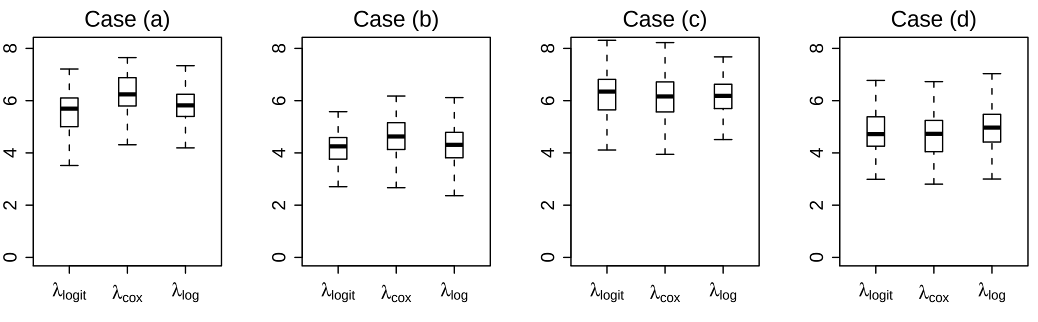

Figure 3. Minimum observed $|Z|$-statistics in the oracle models for the FAST-ISIS simulation study.

We have experimented with other link functions and feature distributions than those described above (results not shown). Generally, we found that Cox-ISIS performs worse than FAST-ISIS for bounded link functions. The observation from Table 3, that FAST-ISIS may improve upon Cox-ISIS even under the link function λ_{cox}, does not necessarily hold when the signal strength is increased. Then Cox-ISIS will be superior, as expected. Changing the feature distribution to one for which the linear regression property (Assumption 2) does not hold leads to a decrease in the overall performance for all three ISIS methods.

6. Application to AML data

The study by Metzeler *et al.* (2008) concerns the development and evaluation of a prognostic gene expression marker for overall survival among patients diagnosed with cytogenetically normal acute myeloid leukemia (CN-AML). A total of 44,754 gene expressions were recorded among 163 adult patients using Affymetrix HG-U133 A1B microarrays. Based the method of supervised principal components (Bair and Tibshirani, 2004), the gene expressions were used to develop an 86-gene signature for predicting survival. The signature was validated on an external test data set consisting of 79 patients profiled using Affymetrix HG-U133 Plus 2.0 microarrays. All data is publicly available on the Gene Expression Omnibus web site (http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE12417. The CN-AML data was recently used by Benner *et al.* (2010) for comparing the performance of variable selection methods.

Median survival time was 9.7 months in the training data (censoring rate 37%) and 17.7 months in the test data (censoring rate 41%). Preliminary to analysis, we followed the scaling approach employed by Metzeler *et al.* (2008) and centered the gene expressions separately within the test and training data set, followed by a scaling of the training data with respect to the test data.

We first applied vanilla FAST-SIS to the $n = 163$ patients in the training data to reduce the dimension from $p = 44{,}754$ to $d = \lfloor n/\log(n) \rfloor = 31$. We then used OS-SCAD to select a final set among these 31 genes. Since the PBIC criterion can be somewhat conservative in practice, we selected the OS-SCAD tuning parameter using 5-fold cross-validation based on the loss function (18). Specifically, using a random split of $\{1,\ldots,163\}$ into folds $F_1,\ldots,F_5$ of approximately equal size, we chose λ as:

$$\hat{\lambda} = \operatorname{argmin}_\lambda \sum_{i=1}^{5} L^{(F_i)}\{\hat{\boldsymbol{\beta}}_{-F_i}(\lambda)\};$$

with $L^{(F_i)}$ the loss function using only observations from F_i and $\hat{\boldsymbol{\beta}}_{-F_i}(\lambda)$ the regression coefficients estimated for a tuning parameter λ, omitting observations from F_i. This approach yielded a set of 7 genes, 5 of which also appeared in the signature of Metzeler *et al.* (2008). For $\hat{\boldsymbol{\beta}}$ the estimated penalized regression coefficients, we calculated a risk score $\boldsymbol{Z}_j^\top \hat{\boldsymbol{\beta}}$ for each patient in the test data. In a Cox model, the standardized risk score had a hazard ratio of 1.69 ($p = 6 \cdot 10^{-4}$; Wald test). In comparison, lasso based on the Lin-Ying model (Leng *et al.* (2007); Martinussen and Scheike (2009)) with 5-fold cross-validation gave a standardized risk score with a hazard ratio of 1.56 ($p = 0.003$; Wald test) in the test data, requiring 5 genes; Metzeler *et al.* (2008) reported a hazard ratio of 1.85 ($p = 0.002$) for their 86-gene signature.

We repeated the above calculations for the three scaled versions of the FAST statistic (15)-(17). Since assessment of prediction performance using only a single data set may be misleading, we also validated the screening methods via leave-one-out (LOO) cross-validation based on the 163 patients in the training data. For each patient j, we used FAST-SIS as above (or Lin-Ying lasso) to obtain regression coefficients $\hat{\boldsymbol{\beta}}_{-j}$ based on the remaining 162 patients and defined the jth LOO risk score as the percentile of $\mathbf{Z}_j^\top \hat{\boldsymbol{\beta}}_{-j}$ among $\{\mathbf{Z}_i^\top \hat{\boldsymbol{\beta}}_{-j}\}_{i \neq j}$. We calculated Wald p-values in a Cox regression model including the LOO score as a continuous predictor. Results are shown in Table 4 while Table 5 shows the overlap between gene sets selected in the training data. There is seen to be some overlap between the different methods, particularly between vanilla FAST-SIS and the lasso, and many of the selected genes also appear in the signature of Metzeler *et al.* (2008). In the test data, the prediction performance of the different screening methods was comparable whereas the lasso had a slight edge in the LOO calculations. Lin-Ying SIS selected only a single gene in the test data and typically selected no genes in the LOO calculations. We found FAST screening to be slightly more sensitive to the cross-validation procedure than the lasso.

We next evaluated the extent to which iterated FAST-SIS might improve upon the above results. From our limited experience with applying ISIS to real data, instability can become an issue when several iterations of ISIS are run; particularly when cross-validation is involved. Accordingly, we ran only a single iteration of ISIS using $|Z|$-FAST-ISIS. The algorithm kept 2 of the genes from the first FAST-SIS round and selected 3 additional genes so that the total number of genes was 5. Calculating in the test data a standardized risk score based on the final regression coefficients, we obtained a Cox hazard ratio of only

Table 4. Prediction performance of FAST-SIS and Lin-Ying lasso in the AML data, evaluated in terms of the Cox hazard ratio for the standardized continuous risk score. The LOO calculations are based on the training data only.

		Screening method				
Scenario	*Summary statistic*	$\mathbf{d}$	$\mathbf{d}^{\mathrm{LY}}$	$\mathbf{d}^{\lvert Z\rvert}$	$\mathbf{d}^{\mathrm{loss}}$	Lasso
Test data	Hazard ratio	1.69	1.59	1.46	1.58	1.54
	p-value	$6 \cdot 10^{-4}$	0.0007	0.01	0.002	0.004
	No. predictors	7	1	3	7	5
LOO	p-value	$4 \cdot 10^{-7}$	0.16	$5 \cdot 10^{-5}$	$4 \cdot 10^{-4}$	$4 \cdot 10^{-8}$
	Median no. predictors	7	0	3	5	5

Table 5. Overlap between gene sets selected by the different screening methods and the signature of Metzeler *et al.* (2008).

	$\mathbf{d}$	$\mathbf{d}^{\mathrm{LY}}$	$\mathbf{d}^{\lvert Z\rvert}$	$\mathbf{d}^{\mathrm{loss}}$	Lasso	Metzeler
$\mathbf{d}$	7	0	1	2	4	5
$\mathbf{d}^{\mathrm{LY}}$		1	0	0	0	0
$\mathbf{d}^{\lvert Z\rvert}$			3	2	2	2
$\mathbf{d}^{\mathrm{loss}}$				7	2	5
Lasso					5	5
Metzeler						86

1.06 ($p = 0.6$; Wald test) which is no improvement over non-iterated FAST-SIS. A similar conclusion was reached for the corresponding LOO calculations in the training data which gave a Cox Wald p-value of 0.001 for the LOO risk score, using a median of 4 genes. None of the other FAST-ISIS methods lead to improved prediction performance compared to their non-iterated counterparts.

FAST-ISIS runs swiftly on this large data set: one iteration of the algorithm (re-recruitment and OSS-SCAD feature selection with 5-fold cross-validation) completes in under 5 seconds on a standard laptop.

Altogether, the example shows that FAST-SIS can compete with a computationally more demanding full-scale variable selection method in the sense of providing similarly sparse models with competitive prediction properties. FAST-ISIS, while computationally very feasible, did not seem to improve prediction performance over simple independent screening in this particular data set.

7. Discussion

Independent screening – the general idea of looking at the effect of one feature at a time – is a well-established method for dimensionality reduction. It constitutes a simple and excellently scalable approach to analyzing high-dimensional data. The SIS property introduced by Fan and Lv (2008) has enabled a basic formal assessment of the reasonableness of general independent screening methods. Although the practical relevance of the SIS property has been subject to scepticism (Roberts, 2008), the formal context needed to develop the SIS property is clearly useful for identifying the many implicit assumptions made when applying univariate screening methods to multivariate data.

We have introduced a SIS method for survival data based on the notably simple FAST statistic. In simulation studies, FAST-SIS performed on par with SIS based on the popular Cox model, while being considerably more amenable to analysis. We have shown that FAST-SIS may admit the formal SIS property within a class of single-index hazard rate models. In addition to assumptions on the feature distribution which are well known in the literature, a principal assumption for the SIS property to hold is that censoring times do not depend on the relevant features nor survival. While such partially random censoring may be appropriate to assume in many clinical settings, it indicates that additional caution is called for when applying univariate screening and competing risks are suspected.

A formal consistency property such as the SIS property is but one aspect of a statistical method and does not make FAST-SIS universally preferable. Not only is the SIS property unlikely to be unique to FAST screening, but different screening methods often highlight different aspects of data (Ma and Song, 2011), making it impossible and undesirable to recommend one generic method. We do, however, consider FAST-SIS a good generic choice of initial screening method for general survival data. Ultimately, the initial choice of a statistical method is likely to be made on the basis of parsimony, computational speed, and ease of implementation. The FAST statistic is about as difficult to evaluate as a collection of correlation coefficients while iterative FAST-SIS only requires solving one linear system of equations. This yields substantial computational savings over methods not sharing the advantage of linearity of estimating equations.

Iterated SIS has so far been studied to a very limited extent in an empirical context. The iterated approach works well on simulated data, but it is not obvious whether this necessarily translates into good performance on real data. In our example involving a large gene expression data set, ISIS did not improve results in terms of prediction accuracy. Several issues may affect the performance of ISIS on real data. First, it is our experience that the 'Rashomon effect', the multitude of well-fitting models (Breiman, 2001), can easily lead to stability issues for this type of forward selection. Second, it is often difficult to choose a good tuning parameter for the variable selection part of ISIS. Using BIC may lead to overly conservative results, whereas cross-validation may lead to overfitting when only the variable selection step – and not the recruitment steps – are cross-validated. He and Lin (2011) recently discussed how to combine ISIS with stability selection (Meinshausen and Bühlmann, 2010) in order to tackle instability issues and to provide a more informative output than the concise 'list of indices' obtained from standard ISIS. Their proposed scheme requires running many subsampling iterations of ISIS, a purpose for which FAST-ISIS will be ideal because of its computational efficiency. The idea of incorporating stability considerations is also attractive from a foundational point of view, being a pragmatic departure from the limiting *de facto* assumption that there is a single, true model. Investigation of such computationally intensive frameworks, alongside a study of the behavior of ISIS on a range of different real data sets, is a pertinent future research topic.

A number of other extensions of our work may be of interest. We have focused on the important case of time-fixed features and right-censored survival times but the FAST statistic can also be used with time-varying features alongside other censoring and truncation mechanism supported by the counting process formalism. Theoretical analysis of such extensions is a relevant future research topic, as is analysis of more flexible, time-dependent scaling strategies for the FAST statistic. Fan *et al.* (2011) recently discussed SIS where features enter in nonparametric, smooth manner, and an extension of their framework to FAST-SIS appears both theoretically and computationally feasible. Lastly, the FAST statistic is closely related to the univariate regression coefficients in the Lin-Ying model which is rather forgiving towards misspecification: under feature independence, the univariate estimator is consistent whenever the particular feature under investigation enters the hazard rate model as a linear function of regression coefficients (Hattori, 2006). The Cox model does not have a similar property (Struthers and Kalbfleisch, 1986). Whether such internal consistency under misspecification or lack hereof affects screening in a general setting is an open question.

Appendix: proofs

In addition to Assumptions 1-4 stated in the main text, we will make use of the following assumptions for the quantities defining the class of single-index hazard rate models (7):

A. $E(Z_{1j}) = 0$ and $E(Z_{1j}^2) = 1$, $j = 1, 2, \ldots, p_n$.
B. $P\{Y_1(\tau) = 1\} > 0$.
C. $\mathrm{Var}(\mathbf{Z}_1^\top \boldsymbol{\alpha}^0)$ is uniformly bounded above.

The details in Assumption A are included mainly for convenience; it suffices to assume that $E(Z_{1j}^2) < \infty$.

Our first lemma is a basic symmetrization result, included for completeness.

Lemma 1. Let X be a random variable with mean μ and finite variance σ^2. For $t > \sqrt{8}\sigma$, it holds that $P(|X-\mu| > t) \leq 4P(|X| > t/4)$.

Proof. First note that when $t > \sqrt{8}\sigma$ we have $P(|X-\mu| > t/2) \leq 1/2$, by Chebyshev's inequality. Let X' be an independent copy of X. Then

$$2P(|X| \geq t/4) \geq P(|X'-X| > t/2) \geq P(|X-\mu| > t \wedge |X'-\mu| \leq t/2). \tag{28}$$

But

$$P(|X-\mu| > t \wedge |X'-\mu| \leq t/2) = P(|X-\mu| > t)P(|X'-\mu| \leq t/2) \geq \frac{1}{2}P(|X-\mu| > t).$$

Combining this with (28), the statement of the lemma follows. □

The next lemma provides a universal exponential bound for the FAST statistic and is of independent interest. It bears some similarity to exponential bounds reported by Bradic *et al.* (2010) for the Cox model.

Lemma 2. Under assumptions A-B there exists constants $C_1, C_2, C_3 > 0$ independent of n such that for any $K > 0$ and $1 \leq j \leq p_n$, it holds that

$$P\{n^{1/2}|d_j - \delta_j| > C_1(1+t)\} \leq 10\exp\{-t^2/(2K^2)\} + C_2\exp(-C_3 n) + nP(|Z_{1j}| > K).$$

Proof. Fix j throughout. Assume first that $|Z_{ij}| \leq K$ for some finite K. Define the random variables

$$A_n := n^{-1}\sum_{i=1}^{n}\int_0^\tau \{Z_{ij} - e_j(t)\}\mathrm{d}N_i(t), \quad B_n := \int_0^\tau \{\bar{Z}_j(t) - e_j(t)\}\mathrm{d}\bar{N}(t);$$

where $\bar{N}(t) := n^{-1}\{N_1(t) + \cdots + N_n(t)\}$ and $e_j(t) = E\{\bar{Z}_j(t)\}$. Then we can write

$$n^{1/2}(d_j - \delta_j) = n^{1/2}\{A_n - E(A_n)\} + n^{1/2}\{B_n - E(B_n)\}.$$

We will deal with each term in the display separately. Since $\mathrm{d}N_i(t) \leq 1$, it holds that

$$|A_n| \leq \max_{1\leq i\leq n}|Z_{ij}| + \|e_j\|_\infty \leq 2K.$$

and Hoeffding's inequality (Hoeffding, 1963) implies

$$P(n^{1/2}|A_n - E(A_n)| > t) \leq 2\exp\{-t^2/(2K^2)\}. \tag{29}$$

Obtaining an analogous bound for $n^{1/2}\{B_n - E(B_n)\}$ requires a more detailed analysis. Since $\mathrm{d}\bar{N}(t) \leq 1$,

$$|B_n| \leq \int_0^\tau |\bar{Z}_j(t) - e_j(t)|\mathrm{d}\bar{N}(t) \leq \|\bar{Z}_j - e_j\|_\infty. \tag{30}$$

We will obtain an exponential bound for the right-hand side via empirical process methods. Define $E^{(k)}(t) := n^{-1}\sum_{i=1}^n Z_{ij}^k Y_i(t)$ and $e^{(k)}(t) := E\{E^{(k)}(t)\}$ for $k = 0, 1$. Denote $m := \inf_{t\in[0,\tau]} e^{(0)}(t)$ and observe that $m > 0$, by Assumption B. Moreover, by Cauchy-Schwartz's inequality,

$$\|e^{(1)}/e^{(0)}\|_\infty \leq m^{-1}\sqrt{E|Z_{1j}|^2 e^{(0)}(t)} \leq m^{-1}.$$

Define $\Omega_n := \{\inf_{t\in[0,\tau]} E^{(0)}(t) \geq m/2\}$ and let 1_{Ω_n} be the indicator of this event. In view of the preceding display, we can write

$$|\bar{Z}_j(t) - e_j(t)|1_{\Omega_n} \leq \frac{1}{E^{(0)}(t)}\Big\{\Big|\frac{e^{(1)}(t)}{e^{(0)}(t)}\Big||e^{(0)}(t) - E^{(0)}(t)| + |E^{(1)}(t) - e^{(1)}(t)|\Big\}1_{\Omega_n} \tag{31}$$

$$\leq 2m^{-2}(\|P_n - P\|_{\mathrm{F}_0} + \|P_n - P\|_{\mathrm{F}_1})1_{\Omega_n} \tag{32}$$

with function classes $\mathrm{F}_k := \{t \mapsto Z^k 1(T \geq t \wedge C \geq t)\}$. We proceed to establish exponential bounds for the empirical process suprema in (32). Each of the F_ks are Vapnik-Cervonenkis subgraph classes, and from Pollard (1989) there exists some finite constant ζ depending only on intrinsic properties of the F_ks such that

$$E(\|P_n - P\|_{\mathrm{F}_k}^2) \leq \zeta n^{-1}E(Z_{1j}^2) = n^{-1}\zeta. \tag{33}$$

In particular, it also holds that $E(\|P_n - P\|_{\mathrm{F}_k}) \leq n^{-1/2}\zeta^{1/2}$. Moreover,

$$|Z_{1j}^k 1(T_1 \geq t \wedge C_1 \geq t) - Z_{1j}^k 1(T_1 \geq s \wedge C_1 \geq s)|^2 \leq K^{2k}, \quad s,t \in [0,\tau].$$

With $k_1 := \zeta^{1/2}$, the concentration theorem of Massart (2000) implies

$$P\{n^{1/2}\|P_n - P\|_{\mathrm{F}_k} > k_1(1+t)\} \leq \exp\{-t^2/(2K^2)\}, \quad k = 0,1. \tag{34}$$

Combining (30)-(32), taking $k_2 := k_1 m^2/2$, we obtain

$$P(\{n^{1/2}|B_n| > k_2(1+t)\} \cap \Omega_n) \leq 2\exp\{-t^2/(2K^2)\}. \tag{35}$$

whereas Cauchy-Schwarz's inequality implies

$$E(B_n^2 1_{\Omega_n}) \leq E\|\bar{Z}_j - e_j\|_\infty^2 1_{\Omega_n} \leq 4m^{-4}E\{(\|P_n - P\|_{\mathrm{F}_0} + \|P_n - P\|_{\mathrm{F}_1})^2\}1_{\Omega_n} \leq 12m^{-4}\zeta n^{-1}.$$

Combining Lemma 1 and (35), there exists nonnegative k_3 (depending only on m and ζ) such that

$$P\{n^{1/2}|B_n - E(B_n)| \geq k_3(1+t)\} \leq 8\exp\{-t^2/(2K^2)\} + P(\Omega_n^c). \tag{36}$$

To bound $P(\Omega_n^c)$, recall that $e^{(0)}(t) \geq m$ by assumption. Consequently,

$$\Omega_n^c \subseteq \{|E^{(0)}(t) - e^{(0)}(t)| > m/2 \text{ for some } t\} \subseteq \{\|P_n - P\|_{\mathrm{F}_0} > m/2\}.$$

By (33), we have $E(\|P_n - P\|_{\mathrm{F}_0}) \leq m/4$ eventually. By another application of the concentration theorem (Massart, 2000), there exists finite k_4 so that $P\{\|P_n - P\|_{\mathrm{F}_0} > m/4(1+t)\} \leq k_4\exp(-nt^2/2)$. Setting $t = 1$,

$$P(\Omega_n^c) \leq P(\|P_n - P\|_{\mathrm{F}_0} > m/2) \leq k_4\exp(-n/2).$$

Substituting this bound in (36) and combining with (29), omitting now the assumption that Z_{ij} is bounded, it follows that there exists constants $C_1, C_2, C_3 > 0$ such that for any $K > 0$ and $t > 0$,

$$P\{n^{1/2}|d_j - \delta_j| > C_1(1+t)\} \leq 10\exp\{-t^2/(2K^2)\} + C_2\exp(-C_3 n) + P\Big(\max_{1\leq i\leq n}|Z_{ij}| > K\Big).$$

The statement of the lemma then follows from the union bound. □

Lemma 3. Suppose that Assumptions A-B hold and that there exists positive constants l_0, l_1, η such that $P(|Z_{1j}| > s) \leq l_0\exp(-l_1 s^\eta)$ for sufficiently large s. If $\kappa < 1/2$ then for any $k_1 > 0$ there exists $k_2 > 0$ such that

$$P\Big(\max_{1\leq j\leq p_n}|d_j - \delta_j| > k_1 n^{-\kappa}\Big) \leq O[p_n\exp\{-k_2 n^{(1-2\kappa)\eta/(\eta+2)}\}]. \tag{37}$$

Suppose in addition that $|\delta_j| > k_3 n^{-\kappa}$ whenever $j \in \mathrm{M}_\delta^n$ and that $\gamma_n = k_4 n^{-\kappa}$ where k_3, k_4 are positive constants and $k_4 \leq k_3/2$. Then

$$P(\mathrm{M}_\delta^n \subseteq \widehat{\mathrm{M}}_d^n) \geq 1 - O[p_n\exp\{-k_2 n^{(1-2\kappa)\eta/(\eta+2)}\}]. \tag{38}$$

In particular, if $\log p_n = o\{n^{(1-2\kappa)\eta/(\eta+2)}\}$ then $P(\mathrm{M}_\delta^n \subseteq \widehat{\mathrm{M}}_d^n) \to 1$ when $n \to \infty$.

Proof. In Lemma 2, take $1 + t = k_1 n^{1/2-\kappa}/C_1$ and $K := n^{(1-2\kappa)/(\eta+2)}$. Then there exists positive constants $\tilde{k}_2, \tilde{k}_3$ such that for each $j = 1, \ldots, p_n$,

$$P(|d_j - \delta_j| > k_1 n^{-\kappa}) \leq 10\exp\{-\tilde{k}_2 n^{(1-2\kappa)\eta/(\eta+2)}\} + nl_0\exp\{-\tilde{k}_3 n^{(1-2\kappa)\eta/(\eta+2)}\}.$$

By the union bound, there exists $k_2 > 0$ such that

$$P\Big(\max_{1\leq j\leq p_n}|d_j - \delta_j| > k_1 n^{-\kappa}\Big) \leq O[p_n\exp\{-k_2 n^{(1-2\kappa)\eta/(\eta+2)}\}];$$

which proves (37). Concerning (38), $k_3 n^{-\kappa} - |d_j| \leq |\delta_j - d_j|$ by assumption and so

$$P\Big(\min_{j\in\mathrm{M}_\delta^n}|d_j| < \gamma_n\Big) \leq P\Big(\max_{j\in\mathrm{M}_\delta^n}|d_j - \delta_j| \geq k_4 n^{-\kappa} - \gamma_n\Big) \leq P\Big(\max_{j\in\mathrm{M}_\delta^n}|d_j - \delta_j| \geq n^{-\kappa}k_3/2\Big);$$

where the last inequality follows since we assume $k_4 \leq k_3/2$. Taking $k_1 = k_3/2$ in (37), we arrive at the desired conclusion:

$$P(\mathrm{M}_\delta^n \subseteq \widehat{\mathrm{M}}_d^n) \geq 1 - P\Big(\min_{j\in\mathrm{M}_\delta^n}|d_j| < \gamma_n\Big) \geq 1 - O[p_n\exp\{-k_2 n^{(1-2\kappa)\eta/(\eta+2)}\}].$$

Finally, $P(\mathrm{M}_\delta^n \subseteq \widehat{\mathrm{M}}_d^n) \to 1$ when $n \to \infty$ follows immediately when $\log p_n = o\{n^{(1-2\kappa)\eta/(\eta+2)}\}$. □

Lemma 4. Let $\mathbf{Z} \in \mathbb{R}^p$ be a random vector with zero mean and covariance matrix $\mathbf{\Sigma}$. Let $\mathbf{b} \in \mathbb{R}^p$ and suppose that $E(\mathbf{Z}|\mathbf{Z}^\top\mathbf{b}) = \mathbf{c}\mathbf{Z}^\top\mathbf{b}$ for some constant vector $\mathbf{c} \in \mathbb{R}^p$. Assume that f is some real function. Then

$$E\{\mathbf{Z}f(\mathbf{Z}^\top\mathbf{b})\} = \mathbf{\Sigma}\mathbf{b}\frac{E\{\mathbf{Z}^\top\mathbf{b}f(\mathbf{Z}^\top\mathbf{b})\}}{\mathrm{Var}(\mathbf{Z}^\top\mathbf{b})}; \tag{39}$$

taking $0/0 := 0$. If moreover f is differentiable and strictly monotonic, there exists $\varepsilon > 0$ such that

$$E|\mathbf{Z}f(\mathbf{Z}^\top\mathbf{b})| \geq \mathbf{\Sigma}\mathbf{b}\varepsilon/\mathrm{Var}(\mathbf{Z}^\top\mathbf{b}). \tag{40}$$

In particular, $E\{Z_j f(\mathbf{Z}^\top\mathbf{b})\} = 0$ iff $\mathrm{Cov}(Z_j, \mathbf{Z}^\top\mathbf{b}) = 0$.

Proof. Set $W := \mathbf{Z}^\top \mathbf{b}$. By standard properties of conditional expectations, it holds that

$$0 = E\{W(\mathbf{Z} - E(\mathbf{Z}|W))\} = \mathbf{\Sigma b} - E\{WE(\mathbf{Z}|W)\} = \mathbf{\Sigma b} - \mathbf{c}E(W^2),$$

implying $E(\mathbf{Z}|W) = \mathbf{\Sigma b}W/\mathrm{Var}(W)$. We then obtain (39):

$$E\{\mathbf{Z}f(\mathbf{Z}^\top \mathbf{b})\} = E\{E(\mathbf{Z}|W)f(W)\} = \mathbf{\Sigma b}E\{Wf(W)\}/\mathrm{Var}(W).$$

To show (40), the mean value theorem implies the existence of some $0 < \tilde{W} < W$ such that

$$E(Wf(W)) = E[W\{f(0) + f'(\tilde{W})W] = E\{W^2 f'(\tilde{W})\}.$$

Then

$$E|W^2 f'(\tilde{W})| \geq E\{|f'(\tilde{W})|W^2 1(W^2 \leq 1)\} \geq \inf_{0\leq x\leq 1} |f'(x)|E\{W^2 1(W^2 \leq 1)\}.$$

Strict monotonicity of f then yields (40). □

Lemma 5. Assume that the survival time T has a general, continuous hazard rate function $\lambda_T(t|Z)$ depending on the random variable $Z \in \mathbb{R}$ and that the censoring time C is independent of Z, T. Then

$$\delta = \int_0^\tau \tilde{e}(t)\mathrm{d}F(t) = E\{\tilde{e}(T \wedge C \wedge \tau)\};$$

where $F(t) := P(T \wedge C \wedge \tau \leq t)$ and $\tilde{e}(t) := E\{ZP(T \geq t|Z)\}/E\{P(T \geq t)\}$.

Proof. Let S_T, S_C denote the survival functions of T, C, conditionally on Z. Using the expression (8) for δ alongside the assumption of random censoring, we obtain

$$\begin{aligned}
\delta &= E\Big[\int_0^\tau \{Z - e(t)\}Y(t)\lambda_T(t|Z)\mathrm{d}t\Big] &(41)\\
&= \int_0^\tau S_C(t)E\{ZS_T(t)\lambda_T(t|Z)\}\mathrm{d}t - \int_0^\tau \frac{E\{ZS_T(t)\}}{E\{Y(t)\}}S_C(t)E\{Y(t)\lambda_T(t|Z)\}\mathrm{d}t &(42)\\
&= -\int_0^\tau \frac{\mathrm{d}}{\mathrm{d}t}\tilde{e}(t)E\{Y(t)\}\mathrm{d}t; &(43)
\end{aligned}$$

where last equality follows since $S_T' = -\lambda_T S_T$. Integrating by parts, we obtain the statement of the lemma:

$$\delta = -\int_0^\tau \frac{\mathrm{d}}{\mathrm{d}t}\tilde{e}(t)E\{Y(t)\}\mathrm{d}t = -\int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t}\tilde{e}(t)E\{P(T \wedge C \wedge \tau \geq t|Z)\}\mathrm{d}t = E\{\tilde{e}(T \wedge C \wedge \tau)\}.$$

□

Proof of Theorem 1. Set $\tilde{e}_j(t) := E\{Z_{1j}S_T(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)\}/E\{S_T(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)\}$ with $S_T(t, \cdot) = \exp\{-\int_0^t \lambda(s, \cdot)\mathrm{d}s\}$. From Assumptions 1-3, Assumption C, and Lemma 4, there exists a universal positive constant k_1 such that

$$|\delta_j| = |\tilde{e}_j(t)| \geq |E\{Z_{1j}S_T(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)\}| \geq k_1|\mathrm{Cov}(Z_{1j}, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)|, \quad j \in \mathrm{M}^n.$$

Then $\mathrm{M}^n \subseteq \mathrm{M}^n_\delta$. The sure screening property follows from Lemma 3 and the assumptions. □

Proof of Theorem 2. Suppose that

$$\|\boldsymbol{\delta}\|^2 = O\{\lambda_{\max}(\mathbf{\Sigma})\}. \tag{44}$$

For any $\varepsilon > 0$, on the set $B_n := \{\max_{1\leq j\leq p_n} |d_j - \delta_j| \leq \varepsilon n^{-\kappa}\}$, it then holds that

$$|\{1 \leq j \leq p_n : |d_j| > 2\varepsilon n^{-\kappa}\}| \leq |\{1 \leq j \leq p_n : |\delta_j| > \varepsilon n^{-\kappa}\}| \leq O\{n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma})\}.$$

Taking $k_1 = 2\varepsilon$ in Lemma 3, we have

$$P[|\widehat{\mathrm{M}}^n_d| \leq O\{n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma})\}] \geq P[|\{j : |d_j| > k_1 n^{-\kappa}\}| \leq O\{n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma})\}] \geq P(B_n).$$

By Lemma 3, $P(B_n) = 1 - O[p_n \exp\{-c_3 n^{(1-2\kappa)\eta/(\eta+2)}\}]$ as claimed. So we need only verify (44).

By Lemma 5, there exists a positive constant c_1 such that $|\delta_j| \le c_1 \int_0^\tau |E\{Z_{1j} S_T(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)\}| \mathrm{d}F(t)$ for $j \in \mathrm{M}^n$ with F the unconditional distribution function of $T_1 \wedge C_1 \wedge \tau$. In contrast, $\delta_j = 0$ for $j \notin \mathrm{M}^n$, by Assumptions 3-4. It follows from Jensen's inequality that there exists a positive constant c_2 such that

$$\|\boldsymbol{\delta}\|^2 \le c_2 \int_0^\tau \|E\{\mathbf{Z}_1 S_T(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)\}\|^2 \mathrm{d}F(t). \tag{45}$$

Lemma 4 implies

$$E\{\mathbf{Z}_1 S_T(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)\} = \frac{E\{\mathbf{Z}_1^\top \boldsymbol{\alpha}^0 S_T(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)\}}{\mathrm{Var}(\mathbf{Z}_1^\top \boldsymbol{\alpha}^0)} \boldsymbol{\Sigma}\boldsymbol{\alpha}^0. \tag{46}$$

By Cauchy-Schwartz's inequality, using that $\|\boldsymbol{\Sigma}\boldsymbol{\alpha}^0\|^2 \le \|\boldsymbol{\Sigma}^{1/2}\|^2 \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}^0\|^2 \le \lambda_{\max}(\boldsymbol{\Sigma}) \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}^0\|^2$,

$$\|E\{\mathbf{Z}_1 S_T(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)\}\|^2 \le \|\boldsymbol{\Sigma}\boldsymbol{\alpha}^0\|^2 / \mathrm{Var}(\mathbf{Z}_1^\top \boldsymbol{\alpha}^0) \le \lambda_{\max}(\boldsymbol{\Sigma}).$$

Inserting this in (45) then yields the desired result (44). Note that this result does not rely on the uniform boundedness of $\mathrm{Var}(\mathbf{Z}_1^\top \boldsymbol{\alpha}^0)$ (Assumption C). □

Lemma 6. Suppose that Assumption A holds and that both the survival time T_1 and censoring time C_1 follow a nonparametric Aalen model (11) with time-varying parameters $\boldsymbol{\alpha}^0$ and $\boldsymbol{\beta}^0$, respectively. Suppose moreover that $\mathbf{Z}_1 = \boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{Z}}_1$ where $\tilde{\mathbf{Z}}_1$ has i.i.d. components and denote by $\phi(x) := E\{\exp(\tilde{Z}_{j1} x)\}$ the moment generating function of $\tilde{Z}_{j1}$. Then

$$\boldsymbol{\delta} = \boldsymbol{\Sigma}^{1/2} \Big[\int_0^\tau \mathrm{diag}\Big\{ \frac{\mathrm{d}}{\mathrm{d}x} \frac{\phi'(x)}{\phi(x)} \Big|_{x=-\Gamma_j^0(t)} \Big\} E\{Y_1(t)\} \boldsymbol{\alpha}^0(t)^\top \mathrm{d}t \Big] \boldsymbol{\Sigma}^{1/2}; \tag{47}$$

where $\boldsymbol{\Gamma}^0(t) := \boldsymbol{\Sigma}^{1/2} \int_0^t \{\boldsymbol{\alpha}^0(s) + \boldsymbol{\beta}^0(s)\} \mathrm{d}s$. In particular, if $\mathbf{Z}_1 \sim \mathrm{N}(0, \boldsymbol{\Sigma})$ then

$$\boldsymbol{\delta} = \boldsymbol{\Sigma} \Big\{ \int_0^\tau \boldsymbol{\alpha}^0(t) E\{Y_1(t)\} \mathrm{d}t \Big\}. \tag{48}$$

Proof. Let Λ_T and Λ_C denote the cumulative baseline hazard functions associated with T_1 and C_1. Combining (8) and (11), we get

$$\begin{aligned}
\boldsymbol{\delta} &= E\Big\{ \int_0^\tau \mathbf{Z}_1 \mathbf{Z}_1^\top Y_1(t) \boldsymbol{\alpha}^0(t) \mathrm{d}t \Big\} - \int_0^\tau E\{\mathbf{Z}_1 Y_1(t)\}^{\otimes 2} E\{Y_1(t)\}^{-1} \boldsymbol{\alpha}^0(t) \mathrm{d}t && (49)\\
&= \int_0^\tau \boldsymbol{\Sigma}^{1/2} \mathbf{H}(t) \boldsymbol{\Sigma}^{1/2} E\{Y_1(t)\} \boldsymbol{\alpha}^0(t) \mathrm{d}t; && (50)
\end{aligned}$$

defining here

$$\mathbf{H}(t) := \frac{E\{Y_1(t)\} E\{\tilde{\mathbf{Z}}_1 \tilde{\mathbf{Z}}_1^\top Y_1(t)\} - E\{\tilde{\mathbf{Z}}_1 Y_1(t)\}^{\otimes 2}}{E\{Y_1(t)\}^2}.$$

Since we have $Y_1(t) = \exp[-\{\Lambda_T(t) + \Lambda_C(t) + \tilde{\mathbf{Z}}_1^\top \boldsymbol{\Gamma}^0(t)\}]$ conditionally on $\tilde{\mathbf{Z}}_1$, independence of the components of $\tilde{\mathbf{Z}}_1$ clearly implies $[\mathbf{H}(t)]_{ij} \equiv 0$ for $i \neq j$. For $i = j$, factor the conditional at-risk indicator as $Y_1(t) = Y_1^{(j)}(t) Y_1^{(-j)}(t)$ where $Y_1^{(j)} := \exp\{-\tilde{Z}_{1j} \Gamma_j^0(t)\}$. Utilizing independence again, we get

$$[\mathbf{H}(t)]_{jj} = \frac{E\{Y_1^{(j)}(t)\} E\{\tilde{Z}_{1j}^2 Y_1^{(j)}(t)\} - E\{Y_1^{(j)}(t) \tilde{Z}_{1j}\}^2}{E\{Y_1^{(j)}(t)\}^2} = \frac{\mathrm{d}}{\mathrm{d}x} \frac{\phi'(x)}{\phi(x)} \Big|_{x=-\Gamma_j^0(t)}$$

This proves (47). To verify (48), simply note that the moment generating function of a standard Gaussian is $\phi(x) = \exp(x^2/2)$ for which $\mathrm{d}/\mathrm{d}x\,(\phi'(x)\phi(x)^{-1}) = 1$. □

From (47), a 'simple' description of $\boldsymbol{\delta}$ (which does not involve factorizing a matrix in terms of $\boldsymbol{\Sigma}^{1/2}$) is available exactly when features are Gaussian. Specifically, it holds for some fixed $K > 0$ that

$$\frac{\mathrm{d}}{\mathrm{d}x} \frac{\phi'(x)}{\phi(x)} = K, \quad \text{and } \phi(0) = 1,$$

iff $\phi(x) = \exp(Kx^2/2)$, the moment generating function of a centered Gaussian random variable.

Proof of Theorem 3. We apply Lemma 6. Denote by $\mathbf{v}_j$ the jth canonical basis vector in $\mathbb{R}^{p_n}$. Integrating by parts in (48), we obtain

$$\delta_j = \mathbf{v}_j^\top \boldsymbol{\Sigma} \int_0^\tau \boldsymbol{\alpha}^0(t) E\{Y_1(t)\} \mathrm{d}t = \mathbf{v}_j^\top \boldsymbol{\Sigma} \int_0^\infty \boldsymbol{\alpha}^0(t) E\{P(T_1 \wedge C_1 \wedge \tau \geq t)\} \mathrm{d}t = \mathbf{v}_j^\top \boldsymbol{\Sigma} E\{\mathbf{A}^0(T_1 \wedge C_1 \wedge \tau)\}.$$

By the assumptions, $|\mathbf{v}_j^\top \boldsymbol{\Sigma} E\{\mathbf{A}^0(T_1 \wedge C_1 \wedge \tau)\}| \geq c_1 n^{-\kappa}$ whenever $j \in \mathrm{M}^n$. Thus $\mathrm{M}^n \subseteq \mathrm{M}^n_\delta$. For Gaussian Z_{1j}, we have $P(|Z_{1j}| > s) \leq \exp(-s^2/2)$, and the SIS property then follows from Lemma 3. □

Proof of Theorem 4. Recall that

$$\boldsymbol{\Delta} = E\Big[\int_0^\tau \{\mathbf{Z}_1 - \mathbf{e}(t)\}^{\otimes 2} Y_1(t) \mathrm{d}t\Big].$$

Then

$$\boldsymbol{\Delta}\boldsymbol{\alpha}^0 = \int_0^\tau \frac{E\{Y_1(t)\}E\{Y_1(t)\mathbf{Z}_1\mathbf{Z}_1^\top\boldsymbol{\alpha}^0\} - E\{Y_1(t)\mathbf{Z}_1^\top\boldsymbol{\alpha}^0\}E\{Y_1(t)\mathbf{Z}_1\}}{E\{Y_1(t)\}} \mathrm{d}t,$$

But by Lemma 4 and the assumption of random censoring,

$$E\{Y_1(t)\mathbf{Z}_1\mathbf{Z}_1^\top\boldsymbol{\alpha}^0\} = \boldsymbol{\Sigma}\boldsymbol{\alpha}^0 \frac{E\{(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0)^2 Y_1(t)\}}{\mathrm{Var}(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0)}, \quad \text{and } E\{\mathbf{Z}_1 Y_1(t)\} = \boldsymbol{\Sigma}\boldsymbol{\alpha}^0 \frac{E\{Y_1(t)\mathbf{Z}_1^\top\boldsymbol{\alpha}^0\}}{\mathrm{Var}(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0)}.$$

So we can construct a function ξ such that $\boldsymbol{\Delta}\boldsymbol{\alpha}^0 = \boldsymbol{\Sigma}\boldsymbol{\alpha}^0 \int_0^\tau \xi(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0, t)\mathrm{d}t$ where $\int_0^\tau \xi(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0, t)\mathrm{d}t \neq 0$, by nonsingularity of $\boldsymbol{\Delta}$. Similarly, using Lemma 5, we may construct a function ζ such that $\boldsymbol{\delta} = \boldsymbol{\Sigma}\boldsymbol{\alpha}^0 \int_0^\tau \zeta(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0, t)\mathrm{d}t$. Taking $\nu := \int_0^\tau \zeta(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0, t)\mathrm{d}t / \int_0^\tau \xi(t, \mathbf{Z}_1^\top\boldsymbol{\alpha})\mathrm{d}t$, $\boldsymbol{\beta}^0 = \nu\boldsymbol{\alpha}^0$ solves $\boldsymbol{\Delta}\boldsymbol{\beta}^0 = \boldsymbol{\delta}$. □

References

Aalen, O. O. (1980) A model for non-parametric regression analysis of counting processes. In *Lecture Notes on Mathematical Statistics and Probability 2* (eds. W. Klonecki, A. Kozek and J. Rosinski), 1–25. Springer-Verlag.

Aalen, O. O. (1989) A linear regression model for the analysis of lifetimes. *Statist. Med.*, **8**, 907–925.

Bair, E. and Tibshirani, R. (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.*, **2**, E108.

Benner, A., Zucknick, M., Hielscher, T. and et al. (2010) High-dimensional Cox models: The choice of penalty as part of the model building process. *Biom. J.*, **52**, 50–69.

Bradic, J., Fan, J. and Jiang, J. (2010) Regularization for Cox's proportional hazards model with NP-dimensionality. Tech. rep., Princeton University.

Bradic, J., Fan, J. and Wang, W. (2011) Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *J. R. Statist. Soc. B*, **73**, 325–349.

Breiman, L. (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statist. Sci.*, **16**, 199–231.

Brillinger, D. R. (1983) A generalized linear model with "Gaussian" regressor variables. In *A Festschrift for Erich L. Lehmann* (eds. P. J. Bickel, K. A. Doksum and J. L. Hodges), 97–114. Wadsworth.

Cheng, K. F. and Wu, J. W. (1994) Adjusted least squares estimates for the scaled regression coefficients with censored data. *J. Amer. Statist. Assoc.*, **89**, 1483–1491.

Fan, J., Feng, Y. and Song, R. (2011) Nonparametric independence screening in sparse ultra-high dimensional additive models. Tech. rep., Princeton University. http://arxiv.org/pdf/1012.4255.

Fan, J., Feng, Y. and Wu, Y. (2010) *Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown*, chap. High-dimensional variable selection for Cox's proportional hazards model. Institute of Mathematical Statistics.

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

Fan, J. and Lv, J. (2008) Sure independence screening for ultra-high dimensional feature space. *J. R. Statist. Soc. B*, **70**, 849–911.

Fan, J. and Lv, J. (2009) Non-concave penalized likelihood with NP-dimensionality. Tech. rep., Princeton University and University of South California.

Fan, J., Samworth, R. and Wu, Y. (2009) Ultrahigh dimensional feature selection: Beyond the linear model. *J. Machine Learning Res.*, **10**, 2013–2038.

Fan, J. and Song, R. (2010) Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.*, **38**, 3567–3604.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007) Pathwise coordinate optimization. *Ann. Appl. Statist.*, **1**, 302–332.

Gorst-Rasmussen, A. (2011) *ahaz – Regularization for semiparametric additive hazards regression*. Department of Mathematical Sciences, Aalborg University. http://cran.r-project.org/web/packages/ahaz/index.html.

Gorst-Rasmussen, A. and Scheike, T. (2011) Efficient penalized estimation in the semiparametric additive hazards model. Tech. rep., Department of Mathematical Sciences, Aalborg University. In preparation.

Hall, P. and Li, K. (1993) On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.*, **21**, 867–889.

Hardin, C. D. (1982) On the linearity of regression. *Z. Wahrsch. verw. Gebiete*, **61**, 293–302.

Hattori, S. (2006) Some properties of misspecified additive hazards models. *Statist. Prob. Letters*, **76**, 1641–1646.

He, Q. and Lin, D. (2011) A variable selection method for genome-wide association studies. *Bioinformatics*, **27**, 1–8.

Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, **58**, 13–30.

Leng, C., Lin, Y. and Wahba, G. (2007) A note on the lasso and related procedures in model selection. *Statist. Sinica*, **16**, 1273–1284.

Li, K. and Duan, N. (1989) Regression analysis under link violation. *Ann. Statist.*, **17**, 1009–1052.

Lin, D. Y. and Ying, Z. (1994) Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61–71.

Ma, S. and Leng, C. (2007) Path consistent model selection in additive risk model via lasso. *Statist. Med.*, **26**, 3753–3770.

Ma, S. and Song, X. (2011) Ranking prognosis markers in cancer genomic studies. *Brief. Bioinform.*, **12**, 33–40.

Martinussen, T. and Scheike, T. H. (2009) Covariate selection for the semiparametric additive risk model. *Scand. J. Statist.*, **36**, 602–619.

Massart, P. (2000) About the constants in Talagrands concentration inequalities for empirical processes. *Ann. Prob.*, **28**, 863–884.

McKeague, I. W. and Sasieni, P. D. (1994) A partly parametric additive risk model. *Biometrika*, **81**, 501–514.

Meinshausen, N. and Bühlmann, P. (2010) Stability selection. *J. R. Statist. Soc. B*, **72**, 417–473.

Metzeler, K. H., Hummel, M., Bloomfield, C. D. and et al. (2008) An 86 probe set gene expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood*, **112**, 4193–4201.

Pollard, D. (1989) Asymptotics via empirical processes. *Statist. Sci.*, **4**, 341–354.

Roberts, C. P. (2008) Discussion of 'sure independence screening for ultrahigh dimensional feature space'. *J. R. Statist. Soc. B*, **70**, 901.

Struthers, C. A. and Kalbfleisch, J. D. (1986) Misspecified proportional hazards models. *Biometrika*, **73**, 363–369.

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

Tibshirani, R. (2009) Univariate shrinkage in the Cox model for high dimensional data. *Stat Appl Genet Mol Biol*, **8**. Article 21.

Wang, H. and Leng, C. (2007) Unified LASSO estimation by least squares approximation. *J. Amer. Statist. Assoc.*, **102**, 1039–1048.

Zhao, S. D. and Li, Y. (2010) Principled sure independence screening for Cox models with ultra-high-dimensional covariates. Tech. rep., Harvard School of Public Health.

Zhu, L., Qian, L. and Lin, J. (2009) Variable selection in a class of single-index models. *Ann. Inst. Statist. Math.* DOI: 10.1007/s10463-010-0287-4.

Zhu, L. and Zhu, L. (2009) Nonconcave penalized inverse regression in single-index models with high dimensional predictors. *J. Multivariate Anal.*, **100**, 862–875.

Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, **101**, 1418–1429.

Zou, H., Hastie, T. and Tibshirani, R. (2007) On the "degrees of freedom" of the lasso. *Ann. Statist.*, **35**, 2173–2192.

Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, **36**, 1509–1533.