α-KIDS: A novel feature evaluation in the ultrahigh-dimensional right-censored setting, with application to Head and Neck Cancer

A Preprint

Atika FArzana Urmi, Ph.D
Department of Biostatistics
Virginia Commonwealth University
VA, USA
urmiaf@vcu.edu

Chenlu Ke, Ph.D[*]
Department of Statistical Sciences
and Operations Research
Virginia Commonwealth University
VA, USA
kec2@vcu.edu

Dipankar Bandyopadhyay, Ph.D
Department of Biostatistics
Virginia Commonwealth University
VA, USA
dbandyop@vcu.edu

August 13, 2024

Abstract

Recent advances in sequencing technologies have allowed collection of massive genome-wide information that substantially enhances the diagnosis and prognosis of head and neck cancer. Identifying predictive markers for survival time is crucial for devising prognostic systems, and learning the underlying molecular driver of the cancer course. In this paper, we introduce α-KIDS, a model-free feature screening procedure with false discovery rate (FDR) control for ultrahigh dimensional right-censored data, which is robust against unknown censoring mechanisms. Specifically, our two-stage procedure initially selects a set of important features with a dual screening mechanism using nonparametric reproducing-kernel-based ANOVA statistics, followed by identifying a refined set (of features) under directional FDR control through a unified knockoff procedure. The finite sample properties of our method, and its novelty (in light of existing alternatives) are evaluated via simulation studies. Furthermore, we illustrate our methodology via application to a motivating right-censored head and neck (HN) cancer survival data derived from The Cancer Genome Atlas, with further validation on a similar HN cancer data from the Gene Expression Omnibus database. The methodology can be implemented via the R package DSFDRC, available in GitHub.

Keywords FDR control; feature screening; head and neck cancer; high dimensional covariates; nonparametric; right-censoring

1 Introduction

Recent advancements in automated data collection techniques has led to an escalating prevalence of ultrahigh-dimensional data in biomedical sciences. Such data frequently exhibits an abundance of features that far surpasses the available number of observations. A salient example is evident in the massive data generated through high-throughput sequencing processes. With the ability to capture a wide spectrum of molecular, genetic, and phenotypic information on a large scale, researchers can now explore complex biological systems with unprecedented granularity and unveil novel

[*] Corresponding Author

insights into precision medicine, biomarker identification, and the exploration of pathways and networks. However, traditional statistical learning methods tend to falter when confronted with ultrahigh-dimensional data - a predicament known as the ‘curse of dimensionality’. The present work was motivated by a study of the head and neck squamous cell carcinomas (HNSCC). Constituting around 4% of all cancer cases in the United States, head and neck cancer predominantly manifests as squamous cell carcinomas[1]. Despite surgery, radiation and chemotherapy, the 5-year survival rate stands at only 40-50% among all patients[30]. Extensive studies have found high biological and clinical heterogeneity in HNSCC patients, underscoring the need for a deeper molecular-level understanding of the disease.

Our goal in this paper is to identify which genes, among hundreds of thousands, contribute to survival prognosis of HNSCC, utilizing data from the HNSCC cohort available in the Cancer Genome Atlas (TCGA) network. The primary endpoint is the time to death (due to HNSCC). Due to loss to follow-up or no event occurrence until the study’s conclusion, more than half of cases were right-censored. As it is widely acknowledged that only a small subset of molecular features are truly relevant to specific clinical outcomes, feature selection has become one of the cornerstones for biomarker identification. While regularization methods such as LASSO [43], SCAD [13], adaptive Lasso [53], and Dantzig selector [6] have been the most popular feature selection tools, they can suffer from various issues including computational expediency, statistical inaccuracy and algorithmic instability [16] when applied to ultrahigh dimensional settings. A pragmatic solution is to perform feature screening before embarking on exact feature selection. A screening procedure applies a coarse filter to individual features to winnow out a significant portion of noise, thus circumventing a concurrent analysis of all features that gives rise to the aforementioned issues. This is frequently achieved through dependence learning between the outcome and individual features, which can be model-based [14, 23, 12], or model-free [52, 32, 35]. Given the challenges of verifying model assumptions in ultrahigh dimensional data, coupled with the risk of overlooking vital features due to model misspecification, opting for model-free screening approaches is a more prudent choice in practice.

In the regime of survival analysis, a plethora of model-free feature screening procedures have been developed for survival outcomes, subject to right censoring [40, 31, 51, 26, 7]. Since the true survival time is not fully observable, these approaches focus on the association between the estimated survival probabilities and individual features through, for examples, inverse-probability-of-censoring-weighted (IPCW) rank correlation [40], a generalized Kolmogorov statistic for covariate-stratified survival distribution [26], and distance correlation [7]. Nonetheless, there is no free lunch in estimating the survival function; assumptions on censoring are imposed, explicitly or implicitly, to ensure proper behavior of the survival estimators. The efficacy of the most widely used Kaplan-Meier estimation requires censoring to be independent of the event, as well as the predictive features for the survival time. Violations of independent censoring are not uncommon; subjects may tend to withdraw from the study due to either favorable or unfavorable prognosis, which can be further complicated by the effect of prognostic covariates. The ability of the IPCW method to adjust for dependent censoring hinges on the assumptions of exchangeability, and correct model specification used to estimate the weights. While a biased estimator undermines screening accuracy, it is extremely difficult to conjecture the censoring mechanism in practice to ensure unbiasedness, given ultrahigh dimensional covariates. Heavy censoring can also lead to less reliable estimators as the equivalent number of subjects at risk decreases at later times. The main impetus of this paper is the general lack of flexible and reliable screening tools for ultrahigh dimensional survival analysis, that are robust to heavy censoring and uncertain censoring mechanism as presented in the TCGA HNSCC dataset.

Moreover, in order to ensure the retention of important features with high confidence, feature screening procedures tend to opt for a conservative threshold to distinguish signal from noise. This approach, however, often leads to an excess of false discoveries. Consequently, it is essential to supplement the feature screening process with a more precise feature selection step to control the false discovery rate (FDR) in biomarker identification, thereby enhancing the accuracy of prognostic modeling.

Traditionally, feature selection and feature screening have been treated as separate domains in the literature. This division persisted until recent groundbreaking contributions in the form of the knockoff methodology [4, 5, 3, 33], which bridged this gap. The knockoff features are strategically designed to replicate the correlation structure inherent in the original variables, serving as negative controls to aid in the identification of truly significant features while controlling the FDR. This innovative approach can be extended to ultrahigh dimensional survival analysis, offering a solution to the longstanding challenge of FDR control in feature screening. In this paper, we propose a novel feature screening procedure with FDR control for ultra-high dimensional right-censored survival outcomes. The proposed method operates in two key stages. First, we find a preliminary set of potentially important features with a dual screening mechanism. Specifically, two filters are implemented to screen out irrelevant information through nonparametric dependence learning between the raw survival outcome and individual features, with no need for intermediate survival function estimation. The contribution of each feature is quantified by reproducing-kernel-based ANOVA statistics [28] in a model-free way. Then, we further identify a refined set of important features under directional FDR through a unified knockoff procedure based on the same utility measures adopted in the initial screening step. Our proposed method enjoys several distinctive advantages. First, it requires no pre-specification of the model structure and minimizes the assumption on the censoring mechanism. As a result, it exhibits more resilience to dependent and heavy censoring, than existing alternatives. Second, it effectively detects both linear and nonlinear features by capitalizing on the kernel-based utility measures. Third, it coherently controls FDR, and thus protects prognostic modeling from excessive noise. Finally, it boasts general applicability to other censored regression settings characterized by ultrahigh dimensional data with easy and fast implementation. All these advantages greatly facilitate the utilization of the proposed method in real applications. We substantiate both theoretically and numerically that the proposed feature evaluation procedure enjoys the sure screening property with rigorous control over FDR. Furthermore, through empirical analysis conducted on the TCGA HNSCC dataset and external validation, we showcase the efficacy and practical utility of the proposed method.

The rest of the paper is organized as follows. In Section 2, we develop our new framework of feature screening for ultrahigh dimensional survival analysis, and provide necessary theoretical justification. Simulation studies assessing finite sample properties of our proposal, and comparisons to existing alternatives are provided in Section 3. We illustrate our proposed methodology via application to the motivating TCGA HNSCC data in Section 4. Finally, Section 5 concludes, with a short discussion. All technical proofs, along with additional simulation results are deferred to the Web Supplement accompanying the paper.

2 Statistical Methods

In this section, we introduce a model-free dual screening framework for ultra-high dimensional right-censored data with FDR control. The proposed method is implemented via two main steps. First, it finds a crude set of potentially important features through a dual screening mechanism. Then, it further identifies a refined set of important features under directional FDR.

2.1 Assumptions and Dual Screening

Let T be the survival time and $\mathbf{X} = (X_1, \dots, X_p)^T$ be a p-dimensional set of covariates. We denote the censoring time by C. In reality, the observable survival response variables are (Y, δ), where $Y = \min\{T, C\}$ and $\delta = I\{T \leq C\}$ with $I\{\cdot\}$ being the indicator function. Ideally, we would like to identify the smallest active set, denoted as $\mathbf{X}_{\mathcal{A}_T} = \{X_j : j \in \mathcal{A}_T\}$, satisfying

$$T \perp\!\!\!\perp \mathbf{X} | \mathbf{X}_{\mathcal{A}_T}. \tag{1}$$

Since T is not fully observable, our focus shifts to identifying the smallest active set relevant to the observable outcome (Y, δ), denoted as $\mathbf{X}_{\mathcal{A}} = \{X_j : j \in \mathcal{A}\}$, ensuring that

$$(Y, \delta) \perp\!\!\!\perp \mathbf{X} | \mathbf{X}_{\mathcal{A}}. \tag{2}$$

Nonetheless, the relation between the two active sets $\mathcal{A}_T$ and $\mathcal{A}$ can be established under a relatively simple condition, as shown in the following proposition.

Proposition 1. *Let $\mathbf{X}_{\mathcal{A}_T}$ and $\mathbf{X}_{\mathcal{A}}$ be the active sets that satisfy (1) and (2), respectively. Assuming that*

$$C \perp\!\!\!\perp \mathbf{X} | (\mathbf{X}_{\mathcal{A}_T}, T), \tag{3}$$

we have $\mathbf{X}_{\mathcal{A}} \subseteq \mathbf{X}_{\mathcal{A}_T}$.

The pair of conditions (1) and (3) is equivalent to $(T, C) \perp\!\!\!\perp \mathbf{X} | \mathbf{X}_{\mathcal{A}_T}$, which implies $(Y, \delta) \perp\!\!\!\perp \mathbf{X} | \mathbf{X}_{\mathcal{A}_T}$ because (Y, δ) is a function of (T, C). It follows immediately that under condition (3), the important predictors for the observable outcome (Y, δ) are also important predictors for the true survival time T. Moreover, it is expected that in practice the equality in $\mathbf{X}_{\mathcal{A}} \subseteq \mathbf{X}_{\mathcal{A}_T}$ will normally hold since proper containment requires carefully balanced conditions. We note that assumption (3) is a mild condition since it allows censoring to vary with the true survival time and all the prognostic features. By contrast, the independent censoring condition,

$$C \perp\!\!\!\perp (T, \mathbf{X}),$$

is more stringent and implies assumption (3). While many existing survival data screening methods assume independent censoring [40, 7, 10, 34], this assumption can be unrealistic in cases with complex censoring mechanisms and a large number of features. Another common assumption for ensuring identifiability [51, 26, 48],

$$C \perp\!\!\!\perp T | \mathbf{X},$$

does not ensure the equivalence of $\mathbf{X}_{\mathcal{A}_T}$ and $\mathbf{X}_{\mathcal{A}}$.

Throughout this paper, we assume that condition 3 holds, based on which a new framework of feature screening for right-censored data is developed. Specifically, we propose a new screening approach that directly applies to the raw survival outcome and thus avoids estimating the survival function. The following proposition lays the cornerstone for our method.

Proposition 2. *The pair of the following conditions* $(b1)$ *and* $(b2)$ *is equivalent to condition* (a)*:*

$$(a)\ X_j \perp\!\!\!\perp (Y, \delta);$$

$$(b1)\ X_j \perp\!\!\!\perp \delta;\ (b2)\ X_j \perp\!\!\!\perp Y|\delta,$$

for $j = 1, \ldots, p$.

According to Proposition 2, the pair of conditions $(b1)$ and $(b2)$ jointly implies the irrelevance of X_j. Or, in other words, important features must be either marginally correlated with δ or conditionally correlated with Y given δ. Since δ is simply binary, the two conditions can be easily assessed by a wide range of independence measures. As a result, feature screening for right-censored data boils down to traditional univariate independence learning on complete data, bypassing the need to estimate the survival function. In this paper, we adopt a recently developed nonparametric independence measure, namely Expected Conditional Characteristic function Based Independence Criterion[28] (ECCFIC), as the filter in our screening procedure. We briefly review ECCFIC in the following subsection and illustrate its advantages over some existing celebrated independence measures.

2.2 Independence Measures

Let $U, V \in \mathbb{R}$ be two random variables. Also, let φ_U denote the characteristic function of U, and $\varphi_{U|V}$ denote the conditional characteristic function of U given V. Elicited by the fact that $U \perp\!\!\!\perp V$ if and only if $\varphi_{U|V} = \varphi_U$, the ECCFIC for quantifying the association between U and V is defined by

$$\mathcal{H}^2_w(U|V) = E_V \int_R |\varphi_{U|V}(t) - \varphi_U(t)|^2 dw(t), \tag{4}$$

where, $w(\cdot)$ is a finite nonnegative Borel measure on $\mathbb{R}$. An equivalent formula to (4) is given by

$$\mathcal{H}^2_K(U|V) = E_V E_{U|V,U'|V} K(U - U') - E_{U,U'} K(U - U'), \tag{5}$$

where (U', V') is an i.i.d. copy of (U, V), $E_{U|v,U'|v}(\cdot)$ denotes conditional expectation $E(\cdot|V = v, V' = v)$, and $K : \mathbb{R} \to \mathbb{C}$ is a translation-invariant positive definite kernel induced by w, such that $K(x) = \int_{\mathbb{R}} e^{-ixt} dw(t)$ for $x \in \mathbb{R}$ by Bochner Theorem [46]. Henceforth, we consider the alternative representation in (5), as it is easier to estimate for a given kernel. As a special case, we have

$$\mathcal{H}^2_K(U|U) = K(0) - E_{U,U'} K(U - U').$$

It can be shown that $0 \leq \mathcal{H}^2_K(U|V) \leq \mathcal{H}^2_K(U|U)$. Moreover, if K is characteristic [19], then $\mathcal{H}^2_K(U|V) = 0$ if and only if $U \perp\!\!\!\perp V$. As a result, we can define an R^2-type statistic as

$$R^2_K(U|V) = \frac{\mathcal{H}^2_K(U|V)}{\mathcal{H}^2_K(U|U)}$$

and $0 \leq R^2_K(U|V) \leq 1$. In particular, $R^2_K(U|V) = 0$ if and only if $U \perp\!\!\!\perp V$ and $R^2_K(U|V) = 1$ if and only if U is a measurable function of V. Intuitively, this kernel-based R^2 statistic can be regarded as a nonlinear generalization of the classical R^2 as it requires no linearity or distributional assumptions for the regression of U on V. It is worth noting that ECCFIC is closely related to a well-known family of measures, called Hilbert-Schmidt independence criterion[20] (HSIC), which includes distance covariances[42] as a special case. To assess the association between U and V, HSIC consider the discrepancy between the joint characteristic function $\varphi_{U,V}$ and the product of the marginals $\varphi_U \varphi_V$ [21], where the two random variables are treated symmetrically. Although HSIC equal zero also indicates independence and vice versa, it is not clear under what circumstances HSIC approaches its upper bound or how the random variables are related when the upper bound is attained. Compared to HSIC, ECCFIC better quantifies the contribution of a feature to the outcome, since it characterizes both independence and functional dependence in a supervised way, thereby making ECCFIC a more appealing alternative for model-free feature screening.

In the similar vein, the marginal effect associated with V (given another variable $Z \in \mathbb{R}$ already contained in the model to explain U) can be measured by the expected conditional characteristic function based conditional independence criterion (ECCFCIC [28]):

$$\mathcal{H}^2_K(U|V;Z) = E_{(V,Z)} E_{U|(V,Z),U'|(V,Z)} K(U - U') - E_Z E_{U|Z,U'|Z} K(U - U').$$

Then, a kernel-based partial R^2 statistic can be defined by

$$R_K^2(U|V;Z) = \frac{\mathcal{H}_K^2(U|V;Z)}{\mathcal{H}_K^2(U|U;Z)},$$

and $0 \leq R_K^2(U|V;Z) \leq 1$ with $R_K^2(U|V;Z) = 0$ if and only if $U \perp\!\!\!\perp V|Z$. Although here we restrict ourselves to translation-invariant kernel for ease of presentation, ECCFIC can be generalized using any positive definite characteristic kernel in the associated reproducing kernel Hilbert space[28]. Examples of characteristic kernel include but not limited to Gaussian, Laplacian, and inverse multiquadric.

2.3 The Screening Procedure

Returning to the context of feature screening for right-censored data, we propose to use the following two utility measures to evaluate each feature based on conditions $(b1)$ and $(b2)$, respectively:

1. $\Omega_{j,1} = R_K^2(X_j|\delta)$;
2. $\Omega_{j,2} = R_K^2(X_j|Y;\delta)$.

The first measure is the kernel-based R^2 for the inverse regression of $X_j|\delta$ and the second measure is the kernel-based partial R^2 for the inverse regression of $X_j|Y$ while adjusting for δ. Here the inverse regression is to facilitate sample estimation as δ is a binary variable. We note again that any appropriate nonparametric dependence measures may be used to access conditions $(b1)$ and $(b2)$. For example, the marginal measure can be replaced by the Kolmogorov filter [35] or the MV-SIS filter [8], while the conditional measure can be replaced by $R_K^2(Y|X_j;\delta)$ by exchanging X_j and Y.

Given sample data $\{(\mathbf{X}_i, Y_i, \delta_i) : i = 1, \dots, n\}$, we develop the estimators for the proposed utility measures. Let $\mathcal{J}_s = \{i : \delta_i = s\}$, $n_s = |\mathcal{J}_s|$, and $w_s = \frac{n_s}{n}$ for $s = 0, 1$. Denote $\overline{K}_j = \frac{1}{n^2}\sum_{i_1,i_2} K(X_{i_1,j} - X_{i_2,j})$ and $\overline{K}_{j,s} = \frac{1}{n_s^2}\sum_{i_1,i_2\in\mathcal{J}_s} K(X_{i_1,j} - X_{i_2,j})$ for $s = 0, 1$. The marginal utility can be estimated as

$$\widehat{\Omega}_{j,1} = \frac{\sum_{s=0,1} w_s\overline{K}_{j,s} - \overline{K}_j}{K(0) - \overline{K}_j}. \tag{6}$$

The conditional utility can be estimated as

$$\widehat{\Omega}_{j,2} = \frac{\sum_{s=0,1} w_s\mathcal{H}^2_{K,G_{h_s},n_s}(X_j|Y;\delta=s)}{K(0) - \sum_{s=0,1} w_s\overline{K}_{j,s}}, \tag{7}$$

where, $\mathcal{H}^2_{K,G_{h_s},n_s}(X_j|Y;\delta=s)$ is the Nadaraya-Watson estimator of $\mathcal{H}_K^2(X_j|Y)$ given $\delta = s$, relying on a smoothing kernel $G : \mathbb{R} \to \mathbb{R}$ and an associated tuning bandwidth $h_s = h_s(n_s) \in \mathbb{R}$. Specifically,

$$\mathcal{H}^2_{K,G_{h_s},n_s}(X_j|Y;\delta=s) = \frac{1}{n_s}\sum_{i_1\in\mathcal{J}_s}\frac{\sum_{i_2,i_3\in\mathcal{J}_s} G_{h_s}(Y_{i_1} - Y_{i_2})G_{h_s}(Y_{i_1} - Y_{i_3})K(X_{i_2,j} - X_{i_3,j})}{\sum_{i_2,i_3\in\mathcal{J}_s} G_{h_s}(Y_{i_1} - Y_{i_2})G_{h_s}(Y_{i_1} - Y_{i_3})} - \overline{K}_{j,s},$$

where, $G_{h_s}(\cdot) = \frac{1}{h_s}G(\cdot/h_s)$.

According to Proposition 2, features making discernibly marginal or conditional contribution to the survival outcome should be retained. Therefore, we estimate the active index set by

$$\widehat{\mathcal{A}} = \{1 \leq j \leq p : \widehat{\Omega}_{j,1} \geq c_1 n^{-\gamma_1} \text{ or } \widehat{\Omega}_{j,2} \geq c_2 n^{-\gamma_2}\},$$

where, c_1, c_2, γ_1 and γ_2 are some threshold values relying on the strength of the true signal, which is to be defined in condition v below. Henceforth, we refer to the proposed screening procedure as *k*ernel-based *i*ndependence *d*ual *s*reening (abbreviated as KIDS). The proposed procedure embraces the sure screening property as well as the rank consistency property, which are established in Theorems 2.3 and 2.3 below.

Let $\mathcal{A}_1 = \{j \in \mathcal{A} : X_j \not\perp\!\!\!\perp \delta\}$ and $\mathcal{A}_2 = \{j \in \mathcal{A} : X_j \not\perp\!\!\!\perp Y|\delta\}$. Then $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$. The following regularity conditions are imposed to facilitate the technical proof, although, they may not be the weakest one.

libel=(C0),itemsep=0pt The characteristic kernel K is bounded.

liibel=(C0),iitemsep=0pt The smoothing kernel $G : \mathbb{R} \to \mathbb{R}$ satisfies $\int_{\mathbb{R}} y^i G(y)dy = I\{i = 0\}$ for $i = 0$ and 1, and $G(y) = O((1 + |y|^4)^{-1})$.

liiibel=(C0),iiitemsep=0pt $h_s \to 0$ and $n_s h_s^2 \to \infty$ as $n_s \to \infty$, for $s = 0, 1$.

livbel=(C0),ivtemsep=0pt The density of Y given $\delta = s$, denoted as $f_{Y,s}(y)$, is bounded away from zero, for $s = 0, 1$. In addition, the first partial derivatives of $f_{Y,s}(y)$ is uniformly bounded by some constant that does not depend on y, for $s = 0, 1$.

lvbel=(C0),vtemsep=0pt There exist $c_1, c_2 > 0$ and $\gamma_1, \gamma_2 \in [0, 1/2)$, such that

$$\min_{j\in\mathcal{A}_1} \Omega_{j,1} \geq 2c_1 n^{-\gamma_1} \text{ and } \min_{j\in\mathcal{A}_2} \Omega_{j,2} \geq 2c_2 n^{-\gamma_2}.$$

lvibel=(C0),vitemsep=0pt There exist $c_3, c_4 > 0$ and $\gamma_3, \gamma_4 \in [0, 1/2)$, such that

$$\min_{j\in\mathcal{A}_1} \Omega_{j,1} - \max_{j\notin\mathcal{A}_1} \Omega_{j,1} \geq 2c_3 n^{-\gamma_3} \text{ and } \min_{j\in\mathcal{A}_2} \Omega_{j,2} - \max_{j\notin\mathcal{A}_2} \Omega_{j,2} \geq 2c_4 n^{-\gamma_4}.$$

Condition i is satisfied for many popular kernels [2]. Conditions ii-iv are commonly assumed for Nadaraya-Watson estimators. Condition v and vi are also standard in the literature of variable screening requiring that the true signal is detectable and is distinguishable from noise.

[Sure Screening] Under conditions i-v,

$$P\left(\mathcal{A} \subset \widehat{\mathcal{A}}\right) \geq 1 - O\left(|\mathcal{A}| \exp\left\{-an^{1-2(\gamma_1 \vee \gamma_2)} + \log n\right\}\right),$$

where, $a > 0$ is some constant.

[Rank Consistency] Under conditions i-iv and vi,

$$\liminf_{n\to\infty} \left\{\min_{j\in\mathcal{A}_1} \widehat{\Omega}_{j,1} - \max_{j\notin\mathcal{A}_1} \widehat{\Omega}_{j,1}\right\} > 0 \text{ and } \liminf_{n\to\infty} \left\{\min_{j\in\mathcal{A}_2} \widehat{\Omega}_{j,2} - \max_{j\notin\mathcal{A}_2} \widehat{\Omega}_{j,2}\right\} > 0$$

almost surely for $\log p = o(n^{1-2(\gamma_3 \vee \gamma_4)})$.

Proofs of Theorems 2.3 and 2.3 appear in Web Supplement 6.2 and 6.3, respectively. Theorem 2.3 suggests that all the important features are selected asymptotically almost surely, and Theorem 2.3 further indicates that active features can be well separated from inactive ones. Both properties hold with NP-dimensionality $\log p = o(n^{1-2\gamma})$ for some $\gamma \in [0, 1/2)$.

There is no established way of determining the threshold values in a finite sample setting. As it is commonly assumed that the cardinality of the truly important set is small, one may specify a model size $d < n$ and select $\widehat{\mathcal{A}}^*(d) = \widehat{\mathcal{A}}_1^*(d_1) \cup \widehat{\mathcal{A}}_2^*(d_2)$, where

$$\widehat{\mathcal{A}}_1^*(d_1) = \{1 \leq j \leq p : \Omega_{j,1} \text{ is among the first } d_1 \text{ largest of all}\},$$
$$\widehat{\mathcal{A}}_2^*(d_2) = \{j \notin \widehat{\mathcal{A}}_1^*(d_1) : \Omega_{j,2} \text{ is among the first } d_2 \text{ largest of all}\},$$

for $d_1 + d_2 = d$. Typical choices of d are $[n/\log(n)]$, $2[n/\log(n)]$, $3[n/\log(n)]$, and $n - 1$ [14, 32]. We can simply set $d_1 = d_2 = [d/2]$, in which case the marginal and conditional utility measures are equally weighted in the selection of $\widehat{\mathcal{A}}^*$. Let $\{r_j^M\}_{j=1}^p$ and $\{r_j^C\}_{j=1}^p$ be the two rankings of variables by $\{\Omega_{j,1}\}_{j=1}^p$ and $\{\Omega_{j,2}\}_{j=1}^p$, respectively. A joint ranking $\{r_j\}_{j=1}^p$ can be acquired by ascending $(r_j^M \wedge r_j^C, r_j^M \vee r_j^C)$. Then selecting the top d variables is identical to the trivial choice of $\widehat{\mathcal{A}}^*$ with $d_1 = d_2$. The sure screening property entails that the probability of selecting all the active predictors is close to one when d is sufficiently large. Inevitably, false discoveries can be inflated simultaneously with a generous choice of d. We address this issue in the next two subsections.

2.4 FDR Control via Knockoff

The most important assumption of ultrahigh dimensional problems is the sparsity principle, which assumes that the cardinality of $\mathcal{A}$ is very small compared to p. In most cases, it is very hard, if not impossible, to recover $\mathcal{A}$ exactly without error. Ensuring all the active predictors are selected with high probability in the preceding screening procedure may introduce too much noise to the downstream analysis in the meanwhile. Therefore, a natural interest is to find a balancing trade-off between the sure screening property and the false discovery rate (FDR). In this subsection, we develop a dual selection procedure for right-censored data with FDR controlling using knockoff features.

We say $\widetilde{\mathbf{X}} \in \mathbb{R}^p$ is a knockoff copy of $\mathbf{X}$ if

1. Swapping X_j with $\widetilde{X}_j$ does not change the joint distribution of $(\mathbf{X}, \widetilde{\mathbf{X}})$;

2. $\widetilde{\mathbf{X}} \perp\!\!\!\perp (Y, \delta) | \mathbf{X}$.

The second condition is trivially achieved as long as $\widetilde{\mathbf{X}}$ is constructed without using (Y, δ). However, if the distribution of $\mathbf{X}$ is unknown, how to obtain exact knockoff copies that satisfy the first condition remains elusive. Nonetheless, we may construct approximate second-order knockoff features, such that $(\mathbf{X}, \widetilde{\mathbf{X}})$ is pairwise exchangeable with respect to the first two moments. Suppose $\boldsymbol{\mu} = E(\mathbf{X})$, $\Sigma = Cov(\mathbf{X})$. Mean invariance can be easily achieved by forcing $E(\widetilde{\mathbf{X}}) = \boldsymbol{\mu}$. The second-order pairwise exchangeable condition is equivalent to

$$Cov(\mathbf{X}, \widetilde{\mathbf{X}}) = G, \text{ where } G = \begin{pmatrix} \Sigma & \Sigma - \text{diag}\{\mathbf{h}\} \\ \Sigma - \text{diag}\{\mathbf{h}\} & \Sigma \end{pmatrix},$$

and $\mathbf{h} \in \mathbb{R}^p$ is a vector that makes G a positive semidefinite covariance matrix. Different approaches are available to select $\mathbf{h}$ [4]. For example, one may find $\mathbf{h}$ by solving

$$\text{argmin} \sum_j |1 - h_j|$$

subject to $h_j \geq 0$ and $2\Sigma - \text{diag}\{\mathbf{h}\}$ being positive semidefinite. If we treat $\mathbf{X}$ as fixed [4] and normalize each feature such that the sample covariance $\widehat{\Sigma} = X^T X$ and $\widehat{\Sigma}_{jj} = 1$ with $X \in \mathbb{R}^{n \times p}$ being the data matrix, then the knockoff data matrix $\widetilde{X} \in \mathbb{R}^{n \times p}$ can be obtained by

$$\widetilde{X} = X(I - \widehat{\Sigma}^{-1}\text{diag}\{\mathbf{h}\}) + \widetilde{U}L,$$

where, $\widetilde{U}$ is an $n \times p$ orthonormal matrix that is orthogonal to the span of X, and $L^T L = 2\text{diag}\{\mathbf{h}\} - \text{diag}\{\mathbf{h}\}\widehat{\Sigma}^{-1}\text{diag}\{\mathbf{h}\}$ is a Cholesky decomposition. In a more general Model-$\mathbf{X}$ setting [5] where $\mathbf{X}$ has an unknown distribution, we can generate approximate knockoff features from conditional normal distribution as

$$\widetilde{\mathbf{X}} | \mathbf{X} \sim N\left(\mathbf{X} - \text{diag}\{\mathbf{h}\}\Sigma^{-1}(\mathbf{X} - \boldsymbol{\mu}), 2\text{diag}\{\mathbf{h}\} - \text{diag}\{\mathbf{h}\}\Sigma^{-1}\text{diag}\{\mathbf{h}\}\right). \tag{8}$$

Note that, if $\mathbf{X}$ is Gaussian, the equivalence of the first two moments implies the equivalence of the joint distribution, such that (8) yields exact knockoff features.

Then, we quantify the contribution of X_j to (Y, δ) by the following two measures:

1. $W_{j,1} = \Omega_{j,1} - \widetilde{\Omega}_{j,1} = R_K^2(X_j|\delta) - R_K^2(\widetilde{X}_j|\delta)$;
2. $W_{j,2} = \Omega_{j,2} - \widetilde{\Omega}_{j,2} = R_K^2(X_j|Y;\delta) - R_K^2(\widetilde{X}_j|Y;\delta)$.

Given sample data $\{\mathbf{X}_i, Y_i, \delta_i\}_{i=1}^n$, we estimate $W_{j,1}$ by $\widehat{W}_{j,1} = \widehat{\Omega}_{j,1} - \widehat{\widetilde{\Omega}}_{j,1}$, where $\widehat{\Omega}_{j,1}$ and $\widehat{\widetilde{\Omega}}_{j,1}$ are calculated using (6). Similarly, we estimate $W_{j,2}$ by $\widehat{W}_{j,2} = \widehat{\Omega}_{j,2} - \widehat{\widetilde{\Omega}}_{j,2}$ using (7). Intuitively, a large value of either $\widehat{W}_{j,1}$ or $\widehat{W}_{j,2}$ indicates the significance of X_j as X_j outperforms $\widetilde{X}_j$. On the other hand, it is expected that irrelevant variables behave similarly to their knockoff counterparts, resulting in small sample utilities that bounce around 0 as shown in the following proposition.

Proposition 3. *Let $\widetilde{\mathbf{X}}$ be an exact knockoff copy of $\mathbf{X}$. Then, for $j \notin \mathcal{A}$,*

1. $W_{j,1} = W_{j,2} = 0$;
2. *Conditioning on* $\left\{\left(|\widehat{W}_{j,1}|, |\widehat{W}_{j,2}|\right) : j \notin \mathcal{A}\right\}$, $I\left\{\widehat{W}_{j,1} > 0\right\}$ *and* $I\left\{\widehat{W}_{j,2} > 0\right\} \overset{i.i.d.}{\sim}$ *Bernoulli(0.5),*

where $I\{\cdot\}$ is the indicator function.

For fixed thresholds $t_1, t_2 > 0$, the false discovery proportion is

$$\text{FDP}(t_1, t_2) = \frac{\#\left\{j \notin \mathcal{A} : \widehat{W}_{j,1} \geq t_1 \text{ or } \widehat{W}_{j,2} \geq t_2\right\}}{\#\left\{j : \widehat{W}_{j,1} \geq t_1 \text{ or } \widehat{W}_{j,2} \geq t_2\right\}},$$

and $\text{FDR}(t_1, t_2) = E[\text{FDP}(t_1, t_2)]$. Note, from Proposition 3,

$$\begin{aligned}\#\left\{j \notin \mathcal{A} : \widehat{W}_{j,1} \geq t_1 \text{ or } \widehat{W}_{j,2} \geq t_2\right\} &\approx \#\left\{j \notin \mathcal{A} : \widehat{W}_{j,1} \leq -t_1 \text{ or } \widehat{W}_{j,2} \leq -t_2\right\} \\ &\leq \#\left\{j : \widehat{W}_{j,1} \leq -t_1 \text{ or } \widehat{W}_{j,2} \leq -t_2\right\},\end{aligned}$$

which leads to a conservative estimator of FDP:

$$\widehat{\text{FDP}}(t_1,t_2)=\frac{1+\#\left\{j:\widehat{W}_{j,1}\leq -t_1 \text{ or } \widehat{W}_{j,2}\leq -t_2\right\}}{\#\left\{j:\widehat{W}_{j,1}\geq t_1 \text{ or } \widehat{W}_{j,2}\geq t_2\right\}}.$$

The proof of Proposition 3 appear in Appendix 6.4. The offset of 1 in the numerator, yielding a slightly more conservative estimator, is necessary both theoretically and empirically to control the FDR. Define $\widehat{W}_{0,1}=\widehat{W}_{0,2}=\infty$ and let $\mathcal{T}=\left\{\left(|\widehat{W}_{j_1,1}|,|\widehat{W}_{j_2,2}|\right):0\leq j_1,j_2\leq p\right\}$. Then $(\mathcal{T},\preceq)$ is a partially ordered set, where $(t_1,t_2)\preceq(t_1',t_2')$ if $t_1\leq t_1'$ and $t_2\leq t_2'$, for $(t_1,t_2),(t_1',t_2')\in\mathcal{T}$. To control FDR at a pre-specified level α, we choose the thresholds $T_{\alpha,1}$ and $T_{\alpha,2}$ as

$$(T_{\alpha,1},T_{\alpha,2})=\min_{\preceq}\left\{(t_1,t_2)\in\mathcal{T}:\widehat{\text{FDP}}(t_1,t_2)\leq\alpha\right\}, \tag{9}$$

where $\min_{\preceq}$ represents the minimal element of the set with respect to $\preceq$. Then, the selected active set is given by

$$\widehat{\mathcal{A}}\left(T_{\alpha,1},T_{\alpha,2}\right)=\left\{j:\widehat{W}_{j,1}\geq T_{\alpha,1} \text{ or } \widehat{W}_{j,2}\geq T_{\alpha,2}\right\}.$$

Note that there can be more than one minimal element in $\mathcal{T}$; so, the choice of $(T_{\alpha,1},T_{\alpha,2})$ may not be unique, leading to different estimates of the active set. In practice, one can choose the minimal element that yields the largest average utility of the selected features, $\text{avg}_{j\in\widehat{\mathcal{A}}(T_{\alpha,1},T_{\alpha,2})}\widehat{W}_j$, where $\widehat{W}_j=\widehat{W}_{j,1}$ if X_j is ranked higher based on the marginal statistic than the conditional statistic and $\widehat{W}_j=\widehat{W}_{j,2}$ vice versa. This approach works well in our simulation studies (see, Section 3).

Although this dual selection procedure controls FDR, it is not readily applicable to ultrahigh dimensional data because constructing knockoff features becomes computationally intractable for large p. However, feature screening and knockoff-based selection naturally complement each other under ultrahigh dimensionality: one can perform screening to reduce p, and then apply the knockoff technique to further control FDR [3, 33]. We elaborate this adaption in the next subsection and show that sure screening is still attainable with FDR under control.

2.5 Refined Screening with FDR Control

Consider splitting n observations into two disjoint groups of size n_1 and $n_2=n-n_1$, denoted as $\{(\mathbf{X}_i^{(1)},Y_i^{(1)},\delta_i^{(1)}):i=1,\ldots,n_1\}$ and $\{(\mathbf{X}_i^{(2)},Y_i^{(2)},\delta_i^{(2)}):i=1,\ldots,n_2\}$. We follow the next two steps:

1. We start with conducting the screening procedure as described in 2.3 using $\{(\mathbf{X}_i^{(1)},Y_i^{(1)},\delta_i^{(1)}):i=1,\ldots,n_1\}$ to select a small index subset of potentially relevant features $\widehat{\mathcal{A}}^*(d)$ for $d<n_2$.
2. Next, we run the knockoff procedure as described in 2.4 on the remaining data $\{(\mathbf{X}_{i,\widehat{\mathcal{A}}^*(d)}^{(2)},Y_i^{(2)},\delta_i^{(2)}):i=1,\ldots,n_2\}$, ignoring features that were not selected in the screening step. Specifically, we first obtain the knockoff matrix $\widetilde{X}_{\widehat{\mathcal{A}}^*(d)}^{(2)}$ for the original design matrix $X_{\widehat{\mathcal{A}}^*(d)}^{(2)}$ on the remaining data. Then, we compute $\widehat{W}_{j,1}$ and $\widehat{W}_{j,2}$ for $j\in\widehat{\mathcal{A}}^*(d)$. For a pre-specified FDR level α, we ultimately select

$$\widehat{\mathcal{A}}\left(T_{\alpha,1},T_{\alpha,2}\right)=\left\{j\in\widehat{\mathcal{A}}^*(d):\widehat{W}_{j,1}\geq T_{\alpha,1} \text{ or } \widehat{W}_{j,2}\geq T_{\alpha,2}\right\},$$

where $T_{\alpha,1}$ and $T_{\alpha,2}$ are determined by solving (9).

Hereafter, we refer to this screening-and-knockoff procedure as α-controlled *k*ernel-based *i*ndependence *d*ual *s*reening (α-KIDS for short). It is critical that the two steps of α-KIDS are performed on distinct data. The following procedure would not control the FDR: we perform the screening step using the full data set to select $\widehat{\mathcal{A}}^*(d)$ and run the knockoff procedure on the dimension-reduced data $\{(\mathbf{X}_{i,\widehat{\mathcal{A}}^*(d)},Y_i,\delta_i):i=1,...,n\}$. The problem is that $\mathbf{X}_{\widehat{\mathcal{A}}^*(d)}$ can be viewed as a function of $(\mathbf{X},Y,\delta)$ because $\widehat{\mathcal{A}}^*(d)$ is selected using all data. As a result, there is no guarantee that $\widetilde{\mathbf{X}}_{\widehat{\mathcal{A}}^*(d)}\perp\!\!\!\perp(Y,\delta)|\mathbf{X}_{\widehat{\mathcal{A}}^*(d)}$, even if $\widetilde{\mathbf{X}}_{\widehat{\mathcal{A}}^*(d)}$ is constructed without using (Y,δ). The loss of FDR control is not merely theoretical; an unimportant feature X_j that is kept by the screening step is generally more likely to appear as a false positive when running the knockoff filter, leading to a much higher FDR [3]. With the data splitting mechanism, as long as the screening step correctly identifies all the relevant features (which happens asymptotically almost surely as shown in Theorem 2.3), the knockoff step will control the FDR as desired. Moreover, the sure screening property is inherited. That is, the α-KIDS procedure achieves a balancing trade-off between type I and type II errors. This appealing property is justified in Theorem 2.5 below.

Denote $\mathcal{E} = \{\mathcal{A} \subseteq \widehat{\mathcal{A}}^*(d)\}$ the event that all the important features are selected in the screening step. We further require that the true signal cannot be too weak to be captured by the knockoff filter:

lvbel=(C0'),vtemsep=0pt There exist $c_5, c_6 > 0$ and $\gamma_5, \gamma_6 \in [0, 1/2)$, such that

$$\min_{j \in \mathcal{A}_1} W_{j,1} \geq 4c_5 n_2^{-\gamma_5} \text{ and } \min_{j \in \mathcal{A}_2} W_{j,2} \geq 4c_6 n_2^{-\gamma_6}.$$

[FDR-Controlled Sure Screening] For any $\alpha \in (0, 1)$, we have

$$\text{FDR} = E\left[\frac{\#\left\{j \notin \mathcal{A} : j \in \widehat{\mathcal{A}}(T_{\alpha,1}, T_{\alpha,2})\right\}}{\#\left\{j : j \in \widehat{\mathcal{A}}(T_{\alpha,1}, T_{\alpha,2})\right\} \vee 1}\middle| \mathcal{E}\right] \leq \alpha.$$

Furthermore, under condition v,

$$P\left(\mathcal{A} \subset \widehat{\mathcal{A}}(T_{\alpha,1}, T_{\alpha,2})\middle| \mathcal{E}\right) \geq 1 - O\left(\exp\left\{-bn_2^{1-2(\gamma_5 \vee \gamma_6)} + 2\log n_2\right\}\right),$$

for $\alpha \geq 1/|\mathcal{A}|$, where $b > 0$ is a constant.

The proof of Theorem 2.5 appear in Appendix 6.5. Although data splitting is a straightforward approach to handle ultrahigh-dimensionality, there is certainly a loss of power since each step only uses part of the data. One solution to the issue is to smartly recycle the data used in the screening step to raise power while retaining the FDR control property for the knockoff procedure [3]. We modify the α-KIDS procedure as follows:

1. The screening step remains the same. Use $\{(\mathbf{X}_i^{(1)}, Y_i^{(1)}, \delta_i^{(1)}) : i = 1, \ldots, n_1\}$ to select a small index subset of potentially relevant features $\widehat{\mathcal{A}}^*(d)$ for $d < n_2$.
2. For the knockoff step, we still start with obtaining the knockoff matrix $\widetilde{X}^{(2)}_{\widehat{\mathcal{A}}^*(d)}$ for the original design matrix $X^{(2)}_{\widehat{\mathcal{A}}^*(d)}$ on the remaining data. Then we concatenate the original design matrix on the first n_1 observations with the knockoff matrix on the next n_2 observations as

$$\widetilde{X}_{\widehat{\mathcal{A}}^*(d)} = \begin{bmatrix} X^{(1)}_{\widehat{\mathcal{A}}^*(d)} \\ \widetilde{X}^{(2)}_{\widehat{\mathcal{A}}^*(d)} \end{bmatrix}.$$

   Now, we calculate $\widehat{W}_{j,1}$ and $\widehat{W}_{j,2}$ using the full data $\left\{\left(\mathbf{X}_{i,\widehat{\mathcal{A}}^*(d)}, \widetilde{\mathbf{X}}_{i,\widehat{\mathcal{A}}^*(d)}, Y_i, \delta_i\right) : i = 1, \ldots, n\right\}$ for $j \in \widehat{\mathcal{A}}^*(d)$, where $\widetilde{\mathbf{X}}_{i,\widehat{\mathcal{A}}^*(d)}$ is the ith row of the knockoff matrix $\widetilde{X}_{\widehat{\mathcal{A}}^*(d)}$. For a pre-specified FDR level α, we ultimately select

$$\widehat{\mathcal{A}}(T_{\alpha,1}, T_{\alpha,2}) = \left\{j \in \widehat{\mathcal{A}}^*(d) : \widehat{W}_{j,1} \geq T_{\alpha,1} \text{ or } \widehat{W}_{j,2} \geq T_{\alpha,2}\right\},$$

   where $T_{\alpha,1}$ and $T_{\alpha,2}$ are determined by solving (9).

Here, we follow the convention to treat $\{(Y_i^{(1)}, \delta_i^{(1)})\}_{i=1}^{n_1}$ as fixed when creating $\widetilde{X}^{(2)}_{\widehat{\mathcal{A}}^*(d)}$ in the knockoff step [3]. In other words, although $\widehat{\mathcal{A}}^*(d)$ was selected using the first portion of the data, we think of $X^{(2)}_{\widehat{\mathcal{A}}^*(d)}$ as being independent of $\{(Y_i^{(1)}, \delta_i^{(1)})\}_{i=1}^{n_1}$. As a result, $\widetilde{X}_{\widehat{\mathcal{A}}^*(d)}$ gives legitimate knockoff features and the procedure controls the directional FDR. On the other hand, there is an inherent gain of power compared to the data splitting approach as the first n_1 observations weigh in. If a feature X_j is important, the first portion of data will contribute to large R^2 or partial R^2 values for both X_j and $\widetilde{X}_j$ since $X_j^{(1)} = \widetilde{X}_j^{(1)}$ by design, and the second portion of data will help separate X_j from its knockoff counterpart, resulting in a positive value of $W_{j,1}$ or $W_{j,2}$.

3 Simulation Studies

In this section, We evaluate the performance of our method on simulated ultra-high dimensional datasets, and make comparisons with several other competing methods, including censored rank independence screening (CRIS) [40], integrated powered density (IPOD) [26], and robust censored distance correlation screening (RCDCS) [7]. Our method is conducted with the Gaussian kernel being the reproducing kernel as well as the smoothing kernel for density estimation. The bandwidths of the two Gaussian kernels are set to heuristic median pairwise distance [22] and $h = 1.06\widehat{\sigma}n^{-1/5}$, where n is the sample size and $\widehat{\sigma}$ is the sample standard deviation [39]. We generate correlated features $\mathbf{X}$ from $N_p(\mathbf{0}, \Sigma)$ with $p = 5,000$ and Σ having a first-order autoregressive (AR) structure, and consider a variety of survival models under independent or dependent censoring. The design of correlated features is to mimic the phenomenon that features tend to be correlated, even purely by chance, in ultrahigh dimensional space [15], which makes it more challenging to distinguish truly important features from spurious ones and achieve exact feature selection. We report the following results based on 200 replicates:

- the τ^{th} quantiles of the minimum model size (MMS), denoted as M_τ, that includes all active features for the screening methods, where the MMS for KIDS is defined as $\min\{M_1+M_2\}$ such that $\mathcal{A} \subseteq \widehat{\mathcal{A}}_1^*(M_1) \cup \widehat{\mathcal{A}}_2^*(M_2)$;
- the proportion of selecting a certain active predictor X_j, denoted as P_j, and the proportion of including all active predictors, denoted as $P_{\mathcal{A}}$, for all the screening methods and α-KIDS;
- the average model size (AMS) determined by α-KIDS;
- and empirical FDR (EFDR) for α-KIDS.

Example 1. In this example, we evaluate the efficacy of KIDS in comparison to the other screening methods. Let $\mathbf{X} \sim N_p(\mathbf{0}, \Sigma)$, where $\Sigma = AR(0.5)$. Given $\mathbf{X}$, the true survival time is generated from the following accelerated failure time (AFT) model and proportional hazard (PH) model:

1. **Model 1:** $\log T = X_1 + X_2 + 1.5X_7^2 + \epsilon$, where $\epsilon \sim N(0,1)$ independently;
2. **Model 2:** $\log(.5(e^{2T} - 1)) = X_1 + X_2 + 1.5X_7^2 + \epsilon$, where ϵ follows the standard extreme value distribution independently, which corresponds to a PH model [27].

For each model, the survival time is subject to two censoring mechanisms:

(a) independent censoring time C generated from uniform distribution on $[0, c_0]$;

(b) dependent censoring time C generated from exponential distribution with mean $c_0 e^{X_1}$,

where, the constant c_0 is chosen to achieve 30% or 50% censoring rate (CR).

The results are summarized in Table 1 for $n = 200$ and $d = [n/\log n] = 38$. In all scenarios, KIDS outperforms the other methods with higher selection proportions, and minimum model sizes closer to the truth, i.e., $|\mathcal{A}| = 3$. The three competitors are not as robust to heavy censoring or dependent censoring as KIDS. In addition, CRIS barely detects the feature (X_7) that is non-linearly related to the endpoint. In the Supplements, we further consider a linear design with varying signal strength (Example 3), and a more complex nonlinear design (Example 4) for both AFT- and PH-type of models under varying censoring mechanisms. Once again, our method performs consistently well compared to the other methods.

Example 2. This example is to verify Theorem 2.5 for the α-KIDS procedure. Similar to Example 1, we generate $\mathbf{X}$ from $N(\mathbf{0}, \Sigma)$ with $\Sigma = AR(0.3)$ and simulate the true survival time from the following two models:

1. **Model 3:** $\log T = \mu(\mathbf{X}) + \epsilon$, where $\mu(\mathbf{X}) = 1.2X_1 + 1.1X_2 + 1.5X_3^2 + X_4 + X_5/\log(1 + |X_5|) + X_6 + 1.5\sin(.5X_7) + X_8 + X_9 + 1.1X_{10}$ and $\epsilon \sim N(0,1)$, independently;
2. **Model 4:** $\log(.5(e^{2T} - 1)) = \mu(\mathbf{X}) + \epsilon$, where $\mu(\mathbf{X})$ is the same as in Model 3 and ϵ follows the standard extreme value distribution, independently.

The censoring time is simulated from:

(a) uniform distribution on $[0, c_0]$;

(b) exponential distribution with mean $c_0 e^{X_1+X_2-X_{10}}$,

where, the constant c_0 is chosen to yield 30% or 50% CR.

Table 1: Quantiles of MMS (M_τ) and selection proportions (P_j's and $P_{\mathcal{A}}$) for models in **Example 1** based on 200 replicates with $n = 200$, $p = 5000$ and $d = [n/\log n] = 38$.

Model	CR	Method	$M_{5\%}$	$M_{25\%}$	$M_{50\%}$	$M_{75\%}$	$M_{95\%}$	P_1	P_2	P_7	$P_{\mathcal{A}}$
1(a)	30%	KIDS	3.0	3.0	3.0	3.0	6.0	0.995	1.000	1.000	0.995
		CRIS	237.8	1255.2	2531.0	3613.0	4640.8	1.000	1.000	0.005	0.005
		IPOD	3.0	3.0	7.0	23.0	143.4	0.955	0.965	0.885	0.810
		RCDCS	3.0	4.0	8.0	17.0	57.0	1.000	1.000	0.915	0.915
	50%	KIDS	3.0	3.0	3.0	3.0	4.0	1.000	1.000	1.000	1.000
		CRIS	460.7	1670.8	3001.0	4157.8	4838.1	0.560	0.530	0.005	0.000
		IPOD	3.0	3.0	6.0	16.0	91.3	0.995	1.000	0.895	0.890
		RCDCS	6.0	18.0	36.0	98.8	359.0	1.000	1.000	0.525	0.525
1(b)	30%	KIDS	3.0	3.0	3.0	3.0	3.0	1.000	1.000	1.000	1.000
		CRIS	103.2	1009.5	2265.0	3348.8	4754.6	1.000	1.000	0.035	0.035
		IPOD	3.0	4.0	11.5	56.0	315.1	1.000	0.920	0.745	0.675
		RCDCS	4.0	9.0	19.0	45.0	120.4	1.000	1.000	0.715	0.715
	50%	KIDS	3.0	3.0	3.0	3.0	4.0	1.000	1.000	1.000	1.000
		CRIS	173.0	1116.8	2086.5	3460.0	4623.5	0.950	0.970	0.005	0.005
		IPOD	3.0	6.0	16.0	59.5	352.2	0.970	0.925	0.760	0.680
		RCDCS	14.0	87.5	196.0	390.0	1180.0	1.000	1.000	0.110	0.110
2(a)	30%	KIDS	3.0	3.0	3.0	3.0	4.0	1.000	1.000	0.995	0.995
		CRIS	144.3	793.8	1994.5	3590.0	4574.0	1.000	1.000	0.020	0.020
		IPOD	3.0	3.0	3.0	3.2	15.1	1.000	1.000	0.975	0.975
		RCDCS	4.0	10.0	23.5	73.0	213.3	1.000	1.000	0.650	0.650
	50%	KIDS	3.0	3.0	3.0	4.0	14.1	1.000	1.000	0.965	0.965
		CRIS	291.8	1373.2	2671.0	3921.2	4663.0	0.975	0.965	0.015	0.015
		IPOD	3.0	3.0	4.0	16.0	133.2	1.000	0.990	0.880	0.870
		RCDCS	11.9	53.8	111.0	297.0	922.2	1.000	0.975	0.190	0.185
2(b)	30%	KIDS	3.0	3.0	3.0	3.0	3.0	1.000	1.000	1.000	1.000
		CRIS	99.6	904.5	2053.0	3530.8	4716.9	1.000	1.000	0.030	0.030
		IPOD	3.0	3.0	3.0	4.0	15.0	1.000	1.000	0.980	0.980
		RCDCS	7.0	28.5	78.0	169.2	508.1	1.000	1.000	0.335	0.335
	50%	KIDS	3.0	3.0	3.0	4.0	9.0	1.000	1.000	0.990	0.990
		CRIS	05.7	604.2	2016.5	3427.8	4630.7	1.000	1.000	0.035	0.035
		IPOD	3.0	3.0	4.0	13.0	56.4	0.995	1.000	0.920	0.915
		RCDCS	26.9	139.2	374.0	985.8	2630.1	1.000	1.000	0.080	0.080

We set $n = 2,000$, $n_1 = 500$, $n_2 = 1,500$, $d = 100$ and vary the nominal level α from .1 to .3[3, 33]. We report the overall selection proportion $P_{\mathcal{A}}^{KIDS}$ for the screening step, whereas, for the knockoff step, we report P_j's, $P_{\mathcal{A}}$, AMS and EFDR given $\mathcal{E} = \{\mathcal{A} \subseteq \widehat{\mathcal{A}}^*(d)\}$. The results are summarized in Table 2. Despite the models involve linear and nonlinear terms of correlated features, the α-KIDS procedure in general controls the FDR at the desired level fairly well and inherits the sure screening property across different censoring settings. Note, at $\alpha = .1$, the procedure has to precisely identify $\mathcal{A}$ (in theory) to control the FDR and maintain power simultaneously because the FDP is exactly .1 with $|\mathcal{A}| = 10$ – a challenging borderline scenario. If $\alpha < .1$, with high probability, the procedure ends up with an empty set.

The choices of n_1/n_2 ratio and the model size d for the screening step in our setting appear to give a favorable balance between finding a sufficiently good screened set at the first stage, and retaining a large enough sample size for powerful inference in the second stage. In practice, we also suggest a split with $n_2 > n_1$ to allow more information for accurate selection via knockoff. On the other hand, d cannot be too small to ensure the coverage of $\mathcal{A}$ for the screening step and to provide adequate amount of noise as reference to control FDR for the knockoff step, while the computational cost for knockoff may prevent us from choosing an arbitrarily large d. Whether we can determine theoretically the optimal split and model size for making the most discoveries is worthy of future research.

Table 2: Selection proportions ($P_{\mathcal{A}}^{KIDS}$, P_j's and $P_{\mathcal{A}}$), AMS and EFDR for models in **Example 2** based on 200 replicates with $n_1 = 500$, $n_2 = 1500$, $p = 5000$ and $d = 100$.

Model	CR	$P_{\mathcal{A}}^{KIDS}$	α	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	$P_{\mathcal{A}}$	AMS	EFDR
3(a)	30%	0.970	0.10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	11.546	0.116
			0.15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	12.335	0.161
			0.20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	12.711	0.176
			0.25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	13.629	0.227
			0.30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	14.093	0.237
	50%	0.980	0.10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	11.561	0.120
			0.15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	12.327	0.159
			0.20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	12.643	0.173
			0.25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	13.622	0.225
			0.30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	14.214	0.248
3(b)	30%	0.985	0.10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	11.523	0.117
			0.15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	12.041	0.148
			0.20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	12.406	0.168
			0.25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	13.310	0.221
			0.30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	13.888	0.241
	50%	0.995	0.10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	11.417	0.108
			0.15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	12.201	0.152
			0.20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	12.397	0.162
			0.25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	13.236	0.212
			0.30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	13.618	0.223
4(a)	30%	0.890	0.10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	11.546	0.113
			0.15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	12.382	0.163
			0.20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	12.652	0.176
			0.25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	13.438	0.223
			0.30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	14.011	0.242
	50%	0.930	0.10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	11.452	0.112
			0.15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	12.134	0.150
			0.20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	12.360	0.161
			0.25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	13.323	0.218
			0.30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	13.930	0.236
4(b)	30%	0.960	0.10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	11.604	0.121
			0.15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	12.281	0.160
			0.20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	12.578	0.175
			0.25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	13.432	0.223
			0.30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	13.880	0.236
	50%	0.850	0.10	0.994	0.994	0.994	0.988	0.994	0.988	0.994	0.994	0.994	0.994	0.982	11.594	0.125
			0.15	1.000	1.000	1.000	0.988	1.000	0.994	1.000	1.000	1.000	1.000	0.982	12.353	0.164
			0.20	1.000	1.000	1.000	0.988	1.000	1.000	1.000	1.000	1.000	1.000	0.988	12.729	0.185
			0.25	1.000	1.000	1.000	0.994	1.000	1.000	1.000	1.000	1.000	1.000	0.994	13.500	0.230
			0.30	1.000	1.000	1.000	0.988	1.000	1.000	1.000	1.000	1.000	1.000	0.988	13.853	0.239

4 Application: Head And Neck Cancer Data

We investigated the head and neck squamous cell carcinoma (HNSCC) cohort in the Cancer Genome Atlas (TCGA) network. Upper quartile normalized RSEM TPM mRNA expression values for 518 primary-solid tumor samples with matched clinical information were obtained using the `R` package `curatedTCGAData`. Genes with low expression, as indicated by a zero interquartile range, were excluded from the analysis. The remaining genes were log-transformed. The endpoint of interest in our study was the number of days to death, which was subject to right censoring either due to loss of follow-up or no event occurrence until the end of the study. The observed survival time ranged between 2 to 6417 days with a median of 649.5 days and with 57.53% censoring rate. For external validation of our findings, gene expressions and clinical data for 253 HNSCC primary tumor samples were acquired from the Gene Expression

Omnibus (GEO) database (accession number GSE65858 [47]). After preliminary data processing, a total of 15,887 genes commonly available in the TCGA and the GSE65858 datasets, along with important clinical covariates (as summarized in Table 8 in the Web Supplement), were included for our analysis. The `R` package `DSFDRC` available in `GitHub` (`https://github.com/urmiaf/DSFDRC`) implements the methodology.

4.1 Model Selection

The TCGA samples were partitioned into training and testing subgroups using a 4:1 ratio. The training cohort consisted of 414 samples, while the testing cohort had 104 samples, and both cohorts shared the same censoring proportion. We employed various competing methods on the training data to identify prognostic gene signatures for HNSCC, and subsequently evaluated their performance using the testing samples. The models were as follows:

- α-KIDS, with specific parameters set to $n_1 = 200$, $d = 100$ and $\alpha = 0.1$, followed by a Cox proportional hazard model built on the selected genes;
- KIDS/CRIS/IPOD/RCDCS to pre-select $d = [104/\log(104)] = 22$ candidate genes, followed by a penalized Cox model (CoxNet) applied on the dimension-reduced data for further gene selection and prognostic modeling;
- α-KIDS followed by a Cox gradient boosting machine (CoxGBM [18]) built on the selected genes;
- KIDS/CRIS/IPOD/RCDCS to pre-select $d = 22$ candidate genes, followed by the double-slicing assisted procedure (DS [9]) for further gene selection and finally a prognostic CoxGBM applied on the dimension-reduced data after the screening and selection steps.

To ensure a fair comparison between α-KIDS (which performs both screening and selection), and the screening-only procedures (KIDS, CRIS, IPOD, and RCDCS), we augmented the screening procedures with more precise selection techniques, namely CoxNet, a model-based approach, and DS, a model-free method. The DS procedure identifies low-dimensional sparse linear combinations of features $\Gamma^T\mathbf{X}$, such that $(Y,\delta) \perp\!\!\!\perp \mathbf{X}|\Gamma^T\mathbf{X}$, where Γ is a $p \times q$ matrix with q (the number of linear combinations) usually being much smaller than p. It achieves simultaneous feature selection through regularization without assuming any parametric distribution of (Y,δ) or linear relation between (Y,δ) and $\Gamma^T\mathbf{X}$. For the purpose of gene selection, we only leveraged the ability of DS to extract relevant genes rather than utilizing the linear combinations it produced. Both linear Cox model and nonlinear CoxGBM were used to construct prognostic signatures based on the selected genes. Optimal tuning parameters for CoxNet, DS, and CoxGBM were determined through cross-validation.

A patient's gene signature loading was calculated as the linear predictor for the fitted Cox/CoxNet model, or the link function value of the fitted CoxGBM model, which can also be viewed as a risk score. Subsequently, patients were classified into high-risk and low-risk groups, using the median risk score of the training cohort as the cutoff. The log-rank tests were conducted to compare the survival functions of the two risk groups and the p-values are reported in Table 3. Furthermore, we assessed the gene signatures by the time-dependent dynamic receiver operating characteristic (ROC) curves [24] at 1, 3 and 5 years. The corresponding area under curve (AUC) values are summarized in Table 3. According to the log-rank tests and the ROC curves, relevant genes selected by α-KIDS and KIDS led to more informative prognostic signatures for HNSCC in terms of risk stratification and survival prediction. Notably, the α-KIDS+CoxGBM model demonstrated the most favorable overall performance on the testing samples. We then proceeded to refit the model to the entire TCGA dataset, and validated the resulting gene signature with the external data.

Table 3: Gene signature sizes (number of genes in a signature), p-values for the log-rank tests on the risk stratification of the TCGA samples, and AUCs for 1-, 3-, 5-year ROC curves of the risk scores across competing methods.

Model	**Size**	**Training**				**Testing**			
		Log-rank p-value	**1-year AUC**	**3-year AUC**	**5-year AUC**	**Log-rank p-value**	**1-year AUC**	**3-year AUC**	**5-year AUC**
α-KIDS+Cox	11	<0.001	0.676	0.664	0.649	0.013	0.597	0.604	0.580
KIDS+CoxNet	12	<0.001	0.703	0.685	0.669	0.091	0.604	0.617	0.610
CRIS+CoxNet	9	<0.001	0.609	0.575	0.567	0.320	0.526	0.521	0.510
IPOD+CoxNet	18	<0.001	0.687	0.674	0.653	0.102	0.595	0.587	0.577
RCDCS+CoxNet	12	<0.001	0.663	0.645	0.663	0.750	0.545	0.543	0.542
α-KIDS+CoxGBM	11	<0.001	0.740	0.719	0.685	0.004	0.633	0.624	0.598
KIDS+DS+CoxGBM	7	<0.001	0.745	0.707	0.689	0.016	0.604	0.606	0.600
CRIS+DS+CoxGBM	3	<0.001	0.583	0.579	0.569	0.235	0.513	0.511	0.507
IPOD+DS+CoxGBM	11	<0.001	0.757	0.736	0.710	0.069	0.573	0.559	0.547
RCDCS+DS+CoxGBM	6	<0.001	0.686	0.658	0.674	0.080	0.568	0.559	0.565

4.2 10-gene Signature and External Validation

The α-KIDS procedure was applied to the full TCGA dataset, resulting in the selection of 10 genes: OLR1, SPOCK1, DDX19A, FADS3, P2RX6, C9ORF4, C15ORF21, TMED6, TFB2M and C22ORF15. A 10-gene prognostic signature was constructed using CoxGBM subsequently. The GEO platform data was used to validate the effectiveness of the gene signature. Patients were classified as having a high-risk gene signature or a low-risk gene signature on the basis of the link function values, with the median score of the TCGA samples as the threshold. Patients with a high-risk 10-gene signature exhibited significantly lower median survival compared to those with a low-risk gene signature in both the TCGA cohort (727 days vs. 2717 days) and the validation cohort (1068 days vs. 1962 days), as supported by the Kaplan-Meier curves and the log-rank test p-values (Figure 1). An interesting finding is that patients with HPV infection were associated with lower risk scores in both cohorts (p-values < 0.001 based on two-sample t-tests). This aligns with previous reports that patients with HPV-positive cancers generally experience better prognoses than those with HPV-negative cancers, particularly for tumors arising in the oropharynx [29]. Therefore, the gene signature may offer insights into the underlying molecular mechanisms of the HPV heterogeneity.

Additionally, multivariate Cox proportion-hazards regression analysis was used to evaluate independent prognostic factors associated with survival, and the 10-gene signature, age, sex, tumor stage, HPV status, alcohol history and smoking history were used as covariates. The fitted models are summarized in Table 4. For the TCGA cohort, the 10-gene signature was a strong predictor with an hazard ratio of 7.62 (p-value < 0.001), after adjusting for other clinical covariates. There was a 2% increase in the expected hazard relative to a one year increase in age (p-value < 0.001). Patients with IV-stage cancer experienced a remarkable 98% increase in hazard (p-value < 0.001) compared to those in early stages (I/II). Similar results were observed in the validation cohort, indicating the potential clinical utility of the 10-gene signature in enhancing prognostic assessments and guiding personalized treatment decisions for HNSCC patients beyond conventional phenotype-based predictors.

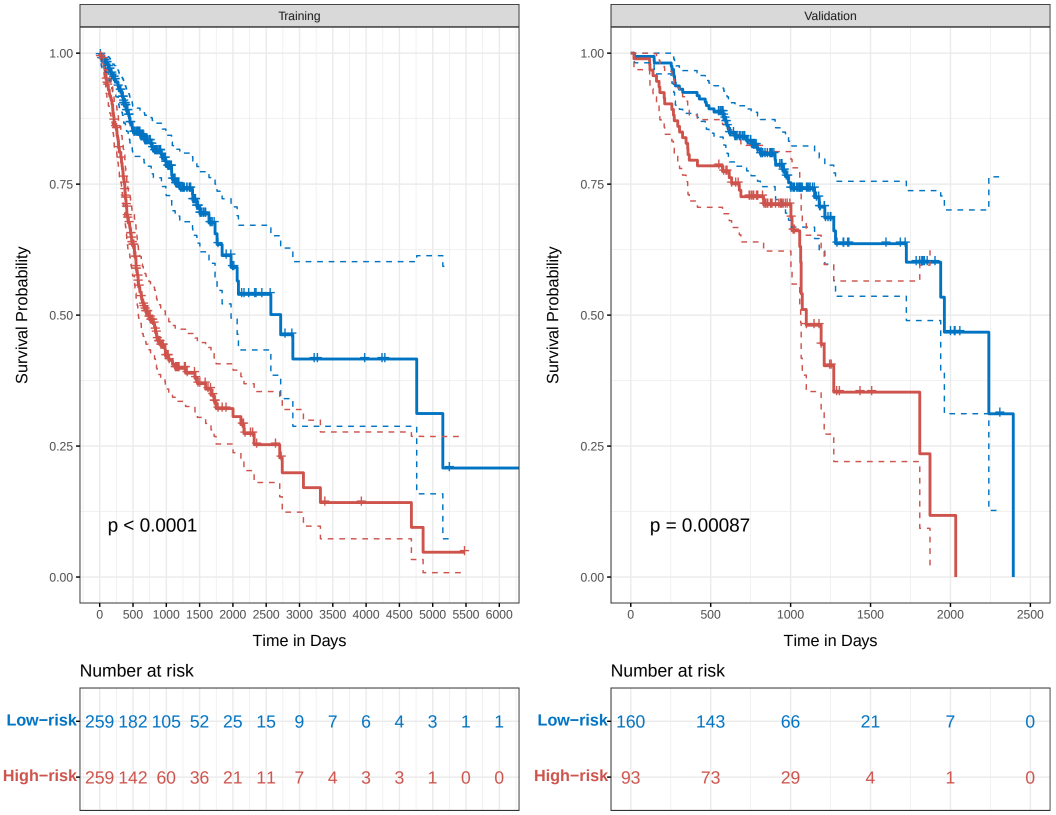

Figure 1: Kaplan–Meier estimates of overall survival (solid) with 95% confidence interval (dash) and log-rank test p-values for risk stratification of training (TCGA) and validation (GEO) samples according to the 10-gene signature identified by the α-KIDS+CoxGBM model.

Table 4: Multivariate Cox regression analysis based on the 10-gene signature and other clinical covariates. CI denotes confidence interval.

Variable	**Training (TCGA)**		**Validation (GEO)**	
	Hazard Ratio (95% CI)	**P-value**	**Hazard Ratio (95% CI)**	**P-value**
10-gene signature	7.62 (4.59, 12.66)	<0.001	3.39 (1.69, 6.79)	<0.001
Age	1.02 (1.01, 1.04)	<0.001	1.03 (1.00, 1.05)	0.030
Gender				
Male	1.00 (0.70, 1.43)	0.989	0.92 (0.50, 1.67)	0.773
Female	-	-	-	-
Tumor stage				
III	1.55 (0.92, 2.60)	0.099	1.07 (0.32, 3.57)	0.915
IV	1.98 (1.32, 2.98)	<0.001	5.34 (2.29, 12.47)	<0.001
I/II	-	-	-	-
HPV status				
Positive	1.53 (0.96, 2.47)	0.353	0.51 (0.29, 0.91)	0.021
Negative	-	-	-	-
Alcohol history				
Yes	0.86 (0.61, 1.19)	0.915	1.19 (0.57,2.50)	0.639
No	-	-	-	-
Smoking history				
Yes	1.02 (0.70, 1.50)	0.854	0.99 (0.52,1.90)	0.977
No	-	-	-	-

Table 5: P-values for the log-rank tests on the risk stratification, and AUCs for 1-, 3-, 5-year ROC curves of the risk scores across three competing CoxGBMs.

Variables	**Model size**	**Training**				**Validation**			
		Log-rank p-value	**1-year AUC**	**3-year AUC**	**5-year AUC**	**Log-rank p-value**	**1-year AUC**	**3-year AUC**	**5-year AUC**
Clinical-only	6	<0.001	0.626	0.606	0.611	0.012	0.642	0.605	0.613
Genetic-only	10	<0.001	0.670	0.674	0.664	<0.001	0.621	0.619	0.579
Composite	16	<0.001	0.720	0.678	0.665	<0.001	0.698	0.693	0.620

4.3 Integrated Clinicogenomics Modeling

Results from the above analysis motivated us to consider a CoxGBM combining the 10 genes selected by α-KIDS and important clinical covariates, namely age, sex, tumor stage, HPV status, alcohol history and smoking history. The composite model was compared against two other CoxGBM models: the one based solely on the 10 selected genes (which was used to discover the 10-gene signature in the previous subsection) and the other based solely on the clinical covariates. Again, the TCGA cohort was utilized as the training data and the GEO platform data served as the external validation data. Differences in survival between the high-risk group and the low-risk group were analyzed with the log-rank test. ROC curves and associated AUCs were calculated to assess time-dependent predictive performance of the three models. The results, as summarized in Table 5, revealed that the model integrating both clinical and genetic information had improved prognostic accuracy over the other two models.

Finally, we highlight some biological implications of the genes selected by α-KIDS. OLR1 is a scavenger receptor for oxidized low-density lipoprotein (LDL) on endothelial cells and other cell types. OLR1 up-regulation in different tumors has evidenced its involvement in cancer onset, progression and metastasis, including HNSCC [36, 50]. High expression of FADS3, located at the cancer genomic hotspot 11q13 locus, has been reported to predict poor prognosis in HNSCC [41]. The oncogenic functions of SPOCK1, C15orf21, and TMED6 have also been investigated in several cancer cells [38, 17, 44, 11, 49].

5 Discussion

Large scale collaborative effort, such as TCGA, have allowed researchers with access to vast and curated data, enabling investigations into the underlying molecular mechanisms of HNSCC prognosis at various levels of complexity. A notable characteristic of such datasets is their ultra-high dimensionality, which places particular demands on the methods used to build prognostic models - they must be able to handle data where the number of features far exceeds the number of observations. Moreover, in the context of survival analysis, how to handle censoring appropriately is paramount to avoid biased estimations and drawing incorrect conclusions, especially in the presence of heavy censoring. Feature screening emerges as a crucial step to efficiently reduce dimension before undertaking more accurate analyses. However, existing methods often impose explicit or implicit assumptions on censoring that are rather difficult to verify given the large number of features, creating impediments to their practical uses. Our proposed novel feature screening procedure quickly reduces irrelevant information under ultrahigh-dimensional right-censored settings, along with a unified selection procedure to control FDR. The proposed framework requires no pre-specification of the model structure and has the minimal assumption on the censoring mechanism. The flexibility is achieved by direct nonparametric learning of the survival outcome, without the need for intermediate estimation of survival probabilities. Our methodology is also readily generalizable for feature evaluation in other cancer types.

We remark that even if our assumption in equation 3 is not met, our procedure still serves the purpose of feature screening by identifying $\mathcal{A}_{(T,C)}$, the active set for (T, C), jointly, which inherently contains $\mathcal{A}_T$. To further isolate the important features for T from the estimated active set, more precise feature selection methods [9] that are tailored for lower dimensional data, can be further applied. The FDR control step ensures that only the most informative features enter the downstream analyses to construct accurate prognostic models. Although our initial motivation was to address the challenges lying within the TCGA HNSCC dataset, the developed method is generally applicable to devise robust prognostic systems for new patient cohorts and other cancers. Along the line, future research should explore how to account for sample heterogeneity and integrate domain knowledge into the feature screening procedure, especially for HNSCC data encompassing samples from diverse sites and HPV subtypes. Additional avenues for future research include extending the methodology to more complex (cancer) endpoints, such as interval-censoring, and multistate models, etc. Our current exploration only considers patient mRNA genomic information. However, an integrative approach that analyzes and combines multiple -omics data, such as genomic, transcriptomic, and methylome data via identifying and validating a multi-omics signature may enhance HNSCC prognosis[37]. An extension of our current methodology for the multi-omics case, although non-trivial, is relevant.

Supplementary Materials

6 Theorems and Proofs

6.1 Lemmas and Proofs

In this subsection, we first show some useful lemmas as preliminary results for Supplements 6.2 and 6.3.

Lemma 1 (Deviation bound for U-statistics, [25]). *Let $g(\mathbf{U}_1, ..., \mathbf{U}_r)$ be a kernel of a U-statistic U_n, i.e., $U_n := \frac{1}{(n)_r} \sum_{i_r^n} g(\mathbf{U}_{i_1}, .., \mathbf{U}_{i_r})$, where $n > r$, $(n)_r := \frac{n!}{(n-r)!}$ and $\sum_{i_r^n}$ is taken over all r-tuples $\{i_1, ..., i_r\}$ drawn without replacement from $\{1, ..., n\}$. If $b_1 \le g(\mathbf{U}_1, .., \mathbf{U}_r) \le b_2$, then for any $\epsilon > 0$, the following bound holds:*

$$P\left\{|U_n - EU_n| \ge \epsilon\right\} \le 2\exp\{-2w\epsilon^2/(b_2 - b_1)^2\},$$

where $w := [n/r]$, the largest integer contained in n/r.

This lemma gives a uniform bound for any U-statistic of arbitrary dimensional data, as long as the associated kernel is bounded. We repeatedly use this result to prove the next two lemmas.

Lemma 2 (Deviation bound for marginal utilities). *Under condition i, for any $\epsilon \in (0, 1)$,*

$$P\{|\widehat{\Omega}_{j,1} - \Omega_{j,1}| \ge \epsilon\} \le 4\exp\left\{-a_1 n\epsilon^2\right\},$$

where $j = 1, ..., p$, and $a_1 > 0$ is a constant.

Proof. We aim to show the uniform consistency of the denominator and the numerator of $\Omega_{j,1}$ under regularity conditions respectively. Because the denominator of $\Omega_{j,1}$ has a similar form as the numerator, we deal with its numerator only below. Let

$$\begin{aligned}
\widehat{\mathcal{H}} &:= \sum_{s=0,1} \frac{n_s}{n} \frac{1}{n_s^2} \sum_{i_1,i_2=1}^{n_s} K(X_{i_1,j}^{(l)} - X_{i_2,j}^{(l)}) - \frac{1}{n^2} \sum_{i_1,i_2=1}^{n} K(X_{i_1,j} - X_{i_2,j}) \\
&:= \sum_{s-0,1} P_s V_{n_s}^{(s)} - V_n^{(0)},
\end{aligned}$$

where $V_{n_s}^{(s)}$ $(s = 0, 1)$ are V-statistics. Let $U_{n_s}^{(s)}$ $(s = 0, 1)$ be corresponding U-statistics with $E_s := EU_{n_s}^s$ $(s = 0, 1)$. Under condition i, without loss of generality, we assume that the kernel K is bounded above by 1. Hence, $0 \le E_l \le 1$ for $s = 0, 1$. Denote $\mathcal{H} := \sum_{l=1}^{L} P_s E_s - E_0$, where $P_s = P(\delta = s)$. For any $\epsilon \in (0, 1)$,

$$\begin{aligned}
&P\left\{|\widehat{\mathcal{H}} - \mathcal{H}| \ge \epsilon\right\} \\
=&P\left\{\left|\sum_{s=0,1} \hat{P}_s \left(V_{n_s}^{(s)} - E_s\right) + \sum_{s=0,1} \left(\hat{P}_s - P_s\right) E_s - \left(V_n^{(0)} - E_0\right)\right| \ge \epsilon\right\} \\
\le&P\left\{\sum_{s=0,1} \hat{P}_s \left|V_{n_s}^{(s)} - E_s\right| \ge \frac{\epsilon}{3}\right\} + P\left\{\sum_{s=0,1} \left|\hat{P}_s - P_s\right| E_s \ge \frac{\epsilon}{3}\right\} + P\left\{\left|V_n^{(0)} - E_0\right| \ge \frac{\epsilon}{3}\right\} \\
:=&T_1 + T_2 + T_3.
\end{aligned}$$

Let us consider T_1 first.

$$\begin{aligned}
T_1 \leq & P\left\{2\max_s \hat{P}_s\left|V_{n_s}^{(s)} - E_s\right| \geq \frac{\epsilon}{3}\right\} \\
\leq & P\left\{\max_s\left|V_{n_s}^{(s)} - E_s\right| \geq \frac{\epsilon}{6},\ n_s > \frac{P_s n}{2}\right\} \text{ for } n \text{ sufficiently large} \\
\leq & \sum_s P\left\{\left|V_{n_s}^{(s)} - E_s\right| \geq \frac{\epsilon}{6}, n_s > \frac{P_s n}{2}\right\} \\
= & \sum_s P\left\{\left|\frac{n_s - 1}{n_s}U_{n_s}^{(s)} + \frac{1}{n_s}K(0) - E_s\right| \geq \frac{\epsilon}{6}, n_s > \frac{P_s n}{2}\right\} \\
\leq & \sum_s P\left\{\left|U_n^{(s)} - E_s\right| \geq \frac{\epsilon}{6} - \frac{2}{n_s}, n_s > \frac{P_s n}{2}\right\} \\
\leq & \sum_s P\left\{\left|U_n^{(s)} - E_s\right| \geq \frac{\epsilon}{12}, n_s > \frac{P_s n}{2}\right\} \\
\leq & 4\exp\left\{-\frac{P_s n\epsilon^2}{288}\right\},
\end{aligned}$$

where the last inequality follows from Lemma 1. Also,

$$T_2 \leq P\left\{\max_s\left|\hat{P}_s - P_s\right| \geq \frac{\epsilon}{6}\right\} \leq 4\exp\left\{-\frac{2n\epsilon^2}{36}\right\}$$

and $T_3 \leq 2\exp\left\{-\frac{n\epsilon^2}{36}\right\}$. Combining T_1, T_2 and T_3, we have

$$P\left\{\left|\widehat{\mathcal{H}} - \mathcal{H}\right| \geq \epsilon\right\} \leq 4\exp\left\{-a_1 n\epsilon^2\right\},$$

for some $a_1 > 0$.

Lemma 3 (Deviation bound for conditional utilities). *Under conditions i-iv, for any $\epsilon \in (0,1)$,*

$$P\left\{|\widehat{\Omega}_{j,2} - \Omega_{j,2}| \geq \epsilon\right\} \leq 4n\exp\left\{-a_2 n\epsilon^2\right\},$$

where $j = 1, ..., p$, and $a_2 > 0$ is a constant.

Proof. For a given $j \in \{1, ..., p\}$ and $s \in \{0, 1\}$, let $\gamma_s(y) := E(d_{i_2i_3i_4i_5}|Y_{i_2} = y, Y_{i_3} = y, \delta = s)$, where $d_{i_2i_3i_4i_5} = K(X_{i_2j} - X_{i_3j}) - K(X_{i_3j} - X_{i_5j})$, then $\mathcal{H}_s := \mathcal{H}_K^2(X_j|Y; \delta = s) = E\gamma_s(Y)$. The kernel regression estimator

$$\begin{aligned}
\widehat{\mathcal{H}}_s := & \mathcal{H}_{K,G_h,n_s}^2(X_j|Y; \delta = s) \\
= & \frac{1}{n_s^5}\sum_{i_1,\ldots,i_5=1}^{n_s}\frac{G_{i_1i_2}G_{i_1i_3}d_{i_2i_3i_4i_5}}{\hat{f}_{Y,s}^2(y_{i_1})} \\
= & \frac{1}{n_s}\sum_{i_1=1}^{n_s}\frac{f_{Y,s}^2(y_{i_1})}{\hat{f}_{Y,s}^2(y_{i_1})}\widehat{\gamma}_s(y_{i_1}),
\end{aligned}$$

where $G_{i_1i_2} = G_h(Y_{i_1} - Y_{i_2})$, $f_{Y,s}(\cdot)$ is the density function of Y given $\delta = s$, $\hat{f}_{Y,s}(y_{i_1}) := \frac{1}{n_s}\sum_{t=1}^{n_s} G_{i_1t}$ and

$$\widehat{\gamma}_s(y_{i_1}) := \frac{1}{n_s^4}\sum_{i_2,i_3,i_4,i_5=1}^{n_s}\frac{G_{i_1i_2}G_{i_1i_3}d_{i_2i_3i_4i_5}}{f_{Y,s}^2(y_{i_1})}.$$

Without loss of generality, we assume that $f_{Y,s}(y)$ is bounded below by some $L > 0$ in condition iv. We first show some intermediate results.

libel=(R0),leftmirgin=2 $P\{|\hat{f}_{Y,s}(y_{i_1}) - f_{Y,s}(y_{i_1})| \geq \epsilon\} \leq 2\exp\{-n_s\epsilon^2/2\}$.
Note that $\hat{f}_{Y,s}(y_{i_1}) = \frac{1}{n_s h}G(0) + \frac{n_s - 1}{n_s}\left(\frac{1}{n_s - 1}\sum_{t \neq i_1} G_{i_1t}\right)$ and $\frac{1}{n_s h}G(0) = o(1)$ by

conditions ii and iii. Denote $U_{n_s-1} := \frac{1}{n_s-1}\sum_{t\neq i_1} G_{i_1 t}$. Then

$$\begin{aligned} EU_{n_s-1} &= \int h^{-1}G(\frac{y_{i_1}-y}{h})f_{Y,s}(y)dy \\ &= \int G(u)f_{Y,s}(y_{i_1}+hu)du = f_{Y,s}(y_{i_1}) + O(h^2) \end{aligned}$$

by Taylor expansion and conditions ii and iv. Hence,

$$\begin{aligned} &P\left\{\left|\hat{f}_{Y,s}(y_{i_1}) - f_{Y,s}(y_{i_1})\right| \geq \epsilon\right\} \\ =&P\left\{\left|\frac{n_s-1}{n_s}U_{n_s-1} - EU_{n_s-1} + O(h^2)\right| \geq \epsilon\right\} \\ =&P\left\{\left|\frac{n_s-1}{n_s}(U_{n_s-1} - EU_{n_s-1}) + O(h^2)\right| \geq \epsilon\right\} \\ \leq&P\left\{|U_{n_s-1} - EU_{n_s-1}| \geq \frac{\epsilon}{2}\right\} \text{ for } n_s \text{ sufficiently large} \\ \leq&2\exp\left\{-\frac{n_s\epsilon^2}{2}\right\} \text{ by Lemma 1.} \end{aligned}$$

liibel=(R0),leftmiirgiin=2 $P\{|\frac{1}{n_s}\sum_{i_1=1}^{n_s}\widehat{\gamma}_s(y_{i_1}) - \mathcal{H}_s| \geq \epsilon\} \leq 2\exp\{-n_s\epsilon^2/8\}$.

Denote the corresponding U-statistic of $\frac{1}{n_s}\sum_{i_1=1}^{n_s}\widehat{\gamma}_s(y_{i_1})$ as $\widetilde{\mathcal{H}}_s$, that is,

$$\widetilde{\mathcal{H}}_s := C_{n_s}^5 \sum_{i_1<\ldots<i_5}\frac{1}{5!}\sum_{\pi} g_{i_1i_2i_3i_4i_5},$$

where $g_{i_1i_2i_3i_4i_5} := G_{i_1i_2}G_{i_1i_3}d_{i_2i_3i_4i_5}/f_{Y,s}^2(y_{i_1})$ and $\sum_\pi$ represents summation over the $5!$ permutations of $(i_1,...,i_5)$. Under conditions ii and iii, $\frac{1}{n_s}\sum_{i_1=1}^{n_s}\widehat{\gamma}_s(y_{i_1}) = \widetilde{\mathcal{H}}_s + o(1)$. We will show in the next that $E\widetilde{\mathcal{H}}_s = \mathcal{H}_s + o(1)$ in two parts. Firstly,

$$\begin{aligned} \Gamma_1 :=& \int h^{-2}G(\frac{y_{i_1}-y_{i_2}}{h})G(\frac{y_{i_1}-y_{i_3}}{h})K(x_{i_2}-x_{i_3}) \\ &f_{Y,s}^{-1}(y_{i_1})f_{X_jY,s}(x_{i_2},y_{i_2})f_{X_jY,s}(x_{i_3},y_{i_3})dx_{i_2}dx_{i_3}dy_{i_1}dy_{i_2}dy_{i_3} \\ =& \int K(x_{i_2}-x_{i_3})f_{X_j|Y,s}(x_{i_2}|y_{i_1}+hu)f_{X_j|Y,s}(x_{i_3}|y_{i_1}+hv)dx_{i_2}dx_{i_3} \\ &G(u)G(v)f_{Y,s}(y_{i_1}+hu)f_{Y,s}(y_{i_1}+hv)dudvf_{Y,s}^{-1}(y_{i_1})dy_{i_1} \\ =& \int K(x_{i_2}-x_{i_3})f_{X_j|Y,s}(x_{i_2}|y_{i_1})f_{X_j|Y,s}(x_{i_3}|y_{i_1})dx_{i_2}dx_{i_3}f_{Y,s}(y_{i_1})dy_{i_1} + O_p(h^2) \end{aligned}$$

by Taylor expansion and conditions ii and iv. Similarly, we can show

$$\begin{aligned} \Gamma_2 :=& \int K(x_{i_4}-x_{i_5})f_{Y,s}(x_{i_4})f_{Y,s}(x_{i_5})dx_{i_4}dx_{i_5} \\ &h^{-2}G(\frac{y_{i_1}-y_{i_2}}{h})G(\frac{y_{i_1}-y_{i_3}}{h})f_{Y,s}^{-1}(y_{i_1})f_{Y,s}(y_{i_2})f_{Y,s}(y_{i_3})dy_{i_1}dy_{i_2}dy_{i_3} \\ =& \int K(x_{i_4}-x_{i_5})f_{Y,s}(x_{i_4})f_{Y,s}(x_{i_5})dx_{i_4}dx_{i_5}f_{Y,s}(y_{i_1})dy_{i_1} + O_p(h^2). \end{aligned}$$

Therefore, $E\widetilde{\mathcal{H}}_s = \Gamma_1 + \Gamma_2 = \mathcal{H}_s + o(1)$. Then

$$\begin{aligned} &P\left\{\left|\frac{1}{n_s}\sum_{i_1=1}^{n_s}\widehat{\gamma}_s(y_{i_1}) - \mathcal{H}_s\right| \geq \epsilon\right\} \\ =&P\left\{\left|\widetilde{\mathcal{H}}_s - E\widetilde{\mathcal{H}}_s + o(1)\right| \geq \epsilon\right\} \\ \leq&P\left\{\left|\widetilde{\mathcal{H}}_s - E\widetilde{\mathcal{H}}_s\right| \geq \frac{\epsilon}{2}\right\} \text{ for } n_s \text{ sufficiently large} \\ \leq&2\exp\left\{-\frac{n_s\epsilon^2}{8}\right\} \text{ by Lemma 1.} \end{aligned}$$

Now, for arbitrary $\epsilon \in (0, 1)$,

$$\begin{aligned}
&P\left\{|\widehat{\mathcal{H}}_s - \mathcal{H}_s| \geq \epsilon\right\} \\
\leq &P\left\{\left|\frac{1}{n_s}\sum_{i_1=1}^{n_s}\frac{f^2_{Y,s}(y_{i_1})}{\hat{f}^2_{Y,s}(y_{i_1})}\widehat{\gamma}_s(y_{i_1}) - \mathcal{H}_s\right| \geq \epsilon\right\} \\
\leq &P\left\{\left|\frac{1}{n_s}\sum_{i_1=1}^{n_s}\widehat{\gamma}_s(y_{i_1}) - \mathcal{H}_s + \frac{1}{n_s}\sum_{i_1=1}^{n_s}\left(\frac{f^2_{Y,s}(y_{i_1})}{\hat{f}^2_{Y,s}(y_{i_1})} - 1\right)\widehat{\gamma}_s(y_{i_1})\right| \geq \epsilon\right\} \\
\leq &P\left\{\left|\frac{1}{n_s}\sum_{i_1=1}^{n_s}\widehat{\gamma}_s(y_{i_1}) - \mathcal{H}_s\right| \geq \frac{\epsilon}{2}\right\} + P\left\{\left|\frac{1}{n_s}\sum_{i_1=1}^{n_s}\left(\frac{f^2_{Y,s}(y_{i_1})}{\hat{f}^2_{Y,s}(y_{i_1})} - 1\right)\widehat{\gamma}_s(y_{i_1})\right| \geq \frac{\epsilon}{2}\right\} \\
:= &T_1 + T_2.
\end{aligned}$$

By (R2), $T_1 \leq 2\exp\{-n_w\epsilon^2/32\}$. Moreover,

$$\begin{aligned}
T_2 \leq &P\left\{\max_{i_1}\left|\left(\frac{f^2_{Y,s}(y_{i_1})}{\hat{f}^2_{Y,s}(y_{i_1})} - 1\right)\widehat{\gamma}_s(y_{i_1})\right| \geq \frac{\epsilon}{2}\right\} \\
\leq &P\left\{\max_{i_1}\left|\left(\frac{f^2_{Y,s}(y_{i_1})}{\hat{f}^2_{Y,s}(y_{i_1})} - 1\right)\widehat{\gamma}_s(y_{i_1})\right| \geq \frac{\epsilon}{2}, \min_{i_1}\hat{f}_{Y,s}(y_{i_1}) \geq \frac{L}{2}\right\} + P\left\{\min_{i_1}\hat{f}_{Y,s}(y_{i_1}) < \frac{L}{2}\right\} \\
\leq &P\left\{\max_{i_1}\left|\widehat{\gamma}_s(y_{i_1})[f^2_{Y,s}(y_{i_1}) - \hat{f}^2_{Y,s}(y_{i_1})]\right| \geq \frac{L^2\epsilon}{8}\right\} + P\left\{\max_{i_1}\left|f_{Y,s}(y_{i_1}) - \hat{f}_{Y,s}(y_{i_1})\right| \geq \frac{L\epsilon}{2}\right\} \\
:= &T_{21} + T_{22}.
\end{aligned}$$

By (R1), $T_{22} \leq 2n_s\exp\{-L^2n_s\epsilon^2/2\}$. Let $\widehat{m}_s(y_{i_1}) := \widehat{\gamma}_s(y_{i_1})[f^2_{Y,s}(y_{i_1}) - \hat{f}^2_{Y,s}(y_{i_1})]$ and $\widehat{m}^U_s(y_{i_1})$ be the corresponding U-statistic. Similar to (R2), we can show that

$$\widehat{m}_s(y_{i_1}) = \widehat{m}^U_s(y_{i_1}) + o(1) \text{ and } E_{i_2i_3i_4i_5}\widehat{m}^U_s(y_{i_1}) = O(h^2).$$

Hence, for n_s sufficiently large,

$$T_{21} \leq P\left\{\max_{s_1}\left|\widehat{m}^U_s(y_{i_1}) - E_{i_2i_3i_4i_5}\widehat{m}^U_s(y_{i_1})\right| \geq \frac{L^2\epsilon}{16}\right\} \leq 2n_s\exp\left\{-\frac{L^4n_s\epsilon^2}{512}\right\}.$$

Finally, we have

$$P\left\{|\widehat{\mathcal{H}}_s - \mathcal{H}_s| \geq \epsilon\right\} \leq 2n_s\exp\left\{-\tilde{a}_2n_s\epsilon^2\right\},$$

where $\tilde{a}_2$ is some constant depending on L. Consequently,

$$P\left\{|\widehat{\Omega}_{j,2} - \Omega_{j,2}| \geq \epsilon\right\} \leq 4n\exp\left\{-a_2n\epsilon^2\right\},$$

for some $a_2 > 0$.

6.2 Proof of Theorem 2.3

Proof. Following from **Lemma 2** and **3**,

$$P\left\{\max_{j\in\mathcal{A}_1}|\widehat{\Omega}_{j,1} - \Omega_{j,1}| > c_1n^{-\gamma_1}\right\} \leq O\left(|\mathcal{A}_1|\exp\left\{-b_1n^{1-2\gamma_1}\right\}\right),$$

and

$$P\left\{\max_{j\in\mathcal{A}_2}|\widehat{\Omega}_{j,2} - \Omega_{j,2}| > c_2n^{-\gamma_2}\right\} \leq O\left(|\mathcal{A}_2|\exp\left\{-b_2n^{1-2\gamma_2} + \log n\right\}\right).$$

Under condition v, if $\mathcal{A} \not\subseteq \widehat{\mathcal{A}}$, there must exist some $j \in \mathcal{A}_1$ such that $\Omega_{j,1} \geq 2c_1 n^{-\gamma_1}$ or some $j \in \mathcal{A}_2$ such that $\Omega_{j,2} \geq 2c_2 n^{-\gamma_2}$ but $\widehat{\Omega}_{j,1} < c_1 n^{-\gamma_1}$ and $\widehat{\Omega}_{j,2} < c_2 n^{-\gamma_2}$. Therefore,

$$\begin{aligned}
P\left\{\mathcal{A} \not\subseteq \widehat{\mathcal{A}}\right\} \leq & P\left\{\left|\widehat{\Omega}_{j,1} - \Omega_{j,1}\right| > c_1 n^{-\gamma_1} \text{ for some } j \in \mathcal{A}_1\right\} \\
& + P\left\{\left|\widehat{\Omega}_{j,2} - \Omega_{j,2}\right| > c_2 n^{-\gamma_2} \text{ for some } j \in \mathcal{A}_2\right\} \\
\leq & P\left\{\max_{j \in \mathcal{A}_1}\left|\widehat{\Omega}_{j,1} - \Omega_{j,1}\right| > c_1 n^{-\gamma_1}\right\} + P\left\{\max_{j \in \mathcal{A}_2}\left|\widehat{\Omega}_{j,2} - \Omega_{j,2}\right| > c_2 n^{-\gamma_2}\right\} \\
\leq & O\left(|\mathcal{A}_1| \exp\left\{-b_1 n^{1-2\gamma_1}\right\}\right) + O\left(|\mathcal{A}_2| \exp\left\{-b_2 n^{1-2\gamma_2} + \log n\right\}\right) \\
\leq & O\left(|\mathcal{A}| \exp\left\{-b n^{1-2\gamma} + \log n\right\}\right),
\end{aligned}$$

where b is a constant depending on c_1 and c_2, and $\gamma = \max\{\gamma_1, \gamma_2\}$. In other words,

$$P\left\{\mathcal{A} \subseteq \widehat{\mathcal{A}}\right\} \geq 1 - O\left(|\mathcal{A}| \exp\left\{-b n^{1-2\gamma} + \log n\right\}\right).$$

6.3 Proof of Theorem 2.3

Proof. By condition vi and **Lemma 2**,

$$\begin{aligned}
& P\left\{\left(\min_{j \in \mathcal{A}_1} \widehat{\Omega}_{j,1} - \max_{j \notin \mathcal{A}_1} \widehat{\Omega}_{j,1}\right) < c_3 n^{-\gamma_3}\right\} \\
\leq & P\left\{\left(\min_{j \in \mathcal{A}_1} \widehat{\Omega}_{j,1} - \max_{j \notin \mathcal{A}_1} \widehat{\Omega}_{j,1}\right) - \left(\min_{j \in \mathcal{A}_1} \Omega_{j,1} - \max_{j \notin \mathcal{A}_1} \Omega_{j,1}\right) < -c_3 n^{-\gamma_3}\right\} \\
\leq & P\left\{\left|\left(\min_{j \in \mathcal{A}_1} \widehat{\Omega}_{j,1} - \max_{j \notin \mathcal{A}_1} \widehat{\Omega}_{j,1}\right) - \left(\min_{j \in \mathcal{A}_1} \Omega_{j,1} - \max_{j \in \mathcal{A}_1} \Omega_{j,1}\right)\right| > c_3 n^{-\gamma_3}\right\} \\
\leq & P\left\{\max_{j}\left|\widehat{\Omega}_{j,1} - \Omega_{j,1}\right| > \frac{c_3 n^{-\gamma_3}}{2}\right\} \\
\leq & 4p \exp\left\{-a_3 n^{1-2\gamma_3}\right\}
\end{aligned}$$

for some $a_3 > 0$ depending on c_3. Since $\log(p) = o(n^{1-2\gamma_3})$, we have $\log(p) \leq a_3 n^{1-2\gamma_3}/2$ for n sufficiently large. For some n_0 sufficiently large,

$$\sum_{n=n_0}^{+\infty} p \exp\{-a_3 n^{1-2\gamma_3}\} \leq \sum_{n=n_0}^{+\infty} n^{-2} < +\infty.$$

By Borel-Cantelli Lemma,

$$\liminf_{n \to \infty}\left\{\min_{j \in \mathcal{A}_1} \widehat{\Omega}_{j,1} - \max_{j \notin \mathcal{A}_1} \widehat{\Omega}_{j,1}\right\} \geq c_3 n^{-\gamma_3} > 0 \text{ a.s.}$$

We can derive similarly that

$$\liminf_{n \to \infty}\left\{\min_{j \in \mathcal{A}_2} \widehat{\Omega}_{j,2} - \max_{j \notin \mathcal{A}_2} \widehat{\Omega}_{j,2}\right\} \geq c_4 n^{-\gamma_4} > 0 \text{ a.s.}$$

6.4 Proof of Proposition 3

Proof. For any j, let $(\mathbf{X}, \widetilde{\mathbf{X}})_{(j)}$ be the vector by swapping the entries X_j and $\widetilde{X}_j$ in $(\mathbf{X}, \widetilde{\mathbf{X}})$; let $(\mathbf{x}, \tilde{\mathbf{x}})_{(j)}$ be the vector by swapping the entries x_j and $\tilde{x}_j$ in $(\mathbf{x}, \tilde{\mathbf{x}}) \in \mathbb{R}^{2p}$; and let $\mathbf{x}_{-j}$ denote the vector of $\mathbf{x}$ excluding x_j. Let $f_{\mathbf{U}|\mathbf{V}}(\mathbf{u}|\mathbf{v})$ denote the conditional distribution of $\mathbf{U}$ given $\mathbf{V} = \mathbf{v}$. For $j \notin \mathcal{A}_1$,

$$\begin{aligned}
f_{(Y,\delta)|((\mathbf{X},\widetilde{\mathbf{X}})_{(j)})}(y, s|(\mathbf{x}, \tilde{\mathbf{x}})) &= f_{(Y,\delta)|(\mathbf{X},\widetilde{\mathbf{X}})}(y, s|(\mathbf{x}, \tilde{\mathbf{x}})_{(j)}) \\
&= f_{(Y,\delta)|\mathbf{X}}(y, s|x_1, \cdots, x_{j-1}, \tilde{x}_j, x_{j+1}, \cdots, x_p) \\
&= f_{(Y,\delta)|\mathbf{X}_{-j}}(y, s|\mathbf{x}_{-j}) \\
&= f_{(Y,\delta)|\mathbf{X}}(y, s|\mathbf{x}) \\
&= f_{(Y,\delta)|(\mathbf{X},\widetilde{\mathbf{X}})}(y, s|\mathbf{x}, \tilde{\mathbf{x}}),
\end{aligned}$$

where the second and the last equations are due to $(Y,\delta) \perp\!\!\!\perp \widetilde{\mathbf{X}}|\mathbf{X}$, and the two equations in between follow from $(Y,\delta) \perp\!\!\!\perp X_j|\mathbf{X}_{-j}$ for $j \notin \mathcal{A}$. That is,

$$(Y,\delta)|((\mathbf{X},\widetilde{\mathbf{X}})_{(j)}) \overset{d}{=} (Y,\delta)|(\mathbf{X},\widetilde{\mathbf{X}}).$$

Since $(\mathbf{X},\widetilde{\mathbf{X}}) \overset{d}{=} (\mathbf{X},\widetilde{\mathbf{X}})$ by the definition of knockoff copies, it follows that

$$(Y,\delta,(\mathbf{X},\widetilde{\mathbf{X}})_{(j)}) \overset{d}{=} (Y,\delta,\mathbf{X},\widetilde{\mathbf{X}}),$$

which implies that $(\delta, X_j) \overset{d}{=} (\delta,\widetilde{X}_j)$ and $(Y,X_j)|\delta \overset{d}{=} (Y,\widetilde{X}_j)|\delta$. Therefore, $W_{j,1} = W_{j,2} = 0$.

In fact, we can show by repeating the above arguments that for any $\mathcal{S} \subset \mathcal{A}^c$,

$$(Y,\delta,(\mathbf{X},\widetilde{\mathbf{X}})_{\mathcal{S}}) \overset{d}{=} (Y,\delta,\mathbf{X},\widetilde{\mathbf{X}}),$$

where $(\mathbf{X},\widetilde{\mathbf{X}})_{\mathcal{S}}$ is the vector by swapping the entries X_j and $\widetilde{X}_j$ in $(\mathbf{X},\widetilde{\mathbf{X}})$ for all $j \in \mathcal{S}$. Let $\widehat{\mathbf{W}}_1 = (\widehat{W}_{1,1},\cdots,\widehat{W}_{p,1})$ and let $\mathbf{g}_1(\cdot):\mathbb{R}^{2p+1}\to\mathbb{R}^p$ be a function such that $\widehat{\mathbf{W}}_1 = \mathbf{g}_1(\delta,\mathbf{X},\widetilde{\mathbf{X}})$. Define $\epsilon_1,\cdots,\epsilon_p$ such that $\epsilon_j = 1$ for $j \in \mathcal{A}_1$ and ϵ_j is i.i.d. coin flip of $\{+1,-1\}$ for $j \notin \mathcal{A}_1$. Consider $\mathcal{S} = \{j : \epsilon_j = -1\} \subset \mathcal{A}_1^c$, then

$$(\widehat{W}_{1,1},\cdots,\widehat{W}_{p,1}) = \mathbf{g}(\delta,\mathbf{X},\widetilde{\mathbf{X}}) \overset{d}{=} \mathbf{g}(\delta,(\mathbf{X},\widetilde{\mathbf{X}})_{\mathcal{S}}) = (\epsilon_1\widehat{W}_{1,1},\cdots,\epsilon_p\widehat{W}_{p,1}).$$

The statement for $\left\{\widehat{W}_{j,2}\right\}_{j=1}^p$ can be shown analogously.

6.5 Proof of Theorem 2.5

Throughout this proof, we restrict ourselves to the event $\mathcal{E} = \{\mathcal{A} \subseteq \widehat{\mathcal{A}}^*(d)\}$. Let $\mathcal{B} = \mathcal{A}^c \cap \widehat{\mathcal{A}}^*(d)$. Denote by $|\widehat{W}_{(k),1}|$ the kth largest absolute value of the marginal statistics $\left\{\widehat{W}_{j,1} : j \in \widehat{\mathcal{A}}^*(d)\right\}$, $k = 1,...,d$. Define $|\widehat{W}_{(k),2}|$ analogously for the conditional statistics. For ease of presentation, we further define $\widehat{W}_{(d+1),1} = \widehat{W}_{(d+1),2} = 0$. Then we have

$$\begin{aligned}
\text{FDR} &= E\left[\frac{\#\left\{j \in \mathcal{B} : \widehat{W}_{j,1} \geq T_{\alpha,1} \text{ or } \widehat{W}_{j,2} \geq T_{\alpha,2}\right\}}{\#\left\{j : j \in \widehat{\mathcal{A}}\left(T_{\alpha,1},T_{\alpha,2}\right)\right\} \vee 1}\right] \\
&= E\left[\frac{\#\left\{j \in \mathcal{B} : \widehat{W}_{j,1} \geq T_{\alpha,1} \text{ or } \widehat{W}_{j,2} \geq T_{\alpha,2}\right\}}{1+\#\left\{j \in \mathcal{B} : \widehat{W}_{j,1} \leq -T_{\alpha,1} \text{ or } \widehat{W}_{j,2} \leq -T_{\alpha,2}\right\}}\right. \\
&\qquad \left.\times \frac{1+\#\left\{j \in \mathcal{B} : \widehat{W}_{j,1} \leq -T_{\alpha,1} \text{ or } \widehat{W}_{j,2} \leq -T_{\alpha,2}\right\}}{\#\left\{j : j \in \widehat{\mathcal{A}}\left(T_{\alpha,1},T_{\alpha,2}\right)\right\} \vee 1}\right] \\
&\leq E\left[\frac{\#\left\{j \in \mathcal{B} : \widehat{W}_{j,1} \geq T_{\alpha,1} \text{ or } \widehat{W}_{j,2} \geq T_{\alpha,2}\right\}}{1+\#\left\{j \in \mathcal{B} : \widehat{W}_{j,1} \leq -T_{\alpha,1} \text{ or } \widehat{W}_{j,2} \leq -T_{\alpha,2}\right\}}\right. \\
&\qquad \left.\times \frac{1+\#\left\{j \in \widehat{\mathcal{A}}^*(d) : \widehat{W}_{j,1} \leq -T_{\alpha,1} \text{ or } \widehat{W}_{j,2} \leq -T_{\alpha,2}\right\}}{\#\left\{j : j \in \widehat{\mathcal{A}}\left(T_{\alpha,1},T_{\alpha,2}\right)\right\} \vee 1}\right] \\
&\leq E\left[\frac{\#\left\{j \in \mathcal{B} : \widehat{W}_{j,1} \geq T_{\alpha,1} \text{ or } \widehat{W}_{j,2} \geq T_{\alpha,2}\right\}}{1+\#\left\{j \in \mathcal{B} : \widehat{W}_{j,1} \leq -T_{\alpha,1} \text{ or } \widehat{W}_{j,2} \leq -T_{\alpha,2}\right\}} \cdot \alpha\right].
\end{aligned}$$

The first inequality holds since $\mathcal{B} \subseteq \widehat{\mathcal{A}}^*(d)$ and the second inequality is due to the definition of $(T_{\alpha,1},T_{\alpha,2})$ in (9). Consider a partially-ordered discrete time process

$$\begin{aligned}
M(k_1,k_2) &= \frac{\#\left\{j \in \mathcal{B} : \widehat{W}_{j,1} \geq |\widehat{W}_{(k_1),1}| \text{ or } \widehat{W}_{j,2} \geq |\widehat{W}_{(k_2),2}|\right\}}{1+\#\left\{j \in \mathcal{B} : \widehat{W}_{j,1} \leq -|\widehat{W}_{(k_1),1}| \text{ or } \widehat{W}_{j,2} \leq -|\widehat{W}_{(k_2),2}|\right\}} \\
&= \frac{V_+(k_1,k_2)}{1+V_-(k_1,k_2)}
\end{aligned}$$

for $\{(k_1,k_2)\in\mathbb{Z}_+^2 : d+1\geq k_1,k_2\geq 1\}$, where

$$V_+(k_1,k_2)=\#\left\{j\in\mathcal{B}:\widehat{W}_{j,1}\geq|\widehat{W}_{(k_1),1}|\text{ or }\widehat{W}_{j,2}\geq|\widehat{W}_{(k_2),2}|\right\},$$
$$V_-(k_1,k_2)=\#\left\{j\in\mathcal{B}:\widehat{W}_{j,1}\leq-|\widehat{W}_{(k_1),1}|\text{ or }\widehat{W}_{j,2}\leq-|\widehat{W}_{(k_2),2}|\right\}.$$

Let $\mathcal{F}(k_1,k_2)$ be the σ-field generated by knowing $\{V_\pm(j_1,j_2): d+1\geq j_1\geq k_1, d+1\geq j_2\geq k_2\}$ as well as all the non-null statistics. The collection $\{\mathcal{F}(k_1,k_2): k_1,k_2=d+1,d,...,1\}$ of σ-fields is monotonic (thus a filtration) since $\mathcal{F}(k_1,k_2)\subseteq\mathcal{F}(k_1',k_2')$ for $k_1\geq k_1'$ and $k_2\geq k_2'$. In the next we show that $\{M(k_1,k_2):k_1,k_2=d+1,d,...,1\}$ is a supermartingale (running backward) with respect to $\{\mathcal{F}(k_1,k_2):k_1,k_2=d+1,d,...,1\}$. In other words, $E\left[M(k_1-1,k_2)|\mathcal{F}(k_1,k_2)\right]\leq M(k_1,k_2)$ and $E\left[M(k_1,k_2-1)|\mathcal{F}(k_1,k_2)\right]\leq M(k_1,k_2)$, $\forall k_1,k_2$.

Suppose that $|\widehat{W}_{j_{k_1},1}|=|\widehat{W}_{(k_1),1}|$ for $j_{k_1}\in\widehat{\mathcal{A}}^*(d)$. The filtration $\mathcal{F}(k_1,k_2)$ informs us about whether $j_{k_1}\in\mathcal{A}$ or not. On the one hand, if $j_{k_1}\in\mathcal{A}$ or $\widehat{W}_{j_{k_1},2}\geq|\widehat{W}_{(k_2),2}|$, then $M(k_1-1,k_2)=M(k_1,k_2)$. On the other hand, if $j_{k_1}\in\mathcal{B}$ and $\widehat{W}_{j_{k_1},2}<|\widehat{W}_{(k_2),2}|$, then

$$M(k_1-1,k_2)=\frac{V_+(k_1,k_2)-I_{j_{k_1}}}{1+V_-(k_1,k_2)-(1-I_{j_{k_1}})}=\frac{V_+(k_1,k_2)-I_{j_{k_1}}}{(V_-(k_1,k_2)+I_{j_{k_1}})\vee 1},$$

where $I_{j_{k_1}}=I\left\{\widehat{W}_{j_{k_1},1}>0\right\}$. Since $I_j\overset{\text{i.i.d}}{\sim}$ Bernoulli(0.5) for $j\in\mathcal{B}$ by Proposition 3, it follows that $P(I_{j_{k_1}}=1)=V_+(k_1,k_2)/[V_+(k_1,k_2)+V_-(k_1,k_2)]$ given $\mathcal{F}(k_1,k_2)$. As a result,

$$\begin{aligned}E\left[M(k_1-1,k_2)|\mathcal{F}(k_1,k_2)\right]&=\frac{V_+(k_1,k_2)-1}{V_-(k_1,k_2)+1}P(I_{j_{k_1}}=1)+\frac{V_+(k_1,k_2)}{V_-(k_1,k_2)\vee 1}\left[1-P(I_{j_{k_1}}=1)\right]\\&=\begin{cases}V_+(k_1,k_2)/[V_-(k_1,k_2)+1]=M(k_1,k_2), & \text{if } V_-(k_1,k_2)>0;\\ V_+(k_1,k_2)-1=M(k_1,k_2)-1, & \text{if } V_-(k_1,k_2)=0.\end{cases}\end{aligned}$$

Therefore, $E\left[M(k_1-1,k_2)|\mathcal{F}(k_1,k_2)\right]\leq M(k_1,k_2)$. We can show that $E\left[M(k_1,k_2-1)|\mathcal{F}(k_1,k_2)\right]\leq M(k_1,k_2)$ in the similar vein.

In this process, $(T_{\alpha,1},T_{\alpha,2})$ can be regarded as a stopping time with respect to the filtration $\mathcal{F}(k_1,k_2)$ as $\left\{k_{T_{\alpha,1}}\geq k_1, k_{T_{\alpha,2}}\geq k_2\right\}\in\mathcal{F}(k_1,k_2)$, where $k_{T_{\alpha,1}}$ and $k_{T_{\alpha,2}}$ denote the indices such that $|\widehat{W}_{k_{T_{\alpha,1}},1}|=T_{\alpha,1}$ and $|\widehat{W}_{k_{T_{\alpha,2}},2}|=T_{\alpha,2}$. According to the optional sampling theorem [45] and Proposition 3, we deduce

$$\begin{aligned}E\left[M(k_{T_{\alpha,1}},k_{T_{\alpha,2}})\right]&\leq E\left[M(d+1,d+1)\right]\\&=E\left[\frac{\#\left\{j\in\mathcal{B}:\widehat{W}_{j,1}>0\text{ or }\widehat{W}_{j,2}>0\right\}}{1+\#\left\{j\in\mathcal{B}:\widehat{W}_{j,1}\leq 0\text{ or }\widehat{W}_{j,2}\leq 0\right\}}\right]\\&=E\left[\frac{d_0-Y_1}{d_0-Y_2+1}\right],\end{aligned}$$

where $d_0=|\mathcal{B}|$, $Y_1=\#\left\{j\in\mathcal{B}:\widehat{W}_{j,1}\leq 0\text{ and }\widehat{W}_{j,2}\leq 0\right\}$, $Y_2=\#\left\{j\in\mathcal{B}:\widehat{W}_{j,1}>0\text{ and }\widehat{W}_{j,2}>0\right\}$. Let $Y_3=\#\left\{j\in\mathcal{B}:\widehat{W}_{j,1}\leq 0\text{ and }\widehat{W}_{j,2}>0\right\}$ and $Y_4=\#\left\{j\in\mathcal{B}:\widehat{W}_{j,1}>0\text{ and }\widehat{W}_{j,2}\leq 0\right\}$. Then $(Y_1,...,Y_4)$ follow a multinomial distribution with equal event probabilities and $\sum_{j=1}^{4} Y_j = d_0$. It follows that

$$\begin{aligned}
&E\left[\frac{d_0 - Y_1}{d_0 - Y_2 + 1}\right] \\
=&\sum_{y_2=0}^{d_0}\sum_{y_1=0}^{d_0-y_2}\frac{d_0-y_1}{d_0-y_2+1}\binom{d_0}{y_2}\binom{d_0-y_2}{y_1}\left(\frac{1}{4}\right)^{y_1}\left(\frac{1}{4}\right)^{y_2}\left(\frac{1}{2}\right)^{d_0-y_1-y_2} \\
=&\sum_{y_2=0}^{d_0}\frac{1}{d_0-y_2+1}\binom{d_0}{y_2}\left(\frac{1}{4}\right)^{y_2}\left[\sum_{y_1=0}^{d_0-y_2}(d_0-y_1)\binom{d_0-y_2}{y_1}\left(\frac{1}{4}\right)^{y_1}\left(\frac{1}{2}\right)^{d_0-y_1-y_2}\right] \\
=&\sum_{y_2=0}^{d_0}\frac{1}{d_0-y_2+1}\binom{d_0}{y_2}\left(\frac{1}{4}\right)^{y_2}\left[d_0\left(\frac{1}{4}+\frac{1}{2}\right)^{d_0-y_2}\right. \\
&\left.-(d_0-y_2)\sum_{y_1=1}^{d_0-y_2}\binom{d_0-y_2-1}{y_1-1}\left(\frac{1}{4}\right)^{y_1}\left(\frac{1}{2}\right)^{d_0-y_1-y_2}\right] \\
=&\sum_{y_2=0}^{d_0}\frac{1}{d_0-y_2+1}\binom{d_0}{y_2}\left(\frac{1}{4}\right)^{y_2}\left[d_0\left(\frac{1}{4}+\frac{1}{2}\right)^{d_0-y_2}\right. \\
&\left.-(d_0-y_2)\sum_{y_1=0}^{d_0-y_2-1}\binom{d_0-y_2-1}{y_1}\left(\frac{1}{4}\right)^{y_1+1}\left(\frac{1}{2}\right)^{d_0-y_1-y_2-1}\right] \\
=&\sum_{y_2=0}^{d_0}\frac{1}{d_0-y_2+1}\binom{d_0}{y_2}\left(\frac{1}{4}\right)^{y_2}\left[d_0\left(\frac{1}{4}+\frac{1}{2}\right)^{d_0-y_2}-(d_0-y_2)\left(\frac{1}{4}\right)\left(\frac{1}{4}+\frac{1}{2}\right)^{d_0-y_2-1}\right] \\
=&E\left[\frac{\frac{2}{3}d_0+\frac{1}{3}Y_2}{d_0-Y_2+1}\right].
\end{aligned}$$

Since $Y_2 \sim \text{Binomial}\left(d_0, \frac{1}{4}\right)$ by Proposition 3, we have

$$E\left[\frac{1}{d_0-Y_2+1}\right] = \frac{1}{d_0+1}\sum_{y_2=0}^{d_0}\binom{d_0+1}{y_2}\left(\frac{1}{4}\right)^{y_2}\left(\frac{3}{4}\right)^{d_0-y_2} = \frac{4}{3(d_0+1)}\left[1-\left(\frac{1}{4}\right)^{d_0+1}\right]$$

and

$$\begin{aligned}
E\left[\frac{Y_2}{d_0-Y_2+1}\right] &= \sum_{y_2=1}^{d_0}\binom{d_0}{y_2-1}\left(\frac{1}{4}\right)^{y_2}\left(\frac{3}{4}\right)^{d_0-y_2} \\
&= \frac{1}{3}\sum_{y_2=0}^{d_0-1}\binom{d_0}{y_2}\left(\frac{1}{4}\right)^{y_2}\left(\frac{3}{4}\right)^{d_0-y_2} \\
&= \frac{1}{3}\left[1-\left(\frac{1}{4}\right)^{d_0}\right]
\end{aligned}$$

Therefore,

$$E\left[\frac{\frac{2}{3}d_0+\frac{1}{3}Y_2}{d_0-Y_2+1}\right] \le \frac{8d_0}{9(d_0+1)}+\frac{1}{9} \le 1.$$

As a consequence, FDR$\le \alpha$.

In the next, we show the sure screening property. We can deduce from Lemma 2 that

$$P\left\{|\widehat{W}_{j,1} - W_{j,1}| \ge 2c_5 n_2^{-\gamma_5}\right\} \le O\left(\exp\left\{-b_5 n_2^{1-2\gamma_5}\right\}\right),$$

for some $b_5 > 0$. Furthermore, since $W_j = 0$ for $j \notin \mathcal{A}$ and $d < n_2$,

$$P\left\{\max_{j\in\mathcal{B}} |\widehat{W}_{j,1}| \le 2c_5 n_2^{-\gamma_5}\right\} \ge 1 - O\left(n_2 \exp\left\{-b_5 n_2^{1-2\gamma_5}\right\}\right).$$

Also, since $\min_{j\in\mathcal{A}_1} W_{j,1} \geq 4c_5 n_2^{-\gamma_5}$,

$$P\left\{\min_{j\in\mathcal{A}_1} \widehat{W}_{j,1} \geq 2c_5 n_2^{-\gamma_5}\right\} \geq 1 - O\left(n_2 \exp\left\{-b_5 n_2^{1-2\gamma_5}\right\}\right).$$

As a result,

$$P\left\{\min_{j\in\mathcal{A}_1} \widehat{W}_{j,1} \geq \max_{j\in\mathcal{B}} |\widehat{W}_{j,1}|\right\} \geq 1 - O\left(n_2 \exp\left\{-b_5 n_2^{1-2\gamma_5}\right\}\right).$$

Similarly, for some $b_6 > 0$,

$$P\left\{\min_{j\in\mathcal{A}_2} \widehat{W}_{j,2} \geq \max_{j\in\mathcal{B}} |\widehat{W}_{j,2}|\right\} \geq 1 - O\left(n_2^2 \exp\left\{-b_6 n_2^{1-2\gamma_6}\right\}\right).$$

That is, important features are ranked above unimportant ones with probability approaching 1. Given $\min_{j\in\mathcal{A}_1} \widehat{W}_{j,1} \geq \max_{j\in\mathcal{B}} |\widehat{W}_{j,1}|$ and $\min_{j\in\mathcal{A}_2} \widehat{W}_{j,2} \geq \max_{j\in\mathcal{B}} |\widehat{W}_{j,2}|$, the knockoff procedure stops at $T_{\alpha,1} \leq \min_{j\in\mathcal{A}_1} \widehat{W}_{j,1}$ and $T_{\alpha,2} \leq \min_{j\in\mathcal{A}_2} \widehat{W}_{j,2}$ as $\widehat{\mathrm{FDP}}(\min_{j\in\mathcal{A}_1} \widehat{W}_{j,1}, \min_{j\in\mathcal{A}_2} \widehat{W}_{j,2}) = \frac{1+0}{|\mathcal{A}|} \leq \alpha$, in which case $\widehat{\mathcal{A}}\left(T_{\alpha,1}, T_{\alpha,2}\right) \supseteq \mathcal{A}$.

7 Additional Simulation Results

As additional simulation studies, we further consider a linear design with varying signal strength (Example 3) and a more complex nonlinear design (Example 4) for both AFT- and PH-type of models under different censoring mechanisms.

Example 3. Let $\mathbf{X} \sim N_p(\mathbf{0}, \Sigma)$, where $\Sigma = AR(0.5)$. Given $\mathbf{X}$, the true survival time is generated from the following accelerated failure time (AFT) model and proportional hazard (PH) model:

1. **Model 1:** $\log T = 2X_1 + .8X_2 + .9X_3 + X_4 + 2X_5 + \epsilon$, where $\epsilon \sim N(0,1)$ independently;
2. **Model 2:** $\log(.5(e^{2T} - 1)) = 2X_1 + .8X_2 + .9X_3 + X_4 + 2X_5 + \epsilon$, where ϵ follows the standard extreme value distribution independently.

For each model, the survival time is subject to two censoring mechanisms:

1. independent censoring time C generated from uniform distribution on $[0, c_0]$;
2. dependent censoring time C generated from exponential distribution with mean $c_0 e^{X_3}$,

where the constant c_0 is chosen to achieve 30% or 50% censoring rate (CR). The results are summarized in Table 6 for $n = 200$ and $d = [n/\log n] = 38$.

Example 4. The setup of this example is identical to Example 3, except that the true survival time is generated from the following accelerated failure time (AFT) model and proportional hazard (PH) model:

1. **Model 3:** $\log T = g_1 + g_2 + g_3 + g_{10} + \epsilon$,where $\epsilon \sim N(0,1)$ independently and $g_1 = X_1$, $g_2 = -X_2^2 + X_2$, $g_3 = 2[exp(-3(X_3-1)^2) + exp(-4(X_3-3)^2)]$ and $g_{10} = X_{10}^2 + |X_{10}|$;
2. **Model 4:** $\log(.5(e^{2T} - 1)) = g_1 + g_2 + g_3 + g_{10} + \epsilon$, where ϵ follows the standard extreme value distribution independently.

The results are summarized in Table 7 for $n = 200$ and $d = [n/\log n] = 38$.

Table 6: Quantiles of MMS (M_τ) and selection proportions (P_j's and $P_{\mathcal{A}}$) for models in **Example 3** based on 200 replicates with $n = 200$, $p = 5000$ and $d = [n/\log n] = 38$.

Model	CR	Method	$M_{5\%}$	$M_{25\%}$	$M_{50\%}$	$M_{75\%}$	$M_{95\%}$	P_1	P_2	P_3	P_4	P_5	$P_{\mathcal{A}}$
1(a)	30%	KIDS	5.0	5.0	5.0	5.0	6.0	1.000	1.000	0.995	1.000	1.000	0.995
		CRIS	5.0	5.0	5.0	5.0	6.0	1.000	1.000	1.000	1.000	1.000	1.000
		IPOD	5.0	5.0	7.0	14.0	91.2	0.985	0.960	0.975	0.980	1.000	0.915
		RCDCS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
	50%	KIDS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
		CRIS	246.4	2212.0	4149.0	4969.5	5000.0	0.090	0.045	0.065	0.085	0.146	0.020
		IPOD	5.0	5.0	5.0	5.0	6.0	0.995	1.000	1.000	0.995	0.995	0.985
		RCDCS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
1(b)	30%	KIDS	5.0	5.0	5.0	5.0	6.0	1.000	0.995	0.995	1.000	1.000	0.995
		CRIS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
		IPOD	5.0	6.0	10.0	35.2	241.6	0.980	0.920	0.890	0.945	0.965	0.765
		RCDCS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
	50%	KIDS	5.0	5.0	5.0	5.0	6.0	1.000	0.995	1.000	1.000	1.000	0.995
		CRIS	5.0	11.0	159.0	1675.5	4583.1	0.510	0.505	0.545	0.670	0.710	0.385
		IPOD	5.0	5.0	10.0	26.5	467.2	0.990	0.915	0.925	0.965	0.980	0.805
		RCDCS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
2(a)	30%	KIDS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
		CRIS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
		IPOD	5.0	5.0	5.0	5.0	6.0	1.000	1.000	1.000	1.000	1.000	1.000
		RCDCS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
	50%	KIDS	5.0	5.0	5.0	5.0	6.0	1.000	1.000	1.000	1.000	1.000	1.000
		CRIS	5.0	5.0	6.5	55.5	1147.6	0.860	0.830	0.835	0.885	0.920	0.720
		IPOD	5.0	5.0	5.0	5.0	6.0	1.000	1.000	0.995	1.000	1.000	0.995
		RCDCS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
2(b)	30%	KIDS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
		CRIS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
		IPOD	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
		RCDCS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
	50%	KIDS	5.0	5.0	5.0	5.0	7.0	1.000	1.000	1.000	1.000	1.000	1.000
		CRIS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
		IPOD	5.0	5.0	5.0	5.0	9.0	1.000	0.985	1.000	1.000	1.000	0.985
		RCDCS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000

Table 7: Quantiles of MMS (M_τ) and selection proportions (P_j's and $P_{\mathcal{A}}$) for models in **Example 4** based on 200 replicates with $n = 200$, $p = 5000$ and $d = [n/\log n] = 38$.

Model	CR	Method	$M_{5\%}$	$M_{25\%}$	$M_{50\%}$	$M_{75\%}$	$M_{95\%}$	P_1	P_2	P_3	P_{10}	$P_{\mathcal{A}}$
3(a)	30%	KIDS	4.0	4.0	7.5	24.2	139.4	0.945	0.985	0.790	1.000	0.745
		CRIS	90.0	1270.8	2333.5	3871.8	4765.5	1.000	1.000	0.785	0.030	0.030
		IPOD	7.0	34.8	122.5	434.5	1544.6	0.740	0.790	0.420	0.965	0.285
		RCDCS	4.0	7.0	11.0	30.2	79.1	1.000	1.000	1.000	0.820	0.820
	50%	KIDS	4.0	4.0	4.0	6.0	28.4	0.995	1.000	0.940	1.000	0.935
		CRIS	1031.2	3039.5	4219.5	4909.0	4999.0	0.190	0.265	0.055	0.000	0.000
		IPOD	6.0	22.8	44.5	168.2	713.3	0.870	0.935	0.585	0.890	0.460
		RCDCS	11.0	40.5	98.0	219.5	611.0	1.000	1.000	0.965	0.260	0.245
3(b)	30%	KIDS	4.0	4.0	5.5	10.0	49.2	0.940	0.980	0.965	1.000	0.900
		CRIS	47.7	736.8	1978.5	3849.2	4726.1	1.000	1.000	0.935	0.045	0.045
		IPOD	7.0	54.8	223.0	628.5	1801.6	0.540	0.650	0.560	0.865	0.210
		RCDCS	5.0	9.0	20.0	49.5	184.4	1.000	1.000	1.000	0.670	0.670
	50%	KIDS	4.0	4.0	4.0	6.0	16.0	0.980	1.000	1.000	0.995	0.975
		CRIS	269.7	1161.8	2518.5	3693.8	4679.7	0.855	0.825	0.430	0.020	0.005
		IPOD	10.0	83.5	257.5	827.2	2239.7	0.550	0.635	0.440	0.745	0.155
		RCDCS	17.9	80.5	183.5	454.8	1055.3	1.000	1.000	1.000	0.120	0.120
4(a)	30%	KIDS	4.0	4.0	6.0	21.8	113.1	0.985	1.000	0.760	0.995	0.750
		CRIS	80.9	837.0	2080.0	3473.0	4675.1	1.000	1.000	0.985	0.020	0.020
		IPOD	4.0	4.0	8.0	25.2	188.2	0.995	0.995	0.830	0.975	0.800
		RCDCS	5.0	16.0	40.0	126.8	346.0	1.000	1.000	0.945	0.520	0.490
	50%	KIDS	4.0	5.0	7.0	18.2	90.0	0.995	1.000	0.845	0.950	0.795
		CRIS	159.8	1249.8	2190.0	3454.2	4686.6	0.995	0.975	0.780	0.025	0.015
		IPOD	4.0	6.8	17.0	64.5	405.0	0.975	0.995	0.760	0.935	0.695
		RCDCS	14.9	59.5	178.5	365.5	918.9	0.995	1.000	0.920	0.200	0.185
4(b)	30%	KIDS	4.0	4.0	4.0	8.5	154.3	1.000	1.000	0.840	1.000	0.840
		CRIS	86.4	669.2	2187.5	3419.0	4744.1	1.000	1.000	0.985	0.025	0.025
		IPOD	4.0	4.0	6.5	15.0	252.9	0.995	1.000	0.830	0.995	0.825
		RCDCS	7.0	21.8	63.5	147.2	390.0	1.000	1.000	0.970	0.405	0.395
	50%	KIDS	4.0	5.0	11.0	36.5	189.3	1.000	1.000	0.730	0.875	0.630
		CRIS	87.0	761.8	1901.5	3381.8	4633.0	1.000	0.985	0.905	0.015	0.015
		IPOD	5.0	12.8	51.5	149.2	757.8	0.920	0.980	0.575	0.905	0.460
		RCDCS	43.0	233.8	533.5	954.8	2620.9	1.000	1.000	0.860	0.045	0.030

8 Clinical Characteristics of The TCGA and GSE65858 Primary Tumor Samples.

A summary of the clinical characteristics of the TCGA and GSE65858 primary tumor samples is presented in Table 8.

Table 8: Subgroup frequency (and percentage in parentheses) for clinical characteristics of the TCGA and GSE65858 Primary Tumor Samples.

Variable	**TCGA (n=518)**	**GSE65858 (n=253)**
Age		
$<= 50$	94 (18.15%)	40 (15.81%)
> 50	424 (81.85%)	213 (84.19%)
Gender		
Male	383 (73.94%)	210 (83.00%)
Female	135 (26.06%)	43 (17.00%)
Tumor stage		
I/II	100 (22.42%)	49 (19.36%)
III	81 (18.16%)	33 (13.04%)
IV	265 (59.42%)	171 (67.59%)
HPV status		
Positive	97 (18.80%)	73 (28.97%)
Negative	419 (81.20%)	179 (71.03%)
Alcohol history		
Yes	345 (68.05%)	226 (89.33%)
No	162 (31.95%)	27 (10.67%)
Smoking history		
Yes	389 (76.88%)	209 (82.61%)
No	117 (23.12%)	44 (17.39%)

Acknowledgements

This research was partially funded by grants P20CA252717, P20CA264067 and R21DE031879 from the United States National Institutes of Health (NIH), and the VCU Quest fund. Services and products in support of the research project were generated by the VCU Massey Comprehensive Cancer Center Biostatistics Shared Resource, supported, in part, with funding from NIH-NCI Cancer Center Support Grant P30CA016059.

References

[1] Athanassios Argiris, Michalis V Karamouzis, David Raben, and Robert L Ferris. Head and Neck Cancer. *The Lancet*, 371(9625):1695–1709, 2008.

[2] Krishnakumar Balasubramanian, Bharath Sriperumbudur, and Guy Lebanon. Ultrahigh dimensional feature screening via rkhs embeddings. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 126–134. PMLR, 2013.

[3] Rina Foygel Barber and Emmanuel J Candès. A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5):2504–2537, 2019.

[4] Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055 – 2085, 2015.

[5] Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

[6] Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[7] Xiaolin Chen, Xiaojing Chen, and Hong Wang. Robust feature screening for ultra-high dimensional right censored data via distance correlation. *Computational Statistics & Data Analysis*, 119:118–138, 2018.

[8] Hengjian Cui, Runze Li, and Wei Zhong. Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, 110(510):630–641, 2015.

[9] Shanshan Ding, Wei Qian, and Lan Wang. Double-slicing assisted sufficient dimension reduction for high-dimensional censored data. *The Annals of Statistics*, 48(4):2132 – 2154, 2020.

[10] Dominic Edelmann, Manuela Hummel, Thomas Hielscher, Maral Saadati, and Axel Benner. Marginal variable screening for survival endpoints. *Biometrical Journal*, 62(3):610–626, 2020.

[11] João Fadista, Petter Vikman, Emilia Ottosson Laakso, Inês Guerra Mollet, Jonathan Lou Esguerra, Jalal Taneera, Petter Storm, Peter Osmark, Claes Ladenvall, Rashmi B Prasad, et al. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proceedings of the National Academy of Sciences*, 111(38):13924–13929, 2014.

[12] Jianqing Fan, Yang Feng, and Rui Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557, 2011.

[13] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[14] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[15] Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

[16] Jianqing Fan, Richard Samworth, and Yichao Wu. Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research*, 10:2013–2038, 2009.

[17] Li-Ching Fan, Yung-Ming Jeng, Yueh-Tong Lu, and Huang-Chun Lien. Spock1 is a novel transforming growth factor-β–induced myoepithelial marker that enhances invasion and correlates with poor prognosis in breast cancer. *PLoS One*, 11(9):e0162933, 2016.

[18] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[19] Kenji Fukumizu, Arthur Gretton, Gert Lanckriet, Bernhard Schölkopf, and Bharath K. Sriperumbudur. Kernel choice and classifiability for rkhs embeddings of probability distributions. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22, pages 1750–1758. Curran Associates, Inc., 2009.

[20] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory: 16th International Conference, ALT 2005, Singapore, October 8-11, 2005. Proceedings 16*, pages 63–77. Springer, 2005.

[21] Arthur Gretton, Kenji Fukumizu, and Bharath K Sriperumbudur. Discussion of: Brownian distance covariance. *The annals of applied statistics*, 3(4):1285–1294, 2009.

[22] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, page 585–592. MIT Press, 2008.

[23] Peter Hall and Hugh Miller. Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18(3):533–550, 2009.

[24] Patrick J Heagerty and Yingye Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105, 2005.

[25] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[26] Hyokyoung G Hong, Xuerong Chen, David C Christiani, and Yi Li. Integrated powered density: Screening ultrahigh dimensional covariates with survival outcomes. *Biometrics*, 74(2):421–429, 2018.

[27] John D Kalbfleisch and Ross L Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2011. p. 241.

[28] Chenlu Ke and Xiangrong Yin. Expected conditional characteristic function-based measures for testing independence. *Journal of the American Statistical Association*, 115(530):985–996, 2020.

[29] Randall J Kimple and Paul M Harari. The prognostic value of hpv in head and neck cancer patients undergoing postoperative chemoradiotherapy. *Annals of translational medicine*, 3(Suppl 1), 2015.

[30] C René Leemans, Boudewijn JM Braakhuis, and Ruud H Brakenhoff. The molecular biology of head and neck cancer. *Nature reviews cancer*, 11(1):9–22, 2011.

[31] Jialiang Li, Qi Zheng, Limin Peng, and Zhipeng Huang. Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes. *Biometrics*, 72(4):1145–1154, 2016.

[32] Runze Li, Wei Zhong, and Liping Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139, 2012.

[33] Wanjun Liu, Yuan Ke, Jingyuan Liu, and Runze Li. Model-free feature screening and fdr control with knockoff features. *Journal of the American Statistical Association*, 117(537):428–443, 2022.

[34] Yi Liu, Xiaolin Chen, and Gang Li. A new joint screening method for right-censored time-to-event data with ultra-high dimensional covariates. *Statistical methods in medical research*, 29(6):1499–1513, 2020.

[35] Qing Mai and Hui Zou. The kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, 100(1):229–234, 2013.

[36] M Murdocca, C De Masi, S Pucci, R Mango, G Novelli, C Di Natale, and F Sangiuolo. Lox-1 and cancer: an indissoluble liaison. *Cancer gene therapy*, 28(10-11):1088–1098, 2021.

[37] Ilda Patrícia Ribeiro, Luísa Esteves, Francisco Caramelo, Isabel Marques Carreira, and Joana Barbosa Melo. Integrated multi-omics signature predicts survival in head and neck cancer. *Cells*, 11(16):2536, 2022.

[38] Yi-Jun Shu, Hao Weng, Yuan-Yuan Ye, Yun-Ping Hu, Run-Fa Bao, Yang Cao, Xu-An Wang, Fei Zhang, Shan-Shan Xiang, Huai-Feng Li, et al. Spock1 as a potential cancer prognostic marker promotes the proliferation and metastasis of gallbladder cancer cells by activating the pi3k/akt pathway. *Molecular cancer*, 14(1):1–14, 2015.

[39] Bernard W Silverman. *Density Estimation for Statistics and Data Analysis*. Routledge, 2018.

[40] Rui Song, Wenbin Lu, Shuangge Ma, and X Jessie Jeng. Censored rank independence screening for high-dimensional survival data. *Biometrika*, 101(4):799–814, 2014.

[41] Kuiwei Su, Ying Wang, Hefeng Gu, Lan Ma, and Guihong Xuan. Overexpression of fatty acid desaturase 3 predicts poor prognosis in head and neck squamous cell carcinoma. *Medicine*, 101(49), 2022.

[42] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769 – 2794, 2007.

[43] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[44] Scott A Tomlins, Bharathi Laxman, Saravana M Dhanasekaran, Beth E Helgeson, Xuhong Cao, David S Morris, Anjana Menon, Xiaojun Jing, Qi Cao, Bo Han, et al. Distinct classes of chromosomal rearrangements create oncogenic ets gene fusions in prostate cancer. *Nature*, 448(7153):595–599, 2007.

[45] Robert Buchanan Washburn. *The optional sampling theorem for partially ordered time processes and multiparameter stochastic calculus*. PhD thesis, Massachusetts Institute of Technology, 1979.

[46] Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

[47] Gunnar Wichmann, Maciej Rosolowski, Knut Krohn, Markus Kreuz, Andreas Boehm, Anett Reiche, Ulrike Scharrer, Dirk Halama, Julia Bertolini, Ulrike Bauer, et al. The role of hpv rna transcription, immune response-related gene expression and disruptive tp53 mutations in diagnostic and prognostic profiling of head and neck cancer. *International journal of cancer*, 137(12):2846–2857, 2015.

[48] Jinfeng Xu, Wai Keung Li, and Zhiliang Ying. Variable screening for survival data in the presence of heterogeneous censoring. *Scandinavian Journal of Statistics*, 47(4):1171–1191, 2020.

[49] Lijing Yao, Yu Gyoung Tak, Benjamin P Berman, and Peggy J Farnham. Functional annotation of colon cancer risk snps. *Nature communications*, 5(1):5114, 2014.

[50] Peng Zhang, Yan Zhao, Xin Xia, Song Mei, Yixuan Huang, Yingying Zhu, Shuting Yu, and Xingming Chen. Expression of olr1 gene on tumor-associated macrophages of head and neck squamous cell carcinoma, and its correlation with clinical outcome. *Oncoimmunology*, 12(1):2203073, 2023.

[51] Tingyou Zhou and Liping Zhu. Model-free feature screening for ultrahigh dimensional censored regression. *Statistics and Computing*, 27(4):947–961, 2017.

[52] Li-Ping Zhu, Lexin Li, Runze Li, and Li-Xing Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475, 2011.

[53] Hui Zou. The adaptive lasso and its Oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.