Variable Screening for Survival Data in the Presence of Heterogeneous Censoring

Jinfeng Xu,[1] Waikeung Li[1] and Zhiliang Ying[2]

[1]Department of Statistics and Actuarial Science, University of Hong Kong

[2]Department of Statistics, Columbia University

Abstract: Variable screening for censored survival data is most challenging when both survival and censoring times are correlated with a ultrahigh-dimensional vector of covariates. Existing approaches to handling censoring often make use of inverse probability weighting by assuming independent censoring with both survival time and covariates. This is a convenient but rather restrictive assumption which may be unmet in real applications, especially when the censoring mechanism is complex and the number of covariates is large. To accommodate heterogeneous (covariate-dependent) censoring that is often present in high-dimensional survival data, we propose a Gehan-type rank screening method to select features that are relevant to the survival time. The method is invariant to monotone transformations of the response and of the predictors, and works robustly for a general class of survival models. We establish the sure screening property of the proposed methodology. Simulation studies and a lymphoma data analysis demonstrate its favorable performance and practical utility.

Key words: Gehan-type rank statistics; High-dimensional survival data; Heterogeneous censoring; Sure screening property; U-statistic; Variable screening.

1 Introduction

Variable screening has been proven to be a useful tool in ultrahigh-dimension data analysis. With many more features than observations, a screening procedure can first help to filter out a majority of noise variables. The analysis with remaining variables would then become much easier and more effective. In their pioneering work, Fan & Lv (2008) proposed a Pearson correlation-based independence screening for ultrahigh-dimensional linear models and established its desirable sure screening property. Such a screening procedure has been further extended to generalized linear models (Fan & Song, 2010) and additive models (Fan et al., 2011).

In biomedical applications, the response variable of interest is often a possibly right-censored survival outcome. For such censored data, the approach of Fan & Lv (2008) may not be applied directly, especially for model-free approaches. An important recent work of variable screening for ultrahigh-dimensional variable selection with survival outcome was done by Song et al. (2014). Assuming that the censoring time independent of both failure and covariates, they proposed a novel model-free censored rank screening procedure based on an inverse probability-of-censoring weighted Kendall's τ. Under suitable regularity conditions, they established its sure screening property. For other approaches to ultrahigh-dimensional feature screening with survival outcomes data, we refer to Fan et al. (2010) and Zhao & Li (2012) for the Cox proportional hazards model-based methods, to Gorst-Rasmussen & Scheike (2013) for the additive hazards model-based method and to He et al. (2013) and Wu & Yin (2015) for quantile-based methods. However, all these methods are developed for the situation where the censoring time does not depend on the covariates, a rather restrictive assumption which may be unmet in practice as demonstrated in our real data example in Section 6.

In this paper, we propose a Gehan-type rank screening approach for variable screening with survival outcome data. Compared with existing methods, it has several advantages. First, it naturally incorporates censoring without the need to know its conditional survival function, thus allowing heterogeneous censoring and bypassing the difficulty of the local Kaplan-Meier estimation. Second, the method is invariant under monotone transformations of the response and predictors and does not require any finite-moment assumption for the covariates. Third, the proposed method is a model-free approach and works robustly in a

general class of survival models. Numerical studies demonstrate its favorable finite-sample performance especially when the censoring is heterogeneous or the censoring rate is high.

The rest of the paper is organized as follows. The next section introduces the main method. Section 3 examines the screening properties of the proposed method under certain conditions in a general class of transformation models. Section 4 establishes a general result about the sure screening property for the proposed Gehan-type rank screening. Simulation results are reported in Section 5, while in Section 6, the method is applied to a real data set. Section 7 gives some concluding remarks. All technical proofs are given in the Appendix.

2 Gehan-type Rank Screening

Let Y denote the survival time, C the censoring time, and $Z = (Z_1, ..., Z_{p_n})^T$ the p_n-dimensional vector of covariates. We observe an independent and identically distributed sample $\{X_i, \Delta_i, (Z_{i1}, ..., Z_{ip_n})^T : i = 1, ..., n\}$, where $X_i = \min(Y_i, C_i)$, $\Delta_i = I(Y_i \leq C_i)$, and $I(\cdot)$ denotes the indicator function. We assume that Y and C are conditionally independent given Z. Such conditional independence is commonly assumed in regression analysis of survival data. The dimensionality of p_n is allowed to increase very rapidly with the sample size as we are considering variable screening procedures in an ultrahigh-dimensional setting.

Following Fan & Lv (2008) and Song et al. (2014), let $\mathcal{M}_*$ denote the index set of the active variables for the survival time Y:

$$\mathcal{M}_* = \{k : \text{pr}(Y > y|Z) \quad \text{depends functionally on} \quad Z_k, k = 1, ..., p_n\}.$$

Likewise, let $\mathcal{M}_*^C$ denote the index set of the active variables for the censoring time C:

$$\mathcal{M}_*^C = \{k : \text{pr}(C > y|Z) \quad \text{depends functionally on} \quad Z_k, k = 1, ..., p_n\}.$$

We consider variable screening procedures to estimate $\mathcal{M}_*$ while allowing $\mathcal{M}_*^C$ to be arbitrary and completely unspecified. The existing approaches require that $\mathcal{M}_*^C$ is an empty or known singleton set, a convenient but rather stringent assumption. Intuitively speaking, when the censoring time does not depend on the covariates, the active set for Y is the same as those for X or Δ and variable screening can be conveniently done based on either of them. Variable screening is most challenging when the survival-covariate dependence and the censoring-covariate dependence are intertwined with each other. This is the setting to be investigated in this paper.

The censored rank screening (Song et al., 2014) uses Kendall's τ (Kendall, 1962) to select active variables, where marginal Kendall rank correlation coefficient for the kth variable, denoted by τ_k, is defined as $P(Y_1 < Y_2, Z_{1k} < Z_{2k}) - 1/4$. As some Ys are unobserved due to censoring, it adopts the inverse probability-of-censoring weighting technique to estimate τ_k by making the homogeneous censoring assumption, i.e. C is independent of Z and Y, that justifies the use of the Kaplan-Meier estimation of the survival function of the censoring time.

We propose a different approach to screening covariates that can bypass the inverse probability weighting, thereby avoiding the restrictive homogeneous censoring assumption. Our method is motivated by the Gehan extension of the Wilcoxon-Mann-Whitney rank test statistic to cover censored data (Gehan, 1965)

$$\hat{W}_k = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \{\Delta_i I(X_i \leq X_j)(Z_{ik} - Z_{jk})\}. \tag{2.1}$$

The Gehan statistic may be viewed as an extension of the Mann-Whitney pairwise comparisons, by replacing all pairs with all comparable pairs. Unlike the rank statistic used in Song et al. (2014), it naturally incorporates the information of censoring without the need to know the conditional survival function of censoring. There is also a close connection between (1) and the class of rank-based estimating functions for censored survival data (Jin et al., 2003).

To allow possible nonlinear covariate-survival relationship and to make our method less sensitive to extreme values of Z_k, we dichotomize Z_k to $Z_k(t) = I(Z_k \leq t)$ according to a threshold value t. The difference in survival between groups with $Z_k \leq t$ and $Z_k > t$ is assessed by the following Gehan test statistic:

$$\hat{S}_k(t) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} [\Delta_i I(X_i \leq X_j)\{Z_{ik}(t) - Z_{jk}(t)\}].$$

When Y and Z_k are independent, $\hat{S}_k(t)$ is expected to fluctuate around zero as it has a zero mean. In fact, it is not difficult to show that it is of order $O_p(n^{-1/2})$ as $n^{1/2}\hat{S}_k(t)$ converges weakly to a zero-mean Gaussian process. Note that $\hat{S}_k(t)$ captures the overall survival difference when the subjects are dichotomized into two groups according to the kth covariate Z_k into $(Z_k \leq t)$ and $(Z_k > t)$ groups. Here t can be any value in the range of the kth covariate Z_k. From data, we observe n values of Z_k: $Z_{\ell k}, \ell = 1, ..., n$. It is natural

to use $\hat{S}_k(Z_{\ell k})$ to capture the discrepancy when Z_k is dichotomized by $Z_{\ell k}$, for $\ell = 1, ..., n$. Therefore, to capture the overall covariate-survival association across the whole spectrum of Z_k, we propose the following Cramér-von Mises-type marginal utility for the kth predictor:

$$\hat{D}_k = n^{-1} \sum_{\ell=1}^{n} \hat{S}_k^2(Z_{\ell k}).$$

It is straightforward to see that the above defined marginal utility $\hat{D}_k$ is invariant under monotone transformations of the response as well as the predictor. Alternatively, we can write $\hat{D}_k = \int \hat{S}_k^2(t) d\hat{L}_k(t)$, where $\hat{L}_k$ denotes the empirical distribution of Z_k. Let $\tilde{S}_k(t) = n\hat{S}_k(t)/(n-1)$ and $S_k(t) = E\{\tilde{S}_k(t)\}$. It is easy to see that

$$S_k(t) = E[\Delta_1 I(X_1 \leq X_2)\{Z_{1k}(t) - Z_{2k}(t)\}].$$

Define $D_k = E\{S_k^2(Z_k)\}$, which is clearly the limit of $\hat{D}_k$ as $n \to \infty$. When Y and Z_k are independent, $S_k(t) = 0$ for all t, implying $D_k = 0$. On the other hand, if $S_k(t) \neq 0$ for some t in the support of Z_k, then $D_k \neq 0$. It follows that the departure of $\hat{S}_k$ from zero is related to the dependency of Z_k on Y and that the value of $\hat{D}_k$ indicates the level of importance of the corresponding predictor variable in its relation to the survival time. Therefore, we propose to select the set

$$\hat{\mathcal{M}}_{\gamma_n} = \{k : \hat{D}_k \geq \gamma_n, k = 1, ..., p_n\},$$

as an approximation to $\mathcal{M}_*$, where γ_n is a pre-specified threshold that depends on n. In the next section, we investigate screening properties of the proposed method in a general class of transformation models.

3 Gehan Screening In A General Class of Transformation Models

For a variable screening procedure to exhibit desirable sure screening properties, the magnitude of the screening utility for the active set must be of a non-vanishing order to be distinguishable from those for the inactive set as stated in the following regularity condition.

Condition A. For some $0 \leq \alpha < 1/2$ and $c_0 > 0$, $\min_{k \in \mathcal{M}_*} D_k \geq c_0 n^{-\alpha}$.

To quantify the magnitude of the Gehan screening utility, consider a general class of transformation models given by

$$H(Y) = m(Z_1, ..., Z_{p_n}) + \epsilon, \tag{3.2}$$

where H is an increasing transformation function, $m(\cdot)$ is unspecified and ϵ is independent of Z and has a density function. Clearly, $\mathcal{M}_*$ is the smallest subset $\mathcal{M}$ such that $m(\cdot)$ is a function only of covariates in $\mathcal{M}$. Let $\mathcal{M}_0 = \{1, ..., p_n\}$ and $Z_{\mathcal{B}} = \{Z_j : j \in \mathcal{B}\}$ for a set $\mathcal{B} \subset \mathcal{M}_0$. Write $Z_{-k} = Z_{\mathcal{M}_0/\{k\}}$. Define $m_k(Z_k) = E\{m(Z)|Z_k\}$, and $m_{-k}(Z) = m(Z) - m_k(Z_k)$. Without loss of generality, assume $E(\epsilon) = 0$ and $E\{m(Z)\} = 0$. Otherwise, they can be absorbed into H since H is an unspecified monotone function. Likewise, $E\{m_k(Z_k)\} = 0$ and $E\{m_{-k}(Z)\} = 0$. As in Song et al. (2014), assume that $m_k(x)$ is monotone in x for each k in $\mathcal{M}_*$. Model (3.2) includes many popular regression models in survival analysis such as the linear transformation model (Clayton & Cuzick, 1985) and the accelerated failure time model (Kablfleisch & Prentice, 2002).

The following conditions are sufficient to ensure that Condition A holds in model (3.2).

Condition 1. For any $k \in \mathcal{M}_*$, there exists $\tilde{Z}_{-k}$ independent of Z_k and $m(Z) = m_k(Z_k) + \tilde{m}_{-k}(\tilde{Z}_{-k})$, for some $\tilde{m}_{-k}(\cdot)$.

Condition 2. There exists a positive constant η such that the variance of $H(Y) - m_k(Z_k)$ is uniformly bounded above by η^2, for all $k \in \mathcal{M}_*$.

Condition 3. For any $k \in \mathcal{M}_*$ and $\tilde{Z}_{-k}$ given in Condition 1, the conditional survival distribution of C given Z, denoted by $\tilde{G}(t|Z)$, satisfies $-\log \tilde{G}(t|Z) = a_k(t, Z_k) + b_k(t, \tilde{Z}_{-k})$ for some functions $a_k(\cdot)$ and $b_k(\cdot)$.

Condition 4. There exists a survival function $G_0(\cdot)$ such that for any $x \in \mathcal{R}$,

$$\inf_z \text{pr}(H(C) \geq x|Z = z) \geq G_0(x).$$

Condition 5. For G_0 given in Condition 4, there exists $\tau > 0$ such that $E[G_0\{H(Y)\}] \geq \tau$.

Condition 6. Let Q_k be the distribution function of $m_k(Z_k)$ and f be the density function of ϵ. Then, for some $0 \leq \alpha < 1/2$ and $c > 0$, Q_k and f are Lipschitz with constants ξ_k and ψ such that $\max_{k \in \mathcal{M}_*} \xi_k \leq (216c\eta)^{-1}\psi^{-1/2}\tau^{7/2}n^{\alpha/2}$, where η and τ are given in Conditions 2 and 5.

Condition 1 means that each variable in the active set influences the regression function through its marginal projection. Note that in the general transformation model, the effect of Z is assumed to be $m(Z)$. Condition 1 is hence similar to the additivity assumption in the additive regression model and can be checked by fitting two separate nonparametric regression models with and without additivity separately and then examine the discrepancy between two fitted models. Conditions 2 and 6 control the magnitude of detectable signals. Condition 2 assumes an upper bound of the variance of the difference between $H(T)$ and $m_k(Z_k)$. This is generally true when the underlying functions H and m_k satisfy certain mild regularity conditions. Condition 3 says that there is no interaction between variables in the conditional cumulative hazard function of censoring given covariates. This condition can be similarly checked by fitting two nonparametric models with and without interaction assumptions and then examine the discrepancy between two fitted models. This assumption will be restrictive in some applications when the interaction is present but substantially relaxes the homogeneous censoring assumption and allows heterogeneous censoring to some extent. Condition 4 assumes a uniform lower bound for the conditional survival function of $H(C)$ given covariates Z. Condition 5 states that this bound is not degenerate. In the homogeneous censoring case, Condition 4 is trivially satisfied with G_0 taken as the common survival function of $H(C)$; furthermore, Condition 5 is equivalent to the following condition:

Condition 5'. There exists $\tau > 0$ such that $\text{pr}(Y \leq C) \geq \tau$.

The above condition ensures that the proportion of uncensored observations does not decay to zero. This is quite reasonable because when the proportion of uncensored observations tends to zero, it is unlikely for us to obtain useful information regarding relevant predictors even if the other model assumptions are appropriate. To get a better understanding of Conditions 1-6, in the following, we also illustrate it in more detail in the model (3.3), as

we will see, often in a more specific model, the conditions 1 to 6 will become much less restrictive and easier to check than in the general setting.

Propostion 1. Conditions 1-6 imply Condition A.

In the aforementioned conditions, Conditions 3-5 are for the censoring mechanism, while Conditions 1, 2 and 6 are for the underlying model of the survival time. To better illustrate them, we next consider the variable-transformation linear normal model (Mai & Zhou, 2015) given by

$$H(Y) = \mathbf{g}^T(\mathbf{Z})\beta + \epsilon, \tag{3.3}$$

where $\mathbf{g} = (g_1, ..., g_{p_n})^T$ and $H, g_1, ..., g_{p_n}$ are strictly monotone univariate transformations. It is also assumed that $\mathbf{g}(\mathbf{Z}) \sim N(0, \mathbf{\Sigma})$ with $\Sigma_{jj} = 1$ for $j = 1, ..., p_n$, and $\epsilon \sim N(0, \sigma^2)$ is independent of Z. This model has close connections to many transformation models in the literature; for example, see Breiman & Friedman (1985) and He & Shen (1997). Without loss of generality, assume that $\beta = (\beta_{\mathcal{D}}, 0)$, where $\mathcal{D} = \{1, ..., s_n\} = \mathcal{M}_* \subset \mathcal{M}_0$.

Propostion 2. For model (3.3), we have the following results.
(i) Condition 1 always holds.
(ii) If $\beta_{\mathcal{D}}^T \Sigma_{\mathcal{D}\mathcal{D}} \beta_{\mathcal{D}} + \sigma^2 \leq \eta^2$, then Condition 2 holds.
(iii) If $\min_{k \in \mathcal{M}_*} |\omega_k| \geq 108 c\eta\tau^{-7/2}\pi^{-3/4}\sigma^{-1}n^{-\alpha/2}$, where $\omega = \mathbf{\Sigma}\beta$, then Condition 6 holds.

Propositions 1 and 2 suggest that under mild conditions, Condition A holds and the magnitude of the Gehan screening utility for the active set is distinguishable from those for the inactive set. In the next section, we show that the proposed method enjoys the sure screening property whenever Condition A holds.

4 Sure Screening Property

We present the following theorem establishing the sure screening property of the Gehan screening procedure.

Theorem 1. (i) For any $0 \leq \alpha < 1$ and $c > 0$, if $n \geq (20/c)^{1/(1-\alpha)} + 10$, then

$$\text{pr}(\max_{1 \leq k \leq p_n} |\hat{D}_k - D_k| > 2cn^{-\alpha}) \leq 2p_n e^{-c_1 n^{1-2\alpha}},$$

where $c_1 = c^2/20$.

(ii) Suppose that Condition A holds and $\gamma_n \leq c_0 n^{-\alpha}/2$. Then, for all $n \geq (80/c_0)^{1/(1-\alpha)}+10$,

$$\text{pr}(\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n}) \geq 1 - 2s_n e^{-c_2 n^{1-2\alpha}},$$

where $c_2 = c_0^2/320$ and s_n is the cardinality of $\mathcal{M}_*$.

Theorem 1 implies that our screening procedure can handle nonpolynomial dimensionality of order $\log p_n = o(n^{1-2\alpha})$ with $\alpha \in [0, 1/2)$. No tail probability conditions for covariates are needed due to the dichotomization of covariates in constructing the screening utility. It is easy to see that this Gehan rank screening method is invariant under monotone transformations of the response and covariates. It enjoys the desirable sure screening property as long as Condition A holds, without the need to further specify a model structure. The results in Section 3, however, from model-based perspectives, illustrate that the sure screening property indeed holds with mild regularity conditions in a general class of survival models.

In practice, it is important to specify the threshold values γ_n in screening procedures to select the size of the important set. As in Fan & Song (2010) and Song et al. (2014), we show that the size of set $\hat{\mathcal{M}}$ can be controlled for the variable-transformation linear normal model. Let $\alpha \in [0, 1/2)$.

Theorem 2. Consider model (3.3) and let $\omega = \mathbf{\Sigma}\beta$. Suppose that Conditions 3-5 hold, $\beta_{\mathcal{D}}^T \Sigma_{\mathcal{D}\mathcal{D}} \beta_{\mathcal{D}} + \sigma^2 \leq \eta^2$, and $\min_{k \in \mathcal{M}_*} |\omega_k| \geq 108c\eta\tau^{-7/2}\pi^{-3/4}\sigma^{-1}n^{-\alpha/2}$. Then, for $\gamma_n = 2c_0 n^{-\alpha}$, there exists a constant c_4 such that for $n \geq (40/c_0)^{1/(1-\alpha)} + 10$,

$$\text{pr}\{|\hat{\mathcal{M}}_{\gamma_n}| \leq c_4 n^{\alpha} \lambda_{\max}(\Sigma)\} \geq 1 - 2p_n e^{-c_3 n^{1-2\alpha}},$$

where $c_3 = c_0^2/40$, $\lambda_{\max}(\Sigma)$ is the largest eigenvalue of Σ, and c_0 is given in Condition A.

The above results shows that if the largest eigenvalue of Σ is of polynomial order, then the model after screening is also of polynomial size with probability tending to 1, indicating that the size of the selected set can be effectively controlled.

5 Simulation Studies

To implement the Gehan screening method for its practical use, we propose the following procedure. Data consist of $\{X_i, \Delta_i, (Z_{i1}, ..., Z_{ip_n})^T : i = 1, ..., n\}$.

The Gehan Screening procedure:

Step 1: For each $k = 1, ..., p_n$, calculate the marginal Gehan score $\hat{D}_k$.

Step 2: Rank $\hat{D}_k, k = 1, ..., p_n$ from the largest to the smallest:

$\hat{D}_{k_1} > \hat{D}_{k_2} > ... > \hat{D}_{k_{p_n}}$.

Step 3: If the size of the model is chosen to be m, then the set of predictors are $\{Z_{k_1}, ..., Z_{k_m}\}$.

Step 4: In practice, for parsimony and by the sparsity assumption, we let $m \leq n/\log n$.

For prediction purpose, the size of the final model is chosen by using five-fold cross validation to maximize the number of concordant pairs. Let $Z_{(m)} = (Z_{k_1}, ..., Z_{k_m})^T$. In the bth random split of data, let $\hat{\beta}_m^b$ be the estimated regression coefficients from the model relating T to Z by $H(T)- = \beta^T Z + \epsilon$ with the training data. The validated prediction accuracy is evaluated by the number of concordant pairs:

$$C(m) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i I(X_i < X_j) I(Z_{(m),i}^T \hat{\beta}_m^b > Z_{(m),j}^T \hat{\beta}_m^b),$$

where B is the number of splits in the five fold cross validation. The size of the final model $\hat{m}$ is chosen to maximize $C(m)$ for $m = 1, ..., \lfloor n/\log n \rfloor$. Note that using thresholding constants γ_n or the size of the final model m is equivalent once the marginal Gehan score is obtained and ranked. Furthermore, instead of using model (3.2) for estimation, for simplicity, we use linear transformation model because general nonparametric transformation model is difficult to estimate when there are multiple predictors.

We conduct simulations to investigate the performance of the proposed Gehan rank independence screening procedure. For comparison, we consider three alternative methods: feature aberration at survival times screening (Gorst-Rasmussen & Scheike, 2013), partial likelihood ratio screening, and censored rank independence screening (Song et al., 2014). For partial likelihood ratio screening, we fit a marginal Cox model for each covariate and the screening is based on the marginal partial likelihood ratio test statistic.

Example 5.1 We consider the following model adapted from Song et al. (2014):

$$H(Y) = -\beta^T Z + \sigma\epsilon,$$

where the ultrahigh-dimensional covariates $Z = (Z_1, ..., Z_{p_n})^T$ follow a multivariate normal distribution with mean zero and correlation matrix $\Sigma = (0.8^{|i-j|})(i, j = 1, ..., p_n)$, $H(t) = \log\{0.5(e^{2t} - 1)\}$, $\beta = (1.8_5^T, 0_5^T, -1.8_5^T, 0_{p_n-15}^T)^T$ and $\sigma = 0.5$. We considered the sample sizes $n = 100$ and 200, and set the number of covariate p_n to 2000. Three error distributions were considered: the standard extreme value distribution, which corresponds to a proportional hazards model; the standard logistic distribution, which corresponds to a proportional odds model; and the standard normal distribution, which corresponds to a normal transformation model. The censoring time was generated from a uniform distribution on $[0, \theta e^{\kappa^T Z}]$, where θ was chosen to yield a censoring rate of 20% or 40% and κ was set to be 0_{p_n} or $(3_3^T, 0_4^T, 3, -3, 0, -3_3^T, 0_{p_n-13}^T)^T$, corresponding to the scenarios for homogeneous censoring and heterogeneous censoring, respectively.

To assess the performance of the screening procedures, we first report the minimum model size which is the smallest number of covariates needed to include all the active predictors. The smaller the minimum model size a screening procedure has, the better it performs as it results in a more parsimonious model. We present the median and interquartile range of the minimum model size over 100 replications. Secondly, we calculate the proportion, out of the 100 replications, that all of the active predictors are selected for a given model size $\lfloor n/\log n \rfloor$, where $\lfloor x \rfloor$ denotes the integer part of x. We denote this proportion by $\mathcal{P}_{All}$. A screening procedure yielding $\mathcal{P}_{All}$ closer to 1 is considered to be more effective.

The simulation results are summarized in Tables 1 and 2. Based on the results, the performances of four procedures are affected differently by heterogeneous censoring or high censoring rate. The censored rank screening deteriorated greatly when the censoring rate is high while the feature aberration at survival times screening and partial likelihood ratio screening is much less effective when censoring is heterogeneous. In general, our proposed method performs robustly and substantially better than the other three procedures in terms of $\mathcal{P}_{All}$ and the minimum model size.

Example 5.2 To examine the performance of the screening procedures in the nonlinear scenario, we consider the following nonlinear transformation model,

$$H(Y) = -Z_1^2 - \cos(Z_2) - Z_6^3 - \sin(Z_7) - Z_8 + \sigma\epsilon.$$

The censoring time was generated from a uniform distribution on $[0, \theta e^{\kappa^T Z}]$, where κ was chosen to be 0_{p_n} or $(1, 0_6^T, 1, 1, 0_{p_n-9}^T)^T$, and the rest of the set-up is the same as in Example 1. The simulation results are summarized in Table 3. As in Example 1, the proposed method exhibits favorable performance over the three existing approaches in terms of both the minimum model size and the selection proportion, especially when censoring is heterogeneous or the censoring rate is high.

It is important to note that although we relax the assumption on the censoring mechanism, the screening test statistics $\hat{D}_k, k = 1, ..., p_n$ involve $\delta_i, i = 1, ..., n$. Hence, the censoring mechanism affects the performance of the proposed method. For example, as pointed out by an anonymous referee, if two breast cancer populations with the same set of relevant genes were observed, the information available to us will differ for both identifying relevant genes and making predictions as the censoring mechanism will play an important role. This dependence can be similarly compared with the sample size. If we have the same breast cancer population with two different sample sizes, the chance of identifying relevant genes and obtaining accurate prediction will be higher if we have a larger sample size. For a homogeneous censoring mechanism which does not depend on the covariates, this is easy to understand, the higher censoring rate will be equivalent to the smaller sample size. For the covariate-dependent censoring, the dependence pattern may not be that direct. To assess the impact of the censoring mechanism on the screening results, we report the results for the Gehan method for varying censoring mechanisms and censoring proportions. We let $\kappa = \gamma(1, 0_6^T, 1, 1, 0_{p_n-9}^T)^T$. We let $\gamma = 0, 0.3, 0.6$ and 1 to vary different censoring mechanisms. The results are summarized in Table 4. It can be seen that the performance of the proposed method is indeed affected by the sample size, the censoring mechanism and the censoring proportion.

6 Real Data Example

We apply the proposed screening method to the diffuse large B cell lymphoma (DLBCL) microarray data of Lenz, et al. (2008). The dataset contains the survival times of 181 patients treated with a combination chemotherapy with cyclophosphamide, doxorubicin, vincristine and prednisone (CHOP). The primary goal of the study is to identify genes that

are associated with the survival of lymphoma patients. The original data have 54675 probe sets or covariates. A pre-selection procedure was conducted to filter out the genes with lower variations if a sample variance for a gene was smaller than the 10th percentile for that gene. There are 3833 genes after the filtering process. The median survival time was 3.58 years. During the follow-up, 105 patients died of lymphoma, and the other 76 patients were censored, yielding a censoring rate of 41.9%.

The top 10 selected genes are reported in Table 5. There are 28 genes selected by at least one screening method. Out of the 28 genes, none were selected by all four screenings. Five genes, 1554413_s_at, 1569344_a_at, 229839_at, 231049_at, and 240898_at, were selected by only three screenings and the censored rank screening did not select any of them. Four genes were selected by only two screenings and the rest 19 genes were selected by only one screening. To check whether censoring is heterogeneous, we fit the Cox proportional hazards model relating the censoring time to the aforementioned five genes which were selected by at least three screenings under a given model size of 10. Three genes, 1554413_s_at, 1569344_a_at, and 229839_at, are statistically significant with p-values $0.043, 0.001$, and < 0.001, respectively. This indicates the presence of heterogeneous censoring and it is better to employ screening procedures which can accommodate this phenomenon. In Table 6, we present the selection results for seven genes, LOC283922/PDPR, MFAP4, UPB1, SCARA5, TMEM56, DNM1L, and JMJD1A, for a given model size of 9 or 34. Note that $n = 181$ and $\lfloor n/\log n \rfloor = 34$. Hence, 34 is the largest model size we consider. We also consider model size 9 since for the Gehan method, the model size with 9 achieves the maximum concordance based on the five-fold cross-validation and the linear transformation model. These biologically relevant genes have been reported to be associated with DLBCL (Wang & Wang, 2010). The minimum model sizes for including all seven genes are 139, 3767, 308, and 111, for the Gehan rank screening, censored rank screening, feature aberration at survival times screening, and partial likelihood ratio screening, respectively. The results show that the proposed Gehan-type rank screening and the Cox regression-based partial likelihood screening can more easily detect these genes, requiring model sizes smaller than the other two screenings. The results also show that the proposed Gehan-type rank screening behaves quite differently from Song et al. (2014), even though both are rank based methods.

7 Remarks

This paper proposes a variable screening for ultrahigh-dimensional survival time data. It is based on the well-known Gehan test statistic, which is an extension of the Mann-Whitney rank test statistic. The new method is easy to implement and requires only conditional independence between survival and censoring times given covariates, thus allowing heterogeneous censoring. Both simulation studies and a real data analysis show that the method performs quite well, even with moderate sample sizes and substantial censoring proportions. The real data analysis also shows that the proposed method can behave quite differently from Song et al. (2014), which is also a rank-based screening but uses inverse probability weighting.

An important aspect in classical regression analysis of survival data is the incorporation of time-dependent covariates. The proposed approach does not allow for time-dependent covariates. However, this may not be an issue of significance under the current setting of high-dimensional variable screening in which very few if any covariates are time-dependent. Furthermore, the main purpose of the screening is to narrow down the number of covariates to a manageable level.

The Gehan rank test statistic is only a member of the class of weighted log-rank test statistics (Kalbfleisch & Prentice, 2002). Suitable choice of the weight function can increase the power and efficiency of the corresponding test. In view of Jin et al. (2003), it may be possible to extend the proposed Gehan rank screening to a more general approach that is based on the class of weighted log-rank test statistics.

Appendix: Proofs

Proof of Theorem 1. For part (i), by definition,

$$\hat{D}_k = n^{-5} \sum_{i,j,h,m,\ell=1}^{n} \phi_k(i,j,h,m,\ell) \tag{A1}$$

is the von Mises statistic associated with the kernel ϕ_k, where

$$\phi_k(i,j,h,m,\ell) = \Delta_i I(X_i \le X_j)\{Z_{ik}(Z_{\ell k}) - Z_{jk}(Z_{\ell k})\}\Delta_h I(X_h \le X_m)\{Z_{hk}(Z_{\ell k}) - Z_{mk}(Z_{\ell k})\}.$$

Define the symmetric kernel

$$\phi_{1k}(i,j,h,m,\ell) = (5!)^{-1} \sum_p \phi_k(i_1,i_2,i_3,i_4,i_5),$$

where $\sum_p$ denotes summation over the 5! permutations (i_1,i_2,i_3,i_4,i_5) of (i,j,h,m,ℓ). For ϕ_{1k}, the corresponding U-statistic is

$$U_{nk} = \binom{n}{5}^{-1} \sum_{1\le i<j<h<m<\ell\le n} \phi_{1k}(i,j,h,m,\ell)$$

$$= \{n(n-1)(n-2)(n-3)(n-4)\}^{-1} \sum_* \phi_k(i,j,h,m,\ell),$$

where $\sum_*$ denotes summation over all distinct i,j,h,m,ℓ from $\{1,2,...,n\}$. It is easy to see that U_{nk} is an unbiased estimate of D_k. Note that $-1 \le \phi_k(i,h,j,k,\ell) \le 1$. As in the proof of Lemma 5.7.3 (page 206, Serfling, 1980), we have $|\hat{D}_k - U_{nk}| \le 2\{n^5 - n(n-1)(n-2)(n-3)(n-4)\}n^{-5} \le 20/n$. By Hoeffding's inequality (page 201, Serfling, 1980; Lemma A2, Song et al., 2014), for $t > 0$,

$$\mathrm{pr}(|U_{nk} - D_k| \ge t) \le 2e^{-\lfloor n/5 \rfloor t^2/2}.$$

Therefore, for any k,

$$\mathrm{pr}(|\hat{D}_k - D_k| > 2cn^{-\alpha}) \le \mathrm{pr}(|\hat{D}_k - U_{nk}| > cn^{-\alpha}) + \mathrm{pr}(|U_{nk} - D_k| > cn^{-\alpha})$$

$$\le \mathrm{pr}(20/n > cn^{-\alpha}) + 2e^{-c^2[n/5]n^{-2\alpha}/2}.$$

Hence, for $n \ge (20/c)^{1/(1-\alpha)}$,

$$\mathrm{pr}(\max_{1\le k\le p_n} |\hat{D}_k - D_k| > 2cn^{-\alpha}) \le 2p_n e^{-c^2[n/5]n^{-2\alpha}/2}.$$

Let $c_1 = c^2/20$. For $n \ge (20/c)^{1/(1-\alpha)} + 10$,

$$\mathrm{pr}(\max_{1\le k\le p_n} |\hat{D}_k - D_k| > 2cn^{-\alpha}) \le 2p_n e^{-c_1 n^{1-2\alpha}}.$$

For part (ii), note that on the event

$$A_n \equiv \{\max_{k\in\mathcal{M}_*} |\hat{D}_k - D_k| \le c_0 n^{-\alpha}/2\},$$

by Condition A, we have $\hat{D}_k \ge c_0 n^{-\alpha}/2$, for all $k \in \mathcal{M}_*$. With $\gamma_n \le c_0 n^{-\alpha}/2$, we have $\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n}$. Therefore, when $n \ge (80/c_0)^{1/(1-\alpha)} + 10$,

$$\mathrm{pr}(\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n}) \ge \mathrm{pr}(A_n) = 1 - \mathrm{pr}(A_n^c) \ge 1 - 2s_n e^{-c_2 n^{1-2\alpha}}. \quad \square$$

To prove Proposition 1, we need the following lemma. First define $\tilde{G}_{0k}(u) = EG_0\{u + m_k(Z_k)\}$. Let f_k be the density function of $\epsilon + \tilde{m}_{-k}(\tilde{Z}_{-k})$. Note that $EG_0\{H(Y)\} = E[G_0\{\epsilon + \tilde{m}_{-k}(\tilde{Z}_{-k}) + m_k(Z_k)\}] = \int_{-\infty}^{\infty} f_k(u)\tilde{G}_{0k}(u)du$, for any $k \in \mathcal{M}_*$.

Lemma A1. (i). If Conditions 2 and 5 hold, then

$$\min_{k\in\mathcal{M}_*} \int_{-\infty}^{\infty} f_k^2(u)\tilde{G}_{0k}^3(u)du \geq \tau^{7/2}/(9\eta).$$

(ii). If Condition 6 holds, then for any $k \in \mathcal{M}_*$,

$$\int_{-\infty}^{\infty} f_k^2(u)du \leq \psi^{1/2}.$$

Proof of Lemma A1. For part (i), by Markov's inequality and Condition 2, for any $k \in \mathcal{M}_*$,

$$\int\limits_{|u|\geq(7\eta^2/\tau)^{1/2}} f_k(u)\tilde{G}_{0k}(u)du \leq (7\eta^2/\tau)^{-1} \int\limits_{|u|\geq(7\eta^2/\tau)^{1/2}} u^2 f_k(u)du \leq \tau/7.$$

Under Condition 5, for any $k \in \mathcal{M}_*$, $\int_{-\infty}^{\infty} f_k(u)\tilde{G}_{0k}(u)du \geq \tau$. It follows that

$$\int_{-(7\eta^2/\tau)^{1/2}}^{(7\eta^2/\tau)^{1/2}} f_k(u)\tilde{G}_{0k}(u)du \geq 6\tau/7.$$

By Hölder's inequality,

$$\int_{-(7\eta^2/\tau)^{1/2}}^{(7\eta^2/\tau)^{1/2}} f_k^{3/2}(u)\tilde{G}_{0k}^{3/2}(u)du \geq \{\int_{-(7\eta^2/\tau)^{1/2}}^{(7\eta^2/\tau)^{1/2}} 1du\}^{-1/2}\{\int_{-(7\eta^2/\tau)^{1/2}}^{(7\eta^2/\tau)^{1/2}} f_k(u)\tilde{G}_{0k}(u)du\}^{3/2} \geq \tau^{7/4}/(3\eta^{1/2}). \tag{A2}$$

By the Cauchy-Schwarz inequality and (A2),

$$\int_{-\infty}^{\infty} f_k^2(u)\tilde{G}_{0k}^3(u)du \geq \{\int_{-\infty}^{\infty} f_k(u)du\}^{-1}\{\int_{-\infty}^{\infty} f_k^{3/2}(u)\tilde{G}_{0k}^{3/2}(u)du\}^2 \geq \tau^{7/2}/(9\eta).$$

For part (ii), under Condition 6, f_k is Lipschitz with constant ψ. For any $k \in \mathcal{M}_*$,

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_k(u)f_k(v)dvdu \geq \int_{-\infty}^{\infty}\int_{u-\psi^{-1/2}}^{u+\psi^{-1/2}} f_k^2(u)dvdu - \int_{-\infty}^{\infty}\int_{u-\psi^{-1/2}}^{u+\psi^{-1/2}} f_k(u)|f_k(u)-f_k(v)|dvdu$$

$$\geq 2\psi^{-1/2}\int_{-\infty}^{\infty} f_k^2(u)du - \int_{-\infty}^{\infty}\int_{u-\psi^{-1/2}}^{u+\psi^{-1/2}} f_k(u)\psi|v-u|dvdu \geq 2\psi^{-1/2}\int_{-\infty}^{\infty} f_k^2(u)du - 1.$$

Hence, for any $k \in \mathcal{M}_*$,

$$\int_{-\infty}^{\infty} f_k^2(u)du \leq \psi^{1/2}. \quad \square$$

*Proof of Proposition 1.*For $k \in \mathcal{M}_*$ and $t \in \mathcal{R}$,

$$S_k(t) = E[I(Y_1 \leq Y_2 \wedge C_1 \wedge C_2)\{Z_{1k}(t) - Z_{2k}(t)\}]$$

$$= E[I(Z_{1k} \le t, Z_{2k} > t)\{I(Y_1 \le Y_2 \wedge C_1 \wedge C_2) - I(Y_2 \le Y_1 \wedge C_1 \wedge C_2)\}].$$

Without loss of generality, assume $m_k(\cdot)$ is monotone increasing. Define $J = I(Y_1 \le Y_2 \wedge C_1 \wedge C_2) - I(Y_2 \le Y_1 \wedge C_1 \wedge C_2)$, $\mu_{ik} = m_k(Z_{ik}), \nu_{ik} = \tilde{m}_{-k}(\tilde{Z}_{i,-k}), \pi_{ik} = \mu_{ik} + \nu_{ik}$, and $\tilde{C}_i = H(C_i)$, $i = 1, 2$. Let $f(\cdot)$ be the density function of ϵ and $G(\cdot|Z)$ be the conditional survival function of $H(C)$ given Z. Note that

$$J = I(\epsilon_1 + \mu_{1k} + \nu_{1k} \le \tilde{C}_1 \wedge \tilde{C}_2) I(\epsilon_1 + \mu_{1k} + \nu_{1k} \le \epsilon_2 + \mu_{2k} + \nu_{2k})$$

$$-I(\epsilon_2 + \mu_{2k} + \nu_{2k} \le \tilde{C}_1 \wedge \tilde{C}_2) I(\epsilon_2 + \mu_{2k} + \nu_{2k} \le \epsilon_1 + \mu_{1k} + \nu_{1k}).$$

Therefore,

$$E(J|Z_{1k} \le t, Z_{2k} > t, \mu_{1k}, \mu_{2k}, \nu_{1k}, \nu_{2k}) =$$

$$\int\int f(u)f(v)G(u + \pi_{1k}|\mu_{1k}, \nu_{1k})G(u + \pi_{1k}|\mu_{2k}, \nu_{2k})I(u \le v + \mu_{2k} - \mu_{1k} + \nu_{2k} - \nu_{1k})dudv$$

$$- \int\int f(u)f(v)G(u + \pi_{2k}|\mu_{1k}, \nu_{1k})G(u + \pi_{2k}|\mu_{2k}, \nu_{2k})I(u \le v + \mu_{1k} - \mu_{2k} + \nu_{1k} - \nu_{2k})dudv$$

$$= \int\int f(u)f(v)G(u + \pi_{1k}|\mu_{1k}, \nu_{1k})G(u + \pi_{1k}|\mu_{2k}, \nu_{2k})I(u \le v + \mu_{2k} - \mu_{1k} + \nu_{2k} - \nu_{1k})dudv$$

$$- \int\int f(u)f(v)G(u + \pi_{2k}|\mu_{1k}, \nu_{2k})G(u + \pi_{2k}|\mu_{2k}, \nu_{1k})I(u \le v + \mu_{1k} - \mu_{2k} + \nu_{1k} - \nu_{2k})dudv.$$

The last equality holds since for any t, $G(t|\mu_{1k}, \nu_{2k})G(t|\mu_{2k}, \nu_{1k}) = G(t|\mu_{1k}, \nu_{1k})G(t|\mu_{2k}, \nu_{2k})$ by Condition 3. By the exchangeability of ν_{1k} and ν_{2k}, it can be shown that

$$E(J|Z_{1k} \le t, Z_{2k} > t, \mu_{1k}, \mu_{2k}) = E(\tilde{J}|Z_{1k} \le t, Z_{2k} > t, \mu_{1k}, \mu_{2k}),$$

where

$$\tilde{J} = I(\epsilon_1 + \mu_{1k} + \nu_{1k} \le \tilde{C}_1 \wedge \tilde{C}_2) I(\epsilon_1 + \mu_{1k} + \nu_{1k} \le \epsilon_2 + \mu_{2k} + \nu_{2k})$$

$$-I(\epsilon_1 + \mu_{2k} + \nu_{1k} \le \tilde{C}_1 \wedge \tilde{C}_2) I(\epsilon_1 + \mu_{2k} + \nu_{1k} \le \epsilon_2 + \mu_{1k} + \nu_{2k}).$$

On the event $\{Z_{1k} \le t, Z_{2k} > t\}$, we know that $\mu_{2k} - \mu_{1k} \ge 0$; thus

$$\tilde{J} \ge I(\epsilon_1 + \mu_{1k} + \nu_{1k} \le \tilde{C}_1) I(\epsilon_2 + \mu_{2k} + \nu_{2k} \le \tilde{C}_2) I(|\epsilon_1 + \nu_{1k} - \epsilon_2 - \nu_{2k}| < \mu_{2k} - \mu_{1k}).$$

Let E^* denote the expectation with respect to ν_{1k} and ν_{2k} and conditional on the event $\{Z_{1k} \le t, Z_{2k} > t\}$ and μ_{1k}, μ_{2k}. Under Condition 4,

$$E(\tilde{J}|Z_{1k} \le t, Z_{2k} > t, \mu_{1k}, \mu_{2k})$$

$$\ge E^* \int\int f(u)f(v)G_0(u + \mu_{1k} + \nu_{1k})G_0(v + \mu_{2k} + \nu_{2k})I(|u + \nu_{1k} - v - \nu_{2k}| < \mu_{2k} - \mu_{1k})dudv.$$

Hence,

$$S_k(t) \geq E\{I(Z_{1k} \leq t, Z_{2k} > t) I(|\epsilon_1 + \nu_{1k} - \epsilon_2 - \nu_{2k}| < \mu_{2k} - \mu_{1k}) G_0(\epsilon_1 + \pi_{1k}) G_0(\epsilon_2 + \pi_{2k})\}.$$

It suffices to prove the result for large n. Let Z_k be an identical and independent copy of Z_{1k} and Z_{2k}. Note that

$$E\{S_k(Z_k)\}$$

$$\geq E\{I(Z_{1k} \leq Z_k < Z_{2k}) I(|\epsilon_1 + \nu_{1k} - \epsilon_2 - \nu_{2k}| \leq \mu_{2k} - \mu_{1k}) G_0(\epsilon_1 + \mu_{1k} + \nu_{1k}) G_0(\epsilon_2 + \mu_{2k} + \nu_{2k})\}$$

$$\geq \int_{-\infty}^{\infty} \int_{u - cn^{-\alpha/2}}^{u} f_k(u) f_k(v) E\{I(Z_{1k} \leq Z_k < Z_{2k}) I(\mu_{2k} - \mu_{1k} \geq cn^{-\alpha/2}) G_0(u + \mu_{1k}) G_0(u + \mu_{2k})\} du dv$$

$$\geq \int_{-\infty}^{\infty} \int_{u - cn^{-\alpha/2}}^{u} f_k^2(u) E\{I(Z_{1k} \leq Z_k < Z_{2k}) G_0(u + \mu_{1k}) G_0(u + \mu_{2k})\} du dv$$

$$- \int_{-\infty}^{\infty} \int_{u - cn^{-\alpha/2}}^{u} f_k^2(u) E\{I(0 < \mu_{2k} - \mu_{1k} < cn^{-\alpha/2})\} du dv - \int_{-\infty}^{\infty} \int_{u - cn^{-\alpha/2}}^{u} |f_k(v) - f_k(u)| f_k(u) du dv$$

$$\geq 6^{-1} cn^{-\alpha/2} \int_{-\infty}^{\infty} f_k^2(u) \tilde{G}_{0k}^3(u) du - 2\xi_k c^2 n^{-\alpha} \int_{-\infty}^{\infty} f_k^2(u) du - 2^{-1} \psi c^2 n^{-\alpha}$$

$$\geq 6^{-1} cn^{-\alpha/2} \{\tau^{7/2}/(9\eta) - 12\psi^{1/2} \xi_k cn^{-\alpha/2} - 3\psi cn^{-\alpha/2}\} > cn^{-\alpha/2} \tau^{7/2}/(120\eta).$$

The second last inequality holds by Lemma A1 and Condition 6. It follows that $D_k = E\{S_k^2(Z_k)\} \geq [E\{S_k(Z_k)\}]^2 \geq c^2 n^{-\alpha} \tau^7 (120\eta)^{-2}$. Let $c_0 = \tau^7 c^2 (120\eta)^{-2}$. Hence, $\min_{k \in \mathcal{M}_*} D_k \geq c_0 n^{-\alpha}$. □

Proof of Proposition 2. (i). In model (3.3),

$$m_k(Z_k) = E[g^T(Z)\beta | Z_k] = \omega_k g_k(Z_k),$$

and

$$m(Z) = \sum_{j=1}^{s_n} g_j(Z_j) \beta_j.$$

For any $k \in \mathcal{D}$, let $\tilde{Z}_{-k} = (\tilde{Z}_{-k,1}, ..., \tilde{Z}_{-k,k-1}, \tilde{Z}_{-k,k+1}, ..., \tilde{Z}_{-k,p_n})^T$, where $\tilde{Z}_{-k,j} = g_j(Z_j) - \Sigma_{kj} g_k(Z_k)$, for $j \neq k$. Note that $\tilde{Z}_{-k}$ is independent of Z_k. Furthermore, Condition 1 holds since

$$\tilde{m}_{-k}(Z) = m(Z) - m_k(Z_k) = \sum_{j=1}^{s_n} g_j(Z_j) \beta_j - \omega_k g_k(Z_k) = \sum_{j=1, j \neq k}^{s_n} \tilde{Z}_{-k,j} \beta_j = \tilde{m}_{-k}(\tilde{Z}_{-k}).$$

(ii). This follows from the fact that

$$\text{var}(H(Y) - m_k(Z_k)) = \text{var}(\epsilon + \tilde{m}_{-k}(\tilde{Z}_{-k})) \leq \sigma^2 + \text{var}(m(Z)) = \sigma^2 + \beta_{\mathcal{D}}^T \Sigma_{\mathcal{D}\mathcal{D}} \beta_{\mathcal{D}}.$$

(iii). Let Φ and ϕ be the standard normal distribution function and density function, respectively. Note that $Q_k(\cdot) = \Phi(\cdot/\omega_k)$ and $f(\cdot) = \sigma^{-1}\phi(\cdot/\sigma)$. Condition 6 holds since $\xi_k \le (2\pi)^{-1/2}(|\omega_k|)^{-1}$ and $\psi = 2^{-1}\pi^{-1/2}\sigma^{-2}$. $\square$

Proof of Theorem 2. On the event $\{Z_{1k} \le t, Z_{2k} > t\}$,

$$0 \le \tilde{J} \le I(\tilde{C}_1 \wedge \tilde{C}_2 - \mu_{2k} < \epsilon_1 + \nu_{1k} \le \tilde{C}_1 \wedge \tilde{C}_2 - \mu_{1k}) + I(|\epsilon_1 + \nu_{1k} - \epsilon_2 - \nu_{2k}| \le \mu_{2k} - \mu_{1k}),$$

Let $c_5 = 3(2\pi)^{-1/2}\sigma^{-1}$. Note that

$$0 \le E\{\tilde{J}|Z_{1k} \le t, Z_{2k} > t, \mu_{1k}, \mu_{2k}\} \le c_5(\mu_{2k} - \mu_{1k}).$$

Thus,

$$0 \le S_k(t) \le c_5 E\{I(Z_{1k} \le t, Z_{2k} > t)(m_k(Z_{2k}) - m_k(Z_{1k}))\} \le c_5 E\{|m_k(Z_{2k}) - m_k(Z_{1k})|\}.$$

It follows that

$$0 \le D_k \le c_5^2[E\{|m_k(Z_{2k}) - m_k(Z_{1k})|\}]^2 \le c_5^2 E\{m_k(Z_{2k}) - m_k(Z_{1k})\}^2 = 2c_5^2\omega_k^2.$$

If $D_k \ge c_0 n^{-\alpha}$, then $\omega_k^2 \ge c_5^{-2}c_0 n^{-\alpha}/2$. Since $\mathrm{var}\{H(T)\} = \beta_{\mathcal{D}}^T \Sigma_{\mathcal{D}\mathcal{D}}\beta_{\mathcal{D}} + \sigma^2 \le \eta^2$, we have $\sum_{k=1}^{p_n} \omega_k^2 = \beta^T \Sigma \Sigma \beta \le \eta^2 \lambda_{\max}(\Sigma)$. Hence, the size of $\{k : D_k \ge c_0 n^{-\alpha}\} \le 2c_5^2\eta^2\lambda_{\max}(\Sigma)n^{\alpha}/c_0$. Because the size of $\{k : \hat{D}_k \ge 2c_0 n^{-\alpha}\}$ is at most the size of $\{k : D_k \ge c_0 n^{-\alpha}\}$ on the event $\{\max_{1\le k\le p_n} |\hat{D}_k - D_k| \le c_0 n^{-\alpha}\}$. Taking $c_4 = 2c_5^2\eta^2/c_0$, we have

$$\mathrm{pr}[|\hat{\mathcal{M}}_{\gamma_n}| \le c_4 n^{\alpha}\lambda_{\max}(\Sigma)\}] \ge \mathrm{pr}(\max_{1\le k\le p_n} |\hat{D}_k - D_k| \le c_0 n^{-\alpha}).$$

The conclusion follows from part (i) of Theorem 1.

References

Breiman, L. and Friedman, J. H. (1985) Estimating optimal transformations for multiple regression and correlation. *J. Am. Statist. Assoc.* **80**, 580-619.

Clayton, D. and Cuzick, J. (1985) Multivariate generalizations of the proportional hazards model (with discussion). *J. R. Statist. Soc. A* **148** 82-117.

Fan J., Feng Y. and Song R. (2011) Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Am. Statist. Assoc.* **106** 544-57.

Fan J., Feng Y. and Wu Y. (2010) Ultrahigh dimensional variable selection for Cox's proportional hazards model. *IMS Collections* **6** 70-86.

Fan, J. and Lv, J. (2008) Sure Independence Screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc. B* **70** 849-911.

Fan J. and Song R. (2010) Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567-604.

Gehan, E. A. (1965) A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52** 203-23.

Gonzalez-Manteiga, W. and Cadarso-suarez, C. (1994) Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. *J. Nonparam. Statist.* **4** 65-78.

Gorst-Rasmussen, A. and Scheike, T. (2013) Independent screening for single-index hazard rate models with ultrahigh dimensional features. *J. R. Statist. Soc. B* **75** 217-45.

He X. and Shen L. (1997) Linear regression after spline transformation. *Biometrika* **84** 474-81.

He X., Wang L. and Hong H. G. (2013) Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Statist.* **41** 342-69.

Jin, Z., Lin, D. Y., Wei, L. J. and Ying, Z. (2003) Rank-based inference for the accelerated failure time models. *Biometrika* **90** 341-353.

Kalbfleisch, J. D. and Prentice, R. L. (2002) *The Statistical Analysis of Failure Time Data*, 2nd ed. New York: Wiley.

Kendall, M. G. (1962) *Rank Correlation Methods*, 3rd ed. London: Griffin & Co.

Mai, Q. and Zou, H. (2015) The fused Kolmogorov filter: a nonparametric model-free screening method. *Ann. Statist.* **43** 1471-97.

Lenz, G., Wright, G., Dave, S. S., Xiao, W., Powell, J., Zhao, H., Xu, W., Tan, B., Chan, W. C. and Staudt, L. M. (2008) Stromal gene signatures in large-B-cell lymphomas. *N. Engl. J. Med.* **359** 2313-2323.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Song, R., Lu, W., Ma, S. and Jeng, X. J. (2014) Censored rank independence screening for high-dimensional survival data. *Biometrika* **101** 799-814.

Wang, Z. and Wang, C. Y. (2010) Buckley-James boosting for survival analysis with high-dimensional biomarker Data. *Stat. Appl. Genet. Molec. Biol.* **9** 24.

Wu, Y. and Yin, G. (2015) Conditional quantile screening in ultrahigh-dimensional heterogeneous data. *Biometrika* **102** 65-76.

Zhao, D. S. and Li, Y. (2012) Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J. Mult. Anal.* **105** 397-411.

Table 1: Simulation results for Example 5.1 in the case of homogeneous censoring with true model size $p_0 = 10$: reported are the median and interquartile range of the minimum model size needed to include all active predictors, along with the proportion $\mathcal{P}_{\text{All}}$ that all of the active predictors are selected for a given model size

Error	CP (%)	Method	Median	IQR	$\mathcal{P}_{\text{All}}$	Median	IQR	$\mathcal{P}_{\text{All}}$
				n=100			n=200	
Normal	20	GR	12	5.3	0.81	11	2.0	1.00
		CR	12	5.0	0.89	11	2.0	1.00
		FAST	13	6.0	0.80	11	2.0	1.00
		PL	12	5.3	0.87	11	2.0	1.00
	40	GR	13	8.3	0.77	11	2.0	1.00
		CR	138	556.0	0.17	29	96.8	0.54
		FAST	16	14.5	0.66	11	2.0	1.00
		PL	13	7.0	0.82	11	2.0	1.00
Extreme value	20	GR	13	14.0	0.72	11	2.0	1.00
		CR	13	12.0	0.73	11	2.0	1.00
		FAST	17	18.3	0.62	11	2.0	0.99
		PL	13	12.0	0.71	11	2.0	0.99
	40	GR	13	10.5	0.75	11	2.0	1.00
		CR	81	333.3	0.26	14	44.3	0.71
		FAST	15	12.3	0.68	11	2.0	1.00
		PL	13	8.0	0.80	11	2.0	1.00
Logistic	20	GR	13	9.3	0.76	11	2.0	1.00
		CR	13	7.0	0.83	11	2.0	1.00
		FAST	14	17.3	0.67	11	2.0	1.00
		PL	13	7.0	0.79	11	1.0	1.00
	40	GR	14	11.0	0.74	11	2.0	1.00
		CR	133	412.0	0.14	19	60.3	0.63
		FAST	15	13.5	0.71	11	2.0	1.00
		PL	13	6.3	0.85	11	2.0	1.00

IQR, interquartile range of the minimum model size; CP, censoring proportion; GR, proposed Gehan rank screening; CR, censored rank screening of Song et al. (2014); FAST, feature aberration at survival times screening of Gorst-Rasmussen& Scheike (2013), PL, partial likelihood ratio screening.

Table 2: Simulation results for Example 5.1 in the case of heterogeneous censoring with true model size $p_0 = 10$: reported are the median and interquartile range of the minimum model size needed to include all active predictors, along with the proportion $\mathcal{P}_{\text{All}}$ that all of the active predictors are selected for a given model size

Error	CP (%)	Method	Median	IQR	$\mathcal{P}_{\text{All}}$	Median	IQR	$\mathcal{P}_{\text{All}}$
			n=100			n=200		
Normal	20	GR	19	32.5	0.56	12	2.0	0.99
		CR	22	36.3	0.49	12	2.0	0.99
		FAST	49	109.8	0.26	14	7.0	0.89
		PL	33	79.5	0.44	12	3.3	0.96
	40	GR	47	100.5	0.29	13	6.0	0.91
		CR	122	306.8	0.06	17	24.0	0.75
		FAST	158	377.5	0.07	23	47.5	0.62
		PL	98	330.3	0.11	16	24.0	0.76
Extreme value	20	GR	18	21.0	0.58	11	1.0	1.00
		CR	20	28.0	0.56	11	1.0	0.99
		FAST	53	92.0	0.21	13	3.3	0.97
		PL	34	59.5	0.37	12	2.0	0.98
	40	GR	40	78.8	0.23	12	3.0	0.98
		CR	133	249.5	0.11	15	16.5	0.77
		FAST	149	248.8	0.11	23	60.3	0.69
		PL	80	174.8	0.22	15	26.0	0.75
Logistic	20	GR	17	28.3	0.59	11	2.0	0.99
		CR	18	24.3	0.58	12	2.0	0.98
		FAST	40	93.5	0.28	14	9.8	0.89
		PL	23	56.5	0.45	13	5.0	0.93
	40	GR	43	71.0	0.26	13	4.0	0.95
		CR	87	252.3	0.09	17	22.3	0.77
		FAST	128	243.5	0.03	25	48.5	0.61
		PL	84	157.5	0.14	18	31.3	0.71

Table 3: Simulation results for Example 5.2 with sample size $n = 200$ and true model size $p_0 = 5$: reported are the median and interquartile range of the minimum model size needed to include all active predictors, along with the proportion $\mathcal{P}_{\text{All}}$ that all of the active predictors are selected for a given model size

Error	CP (%)	Method	Median	IQR	$\mathcal{P}_{\text{All}}$	Median	IQR	$\mathcal{P}_{\text{All}}$
			homogeneous censoring			heterogeneous censoring		
Normal	20	GR	12	4.0	0.90	12	10.0	0.87
		CR	13	8.3	0.86	19	53.3	0.65
		FAST	16	29.8	0.72	54	212.3	0.42
		PL	13	9.3	0.89	17	76.5	0.62
	40	GR	12	5.3	0.91	21	35.3	0.71
		CR	14	12.0	0.80	75	327.3	0.36
		FAST	19	36.3	0.68	165	351.0	0.24
		PL	15	12.0	0.81	57	141.3	0.42
Extreme value	20	GR	13	5.0	0.93	16	14.3	0.84
		CR	13	10.0	0.86	46	84.3	0.47
		FAST	18	21.0	0.79	131	259.8	0.23
		PL	13	7.0	0.85	46	145.3	0.47
	40	GR	14	10.0	0.87	17	38.3	0.72
		CR	15	21.3	0.80	83	250.5	0.35
		FAST	20	57.5	0.64	129	415.0	0.21
		PL	15	31.0	0.74	51	216.0	0.47
Logistic	20	GR	12	6.0	0.93	13	16.8	0.81
		CR	14	9.3	0.83	30.5	82.5	0.54
		FAST	20	37.3	0.68	89	214.5	0.34
		PL	13	10.5	0.82	29	93.0	0.57
	40	GR	14	20.3	0.80	17	97.3	0.62
		CR	17	29.0	0.74	124	463.3	0.33
		FAST	32	124.5	0.55	232	515.5	0.24
		PL	16	58.5	0.65	75	326.5	0.38

Table 4: Simulation results for varying censoring mechanisms and censoring proportions

Error	CP (%)	γ	Median	IQR	$\mathcal{P}_{\text{All}}$	Median	IQR	$\mathcal{P}_{\text{All}}$
			n=100			n=200		
Normal	20	1	19	32.5	0.56	12	2.0	0.99
		0.6	17	21.5	0.66	12	2.0	0.99
		0.3	15	10.9	0.75	11	2.0	1.00
		0	12	5.3	0.81	11	2.0	1.00
	40	1	47	100.5	0.29	13	6.0	0.91
		0.6	32	75.2	0.35	13	5.0	0.93
		0.3	25	38.9	0.56	12	3.0	0.96
		0	13	8.3	0.77	11	2.0	1.00
Extreme value	20	1	18	21.0	0.58	11	1.0	1.00
		0.6	16	17.0	0.65	11	1.0	1.00
		0.3	14	15.0	0.69	11	1.0	1.00
		0	13	14.0	0.72	11	2.0	1.00
	40	1	40	78.8	0.23	12	3.0	0.98
		0.6	29	50.2	0.37	12	3.0	0.99
		0.3	17	20.6	0.59	11	2.0	0.99
		0	13	10.5	0.75	11	2.0	1.00
Logistic	20	1	17	28.3	0.59	11	2.0	0.99
		0.6	15	16.2	0.63	11	2.0	0.99
		0.3	14	12.7	0.70	11	2.0	1.00
		0	13	9.3	0.76	11	2.0	1.00
	40	1	43	71.0	0.26	13	4.0	0.95
		0.6	31	50.3	0.40	12	3.0	0.97
		0.3	20	25.1	0.61	12	3.0	0.99
		0	14	11.0	0.74	11	2.0	1.00

Table 5: Top 10 selected genes for lymphoma data

Order	GR	CR	FAST	PL
1	229839_at	241647_x_at	236981_at	229839_at
2	206439_at	216233_at	1554413_s_at	1569344_a_at
3	1554413_s_at	240563_at	1569344_a_at	237493_at
4	237493_at	1558999_x_at	1553499_s_at	236981_at
5	231049_at	223864_at	240898_at	240898_at
6	1569344_a_at	1562727_at	231455_at	1553499_s_at
7	226869_at	1565799_at	231049_at	231049_at
8	240898_at	220393_at	1568752_s_at	243713_at
9	212713_at	1553441_at	229839_at	1554413_s_at
10	243713_at	1561765_at	1568751_at	237797_at

Table 6: Selection results for seven biologically relevant probe sets in the lymphoma study; selected genes are indicated by a ✓ symbol

Probe set	Gene symbol	GR	CR	FAST	PL	GR	CR	FAST	PL
		Model size 9				Model size 34			
1558999_x_at	LOC283922/PDPR		✓			✓	✓		✓
212713_at	MFAP4	✓				✓		✓	✓
224043_s_at	UPB1								
229839_at	SCARA5	✓		✓	✓	✓		✓	✓
237515_at	TMEM56						✓		
237797_at	DNM1L					✓			✓
242758_x_at	JMJD1A					✓			✓