Statistica Sinica Preprint No: SS-2021-0265	
Title	Greedy Variable Selection for High-Dimensional Cox Models
Manuscript ID	SS-2021-0265
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0265
Complete List of Authors	Chien-Tong Lin, Yu-Jen Cheng and Ching-Kang Ing
Corresponding Authors	Yu-Jen Cheng
E-mails	ycheng@stat.nthu.edu.tw
Notice: Accepted version subject to English editing.	

Statist. Sinica

GREEDY VARIABLE SELECTION FOR HIGH-DIMENSIONAL COX MODELS

Chien-Tong Lin, Yu-Jen Cheng* and Ching-Kang Ing

National Tsing Hua University

Abstract: We are concerned with the problem of variable selection for high-dimensional sparse Cox models. We propose using a computationally efficient procedure, the Chebyshev greedy algorithm (CGA), to sequentially include variables, and derive its convergence rate under a weak sparsity condition. When a strong sparsity condition is assumed, a high-dimensional information criterion (HDIC) is introduced and used together with CGA to achieve variable selection consistency. We further devise a greedier version of CGA (gCGA). With the help of HDIC, gCGA not only enjoys selection consistency but also exhibits superior finite-sample performance in detecting marginally weak but jointly strong signals over the original CGA and other related high-dimensional methods such as conditional sure independence screening. The proposed methods are illustrated using real data from a cytogenetically normal acute myeloid leukaemia (CN-AML) dataset.

Key words and phrases: Chebyshev greedy algorithm, high-dimensional information criterion, sure screening, variable selection consistency.

1. Introduction

In modern biomedical studies, the excessive number of biomarkers presents technical challenges in the application of existing statistical methods. For example, in the context of genomic research of acute myeloid leukaemia, tens of thousands of gene signatures are measured to predict cancer patients' overall survival (Metzeler et al., 2008). Typically, only a small portion of biomarkers are relevant to the clinical outcome; thus developing a tailored procedure which effectively identifies those relevant biomarkers is essential in the analysis of high-dimensional survival data.

Fan and Lv (2008) have introduced a two-step procedure for high-dimensional variable selection. In the first step, sure independence screening is used to reduce the number of candidate variables to a scalable size. Then, the nonconcave penalized likelihood method is exploited to achieve the oracle property (Fan and Li, 2001). After the seminal work of Fan and Lv (2008), the development of marginal screening methods has proliferated over the past decade and has been extended to various survival models (Fan, Feng and Wu, 2010; Song et al., 2014). Nevertheless, most marginal screening methods hang on the assumption that jointly important variables should also have strong marginal associations with the outcome. Consequently, marginally weak but jointly strong signals are unlikely to be

detected by these methods.

Barut, Fan and Verhasselt (2016) and Hong, Kang and Li (2018) have addressed this problem by implementing sure independence screening (SIS) after conditioning on a known variable set $\mathcal{C}$, which is referred to as conditional SIS (CSIS). They have argued that CSIS asymptotically detects marginally weak but jointly strong signals (variables), provided that $\mathcal{C}$ satisfies some technical assumptions (see Theorem 3 of Barut, Fan and Verhasselt (2016)). However, it seems difficult to show that these assumptions are fulfilled by the commonly used $\mathcal{C}$, which is determined either from biological knowledge or other variable screening methods such as (unconditional) SIS.

To gain further insight into the impact of $\mathcal{C}$ on the performance of CSIS, we conduct a simulation study based on data generated from a sparse Cox model with the hazard function $\lambda(t|\boldsymbol{Z}) = \exp(\boldsymbol{Z}'\boldsymbol{\beta})$. The censoring time is generated from the Uniform$(0, c)$ distribution and the censoring rate is controlled around 30% through the constant c. The sample size is set to 400, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{10000})'$ is the coefficient vector satisfying $\beta_1 = \beta_2 = \beta_3 = 3$ and $\beta_j = 0$ for $4 \leq j \leq 10000$, and $\boldsymbol{Z} = (Z_1, \ldots, Z_{10000})'$ is the covariate vector obeying $Z_1 = W_1 - W_2 - W_3$, $Z_2 = W_2 - W_3$, $Z_3 = 2W_3$, and $Z_j = W_j$ for $4 \leq j \leq 10000$, with $\{W_j\}_{j=1}^{10000}$ being independently and identically distributed as the standard normal distribution. Given this specification,

the relevant variables Z_2 and Z_3 are marginally weak so that they are hardly selected by SIS. Moreover, Z_2 cannot be selected even by CSIS with some commonly used data driven variable set $\mathcal{C}$. To see this, denote $Z_J = (Z_j, j \in J)$ with $J \subseteq \{1, \ldots, 10000\}$ and let $L_{\mathcal{C},j}, j = 1, \ldots, 10000$ be the maximum partial likelihood values obtained under Cox models with covariates $Z_{\mathcal{C} \bigcup \{j\}}$, $j \notin \mathcal{C}$; define $L_{\mathcal{C},4:10000} = \max_{4 \le j \le 10000} L_{\mathcal{C},j}$, which is used to represent the conditional marginal utility of irrelevant variables in the presence of $\mathcal{C}$. Boxplots in Figure 1 display the empirical distributions of $L_{\mathcal{C},j}$, $j = 1, 2, 3$ and $L_{\mathcal{C},4:10000}$ based on 100 replicates. The left panel of Figure 1 shows that $L_{\emptyset,1}$ is much larger than the others and that both $L_{\emptyset,2}$ and $L_{\emptyset,3}$ are largely indistinguishable from $L_{\emptyset,4:10000}$. Therefore, when SIS is used to determine $\mathcal{C}$ for CSIS (as suggested by Barut, Fan and Verhasselt (2016)), $\{1\}$ is likely to be selected. The behavior of CSIS with $\mathcal{C} = \{1\}$ is illustrated in the middle panel of Figure 1: Z_3 is easily detected but Z_2 is not because $L_{\{1\},2}$ is indistinguishable from $L_{\{1\},4:10000}$. These two panels reflect the intrinsic difficulty of using SIS to choose $\mathcal{C}$.

On the other hand, when $\mathcal{C}$ is set to $\{1, 3\}$, the remaining relevant variable Z_2 is readily detected by CSIS because $L_{\{1,3\},2} \gg L_{\{1,3\},4:10000}$, as shown in the right panel of Figure 1. Note that if we select one variable at a time using CSIS and update $\mathcal{C}$ (initialized with $\mathcal{C} = \emptyset$) iteratively by adding

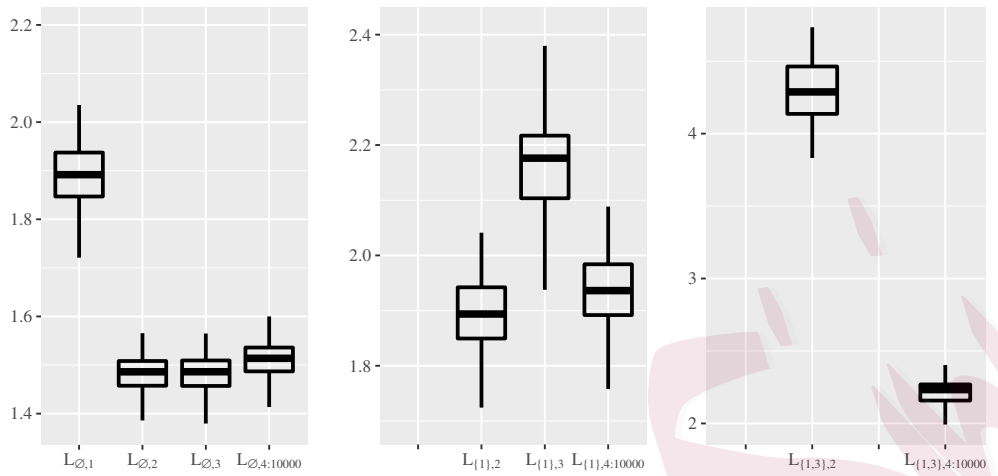

Figure 1: Boxplots of empirical distributions of $L_{\mathcal{C},j}, j = 1, 2, 3$ and $L_{\mathcal{C},4:10000}$ based on 100 replicates, with $\mathcal{C} = \emptyset$ (left panel), $\{1\}$ (middle), and $\{1, 3\}$ (right).

the newly selected variable, then all relevant variables can be included at the third iteration as illustrated in Figure 1. This procedure is exactly forward regression (FR) with partial likelihood pursuit (Hong, Zheng and Li, 2019).

Despite the advantages of FR in terms of selection accuracy, the method has been criticized for its prohibitive computational complexity when the number of candidate variables, p, is large. Greedy algorithms such as L_2-boosting (Bühlmann, 2006), orthogonal greedy algorithm (OGA) (Ing and Lai, 2011), and orthogonal matching pursuit (Tropp and Gilbert, 2007) have

been proposed to alleviate this difficulty by sequentially choosing variables to enter in a linear model with much less computational effort but with the desired accuracy of prediction and selection. Greedy algorithms also have satisfying statistical properties in high-dimensional generalized linear models (Elenberg et al., 2018). However, not much is known about the algorithms when it comes to high-dimensional survival models.

In this paper, we attempt to fill this gap by investigating the Chebyshev greedy algorithm (CGA) (Temlyakov, 2015) in a high-dimensional sparse Cox model in which the number, $p = p_n$, of candidate variables is much larger than the sample size, n. We first derive a uniform error bound for the CGA, which holds uniformly for the number of iterations and can be explained by a bias-variance trade-off between the approximation error and the estimation error. When the model coefficients satisfy a weak sparsity condition, the best compromise between these two errors is achieved by suitably choosing the iteration number, leading to a convergence rate of $(\log p_n/n)^{1/2}$, which coincides with the "minimax-optimal" rate obtained in linear regression models (Raskutti, Wainwright and Yu, 2011).

Moreover, it is shown in Section 4 that the finite-sample performance of CGA in finding the relevant covariates is quite satisfactory in the above example where two marginally weak but jointly strong signals, Z_2 and Z_3,

are present. However, the algorithm's performance deteriorates when the relevant covariates, $Z_1, \ldots, Z_3$, become correlated with the irrelevant ones, $Z_4, \ldots, Z_{10000}$; see Section 3.1. In contrast, FR remains robust, albeit time-consuming. This observation motivates us to develop a greedier variant of CGA (gCGA) that combines the strengths of CGA and FR. We show that gCGA not only shares the same computational efficiency as that of CGA, but it also boasts exceptional finite-sample performance in terms of correctness of selection, in particular, in the difficult case just mentioned. In addition, we establish, under a strong sparsity condition, gCGA's sure screening property (defined in Theorem 2) and selection consistency when it is used in conjunction with a high-dimensional information criterion (HDIC) that can remove all irrelevant covariates included by the algorithm. As far as we know, there is no previous research concerning the selection consistency of greedy-type algorithms in high-dimensional Cox models.

The rest of this paper is organized as follows. We describe CGA and introduce its uniform convergence rate in Section 2. In Section 3, we propose gCGA, present its sure screening property, and establish its selection consistency when used together with HDIC. In Sections 4 and 5, the performance of the proposed methods and those based on CSIS or LASSO is compared using simulated data and a CN-AML dataset. We conclude in

Section 6. All technical proofs and additional simulations are deferred to the Supplementary Material.

We end this section with some notation used throughout the paper. For $\boldsymbol{u} = (u_1, \ldots, u_p)' \in \mathbb{R}^p$, $\boldsymbol{u}^{\otimes 0} = 1, \boldsymbol{u}^{\otimes 1} = \boldsymbol{u}$, $\boldsymbol{u}^{\otimes 2} = \boldsymbol{u}\boldsymbol{u}'$, $\text{supp}(\boldsymbol{u}) = \{j : u_j \neq 0\}$, $\|\boldsymbol{u}\|_q = \{\sum_{j=1}^p |u_i|^q\}^{1/q}$ for $1 \leq q < \infty$, and $\|\boldsymbol{u}\|_0 = \sum_{j=1}^p I(u_j \neq 0)$, $\|\boldsymbol{u}\|_\infty = \max_{1 \leq j \leq p} |u_j|$. For $J \subseteq \{1, \ldots, p\}$, $\boldsymbol{u}_J \in \mathbb{R}^p$ denotes the vector satisfying $u_i = 0, i \in J^c$, $J^c = \{1, \ldots, p\} - J$ is the complement of J, and $|J|$ denotes the cardinality of J. We denote the minimum eigenvalues of a matrix A by $\lambda_{\min}[A]$. Also, $\lfloor a \rfloor$ ($\lceil a \rceil$) denotes the largest (smallest) integer $\leq a$ ($\geq a$).

2. CGA for selecting high-dimensional Cox models

2.1 Preliminaries

There are three popular greedy algorithms for high-dimensional linear regression models: FR (Wang, 2009), L_2-boosting, and OGA. Although FR has desirable theoretical properties, it is very time consuming. This weakness becomes more prominent when the method is generalized to high-dimensional Cox models; see Section 4 and Section S3 of the supplementary material for details. By contrast, while having great computational efficiency, L_2-boosting suffers from very slow convergence (to the true model),

resulting in unsatisfactory performance in estimation and variable selection. As a greedy algorithm lying somewhere between FR and L_2-boosting, OGA can adequately share their advantages. Briefly speaking, it gains computational efficiency by including variables like L_2-boosting, and enjoys an excellent convergence rate and selection accuracy by updating parameters like FR. The outstanding performance of OGA motivates this study to use its nonlinear counterpart, CGA, to choose variables in high-dimensional Cox models.

Let the failure time, the censoring time, and the p-dimensional covariate vector be denoted by T, C, and $\boldsymbol{Z} = (Z_1, \ldots, Z_p)'$ respectively. Assume that T and C are independent given $\boldsymbol{Z}$, and T follows the Cox model

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t) \exp(\boldsymbol{Z}'\boldsymbol{\beta}^*), \tag{2.1}$$

where $\lambda_0(t)$ is the unspecified baseline hazard function and $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is the true coefficient vector. Because of right censorship, we only observe $\{(\boldsymbol{Z}_i, X_i, \delta_i)\}$ for $i = 1, \ldots, n$, where $\boldsymbol{Z}_i = (Z_{i,1}, \ldots, Z_{i,p})'$ is the observed covariate vector, $X_i = \min(T_i, C_i)$ is the observed event time, and $\delta_i = I(T_i \leq C_i)$ is the censoring indicator. For $r = 0, 1, 2$, define

$$S^{(r)}(\boldsymbol{\beta}, t) = n^{-1} \sum_{i=1}^{n} \boldsymbol{Z}_i^{\otimes r} Y_i(t) \exp\{\boldsymbol{Z}_i'\boldsymbol{\beta}\}$$

where $Y_i(t) = I(X_i \geq t)$ is referred to as the at-risk process, $\boldsymbol{\beta} \in \mathbb{R}^p$

and $\bar{\boldsymbol{Z}}_n(\boldsymbol{\beta}, t) = S^{(1)}(\boldsymbol{\beta}, t)/S^{(0)}(\boldsymbol{\beta}, t)$. For a prespecified τ, the negative log-partial likelihood is given by

$$l_n(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\Big(\boldsymbol{Z}_i'\boldsymbol{\beta} - \log S^{(0)}(\boldsymbol{\beta}, t)\Big)dN_i(t),$$

where $N_i(t) = I(X_i \le t, \delta_i = 1)$ is a counting process. Straightforward calculations yield

$$\nabla l_n(\boldsymbol{\beta}) = -n^{-1}\sum_{i=1}^{n}\int_0^{\tau}\big[\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}_n(\boldsymbol{\beta}, t)\big]dN_i(t) \quad \text{and} \quad \nabla^2 l_n(\boldsymbol{\beta}) = \int_0^{\tau} V_n(\boldsymbol{\beta}, t)d\bar{N}(t),$$

where $\bar{N}(t) = n^{-1}\sum_{i=1}^{n} N_i(t)$.

Denote $\nabla l_n(\boldsymbol{\beta})$ by $(\nabla_1 l_n(\boldsymbol{\beta}), \ldots, \nabla_p l_n(\boldsymbol{\beta}))'$. For $J \subseteq \{1, \ldots, p\}$, define

$$\hat{\boldsymbol{\beta}}_J = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}, \mathrm{supp}(\boldsymbol{\beta})=J} l_n(\boldsymbol{\beta}),$$

where $\mathcal{B} \subseteq \mathbb{R}^p$ is the parameter space of interest. The CGA is an iterative algorithm that generates a sequence of nested sets $\{\hat{J}_1, \ldots, \hat{J}_K\}$ in $\{1, \ldots, p\}$, where K is a prescribed upper bound for the iteration number and

$$\hat{J}_k = \hat{J}_{k-1}\bigcup\{\hat{j}_k\}, \ k = 1, \ldots, K, \tag{2.2}$$

with $\hat{J}_0 = \emptyset$, $\hat{j}_k = \arg\max_{1\le j\le p, j\in\hat{J}_{k-1}^c}\big|\nabla_j l_n(\hat{\boldsymbol{\beta}}_{\hat{J}_{k-1}})\big|$, and $\hat{\boldsymbol{\beta}}_\emptyset = \mathbf{0}$. Indeed, the selection criterion (2.2) can be interpreted as choosing the variable having the strongest correlation with the current functional gradient (He et al., 2016), and resembles the variable inclusion method used in L_2-boosting and OGA for linear models.

2.2 Convergence analysis of CGA

Our asymptotic results are mainly built on assumptions concerning the population counterparts of $l_n(\boldsymbol{\beta})$, $\nabla l_n(\boldsymbol{\beta})$, and $\nabla^2 l_n(\boldsymbol{\beta})$:

$$
\begin{aligned}
l(\boldsymbol{\beta}) &= -\int_0^\tau \Big(s^{(1)}(\boldsymbol{\beta}^*, t)'\boldsymbol{\beta} - [\log s^{(0)}(\boldsymbol{\beta}, t)] s^{(0)}(\boldsymbol{\beta}^*, t) \Big) \lambda_0(t) dt, \\
\nabla l(\boldsymbol{\beta}) &= -\int_0^\tau \Big\{ s^{(1)}(\boldsymbol{\beta}^*, t) - \frac{s^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} s^{(0)}(\boldsymbol{\beta}^*, t) \Big\} \lambda_0(t) dt, \\
\nabla^2 l(\boldsymbol{\beta}) &= \int_0^\tau \Big\{ \frac{s^{(2)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} - \Big(\frac{s^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \Big)^{\otimes 2} \Big\} s^{(0)}(\boldsymbol{\beta}^*, t) \lambda_0(t) dt,
\end{aligned}
$$

where $s^{(r)}(\boldsymbol{\beta}, t) = E\{S^{(r)}(\boldsymbol{\beta}, t)\}$ for $r = 0, 1, 2$. Let b_0 be a large constant. The parameter space that we are interested in is the l_1-ball of radius b_0, $\mathcal{B} = \{\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p, \|\boldsymbol{\beta}\|_1 \le b_0\}$, where $p = p_n$ is allowed to approach infinity faster than n. For $J, J' \subseteq \{1, \ldots, p\}$, define $\boldsymbol{\beta}_J = \arg\min_{\boldsymbol{\beta} \in \mathcal{B}, \mathrm{supp}(\boldsymbol{\beta}) = J} l(\boldsymbol{\beta})$ and $\nabla^2_{JJ'} l(\boldsymbol{\beta}) = [\nabla^2_{kl} l(\boldsymbol{\beta})]_{k \in J, l \in J'}$, where $\nabla^2_{kl} l(\boldsymbol{\beta})$ is the (k, l)-th element of $\nabla^2 l(\boldsymbol{\beta})$. The assumptions required in our analysis are listed below:

(C1) $\boldsymbol{\beta}^*$ is an interior point of $\mathcal{B}$; moreover, there exists a positive constant $\bar{D}$ such that for any $|J| \le D_n = \lceil \bar{D}(n/\log p_n)^{1/2} \rceil$, $\boldsymbol{\beta}_J$ is an interior point of $\mathcal{B}$.

(C2) There exists a constant $\eta > 0$ such that $P(\max_{1 \le j \le p_n} |Z_j| > \eta) = 0$.

(C3) There exists a constant $0 < \rho < 1$ such that $\rho := P(Y_1(\tau) = 1)$.

(C4) $\log p_n = O(n^\kappa)$ for some $0 \le \kappa < 1$.

(C5) There exists a constant $\delta_0 > 0$ such that

$$\delta_0 \leq \min_{|J| \leq D_n} \lambda_{\min}[\nabla^2_{JJ} l(\boldsymbol{\beta}_J)].$$

(C6) There is an arbitrarily small $\epsilon > 0$ such that for some $0 < M < \infty$,

$$\max_{|J| \leq D_n, i \in J^c} \sup_{\substack{\boldsymbol{\beta} \in B_\epsilon(\boldsymbol{\beta}_J) \\ \mathrm{supp}(\boldsymbol{\beta}) = J}} \left\| \left\{ \int_0^1 \nabla^2_{iJ} l((1-t)\boldsymbol{\beta}_J + t\boldsymbol{\beta}) dt \right\} \left\{ \int_0^1 \nabla^2_{JJ} l((1-t)\boldsymbol{\beta}_J + t\boldsymbol{\beta}) dt \right\}^{-1} \right\|_1 < M. \tag{2.3}$$

(C7) Let $\mathcal{N} := \mathrm{supp}(\boldsymbol{\beta}^*)$; there exist $0 \leq \theta < (1-\kappa)/4$ and $C_0 > 0$ such that

$$\min_{|J| \leq D_n, \mathcal{N} - J \neq \emptyset} \max_{j \in \mathcal{N} - J} |\nabla_j l(\boldsymbol{\beta}_J)| > C_0 n^{-\theta},$$

where $\nabla_j l(\boldsymbol{\beta})$ denotes the j-th component of $\nabla l(\boldsymbol{\beta})$.

A few comments are in order regarding (C1)–(C7). The first part of (C1) is often referred to as the weak sparsity condition. It allows all components in $\boldsymbol{\beta}^*$ to be non-zero, but requires that they are absolutely summable. The second part of (C1), together with (C5), ensures that for any $j \in J$ and $|J| \leq D_n$, $\boldsymbol{\beta}_J$ is unique and $\nabla_j l(\boldsymbol{\beta}_J) = 0$, which is crucial in our analysis of CGA. To ensure that these two properties hold during the iterations, the iteration number is restricted to $K = K_n < D_n$; see Theorem 1. Conditions (C2) and (C3) are commonly assumed in the literature on high dimensional

survival analysis; see Kong and Nan (2014), Hong, Kang and Li (2018), and Hong, Zheng and Li (2019). Condition (C4) allows p_n to grow exponentially with n. Condition (C5) imposes a lower bound for the minimum eigenvalue of the Hessian matrix of $l(\cdot)$, evaluated at the local minimizer, $\boldsymbol{\beta}_J$, over $\mathcal{B}_J := \mathcal{B} \bigcap \{\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p, \text{supp}(\boldsymbol{\beta}) = J\}$ with $|J| \leq D_n$. The condition is flexible in the sense that it does not introduce any restriction on the maximum eigenvalue of the matrix. Conditions like (C6) are frequently used in the derivations of the convergence rates of greedy-type algorithms under weak sparsity conditions; see Ing and Lai (2011) and Ing (2020). Since the ε in (2.3) can be arbitrarily small, (2.3) is almost equivalent to

$$\max_{|J| \leq D_n, i \in J^c} \left\| \left\{ \nabla^2_{iJ} l(\boldsymbol{\beta}_J) \right\} \left\{ \nabla^2_{JJ} l(\boldsymbol{\beta}_J) \right\}^{-1} \right\|_1 < M,$$

which is further simplified to

$$\max_{|J| \leq D_n, i \in J^c} \left\| cov(z_i, \boldsymbol{Z}_J) var^{-1}(\boldsymbol{Z}_J) \right\|_1 < M \tag{2.4}$$

in the case of the linear model. As argued in Ing and Lai (2011) and Ing (2020), (2.4) holds even when the components in $\boldsymbol{Z}$ are highly correlated. Condition (C7) is closely related to the so-called "beta-min" condition (which requires that the non-zero coefficients are sufficiently large) as well as the signal strength condition in Barut, Fan and Verhasselt (2016). Moreover, (C7) together with (C1) is referred to as the strong sparsity condition,

which stipulates that the number of non-zero coefficients be much smaller than n. In fact, it can be shown (Section S2 in the supplementary materials) that

$$\min_{j\in\mathcal{N}} |\beta_j^*| \geq \frac{C_0}{4\eta^2} n^{-\theta} \text{ and } |\mathcal{N}| \leq 4\eta^2 C_0^{-1} b_0 n^{\theta}, \tag{2.5}$$

provided (C1)–(C3) and (C7) hold true. For more discussion of (C7), see Section 3.1.

We are now in a position to state the main result of this section.

Theorem 1. *Assume (C1)–(C5) and (C6) or (C7). Let* $K_n = \bar{\delta}(n/\log p_n)^{1/2}$, *where* $0 < \bar{\delta} < \bar{D}$ *and may depend on* $b_0, \eta, \rho, \delta_0$, *or* M. *Then*

$$\max_{1\leq k\leq K_n} \frac{l(\hat{\boldsymbol{\beta}}_{\hat{J}_k}) - l(\boldsymbol{\beta}^*)}{k^{-1} + kn^{-1}\log p_n} = O_p(1). \tag{2.6}$$

Note that $l(\hat{\boldsymbol{\beta}}_{\hat{J}_k}) - l(\boldsymbol{\beta}^*)$ is the sum of the approximation error, $l(\boldsymbol{\beta}_{\hat{J}_k}) - l(\boldsymbol{\beta}^*)$, and the estimation error, $l(\hat{\boldsymbol{\beta}}_{\hat{J}_k}) - l(\boldsymbol{\beta}_{\hat{J}_k})$. For the approximation error, we show in the proof of Theorem 1 that

$$\max_{1\leq k\leq K_n} \frac{l(\boldsymbol{\beta}_{\hat{J}_k}) - l(\boldsymbol{\beta}^*)}{k^{-1}} = O_p(1), \tag{2.7}$$

which plays a role similar to (6.17) of Bühlmann (2006) or (3.12) of Ing and Lai (2011) in high-dimensional linear models, in which weak greedy algorithms or OGA is used in place of CGA. Equation (2.7), together with

the uniform bound established for the estimation error,

$$\max_{1\le k\le K_n} \frac{l(\hat{\boldsymbol{\beta}}_{\hat{J}_k}) - l(\boldsymbol{\beta}_{\hat{J}_k})}{kn^{-1}\log p_n} = O_p(1) \tag{2.8}$$

(which is also given in the proof of Theorem 1), suggests that $k_n^* = c_0(n/\log p_n)^{1/2}$, $c_0 > 0$, is an optimal choice of k that achieves the best (up to a constant factor) compromise between the approximation and estimation errors and leads to the the following error bound

$$l(\hat{\boldsymbol{\beta}}_{\hat{J}_{k_n^*}}) - l(\boldsymbol{\beta}^*) = O_p((\log p_n/n)^{1/2}). \tag{2.9}$$

Note that $(\log p_n/n)^{1/2}$ is also the "minimax-optimal" rate for linear models (Raskutti, Wainwright and Yu, 2011). To help better understand (2.9), we also offer a numerical illustration of the equation at different sparsity levels in the supplementary material.

When the $\hat{\boldsymbol{\beta}}_{\hat{J}_{k_n^*}}$ on the left-hand side of (2.9) is replaced by the LASSO estimate, Kong and Nan (2014) have also derived an error bound achieving the optimal balance between the approximation and estimation errors. However, it may seem tricky to recover the $(\log p_n/n)^{1/2}$ convergence rate using their bound when the weak sparsity condition described in (C1) holds. We close this section by mentioning that establishing the sure screening property appears to be more relevant than pursuing the $(\log p_n/n)^{1/2}$ rate when (C7) is assumed. As will be made clear in the next section, (2.7)

plays an indispensable role in developing such a property for CGA and its variants.

3. A greedier variant of CGA and consistent variable selection

Throughout the rest of the paper, we assume that (C7) holds. Motivated by an example in Section 3.1, we first introduce gCGA to combine the advantages of CGA and FR, and then state its sure screening property. In Section 3.2, we establish the selection consistency of gCGA when it is used together with HDIC.

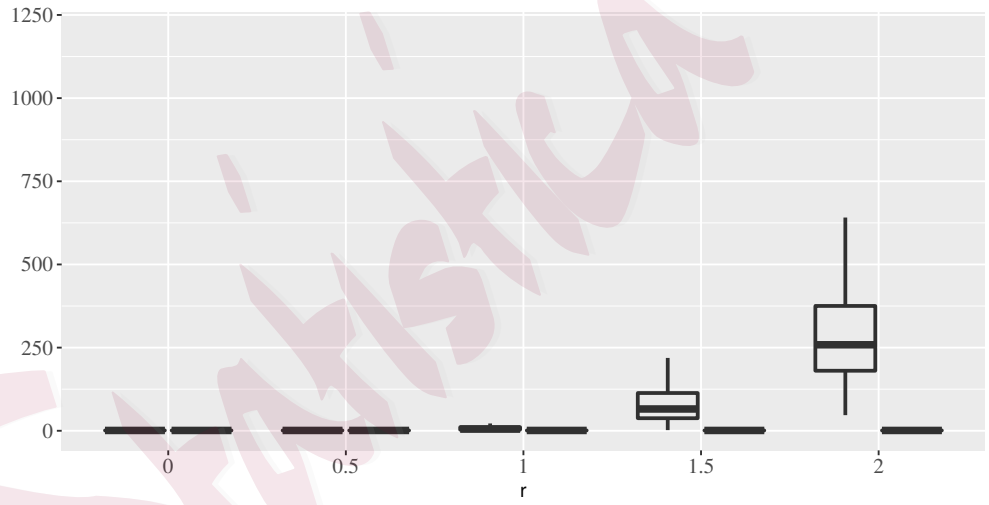

Figure 2: Boxplots for empirical distributions of $k_c(r)$ (left) and $k_f(r)$ (right) based on 100 simulations, where $r \in \mathcal{A}$.

3.1 A greedier variant of CGA and its sure screening property

One salient feature of CGA is reducing computational costs by using only the gradient information while still maintaining the desired convergence rate. In addition, CGA efficiently identifies the relevant covariates in the example of Section 1, which contains two marginally weak but jointly strong signals. For more details, see Section 4. However, CGA's performance deteriorates in the same example, provided the relevant covariates, $Z_1, \ldots, Z_3$, become correlated with the irrelevant ones, $Z_4, \ldots, Z_{10000}$. More specifically, let $Z_1, \ldots, Z_3$ and $W_1, \ldots, W_{10000}$ be defined as in Section 1. Set $Z_j = rW_3 + W_j$, $4 \leq j \leq 10000$, where $r \in \mathcal{A} \equiv \{0, 0.5, 1, 1.5, 2\}$. The larger r is, the higher the correlations between the relevant and irrelevant variables are. Note that $r = 0$ corresponds to the example of Section 1, in which $\{Z_1, \ldots, Z_3\}$ and $\{Z_j, 4 \leq j \leq 10000\}$ are independent. Let $\mathcal{C} = \{1\}$, $\{|\nabla_{c_i} l_n(\hat{\boldsymbol{\beta}}_{\mathcal{C}})|, i = 1, \ldots, p-1\}$ be a non-increasing rearrangement of $\{|\nabla_i l_n(\hat{\boldsymbol{\beta}}_{\mathcal{C}})|, i = 2, \ldots, p\}$, and $\{l_n(\hat{\boldsymbol{\beta}}_{\mathcal{C}\cup\{f_i\}}), i = 1, \ldots, p-1\}$ be a non-decreasing rearrangement of $\{l_n(\hat{\boldsymbol{\beta}}_{\mathcal{C}\cup\{i\}}), i = 2, \ldots, p\}$. Define

$$k_c(r) = \arg \min_{1 \leq j \leq p-1} \{j : \{2, 3\} \cap \{c_1, \ldots, c_j\} \neq \emptyset\},$$

$$k_f(r) = \arg \min_{1 \leq j \leq p-1} \{j : \{2, 3\} \cap \{f_1, \ldots, f_j\} \neq \emptyset\},$$

where $r \in \mathcal{A}$. The boxplots of the empirical distributions of $k_c(r)$ and $k_f(r)$, based on 100 simulations, are presented in Figure 2. The figure shows that for each r, all values of $k_f(r)$ are equal to 1, suggesting that regardless of whether the correlations between $\{Z_1, \ldots, Z_3\}$ and $\{Z_j, 4 \leq j \leq 10000\}$ are high or low, Z_2 or Z_3 is easily chosen by FR at the second iteration once Z_1 has been included at the first iteration. On the other hand, although $k_c(r)$ behaves like $k_f(r)$ when $r \leq 0.5$, the value of $k_c(r)$ is larger than 1 when $r \geq 1$ and grows rapidly as r increases. Therefore, when r is relatively large, it is difficult for CGA to find Z_2 or Z_3, given that Z_1 has been chosen by the algorithm at the first iteration. This numerical experiment reveals that although FR is very time-consuming, it substantially outperforms CGA in terms of selection accuracy in the difficult case where not only do the relevant covariates contain some marginally weak but jointly strong signals, but they also are highly correlated with the irrelevant ones. This observation motivates us to combine the strengths of CGA and FR using a greedier variant of CGA, which we call gCGA.

This gCGA, initiated with $\tilde{J}_0 = \emptyset$, is sequentially updated via

$$\tilde{J}_{k+1} = \tilde{J}_k \bigcup \{\tilde{j}_{k+1}\},$$

where

$$\tilde{j}_{k+1} = \arg\min_{j \in \tilde{\mathcal{M}}_k} l_n(\hat{\boldsymbol{\beta}}_{\tilde{J}_k \bigcup \{j\}})$$

and for some $0 \leq t \leq 1$,

$$\tilde{\mathcal{M}}_k := \{j \in \tilde{J}_k^c : |\nabla_j l_n(\hat{\boldsymbol{\beta}}_{\tilde{J}_k})| \geq t \|\nabla l_n(\hat{\boldsymbol{\beta}}_{\tilde{J}_k})\|_\infty\}.$$

Clearly gCGA includes CGA ($t = 1$) and FR ($t = 0$) as special cases. Since at each iteration k, gCGA implements FR within a "promising" subset, $\tilde{\mathcal{M}}_k$, of $\tilde{J}_k^c$ and since this promising subset is determined solely based on gradient information, the algorithm preserves FR's selection accuracy without much computational effort, provided the t in $\tilde{\mathcal{M}}_k$ is chosen to be close to 1. A practical guideline for determining $\tilde{\mathcal{M}}_k$ is provided in Section 4. The next corollary shows that CGA, gCGA, and FR all share the same convergence rate.

Corollary 1. *Assume (C1)–(C5) and (C7). Then for any $t \in [0, 1]$ and $K_n = \bar{\delta}(n/\log p_n)^{1/2}$, where $\bar{\delta}$ is defined as in Theorem 1, (2.6) holds with $\hat{J}_k$ replaced by $\tilde{J}_k$.*

With the help of Corollary 1, Theorem 2 establishes the sure screening property of gCGA (CGA and FR).

Theorem 2. *Assume (C1)–(C5) and (C7). Then, for any* $t \in [0,1]$ *and* $K_n \geq \lceil C_1 n^{2\theta} \rceil$, *with* C_1 *being a constant depending on* η, b_0, *and* C_0,

$$\lim_{n\to\infty} P(\mathcal{N} \subset \tilde{J}_{K_n}) = 1, \tag{3.10}$$

which is referred to as the sure screening property.

Theorem 2 asserts that gCGA (CGA and FR) enjoys the sure screening property as long as the number of iterations approaches $\lceil C_1 n^{2\theta} \rceil$.

3.2 Variable selection consistency

Although gCGA has the sure screening property when $K_n > C_1 n^{2\theta}$, the model $\tilde{J}_{K_n}$ determined by the algorithm at the end of iteration, K_n, suffers from severe overfitting because, as indicated in (2.5), $|\mathcal{N}| = O(n^{\theta}) \ll K_n$. In this section, we propose using HDIC to overcome this difficulty. Define

$$\mathrm{HDIC}(J) = l_n(\hat{\boldsymbol{\beta}}_J) + |J| w_n \log p_n / n, \tag{3.11}$$

where w_n is some positive constant depending on n. We first restrict our attention to the set of nested models, $\mathcal{J}_{K_n} = \{\tilde{J}_1, \ldots, \tilde{J}_{K_n}\}$, generated during the gCGA iterations, and then find the model $\tilde{J}_{\tilde{k}_n} = \{\tilde{j}_1, \ldots, \tilde{j}_{\tilde{k}_n}\}$ with the smallest HDIC value among $\mathcal{J}_{K_n}$, where

$$\tilde{k}_n = \arg \min_{1 \leq k \leq K_n} \mathrm{HDIC}(\tilde{J}_k). \tag{3.12}$$

We further construct a subset of $\tilde{J}_{\tilde{k}_n}$,

$$\tilde{J}_{\mathrm{Trim}} = \{\tilde{j}_i : 1 \le i \le \tilde{k}_n, \mathrm{HDIC}(\tilde{J}_{\tilde{k}_n} - \{\tilde{j}_i\}) > \mathrm{HDIC}(\tilde{J}_{\tilde{k}_n})\}, \tag{3.13}$$

to exclude (possibly) redundant variables in $\tilde{J}_{\tilde{k}_n}$ by examining the "marginal" contribution of each $Z_{\tilde{j}_i}$, $1 \le i \le \tilde{k}_n$, to HDIC. The asymptotic performance of $\tilde{J}_{\mathrm{Trim}}$ is reported in next theorem.

Theorem 3. *(i) Assume (C1)–(C7). Suppose that* $K_n = \bar{\delta}(n/\log p_n)^{1/2}$, $w_n \to \infty$, *and* $w_n = o(K_n)$. *Then,*

$$\lim_{n\to\infty} P\{\tilde{J}_{Trim} = N_n\} = 1. \tag{3.14}$$

(ii) Assume (C1)–(C5) and (C7) with θ *strengthened to* $0 \le \theta < (1-\kappa)/6$. *Suppose that* $\lceil C_1 n^{2\theta} \rceil \le K_n \le C_2 n^{-\theta+(1-\kappa)/2}$, $w_n \to \infty$, *and* $w_n = o(K_n)$, *where* C_2 *depends on* C_0, η, *and* δ_0. *Then,* (3.14) *holds.*

It would be of interest to compare Theorem 3 with Theorem 4.5 of Bradic, Fan and Jiang (2011), which extends the consistency of SCAD from fixed-dimensional Cox models (Fan and Li, 2002) to high-dimensional ones. Note first that instead of imposing high-level assumptions that require $S^{(r)}(\boldsymbol{\beta}, t), r = 0, 1, 2$, to have probability limits (see Condition 2 (i) of Bradic, Fan and Jiang (2011)), we directly derive concentration inequalities for $S^{(r)}(\boldsymbol{\beta}, t)$ (see Lemma 2 of the supplementary material) under conditions

that can be easily justified. Moreover, Theorem 4.5 of Bradic, Fan and Jiang (2011) demands a maximum eigenvalue condition on the Hessian matrix of $l(\cdot)$, whereas there is no such restriction in Theorem 3. Finally, while Condition (C6) in Theorem 3 (i) is similar but somewhat stronger than Condition 8 in Bradic, Fan and Jiang (2011), Theorem 3 (ii) drops (C6) at the cost of slightly stronger limitations on K_n and θ. Generally speaking, neither of the sets of assumptions used in Theorem 3 (i) (or Theorem 3 (ii)) and Theorem 4.5 of Bradic, Fan and Jiang (2011) is more restrictive than the other. The former set of assumptions, however, allows us to build the selection consistency of greedy-type algorithms in high-dimensional Cox models that does not appear to have been reported in the literature before.

4. Simulations

In this section, we use four simulation scenarios to assess the variable screening performance of gCGA and the variable selection accuracy of $\tilde{J}_{\text{Trim}}$. Note that for a given t, there exists an integer, say m, such that $\tilde{\mathcal{M}}_k$ consists of the variables with the largest m absolute gradients among $\{|\nabla_j l_n(\hat{\boldsymbol{\beta}}_{\tilde{J}_k})|, j = 1, \ldots, p\}$. To facilitate the implementation of gCGA, in the rest of this section, we change gCGA's tuning parameter from t to m, and denote the algorithm by gCGA(m) for a given m. In our simulation study, m is set

to 1, 10, 30, and 50, noting that gCGA(1) reduce to CGA. In addition, K_n, the number of iterations, and w_n, a penalty term of HDIC, are given by $\lfloor 5(n/\log p_n)^{1/2} \rfloor$ and $\log\log n$, respectively. In addition, we suggest a data-driven method to select m,

$$\hat{m} = \arg\min_{m \in \mathcal{Q}} \mathrm{HDIC}(\tilde{J}_{K_n}(m))$$

where $\tilde{J}_{K_n}(m)$ denotes the model chosen by gCGA(m) at the end of iteration and $\mathcal{Q}$, a user chosen subset of $\{1, \ldots, p\}$, is set to $\{1, 10, 30, 50\}$ in our simulation. In the rest of this section, the variable sets $\tilde{J}_{\mathrm{Trim}}$ (see (3.13)) derived from gCGA(m) and gCGA($\hat{m}$), respectively, are referred to as gCGA(m)+Trim and gCGA($\hat{m}$)+Trim.

For the purpose of comparison, we consider three marginal methods:

(a) SIS+SCAD: Uses SIS (Fan, Feng and Wu, 2010) to screen variables and then selects variables using SCAD (Fan and Li, 2002) together with extended BIC (Luo, Xu and Chen, 2015).

(b) CSIS+SCAD: Screens variables by CSIS with the conditioning set given by the set of variables chosen in (a) and then selects variables using SCAD together with extended BIC.

(c) ISIS+SCAD: Performs the same procedure as in (b) except that the conditioning set is replaced by $\mathcal{C}_{10}$, where $\mathcal{C}_1$ is the set of variables

chosen in (a) and for $t \geq 2$, $\mathcal{C}_t$ is that chosen by CSIS+SCAD using conditioning set $\mathcal{C}_{t-1}$.

Note that all screening methods in the above procedures are implemented based on the partial likelihood. In addition, the number of variables included at the screening stage is restricted to $\lceil n/\log n \rceil$ and the tuning constant in the extended BIC is set to $1 - \log n/(3\log p)$, as suggested by Hong, Zheng and Li (2019).

For the sake of completeness, we also consider a regularization method, adaptive LASSO (ALASSO) (Zhang and Lu, 2007), for the Cox model. Since ALASSO uses LASSO (Tibshirani, 1997) as an initial estimator to determine the weights for a second-stage weighted LASSO, we treat LASSO as the screening step of ALASSO, and will compare it with other screening methods mentioned previously. The tuning parameters of LASSO and ALASSO are chosen by five-fold cross validation and extended BIC, respectively.

We conducted 100 replications for $(n, p) = (200, 10000)$ and $(n, p) = (400, 10000)$. For each subject, we generated the survival time T from the Cox model $\lambda(t|\boldsymbol{Z}) = \exp(\boldsymbol{Z}'\boldsymbol{\beta}^*)$, the censoring time from the Uniform$(0, c)$ distribution, the observed time $Y = \min\{T, C\}$, and the censoring indicator $\delta = I(T \leq C)$. The constant c was controlled so that the corresponding

censoring rates were around 20% and 50%. Detailed settings for the covariate vector $\boldsymbol{Z}$ and the coefficient vector $\boldsymbol{\beta}^*$ are given below.

Scenario 1. (AR(1) correlation). The covariate vector $\boldsymbol{Z}$ follows the multivariate normal distribution with zero mean and covariance matrix Σ, where $\Sigma_{jj} = 1$ and $\Sigma_{jk} = 0.5^{|j-k|}$ for $j \neq k$. The coefficients $\{\beta_j^*\}, j \in \{1, 2, 3, 6, 12\}$ are generated from $(4 \log n/\sqrt{n} + |W|/4)U$, in which W follows the standard normal distribution and $P(U = 1) = P(U = -1) = 1/2$, and the other components of $\boldsymbol{\beta}^*$ are fixed to be 0.

Scenario 2. (Equi-correlation). The covariate vector $\boldsymbol{Z}$ follows the multivariate normal distribution with zero mean and covariance matrix Σ, where $\Sigma_{jj} = 1$ and $\Sigma_{jk} = 0.5$ for $j \neq k$. The coefficients $\{\beta_j^*\}, j \in \{1, \ldots, 15\}$ are generated from $(4 \log n/\sqrt{n} + |W|/4)U$, in which W follows the standard normal distribution and $P(U = 1) = P(U = -1) = 1/2$, and β_j^* are fixed to be 0 for $j > 15$.

Scenario 3. (Marginally weak but jointly strong signals I). The covariate vector $\boldsymbol{Z}$ satisfies $Z_1 = W_1 - W_2 - W_3$, $Z_2 = W_2 - W_3$, $Z_3 = 2W_3$, and $Z_j = W_j$ for $j \geq 4$, where $W_1 \sim N(0, 2)$ and $\{W_k\}_{k \geq 2}$ are from i.i.d. standard normal distributions. In addition, $\beta_j^* = 3$ for $j = 1, 2, 3$ and $\beta_j^* = 0$ for $j \geq 4$.

Scenario 4. (Marginally weak but jointly strong signals II). The covariate vector $\boldsymbol{Z}$ satisfies $Z_1 = W_1 - W_2 - W_3$, $Z_2 = W_2 - W_3$, $Z_3 = 2W_3$, and $Z_j = W_3 + G_j$ for $j \geq 4$, where $W_1 \sim N(0, 2)$ and $\{W_2, W_3, G_j\}_{j \geq 4}$ are from i.i.d. standard normal distributions, and the coefficient vector is the same as in Scenario 3.

In Scenarios 1 and 2, an AR(1) correlation and an equi-correlation structures are imposed on the candidate variables, respectively, and the number of the relevant variables in Scenario 2 is considerably larger than that in Scenario 1. In Scenario 3, all candidate variables are uncorrelated with each other except for the relevant ones Z_1, Z_2, and Z_3, in which only Z_1 is correlated with the survival outcome and Z_2 is more difficult to detect than Z_3, as illustrated in Figure 1. The setting of Scenario 4 is same as that of Scenario 3 except that $\{Z_j\}_{j=1}^3$ become correlated with $\{Z_k\}_{k=4}^p$ through W_3. More scenarios and their corresponding simulation results are provided in the supplementary material.

For a given screening method in {gCGA(m), gCGA($\hat{m}$), SIS, CSIS, ISIS, LASSO} and the corresponding model selection method in {gCGA(m)+Trim, gCGA($\hat{m}$)+Trim, SIS+SCAD, CSIS+SCAD, ISIS+SCAD, ALASSO}, define $\hat{\mathcal{S}}_b$ and $\hat{\mathcal{T}}_b$ to be the sets of variables determined by the former and the latter, respectively, in the b-th replication, where $1 \leq b \leq 100$. We

evaluated the performance of the screening method by its true positive rate (TPR) and the frequency of sure screening (Sure):

$$\text{TPR} = 100^{-1}\sum_{b=1}^{100}\frac{|\mathcal{N}\bigcap\hat{\mathcal{S}}_b|}{|\mathcal{N}|}, \quad \text{Sure} = 100^{-1}\sum_{b=1}^{100} I\{\mathcal{N}\subseteq\hat{\mathcal{S}}_b\},$$

and that of the variable selection method by its false discovery rate (FDR), frequency of exactly selecting the true model (Exact), and averaged model size (AMS):

$$\text{FDR} = 100^{-1}\sum_{b=1}^{100}\frac{|\mathcal{N}^c\bigcap\hat{\mathcal{T}}_b|}{|\hat{\mathcal{T}}_b|}, \quad \text{Exact} = 100^{-1}\sum_{b=1}^{100} I\{\mathcal{N}=\hat{\mathcal{T}}_b\}, \quad \text{AMS} = 100^{-1}\sum_{b=1}^{100}|\hat{\mathcal{T}}_b|.$$

These performance measures are summarized in Tables 1 for the case of $(n, p) = (200, 10000)$, and Table 2 for the case of $(n, p) = (400, 10000)$.

As shown in Tables 1 and 2, the performance of gCGA(m), $m \in \{1, 10, 30, 50\}$, is quite satisfactory in Scenarios 1 and 3 because their TPR and Sure values are close to 1. These methods have TPR and Sure values distant from 1 in Scenario 2 with $n = 200$, but equal to 1 as n increases to 400. In Scenario 4, gCGA(1)'s TPR and Sure values are much less than 1 in the case of $n = 200$ and cannot be improved by increasing n. Although gCGA(m)'s performance is also unsatisfactory for $m \in \{10, 30, 50, 60\}$ in Scenario 4 with $n = 200$, it is greatly improved when n grows to 400. This shows that gCGA(m), with $m \geq 10$, indeed borrows from FR's strengths to substantially enhance its screening performance in difficult situations such

as Scenario 4, where gCGA(1) does not work well. Moreover, gCGA(m)'s performance tends to increase with m in all scenarios and gCGA($\hat{m}$) performs equally well as gCGA(50). The screening performance of marginal methods SIS, CSIS, and ISIS is in general inferior to gCGA($\hat{m}$). Their performance, however, improves when the t in $\mathcal{C}_t$ increases (see the item (c) in Section 4). In other words, ISIS is better than CSIS, and CSIS is better than SIS. We therefore focus on the comparison between gCGA($\hat{m}$) and ISIS. Note first that when $n = 200$, the two methods are largely comparable in Scenarios 1 and 4, and Scenario 2 at a censoring rate of 50%, but the former significantly outperforms the latter in all other scenarios. When $n = 400$, ISIS is comparable with gCGA($\hat{m}$) in Scenarios 1 and 3, but its performance is obviously poorer than that of gCGA($\hat{m}$) in Scenarios 2 and 4. The TPR and Sure values of LASSO are close to that of gCGA($\hat{m}$) in Scenarios 1, 2, and 4 with $n = 200$, but are much lower than the latter in Scenario 3 with the same sample size. When n increases to 400, LASSO improves substantially in Scenario 3, and both methods exhibits almost perfect performance in the first three scenarios. In Scenario 4, however, LASSO's TPR and Sure values do not increase with the sample size, resulting in screening performance worse than that of gCGA($\hat{m}$) in the case of $n = 400$.

The selection accuracy of gCGA($\hat{m}$)+HDIC depends mainly on the screening performance of gCGA($\hat{m}$) and on whether HDIC can successfully remove the redundant variables from those included by gCGA($\hat{m}$), while retaining the relevant ones. Tables 1 and 2 suggest that HDIC can indeed do a good job because the Exact value of gCGA($\hat{m}$)+Trim is almost equivalent to the Sure values of gCGA($\hat{m}$). Note that this Sure–Exact equivalence does not occur in any other screening-selection pairs considered in this section. When $n = 400$, the Exact value of gCGA($\hat{m}$)+Trim is equal (or close) to 1 in Scenarios 1–3. The selection performance of the marginal methods is obviously inferior to that of gCGA($\hat{m}$)+Trim. Their Exact values are high only in cases such as CSIS+SCAD and ISIS+SCAD in Scenario 1 at a censoring rate of 20%, and ISIS+SCAD in Scenario 3. ALASSO's selection performance lies between gCGA($\hat{m}$)+Trim and the marginal methods in the first three scenarios. Its Exact value, however, falls to 0 in Scenario 4, which is partly due to the equally low Sure value of LASSO. The Exact values of all methods in the case of $n = 200$ are, in general, smaller than that in the case of $n = 400$. However, gCGA($\hat{m}$)+Trim still performs satisfactorily in Scenarios 1 and 3, even at a censoring rate of 50%.

As pointed out by a reviewer, the proposed gCGA(m) seems applicable to the case when marginal weak but jointly strong signals appear in the

interaction term. To see this, we explore the performance of gCGA(m) on the Cox model involving two-way interaction terms. Denote $\boldsymbol{Z}'\boldsymbol{\beta}^*$ in (2.1) as

$$\beta_1^* Z_1 + \cdots + \beta_p^* Z_p + \beta_{1,2}^* Z_{1,2} + \cdots + \beta_{p,p-1}^* Z_{p,p-1}$$

with $Z_{i,j} = Z_i Z_j$. Under $(n, p) = (400, 200)$ and properly designed $\boldsymbol{\beta}^*$ and $\boldsymbol{Z}$, there are three (out of 20100) relevant variables $Z_1, Z_{1,2}$ and $Z_{1,3}$ in the above Cox model, where the main effect Z_1 and the interaction term $Z_{1,3}$ are marginal weak but jointly strong signals. The details of setting and the result are listed in supplementary material. Note that the true model follows the so-called weak heredity principle because at least one of the main effects is present when an interaction term is included in the model. The result shows that gCGA($\hat{m}$)+Trim can offer satisfying variable selection results and outperform the other methods in high-dimensional Cox models with interaction terms, in which some marginally weak but jointly strong main and interaction effects appear.

To conclude this section, we mention that gCGA($\hat{m}$) and gCGA($\hat{m}$)+Trim have excellent performance in screening and selection that surpasses all other methods under consideration. In particular, when $(n, p) = (400, 10000)$, they perform almost perfectly over Scenarios 1–4, some of which seem very challenging owing to the high correlations between relevant and irrelevant

variables. Furthermore, as displayed in the supplementary the computing time for gCGA(m) grows linearly with m, indicating that our proposed method gCGA($\hat{m}$) with $\hat{m}$ chosen from $\mathcal{Q} \subseteq \{1, \ldots, 50\}$ offers a substantial improvement in speed over gCGA(10000), which is equivalent to FR.

Table 1: Results for $(n, p) = (200, 10000)$ under Scenarios 1–4.

Censor Rate	20%					50%				
	TPR	Sure	FDR	Exact	AMS	TPR	Sure	FDR	Exact	AMS
AR(1) correlation										
gCGA(1)+Trim	1.00	1.00	0.00	1.00	5.00	0.96	0.91	0.00	0.90	4.79
gCGA(10)+Trim	1.00	1.00	0.00	0.99	4.97	0.99	0.97	0.00	0.95	4.90
gCGA(30)+Trim	1.00	1.00	0.00	0.99	4.97	0.99	0.97	0.00	0.95	4.90
gCGA(50)+Trim	1.00	1.00	0.00	0.99	4.97	0.99	0.97	0.00	0.95	4.90
gCGA($\hat{m}$)+Trim	1.00	1.00	0.00	1.00	5.00	0.99	0.98	0.00	0.96	4.92
SIS+SCAD	0.86	0.30	0.10	0.29	4.70	0.81	0.17	0.14	0.11	4.12
CSIS+SCAD	0.99	0.94	0.15	0.42	6.08	0.94	0.72	0.22	0.11	5.89
ISIS+SCAD	1.00	0.99	0.33	0.15	8.15	0.95	0.77	0.22	0.09	5.87
ALASSO	1.00	0.98	0.00	0.97	4.97	0.97	0.83	0.03	0.65	4.86
Equi-correlation										
gCGA(1)+Trim	0.75	0.55	0.16	0.31	5.92	0.28	0.00	0.35	0.00	1.59
gCGA(10)+Trim	0.78	0.63	0.14	0.44	7.81	0.29	0.02	0.39	0.00	1.68
gCGA(30)+Trim	0.78	0.64	0.15	0.43	7.71	0.30	0.03	0.38	0.00	1.70
gCGA(50)+Trim	0.79	0.65	0.15	0.43	7.70	0.30	0.03	0.37	0.00	1.70
gCGA($\hat{m}$)+Trim	0.87	0.78	0.12	0.51	8.75	0.32	0.03	0.36	0.00	1.68
SIS+SCAD	0.34	0.00	0.17	0.00	2.84	0.29	0.00	0.25	0.00	3.04
CSIS+SCAD	0.41	0.00	0.13	0.00	3.06	0.32	0.00	0.23	0.00	3.24
ISIS+SCAD	0.46	0.12	0.14	0.00	4.40	0.32	0.00	0.23	0.00	3.24
ALASSO	0.87	0.55	0.08	0.22	10.96	0.54	0.00	0.17	0.00	2.69
Marginally weak but jointly strong signals I										
gCGA(1)+Trim	0.99	0.99	0.00	0.99	2.98	0.96	0.94	0.00	0.94	2.88
gCGA(10)+Trim	1.00	1.00	0.00	1.00	3.00	0.99	0.99	0.00	0.99	2.98
gCGA(30)+Trim	1.00	1.00	0.00	1.00	3.00	1.00	1.00	0.00	1.00	3.00
gCGA(50)+Trim	1.00	1.00	0.00	1.00	3.00	1.00	1.00	0.00	1.00	3.00
gCGA($\hat{m}$)+Trim	1.00	1.00	0.00	1.00	3.00	1.00	1.00	0.00	1.00	3.00
SIS+SCAD	0.34	0.00	0.04	0.00	1.12	0.34	0.01	0.03	0.01	1.08
CSIS+SCAD	0.67	0.02	0.12	0.02	1.97	0.67	0.02	0.06	0.00	1.50
ISIS+SCAD	0.82	0.45	0.17	0.12	3.14	0.74	0.22	0.11	0.01	2.26
ALASSO	0.63	0.02	0.08	0.02	1.86	0.57	0.00	0.07	0.00	1.48
Marginally weak but jointly strong signals II										
gCGA(1)+Trim	0.67	0.00	0.72	0.00	7.23	0.67	0.00	0.61	0.00	4.98
gCGA(10)+Trim	0.68	0.03	0.71	0.03	7.33	0.67	0.01	0.64	0.01	5.60
gCGA(30)+Trim	0.70	0.11	0.65	0.11	6.90	0.67	0.02	0.64	0.02	5.68
gCGA(50)+Trim	0.72	0.17	0.60	0.17	6.69	0.68	0.05	0.62	0.05	5.57
gCGA($\hat{m}$)+Trim	0.72	0.17	0.61	0.17	6.74	0.68	0.05	0.62	0.05	5.58
SIS+SCAD	0.33	0.00	0.24	0.00	2.25	0.33	0.00	0.20	0.00	1.80
CSIS+SCAD	0.64	0.01	0.30	0.00	4.44	0.57	0.00	0.22	0.00	3.18
ISIS+SCAD	0.71	0.18	0.24	0.11	6.58	0.60	0.06	0.20	0.04	4.68
ALASSO	0.67	0.00	0.06	0.00	1.18	0.67	0.00	0.12	0.00	1.47

Table 2: Results for $(n, p) = (400, 10000)$ under Scenarios 1–4.

Censor Rate	20%					50%				
	TPR	Sure	FDR	Exact	AMS	TPR	Sure	FDR	Exact	AMS
AR(1) correlation										
gCGA(1)+Trim	1.00	1.00	0.00	1.00	5.00	1.00	1.00	0.00	1.00	5.00
gCGA(10)+Trim	1.00	1.00	0.00	1.00	5.00	1.00	1.00	0.00	1.00	5.00
gCGA(30)+Trim	1.00	1.00	0.00	1.00	5.00	1.00	1.00	0.00	1.00	5.00
gCGA(50)+Trim	1.00	1.00	0.00	1.00	5.00	1.00	1.00	0.00	1.00	5.00
gCGA($\hat{m}$)+Trim	1.00	1.00	0.00	1.00	5.00	1.00	1.00	0.00	1.00	5.00
SIS+SCAD	0.88	0.41	0.03	0.41	4.53	0.86	0.33	0.07	0.33	4.54
CSIS+SCAD	1.00	1.00	0.01	0.94	5.06	0.99	0.97	0.09	0.59	5.58
ISIS+SCAD	1.00	1.00	0.02	0.90	5.10	1.00	1.00	0.15	0.46	6.28
ALASSO	1.00	1.00	0.00	1.00	5.00	1.00	1.00	0.00	0.97	5.01
Equi-correlation										
gCGA(1)+Trim	1.00	1.00	0.00	1.00	15.00	1.00	1.00	0.02	0.94	14.34
gCGA(10)+Trim	1.00	1.00	0.00	1.00	15.00	1.00	1.00	0.01	0.98	14.78
gCGA(30)+Trim	1.00	1.00	0.00	1.00	15.00	1.00	1.00	0.00	0.99	14.89
gCGA(50)+Trim	1.00	1.00	0.00	1.00	15.00	1.00	1.00	0.00	0.99	14.89
gCGA($\hat{m}$)+Trim	1.00	1.00	0.00	1.00	15.00	1.00	1.00	0.00	0.98	14.79
SIS+SCAD	0.54	0.00	0.10	0.00	7.14	0.51	0.00	0.14	0.00	5.90
CSIS+SCAD	0.80	0.07	0.03	0.05	10.93	0.69	0.02	0.09	0.02	8.88
ISIS+SCAD	0.91	0.82	0.05	0.38	13.18	0.86	0.69	0.12	0.09	12.74
ALASSO	1.00	1.00	0.00	0.99	15.01	1.00	0.96	0.02	0.65	15.18
Marginally weak but jointly strong signals I										
gCGA(1)+Trim	1.00	1.00	0.00	1.00	3.00	1.00	1.00	0.00	1.00	3.00
gCGA(10)+Trim	1.00	1.00	0.00	1.00	3.00	1.00	1.00	0.00	1.00	3.00
gCGA(30)+Trim	1.00	1.00	0.00	1.00	3.00	1.00	1.00	0.00	1.00	3.00
gCGA(50)+Trim	1.00	1.00	0.00	1.00	3.00	1.00	1.00	0.00	1.00	3.00
gCGA($\hat{m}$)+Trim	1.00	1.00	0.00	1.00	3.00	1.00	1.00	0.00	1.00	3.00
SIS+SCAD	0.34	0.00	0.01	0.00	1.04	0.34	0.00	0.00	0.00	1.02
CSIS+SCAD	0.68	0.05	0.03	0.05	2.14	0.67	0.01	0.04	0.01	2.13
ISIS+SCAD	1.00	1.00	0.00	0.98	3.02	1.00	1.00	0.07	0.76	3.30
ALASSO	1.00	1.00	0.00	1.00	3.00	0.96	0.87	0.00	0.83	2.84
Marginally weak but jointly strong signals II										
gCGA(1)+Trim	0.74	0.21	0.64	0.21	9.09	0.70	0.11	0.67	0.11	7.47
gCGA(10)+Trim	0.94	0.82	0.15	0.82	4.38	0.83	0.50	0.38	0.50	5.69
gCGA(30)+Trim	0.99	0.97	0.02	0.97	3.21	0.89	0.67	0.25	0.67	4.82
gCGA(50)+Trim	1.00	0.99	0.01	0.99	3.07	0.92	0.75	0.19	0.75	4.37
gCGA($\hat{m}$)+Trim	1.00	0.99	0.01	0.99	3.07	0.91	0.74	0.20	0.74	4.44
SIS+SCAD	0.34	0.00	0.32	0.00	3.80	0.34	0.00	0.19	0.00	2.28
CSIS+SCAD	0.72	0.15	0.25	0.15	7.05	0.68	0.05	0.21	0.05	4.91
ISIS+SCAD	0.81	0.43	0.01	0.43	1.88	0.72	0.18	0.10	0.17	3.86
ALASSO	0.67	0.00	0.00	0.00	1.03	0.67	0.00	0.02	0.00	1.06

5. Data Analysis

We apply our proposed method to the study of Metzeler et al. (2008). The primary concern here is to identify the gene signatures relevant to overall survival in patients who are diagnosed with cytogenetically normal acute myeloid leukaemia. In this study, the training cohort consisted of 163 adult patients, from whom a total of 44754 gene signatures is recorded using Affymetrix HG-U133 A+B microarrays. The median survival time in the training cohort is 9.4 months with a censoring rate of 37%. In addition, an independent sample consisting of 79 patients on Affymetrix HG-U133 Plus 2.0 microarrays, is used as the test cohort, whose median survival time is 15.7 months with a censoring rate of 41%. Following Metzeler et al. (2008), all gene expressions are centered and rescaled. This dataset is publicly available on the gene expression omnibus website (`http://www.ncbi.nlm.nih.gov/geo/`) under the accession number GSE12417.

We consider gCGA($\hat{m}$)+Trim, with $\mathcal{Q} = \{1, 10, 30, 50\}$, and other four variable selection methods introduced in Section 4; all methods are applied to the training cohort to select relevant genes. To validate, the concordance statistics (C-statistics) as well as the area under the curve (AUC) developed by Uno et al. (2011) are calculated based on the test cohort. The prediction performance is reported in Table 3, revealing that the resultant 19

Table 3: Summary of prediction performance for gene signatures selected from different methods in CN-AML data.

	gCGA($\hat{m}$)+Trim	SIS+SCAD	CSIS+SCAD	ISIS+SCAD	ALASSO
C-statistic	0.618	0.610	0.579	0.618	0.582
AUC	0.626	0.603	0.544	0.592	0.552
Model size	19	7	10	32	10

gene signatures selected by our method possess more predictive power. In particular, the first 3 genes (SOSC2, AXL, and NCR3LG1) surviving the screening and selection stages of gCGA($\hat{m}$)+Trim deserve a separate look. The first gene signature (SOSC2) is known to be associated with patients' overall survival in CN-AML (Metzeler et al., 2008), but is not discovered by any other methods under consideration. On the other hand, the second gene (AXL) and the third gene (NCR3LG1) are identified simultaneously by CSIS+SCAD, ISIS+SCAD, and ALASSO. Therefore, we conclude that gCGA($\hat{m}$)+Trim yields reliable importance ranking for gene signatures, and leads to an interpretative sparse model with competitive prediction power.

6. Concluding Remarks

In this paper, we propose using CGA, gCGA, and HDIC to select variables for high-dimensional Cox models. The novelty of this paper is threefold:

first, under a weak sparsity condition, this paper shows that the convergence rate of CGA is coincident with the minimax-optimal rate obtained in high-dimensional linear models. Although, as noted by an Associate Editor, this rate is not necessarily minimax-optimal for high-dimensional Cox models, this coincidence suggests that CGA works reasonably well in such models; second, under a strong sparsity condition, this paper shows that gCGA can be used, together with HDIC, to achieve variable selection consistency, a property that has not been established previously for greedy-type algorithms in the case of high-dimensional Cox models; third, gCGA, our newly developed variable screening method, combines CGA's computational efficiency and FR's finite-sample accuracy. In particular, our experimental results show that gCGA($\hat{m}$)+HDIC outperforms ALASSO and marginal methods, and exhibits excellent selection accuracy even in challenging situations where marginally weak but jointly strong signals are present and highly correlated with the irrelevant variables.

On the other hand, the performance of the proposed methods in high-dimensional Cox models with interaction terms is yet to be explored. Model selection for this kind of model can be applied to identify gene-gene interactions associated with patients' overall survival in lung adenocarcinoma (Wu, Huang and Ma, 2018), and merits future research.

Supplementary Materials

The supplementary material contains detailed proofs for the theoretical results, and additional simulations for various settings and for illustrating the time cost of gCGA(m).

Acknowledgements

We would like to thank the associate editor and two anonymous referees for their critical comments and thoughtful suggestions, which led to an improved version of this paper. Lin, Cheng, and Ings research was supported, in part, by grants 111-2118-M-035-007-MY2, 110-2628-M-007-003-MY2, and 109-2118-M-007-007-MY3 from the Ministry of Science and Technology (MOST), Taiwan.

References

Barut, E., Fan, J., and Verhasselt, A. (2016). Conditional Sure Independence Screening. *J. Amer. Statist. Assoc.* **111**, 1266–1277.

Bradic, J., Fan, J., and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.* **39**, 3092–3120.

Bühlmann, P. (2006). Boosting for High-dimensional Linear Models. *Ann. Statist.* **34**, 559–583.

Elenberg, E. R., Khanna, R., Dimakis, A. G., and Negahban, S. (2018). Restricted Strong Convexity Implies Weak Submodularity. *Ann. Statist.* **46**, 3539–3568.

Fan, J., Feng, Y., and Wu, Y. (2010). High-dimensional Variable Selection for Cox's Proportional Hazards Model. In *Borrowing strength: Theory powering applications—a Festschrift for Lawrence D. Brown, Institute of Mathematical Statistics* 70–86. 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 Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30(1)**,74-99.

Fan, J., and Lv, J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space. *J. Roy. Statist. Soc. Ser. B* **70**, 849–911.

He, K., Li, Y., Zhu, J., Liu, H., Lee, J. E., Amos, C. I., and Li, Y. (2016). Component-wise Gradient Boosting and False Discovery Control in Survival Analysis with High-dimensional Covariates. *Bioinformatics* **32**, 50–57.

Hong, H. G., Kang, J., and Li, Y. (2018). Conditional Screening for Ultra-high Dimensional Covariates with Survival Outcomes. *Lifetime Data Anal.* **24**, 45–71.

Hong, H. G., Zheng, Q., and Li, Y. (2019). Forward Regression for Cox Models with High-dimensional Covariates. *J. Multivariate Anal.* **173**, 268–290.

Ing, C. K. (2020). Model Selection for High-dimensional Linear Regression with Dependent Observations. *Ann. Statist.* **48**, 1959–1980.

Ing, C. K., and Lai, T. L. (2011). A Stepwise Regression Method and Consistent Model Selection for High-dimensional Sparse Linear Models. *Statist. Sinica*, 1473–1513.

Kong, S., and Nan, B. (2014). Non-asymptotic Oracle Inequalities for the High-dimensional Cox Regression via Lasso. *Statist. Sinica* **24**, 25–42.

Luo, S., Xu, J., and Chen, Z. (2015). Extended Bayesian Information Criterion in the Cox Model with a High-dimensional Feature Space. *Ann. Inst. Stat. Math.* **67**, 287–311.

Metzeler, K., Hummel, M., Bloomfield, C., Spiekermann, K., Braess, J., Sauerland, M.-C., Heinecke, A., Radmacher, M., Marcucci, G., Whitman, S., Maharry, K., Paschka, P., Larson, R., Berdel, W., Buchner, T., Wormann, B., Mansmann, U., Hiddemann, W., Bohlander, S. and Buske, C. (2008). An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood* **112**, 4193–4201.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57**, 6976–6994.

Song, R., Lu, W., Ma, S., and Jessie Jeng, X. (2014). Censored Rank Independence Screening for High-dimensional Survival Data. *Biometrika* **101**, 799–814.

Temlyakov, V. N. (2015). Greedy Approximation in Convex Optimization. *Constr. Approx.* **41**, 269–296.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in medicine* **16(4)**, 385–395.

Tropp, J. A., and Gilbert, A. C. (2007). Signal Recovery from Random Measurements via Orthogonal Matching Pursuit. *IEEE Trans. Inform. Theory* **53**, 4655–4666.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* **30**, 1105–1117.

Wang, H. (2009). Forward Regression for Ultra-high Dimensional Variable Screening. *J. Amer. Statist. Assoc.* **104**, 1512–1524.

Wu, M., Huang, J., and Ma, S. (2018). Identifying gene-gene interactions using penalized tensor regression. *Stat. Med.* **37**, 598–610.

Zhang, H. H., and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika.* **94(3)**, 691–703.

Department of Statistics, Feng Chia University, Taiwan

E-mail: ctlin@fcu.edu.tw

Institute of Statistics, National Tsing Hua University, Taiwan

E-mail: ycheng@stat.nthu.edu.tw

Institute of Statistics, National Tsing Hua University, Taiwan

E-mail: cking@stat.nthu.edu.tw