The L_q-norm learning for ultrahigh-dimensional survival data:

an integrative framework

H. G. Hong[†], X. Chen[‡], J. Kang[*], and Y. Li[*]

[†] *Michigan State University, USA*

[‡] *Southwestern University of Finance and Economics, China,*

[*] *University of Michigan, USA*

Abstract: In the era of precision medicine, survival outcome data with high-throughput predictors are routinely collected. Models with an exceedingly large number of covariates are either infeasible to fit or likely to incur low predictability because of overfitting. Variable screening is key in identifying and removing irrelevant attributes. Recent years have seen a surge in screening methods, but most of them rely on some particular modeling assumptions. Motivated by a study on detecting gene signatures for multiple myeloma patients' survival, we propose a model-free L_q-norm learning procedure, which includes the well-known Cramér–von Mises and Kolmogorov criteria as two special cases. The work provides an integrative framework for detecting predictors with various levels of impact, such as short- or long-term impact, on censored outcome data. The framework naturally leads to a scheme which combines results from different q to reduce false negatives, an aspect often overlooked by the current literature. We show that our method possesses sure screening properties. The utility of the proposal is confirmed with simulation studies and an analysis of the multiple myeloma study.

Key words and phrases: L_q-norm learning, Kolmogorov statistic, Cramér–von Mises statistic, survival data, variable screening.

1. Introduction

The emergence of high-throughput data arising from genomic, genetic, and clinical stud-

ies has presented unique opportunities for discovering relevant information on patients' survival from massive databases. As the scientific investigation often focuses on discerning lower-dimensional presentations of a high-dimensional feature space that preserve the necessary information to predict survival outcomes, new efficient and reliable methods are needed to select relevant variables. In ultrahigh-dimensional settings where the number of predictors grows exponentially with the sample size, feature screening has become a key analytical step in ensuring computational expediency, statistical accuracy and algorithm stability (Fan et al., 2010). For example, in a motivating clinical study (Avet-Loiseau et al., 2009) on multiple myeloma patients, understanding the molecular etiology of this disease, such as detecting the gene signatures that are relevant to survival, would lead to a more accurate risk classification system and personalized treatment (Mulligan et al., 2007). However, with gene expression measurements on more than 40,000 probe sets, this dataset challenges the existing statistical tools for dimension reduction.

Despite the success of many screening approaches, such as the sure independence screening (Fan and Lv, 2008) and its follow-up works, limited progress has been achieved in developing ultrahigh-dimensional screening tools for survival outcomes. Some related works include a sure screening procedure for proportional hazards models (Fan et al., 2010), a Cox univariate shrinkage estimator (Tibshirani, 2009), a marginal maximum partial likelihood estimator (Zhao and Li, 2012), and a general class of single-index hazard rate statistics (Gorst-Rasmussen and Scheike, 2013). Going beyond marginal regressions, Hong et al. (2018) proposed a conditional screening approach when prior information is available on which variables should be included in the models. However, the validity and usability of these methods often hinge upon some restrictive modeling assumptions.

Model-free screening procedures have recently emerged as a useful tool to avoid these restrictions. Some representative works include a censored rank independence screening method (Song et al., 2014) and a quantile adaptive method (He et al., 2013). These methods are typically robust against outliers in predictors and are applicable to a wide range of survival models. However, they are often computationally intensive and are not designed to handle discrete predictors, which often appear in practice.

The Kolmogorov screening statistic, which compares distribution functions across covariate-defined strata, has been proposed for screening nominal predictors (Mai and Zou, 2015; Zhu et al., 2012). When the outcome data are censored, it is unclear how the method would fare in terms of implementation, interpretation and theoretical justifications. On the other hand, the Cramér–von Mises statistic was developed for detecting distribution differences across various subpopulations in the presence of censoring. For example, Schumacher (1984) demonstrated that the Cramér–von Mises test is superior to log-rank tests when the proportional hazards assumption fails to hold; see also Koziol and Green (1976); Stute (1997); Tamura et al. (2000); Li and Feng (2005) in various contexts. Several authors have also shown that, under general situations such as when a covariate has a long lasting impact on the survival, the Cramér–von Mises statistic may be more powerful than the Kolmogorov statistic in detecting such an impact (Conover and Conover, 1980; Razali et al., 2011; Woodruff and Moore, 1988; Arnold and Emerson, 2011; Chiu and Liu, 2009). However, none of these works explored the Cramér–von Mises statistic for variable screening with censored outcome data.

As the goal of nonparametric screening is to detect the difference between survival functions for the subpopulations or strata defined by each candidate variable, both the Kolmogorov and the Cramér–von Mises statistics are applicable. An often overlooked fact, how-
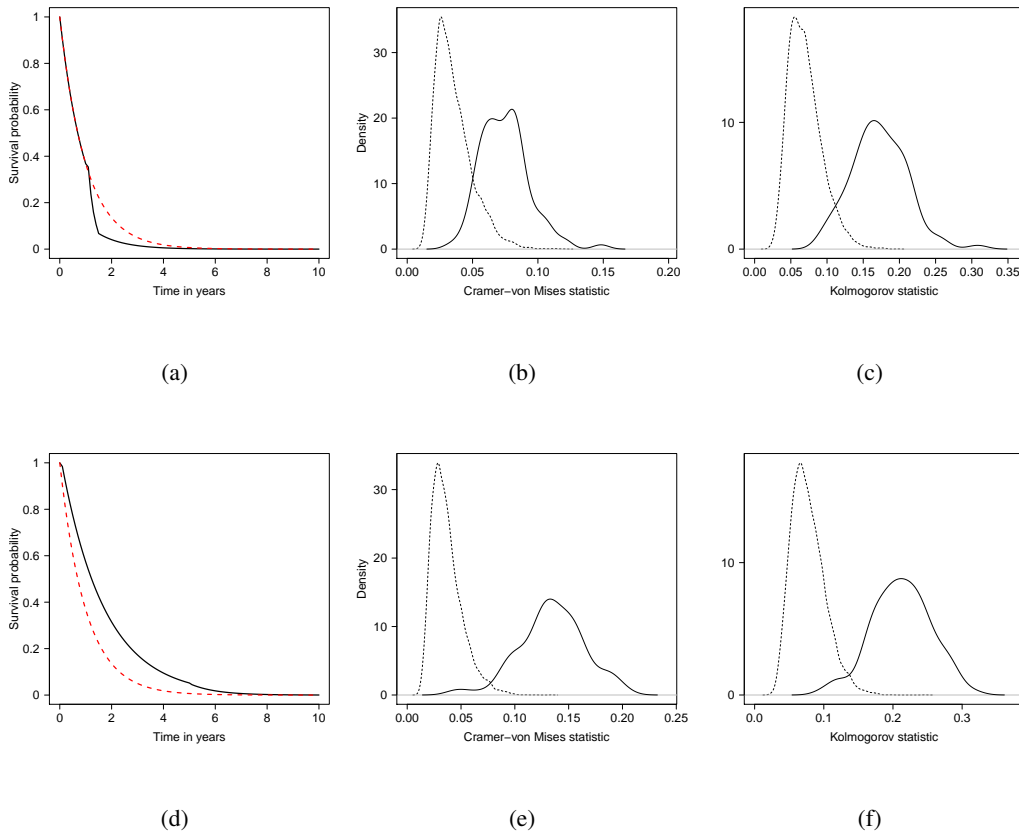
Figure 1: Comparisons of the Cramér–von Mises and Kolmogorov screening statistics in two hypothetical scenarios when only X_1 has an impact. Survival curves for $X_1 = 0$ and $X_1 = 1$ are shown as solid and dashed curves in Figure 1(a): Group $X_1 = 0$ has a constant hazard $h(t) = 1$, and Group $X_1 = 1$ has $h(t) = 4$ for $t \in (1.0, 1.4)$ and $h(t) = 1$, elsewhere. Figure 1(b) shows the density curves of the Cramér–von Mises statistics on the active variable (solid curves) and 100 independent noise variables (dashed curves), based on 100 simulations. Figure 1(c) presents the Kolmogorov statistics. Figure 1(c) indicates a more clear separation than Figure 1(b), meaning that the Kolmogorov statistic was more powerful than the Cramér–von Mises statistic in this setting. In Figure 1(d), Group $X_1 = 0$ has a constant hazard $h(t) = 1$, and Group $X_1 = 1$ has $h(t) = 0.6$ for $t \in (0.01, 5)$ and $h(t) = 1$, elsewhere. Figures 1(e) and 1(f) represent the Cramér–von Mises and Kolmogorov statistics, respectively, under the setting of Figure 1(d). Figure 1(e) shows a more clear separation between active and noise variables than Figure 1(f).

ever, is that the difference patterns may vary across covariates: while some covariates may be impactful during the entire span of follow-up, some covariates may only have short-term impacts such as in the case illustrated in Figure 1. For example, the survival differences between the chemotherapy group and the chemotherapy plus radiation group among childhood cancer patients may only last a short time period. As opposed to the conventional results, in this setting, the Kolmogorov statistic is more powerful than the Cramér–von Mises statistic in detecting such differences; see Figures 1(b) and 1(c). Therefore, given a massive dataset, screening approaches that rely on a single screening criterion, such as the Cramér–von Mises or the Kolmogorov criterion, may not be able to capture different heterogeneous patterns, leading to false discovery as well as false non-discovery.

This paper proposes a class of L_q-norm learning criteria, which include the Cramér–von Mises and Kolmogorov statistics as two special cases with $q = 2$ and $q = \infty$, respectively. The embedded weight q provides a convenient means to detect predictors with short- or long-term impacts on survival. For example, a larger q, which yields statistics more like the Kolmogorov statistic, is useful for detecting predictors with short-term impact, while a smaller q, which generates statistics more like the Cramér–von Mises statistic, is more powerful in other, more general, settings. For a specific dataset, it is unclear which procedure is more likely to miss important predictors with unknown patterns of impact, including short- or long-term impacts, on outcomes. Our framework leads to a natural scheme to combine results obtained from different q in order to reduce false negatives, an aspect often overlooked by literature. The hybrid method proposed in the paper presents a possible path to conduct data-driven integration of different screening procedures, the utility of which is verified theoretically and numerically. In addition, our method is valid without parametric assumptions

or other restrictive conditions that stipulate the dependence between the outcome and predictors and, hence, is applicable to a variety of survival models. Our method is invariant under univariate monotone transformations on survival time or covariates or both. This property is appealing as variable transformation is widely applied in the data processing stage. Finally, as the proposed screening statistic is a function of Kaplan–Meier estimators, its computation is straightforward and scalable for screening ultrahigh-dimensional data.

2. The L_q-norm Learning Criteria

Let $(\Omega, \mathcal{F}, P)$ be the probability space that underlies all the random variables mentioned in this paper, where Ω is the sample space, $\mathcal{F}$ is the σ-algebra and P is the probability measure. Suppose that we have n independent subjects with p covariates, where $p \gg n$. Let i and j index subjects and covariates, respectively. For example, $X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ denotes the covariate vector for subject i and X_{ij} denotes covariate j for subject i. Let T_i be the survival time and C_i be the potential censoring time. We observe that $Y_i = \min\{T_i, C_i\}$ and $\delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ is the indicator function. As a convention, we assume that T_i and C_i are independent given X_i. We further assume that (T_i, C_i, X_i) are independently and identically distributed. In particular, let (T_i, X_{ij}, X_i) $(i = 1, \ldots, n)$ be independent and identically distributed copies of (T, X_j, X), where $X = (X_1, \ldots, X_p)$.

Denote by $S(\cdot)$ the marginal survival function of T and by $S(t \mid X)$ the conditional survival function of T given X. We define the set of active covariates as

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

We assume that the cardinality of $\mathcal{M}$ is small relative to p, as it is not unreasonable to stipulate that only a small number of biomarkers are relevant to patients' survival in biomedical

studies.

The task is to identify $\mathcal{M}$ given the vast number of candidate variables, which can be of mixed types. We propose our method by first considering a categorical variable, say, X_j, with K_j categories such that $X_j \in \{1, 2, \ldots, K_j\}$. Later, we will extend the method to cover continuous covariates.

To proceed, we define the L_q-norm of $g(T)$, where g is a generic function, as

$$\|g(T)\|_q = \{E(|g(T)|^q)\}^{1/q} = \left\{-\int_0^\infty |g(t)|^q dS(t)\right\}^{1/q}, \tag{2.1}$$

where $q \geq 1$ and the last equality holds because $-dS(t) = f(t)dt$.

In order to quantify the relevance of covariate X_j to the survival time T, we compute $S(t \mid X_j)$, the conditional survival function within each category of X_j; afterwards, for every pair of X_j categories, say, $k_1 \neq k_2 \in \{1, \ldots, K_j\}$, we compute the L_q-norm of $S(T \mid X_j = k_1) - S(T \mid X_j = k_2)$, and take the maximum over all pairs of $(k_1, \ k_2)$. More explicitly,

$$\Psi_j^{(q)} = \max_{k_1, k_2 \in \{1, \ldots, K_j\}} \|S(T \mid X_j = k_1) - S(T \mid X_j = k_2)\|_q. \tag{2.2}$$

The rationale of using (2.2) as the screening criterion is that it gauges the survival differences across different subpopulations of X_j, and $\Psi_j^{(q)} = 0$ if and only if T is independent of X_j. Hence, (2.2) measures the relevance of X_j to T. The L_q-norm criteria are general. When $q = 2$, (2.2) is the Cramér–von Mises statistic; when $q = \infty$, it becomes the Kolmogorov statistic:

$$\Psi_j^{(\infty)} = \max_{k_1, k_2 \in \{1, \ldots, K_j\}} \sup_t |S(t \mid X_j = k_1) - S(t \mid X_j = k_2)|. \tag{2.3}$$

Denote by $t_1 < t_2 < \cdots < t_d$ the ordered observed failure times and by $\hat{S}(t)$ the Kaplan–Meier estimate of $S(t)$, the marginal survival function of T at time t. Within each category of a categorical variable, say, using subsamples $\{i : X_{ij} = k\}$, we can compute the

Kaplan–Meier estimate $\hat{S}(t \mid X_j = k)$ of $S(t \mid X_j = k)$. Then, $\Psi_j^{(q)}$ can be estimated by

$$
\begin{aligned}
\widehat{\Psi}_j^{(q)} &= \max_{k_1, k_2 \in \{1, \ldots, K_j\}} \left\{ -\int_0^\infty \left| \hat{S}(t \mid X_j = k_1) - \hat{S}(t \mid X_j = k_2) \right|^q d\hat{S}(t) \right\}^{1/q} \\
&= \max_{k_1, k_2 \in \{1, \ldots, K_j\}} \left[\sum_{l=1}^d \left| \hat{S}(t_l \mid X_j = k_1) - \hat{S}(t_l \mid X_j = k_2) \right|^q \left\{ \hat{S}(t_{l-1}) - \hat{S}(t_l) \right\} \right]^{1/q},
\end{aligned}
\tag{2.4}
$$

where we set $t_0 = 0$ for notational convenience.

Finally, we select the active variables via

$$
\widehat{\mathcal{M}} = \left\{ j : \widehat{\Psi}_j^{(q)} > cn^{-v}, \; j = 1, \ldots, p \right\},
\tag{2.5}
$$

where c and v are constants for pre-determined thresholds defined in Condition 1 in Section 3. As the screening criterion is L_q-norm based, this procedure is termed L_q-norm learning.

The empirical version of $\Psi_j^{(q)}$ in (2.2) is difficult to evaluate when X_j takes infinite values. However, we can find an approximation of $\Psi_j^{(q)}$ by slicing X_j. Without loss of generality, we assume that the support of X_j is the real line $\mathbb{R}$. Let $\tilde{X}_j = k$ if $X_j \in [\hat{Q}_{j(k-1)}, \hat{Q}_{j(k)})$, where $\hat{Q}_{j(k)}$ is the $k/K_j \times 100$th percentile of the empirical distribution of X_j. For notational convenience, we set $\hat{Q}_{j(0)} = -\infty$ and $\hat{Q}_{j(K_j)} = \infty$. We refer to each $[\hat{Q}_{j(k-1)}, \hat{Q}_{j(k)})$ as a slice.

Suppose there are N different ways of slicing a continuous covariate X_j, denoted by Λ_{ju}, $u = 1, \ldots, N$ with each slicing Λ_{ju} containing K_{ju} intervals, i.e.,

$$
\Lambda_{ju} = \left\{ [\hat{Q}_{ju(k-1)}, \hat{Q}_{ju(k)}) : k = 1, \ldots, K_{ju} \text{ and } \bigcup_{k=1}^{K_{ju}} [\hat{Q}_{ju(k-1)}, \hat{Q}_{ju(k)}) = \mathbb{R} \right\}.
$$

We then replace X_j with its discretized version $\tilde{X}_{ju}$ under each Λ_{ju}, i.e., $\tilde{X}_{ju} = k$ if $X_j \in [\hat{Q}_{ju(k-1)}, \hat{Q}_{ju(k)})$. To ensure there are enough samples within each slice for all slicing schemes, one may take $K_{ju} = 3, \ldots, [\log(n)]$, which gives $N = [\log(n) - 2]$ slicing schemes.

Now let $\Psi^{(q)}_{j,\Lambda_{juo}} = \max_{k_1,k_2 \in \{1,\ldots,K_j\}} \|S(t \mid \tilde{X}_{ju} = k_1) - S(t \mid \tilde{X}_{ju} = k_2)\|_q$ be the L_q-norm learning statistic corresponding to the slicing scheme of Λ_{ju} for a continuous covariate j. After slicing, X_j is independent of T if and only if $\Psi^{(q)}_{j,\Lambda_{juo}} = 0$ for all possible choices of Λ_{ju}; see Lemma 1 of Mai and Zou (2015). In addition, although $\Psi^{(q)}_{j,\Lambda_{juo}}$ is used as a surrogate of $\Psi^{(q)}_j$, Lemma 2 of Mai and Zou (2015) shows that $\Psi^{(q)}_{j,\Lambda_{juo}}$ could be a better measure for variable screening than $\Psi^{(q)}_j$.

Finally, we combine the information from all Λ_{ju} using the fused L_q-norm learning statistic

$$\widetilde{\Psi}^{(q)}_j = \sum_{u=1}^N \widehat{\Psi}^{(q)}_{j,\Lambda_{ju}}, \tag{2.6}$$

where

$$\widehat{\Psi}^{(q)}_{j,\Lambda_{ju}} = \max_{k_1,k_2 \in \{1,\ldots,K_{ju}\}} \left[\sum_{l=1}^d \left| \hat{S}(t_l \mid \tilde{X}_{ju} = k_1) - \hat{S}(t_l \mid \tilde{X}_{ju} = k_2) \right|^q \left\{ \hat{S}(t_{l-1}) - \hat{S}(t_l) \right\} \right]^{1/q}, \tag{2.7}$$

leading to the following screening criterion:

$$\widetilde{\mathcal{M}} = \left\{ j : \widetilde{\Psi}^{(q)}_j > \tilde{c}n^{-\tilde{v}}, \ j = 1,\ldots,p \right\}, \tag{2.8}$$

where $\tilde{c}$ and $\tilde{v}$ are two positive constants. Due to favorable numerical experiments, we opt to use the fused method or to use (2.8) as the screening criterion when X_j is continuous.

3. Sure Screening Properties

We establish the sure screening property of the proposed screening method. The following regularity conditions are needed.

Condition 1. *For any $q \geq 1$, there exist $c > 0$ and $v \in [0, 1/2)$ such that $\min_{j\in\mathcal{M}} \Psi^{(q)}_j \geq 2cn^{-v}$ where c and v are the same as those in (2.5).*

Condition 2. *There exist $c_0 > 0$ and $\kappa \in [0, (1-2v)/3)$ such that $K = \max_{1 \leq j \leq p} K_j \leq c_0 n^\kappa$*

for any $n \geq 1$.

Condition 1 stipulates that the minimal signal in the active set should be strong enough. Such an assumption is standard in feature screening literature; see, for example, Condition 3 in Fan and Lv (2008), and Condition C2 in Li et al. (2012); He et al. (2013); Cui et al. (2015); Ni and Fang (2016). Moreover, when the censoring rate is 0 and $q = \infty$, the proposed method includes the Kolmogorov filter in Mai and Zou (2015) as a special case. Indeed, Condition 1 is weaker than condition (C1) of Mai and Zou (2015) and can be satisfied even when the active set is correlated with the inactive set.

Condition 2 allows the number of categories for each covariate to diverge with a certain order. A similar assumption was also made in Condition C3 of Ni and Fang (2016).

Theorem 1. *When all covariates are categorical, for any $q \geq 1$, there exist constants $c_1 > 0$, $c_2 > 0$, κ and v under Conditions 1–2, for a sufficiently large n, we have that*

$$P(\mathcal{M} \subset \widehat{\mathcal{M}}) \geq 1 - c_2 p \exp(-c_1 n^{1-3\kappa-2v} + \kappa \log n),$$

Hence, when $\log p = c_2 n^\alpha$ with $\alpha \in [0, 1-3\kappa-2v)$, L_q-norm learning has the sure screening property.

We next consider when X_j is continuous for some j. We denote by $f_{X_j}(x)$ the probability density of X_j, and replace Condition 2 with the following condition.

Condition 3. *Suppose that $f_{X_j}(x)$ is continuous and bounded on the support of X_j. There exist $c_3 > 0$ and $\rho \in [0, (1 - 2v - 3\kappa)/2)$, such that $\min_{1 \leq k \leq K_j - 1} f_{X_j}(Q_{j(k)}) \geq c_3 n^{-\rho}$.*

This condition implies that the density values among all the slicing points have a lower bound in the order of $n^{-\rho}$, ensuring that there are enough samples within each slice of X_j.

Theorem 2. *When covariates include both continuous and categorical types, for any $q \geq 1$,*

there exist constants $c_3 > 0$, $c_4 > 0$, κ v and ρ under Conditions 1 and 3, for n sufficiently

large, we have that

$$P(\mathcal{M} \subset \widehat{\mathcal{M}}) \geq 1 - c_4 p \exp(-c_3 n^{1-3\kappa-2v-2\rho} + \kappa \log n).$$

Hence, when $\log p = c_4 n^\alpha$ with $\alpha \in [0, 1 - 3\kappa - 2v - 2\rho)$, L_q-norm learning has the sure

screening property.

Fused L_q-norm learning requires additional notations and regularity conditions. Let Λ_{juo}

be the partition using the theoretical $k/K_{ju} \times 100$th percentile of X_j, and denote by $Q_{ju(k)}$

($k = 0, \ldots, K_{ju}$) the slicing points. Denote the true value of the L_q-norm learning statistic

for the partition Λ_{juo} by $\Psi_{j,\Lambda_{juo}}^{(q)}$ and let $\Psi_{jo}^{(q)} = \sum_{u=1}^{N} \Psi_{j,\Lambda_{juo}}^{(q)}$. The following Conditions 4–5

are modified from Conditions 1 and 3.

Condition 4. *For any $q \geq 1$, there exist constants $\tilde{c} > 0$ and $\tilde{v} \in [0, 1/2)$ such that*

$\min_{j \in \mathcal{M}} \Psi_{jo}^{(q)} \geq 2\tilde{c} n^{-\tilde{v}}$, *where $\tilde{c}$ and $\tilde{v}$ are the same as those in (2.8).*

Condition 5. *Suppose that $f_{X_j}(x)$ is bounded and continuous with respect to x. There exist*

constants $\tilde{c}_0 > 0$ and $\tilde{\kappa} \in [0, (1-2\tilde{v})/3)$ such that $\tilde{K} = \max_{1 \leq j \leq p, 1 \leq u \leq N} K_{ju} \leq \tilde{c}_0 n^{\tilde{\kappa}}$. There

exist constants $\tilde{c}_1 > 0$ and $\tilde{\rho} \in [0, (1-2\tilde{v}-3\tilde{\kappa})/2)$ such that $\min_{1 \leq k \leq K_{ju}-1, 1 \leq u \leq N} f_{X_j}(Q_{ju(k)}) \geq$

$\tilde{c}_1 n^{-\tilde{\rho}}$.

Theorem 3. *When covariates include both continuous and categorical types, for any $q \geq 1$,*

there exist $\tilde{c}_2 > 0$, $\tilde{c}_3 > 0$, $\tilde{\kappa}$, $\tilde{v}$ and $\tilde{\rho}$ under Conditions 4–5, for n sufficiently large, we have

that

$$P(\mathcal{M} \subset \widetilde{\mathcal{M}}) \geq 1 - \tilde{c}_3 p \log n \exp\{(-\tilde{c}_2 n^{1-3\tilde{\kappa}-2\tilde{v}-2\tilde{\rho}}/\log n) + \tilde{\kappa} \log n\}.$$

When $\log p = \tilde{c}_3(n^{\tilde{\alpha}}/\log n)$ *and* $\alpha \in [0, 1 - 3\tilde{\kappa} - 2\tilde{v} - 2\tilde{\rho})$, *fused* L_q-*norm learning has the sure screening property.*

4.   Hybrid L_q-norm learning

The performance of the L_q-norm learning depends on q, with an unknown best q for any given dataset. Thus, instead of solely relying on a specific q, we propose to combine the L_q-norm learning results obtained from various q, and show that it has desirable theoretical properties.

Suppose that we perform screening based on various q, say, $1 \le q_1 < \cdots < q_L < \infty$. We define hybrid L_q-norm learning as

$$\widetilde{\mathcal{M}}_h = \bigcup_{l=1}^{L} \widetilde{\mathcal{M}}^{(q_l)}, \tag{4.1}$$

where $\widetilde{\mathcal{M}}^{(q_l)} = \left\{ j : \widetilde{\Psi}_j^{(q_l)} > \tilde{c}_h n^{-v_l}, \ j = 1, \ldots, p \right\}$, v_l is a positive constant that depends on q_l, and $\tilde{c}_h$ is a positive constant not depending on l.

In principle, the range of q_l should be wide enough and covers the Cramér–von Mises and Kolmogorov statistics. One possible choice that may satisfy this principle is the Fibonacci numbers, with every number in the sequence (after the first two) being the sum of the two preceding numbers. That is, $q_l = 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, \ldots$. Since in our numerical experience the L_q-norm statistic, when $q > 30$, is very close to the Kolmogorov statistic (corresponding to $q = \infty$), we may consider a sub-series of Fibonacci with the maximum number being 89 as in our later simulation studies.

To show the sure screening property of hybrid L_q-norm learning, we assume the following regularity conditions:

Condition 6. *There exist $q_l \geq 1$, $\tilde{c}_h > 0$ and $v_l \in [0, 1/2)$ such that $\min_{j \in \mathcal{M}} \Psi_{jo}^{(q_l)} \geq \tilde{c}_h n^{-v_l}$, where $\tilde{c}_h$ and v_l are the same as those in (4.1). There exist constants $c_{0,l} > 0$ and $\kappa_l \in [0, (1 - 2v_l)/3)$ such that $\max_{1 \leq j \leq p, 1 \leq u \leq N} K_{ju} \leq c_{0,l} n^{\kappa_l}$. There exist constants $c_{1,l} > 0$ and $\rho_l \in [0, (1 - 2v_l - 3\kappa_l)/2)$ such that $\min_{1 \leq k \leq K_{ju}-1, 1 \leq u \leq N} f_{X_j}(Q_{ju(k)}) \geq c_{1,l} n^{-\rho_l}$.*

Theorem 4. *When covariates include both continuous and categorical types, there exist constants $q_l \geq 1$, $c_{2,l} > 0$, $c_{3,l} > 0$, κ_l, v_l and ρ_l under Condition 6, for n sufficiently large, we have that*

$$P(\mathcal{M} \subset \widetilde{\mathcal{M}}_h) \geq 1 - c_{3,l} p \log n \exp\{(-c_{2,l} n^{1-3\kappa_l - 2v_l - 2\rho_l} / \log n) + \kappa_l \log n\},$$

when $\log p = c_{3,l}(n^\alpha / \log n)$ with $\alpha \in [0, 1 - 3\kappa_l - 2v_l - 2\rho_l)$, hybrid L_q-norm learning has the sure screening property.

Hybrid L_q-norm learning allows covariates chosen by any q_l $(l = 1, \ldots, L)$ to be included in the selected active set, which guarantees the recovery of the true active set or reduces the false negatives to the extent possible. This may fit the overarching goal of variable screening.

5. Simulation Studies

We performed simulations to compare the finite sample performance of the proposed L_q-learning and competing methods, such as principled sure independence screening (Zhao and Li, 2012), censored rank independence screening (Song et al., 2014), independent screening for single-index hazard rate models (Gorst-Rasmussen and Scheike, 2013), quantile adaptive screening (He et al., 2013), and conditional sure independence screening (Hong et al., 2018). For the proposed L_q-norm learning approach, we considered $q = \{1, 2, 5, 13, 89, \infty\}$ over its

domain. Our preliminary analysis shows that large values of q (≥ 30) give results similar to those of the Kolmogorov statistic, while small or moderate values of q (< 10) resemble the Cramér–von Mises criterion. In practice, the values of q would depend on users' research goals: if the focus is on finding predictors with long-term impact, q should be chosen to be close to 1, whereas if the focus is on finding predictors with short-term effect, large values of q would be preferable.

Binary, categorical, and continuous variables were considered in our simulations. The censoring times C_i were independently generated from a uniform distribution $U[0, c_0]$, with c_0 chosen to give censoring proportions of approximately 20% and 40%.

Example 1. The underlying random vector $x^* = (x_1^*, \ldots, x_p^*)$ was generated from a multivariate normal distribution with a mean vector of 0 and an exchangeable correlation structure with an equal correlation of 0.5. For each j, x_j^* was further dichotomized by its median value and the obtained binary variable $X_j = 0$ if x_j^* is in the lower half, and $X_j = 1$ otherwise. The survival times were generated from an accelerated failure time model with a baseline hazard function $h_0(t) = 0.1(t - 2)^2$, i.e.,

$$h(t \mid X) = h_0\{\exp(\beta^{\mathrm{T}} X)t\} \exp(\beta^{\mathrm{T}} X),$$

where $\beta = (-0.5, -0.5, -0.5, -0.5, -0.5, 0_{p-5}^{\mathrm{T}})^{\mathrm{T}}$. A similar model was considered by Zhang and Peng (2009).

Example 2. The underlying random vector x^* was generated as in Example 1. 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. The survival times were generated from

the proportional hazards model,

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

where $\beta = (1.2, 0, 1, 0, 0.8, 0, 1, 0_{p-5}^{\mathrm{T}})^{\mathrm{T}}$.

Example 3. The survival times were generated with the following hazard,

$$h(t \mid X) = 2t(|X_1| + |X_2|),$$

where all 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 4. For each j, the observed discrete covariate X_j was generated as in Example 1. The survival times were generated with the following hazard,

$$h(t \mid X) = \begin{cases} 1 + 4(X_1 + X_2), & \text{for } t \in (1.3, 1.9] \\ 1, & \text{for } t \in (0, 1.3] \cup (1.9, \infty). \end{cases}$$

For each example, 500 simulated datasets were generated. We considered $n = 400$ and $n = 600$ to explore how the performance of the proposed nonparametric method improved with the sample size. The performance was assessed using the criteria of minimum model size (MMS), probability to include the true model (PIT), and true positive rate (TPR). In Examples 1–4, X_1 was used as the true conditioning set for the conditional screening method (Hong et al., 2018).

Tables 1-4 demonstrate that the proposed L_q-norm learning achieves reasonable minimum model size, probability to include the true model, and true positive rate in the considered scenarios. Its performance improves as the sample size increases, which may not be

true for competing methods. When variables are categorical as in Example 2, the results for the competing methods were poor since these methods were not developed for screening categorical variables. In particular, the B-spline based quantile adaptive method (He et al., 2013) is not applicable to binary covariates. On the other hand, in Example 3, when the marginal correlation between each active variable and the survival time is 0, all the competing methods, including the conditional screening method that assumed one active variable was known, had difficulty in identifying active variables. As we conjectured, the optimal q in L_q-norm learning tends to be data-specific. For example, the minimum model size decreases as q decreases in Example 2, whereas it decreases as q increases in Example 4.

To check the invariance property of the proposed method, we used $X^{1/3}$, in lieu of X, and the log transformed observed survival times in Example 3. The transformed data yielded the same L_q-norm learning statistic, supporting the invariance property of the method. Finally, Table S1 of the Supplementary Material shows that the proposed method is not heavily impacted by the violation of the independent censoring assumption.

6. Analysis of Multiple Myeloma Data

Multiple myeloma is a progressive blood disease, characterized by excessive numbers of abnormal plasma cells in the bone marrow and overproduction of intact monoclonal immunoglobulin. Myeloma patients' survival ranges from a few months to more than 10 years, even within the same stage of cancer. Gene expression profiling offers an effective way to predict the survival of patients with newly diagnosed multiple myeloma. We applied the proposed method to study a multiple myeloma trial, which was designed to identify gene signatures that are relevant to patients' survival (Avet-Loiseau et al., 2009). The study had

independent and comparable training and testing sets. The training dataset had 133 patients with a 56% censoring rate, an average age of 55.2 years, and an average follow-up of 44.2 months. Of them, 45% were female. In the testing dataset, there were 37 patients with a 51% censoring rate, a mean age of 56.2 years, and a mean follow-up of 40.8 months. Among this group of patients, 43% were female. Combining both training and testing samples, the study consisted of 170 patients, each with measurements of 44,280 gene expressions.

As the number of gene expressions overwhelmed the sample size, we first applied the proposed L_q-norm learning, as well as several competing methods, to the training dataset with $n_1 = 133$ to screen out irrelevant genes, and reduced the dimension from $p = 44,280$ to $d = [n_1 / \log(n_1)] = 27$.

Since gene expression levels are continuous, we used the fused approach introduced in Section 2.1. That is, we considered slicing schemes, $\Lambda_{j1}, \Lambda_{j2}, \Lambda_{j3}$, which contain $3, 4, 5$ ($= [\log(133)]$) intervals, respectively. Then, we combined the information from all Λ_{ju}, $u = 1, 2, 3$, using the fused L_q-norm learning statistic in (2.6).

Table 5 reports the numbers of overlapping genes selected by the different methods, showing that the variables selected by L_q-norm learning with different q values did differ and the proposed method helped choose novel genes that were not identified by the existing methods.

We next examined the performance of various methods using the random survival forests approach, which is an extension of a random forests model to right-censored survival data and can be implemented using the R package `randomSurvivalForest` (Ishwaran and Kogalur, 2007).

First, we randomly generated 10 training/testing splits from the full data set of 170 pa-

tients, with 133 in the training set and 37 in the testing set. In each training dataset, we selected top 27 genes by each method and fitted a random survival forests model. When fitting the random forests, a total of 100 trees were generated for each training dataset. Then the fitted "forests" were applied to each testing dataset, for which a c-statistic was computed. The overall c-statistic was the average of c-statistics across all the splits.

Finally, for each method, the average of the c-statistics from all 10 testing datasets was listed in Table 6. In general, our method improved c-statistics, even though the improvement may not reach statistical significance.

To evaluate the impact of choosing different numbers of top genes, in the Supplementary Material, we repeated the same investigation by choosing the top 133 genes selected by each method, and presented the results in Tables S2-S3.

To address the important biological question of which genes are relevant to multiple myeloma patients' survival, we applied hybrid L_q-norm learning to the whole dataset and chose the top 27 genes. Based on those genes, we further fitted a random survival forests model and assessed the top 10 genes based on their contributions to the model. Table 7 lists these genes, which have already been recognized in cancer literature. In particular, probes 213901_x_at, 206150_at, and 206662_at have been known to be clinically significant in multiple myeloma. Moreover, our method also highlights some possible novel candidates for multiple myeloma. For example, though probes 205689_at, 39650_s_at, 218058_at, 216860_s_at, 206267_s_at, and 227894_at have not been identified in multiple myeloma literature, they have been linked to a variety of other cancers, including prostate, lung, breast, head and neck cancers. Their roles in multiple myeloma are worth investigating.

7. Conclusion

This paper proposes a new class of model-free L_q-norm learning approaches for screening ultrahigh-dimensional survival data. The important problem of how to combine results from different screening procedures remains open (Liu et al., 2015). To our knowledge, this paper registers the first attempt to combine the screening results with different q via (4.1). The intuition is that the hybrid learning retains covariates chosen by any of the considered screening procedures, which may help reduce the false negatives to the extent possible, a desirable property of screening procedures. Our framework can facilitate fusion of screening procedures in other ways, such as $\widetilde{\mathcal{M}}_h^* = \bigcap_{l=1}^L \widetilde{\mathcal{M}}^{(q_l)}$ and $\widetilde{\mathcal{M}}_h^{**} = \left\{ j : \widetilde{\Psi}_j^{(q_l)*} > cn^{-v}, \ j = 1, \ldots, p \right\}$, where $\widetilde{\Psi}_j^{(q_l)*} = (\widetilde{\Psi}_j^{(q_l)} - \min_{1 \le l \le L} \widetilde{\Psi}_j^{(q_l)})/(\max_{1 \le l \le L} \widetilde{\Psi}_j^{(q_l)} - \min_{1 \le l \le L} \widetilde{\Psi}_j^{(q_l)})$. Here, $\widetilde{\mathcal{M}}_h^*$ includes common covariates that are selected by all q_l ($l = 1, \ldots, L$). This method can guarantee exclusion of unimportant covariates to the greatest extent, but this rather restrictive criterion may lead to many false negatives, which may not be ideal for knowledge discovery in the exploratory phase. On the other hand, $\widetilde{\mathcal{M}}_h^{**}$ may be a compromise between $\widetilde{\mathcal{M}}_h$ and $\widetilde{\mathcal{M}}_h^*$. Through normalization by rescaling between 0 and 1, it makes screening statistics across q comparable.

We envision that this hybrid framework can help address different needs. When the priority is on controlling the false negatives, we recommend $\widetilde{\mathcal{M}}_h$; when the priority is on controlling false positives, we recommend $\widetilde{\mathcal{M}}_h^*$; when one needs to control both false negatives and false positives, we recommend $\widetilde{\mathcal{M}}_h^{**}$.

A more detailed investigation of the strategy in a broader context or a search for more efficient hybrid algorithms, though beyond the scope of this paper, is worth pursuing and will be reported elsewhere.

Supplementary Material

The online Supplementary Material contains theoretical results, additional simulation studies, and data analysis results.

References

Andrews, J., W. Kennette, J. Pilon, A. Hodgson, A. B. Tuck, A. F. Chambers, and D. I. Rodenhiser (2010). Multi-platform whole-genome microarray analyses refine the epigenetic signature of breast cancer metastasis with gene expression and copy number. *PLOS ONE 5*(1), 1–17.

Arnold, T. B. and J. W. Emerson (2011). Nonparametric goodness-of-fit tests for discrete null distributions. *R Journal 3*(2), 34–39.

Avet-Loiseau, H., C. Li, F. Magrangeas, W. Gouraud, C. Charbonnel, J.-L. Harousseau, M. Attal, G. Marit, C. Mathiot, T. Facon, P. Moreau, K. C. anderson, L. Campion, N. C. Munshi, and S. Minvielle (2009). Prognostic significance of copy-number alterations in multiple myeloma. *Journal of Clinical Oncology 27*(27), 4585–4590.

Bayne, R. A. L., T. Forster, S. T. G. Burgess, M. Craigon, M. J. Walton, D. T. Baird, P. Ghazal, and R. A. Anderson (2008). Molecular profiling of the human testis reveals stringent pathway-specific regulation of RNA expression following gonadotropin suppression and progestogen treatment. *Journal of Andrology 29*(4), 389–403.

Chiu, S. N. and K. I. Liu (2009). Generalized Cramér–von Mises goodness-of-fit tests for multivariate distributions. *Computational Statistics & Data Analysis 53*(11), 3817–3834.

Conover, W. J. and W. J. Conover (1980). *Practical Nonparametric Statistics*. Wiley: New York.

Cui, H., R. Li, and W. Zhong (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association 110*(510), 630–641.

Del Giudice, I., M. Messina, S. Chiaretti, S. Santangelo, S. Tavolaro, M. S. De Propris, M. Nanni, E. Pescarmona, F. Mancini, A. Pulsoni, M. Martelli, A. Di Rocco, E. Finolezzi, F. Paoloni, F. R. Mauro, A. Cuneo, A. Guarini, and R. Foá (2012). Behind the scenes of non-nodal MCL: Downmodulation of genes involved in actin cytoskeleton organization, cell projection, cell adhesion, tumour invasion, TP53 pathway and mutated status of immunoglobulin heavy chain genes. *British Journal of Haematology 156*(5), 601–611.

Fan, J., Y. Feng, and Y. Wu (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 J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of Royal Statistical Society B 70*(5), 849–911.

García-Piñeres, A. J., A. Hildesheim, L. Dodd, T. J. Kemp, J. Yang, B. Fullmer, C. Harro, D. R. Lowy, R. A. Lempicki, and L. A. Pinto (2009). Gene expression patterns induced by HPV-16 L1 virus-like particles in leukocytes from vaccine recipients. *The Journal of Immunology 182*(3), 1706–1729.

Gorst-Rasmussen, A. and T. Scheike (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of the Royal Statistical Society: Series B 75*(2), 217–245.

Gutiérrez, N. C., E. M. Ocio, J. de Las Rivas, P. Maiso, M. Delgado, E. Fermiñán, M. J. Arcos, M. L. Sánchez, J. M. Hernández, and J. F. San Miguel (2007). Gene expression profiling of B lymphocytes and plasma cells from Waldenstrom's macroglobulinemia: Comparison with expression patterns of the same cell counterparts from chronic lympho-

cytic leukemia, multiple myeloma and normal individuals. *Leukemia 21*(3), 541–549.

He, X., L. Wang, and H. G. Hong (2013). Quantile-adaptive model-free variable screening for high-dimensional heteroge-

neous data. *The Annals of Statistics 41(1)*, 342–369.

Hong, H. G., J. Kang, and Y. Li (2018). Conditional screening for ultra-high dimensional covariates with survival outcomes.

Lifetime data analysis 24(1), 45–71.

Huang, C.-C., S. Gadd, N. Breslow, C. Cutcliffe, S. T. Sredni, I. B. Helenowski, J. S. Dome, P. E. Grundy, D. M. Green,

M. K. Fritsch, and E. J. Perlman (2009). Predicting relapse in favorable histology wilms tumor using gene expression

analysis: A report from the renal tumor committee of the children's oncology group. *Clinical Cancer Research 15*(5),

1770–1778.

Ishwaran, H. and U. Kogalur (2007). Random survival forests for R. *Rnews 7*(2), 25–31.

Kassambara, A., D. Hose, J. Moreaux, T. Rème, J. Torrent, A. Kassambara, D. Hose, J. Moreaux, T. Rme, J. Torrent,

J. Rossi, H. Goldschmidt, and B. Klein (2012). Identification of pluripotent and adult stem cell genes unrelated to cell

cycle and associated with poor prognosis in multiple myeloma. *PLOS ONE 7*(7), 1–9.

Koziol, J. A. and S. B. Green (1976). A Cramér–von Mises statistic for randomly censored data. *Biometrika 63*(3), 465–474.

Li, R., W. Zhong, and L. Zhu (2012). Feature screening via distance correlation learning. *Journal of the American Statistical*

Association 107(499), 1129–1139.

Li, Y. and J. Feng (2005). A nonparametric comparison of conditional distributions with non-negligible cure fractions.

Lifetime Data Analysis 11(3), 367–387.

Liu, J., W. Zhong, and R. Li (2015). A selective overview of feature screening for ultrahigh-dimensional data. *Science*

China Mathematics 58(10), 2033–2054.

Lu, X., E. Mu, Y. Wei, S. Riethdorf, Q. Yang, M. Yuan, J. Yan, Y. Hua, B. J. Tiede, X. Lu, B. G. Haffty, K. Pantel,

J. Massagué, and Y. Kang (2011). VCAM-1 promotes osteolytic expansion of indolent bone micrometastasis of breast

cancer by engaging $\alpha 4\beta 1$-positive osteoclast progenitors. *Cancer Cell 20*(6), 701–714.

Mai, Q. and H. Zou (2015). The fused Kolmogorov filter: A nonparametric model-free screening method. *The Annals of Statistics 43*(4), 1471–1497.

Mulligan, G., C. Mitsiades, B. Bryant, F. Zhan, W. Chng, S. Roels, E. Koenig, A. Fergus, Y. Huang, P. Richardson, and W. Trepicchio (2007). Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib. *Blood 109*, 3177–3188.

Ni, L. and F. Fang (2016). Entropy-based model-free feature screening for ultrahigh-dimensional multiclass classification. *Journal of Nonparametric Statistics 28*(3), 515–530.

Rao, X.-M., X. Zheng, S. Waigel, W. Zacharias, K. M. McMasters, and H. S. Zhou (2006). Gene expression profiles of normal human lung cells affected by adenoviral E1B. *Virology 350*(2), 418–428.

Razali, N. M., Y. B. Wah, et al. (2011). Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *Journal of statistical modeling and analytics 2*(1), 21–33.

Rodríguez-Caballero, A., A. C. García-Montero, P. Bárcena, J. Almeida, F. Ruiz-Cabello, M. D. Tabernero, P. Garrido, S. Muñoz-Criado, Y. Sandberg, A. W. Langerak, M. González, A. Balanzategui, and A. Orfao (2008). Expanded cells in monoclonal TCR-$\alpha\beta^+$/CD4$^+$/NKa$^+$/CD8$^{-/+\mathrm{dim}}$ T-LGL lymphocytosis recognize hCMV antigens. *Blood 112*(12), 4609–4616.

Santin, A. D., F. Zhan, E. Bignotti, E. R. Siegel, S. Cané, S. Bellone, M. Palmieri, S. Anfossi, M. Thomas, A. Burnett, H. H. Kay, J. J. Roman, T. J. O'Brien, E. Tian, M. J. Cannon, J. Shaughnessy Jr., and S. Pecorelli (2005). Gene expression profiles of primary HPV16- and HPV18-infected early stage cervical cancers and normal cervical epithelium: Identification of novel candidate molecular markers for cervical cancer diagnosis and therapy. *Virology 331*(2), 269–291.

Schumacher, M. (1984). Two-sample tests of Cramér–von Mises-and Kolmogorov–Smirnov-type for randomly censored

data. *International Statistical Review/Revue Internationale de Statistique 52*(3), 263–281.

Song, R., W. Lu, S. Ma, and X. J. Jeng (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika 101*(4), 799–814.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics 25*(2), 613–641.

Tamura, R. N., D. E. Faries, and J. Feng (2000). Comparing time to onset of response in antidepressant clinical trials using the cure model and the Cramér–von Mises test. *Statistics in Medicine 19*(16), 2169–2184.

Tibshirani, R. J. (2009). Univariate shrinkage in the Cox model for high dimensional data. *Statistical Applications in Genetics and Molecular Biology 8*(1), Article21.

Woodruff, B. W. and A. H. Moore (1988). 7 application of goodness-of-fit tests in reliability. *Handbook of Statistics 7*, 113–120.

Yu, J. X., A. M. Sieuwerts, Y. Zhang, J. W. Martens, M. Smid, J. G. Klijn, Y. Wang, and J. A. Foekens (2007). Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer 7*(1), 182.

Zhang, J. and Y. Peng (2009). Crossing hazard functions in common survival models. *Statistics & Probability Letters 79*, 2124–2130.

Zhao, S. D. and Y. Li (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of multivariate analysis 105*(1), 397–411.

Zhu, L., L. Li, R. Li, and L. Zhu (2012). Model-free feature screening for ultrahigh dimensional data. *Journal of the American Statistical Association 106*(496), 1464–1475.

Department of Statistics and Probability, Michigan State University, East Lansing, Michigan 48823, USA

E-mail: hhong@msu.edu

Center of Statistical Research, Southwestern University of Finance and Economics, China

E-mail: chenxuerong@swufe.edu.cn

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA

E-mail: jiankang@umich.edu

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA

E-mail: yili@umich.edu

Table 1: Performance of different variable screening methods for Examples 1–2 with 20% CR.

| | Example 1 | | | | | | Example 2 | | | | | |
| | $n = 400$ | | | $n = 600$ | | | $n = 400$ | | | $n = 600$ | | |
	MMS	TPR	PIT	MMS	TPR	PIT	MMS	TPR	PIT	MMS	TPR	PIT
PSIS	6	1.00	0.99	5	1.00	1.00	910	0.07	0.00	851	0.09	0.00
CRIS	1000	0.00	0.00	1000	0.00	0.00	841	0.06	0.00	829	0.11	0.00
FAST	5	1.00	0.99	5	1.00	1.00	923	0.05	0.00	871	0.06	0.00
QA	-	-	-	-	-	-	39	0.05	0.00	16	0.06	0.00
CS	5	1.00	1.00	5	1.00	1.00	799	0.29	0.00	788	0.30	0.00
L_1	9	0.99	0.97	5	1.00	1.00	9	0.97	0.88	4	0.99	0.97
L_2	7	1.00	0.99	5	1.00	1.00	10	0.97	0.88	4	0.99	0.97
L_5	7	1.00	0.98	5	1.00	1.00	13	0.97	0.87	4	0.99	0.97
L_{13}	7	0.99	0.97	5	1.00	1.00	16	0.96	0.83	4	0.99	0.96
L_{89}	9	0.99	0.95	5	1.00	1.00	11	0.95	0.79	5	0.99	0.95
L_∞	9	0.99	0.94	5	1.00	1.00	11	0.94	0.78	6	0.99	0.95
Hybrid	9	0.99	0.97	5	1.00	1.00	12	0.95	0.81	5	0.99	0.96

Table 2: Performance of different variable screening methods for Examples 3–4 with 20% CR.

| | Example 3 | | | | | | Example 4 | | | | | |
| | $n = 400$ | | | $n = 600$ | | | $n = 400$ | | | $n = 600$ | | |
	MMS	TPR	PIT	MMS	TPR	PIT	MMS	TPR	PIT	MMS	TPR	PIT
PSIS	730	0.05	0.00	737	0.06	0.01	110	0.56	0.36	54	0.79	0.66
CRIS	735	0.04	0.00	724	0.06	0.00	455	0.18	0.05	458	0.34	0.12
FAST	722	0.06	0.00	727	0.07	0.01	97	0.61	0.43	47	0.79	0.66
QA	8	0.95	0.90	4	0.01	0.00	-	-	-	-	-	-
CS	481	0.55	0.09	440	0.56	0.11	52	0.78	0.55	20	0.79	0.66
L_1	2	1.00	1.00	2	1.00	1.00	259	0.23	0.07	190	0.42	0.21
L_2	2	1.00	1.00	2	1.00	1.00	121	0.42	0.22	60	0.80	0.66
L_5	2	1.00	1.00	2	1.00	1.00	38	0.81	0.72	10	0.98	0.98
L_{13}	2	1.00	1.00	2	1.00	1.00	17	0.85	0.78	4	0.99	0.99
L_{89}	2	1.00	1.00	2	1.00	1.00	11	0.87	0.80	3	0.99	0.99
L_∞	2	1.00	1.00	2	1.00	1.00	11	0.88	0.82	3	0.99	0.99
Hybrid	2	1.00	1.00	2	1.00	1.00	16	0.86	0.77	5	0.99	0.98

Table 3: Performance of different variable screening methods for Examples 1–2 with 40% CR.

| | Example 1 | | | | | | Example 2 | | | | | |
| | $n = 400$ | | | $n = 600$ | | | $n = 400$ | | | $n = 600$ | | |
	MMS	TPR	PIT	MMS	TPR	PIT	MMS	TPR	PIT	MMS	TPR	PIT
PSIS	10	0.99	0.94	5	1.00	1.00	861	0.05	0.00	882	0.07	0.00
CRIS	1000	0.00	0.00	1000	0.00	0.00	862	0.06	0.00	879	0.06	0.00
FAST	10	0.99	0.95	5	1.00	1.00	873	0.02	0.00	889	0.04	0.00
QA	-	-	-	-	-	-	250	0.70	0.20	201	0.73	0.29
CS	5	1.00	0.99	5	1.00	1.00	823	0.27	0.00	786	0.28	0.00
L_1	14	0.97	0.88	5	1.00	1.00	4	1.00	0.99	4	1.00	1.00
L_2	10	0.99	0.93	5	1.00	1.00	4	1.00	0.99	4	1.00	1.00
L_5	10	0.99	0.93	5	1.00	1.00	4	1.00	0.99	4	1.00	1.00
L_{13}	11	0.98	0.91	5	1.00	1.00	5	1.00	0.99	4	1.00	1.00
L_{89}	15	0.97	0.86	5	1.00	1.00	7	1.00	0.99	4	1.00	0.99
L_∞	15	0.97	0.86	5	1.00	1.00	7	1.00	0.99	4	1.00	0.99
Hybrid	13	0.98	0.90	5	1.00	1.00	5	0.99	0.97	4	1.00	1.00

Table 4: Performance of different variable screening methods for Examples 3–4 with 40% CR.

| | Example 3 | | | | | | Example 4 | | | | | |
| | $n = 400$ | | | $n = 600$ | | | $n = 400$ | | | $n = 600$ | | |
	MMS	TPR	PIT	MMS	TPR	PIT	MMS	TPR	PIT	MMS	TPR	PIT
PSIS	711	0.07	0.01	743	0.06	0.01	532	0.15	0.03	475	0.24	0.08
CRIS	696	0.05	0.00	762	0.04	0.00	531	0.14	0.02	497	0.23	0.06
FAST	705	0.08	0.01	740	0.06	0.01	532	0.16	0.03	473	0.25	0.09
QA	56	0.74	0.55	69	0.01	0.00	-	-	-	-	-	-
CS	447	0.55	0.10	423	0.55	0.10	381	0.56	0.13	311	0.61	0.23
L_1	2	1.00	1.00	2	1.00	1.00	429	0.14	0.02	337	0.26	0.03
L_2	2	1.00	1.00	2	1.00	1.00	317	0.21	0.04	202	0.44	0.21
L_5	2	1.00	1.00	2	1.00	1.00	189	0.40	0.17	76	0.75	0.58
L_{13}	2	1.00	0.99	2	1.00	1.00	149	0.50	0.27	54	0.83	0.68
L_{89}	2	0.99	0.99	2	1.00	1.00	138	0.54	0.31	43	0.84	0.70
L_∞	2	0.99	0.99	2	1.00	1.00	132	0.54	0.30	46	0.84	0.71
Hybrid	2	1.00	1.00	2	1.00	1.00	174	0.45	0.21	56	0.81	0.65

Table 5: The numbers of overlapping genes among top 27 genes selected by various screening methods on the multiple myeloma training dataset.

	PSIS	CRIS	FAST	CS	QA	L_1	L_2	L_5	L_{13}	L_{89}	L_∞	Hybrid
PSIS	27	6	4	2	0	2	4	1	2	1	1	2
CRIS		27	1	3	0	4	5	3	2	0	0	2
FAST			27	2	0	1	2	0	0	0	0	0
CS				27	0	3	3	5	6	6	6	6
QA					27	0	0	0	0	0	0	0
L_1						27	22	14	8	5	4	12
L_2							27	17	11	5	4	13
L_5								27	20	15	13	20
L_{13}									27	21	20	21
L_{89}										27	23	17
L_∞											27	16
Hybrid												27

Table 6: Comparisons of the average c-statistics along with its 95% confidence interval based on 10 random testing datasets of multiple myeloma.

PSIS	CRIS	FAST	CS	QA	Hybrid
0.61 (0.53, 0.68)	0.59 (0.46, 0.72)	0.55 (0.44, 0.66)	0.59 (0.48, 0.70)	0.53 (0.40, 0.66)	0.63 (0.55, 0.72)

Table 7: The 10 most important genes selected by hybrid L_q-norm learning.

Probes	Description
213901_x_at	average expression differed by$>$ 2.5-fold comparing Adwt with Adhz60 infection (Rao et al., 2006); overexpressed in MMCs or in HMCLs compared to normal counterparts (Kassambara et al., 2012)
206150_at	significant in the apoptosis pathway in ER-positive tumors (Yu et al., 2007); genes exclusively deregulated in PC from MM but with a similar expression profile in WM-PC and NPC (Gutiérrez et al., 2007)
205689_at	concordantly differentially expressed within reported genetic regions of gain or loss in relapses in favorable histology Wilms' tumor (Huang et al., 2009)
39650_s_at	hypomethylated and increased in expression (Andrews et al., 2010)
218058_at	differentially expressed between the dormant SCP6 cell line and related non-metastatic or low-metastatic cell lines, and highly bonemetastatic PD cell lines (Lu et al., 2011)
206662_at	up-regulated genes expressed at least twofold higher in NCK compared with CVX (Santin et al., 2005); genes exclusively deregulated in PC from MM but with a similar expression profile in WM-PC and NPC (Gutiérrez et al., 2007)
216860_s_at	differentially expressed after vaccination (García-Piñeres et al., 2009)
206267_s_at	gene expression in monoclonal CD4 T-LGL cells significantly ($p <$ .006) changed after short-term in vitro hCMV stimulation (Rodríguez-Caballero et al., 2008)
207598_x_at	pathway/response to DNA damage (Del Giudice et al., 2012)
227894_at	genes showing expression profiles similar to genes identified as statistically significant (Bayne et al., 2008)

NOTE: The genes selected by hybrid L_q-norm learning were reordered based on variable importance ranking assessed by a random survival forests model.