Published in final edited form as:

Stat Probab Lett. 2018 April ; 135: 20–25. doi:10.1016/j.spl.2017.11.014.

Screening group variables in the proportional hazards model

Kwang Woo Ahn, Natasha Sahr, and Soyoung Kim

Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53132 USA

Abstract

We propose a method to screen group variables under the high dimensional group variable setting for the proportional hazards model. We study the sure screening property of the proposed method for independent and clustered survival data. The simulation study shows that the proposed method performs better for group variable screening than some existing procedures.

Keywords

Proportional hazards model; Sure screening property; Joint screening; High dimensional data

1. Introduction

There is rich literature on group and within-group variable selection with possibly overlapping groups for the proportional hazards (PH) model. Huang et al. (2014) proposed the group bridge for bi-level selection when the number of covariates p is less than the number of observations n . An alternative approach by Wu and Wang (2013) considered double penalties: lasso and group lasso penalties for $p > n$. However, it lacks theoretical justification. Wang et al. (2009) proposed a hierarchically penalized Cox regression. They used an adaptive penalty to achieve the bi-level selection consistency when $p < n$. For $p > n$, they proposed to use the ridge regression to construct the adaptive penalty without theoretical justification. Although these penalized regression methods may work when p is relatively larger than n , they may be computationally expensive and unstable when p is much larger than n ($p \gg n$).

Screening techniques for the PH model with $p \gg n$ have recently been given much attention (Zhao and Li, 2012; Yang et al., 2016). However, these existing methods are all limited to individual variable screening for independent survival data to the best of the authors' knowledge. Therefore, motivated by Yang et al. (2016), we propose a sure group joint screening (SGJS) procedure to screen group variables for the PH models when $p \gg n$. We show the SGJS enjoys the sure screening property for group variable screening under

Correspondence to: Kwang Woo Ahn.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

independent and clustered survival data. We also propose two ways to handle overlapping group variables for screening. Simulation study is also conducted.

2. Sure group joint variable screening

2.1. Notations and assumptions

We first define the notations used throughout the rest of the paper. We assume there are m clusters and the i th cluster has L individuals. The clusters may have different sizes by defining censoring times as zero when observed times are missing (Spiekerman and Lin, 1998). We denote $n = mL$ as the total sample size. It is assumed that there are p number of covariates. Let T_{ij} , C_{ij} , and $\mathbf{Z}_{ij} = (Z_{ij1}, \dots, Z_{ijp})^T$ be the event time, censoring time, and covariate vector of individual j in cluster i , respectively, for $i = 1, \dots, m$ and $j = 1, \dots, L$. The covariate vector $\mathbf{Z}_{ij}$ may depend on time t . We omit dependency on t in $\mathbf{Z}_{ij}(t)$ when the context is clear. Without loss of generality, we assume that $\mathbf{Z}_{ij}$'s are standardized. The parameter vector of interest is denoted as $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$. Define $\boldsymbol{\beta}_0 = (\beta_{10}, \dots, \beta_{p0})^T$ as the true parameter vector. Let $\mathbf{T}_i = \{T_{ij}, j = 1, \dots, L\}$, $\mathbf{C}_i = \{C_{ij}, j = 1, \dots, L\}$, and $\mathbf{Z}_i = \{\mathbf{Z}_{ij}, j = 1, \dots, L\}$. Suppose that $(\mathbf{T}_i, \mathbf{C}_i, \mathbf{Z}_i)$ are independent and identically distributed. We assume that the T_{ij} 's are independent of the C_{ij} 's given $\mathbf{Z}_{ij}$ for $i = 1, \dots, m$ and $j = 1, \dots, L$.

Let $X_{ij} = T_{ij} \wedge C_{ij}$ be the observed time and $\delta_{ij} = \mathbb{I}(T_{ij} < C_{ij})$, where $a \wedge b = \min(a, b)$ and $\mathbb{I}(\cdot)$ is an indicator function. Let $N_{ij}(t) = \mathbb{I}(X_{ij} \geq t, \delta_{ij} = 1)$ and $Y_{ij}(t) = \mathbb{I}(X_{ij} \geq t)$. The study period is assumed to be $[0, \tau]$. Given covariate vector $\mathbf{Z}$, the hazard function $\lambda(t | \mathbf{Z})$ is defined as $\lambda(t | \mathbf{Z}) = \lambda_0(t) \exp(\mathbf{Z}^T \boldsymbol{\beta})$, where $\lambda_0(t)$ is an unspecified baseline hazard function at time t . Then, Spiekerman and Lin (1998) proposed the marginal log-partial likelihood function for clustered survival data as follows:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^m \sum_{j=1}^L \Delta_{ij} \left[\boldsymbol{\beta}^T \mathbf{Z}_{ij} - \log \left\{ \sum_{a=1}^m \sum_{b=1}^L Y_{ab}(X_{ij}) \exp(\boldsymbol{\beta}^T \mathbf{Z}_{ab}) \right\} \right]. \quad (1)$$

Next, we define some notations on group variables and their memberships. Assume that we have K groups of variables. Let $A_1, \dots, A_K$ be subsets of $\{1, \dots, p\}$ representing group memberships of variables. Define $\boldsymbol{\beta}_A = (\beta_j, j \in A)^T$ and $\boldsymbol{\beta}_{A,0} = (\beta_{j0}, j \in A)^T$ for a set A . Denote $|A|$ as the cardinality of a set A . Without loss of generality, for disjoint groups we assume that β_j 's are on the order of groups such that $\boldsymbol{\beta} = (\boldsymbol{\beta}_{A_1}^T, \dots, \boldsymbol{\beta}_{A_K}^T)^T$, where $\boldsymbol{\beta}_{A_1} = (\beta_1, \dots, \beta_{|A_1|})^T$ and $\boldsymbol{\beta}_{A_k} = (\beta_{|A_1|+1}, \dots, \beta_{|A_1|+|A_k|})^T$ for $k = 2, \dots, K$.

2.2. Sure group joint variable screening for disjoint groups

In this section, we study the sure group joint variable screening when A_k 's are disjoint. To screen zero group variables for the PH model, we propose the following penalized partial likelihood:

$$\arg \max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) \quad \text{subject to number of non-zero groups} \leq q, \quad (2)$$

where q is pre-specified. We call this as the sure group joint screening (SGJS). Because $p > n$, it is not feasible to directly solve (2). Thus, motivated by the Taylor expansion of $\ell(\boldsymbol{\alpha})$ at $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\alpha}$, Yang et al. (2016) considered the following function to approximate the log-partial likelihood function:

$$g(\boldsymbol{\alpha} | \boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + (\boldsymbol{\alpha} - \boldsymbol{\beta})^T \ell'(\boldsymbol{\beta}) - \frac{u}{2} (\boldsymbol{\alpha} - \boldsymbol{\beta})^T \mathbf{W}(\boldsymbol{\beta})(\boldsymbol{\alpha} - \boldsymbol{\beta}), \quad (3)$$

where u is a pre-specified constant and $\mathbf{W}(\boldsymbol{\beta})$ is a diagonal matrix consisting of the diagonal elements of $-\ell''(\boldsymbol{\beta})$. Yang et al. (2016) used $g(\boldsymbol{\alpha} | \boldsymbol{\beta})$ in the optimization procedure to maximize $\ell(\boldsymbol{\beta})$ with L_0 penalty.

The function $g(\boldsymbol{\alpha} | \boldsymbol{\beta})$ in (3) may also be used to solve (2). However, for group structured variables, variables within the same group often tend to be more correlated than variables between different groups. Motivated by this, we consider a block diagonal matrix $\mathbf{W}^*(\boldsymbol{\beta})$ instead of the diagonal matrix $\mathbf{W}(\boldsymbol{\beta})$ in (3), where each block in $\mathbf{W}^*(\boldsymbol{\beta})$ consists of the sub-square matrix of $-\ell''(\boldsymbol{\beta})$ corresponding to each group. To elaborate $\mathbf{W}^*(\boldsymbol{\beta})$, define $|A_0| = 0$. For $k = 1, \dots, K$, we define $\mathbf{W}_k^*(\boldsymbol{\beta}_{A_k})$ as the sub-square matrix of $-\ell''(\boldsymbol{\beta})$ corresponding to $\boldsymbol{\beta}_{A_k}$ as follows:

$$\mathbf{W}_k^*(\boldsymbol{\beta}_{A_k}) = - \begin{pmatrix} \ell''(\boldsymbol{\beta})_{|A_{k-1}|+1, |A_{k-1}|+1} & \dots & \ell''(\boldsymbol{\beta})_{|A_{k-1}|+1, |A_k|} \\ \vdots & \ddots & \vdots \\ \ell''(\boldsymbol{\beta})_{|A_k|, |A_{k-1}|+1} & \dots & \ell''(\boldsymbol{\beta})_{|A_k|, |A_k|} \end{pmatrix},$$

where $\ell''(\boldsymbol{\beta})_{a,b}$ is the (a, b) th entry of $\ell''(\boldsymbol{\beta})$. Define $\mathbf{W}^*(\boldsymbol{\beta})$ as follows:

$$\mathbf{W}^*(\boldsymbol{\beta}) = \begin{pmatrix} \mathbf{W}_1^*(\boldsymbol{\beta}_{A_1}) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2^*(\boldsymbol{\beta}_{A_2}) & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{W}_K^*(\boldsymbol{\beta}_{A_K}) \end{pmatrix},$$

where all non-block-diagonal elements are zeros. For group variable screening, we propose the following approximation at $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\alpha}$:

$$g^*(\boldsymbol{\alpha} | \boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + (\boldsymbol{\alpha} - \boldsymbol{\beta})^T \ell'(\boldsymbol{\beta}) - \frac{u}{2} (\boldsymbol{\alpha} - \boldsymbol{\beta})^T \mathbf{W}^*(\boldsymbol{\beta})(\boldsymbol{\alpha} - \boldsymbol{\beta}). \quad (4)$$

For group variable screening, we propose to maximize $g^*(\boldsymbol{\alpha} | \boldsymbol{\beta})$ as follows:

$$\max_{\alpha} g^*(\alpha | \beta) \quad \text{subject to number of non-zero groups} \leq q. \quad (5)$$

Because of the block-diagonal structure in $\mathbf{W}^*(\beta)$, $g^*(\alpha | \beta)$ is maximized at $\tilde{\alpha} = \beta + u^{-1} \{\mathbf{W}^*(\beta)\}^{-1} \ell'(\beta)$. Let $\tilde{\alpha} = (\tilde{\alpha}_{A_1}^T, \dots, \tilde{\alpha}_{A_K}^T)^T$, where $\tilde{\alpha}_{A_k} = (\tilde{\alpha}_{|A_k-1|+1}, \dots, \tilde{\alpha}_{|A_k-1|+|A_k|})^T$ for $k = 1, \dots, K$. Define $a_k = \tilde{\alpha}_{A_k}^T \mathbf{W}_{A_k}^*(\beta_{A_k}) \tilde{\alpha}_{A_k}$ for $k = 1, \dots, K$. We sort $|a_k|$'s in decreasing order such as $|a_{(1)}| \geq \dots \geq |a_{(K)}|$, where $|a_{(j)}|$ is the j th greatest among $|a_k|$'s. Then, the solution to (5) is

$$\hat{\alpha}_{A_k} = \tilde{\alpha}_{A_k} I\{|a_k| > |a_{(q+1)}|\}, k = 1, \dots, K.$$

To obtain a solution to (2), we use the following algorithm:

1. Set the initial vector $\beta_{(0)} = \mathbf{0}$.
2. Given $\beta^{(r)}$, solve (5) and obtain $\hat{\alpha}_{A_k}^{(r)} = \tilde{\alpha}_{A_k}^{(r)} I\{|a_k^{(r)}| > |a_{(q+1)}^{(r)}|\}$ for $k = 1, \dots, K$.
3. Set $G^{(r)} = \{k: \hat{\alpha}_k^{(r)} \neq \mathbf{0}\}$ as the index set of the non-zero groups.
4. Obtain $\beta^{(r+1)}$ as follows:
 - a. For $k \notin G^{(r)}$, set $\beta_{A_k}^{(r+1)}$ to $\mathbf{0}$;
 - b. For $k \in G^{(r)}$, $\beta_{A_k}^{(r+1)}$'s are estimated by maximizing the marginal partial likelihood (1) with parameters $\{\beta_{A_k}; k \in G^{(r)}\}$.
5. Repeat 2 to 4 until $\|\beta^{(r+1)} - \beta^{(r)}\| < 10^{-4}$.

One may set $\beta_{A_k}^{(r+1)} = \hat{\alpha}_{A_k}^{(r)}$ for $k \in G^{(r)}$ in Step 4 (b) instead of maximizing (1). However, when $G^{(r)}$ and $G^{(r+1)}$ are the same, $\beta^{(r+1)}$ will be the same as $\beta^{(r)}$ in the proposed algorithm so that the optimization ends at the $(r+1)$ th iteration. On the other hand, using $\hat{\alpha}_{A_k}^{(r)}$ for $k \in G^{(r)}$ in Step 4 (b) may lead to (slightly) different $\beta^{(r)}$ and $\beta^{(r+1)}$ so that one may need extra iterations to obtain convergence.

Next, we study the main theorems of the SGJS. Denote an index set of groups G as a subset of $\{1, \dots, K\}$. Let $G_0 = \{k: \beta_{A_k,0} \neq \mathbf{0}\}$ be the index set of the true non-zero groups. For the parameter estimator $\beta^{(r)}$ at the r th iteration in the optimization algorithm, let

$$\kappa^{(r)} = \sup_{\beta} \left(\lambda_{\max} \left[-\{\mathbf{W}^*(\beta^{(r)})\}^{-1/2} \ell''(\beta) \{\mathbf{W}^*(\beta^{(r)})\}^{-1/2} \right] \right),$$

where $\lambda_{\max}(\mathbf{A})$ is the largest eigenvalue of a matrix $\mathbf{A}$. Then, we have the following theorem:

Theorem 1—Assume (C1)–(C4) in the Appendix. If $u = \kappa^{(r)}$, then $\ell(\boldsymbol{\beta}^{(r+1)}) = \ell(\boldsymbol{\beta}^{(r)})$ in the optimization algorithm.

Theorem 1 can be shown similarly to Theorem 1 of Yang et al. (2016) and thus it is omitted. It shows the ascent property of the proposed algorithm. To examine whether u is large enough, one can start with $u = 10^6$. At each iteration, one can calculate

$$\kappa_*^{(r)} = \left(\lambda_{\max} \left[-\{\mathbf{W}^*(\boldsymbol{\beta}^{(r)})\}^{-1/2} \ell''(\boldsymbol{\beta}^{(r)}) \{\mathbf{W}^*(\boldsymbol{\beta}^{(r)})\}^{-1/2} \right] \right)$$

to see whether 10^6 is larger than $\kappa_*^{(r)}$ for all r . A significantly increased value can be used for u if $\kappa_*^{(r)}$ for some r is larger than or close to 10^6 . The algorithm is not guaranteed to converge to the global maximizer. However, our simulation results show the algorithm with $u = 10^6$ worked proficiently under various settings.

Let $\hat{G}$ be the index set of the q non-zero groups selected by the SGJS. The following theorem shows the sure screening property of the SGJS:

Theorem 2—Under (C1)–(C8) in the Appendix, we have $P(G_0 \subset \hat{G}) \rightarrow 1$ as $m \rightarrow \infty$.

The proof of Theorem 2 is provided in the Appendix. One has to pre-specify q to apply the SGJS. Individual variable screening techniques often retain $Ln/\log n$ number of individual variables for independent data, see Fan and Lv (2008) and Yang et al. (2016). Motivated by this, we propose to set $Lq = m/(a_g \log m)$, where a_g is the average group size.

2.3. Sure group joint variable screening for groups with overlaps

We consider overlapping groups in this section. Without loss of generality, we assume that only two groups, say, Group 1 and Group 2, overlap and the other Groups 3, ..., K are disjoint. For $b < a$, let $\boldsymbol{\beta}_{A_1} = (\beta_1, \dots, \beta_a)^T$ and $\boldsymbol{\beta}_{A_2} = (\beta_{a-b+1}, \dots, \beta_a, \beta_{a+1}, \dots, \beta_{a+c})^T$. Thus, $\{\beta_{a-b+1}, \dots, \beta_a\}$ belong to both of Groups 1 and 2. For the purpose of screening group variables, we consider two methods, denoted by M1 and M2. For M1, we combine Groups 1 and 2. So we consider the combined group as one group. For M2, we make three disjoint groups from Groups 1 and 2 by making the overlapping coefficients $\{\beta_{a-b+1}, \dots, \beta_a\}$ as a separate group. Thus, the three disjoint groups consist of $\{\beta_1, \dots, \beta_{a-b}\}$, $\{\beta_{a-b+1}, \dots, \beta_a\}$, and $\{\beta_{a+1}, \dots, \beta_{a+c}\}$, respectively. For M1, if the combined group is selected by the SGJS, Groups 1 and 2 will be examined in the second step. For M2, if either $\{\beta_1, \dots, \beta_{a-b}\}$ or $\{\beta_{a+1}, \dots, \beta_{a+c}\}$ is solely selected, only either Group 1 or Group 2 will be examined in the second step. If $\{\beta_{a-b+1}, \dots, \beta_a\}$ is selected, both of Groups 1 and 2 will be examined in the second step. Last, if at least two groups out of the three groups are selected, both of Groups 1 and 2 will be examined in the second step. The main purpose of M1 and M2 is that we regroup Groups 1 and 2 so that all groups to be evaluated in the SGJS can be disjoint. Thus, we can apply the SGJS of Section 2.2 for screening regrouped variables. We can generalize this procedure to disjointly regroup the variables when some variables belong to more than 2 groups.

Compared to M2, M1 may boost a chance that the true non-zero groups having some overlapping groups are selected. However, zero group variables having overlaps with the true non-zero groups are also selected in that case. Thus, M1 may increase the number of variables to examine in the second step when the combined group size is large. On the other hand, M2 may be useful when the size of the combined group is relatively large for m .

3. Simulation

We performed a simulation study with 200 iterations, cause 1 of 40%, cause 2 of 30%, and overall censoring rate of 30%. Clustered survival times were generated from survival distribution $\exp\{-w \exp(\beta^T Z_{ij})t\}$, where random effect w were generated from a positive stable frailty distribution with parameter $0 < \psi < 1$. Smaller ψ was used to generate a larger correlation between event times within the cluster. Independent event times were generated with $\psi = 1$. Independent of survival times, clustered censoring times were generated from censoring distribution $\exp(-wt)$.

There were three group sizes: 2, 4, and 6. All variables were generated from multivariate normal distribution with mean 0, $\text{var}(Z_{ij}) = 1$, $\text{cov}(Z_{ij}, Z_{ij'}) = 0.5$ if $j = j'$, and $\text{cov}(Z_{ij}, Z_{i'j'}) = 0.2$ if $i = i'$. For overlapping group structure, we considered

$$(\beta_1, \dots, \beta_9) = \underbrace{(0, 0, 0, 0, 0, 0, 0, 0, 0)}_{\text{Group 1}}, \underbrace{(0, 0, 0, 0, 0, 0, 0, 0, 0)}_{\text{Group 2}}, \underbrace{(0, 0, 0, 0, 0, 0, 0, 0, 0)}_{\text{Group 3}}.$$

The other groups were disjoint. Among disjoint groups, there were three non-zero group variables with the true parameter vectors $(0.8, 0)^T$ for group size 2, $(0.7, -0.7, 0, 0)^T$ for group size 4, and $(0.8, -0.8, 0, 0, 0)^T$ for group size 6. All other group variables were zero. Thus, there were seven non-zero individual variables. Two $n = 200$ and 500 were considered with the number of disjoint groups 50 and 200 for each size of group, respectively, in addition to the overlapping Groups 1 to 3. Thus, there were $p = 609$ and 2409 for $n = 200$, and 500, respectively. Two $\psi = 0.25$ and 1 were examined for each n . For $\psi = 0.25$, the number of subjects for each cluster was 2.

As we discussed in Section 2.3, we examined two ways to handle overlapping groups. For M1, we treated $\beta_1, \dots, \beta_9$ as one group. For M2, we regrouped $\beta_1, \dots, \beta_9$ to 5 different groups: $\{\beta_1\}$, $\{\beta_2\}$, $\{\beta_3\}$, $\{\beta_4, \beta_5\}$, and $\{\beta_6, \dots, \beta_9\}$. We compared the SGJS with $\mathbf{W}^*(\beta)$ and M1/M2 to the SGJS with $\mathbf{W}(\beta)$ and M1/M2, the sure joint screening (SJS) of Yang et al. (2016), and the principled sure independence screening (PSIS) of Zhao and Li (2012). For the SJS and the PSIS, we retained $\lfloor n/\log n \rfloor$ number of individual variables for $\psi = 1$ as suggested in Fan and Lv (2008) and $\lfloor m/\log m \rfloor$ number of individual variables for $\psi = 0.25$. For the SGJS, we retained $q = \lfloor n/(4 \log n) \rfloor$ number of groups for $\psi = 1$ and $q = \lfloor m/(4 \log m) \rfloor$ number of groups for $\psi = 0.25$, where 4 is the average group size. The hierarchically penalized Cox regression of Wang et al. (2009) with $\nu = 1$ was used to further remove zero variables in the second step.

The simulation results are summarized in Table 1. P_1 is the proportion that all true non-zero individual variables are retained by each screening method in the first step. P_2 is the proportion that the method of Wang et al. (2009) correctly identified the 7 true non-zero individual variables in the second step. ‘Model Size’ indicate the mean number of selected individual variables in the second step. ‘SGJS*-M1’, ‘SGJS*-M2’, ‘SGJS-M1’, and ‘SGJS-M2’ indicate the SGJS procedures with $\mathbf{W}^*$ and M1, $\mathbf{W}^*$ and M2, $\mathbf{W}$ and M1, and $\mathbf{W}$ and M2, respectively. Since the result of SJS is almost identical to PSIS, it is omitted. Column P_1 of Table 1 shows that the SGJS with $\mathbf{W}^*$ and M1/M2 retained the true non-zero variables much better than the other methods. Compared to using M2, using M1 had slightly higher probability of retaining the true non-zero variables in general. As n increased, the SGJS with $\mathbf{W}^*$ and Wang et al. (2009) identified the true non-zero variables better as in Column P_2 . Its model size also became closer to the true model size 7 as n grew. The SGJS tended to select larger groups than smaller groups under our simulation settings. The adjustment for the group size would be an interesting future research problem and may improve the performance of the SGJS.

Acknowledgments

This work was supported in part by Institutional Research Grant #14-247-29 from the American Cancer Society and the MCW Cancer Center, and the US National Cancer Institute (U24CA076518). The authors would like to thank the two reviewers for their helpful comments.

References

- Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B*. 2008; 70:849–911.
- Huang J, Li L, Liu Y, Zhao X. Group selection in the Cox model with a diverging number of covariates. *Statistica Sinica*. 2014; 24:1787–1810.
- Huang J, Sun T, Ying Z, Yu Y, Zhang CH. Oracle inequalities for the lasso in the Cox model. *Annals of Statistics*. 2013; 41:1142–1165. [PubMed: 24086091]
- Spiekerman CF, Lin DY. Marginal regression models for multivariate failure time data. *Journal of the American Statistical Association*. 1998; 93:1164–1175.
- Wang S, Nan B, Zhou N, Zhu J. Hierarchically penalized Cox regression with grouped variables. *Biometrika*. 2009; 96:307–322.
- Wu TT, Wang S. Doubly regularized Cox regression for high-dimensional survival data with group structures. *Statistics and Its Interface*. 2013; 6:175–186.
- Yang G, Yu Y, Li R, Buu A. Feature screening in ultrahigh dimensional Cox’s model. *Statistica Sinica*. 2016; 26:881–901. [PubMed: 27418749]
- Zhao S, Li Y. Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*. 2012; 105:397–411. [PubMed: 22408278]

Appendix

For $r = 0, 1, 2$, we define

$$\begin{aligned}
\mathbf{s}^{(r)}(\boldsymbol{\beta}, t) &= \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^L Y_{ij}(t) \mathbf{Z}_{ij}(t) \otimes e^{r \boldsymbol{\beta}^T \mathbf{Z}_{ij}(t)}, \mathbf{s}^{(r)}(\boldsymbol{\beta}, t) = \lim_{m \rightarrow \infty} \mathbf{s}^{(r)}(\boldsymbol{\beta}, t), \\
\mathbf{H}(\boldsymbol{\beta}, t) &= \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)}, \mathbf{V}(\boldsymbol{\beta}, t) = \frac{\mathbf{S}^{(2)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} - \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t) \otimes 2}{s^{(0)}(\boldsymbol{\beta}, t)^2}, \\
\mathbf{e}(\boldsymbol{\beta}, t) &= \frac{s^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)}, \mathbf{v}(\boldsymbol{\beta}, t) = \frac{s^{(2)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} - \frac{s^{(1)}(\boldsymbol{\beta}, t) \otimes 2}{s^{(0)}(\boldsymbol{\beta}, t)^2},
\end{aligned}$$

where $\mathbf{a}^{\otimes 0} = \mathbf{1}$, $\mathbf{a}^{\otimes 1} = \mathbf{a}$, and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$. The parameters and the true parameters associated with G are represented by $\boldsymbol{\beta}^G = (\boldsymbol{\beta}_{A_k} : k \in G)^T$ and $\boldsymbol{\beta}_0^G = (\boldsymbol{\beta}_{A_k, 0} : k \in G)^T$, respectively. Let

$F_+^q = \{G : G_0 \subset G, |G| \leq q\}$ and $F_-^q = \{G : G_0 \not\subset G, |G| \leq q\}$ be the collections of the over-fitted models and the under-fitted models, respectively. Let $p_G = \sum_{k \in G} |A_k|$ be the total number of individual variables within the groups belong to a set G . The true number of non-zero groups is assumed to be $|G_0| = g_0$. We assume

- (C1) $\int_0^\tau \lambda_0(u) du < \infty$ and $|Z_{ijl}(0)| + \int_0^\tau |dZ_{ijl}(u)| \leq M$ for all i, j, l , where M is a positive constant;
- (C2) There exists a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}_0$ and $\sup_{t \in [0, \tau], \boldsymbol{\beta} \in \mathcal{B}} \|\mathbf{S}^{(r)}(\boldsymbol{\beta}, t) - \mathbf{s}^{(r)}(\boldsymbol{\beta}, t)\|$ converges in probability to 0 for $r = 0, 1, 2$;
- (C3) For $r = 0, 1, 2$, $\mathbf{s}^{(r)}(\boldsymbol{\beta}, t)$'s are continuous in $\boldsymbol{\beta} \in \mathcal{B}$ uniformly in $t \in [0, \tau]$ and are bounded on $\mathcal{B} \times [0, \tau]$. $\mathbf{s}^{(0)}(\boldsymbol{\beta}, t)$ is bounded away from 0 on $\mathcal{B} \times [0, \tau]$.
- (C4) Let $\boldsymbol{\Omega} = \int_0^\tau v(\boldsymbol{\beta}_0^G, u) s^{(0)}(\boldsymbol{\beta}_0^G, u) \lambda_0(u) du$, where $\boldsymbol{\Omega}$ is positive definite;
- (C5) There exists $c_1, c_2 > 0$ and $d_1, d_2 \geq 0$ such that $d_1 + d_2 < 1/2$, $\min_{k \in G_0} \|\boldsymbol{\beta}_{A_k, 0}\| \geq c_1 m^{-d_1}$ and $g_0 < q - p_G < c_2 m^{d_2}$ for any G with $|G| = q$;
- (C6) $\log(KL) = O(m^h)$ for some $0 < h < 1 - 2(d_1 + d_2)$;
- (C7) There exists $c_3 > 0$ and $\delta > 0$ such that for sufficiently large m , $\lambda_{\min}\{-m^{-1} \ell''(\boldsymbol{\beta}^G)\} \geq c_3$ for $\boldsymbol{\beta}^G \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta}^G - \boldsymbol{\beta}_0^G\| \leq \delta, G \in F_+^{2q}\}$, where $\lambda_{\min}(\mathbf{B})$ indicates the smallest eigenvalue of a matrix $\mathbf{B}$;
- (C8) There exists $c_4, c_5 > 0$ such that $\max_{ijl} |Z_{ijl}| < c_4$ and $\max_{ij} |\mathbf{Z}_{ij}^T \boldsymbol{\beta}_0| < c_5$.

Next, we prove Theorem 2. We are motivated by the steps of the proof of Theorem 2 from Yang et al. (2016). However, clustered failure times and group structure allowed in the SGJS method required a detailed proof. Let $\hat{\boldsymbol{\beta}}^G$ be the maximum partial likelihood estimate of $\boldsymbol{\beta}^G = (\boldsymbol{\beta}_{A_k}, k \in G)^T$ for a subset $G \in \{1, \dots, K\}$. We need to show $P(\hat{G} \in F_+^q) \rightarrow 1$, that is,

$$P \left\{ \max_{G \in F_-^q} \ell(\hat{\boldsymbol{\beta}}^G) \geq \min_{G \in F_+^q} \ell(\hat{\boldsymbol{\beta}}^G) \right\} \rightarrow 0 \text{ as } m \rightarrow \infty.$$

For any $G \in F_-^q$, let $G_1 = G \cup G_0 \in F_+^{2q}$. Consider β^{G_1} such that $\|\beta^{G_1} - \beta_0^{G_1}\| = c_1 m^{-d_1}$ for some $c_1 > 0$ and $d_1 > 0$. For sufficiently large m , similarly to Yang et al. (2016), we can show

$$\ell(\beta^{G_1}) - \ell(\beta_0^{G_1}) \leq c_1 m^{-d_1} \|\ell'(\beta_0^{G_1})\| - c_1^2 (c_3/2)m^{1-2d_1}.$$

Then,

$$\begin{aligned} P\{\ell(\beta^{G_1}) - \ell(\beta_0^{G_1}) \geq 0\} &\leq P\{\|\ell'(\beta_0^{G_1})\| \geq c_1 c_3 m^{1-d_1/2}\} \\ &= P\left\{\sum_{a \in G_1} \sum_{b \in A_a} \ell'_b(\beta_0^{G_1})^2 \geq (c_1 c_3/2)^2 m^{2-2d_1}\right\} \\ &\leq \sum_{a \in G_1} \sum_{b \in A_a} P\left\{\ell'_b(\beta_0^{G_1})^2 \geq (p_{G_1})^{-1} (c_1 c_3/2)^2 m^{2-2d_1}\right\}. \end{aligned}$$

We first consider $P\{\ell'_b(\beta_0^{G_1}) \geq (p_{G_1})^{-1/2} (c_1 c_3/2)m^{1-d_1}\}$. Because $p_{G_1} < 2c_2 m^{d_2}$ from Condition 5, we have

$$P\{\ell'_b(\beta_0^{G_1}) \geq (p_{G_1})^{-1/2} (c_1 c_3/2)m^{1-d_1}\} \leq P\{\ell'_b(\beta_0^{G_1}) \geq c m^{1-d_1-d_2/2}\}, \quad (6)$$

where $c = (2c_2)^{-1/2} (c_1 c_3/2)$. Define a martingale expression

$$M_{ij}(t) = N_{ij}(t) - \int_0^t Y_{ij}(s) \exp\{\beta_0^T \mathbf{Z}_{ij}(s)\} \lambda_0(s) ds.$$

Because of the interclass dependence, $M_{ij}(t)$ is not martingale with respect to $\mathcal{F}_t = \bigvee_{i=1}^n \bigvee_{j=1}^L \mathcal{F}_{t,ij}$, where $\mathcal{F}_{t,ij} = \sigma\{N_{ij}(s), Y_{ij}(s), \mathbf{Z}_{ij}(s); 0 \leq s \leq t\}$. However, for fixed j , $M_{ij}(t)$ is martingale with respect to $\mathcal{F}_{t,j} = \bigvee_{i=1}^n \mathcal{F}_{t,ij}$ because $\{T_{ij}, C_{ij}, \mathbf{Z}_{ij}\}$'s are independent and identically distributed given j . Define C_0 such that $C_0^2 m$ is less than or equal to the number of events. By Lemma 3.3 of Huang et al. (2013), for given j we have

$$P\left[\left|\sum_{i=1}^m \int_0^\tau \{Z_{ija} - H_a(\beta, s)\} dM_{ij}(s)\right| > m C_0 x\right] \leq 2 \exp(-mx^2/2),$$

where $H_a(\boldsymbol{\beta}, s)$ is the a th entry of $H(\boldsymbol{\beta}, s)$ for $a = 1, \dots, p$. Observe

$$\ell'_a(\boldsymbol{\beta}_0^{G_1}) = \sum_{j=1}^L \sum_{i=1}^m \int_0^\tau \{Z_{ija} - H_a(\boldsymbol{\beta}_0^{G_1}, s)\} dM_{ij}(s).$$

Then,

$$P\{\ell'_a(\boldsymbol{\beta}_0^{G_1}) \geq cm^{1-d_1-d_2/2}\} \leq L \exp(-m^{1-2d_1-d_2/2}). \quad (7)$$

Similarly, we can show

$$P\{\ell'_a(\boldsymbol{\beta}_0^{G_1}) \leq -cm^{1-d_1-d_2/2}\} \leq L \exp(-m^{1-2d_1-d_2/2}). \quad (8)$$

By (7) and (8), we have $P\{\ell(\boldsymbol{\beta}^{G_1}) - \ell(\boldsymbol{\beta}_0^{G_1}) \geq 0\} \leq 2p_{G_1} L \exp(-m^{1-2d_1-d_2/2})$. Define c_0 such that $\log(KL) \leq c_0 m^h$. By the Bonferroni inequality and Conditions 5 and 6, as $m \rightarrow \infty$

$$\begin{aligned} P\left\{\max_{G \in F_-^q} \ell(\boldsymbol{\beta}^{G_1}) \geq \ell(\boldsymbol{\beta}_0^{G_1})\right\} &\leq \sum_{G \in F_-^q} P\{\ell(\boldsymbol{\beta}^{G_1}) \geq \ell(\boldsymbol{\beta}_0^{G_1})\} \\ &\leq 2p_{G_1} L K^q \exp(-m^{1-2d_1-d_2/2}) = 2 \exp(\log p_{G_1} + \log L + q \log K - m^{1-2d_1-d_2/2}) \\ &\leq 2 \exp(\log p_{G_1} + q \log(KL) - m^{1-2d_1-d_2/2}) \\ &\leq 2 \exp\{\log(2c_2) + d_2 \log m + c_0 c_2 m^{d_2} m^h - m^{1-2d_1-d_2/2}\} \\ &= 4c_2 \exp\{d_2 \log m + c_0 c_2 m^{d_2} m^h - m^{1-2d_1-d_2/2}\} = o(1). \end{aligned}$$

(9)

In other words, $P\left\{\max_{G \in F_-^q} \ell(\boldsymbol{\beta}^{G_1}) < \ell(\boldsymbol{\beta}_0^{G_1})\right\} \rightarrow 1$ as $n \rightarrow \infty$. In addition, the log-partial likelihood function $\ell(\boldsymbol{\beta}^{G_1})$ is concave in $\boldsymbol{\beta}^{G_1}$, see Huang et al. (2013). Thus, for any $\boldsymbol{\beta}_*^{G_1}$ such

that $\|\beta_*^{G_1} - \beta_0^{G_1}\| > c_1 m^{-d_1}$, $\max_{G \in F_-^q} \ell(\beta_*^{G_1}) \leq \max_{G \in F_-^q} \ell(\beta^{G_1}) < \ell(\beta_0^{G_1})$ in probability,

that is, (9) holds for any $\beta_*^{G_1}$. For $G \in F_-^q$, let $\tilde{\beta}^{G_1}$ be $\hat{\beta}^G$ augmented with zeros corresponding to the elements in $G_1 \setminus G_0$. Then, $\ell(\tilde{\beta}^{G_1}) = \ell(\hat{\beta}_G)$ for $G \in F_-^q$. By the same argument of Yang et al. (2016), we can show $\|\tilde{\beta}^{G_1} - \beta_0^{G_1}\| \geq c_1 m^{d_1}$. By the definition of $\hat{\beta}^G$, for any $G \in F_+^q$, $\ell(\hat{\beta}^G) \geq \ell(\beta_0^G)$. In addition, $\ell(\beta_0^{G_1}) = \ell(\beta_0^G)$ for $G \in F_+^q$ because $G_0 \subset G_1$. Therefore,

$$P \left\{ \max_{G \in F_-^q} \ell(\hat{\beta}^G) \geq \min_{G \in F_+^q} \ell(\hat{\beta}^G) \right\} \leq P \left\{ \max_{G \in F_-^q} \ell(\tilde{\beta}^{G_1}) \geq \ell(\beta_0^{G_1}) \right\} = o(1),$$

which completes the proof of Theorem 2.

Table 1

Simulation results with overlapping groups

		Independent data ($\psi = 1$)			Clustered data ($\psi = 0.25$)		
n	Method	P_1	P_2	Model Size	P_1	P_2	Model Size
200	SGJS-M1*	0.925	0.615	7.235	0.745	0.605	6.670
	SGJS-M2*	0.905	0.605	7.150	0.725	0.560	6.575
	SGJS-M1	0.105	0.075	4.050	0.030	0.025	2.925
	SGJS-M2	0.065	0.045	3.325	0.020	0.020	2.370
500	PSIS	0	0	2.455	0	0	2.190
	SGJS-M1*	1	0.860	7.115	1	0.935	7.070
	SGJS-M2*	1	0.880	7.125	1	0.955	7.045
	SGJS-M1	0.315	0.275	5.310	0.175	0.155	4.620
	SGJS-M2	0.245	0.220	4.740	0.140	0.120	3.950
	PSIS	0.005	0.005	3.375	0.005	0.005	3.420