 DOI: 10.1111/biom.12820

Integrated Powered Density: Screening Ultrahigh Dimensional Covariates with Survival Outcomes

Hyokyoung G. Hong[1,*] **Xuerong Chen**[2] **David C. Christiani**[3] **and Yi Li**[4]

[1]Department of Statistics and Probability, Michigan State University, East Lansing, Michigan, U.S.A.
[2]Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan, China
[3]Department of Environmental Science, Harvard TH Chan School of Public Health, Boston, Massachusetts, U.S.A.
[4]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, U.S.A.
[*]*email:* hhong@msu.edu

Summary. Modern biomedical studies have yielded abundant survival data with high-throughput predictors. Variable screening is a crucial first step in analyzing such data, for the purpose of identifying predictive biomarkers, understanding biological mechanisms, and making accurate predictions. To nonparametrically quantify the relevance of each candidate variable to the survival outcome, we propose integrated powered density (IPOD), which compares the differences in the covariate-stratified distribution functions. The proposed new class of statistics, with a flexible weighting scheme, is general and includes the Kolmogorov statistic as a special case. Moreover, the method does not rely on rigid regression model assumptions and can be easily implemented. We show that our method possesses sure screening properties, and confirm the utility of the proposal with extensive simulation studies. We apply the method to analyze a multiple myeloma study on detecting gene signatures for cancer patients' survival.

Key words: Integrated powered density; Kolmogorov statistic; Survival analysis; Variable screening.

1. Introduction

As a key aspect of scientific discovery is to identify low-dimensional presentations of predictive features in a high-dimensional space, there is an urgent need to develop a fast but crude method to screen relevant variables, and ensure that the unselected variables are indeed irrelevant (Fan and Lv, 2008). In the context of integrative analysis, screening steps are crucial in simultaneously reducing dimensionality and ensuring estimation accuracy (Fan and Lv, 2010). More broadly, screening has found wide applications, ranging from quality control in the data processing step for genome/genetic studies (Beyene et al., 2009), to identifying predictive biomarkers for understanding biological mechanisms (Heinzel et al., 2014). This article is motivated by a clinical study (Shaughnessy et al., 2007) on multiple myeloma, the second-most common hematological cancer, which often results in bone lesions, immunological disorders, and renal failure. A deeper understanding of the molecular etiology of this disease, such as through detecting the gene signatures that are relevant to cancer patients' survival, would lead to novel therapeutic targets and more accurate risk classification systems (Mulligan et al., 2007). However, with expression level measurements on more than 50,000 probe sets, this dataset presents substantial challenges that defy the existing statistical tools for dimension reduction.

While screening approaches, including sure independence screening (Fan and Lv, 2008), have been actively pursued for fully observed outcomes, the development of high-dimensional screening tools with survival outcomes has been less fruitful. Limited works include a sure screening procedure for Cox's proportional hazards model (Fan et al., 2010), a Cox univariate shrinkage estimator (Tibshirani, 2009), a marginal maximum partial likelihood estimator (Zhao and Li, 2012), a general class of single-index hazard rate models (Gorst-Rasmussen and Scheike, 2013), and a conditional screening with prior information (Hong et al., 2016). As successful as these methods have been, their validity often hinges upon modeling assumptions between outcomes and predictors (Lin and Halabi, 2013), violations of which can lead to inflated false discoveries or nondiscoveries.

There has been a surge of effort in developing model-free screening procedures that achieve sure screening properties under weak conditions; see Zhu et al. (2011), Li et al. (2012), Liu et al. (2014), Shao and Zhang (2014), and Mai and Zou (2015). However, the extension of these nonparametric works to accommodate censored outcome data is elusive and nontrivial. Limited works include a quantile adaptive method (He et al., 2013) and a censored rank independence screening method (Song et al., 2014), and a survival impact index procedure (Li et al., 2016).

In a survival setting, nonparametric variable screeners have focused on discerning how each candidate variable influences overall survival functions. Studying the variability of survival functions for strata defined by each variable is one possible way. We note that such survival differences may occur either during the early or late period in the follow-up due to disease-related characteristics. Therefore, screening approaches that rely on a single screening criterion may not be able to capture

the complex difference patterns and may lead to false nondiscovery.

Our article proposes a model-free method for screening ultrahigh dimensional predictors when the outcome is randomly right censored. We propose an integrated powered density (IPOD) criterion to screen predictors that are relevant to the survival outcome. With an embedded weighting scheme, IPOD flexibly compares the differences in covariate-stratified distribution functions that occur at various time points. The proposed framework is general, including the Kolmogorov filter (Mai and Zou, 2015) as a special case.

Compared to the other screening methods, our work presents several novel contributions. First, we have introduced a general nonparametric screening framework to accommodate survival outcomes and provided a convenient means to select ultrahigh dimensional predictors of mixed types, discrete, or continuous. Second, by embedding our screening criteria in a new class of statistics, our method provides more flexibility and power in efficiently selecting signals out of ultrahigh dimenionsal predictors. We have established sure screening properties, and conducted simulations to validate the method.

The remainder is organized as follows. In Section 2, we introduce the concept of integrated powered density and propose a model-free screening procedure to screen covariates of different types (e.g., discrete or continuous). We present in Section 3 the sure screening property. In Section 4, we evaluate the finite-sample performance via simulation studies. In Section 5, we apply the method to analyze a multiple myeloma study and identify gene signatures that are relevant to patients' survival. We conclude with some final remarks in Section 6. Technical details are deferred to the Supplementary Web Materials.

2. Integrated Powered Density (IPOD) for Survival Outcomes

Suppose we have n observations with p covariates, where $p \gg n$. Denote by X_{ij} the jth covariate for subject i, and write $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$. Let T_i be the underlying survival time and C_i be the potential censoring time. We observe $Y_i = \min\{T_i, C_i\}$ and $\delta_i = I(T_i \le C_i)$, where $I(\cdot)$ is the indicator function. We assume that T_i and C_i are independent given $\mathbf{X}_i$, and $(Y_i, \delta_i, \mathbf{X}_i)$ are independently and identically distributed (i.i.d). In particular, we assume (T_i, X_{ij}), $i = 1, \ldots, n$, are i.i.d copies of (T, X_j), the random variables that underlie the survival time and covariates. To ensure the estimability of the survival function, we assume that there exists a $\tau > 0$ such that $P(Y_i > \tau|\mathbf{X}_i) > 0$ and restrict our analysis to $[0, \tau]$, a common practice in survival analysis (Zeng and Lin, 2007). In practice, τ is often chosen to be the study duration.

Denote by $S(\cdot)$ the marginal survival function of T and by $S(t|\mathbf{X})$ the conditional survival function of T given $\mathbf{X}$. To facilitate the selection of covariates that are relevant to T, we define the set of active covariates as

$$\mathcal{M} = \{j : S(t|\mathbf{X}) \text{ functionally depends on } X_j \text{ for some } t \in (0, \infty)\}.$$

In biomedical studies, it is not unreasonable to stipulate a sparsity condition that only a small number of biomarkers are relevant to the disease-specific survival. That is, the cardinality of $\mathcal{M}$ is small relative to p. Our goal is to identify $\mathcal{M}$. As the candidate variables can be of mixed types, we start by considering a categorical variable, say, X_j, with R_j categories such that $X_j \in \{1, 2, \ldots, R_j\}$. Later we will extend the method to cover continuous covariates.

For a generic density function corresponding to a (continuous) survival time, denoted by $f(\cdot)$, and for $t \in (0, \infty)$, we define the integrated powered density (IPOD) as $\int_0^t f^\gamma(s)ds$, for $\gamma > 0$. IPOD resembles the cumulative density function (CDF) and satisfies the basic properties of CDFs, except that it does not necessarily approach to one when $t \to \infty$. This unique property is advantageous for using IPOD to detect distributional differences, as exemplified in Figure 1.

When $\gamma = 1$, IPOD is a CDF. When $\gamma \neq 1$, IPOD is closely related to the Renyi entropy with a power index γ (Cover and Thomas, 2012). To study the relevance of covariate X_j to the survival time T, we propose to characterize the variability of IPOD across different categories of X_j. Specifically, for each pair of categories, say, $X_j = r_1$ and $X_j = r_2$ $(r_1, r_2 \in \{1, \ldots, R_j\})$, we compute the absolute difference of IPOD, take the maximum over all pairs of r_1, r_2 and use it as the screening criterion:

$$\mathcal{I}_j^{(\gamma)} = \max_{r_1, r_2 \in \{1, \ldots, R_j\}} \sup_{t \in [0,\tau]} \left| \int_0^t f^\gamma_{T|X_j}(s|X_j = r_1)ds - \int_0^t f^\gamma_{T|X_j}(s|X_j = r_2)ds \right|, \quad (1)$$

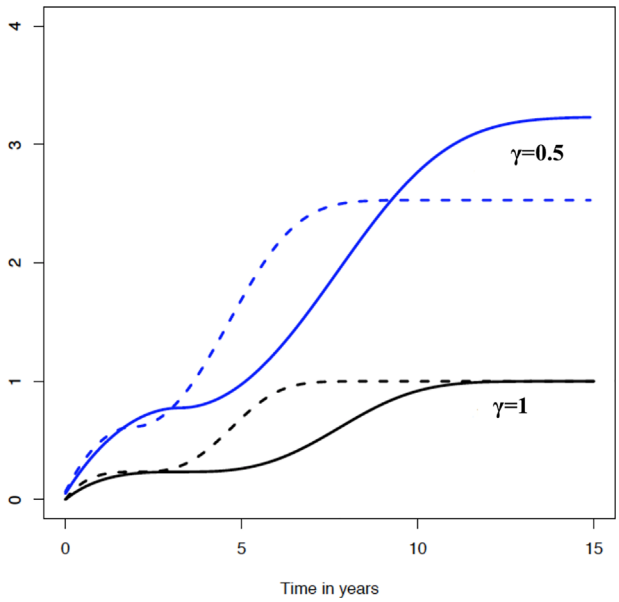

Figure 1. Integrated Powered Density: $\int_0^t f^\gamma(s|x)ds$ is shown for $x = 1$ (solid line) and $x = 0$ (dashed line) with $\gamma = 1$ and $\gamma = 0.5$, respectively. Here, $f(s|x)$ is the density function corresponding to the crossing hazard function of $\lambda(s|x) = 0.1\exp(-0.5x)\{s\exp(-0.5x) - 2\}^2$, which is adapted from Zhang and Peng (2009). This figure appears in color in the electronic version of this article.

where $f_{T|X_j}(s|X_j = r)$ denotes the conditional density function of T given $X_j = r$. The rationale is that $\mathcal{I}_j^{(\gamma)} = 0$ implies T and X_j are independent and so does the converse. Hence, an estimated $\mathcal{I}_j^{(\gamma)}$ empirically gauges the relevance of X_j to T. When $\gamma = 1$, (1) is simply the classical Kolmogorov difference, $\max_{r_1,r_2\in\{1,\ldots,R_j\}} \sup_{t\in[0,\tau]} |F_{T|X_j}(t|X_j = r_1) - F_{T|X_j}(t|X_j = r_2)|$. However, the added power index γ in (1) may inflate early ($\gamma > 1$) or late differences ($\gamma < 1$) and thus gives more flexibility to detect the distributional differences. For example, Figure 1 illustrates that IPOD with $\gamma = 0.5$ amplifies the late differences in CDF (corresponding to $\gamma = 1$), which exemplifies the possible role of γ in differentiating distributions.

This leads to a natural question of which γ should be chosen in analysis. For unbiased analysis, it is not reasonable to look at the survival curves first, and then choose weights. However, prior medical knowledge might guide the choice of γ. For example, late differences are more likely to occur in childhood cancers (National Cancer Policy Board, 2003), while early treatment differences are more prominent in late stage lung cancer remission (Fossella et al., 2000).

2.1. *Estimation of IPOD*

To estimate (1) for a given $\gamma > 0$, we need to first reliably estimate the density of the survival time. We adopt a kernel type density estimation. We illustrate the idea by estimating the marginal density of T.

Denote by $0 \equiv t_0 < t_1 < t_2 < \cdots < t_M$ the ordered observed failure times in the data, and by $\hat{S}_T(t)$ the Kaplan–Meier estimate of $S_T(t)$, the marginal survival function of T at time t. Then the estimated kernel density function for T is

$$\hat{f}_T(t) = -\int K\left(\frac{t-s}{h_n}\right) d\hat{S}_T(s) = \sum_{i=1}^{M} K\left(\frac{t-t_i}{h_n}\right)(\hat{S}_T(t_{i-1}) - \hat{S}_T(t_i)),$$

where $h_n > 0$ is the bandwidth and $K(\cdot)$ is a kernel function.

Similarly, restricting samples to $X_j = r$, we can compute $\hat{f}_{T|X_j}(t|X_j = r)$, the estimate of the conditional density function given $X_j = r$. Thus, (1) can be estimated by

$$\widehat{\mathcal{I}}_j^{(\gamma)} = \max_{r_1,r_2} \sup_{t\in[0,\tau]} \left| \int_0^t \hat{f}_{T|X_j}^{\gamma}(s|X_j = r_1)ds - \int_0^t \hat{f}_{T|X_j}^{\gamma}(s|X_j = r_2)ds \right|. \quad (2)$$

When X_j is continuous, without loss of generality, we assume the support of X_j is the real line $\mathbb{R}$. We can discretize X_j into R_j slices by using the percentiles of the empirical distribution of X_j. That is, $\tilde{X}_j = r$ if $X_j \in [\hat{q}_{j(r-1)}, \hat{q}_{j(r)})$, where $\hat{q}_{j(r)}$ is the r/R_jth percentile of the empirical distribution of X_j. For notational convenience, we set $\hat{q}_{j(0)} = -\infty$ and $\hat{q}_{j(R_j)} = \infty$. We then replace X_j by its discretized version $\tilde{X}_j$ in (2) and compute the corresponding IPOD statistic, which sheds light on the dependence between T and X_j, even when the latter is continuous. As Mai and Zou (2015) noted, discretization could be more preferable than using the continuous version, not only for computational convenience but also for added discriminative power. Moreover, without discretization, the criterion (2) requires the calculation of the pairwise-difference of the conditional density functions for all distinct X_j values, which is computationally intensive even for low-dimensional covariates, let alone for the ultrahigh dimensional settings.

Finally, we use the following criterion to select active variables

$$\widehat{\mathcal{M}}_1 = \left\{ j : \widehat{\mathcal{I}}_j^{(\gamma)} > cn^{-v},\ j = 1, \ldots, p \right\},$$

where $c > 0$ is a pre-specified constant, and term the procedure as the IPOD screening.

When X_j is continuous with infinitely many possible values, the slicing scheme may be driven by the consideration of retaining enough samples within each slice to control the estimation variance. Mai and Zou (2015) noted that the choice of slices does not affect variable screening results much, but fusion can achieve significant improvement. We consider a fusion-based IPOD screening as follows. Suppose there are N different ways of slicing X_j, denoted by $\Lambda_{ju}, u = 1, \ldots, N$, with each slicing Λ_{ju} containing R_{ju} intervals. Specifically, $\Lambda_{ju} = \left\{ [\hat{q}_{ju(r-1)}, \hat{q}_{ju(r)}) : r = 1, \ldots R_{ju}, \text{and } \cup_{r=1}^{R_{ju}} [\hat{q}_{ju(r-1)}, \hat{q}_{ju(r)}) = \mathbb{R} \right\}$, where the slicing point $\hat{q}_{ju(r)}$ is the r/R_{ju}th percentile of the empirical distribution of X_j under partition Λ_{ju}. We then replace X_j by its discretized version $\tilde{X}_{ju}$ under Λ_{ju}. That is, $\tilde{X}_{ju} = r$ if $X_j \in [\hat{q}_{ju(r-1)}, \hat{q}_{ju(r)})$. For example, we can take $N = [\log(n)] - 2$, and $R_{ju} = u + 2, u = 1, \ldots, N$ to ensure there are enough samples within each slice for all slicing schemes. Our numerical experiments suggest that at least 30 samples are needed within each slice for a reasonable estimate.

Let $\widehat{\mathcal{I}}_{j,\Lambda_{ju}}^{(\gamma)}$ be the IPOD screening statistic corresponding to the slicing scheme of Λ_{ju} for covariate X_j such that $\widehat{\mathcal{I}}_{j,\Lambda_{ju}}^{(\gamma)} = \max_{r_1,r_2} \sup_{t\in[0,\tau]} \left| \int_0^t \hat{f}_{T|\tilde{X}_j}^{\gamma}(s|\tilde{X}_{ju} = r_1)ds - \int_0^t \hat{f}_{T|\tilde{X}_j}^{\gamma}(s|\tilde{X}_{ju} = r_2)ds \right|$. Then, the fused IPOD screening statistic $\widetilde{\mathcal{I}}_j^{(\gamma)}$ is

$$\widetilde{\mathcal{I}}_j^{(\gamma)} = \sum_{u=1}^{N} \widehat{\mathcal{I}}_{j,\Lambda_{ju}}^{(\gamma)},$$

leading to the following screening criterion:

$$\widehat{\mathcal{M}}_2 = \left\{ j : \widetilde{\mathcal{I}}_j^{(\gamma)} > cn^{-v},\ j = 1, \ldots, p \right\}, \quad (3)$$

where $c > 0$ is a constant. As our numerical experiment suggests that the fused method performs better than the single slicing-based, we opt to use (3) as the screening criterion in practice.

3. Sure Screening Properties

To establish the sure screening property, we need to set regularity conditions. First, we stipulate the conditions for when all the variables are categorical.

(C1) $P(T > \tau|X_j) > \theta_1 > 0$ for $1 \le j \le p$, where θ_1 is a positive constant.

(C2) For any $t \in [0, \tau]$, $f_{T|X_j}(t|r)$ is greater than a positive constant $\tilde{c}_0$, and has a bounded second order derivative for $r \in \{1, \ldots, R_j\}$, $j \in \mathcal{M}$.

(C3) There exist $c > 0$ and $0 < v < 1/2$ such that $\min_{j\in\mathcal{M}} \mathcal{I}_j^{(\gamma)} \geq 2cn^{-v}$ for a specific γ.

(C4) $K(\cdot)$ is a symmetric probability density function defined on a bounded support with bounded variation $V_K < \infty$ and bandwidth $h_n = O(n^{-\mu})$, where $\mu > 0$ is a constant.

(C5) $R = \max_{1\leq j\leq p} R_j = O(n^{\kappa})$, where R_j is the number of categories for X_j $(j = 1, \ldots, p)$, $\kappa \geq 0$, and $2v + 3\kappa + 2\mu < 1$.

Condition (C1) is imposed to avoid problems with estimating the tail of the conditional survival functions, which is common in survival analysis; see Dabrowska (1989). Conditions (C2) and (C4) are assumed for the kernel function involved in conditional density estimation of censored data, which are common in the nonparametric literature; see Lo et al. (1989) and Chen et al. (2015). Condition (C3) is typical in the feature screening literature; see Condition 3 in Fan and Lv (2008) and Ni and Fang (2016). Condition (C5) allows the number of the classes for covariates diverge with a specific order. A similar assumption was also made in Ni and Fang (2016).

THEOREM 1. *If all the covariates are categorical, under conditions (C1)–(C5), we have*

$$P(\mathcal{M} \subset \widehat{\mathcal{M}}_1) \geq 1 - O(p \exp(-b_0 n^{1-3\kappa-2v-2\mu} + \kappa \log n)),$$

where b_0 is a positive constant. Hence, if $\log p = O(n^{\alpha})$ where $0 < \alpha < 1 - 3\kappa - 2v - 2\mu$, IPOD has the sure screening property.

COROLLARY 1. *Under the conditions of Theorem 1, and assuming $\sum_{j=1}^{p} \mathcal{I}_j^{(\gamma)} = O(n^{\zeta})$ for some $\zeta > 0$, we have*

$$P\{|\widehat{\mathcal{M}}_1| \leq O(n^{\zeta+v})\} \geq 1 - O(p \exp(-b_0 n^{1-3\kappa-2v-2\mu} + \kappa \log n)).$$

If X_j is continuous, we replace condition (C5) by the following:

(C6) $R = \max_{1\leq j\leq p} R_j = O(n^{\kappa})$, where R_j is the number of slices for X_j. Moreover, there exist a positive constant c_1 and $0 \leq \rho < 1/2$ such that $2v + 3\kappa + 2\mu + 2\rho < 1$ and $f_{X_j}(x) \geq c_1 n^{-\rho}$ for any $1 \leq j \leq p$, where $f_{X_j}(x)$ is continuous and bounded from above on the support of X_j.

THEOREM 2. *When the covariates include both continuous and categorical types, under conditions (C1)–(C4) and (C6), we have*

$$P(\mathcal{M} \subset \widehat{\mathcal{M}}_1) \geq 1 - O(p \exp(-b_1 n^{1-3\kappa-2v-2\mu-2\rho} + \kappa \log n)),$$

where b_1 is a positive constant. Hence, if $\log p = O(n^{\alpha})$ and $0 < \alpha < 1 - 3\kappa - 2v - 2\mu - 2\rho$, IPOD has the sure screening property.

COROLLARY 2. *Under the conditions of Theorem 2 and assuming $\sum_{j=1}^{p} \mathcal{I}_j^{(\gamma)} = O(n^{\zeta})$ for some $\zeta > 0$, we have*

$$P(|\widehat{\mathcal{M}}_1| \leq O(n^{\zeta+v})) \geq 1 - O(p \exp(-b_1 n^{1-3\kappa-2v-2\mu-2\rho} + \kappa \log n)).$$

Let Λ_{juo} be the partition using the theoretical quantiles $q_{ju(r)}, r = 0, \ldots, R_{ju}$ of X_j as the slicing points. Denote the true value of IPOD for the specific partition Λ_{juo} by $\mathcal{I}_{j,\Lambda_{juo}}^{(\gamma)}$ and let $\mathcal{I}_{jo}^{(\gamma)} = \sum_{u=1}^{N} \mathcal{I}_{j,\Lambda_{juo}}^{(\gamma)}$. The fused IPOD screening method needs some different regularity conditions:

(C7) There exist a $c > 0$ and $0 < v < 1/2$ such that $\min_{j\in\mathcal{M}} \mathcal{I}_{jo}^{(\gamma)} \geq 2cn^{-v}$ for a specific γ.

(C8) Let $R = \max_{1\leq j\leq p, 1\leq u\leq N} R_{ju}$ and assume $R = O(n^{\kappa})$. There exist a positive constant c_3 and $0 \leq \rho < 1/2$ such that $2v + 3\kappa + 2\mu + 2\rho < 1$, and $f_{X_j}(x) \geq c_3 n^{-\rho}$ for any $1 \leq j \leq p$ and $f_{X_j}(x)$ is bounded from above and continuous with respect to x.

THEOREM 3. *When the covariates include both continuous and categorical types, under conditions (C1)–(C2), (C4), and (C7)–(C8), we have*

$$P(\mathcal{M} \subset \widehat{\mathcal{M}}_2) \geq 1 - O(Np \exp(-b_2 n^{1-3\kappa-2v-2\mu-2\rho} + \kappa \log n)),$$

where b_2 is a positive constant. If $N = O(\log n)$ and $\log p = O(n^{\alpha})$ where $0 < \alpha < 1 - 3\kappa - 2v - 2\mu - 2\rho$, then the fused IPOD has the sure screening property.

COROLLARY 3. *Under the conditions of Theorem 3 and assuming $\sum_{j=1}^{p} \mathcal{I}_{jo}^{(\gamma)} = O(n^{\zeta})$ for some $\zeta > 0$, we have*

$$P\{|\widehat{\mathcal{M}}_2| \leq O(n^{\zeta+v})\} \geq 1 - O(Np \exp(-b_2 n^{1-3\kappa-2v-2\mu-2\rho} + \kappa \log n)).$$

4. Simulation Studies

The finite sample performance of the proposed method was assessed by comparing it with the following methods that are often used for screening survival data.

- PSIS: the principled sure independence screening for Cox models by Zhao and Li (2012).
- CRIS: the censored rank independence screening proposed by Song et al. (2014).
- CS: the conditional screening for survival outcome by Hong et al. (2016).
- SII: the survival impact index by Li et al. (2016) with the uniform weight $W(t, x) = 1$.
- IPOD (γ): the proposed screening with power index γ. Note that $\gamma = 1$ corresponds to the Kolmogorov statistic.

EXAMPLE 1. *The survival time was generated from the proportional hazards model,*

$$h(t|\boldsymbol{X}) = 0.1 \exp\left\{\sum_{j=1}^{p} \beta_j I(X_j \in \{2, 3\})\right\},$$

where $\beta = (\mathbf{0}.5_5, \mathbf{0}_{p-5}^{\mathrm{T}})^{\mathrm{T}}$. *The covariates* $x^\star$ *underlying these discrete variables were generated from a multivariate normal distribution with mean 0 and a covariance matrix* $\Sigma = (\sigma_{jj'})_{p\times p}$, *where* $\sigma_{jj} = 1$ *and* $\sigma_{jj'} = 0.5$ *for* $j \neq j'$. *For each* j, x_j^* *was further quarterized by its quartile values: the obtained quarterly variable* $X_j = 1$ *if* x_j^* *is less than the lower quartile, 2 if between the lower quartile and the median, 3 if between the median and the upper quartile, and* 4 *otherwise.*

EXAMPLE 2. *The survival time was generated from the Cox model* $\lambda(t|\boldsymbol{X}) = 2t(|X_1| + |X_2|)$, *where all the covariates* X_j, $j = 1, \ldots, p$, *were generated from an independent standard normal distribution. In this case, the marginal correlation between each of the active variables,* X_1 *and* X_2, *and the survival time is 0.*

EXAMPLE 3. *The survival time was generated from* $\log(t) = g_1(X_1) + g_2(X_2) + g_3(X_3) + g_4(X_4) + \epsilon$, *where* $g_1(x) = 5x$, $g_2(x) = -4x(1-x)$, $g_3(x) = 10[\exp\{-3(x-1)^2\} + \exp\{-4(x-3)^2\}] - 1.5$, *and* $g_4(x) = 4\sin(2\pi x)$. *The vector of covariates* $\boldsymbol{X}$ *was generated from the multivariate normal distribution with mean 0 and a covariance matrix* $\Sigma = (\sigma_{jj'})_{p\times p}$, *with* $\sigma_{jj} = 1$ *and* $\sigma_{jj'} = \rho^{|j-j'|}$ *for* $j \neq j'$, *and* $\epsilon \sim N(0,1)$ *is independent of* $\boldsymbol{X}$. *The censoring time* C *was generated from a three-component normal mixture distribution* $N(0,4) - N(5,1) + 0.5N(25,1)$.

EXAMPLE 4. *The survival time was generated from* $\log(t) = g_1(X_1) + g_2(X_2) + g_3(X_3) + g_4(X_4) + 0.3(X_5 + X_6 + X_7 + X_8 + X_9 + X_{10}) + \epsilon$, *where all other conditions are the same as in Example 3.*

In all examples, we used $p = 1000$ and $n = 300$ and 500. In Examples 1–2, the censoring times C_i were independently generated from a uniform distribution $U[0, c]$, with c chosen to give approximately 20% and 50% of censoring proportions.

Table 1

Comparisons of competing methods with $(n, p) = (300, 1000)$ *and* $(n, p) = (500, 1000)$ *in terms of the minimum model size to ensure inclusion of the true model (MMS) with interquartile range in parentheses, the true positive rate (TPR), and the probability of including all active variables (PIT)*

	MMS	TPR	PIT	MMS	TPR	PIT	MMS	TPR	PIT	MMS	TPR	PIT
Method	$(n, p) = (300, 1000)$			$(n, p) = (500, 1000)$			$(n, p) = (300, 1000)$			$(n, p) = (500, 1000)$		
Example 1			CR = 20%						CR= = 50%			
IPOD ($\gamma = .8$)	201 (254)	0.63	0.11	59 (92)	0.91	0.60	298 (279)	0.50	0.02	115 (165)	0.82	0.38
IPOD ($\gamma = 1$)	161 (218)	0.70	0.16	41 (71)	0.94	0.71	254 (277)	0.57	0.06	182 (135)	0.87	0.49
IPOD ($\gamma = 1.2$)	159 (217)	0.71	0.17	38 (65)	0.94	0.73	234 (273)	0.59	0.08	72 (112)	0.89	0.56
PSIS	855 (211)	0.06	0.00	858 (217)	0.09	0.00	874 (91)	0.05	0.00	849 (198)	0.08	0.00
CRIS	919 (133)	0.03	0.00	921 (128)	0.04	0.00	913 (155)	0.03	0.00	914 (135)	0.04	0.00
CS	834 (245)	0.24	0.00	830 (230)	0.25	0.00	842 (218)	0.24	0.00	837 (219)	0.26	0.00
SII	258 (274)	0.51	0.02	82 (110)	0.86	0.49	327 (277)	0.41	0.01	136 (181)	0.75	0.28
Example 2			CR = 20%						CR = 50%			
IPOD ($\gamma = .8$)	2 (2)	0.99	0.99	2 (0)	1.00	1.00	5 (10)	0.94	0.89	2 (0)	1.00	1.00
IPOD ($\gamma = 1$)	2 (3)	0.99	0.99	2 (0)	1.00	1.00	4 (11)	0.94	0.89	2 (0)	1.00	1.00
IPOD ($\gamma = 1.2$)	2 (4)	0.99	0.99	2 (0)	1.00	1.00	5 (15)	0.95	0.90	2 (1)	1.00	1.00
PSIS	721 (330)	0.03	0.00	727 (349)	0.06	0.01	700 (392)	0.07	0.01	714 (371)	0.08	0.01
CRIS	738 (339)	0.04	0.00	707 (325)	0.07	0.01	732 (324)	0.06	0.00	706 (324)	0.06	0.00
CS	461 (532)	0.54	0.09	426 (502)	0.55	0.11	457 (571)	0.56	0.12	459 (557)	0.56	0.12
SII	19 (27)	0.90	0.84	3 (4)	1.00	1.00	67 (119)	0.63	0.43	11 (32)	0.95	0.92
Example 3			$\rho = 0$						$\rho = 0.8$			
IPOD ($\gamma = 0.8$)	84 (173)	0.85	0.39	22 (51)	0.96	0.83	6 (10)	0.98	0.93	4 (0)	1.00	1.00
IPOD ($\gamma = 1$)	114 (226)	0.82	0.30	35 (84)	0.92	0.69	8 (19)	0.97	0.90	4 (1)	1.00	1.00
IPOD ($\gamma = 1.2$)	158 (280)	0.81	0.24	61 (132)	0.89	0.57	14 (36)	0.95	0.80	4 (3)	0.99	0.98
PSIS	560 (466)	0.61	0.03	539 (429)	0.67	0.03	330 (624)	0.50	0.17	206 (454)	0.66	0.36
CRIS	619 (415)	0.57	0.02	579 (505)	0.62	0.03	314 (591)	0.53	0.16	183 (384)	0.73	0.37
CS	703 (505)	0.55	0.03	659 (518)	0.62	0.04	886 (993)	0.74	0.38	27 (109)	0.92	0.69
SII	376 (405)	0.69	0.04	296 (330)	0.77	0.12	24 (82)	0.90	0.64	8 (14)	0.99	0.98
Example 4			$\rho = 0$						$\rho = 0.8$			
IPOD ($\gamma = 0.8$)	867 (166)	0.38	0.00	863 (178)	0.47	0.00	57 (166)	0.89	0.49	13 (19)	0.99	0.90
IPOD ($\gamma = 1$)	876 (171)	0.37	0.00	872 (156)	0.45	0.00	96 (253)	0.84	0.36	22 (47)	0.97	0.81
IPOD ($\gamma = 1.2$)	881 (171)	0.36	0.00	868 (178)	0.43	0.00	155 (347)	0.78	0.23	44 (110)	0.94	0.64
PSIS	895 (162)	0.29	0.00	882 (182)	0.36	0.00	102 (267)	0.84	0.37	31 (80)	0.96	0.73
CRIS	925 (121)	0.25	0.00	926 (112)	0.28	0.00	111 (277)	0.82	0.32	24 (51)	0.97	0.80
CS	892 (150)	0.31	0.00	877 (167)	0.38	0.00	774 (305)	0.51	0.00	689 (366)	0.68	0.01
SII	870 (163)	0.31	0.00	874 (192)	0.38	0.00	33 (95)	0.92	0.58	12 (11)	0.99	0.93

Example 3 was adopted from Li et al. (2016). The censoring proportions of Examples 3–4 were set around 35%. In addition, we explored the simulation studies when the censoring times C_i were dependent on covariates; see Table S1 in the Supplementary Web Materials.

The bandwidth h_n was chosen to be $h_0 n^{-1/5}$ with $h_0 = 2$ based on the exploratory analysis reported in Figure S1 in the Supplementary Web Materials; when covariate X_j was continuous, we took various slicings $(3, \ldots, \lceil \log(n) \rceil)$ and reported the fused results.

As the conditional screening (Hong et al., 2016) requires prior information, we chose X_1 as the conditioning variable for Examples 1–4. For each configuration, a total of 500 simulated datasets were generated. We considered the minimum model size (MMS) to ensure inclusion of the true model, the true positive rate (TPR), and the probability of including all active variables (PIT) as metrics to compare the performance between different methods. To compute the TPR and PIT, we selected the first $[n/\log n]$ variables in relation to selection criteria. In general, smaller MMS and larger TPR and PIT indicate better performance of a method in efficiently discovering true signals.

Table 1 demonstrates that the proposed IPOD method worked well in a variety of settings, with improving performance as the sample size increased. This is an appealing feature not necessarily shared by the competing methods. For IPOD, we investigated $\gamma = 0.8, 1, 1.2, 1.5, 1.7$, but we only presented here the results of $\gamma = 0.8, 1, 1.2$ as they well represented the overall results. More results can be found in Figure S1 of the Supplementary Web Materials.

With active variables all being categorical as in Example 1, the results for all competing methods were poor since these methods were not originally developed for screening categorical variables. When the marginal correlations between the active variables and the survival time were 0 as in Examples 2, all the competing methods, including CS that even assumed one active variable was known, had difficulty recruiting active variables. In Example 3, where the active covariates have nonlinear relationships with the response variable, IPOD worked better than the other methods, especially when ρ was high. When the size of active variables was moderate as in Example 4, the performance was poor for all methods when the covariates were independent of each other, but was better with higher correlations among the covariates.

Finally, it is of interest to note that the "optimal" γ varied across examples. However, this optimal γ is rarely known in reality. Illustrated by a real example in the next section, we briefly discuss how to combine results from the IPOD with different γ's.

5. An Application

We applied the proposed methods to analyze a multiple myeloma (MM) study (Shaughnessy et al., 2007), concerning the development and evaluation of a gene-based prognostic tool among 554 newly diagnosed MM patients treated on two separate but similar protocols, namely, total treatment 2 (TT2) and total treatment 3 (TT3). The former served as the training set ($n_1 = 340$ patients), while the latter served as the validation set ($n_2 = 214$ patients). The outcome for our investigation is event-free survival. Gene expressions on 54,675 probe sets were measured for each subject using Affymetrix U133Plus2.0 microarrays. Prior to analysis, we standardized the gene expressions.

For comparisons, we also applied the other competing methods, including PSIS, CRIS, SII, and CS to screen genes that may be relevant to event-free survival. Based on our investigation in the simulation setup, we chose $\gamma = 0.7, 1.0, 1.3, 1.5, 1.7$ for our IPOD method. In addition, we also considered the overlapping genes selected by all these γ's and termed the way of obtaining these genes as composite IPOD. We took the bandwidth to be $2n_1^{-1/5}$. Since the gene expressions were continuous, we chose the combination of $\Lambda = 3, \ldots, \lceil \log(340) \rceil = 6$ for slicing in the proposed IPOD. Moreover, as CS requires a pre-specified conditioning set, we identified the candidates as the overlapping genes selected by composite IPOD, PSIS, SII, and CRIS. These were Probes 225834, 218595, 206332, 208965, and 208966. In our analysis, we used various subsets of these genes as the conditioning sets. We found that CS was sensitive to the choice, leading to quite variable predictive performance; see Table 3.

In a finite sample setting, there are no established ways to select v in (3). Instead, practitioners commonly use the screening statistic to rank variables and select the top variables. We used $[n/\log(n)]$ as suggested by Fan and Lv (2008), which may

Table 2
Numbers of overlapping genes selected by different screening methods on the multiple myeloma training set

	IPOD									
	$\gamma = 0.7$	$\gamma = 1$	$\gamma = 1.3$	$\gamma = 1.5$	$\gamma = 1.7$	Composite	PSIS	CRIS	CS	SII
IPOD										
$\gamma = 0.7$	58									
$\gamma = 1$	40	58								
$\gamma = 1.3$	29	46	58							
$\gamma = 1.5$	24	39	51	58						
$\gamma = 1.7$	20	34	46	53	58					
Composite	20	20	20	20	20	20				
PSIS	18	23	20	17	16	10	58			
CRIS	0	0	1	1	1	0	1	58		
CS	11	10	10	8	7	6	18	5	58	
SII	6	5	5	5	5	5	7	5	15	58

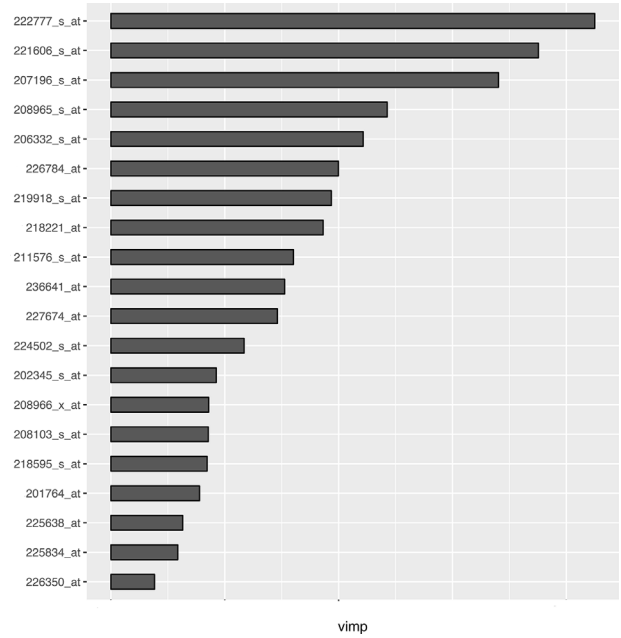

Figure 2. The ranking of variable importance (vimp) among those selected by the composite IPOD method for the multiple myeloma training set.

have become a convention in the follow-up screening works; see Zhao and Li (2012), He et al. (2013), Gorst-Rasmussen and Scheike (2013), and Li et al. (2016). To proceed, we first applied each screening method to the training data to reduce the dimension from $p = 54{,}675$ to $d = [n_1/\log(n_1)] = 58$.

Table 2 shows the numbers of overlapping genes selected by different methods and reveals that IPOD with different γ's ended up with various genes.

We next examined how variable screening helps predict the response variable by using random survival forest (Ishwaran et al., 2008), a nonparametric machine learning strategy for building a predictive survival model. We fitted a random survival forest model, based on a total of 1000 replicate trees by using the `R` package, namely, `randomSurvivalForest` (Ishwaran and Kogalur, 2007). In the interest of space, we reported the results of the composite IPOD method, for which the importance of each selected variable in the "forest" was shown in Figure 2.

To measure the prediction accuracy, we reported the C-index, which gauges the agreement between the observed and predicted responses. The average C-index based on 100 bootstrap replicates of the validation data is listed in Table 3.

Overall, IPOD performed better than or at least on par with the other competing methods, and the IPOD with a larger γ tends give better predictions for this particular dataset.

Finally, we highlight some biological insight offered by our method. Probe 222777 (WHSC1), the most important gene as identified in Figure 2, has been found involved in the chromosomal translocation in multiple myeloma (Xie et al., 2013). Over-expression of Probe 221606 (HMGN5), the second on the list, has been linked to bladder cancer (Gan et al., 2015), prostate cancer (Guo et al., 2015) and breast cancer (Weng et al., 2015). On the other hand, neoplastic B-cell growth is regulated by gene TNIP1 (Probe 207196, the third gene of importance), whose dysfunction may cause multiple myeloma (Naji et al., 2012). Moreover, reduction or loss of IFI16 expression (Probe 208965, the fourth gene in the list) in cells is associated with the development of certain cancers, such as breast and prostate cancer (Choubey et al., 2008). Finally, CHML (Probe 226350, the last gene shown in Figure 2) is able to induce apoptosis or programmed cancer cell death and suppress tumor cell growth in multiple tumor lines (Zhan et al., 2001).

6. Discussion

Motivated by a multiple myeloma genomic study, we introduced a new framework for variable screening based on IPOD. As the method is model-free, it can be applicable in a variety of parametric, semiparametric, and nonparametric settings. In addition, it is theoretically justifiable, and computationally efficient. Using the proposed method, we identified a predictive gene signature model which was more accurate than the models obtained by using other screening methods.

Our work enlightens a few future directions. As our simulations revealed, the optimal γ may vary by the particular simulation configuration. Without knowing the true model in reality, it is challenging, conceptually and computationally, to identify an optimal γ, even in a data-driven way. However, the established framework may lead to some systematic ways of combining the results from IPOD with various γ's. For example, we are currently investigating the means of combining the results of L possible values of γ's using the following

Table 3

Comparisons of the C-index (standard errors) in the multiple myeloma validation set based on 100 bootstraps

IPOD					
$\gamma = .7$	$\gamma = 1$	$\gamma = 1.3$	$\gamma = 1.5$	$\gamma = 1.7$	Composite
0.639 (0.003)	0.638 (0.003)	0.655 (0.003)	0.651 (0.003)	0.657 (0.003)	0.652 (0.003)
PSIS	CRIS	CS	SII		
0.620 (0.003)	0.602 (0.004)	0.641 (0.004)	0.638 (0.003)		

composite screening statistics:

$$\mathcal{I}_{1j} = \max_{1\le l\le L} \tilde{\mathcal{I}}_j^{(\gamma_l)},\ \mathcal{I}_{2j} = \min_{1\le l\le L} \tilde{\mathcal{I}}_j^{(\gamma_l)},\ \mathcal{I}_{3j} = \frac{\tilde{\mathcal{I}}_j^{(\gamma)} - \mathcal{I}_{2j}}{\mathcal{I}_{1j} - \mathcal{I}_{2j}},$$

where $\gamma_l,\ l = 1, \ldots, L$. Each option has pros and cons. In $\mathcal{I}_{1j}$, covariates selected by any $\gamma_l,\ l = 1, \ldots, L$ are to be included in the final selected set. This method could guarantee recovery of the true active set to the greatest extent in theory, though at the cost of inflating false discoveries. In $\mathcal{I}_{2j}$, only the covariates which are selected by all $\gamma_l,\ l = 1, \ldots, L$ could be included in the final selected set. This method could guarantee exclusion of the unimportant covariates to the greatest extent. However, the rather restrictive criterion may lead to many false negatives, which may not be ideal for knowledge discovery at the exploratory phase. The third option $\mathcal{I}_{3j}$ may be a compromise between these two extreme cases. Through rescaling to be a number between 0 and 1, it makes screening statistics across γ comparable. Comprehensively evaluating and studying all these proposals are currently undergoing and may be out of scope of this current article. We expect to report the results elsewhere.

7. Supplementary Materials

Web Appendices and Tables referenced in Sections 1 and 4 are available with this article at the *Biometrics* website on Wiley Online Library. The multiple myeloma dataset that we used for analysis is publicly available at the Gene Expression Omnibus website (`http://www.ncbi.nlm.nih.gov/geo/`) under GSE24080.

Acknowledgements

We thank our editorial assistant, Ms. Martina Fu, for proofreading the manuscript. We thank the editor, the AE and an anonymous referee for insightful comments. The research was supported in part by the grants from the National Security Agency (H98230-15-1-0260, Hong), the Fundamental Research Funds for the Central Universities (JBK140507, JBK120509, Chen), the National Natural Science Foundation of China (11501461, Chen; 11528102, Li), and the National Institutes of Health (U01CA209414, Christiani and Li).

References

Beyene, J., Atenafu, E. G., Hamid, J. S., To, T., and Sung, L. (2009). Determining relative importance of variables in developing and validating predictive models. *BMC Medical Research Methodology* **9**, 64.

Chen, X., Wan, A. T., and Zhou, Y. (2015). Efficient quantile regression analysis with missing observations. *Journal of the American Statistical Association* **110**, 723–741.

Choubey, D., Deka, R., and Ho, S. (2008). Interferon-inducible IFI16 protein in human cancers and autoimmune diseases. *Frontiers in Bioscience* **1**, 598–608.

Cover, T. M. and Thomas, J. A. (2012). *Elements of Information Theory*. New York: John Wiley & Sons.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. *Annals of Statistics* **17**, 1157–1167.

Fan, J., Feng, Y., and Wu, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model. *IMS Collections Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown* **6**, 70–86.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of Royal Statistical Society B* **70**, 849–911.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20**, 101–148.

Fossella, F. V., DeVore, R., Kerr, R. N., Crawford, J., Natale, R. R., Dunphy, F., et al. (2000). Randomized phase III trial of docetaxel versus vinorelbine or ifosfamide in patients with advanced non-small-cell lung cancer previously treated with platinum-containing chemotherapy regimens. *Journal of Clinical Oncology* **18**, 2354–2362.

Gan, Y., Tan, J., Yang, J., Zhou, Y., Dai, Y., He, L., et al. (2015). Knockdown of HMGN5 suppresses the viability and invasion of human urothelial bladder cancer 5637 cells in vitro and in vivo. *Medical Oncology* **32**, 136.

Gorst-Rasmussen, A. and Scheike, T. (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of the Royal Statistical Society B* **75**, 217–245.

Guo, Z., Zhang, X., Li, X., Xie, F., Su, B., Zhang, M., et al. (2015). Expression of oncogenic HMGN5 increases the sensitivity of prostate cancer cells to gemcitabine. *Oncology Reports* **33**, 1519–1525.

He, X., Wang, L., and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics.* **41**, 342–369.

Heinzel, A., Perco, P., Mayer, G., Oberbauer, R., Lukas, A., and Mayer, B. (2014). From molecular signatures to predictive biomarkers: modeling disease pathophysiology and drug mechanism of action. *Frontiers in Cell and Developmental Biology* **2**, 37.

Hong, H. G., Kang, J., and Li, Y. (2016). Conditional screening for ultra-high dimensional covariates with survival outcomes. *Lifetime Data Analysis* https://doi.org/10.1007/s10985-016-9387-7.

Ishwaran, H. and Kogalur, U. (2007). Random survival forests for R. *Rnews* **7**, 25–31.

Ishwaran, H., Kogalur, U., Blackstone, E., and Lauer, M. (2008). Random survival forests. *Annals of Applied Statistics* **2**, 841–860.

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012). Robust rank correlation based screening. *Annals of Statistics* **40**, 1846–1877.

Li, J., Zheng, Q., Peng, L., and Huang, Z. (2016). Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes. *Biometrics* **72**, 1145–1154.

Lin, C.-Y. and Halabi, S. (2013). On model specification and selection of the cox proportional hazards model. *Statistics in Medicine* **32**, 4609–4623.

Liu, J., Li, R., and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh dimensional covariates. *Journal of the American Statistical Association* **109**, 266–274.

Lo, S. H., Mack, Y. P., and Wang, J. L. (1989). Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator. *Probability Theory and Related Fields* **80**, 461–473.

Mai, Q. and Zou, H. (2015). The fused Kolmogorov filter: A nonparametric model-free screening method. *Annals of Statistics* **43**, 1471–1497.

Mulligan, G., Mitsiades, C., Bryant, B., Zhan, F., Chng, W., Roels, S., et al. (2007). Gene expression profiling and correlation

with outcome in clinical trials of the proteasome inhibitor bortezomib. *Blood* **109**, 3177–3188.

Naji, A., Menier, C., Maki, G., Carosella, E. D., and Rouas-Freiss, N. (2012). Neoplastic B-cell growth is impaired by HLA-G/ILT2 interaction. *Leukemia* **26**, 1889–1892.

National Cancer Policy Board. (2003). Late effects of childhood cancer. In *Childhood Cancer Survivorship: Improving Care and Quality of Life*, M. Hewitt, S. Weiner, and J. Simone (eds). Washington, DC: National Academies Press.

Ni, L. and Fang, F. (2016). Entropy-based model-free feature screening for ultrahigh-dimensional multiclass classification. *Journal of Nonparametric Statistics* **28**, 515–530.

Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Annals of Statistics* **109**, 1302–1318.

Shaughnessy, J. D., Zhan, F., Burington, B. E., Huang, Y., Colla, S., Hanamura, I., et al. (2007). A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood* **109**, 2276–2284.

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

Tibshirani, R. J. (2009). Univariate shrinkage in the Cox model for high dimensional data. *Statistical Applications in Genetics and Molecular Biology* **8**, 21.

Weng, M., Song, F., Chen, J., Wu, J., Qin, J., Jin, T., et al. (2015). The high-mobility group nucleosome-binding domain 5 is highly expressed in breast cancer and promotes the proliferation and invasion of breast cancer cells. *Tumor Biology* **36**, 959–966.

Xie, Z., Gunaratne, J., Cheong, L. L., Liu, S. C., Koh, T. L., Huang, G., et al. (2013). Plasma membrane proteomics identifies biomarkers associated with MMSET overexpression in T(4;14) multiple myeloma. *Oncotarget* **4**, 1008–1018.

Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society, Series B* **69**, 507–564.

Zhan, Q., Zhao, S., and Xu, Z. (2001). Antitumor activity of cytotropic heterogeneous molecular lipids (CHML) on human breast cancer xenograft in nude mice. *Anticancer Research* **21**, 2477–2482.

Zhao, S. D. and Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis* **105**, 397–411.

Zhu, L., Li, L., Li, R., and Zhu, L. (2011). Model-free feature screening for ultrahigh dimensional data. *Journal of the American Statistical Association* **106**, 1464–1475.

Received April 2017. *Revised September* 2017.
Accepted October 2017.